UNIVERSIDAD NACIONAL DEL LITORAL

FACULTAD DE BIOQUÍMICA Y CIENCIAS BIOLÓGICAS

Tesis presentada para acceder al grado académico de Doctor en Ciencias Biológicas

"Desarrollo y aplicación de herramientas quimiométricas para resolución de muestras de origen biológico-químico"

Gabriel Siano

Director: Dr. Héctor Goicoechea

Cátedra de Química Analítica I- Laboratorio de Desarrollo Analítico y Quimiometría

-2013-

Agradecimientos

A Héctor Goicoechea, por respetar mi forma de trabajar, por la libertad de poder militar ideas, por dejarme ir por donde y cuando quisiera, y por los brazos abiertos. En su nombre, a la Facultad de Bioquímica y Ciencias Biológicas, a la UNL y al CONICET.

A todos mis compañeros de la Cátedra de Química Analítica I, por lo compartido y por los buenos momentos. A Julia y Pablito, para que (como yo) sientan el placer de ver sus nombres impresos en los inicios de un agradecimiento; a Luciana, por apreciar y compartir momentos reflexivos; a Gonzalo, por ser mi amigo más diferente; a Quela y Gabo, por su participación en el final. Mención especial para los miembros fundadores, visitantes extranjeros y demás investigadores de jerarquía alta e indemostrable que frecuentaban el antro laboral vilmente apodado como "La Kutxa" ("*..donde los malandras se empeñan revoleando los tinteros para que se cumpla mejor el divino propósito del Universo..*" Dolina). Las fuerzas del orden nos han desterrado, pero volveremos.

A Isidro Sánchez Pérez (jolines!), amigo, hermano y compañero de laburo y aventuras, por anular al "Alántico" en un instante cada vez que se lo recuerda. A María y Mariola, por la oportunidad y fundamentalmente por la sensación en la última cena. Al resto de mis queridos ibéricos, por la absoluta integración que me brindaron (esto excluye al oficial que me detuvo durante 3 horas en migraciones): Familia Sánchez Pérez, por tratarme como a un hijo, Isidro padre, gracias por sacarme del aeropuerto, Maruja, gracias por los *tuppers* de comida rotulados a mi nombre y según mis gustos; Paco y Menta (y Coso!), por la sabiduría, por el apoyo incondicional, por lo vivido en medio de camisetas, muñecas "no Barbie" y demás objetos de despedida, y por todo aquello que no se pueda escribir en los agradecimientos de una tesis doctoral; Familia Casado Martínez, por el trato en Almería y por los viajes en tren; María Docavo, por sus rezos aeroportuarios y por encargarse de mi vida en Madrid; a los Cenácolos y al resto en general.

A John Kalivas, por provocar la pérdida de mis referencias más profundas por primera vez y por haber propiciado el inicio de una transformación. Gracias, ahora aprecio más lo que veo cerca.

A mis concubinos/as de soltería doctoral, por enésimos momentos: Albano, Isidro, Mat y Sil, un lujo y un acierto haber vivido con cada uno de ustedes. Mención especial para el semental Samuel Bernardo, para el indomable "Tilo" Fresco y para el inexplicable Ramón Bru (y asociados), cuyas existencias no humanas han sido una grata compañía.

A Chango, Patri (y Felipe!), Bernardo y al recordado Giussepe, por sus presencias y sostenes durante el verano de la chiripiorca. Muchas gracias de corazón.

A los miembros del inolvidable Movimiento de Unión al *Tupper* (MUT), por el nefasto 19 a 2 en rectorado; a los del barrio Las Lomas, por cada vez que sonrieron; a los de la Cooperativa Axón, por cada idea que fracasó. Gracias por dejarme transitar junto a ustedes. Me enorgullece todo lo hecho.

A los caballeros de la "Noche de Hombres", por la veladas de filosofía, catarsis, Bichi y estimulación. Mención especial para el miembro permanente de menor antigüedad y mayor longevidad, cuyo rincón es el último bastión de soltería y cuyo nombre no osaría escribir en este texto.

A mis amigos/as de Sunchales, siempre presentes e independientes de las frecuencias.

A mi amigo Gabo, por sus años de okupa honorario, por la mecánica y los enchastres, por la búsqueda conjunta de cierta apertura mental, intelectual y espiritual, por el buscavidismo de vinilos viajeros (gracias Jezy y Rafita!), escuelas de Arte y demás yerbas. Bendita sea tu actitud MacGiverezca.

A Víctor Mantovani y Mauro Lucero, referentes, espíritus sabios y poderosos, siempre dispuestos a escuchar y aconsejar.

A Lili, Edgardo y a la planta, por sus llegadas oportunas durante la llegada más esperada y menos prevista.

A Ale, Ioia, Juano y Chiti, por la forma en que me adoptaron y por la gran ayuda que nos han brindado mientras esta tortuosa escritura terminaba.

A mi mamá Olinda y a mi papá Néstor, por haber estado y por seguir haciéndolo. A mi hermana Vanesa, porque de sus decisiones obtuve valentía, por su determinación y por optar ser lo que es.

A Vir, por el amor y todo lo que se decanta de éste. De ninguna manera esto hubiese sido posible sin tu apoyo. Fin de los cálculos, paso la posta. Te bogats.

a mi amada Violeta Ainhoa

"..Sólo porque prefiero lo positivo a lo negativo. Pero en este juego que estamos jugando no podemos ganar. Unas clases de fracaso son quizá mejores que otras, eso es todo.."

1984, George Orwell

Publicaciones

Durante el transcurso de la carrera de Doctorado en Ciencias Biológicas (FBCB-UNL) y en relación a los trabajos realizados en la presente tesis, tuvieron lugar las siguientes publicaciones:

G.G. Siano y H.C. Goicoechea (2007) *Representative subset selection and standardization techniques. A comparative study using NIR and a simulated fermentative process UV data.* Chemom. Intell. Lab. Syst. 88, 204–212

M.M. De Zan; M.D. Gil García; M.J. Culzoni; G.G. Siano; H.C. Goicoechea y M. Martinez Galera (2008) *Solving matrix-effects exploiting the second order advantage in the resolution and determination of eight tetracycline antibiotics in effluent wastewater by modelling liquid chromatography data with multivariate curve resolution-alternating least squares and unfolded-partial least squares followed by residual bilinearization algorithms. I. Effect of signal pre-treatment*, J. Chromatogr. A 1179, 106–114

M.D. Gil García; F. Cañada Cañada; M.J. Culzoni; L. Vera-Candioti; G.G. Siano; H.C. Goicoechea y M. Martínez Galera (2009) *Chemometric tools improving the determination of anti-inflammatory and antiepileptic drugs in river and wastewater by solid-phase microextraction and liquid chromatography diode array detection*, J Chromatogr A. 29, 5489–5496

J.H. Kalivas; G.G. Siano; E. Andries y H.C. Goicoechea (2009) *Calibration Maintenance and Transfer Using Tikhonov Regularization Approaches*, Appl. Spectrosc. 63, 800-809.

I. Sánchez Pérez; M.J. Culzoni; G.G. Siano; M.D. Gil García; H.C. Goicoechea y M. Martínez Galera (2009) *Detection of unintended stress effects based on a metabonomic study in tomato fruits after treatment with carbofuran pesticide. Capabilities of MCR-ALS applied to LC-MS three-way data arrays*, Anal. Chem. 81, 8335-8346.

G.G. Siano; I. Sánchez Pérez; M.D. Gil García; M.Martínez Galera y H.C. Goicoechea (2011) *Multivariate curve resolution modelling of liquid chromatography-mass spectrometry data in a comparative study of the different endogenous metabolites behaviour in two tomato cultivars treated with carbofuran pesticide,* Talanta 85, 264-275.

Índice de Contenidos

Abreviaturas y Símbolos

Abreviatura o símbolo	Significado
+	Cálculo de pseudoinversa de Moore-Penrose para una matriz (como superíndice)
-1	Cálculo de inversa para una matriz (como superíndice)
%LOF (EXP)	Porcentaje de Falta de Ajuste (Experimental) en MCR-ALS
$\%R^2$	Porcentaje de Varianza Explicada en MCR-ALS
a	Escalar a
A	Matriz A
a	Vector a
ALS	Mínimos Cuadrados Alternantes (*Alternating Least Squares*)
b	Vector de regresión
CCC	Clasificaciones Correctas en Calibración según PLS-DA
CCV	Clasificaciones Correctas en Validación según PLS-DA
CMV	Calibración Multivariada
cnom	Concentración nominal
CPF	Ciprofloxacina
cpred	Concentración predicha
CV	Validación cruzada (*Cross Validation*)
CWT	Transformada Wavelet Continua
DA	Análisis Discriminante (*Discriminant Analysis)*
DAD	Detección con Arreglo de Diodos
DNF	Danofloxacina
DR	Doble Regularización de Tikhonov
DWT	Transformada Wavelet Discreta
ESI	Ionización por Electrospray
F	Clase de Tomates RAF
Fb	Clase de Tomates RAF blanco
Fcr	Datos de Fluorescencia Crudos
Fss	Datos de Fluorescencia procesados con interpolación (*spline*) y suavizados mediante polinomios de Savitsky-Golay
HPLC	Cromatografía Líquida de Alto Rendimiento
InvSenAn	Inversa de Sensibilidad Analítica
IR	Infrarrojo
iter	Iteraciones en MCR-ALS
IWT	Transformada Wavelet Inversa
KS	Algoritmo de Kennard-Stone
lam	Conjunto de valores de λ^2 en DR

Abreviatura o símbolo	Significado
LC	Cromatografía Líquida
LOD	Límite de Detección
LOQ	Límite de Cuantificación
LV	Variable Latente
m/z	Relación masa/carga
MB	Muestra Blanco (sin Carbofurano)
MCn	Estrategia de Centrado n en DR
MCR	Resolución Multivariada de Curvas
mRec%	media de Recuperaciones porcentuales
MS	Espectrometría de Masa
MT	Muestra Tratada (con Carbofurano)
nb o $\|\|\mathbf{b}\|\|$	Norma Euclidiana de un vector de regresión **b**
ncomp	Número de componentes modelados en MCR-ALS
NIR	Infrarrojo Cercano
OFL	Ofloxacina
PCA	Análisis de Componentes Principales (*Principal Component Analysis)*
PCR	Regresión en Componentes Principales (*Principal Component Regresion*)
PLS	Mínimos Cuadrados Parciales (*Partial Least Squares)*
R	Clase de Tomates Rambo
Rb	Clase de Tomates Rambo blanco
REP	Error Relativo en Predicciones (*Relative Error of Prediction)*
RMSE	Raíz cuadrada del Error Cuadrático Medio (*Root Mean Square Error)*
RR	Regresión "Ridge" (*Ridge Regression)*
SenMCR	Sensibilidad para calibraciones pseudo-univariadas basadas en áreas resueltas con MCR-ALS
SVD	Descomposición en Valores Singulares (*Singular Value Decomposition)*
T	Cálculo de transpuesta de una matriz (como superíndice)
tau	Conjunto de valores de τ^2 en DR
TR	Regularización de Tikhonov
UV	Ultravioleta
WT	Transformada Wavelet
WT2	Transformada Wavelet Bidimensional
λ	Ponderador de error para muestras de transferencia en DR
τ	Regulador de norma vectorial en DR

Índice de tablas

Índice de figuras

Resumen

El presente trabajo estuvo relacionado a la resolución de distintas problemáticas en laboratorios de Química Analítica. Para esto se desarrollaron y/o aplicaron herramientas quimiométricas, las cuales permitieron realizar diversos análisis sobre datos provenientes de muestras biológicas y químicas. El trabajo se dividió en tres capítulos con objetivos diferentes.

El capítulo 1 fue destinado a resolver problemas con datos de orden 1, específicamente relacionados a la desactualización de modelos de Calibración Multivariados y a la necesidad de realizar procedimientos de Transferencia de Calibración. Se presenta un algoritmo desarrollado a tal fin, derivado de la Regularización de Tikhonov, el cual fue denominado Doble Regularización de Tikhonov. Tras análisis iniciales generalistas y finales en detalle, se expone información relativa a la cuantificación de Etanol en muestras ternarias a través de espectros Infrarrojos obtenidos en dos temperaturas diferentes, y de contenido proteico en muestras de maíz mediante espectros del mismo tipo obtenidos en dos instrumentos distintos. Los resultados demuestran la utilidad del algoritmo para realizar la transferencia y sus potenciales ventajas.

En el capítulo 2 se aplicaron algoritmos quimiométricos de pre-procesamiento de señales, de resolución multivariada de curvas y de clasificación para generar estrategias con el objetivo de realizar análisis metabonómicos de muestras provenientes de frutos de tomate en busca de efectos de *stress* derivados de la aplicación del pesticida Carbofurano. Las muestras se analizaron mediante Cromatografía Líquida acoplada a Espectrometría de Masa, y los datos de orden 2 derivados fueron procesados mediante transformada Wavelet, resueltos con MCR-ALS y clasificados con PLS-DA. Los resultados sugieren que las estrategias propuestas son válidas para detectar este tipo de efectos de *stress*.

En el capítulo 3 se utilizaron herramientas quimiométricas no sólo en su forma convencional, sino también para evaluar el desempeño de un dispositivo recolector de muestras y el de interfaces de comunicación con el recolector y con un fluorímetro. Para la mayor parte de estas tareas se utilizó tanto software como hardware de código abierto, y en la construcción del recolector se reciclaron muchos componentes de tecnologías en desuso. Las muestras analizadas contuvieron distintas proporciones de tres fluoroquinolonas, y de éstas se derivaron datos de orden 2 a través de Cromatografía Líquida acoplada a Espectroscopia Ultravioleta, con posterior recolección en placas de ELISA y obtención de matrices de Excitación-Emisión en el fluorímetro. Los resultados dejan ver la potencial utilidad de incluir tecnologías de código abierto en el laboratorio analítico.

Abstract

The present work was related to the resolution of various problems in Analytical Chemistry laboratories. To this were developed and/or applied chemometric tools, which allowed for various analyzes of data from biological and chemical samples. The work was divided into three chapters with different objectives.

Chapter 1 was designed to solve problems with order 1 data, specifically related to the obsolescence of Multivariate Calibration models and to the need for Calibration Transfer procedures. It is presented an algorithm developed for this purpose, derived from the Tikhonov Regularization, which was called Double Tikhonov Regularization. After initial analysis in a general way and final ones in detail it is presented information concerning to the quantization of Ethanol in ternary samples through Infrared spectra obtained at two different temperatures, and protein content in maize samples using the same kind of spectra obtained in two different instruments . The results demonstrate the utility of the algorithm to performing the transfer and its potential advantages.

In Chapter 2 chemometric algorithms were applied for pre-processing of signals, multivariate curve resolution and classification to generate strategies in order to develop metabonomics analysis of samples from tomato fruit looking for stress effects arising from the application of Carbofuran pesticide. The samples were analyzed by Liquid Chromatography coupled to Mass Spectrometry, and the derived order 2 data were processed using wavelet transform, resolved by MCR-ALS and classified by PLS-DA. The results suggest that the proposed strategies are valid to detect such effects of stress.

In Chapter 3 chemometric tools were used not only in its conventional form, but also to evaluate the performance both of a sample collection device and of communication interfaces with the collector and a fluorometer. For most of these tasks it was used both software and hardware of open source code, and for the construction of the collector several components were recycled from obsolete technologies. Samples tested contained different proportions of three fluoroquinolones, and from these samples order 2 data were derived through Liquid Chromatography coupled with Ultraviolet Spectroscopy, with subsequent collection in ELISA plates and obtention of Excitation-Emission matrices in the fluorometer. The results reveal the potential utility of including open source technologies in the analytical laboratory.

CAPÍTULO 1: Transferencia de modelos de Calibración Multivariada de primer orden mediante Doble Regularización de Tikhonov.

1.1 Resumen

El mantenimiento de modelos de Calibración Multivariada es esencial y envuelve operaciones tales que los modelos desarrollados en una situación original o primaria puedan ser re-adaptados para predecir correctamente a muestras que provengan de una situación nueva o secundaria, con nuevas fuentes de varianza no modeladas originalmente. Esta re-adaptación en situaciones recibe el nombre de transferencia de Calibración.

En este trabajo, los procedimientos de actualización necesarios para datos de orden 1 fueron llevados a cabo en el marco de la Regularización de Tikhonov (TR). Ya que la aplicación directa de la teoría de TR no siempre es posible en transferencia de modelos debido a problemas de inversión matriciales, en este trabajo se desarrolló una variante que en general no tendrá inconvenientes a la hora de ser utilizada, a la cual se la denominó Doble Regularización de Tikhonov.

La selección de modelos para su análisis estuvo basada en criterios de armonía que relacionaron los errores de ajuste de distintos conjuntos de muestras con la norma de los vectores de regresión pertinentes, siendo que la última representa un criterio de varianza en las predicciones de muestras futuras. Dado que ambos criterios suelen competir en circunstancias, múltiples diagramas del tipo Pareto permitieron discernir entre diversas condiciones, de forma tal de actualizar los modelos aceptablemente.

Se evaluaron dos estrategias básicas para las transferencias. En una de ellas se utilizaron espectros secundarios para actualizar modelos primarios, y en la otra los últimos fueron actualizados con diferencias de espectros de muestras equivalentes provenientes de ambos dominios. También fueron puestas a prueba 4 estrategias de centrado de datos. A su vez, se evaluaron los procedimientos con cantidades variables de muestras de transferencia.

Dos conjuntos de datos sirvieron para el estudio. En uno de ellos el analito de interés fue Etanol en mezclas ternarias, cuyos espectros IR fueron obtenidos a dos temperaturas diferentes. En el otro caso se cuantificó contenido proteico en muestras de maíz, a través de espectros IR provenientes de 2 instrumentos distintos.

Los resultados obtenidos, también comparados con los de otros métodos como PLS y PDS, indican que efectivamente es posible actualizar modelos primarios con escasas muestras de transferencia a través de DR y estrategias de centrado apropiadas.

1.2 Introducción

En términos generales, una Calibración Multivariada (CMV) tiene como objetivo encontrar relaciones entre una o más variables dependientes, pudiendo ser éstas propiedades físicas o químicas de los sistemas, respecto de variables independientes obtenidas de éstos, como pueden ser espectros de diferentes tipos. Dichas obtenciones, según limitaciones o posibilidades instrumentales, darán origen a datos de distinto orden, término reservado para hacer referencia a un conjunto múltiple de variables que conservan algún tipo de relación funcional entre sí. A modo de ejemplo, una muestra compleja puede ser separada en sus componentes por aplicación de algún tipo de cromatografía, lo cual originará datos en el orden del tiempo de elusión. Si a los componentes aislados en cada tiempo de separación se les aplicaran estímulos variables en busca de respuestas relativas a dichos estímulos, se obtendrían múltiples datos en este otro orden, generando así datos de segundo orden por muestra analizada.

Similarmente, una CMV de primer orden intentará encontrar una relación entre variables dependientes de la composición muestral (Analitos/Propiedades de Interés, A/PI) y vectores de respuestas en múltiples variables, siempre afectadas por la técnica y por el entorno. En estos casos, la relación puede plantearse matemáticamente de la siguiente manera:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \quad (1)$$

donde **y** denota un vector de dimensiones *m* x 1 que contiene información cuantitativa referida a un A/PI en *m* muestras de Calibración (si fueran *N* A/PI, **Y** sería de *m* x *N*); **X** representa una matriz de Calibración, cuyo tamaño será de *m* x *n*, la cual contendrá las respuestas de las *m* muestras de Calibración dadas *n* variables para la predicción; y **b** constituye un vector de dimensiones *n* x 1, el cual contendrá los coeficientes de regresión para el modelo de Calibración una vez que éstos hayan sido estimados a partir de la información experimental contenida en **X** e **y**. Finalmente, **e** hace referencia a un vector de dimensiones *m* x 1, el cual se espera que contenga errores normalmente distribuidos con media cero y matriz de covarianza $\sigma^2\mathbf{I}$ (**I** es una matriz identidad de dimensiones compatibles con los operandos de cada caso). Al respecto, en lo sucesivo podrá obviarse el término **e** en diferentes ecuaciones con el objetivo de simplificar el entendimiento a través de los tratamientos matemáticos, sin asumir con esto la ausencia de error. La estimación de **b** puede realizarse de múltiples maneras. Entre los métodos comúnmente usados para tal fin, pueden nombrarse PLS (del inglés *Partial Least Squares*), PCR (del inglés *Principal Component*

Regression) y RR (del inglés *Ridge Regression*) (Kalivas, 2001), entre otros. En general, dichos métodos son aplicados cuando $n>>m$, en referencia al tamaño de la matriz de Calibración. A su vez, los métodos también son válidos cuando $n \leq m$, pero en dichos casos también puede aplicarse MLR (del inglés *Multiple Linear Regression*) para obtener los coeficientes de regresión. Independientemente del método utilizado, el objetivo primordial de una CMV es encontrar una estimación apropiada para **b** de manera tal de obtener exactitud (mínimo sesgo) y precisión (mínima varianza) en la predicción del A/PI calibrado para muestras futuras. Dicha estimación puede ser obtenida a través de $\hat{y} = \mathbf{x}^t \hat{\mathbf{b}}$, donde $\mathbf{x}^t$ representa la transposición del vector de respuestas para una muestra en las *n* variables independientes, e $\hat{y}$ el valor estimado por el modelo de Calibración para el A/PI en cuestión (Næs y col., 2002; Hastie y col., 2001).

Una vez que un modelo ha sido definido, la duración de su aplicabilidad se convierte en un factor relevante, lo cual da origen a la idea de mantenimiento o actualización de un modelo. A excepción de situaciones extremadamente controladas o de sistemas prácticamente invariantes, no es lógico pensar que una situación modelada en cierto tiempo contemplará todas las variantes posibles para tiempos futuros, sino que en general sucederá lo contrario, dados los cambios habituales en los que los sistemas dinámicos suelen incurrir. Dentro del amplio conjunto de efectos que llevarían al fallo de una Calibración, puede nombrarse la aparición de características no modeladas en las señales de muestras obtenidas en tiempos posteriores. Esto podría deberse a concentraciones de analitos u otra propiedad de interés por fuera del rango calibrado, a la inclusión de muestras con componentes que respondan al mismo estímulo de las señales modeladas pero que no hubieran participado de la Calibración y a cambios instrumentales (desplazamientos, cambios de fuentes, detectores, otros componentes o del instrumento en sí), como así también físicos, químicos y/o ambientales, como ser cambios de viscosidad, tamaño de las partículas, textura de las superficies, pH, temperatura, presión, humedad, etc. Así pues, es razonable concebir que al menos uno de estos eventos tendrá lugar con el discurrir del tiempo, lo cual invalidaría las conclusiones que del modelo pudieran obtenerse. Por ende, deben existir mecanismos para corregir los efectos que pudieran estar fuera del dominio multivariado calibrado.

En relación a lo anterior existe el denominado problema de transferencia de Calibración. En ocasiones, este procedimiento suele ser referido con denominaciones como actualización o mantenimiento de modelos. No obstante, en el contexto de este escrito se utilizará genéricamente la palabra “transferencia” para representar la idea de modelar un nuevo espacio multivariado, transfiriendo la información (señales y valores de referencia) original o primaria hacia un

hiperespacio apto para predecir muestras de 1 ó más dominios nuevos o secundarios, a través de la inclusión conjunta de información proveniente de unas pocas muestras actuales y conocidas (denominadas muestras de transferencia) durante las etapas de modelado. El hecho de que se pretenda realizar el procedimiento con pocas muestras de transferencia, sumado al hecho de que la intención también radica en reutilizar la información primaria desactualizada pero generalmente en mayor número, indican que el procedimiento estará destinado a ahorrar recursos para corregir las situaciones problemáticas. A su vez, se resalta que luego de una transferencia se obtendrá un modelo nuevo, diferente al original. Esta última aclaración se realiza para diferenciar al procedimiento de otros también relacionados a estas problemáticas pero con características distintas. Por ejemplo, en los procedimientos de estandarización espectral el problema se intenta corregir a través de la transformación de espectros para que éstos puedan ser predichos por modelos originales que no son modificados en sí.

El problema relacionado a la inutilización de modelos ha sido estudiado y documentado en varias ocasiones (de Noord, 1994; Fearn, 2001; Feudale y col., 2002; Cogdill y col., 2005). Existen 3 modos generales para realizar correcciones o evitar desactualizaciones. Uno de estos consiste en realizar inicialmente un modelo robusto, lo cual puede lograrse aplicando preprocesamientos a las señales, como MSC (del inglés *Multiplicative Scatter Correction*), filtros FIR (del inglés *Finite Impulse Response*), derivadas, Wavelets, selección de longitudes de onda, etc. Demás está decir que el conjunto de preprocesamientos utilizados para las señales de Calibración deberá ser reutilizado con las señales provenientes de muestras posteriores antes de obtener sus predicciones, o bien deberá existir algún tipo de estrategia diferencial para las incógnitas, pero ya definida en las etapas de Calibración y Validación. Un mecanismo alternativo para formar un modelo robusto es realizar una Calibración global, la cual incluiría todos los efectos potenciales, sean químicos, físicos, ambientales y/o instrumentales, en el modelo original. En otras palabras, se obtendrían en simultáneo señales de muestras que contemplen todos los cambios futuros posibles, por ejemplo muestras a diferentes pH y temperaturas si se esperan cambios en estas condiciones. Sin embargo, la dificultad de este mecanismo radica en la gran cantidad de muestras necesarias para cubrir todos los efectos potenciales futuros y en que, para cada muestra, el valor de referencia para el A/PI deberá ser determinado con el fin de formar **y**, lo cual en general insumirá recursos. Una opción también radica en agregar columnas a **X** (variables) con valores para las distintas condiciones, como ser temperatura o tiempos de desplazamiento (Kalivas y Kowalski, 1982; Wülfert y col., 2000a).

Otro de los modos generales para realizar correcciones consiste en ajustar o transformar las

señales de muestras provenientes de uno o más instrumentos (o situaciones), de manera tal que el producto sea hipotéticamente equivalente a las señales que se hubiesen obtenido de las mismas muestras, pero en otro instrumento de referencia. Considerando al último como del tipo primario, el proceso haría que las señales fueran aptas para ser predichas por un modelo de Calibración originalmente desarrollado en el instrumento de referencia. En este caso el modelo de Calibración no sería transferido, sino reutilizado tal y como fue diseñado (es decir, el vector de regresión no sería modificado en absoluto). Una vez encontrada la forma de transformar la información desde los dominios secundarios hacia el primario y teniendo en cuenta que el modelo para realizar las predicciones ya estaría elaborado, el procedimiento solamente involucraría a las señales de los instrumentos no primarios, en cuyo caso se torna apropiado hablar de estandarización de señales. Por otro lado, la transformación podría realizarse desde el dominio de las señales primarias hacia uno o más dominios secundarios, estableciendo luego nuevos modelos de Calibración que utilicen la información primaria transformada. Independientemente de la elección, estos métodos subyacen tras el marco general del método estadístico denominado Análisis de Procrustes (PA), con el cual se pretende encontrar una función de mapeo matemático para información proveniente de diferentes dominios (Anderson y Kalivas, 1999). En el contexto de este escrito el término “dominios” hace referencia a situaciones diferentes (instrumentos distintos, temperaturas distintas, etc.), mientras que el término “información” refiere a las señales obtenidas de éstas y a sus valores de referencia respectivos. La transformación a través de PA involucrará determinar pasos apropiados de rotación, translación y contracción/expansión en las señales de un subgrupo de muestras provenientes de un dominio, tales que las señales transformadas se asemejen a las provenientes de otro dominio objetivo. Debe entenderse que el pasaje entre dominios tiene entre sus metas reutilizar información confiable para evitar tener que obtenerla nuevamente. En términos prácticos, esto estará asociado a un ahorro de recursos. Así, en el primer caso descripto se reutiliza un modelo primario (e indirectamente la información primaria que le dio origen) para predecir muestras secundarias, mientras que en el segundo el nuevo modelo de Calibración podrá estar basado en la información primaria transformada. Por lo tanto, no sería compatible con esta idea el tener que generar mucha información nueva para suplir la necesidad propia del cálculo de las transformaciones. Dicho de otra manera, el proceso se realizará por medio de información obtenida solamente de unas pocas muestras en los distintos dominios, pues si hubiera posibilidad de obtener más, no se pensaría en estandarización ni transferencia de modelos, sino más bien en re-Calibración directa. A su vez, las variables registradas podrían diferir para los dominios. Por ejemplo, el mismo subgrupo de muestras

podría medirse en un intervalo de longitudes de onda diferente al cambiar instrumentos, siempre y cuando esto tuviera fundamentos que respaldaran ese accionar. Uno de los métodos más populares para realizar estandarización de señales es el denominado PDS (del inglés *Piecewise Direct Standardization*) (Wang y col., 1995). Para su realización, un grupo de muestras proveerá de señales en (al menos) dos dominios, uno de los cuales será definido como primario u objetivo (antiguamente llamado amo), y el otro como secundario (antiguamente llamado esclavo). En el último se definirá una cantidad de variables, llamada ventana, para realizar modelos de regresión multivariados localizados. Usualmente la ventana contendrá un número impar de elementos, y el elemento central indicará la variable primaria contra la que se quieren ajustar las variables de una ventana secundaria. El movimiento de la ventana a través del conjunto total de variables secundarias dará origen a todos los modelos que sean necesarios, los cuales serán ensamblados en una matriz. Dado que los modelos localizados suelen calcularse con PCR o PLS, será necesario indicar un número fijo de factores (Componentes Principales/Variables Latentes) para todos los modelos, o bien un parámetro de tolerancia que indique, para cada localización, cuántos factores deberán ser usados en el cálculo.

El tercer modo general de realizar correcciones consiste en rehacer un modelo de Calibración para predecir apropiadamente señales de un dominio diferente al original, pero sin transformar a las señales. En este escrito, este modo coincide con la idea de transferencia de Calibración. Ya que este tercer modo ha sido objeto de estudio en el presente trabajo, y teniendo en cuenta que se han utilizado datos espectroscópicos, desde ahora en más se hablará de espectros para representar lo que, de forma genérica, podrían ser distintos tipos de señales. Habiendo detectado una fuente de varianza no modelada presente en nuevas muestras, una opción consiste en adicionar a cada espectro de la matriz de Calibración primaria, **X**, las formas espectrales pertinentes a los efectos no modelados (espectros puros, por ejemplo) y ausentes en **X**, tales como corrimientos, nuevas sustancias químicas, temperatura, entre otros (Haaland, 2000). Luego de la adición, un método de regresión debe ser aplicado para estimar un nuevo modelo actualizado. Por ejemplo, en un diseño relacionado a esta idea la matriz **X** se obtuvo a partir de espectros sintéticos basados en una descripción matemática de la propagación de la luz a través de la piel, y a cada uno de estos espectros se le adicionó un espectro característico de la piel de cada sujeto en estudio, con lo cual se realizó un modelo para glucosa por sujeto (Maruo y col., 2006). Esta metodología de adición de un espectro de las nuevas condiciones a cada espectro primario de Calibración requiere una única determinación de referencia, sólo para la muestra en las nuevas condiciones que es adicionada al

resto. Si la nueva muestra no contiene al analito (por ejemplo una corrección de deriva espectral) entonces no es necesario un nuevo análisis de referencia.

En lugar de adicionar espectros provenientes de la nueva condición a los espectros previamente existentes (reales o simulados), una alternativa para formar un modelo actualizado es aumentar el conjunto de Calibración original (**X** e **y**) con muestras adicionales conteniendo las nuevas variaciones químicas o instrumentales. En este caso, la ecuación (1) puede escribirse de la siguiente forma (ignorando el término de error, **e**):

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{y_L} \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \mathbf{L} \end{pmatrix} \mathbf{b} \quad (2)$$

donde **L** representa una matriz de espectros de tamaño *l* x *n* proveniente de *l* muestras en las nuevas condiciones, e $\mathbf{y_L}$ denota las concentraciones (o propiedades de interés) respectivas. Aplicando un método de regresión a la ecuación (2), puede obtenerse una estimación de **b** correspondiente a un vector actualizado a las nuevas condiciones. Con esta forma de proceder será necesaria la determinación de los valores de referencia ($\mathbf{y_L}$) para las muestras en **L**, por lo cual el procedimiento crecerá en complejidad si deben utilizarse métodos de referencia complejos para las determinaciones y a su vez si el número de muestras en **L** es considerable.

Usar solamente unas pocas muestras en **L** para caracterizar las nuevas condiciones ha sido propuesto y estudiado (Westerhaus, 1991; Wang y col., 1991). En todos los casos en que se pretenda realizar transferencia de Calibración usando pocas muestras será crítica la selección de dichas muestras (Capron y col., 2005). Este pequeño subconjunto de muestras suele ser referido como subconjunto (o subset) de transferencia y se requiere que, en la medida de lo posible, éste cubra totalmente el nuevo espacio multivariado de una manera adecuada para describir las nuevas fuentes de varianza.

Si las señales del subconjunto de transferencia pueden ser obtenidas cuando se realiza el modelo de Calibración primario (o bien luego en el tiempo, pero en las mismas condiciones) y también bajo las nuevas condiciones o instrumentos secundarios, entonces pueden utilizarse diferencias de señales para **L**, para lo cual $\mathbf{y_L} = 0$ (Westerhaus, 1991). Esta metodología cuenta con la ventaja de no necesitar métodos de referencia para obtener valores para $\mathbf{y_L}$. Sin embargo, si se quieren realizar actualizaciones en distintos tiempos futuros, será necesario que la composición de las muestras del subconjunto de transferencia sea estable a largo plazo, pues en caso contrario se estarían equiparando señales secundarias obtenidas de muestras cuya composición no sería la misma que a nivel primario. Otra posibilidad radica en utilizar muestras sin el A/PI y nuevamente $\mathbf{y_L}$ valdría 0,

aunque esto puede cuestionarse, por cuanto infinitas muestras podrían cumplir esa característica, muchas de las cuales no representarían correctamente a la información necesaria para transferir. Un intento de este tipo fue realizado con señales provenientes de blancos para **L**, las cuales fueron obtenidas durante la etapa de precalentamiento del instrumento primario, con el fin de realizar ajustes para un nuevo perfil instrumental y corregir cualquier desplazamiento que pudiera haber ocurrido (Kramer y Small, 2007). Otra forma en que se ha aplicado la ecuación (2) consistió en usar para **X** espectros medidos en laboratorio a partir de soluciones preparadas y para **L** un pequeño conjunto de muestras medidas en las nuevas condiciones en las cuales el modelo sería usado (Riley y col., 1998). Al usar un conjunto de muestras de laboratorio bien diseñado, el modelo puede caracterizar mejor la información dependiente del A/PI, mientras que **L** permite al método de regresión realizar las correcciones necesarias para las nuevas condiciones, que en ese caso estaban referidas a un medio de cultivo. Le ecuación (2) también fue utilizada en otro trabajo de transferencia de Calibración, en el que se aumentó una matriz **X** de espectros puros para componentes simulados, con una matriz **L** que contenía espectros reales (Sulub y Small, 2007).

El concepto de aumentar las muestras de Calibración originales con información que cubra las nuevas condiciones se ha aplicado a transferencia de modelos de Calibración por medio de métodos híbridos de predicción (Wehlburg y col., 2002a; Wehlburg y col., 2002b). La caracterización de las nuevas condiciones se lleva a cabo midiendo repetidamente el espectro de una única muestra, seleccionada a partir del centro del espacio de concentraciones. Con este método, el valor de referencia para $\mathbf{y}_L$ necesita ser determinado sólo una vez. Otra posibilidad para insertar en **L** efectos espectrales y desensibilizar un modelo primario consiste en incluir representaciones matemáticas de desplazamientos, interferencias espectrales conocidas, picos de solventes, señales de fondo, entre otras posibilidades, y luego en asociar cada una a valores de referencia de 0, con lo se obtendrán modelos ortogonales a estas características.

Un problema general con el uso de la ecuación (2) está relacionado al tamaño muestral del conjunto de transferencia. Si éste es pequeño, lo cual es normal, entonces los espectros originales de Calibración en **X** tendrán una mayor influencia por el simple hecho de contener más información debido a la mayor cantidad de muestras. Por lo tanto, un esquema de ponderación ha sido propuesto, el cual modifica la ecuación (2):

$$\begin{pmatrix} \mathbf{y} \\ \lambda\,\mathbf{y}_L \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \lambda\,\mathbf{L} \end{pmatrix} \mathbf{b} \quad (3)$$

donde λ simboliza un valor de peso o ponderación (Stork y Kowalski, 1999). El uso de métodos

de regresión como PLS, PCR o MLR para estimar **b** en la ecuación (3) requiere que se determinen los respectivos meta-parámetros. Sin aumentar, es decir, usando la ecuación (1), el meta-parámetro a determinar para PLS y PCR es el número de vectores base (Variables Latentes y Componentes Principales, respectivamente), mientras que para MLR son el número e identidad de las variables disponibles. Con la ecuación (3), existirá ahora un nuevo meta-parámetro de peso, λ, que deberá ser fijado. Si éste último es demasiado grande, el modelo actualizado dará mucha importancia a la información contenida en **L** e $\mathbf{y}_L$, y se degradará al perder varianza relevante proveniente de las muestras originales de Calibración. En cambio, si es demasiado pequeño, la información extra a la original básicamente no será tenida en cuenta. Por ende, una desventaja de la ecuación (3) es que se carece de una metodología obvia para determinar apropiadamente las ponderaciones. Este proceso, por ejemplo, ha sido realizado tomando como base medidas replicadas para las muestras del subset de transferencia (Capron y col., 2005; Stork y Kowalski, 1999). Así pues, si λ = 1, entonces no se usan replicados, mientras que si λ = 2, se usan duplicados (obviamente λ podría tomar cualquier otro valor). Esto no ha sido satisfactorio en todos los casos.

En lugar de usar múltiples copias de ciertos espectros como una manera de dar un valor para λ, pueden realizarse perturbaciones a las señales de las muestras de Calibración originales (o a las del subset de transferencia) aplicando ruido al azar en varias combinaciones. De esta manera puede aumentarse **X** con una sola matriz **L** o bien con múltiples matrices **L** para diferentes perturbaciones (Sáiz-Abajo y col., 2005). Esta forma de aumentar **X** con múltiples arreglos de **L**, provenientes de la modificación con ruido de las señales de las muestras en **X**, se conoce como método de agrupación (*ensemble method*) (Zhu, 2008). Con este método λ no es necesaria y la ecuación (2) es resuelta con un método de regresión para estimar **b**, aunque debe decidirse qué número de señales perturbadas será usado para aumentar **X** y cómo deberán ser realizadas dichas perturbaciones. Cuanto más se parezca la estructura de ruido modelada en **L** al ruido en **X**, mayor será la desensibilización del modelo con respecto al ruido. Por ejemplo **X** fue medida a 36ºC y con la diferencia media de espectros a 34ºC y 38ºC, para distintas muestras, se formaron arreglos de múltiples **L** que fueron agregados a los espectros en **X** (Mevik y col., 2004).

Para finalizar, se hace notar que la ecuación (3) es una representación de la Regularización de Tikhonov (TR), tema sujeto a estudio en el presente trabajo en el contexto de transferencia de modelos de Calibración multivariada de primer orden. Por lo tanto, la idea de ponderar la información de las muestras en **L** será desarrollada a través de las bases lógicas y matemáticas aportadas por la TR. Más detalles al respecto serán presentados en la sección Teoría.

1.3 Objetivos

- Someter a evaluación estrategias y algoritmos derivados de la Regularización de Tikhonov para realizar transferencia de modelos CMV de orden 1
- Reutilizar las bases teóricas de la TR para interpretar los efectos de la ponderación parcial de la información en modelos CMV de orden 1 transferidos
- Evaluar diferentes estrategias de centrado como únicos procesamientos previos a las transferencias
- Evaluar el efecto del número de muestras de transferencia y obtener resultados aceptables a partir de transferencias realizadas con pocas muestras
- Evaluar efectos provenientes de la representatividad de las muestras de transferencia en relación a muestras futuras
- Evaluar la conveniencia de reutilizar la información primaria o descartarla, y si es recomendable utilizar la información secundaria disponible para realizar calibraciones secundarias directamente o bien para transferencias
- Evaluar efectos provenientes del cambio en los meta-parámetros de transferencia
- Contrastar los resultados con los obtenidos por medio de otros algoritmos relacionados a problemáticas similares
- Definir estrategias en base a las evaluaciones realizadas con el objetivo de predecir resultados de calidad aceptable luego de las transferencias

1.4 Teoría

1.4.1 Regularización de Tikhonov (TR) y variantes

La ecuación (3) es en realidad una representación de la Regularización de Tikhonov (TR) (Tikhonov, 1943) (Tikhonov, 1963). Este procedimiento está relacionado a problemas que no pueden ser bien determinados, es decir, sin una solución única o sin solución. Dado un problema del tipo **Xb = y**, siendo **b** e **y** vectores de *n* y *m* componentes, respectivamente, y **X** una matriz de tamaño *m* x *n*, el criterio convencional de mínimos cuadrados intentará minimizar los residuos de:

$$\left\|\mathbf{Xb}-\mathbf{y}\right\|_2^2 \quad (4)$$

donde $\|\ \|_2^2$ representa a la norma euclidiana (por simplicidad y para evitar confusiones con operaciones de potenciación, en lo que resta de este manuscrito se hará referencia a cualquier norma solamente con el subíndice). Por otro lado, la TR dará preferencia a alguna solución con propiedades deseadas, incluyendo términos de regularización extras. Funcionalmente, la regularización mejora el condicionamiento del problema, lo cual facilita encontrar soluciones numéricas. En un sentido amplio, la formulación más general de la TR se expresa identificando los coeficientes de regresión en **b** tales que:

$$\min\left(\left\|\mathbf{Xb}-\mathbf{y}\right\|_g+\lambda^2\left\|\mathbf{L}\left(\mathbf{b}-\mathbf{b}^*\right)\right\|_h\right) \quad (5)$$

donde *g* y *h* representan la misma o diferentes normas en el rango $1 \leq g$, $h < \infty$, **L** corresponde a un operador de regulación que fuerza la estimación de **b** de manera tal que se corresponda con un sub-espacio en particular, **b*** simboliza los verdaderos coeficientes del modelo (para un analito por ejemplo) y λ representa un meta-parámetro de regularización que controla la ponderación dada al segundo término de la ecuación, el cual se corresponde con el criterio de mínimos cuadrados para el caso en que *g* = 2 (Hansen, 1998; Aster y col., 2005; Dax, 1992). En la expresión (5), el término de la izquierda indica exactitud, mientras que el de la derecha representa el tamaño del modelo. Para el caso en que *h* = 2, dicho tamaño representado por la norma euclidiana del vector actúa como un indicador de varianza y precisión (Forrester y Kalivas, 2004), tal que a partir de cierto tamaño, cuanto mayor sea este, existirá mayor varianza en las predicciones futuras. Las opciones para *g* y *h* son variadas, y cuando ambos toman el valor 2, la solución a la expresión (5) es:

$$\hat{\mathbf{b}}=\left(\mathbf{X}^t\mathbf{X}+\lambda^2\mathbf{L}^t\mathbf{L}\right)^{-1}\left(\mathbf{X}^t\mathbf{y}+\lambda^2\mathbf{L}^t\mathbf{L}\,\mathbf{b}^*\right) \quad (6)$$

lo cual también es solución de:

$$\begin{pmatrix} \mathbf{y} \\ \lambda\,\mathbf{L}\mathbf{b}^{*} \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \lambda\,\mathbf{L} \end{pmatrix} \mathbf{b} \quad (7)$$

Es normal que **b*** no sea conocido, tal y como suele ocurrir en el análisis espectroscópico, donde **L** estaría compuesto de espectros o derivados. En tal caso, $\mathbf{y}_L \approx \mathbf{L}\mathbf{b}^{*}$ y la expresión (5) se reduce a:

$$\min\left(\left\|\mathbf{X}\mathbf{b}-\mathbf{y}\right\|_2+\lambda^2\left\|\mathbf{L}\mathbf{b}-\mathbf{y}_L\right\|_2\right) \quad (8)$$

En ese caso, la expresión (7) se convierte en la (3), con la siguiente solución:

$$\hat{\mathbf{b}}=\left(\mathbf{X}^t\mathbf{X}+\lambda^2\mathbf{L}^t\mathbf{L}\right)^{-1}\left(\mathbf{X}^t\mathbf{y}+\lambda^2\mathbf{L}^t\mathbf{y}_L\right) \quad (9)$$

Así pues, el valor de peso determinado empíricamente en la ecuación (3) en trabajos previos es en realidad el meta-parámetro normal de la TR (Hansen, 1998; Aster y col., 2005; Forrester y Kalivas, 2004; Lawson, 1995).

1.4.2 Transferencia de modelos de Calibración con TR

El objetivo radica en actualizar un modelo existente a nuevas condiciones, como por ejemplo la aparición de nuevas especies con respuesta a nivel espectral o las derivadas del reemplazo de partes de un instrumento. En estas situaciones, **L** contendrá espectros bajo el nuevo escenario experimental, o derivaciones de éste. Cuando los espectros en **L** provengan de muestras en las cuales haya A/PI presentes, los valores de referencia serán necesarios en $\mathbf{y}_L$. Al aplicar TR, la estimación de **b** estará dirigida para ser también ortogonal a los interferentes espectrales de las nuevas muestras presentes en **L**, responsables de las predicciones inexactas para el A/PI. De esta manera, el modelo será desensibilizado respecto de los interferentes secundarios. En simultáneo, podría ser necesario que el nuevo vector prediga con exactitud a las muestras de Calibración originales que no tienen a los nuevos interferentes presentes. Dado que los objetivos incluyen actualizar a nuevas condiciones y reutilizar información valiosa previamente obtenida, esta exigencia extra de predecir correctamente a las muestras primarias puede dificultar la elección de un modelo final y quizá sea conveniente no sentar demasiada relevancia en esto, ya que de ser necesario predecir muestras primarias, bien podría utilizarse el vector de Calibración que se hubiera calculado antes de la aparición de los interferentes. Debe entenderse que, aun cuando las muestras originales no logren ser predichas con la misma performance con la que eran predichas originalmente, el objetivo de reutilizar su información estará cumplido de todas maneras y la

actualización tendrá lugar. Por otro lado, el cálculo de nuevos vectores deberá contemplar cuestiones relativas al tamaño de estos. Es esperable que, al aumentar un conjunto de Calibración primario con nuevas muestras de transferencia, la norma de los nuevos vectores (u otro indicador relativo al tamaño) tienda a crecer, siendo esto un efecto resultante del agregado de nueva información y de la necesidad de su contemplación con cierto grado de ajuste. Si el tamaño de los vectores se incrementa de manera abultada, aumentarán las probabilidades de obtener un modelo sobreajustado, con lo cual la varianza de las predicciones podría crecer también.

Si se cuenta con un subconjunto estable de muestras de la Calibración originales, las señales de estas muestras podrán ser obtenidas bajo las nuevas condiciones, formando el par primario/secundario. Similarmente, si las muestras no pertenecen al dominio calibrado originalmente, pero existe la posibilidad actual de realizar mediciones en dicho dominio (por ejemplo, se cuenta con el instrumento primario) y en nuevos dominios, también se podrá establecer el par primario/secundario. En ambos casos, **L** podrá contener diferencias de señales primarias y secundarias, al mismo tiempo que los valores en $\mathbf{y}_L$ no serán necesarios (en el primer caso planteado, igualmente serán conocidos por ser las muestras del dominio calibrado, mientras que en el segundo caso podrían ser totalmente desconocidos, siempre y cuando existan garantías sobre la utilidad de la información contenida en las muestras) y en su lugar se dispondrá de un vector de ceros de tamaño igual al número de diferencias que se hayan obtenido. En esta situación, la expresión (8) y las ecuaciones (3) y (9) se convierten, respectivamente, en:

$$\min\left(\left\|\mathbf{Xb}-\mathbf{y}\right\|_2+\lambda^2\left\|\mathbf{Lb}\right\|_2\right) \quad (10)$$

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix}=\begin{pmatrix}\mathbf{X}\\ \lambda\,\mathbf{L}\end{pmatrix}\mathbf{b} \quad (11)$$

$$\hat{\mathbf{b}}=\left(\mathbf{X}^t\mathbf{X}+\lambda^2\mathbf{L}^t\mathbf{L}\right)^{-1}\left(\mathbf{X}^t\mathbf{y}\right) \quad (12)$$

En el caso en que para **L** se utilicen muestras sin A/PI, como podrían ser blancos de muestra, solvente, espectros puros para componentes interferentes, entre otros, entonces la expresión (10) y las ecuaciones (11) y (12) también serán aplicables. Para corrimientos podría ser posible el uso de pseudo-espectros en **L**, como podrían ser constantes, rectas, parábolas o funciones de orden superior respecto de las variables. Esto ha sido aplicado con PCR, en cuyo caso los pseudo-espectros actuaron como pseudo-Componentes Principales en el set aumentado de autovectores (Vogt y col., 2000; Vogt y col., 2004). Similarmente, autovectores clave provenientes de la SVD (del inglés *Singular Value Decomposition*) de muestras con analito constante o bien ausente, tal como

espectros de una sola muestra medidos repetidamente, pudieron ser usados en **L** con $\mathbf{y}_L = \mathbf{0}$ (Wehlburg y col., 2002a; Wehlburg y col., 2002b).

Tal y como con cualquier método de regresión, se requiere que el vector sea ortogonal a la información en **X** que no sea del A/PI. De la expresión (10) y de la ecuación (11) es sencillo notar que el vector deseado también necesita ser ortogonal a las nuevas condiciones (químicas, físicas, ambientales, instrumentales, etc.) caracterizadas en **L**. Puede apreciarse que el producto $\lambda\mathbf{Lb}$ es igualado a cero, por lo cual el vector **b** será ortogonal a la información contenida en **L**. Uno de los objetivos radicará pues en lograr lo anterior con la menor cantidad posible de muestras en **L**.

Lo dicho para **L** e $\mathbf{y}_L$ también sería válido en el caso de tener una instrumento primario y varios secundarios (por ejemplo un laboratorio central mejor equipado y varios laboratorios satélites), y la decisión radicaría en conformar una actualización propia para cada instrumento secundario, o bien en que el conjunto aumentado en **L** pueda provenir de espectros de múltiples instrumentos en simultáneo.

1.4.3 Armonía como compromiso entre exactitud y precisión

Cuando **L** = **I**, y tanto *g* como *h* toman el valor 2 (normas euclidianas), se dice que TR está en su forma estándar, conocida también como RR (del inglés *Ridge Regression*), con la siguiente solución:

$$\hat{\mathbf{b}} = (\mathbf{X}^t\mathbf{X} + \lambda^2\mathbf{I})^{-1}(\mathbf{X}^t\mathbf{y}) \quad (13)$$

para la ecuación

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\mathbf{X}\\ \lambda\mathbf{I}\end{pmatrix}\mathbf{b} \quad (14)$$

o para la expresión

$$\min\left(\|\mathbf{Xb}-\mathbf{y}\|_2 + \lambda^2\|\mathbf{b}\|_2\right) \quad (15)$$

Debe apreciarse que en este caso λ actúa como regulador de la norma de **b**.

Lo descripto corresponde a la metodología con la que puede obtenerse un modelo de Calibración primario con RR, sin ningún tipo de transferencia (recordar que **L** = **I**, no hay información secundaria aún). En la expresión (15), es claro que el resultado será mínimo cuanto mejor sean las predicciones para **X**, y esa será una de las fuerzas impulsoras en el cálculo. Por otro lado y en relación al término derecho de la minimización, trabajos previos sugieren que la norma euclidiana del vector de regresión en la expresión (15) es proporcional a la varianza de las

predicciones (Forrester y Kalivas, 2004; Faber y Kowalski, 1996; Bechtel, 1997; Faber y col., 2003; Fernández Pierna y col., 2003). Entonces, en el caso de TR en su forma estándar, la optimización de la expresión (15) concierne a minimizar simultáneamente indicadores de exactitud y varianza en las predicciones futuras. Es esperable que estos criterios se comporten como dos fuerzas impulsoras de sentido opuesto. La exactitud en las predicciones futuras estará determinada por la exactitud con la que fueron predichos los valores Calibración. Si la última es escasa, no hay por qué esperar otra cosa para las nuevas muestras. Si la exactitud en las predicciones de Calibración es tenida en cuenta con mucha relevancia, será probable llegar a situaciones de sobreajuste. En este escenario, las predicciones de nuevas muestras podrían tener cierto grado de exactitud, pero ante pequeñas variaciones en los espectros existirán grandes variaciones en las predicciones. Esto es así porque si los modelos llegan a estar sobreajustados, los coeficientes de sus vectores no sólo habrán obtenido sus valores a partir de información útil y generalista, sino también a partir de detalles sólo presentes en los calibradores. Por lo tanto, la ausencia de estos detalles o su presencia parcial en las muestras futuras conllevará mayor variabilidad en las predicciones. Existirá por tal una región de exactitud intermedia que permitirá obtener predicciones aceptables en todos los casos y en ese sentido deberá estar orientada la optimización.

Una optimización de este tipo implicará por tal un equilibrio entre criterios opuestos de exactitud y varianza, donde la minimización de uno llevará al aumento del otro y viceversa. De entre todos los modelos posibles de ser obtenidos variando las exigencias sobre ambos criterios, el modelo buscado deberá ser el más armónico, o dicho de otra forma, deberá ser un modelo óptimo en términos de Pareto.

En una gráfica de Pareto (Kalivas y Green, 2001; Censor, 1977; Da Cunha y Polak, 1967) se evalúa más de un criterio de optimización en simultáneo y cada uno es representado convenientemente en un eje. Habiendo hallado más de una solución a un problema dado, los valores para cada criterio son graficados para cada solución. Si la optimización se corresponde con una minimización, existirá un borde o frontera que tenderá a acercarse a los ejes de coordenadas, conteniendo múltiples soluciones. Todas las soluciones presentes en un borde para este tipo de gráficas son denominadas superiores y óptimas en términos de Pareto (aunque solo unas pocas puedan ser consideradas armónicas), dado que no se podrían encontrar soluciones donde todos los criterios decrecieran en simultáneo, relativos a los valores de los criterios para otras soluciones del borde. O bien, dicho de otra manera, cualquier otra solución podría mejorar algunos criterios, pero al hacerlo indefectiblemente empeoraría en otros. Las soluciones a la derecha y arriba de un borde

de Pareto son denominadas inferiores, ya que sí sería posible encontrar otras soluciones en las cuales todos los criterios serían menores en relación a las soluciones inferiores (Kalivas, 2004). De entre todas las soluciones superiores de un borde, las armónicas serán aquellas que estén lo más cercanas posibles del origen de coordenadas.

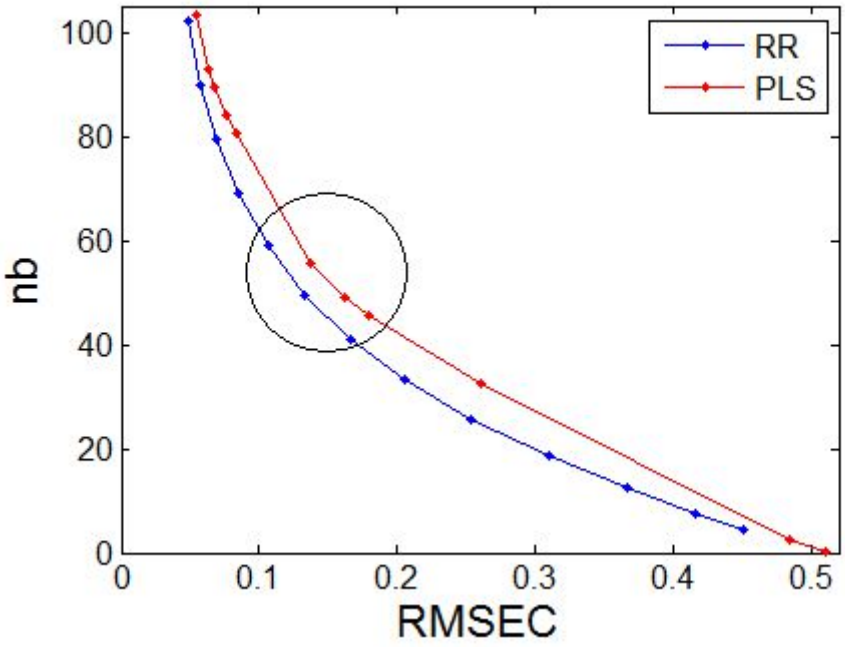


*Figura 1: Ejemplo de gráficas de Pareto para la evaluación simultánea de las cifras RMSEC (error en **X**) y nb (norma de los vectores de regresión **b**) calibrando con RR y PLS*

Referencias: Los puntos en la curva RR señalan modelos provenientes de distintos valores para el meta-parámetro de regulación de la norma. Los puntos en PLS señalan modelos con distinto número de Variables Latentes. El círculo negro señala una zona de interés para su análisis (ver texto).

En la figura 1 pueden observarse dos gráficas de Pareto, una para RR y otra para PLS, para un conjunto de Calibración determinado (del cual no vale la pena dar detalles en este momento). En la curva de PLS puede apreciarse el efecto clásico de elevar el número de Variables Latentes tenidas en cuenta. A medida que éste número aumenta, el error de Calibración (que bien podría haber sido otro como el clásico de Validación Cruzada) se hace cada vez menor, a la vez que la norma de los vectores resultantes tiende a crecer. También puede apreciarse que el cambio de RMSEC en relación al de nb no es constante, sino que depende de las Variables Latentes de los modelos comparados. Este tipo de curvas puede resultar útil a la hora de determinar el número óptimo de Variables Latentes para PLS (o de Componentes Principales para PCR). Similarmente, la curva de RR

muestra modelos que se diferenciaron en el valor utilizado para regular la norma de los vectores de regresión resultantes. Desde ya puede establecerse un paralelismo entre Variables Latentes para PLS y "meta-parámetro de regulación de norma" para RR, sólo que los últimos pueden tomar cualquier valor mayor que cero y formarán una curva más suave que la de PLS (tanto más suave cuanto mayor sea el número de valores reguladores probados). Comparando ambas curvas, se observan las mismas tendencias vistas para PLS. Lo importante en este estudio es destacar la zona remarcada con un círculo. Allí se señalaron los modelos que podrían considerarse como los más armónicos o con los mejores compromisos entre exactitud y varianza, y será la actualización de este tipo de modelos la que acuerde con los objetivos de este estudio. En ocasiones las curvas toman otras formas y a veces el frente tiene lo que comúnmente se denominaría "forma de L", en cuyo caso la zona armónica se identifica fácilmente como la más cercana al vértice del ángulo recto de una "L". Vale destacar que la expresión proviene solamente de la forma de la letra "L" y nada tiene que ver con el uso que se le ha dado a **L** hasta aquí (portador de información para actualizar los modelos).

Cuando se habla de cercanía al origen existe una dificultad importante asociada a la medida de distancia. Si no se tienen valores máximos y mínimos apropiados por cada eje, no es posible escalar los valores obtenidos (a no ser por una elección trivial). Al no poder escalar, la medida de distancia al origen será función de cada coordenada, y éstas podrían estar en unidades sin relación aparente (por ejemplo, la norma de un vector y su RMSEC no tienen las mismas unidades). Por lo tanto, las distancias en sí quizá no sean infalibles aproximadores de la armonía de un modelo. No obstante, a nivel gráfico y confiando en la capacidad visual del observador de estas curvas, será posible determinar al menos una zona aproximadamente armónica y desde allí debería extraerse un modelo apropiado según estos criterios.

1.4.4 Modificación de la TR para transferencia de modelos de Calibración: Doble Regularización de Tikhonov (DR)

En transferencia de Calibración, $\mathbf{L} \neq \mathbf{I}$ y las estructuras de **X** y **L** pueden tener un impacto no deseado en las ecuaciones (9) y (12). Específicamente, la operación de inversión no será estable y estará pobremente definida si los espectros en **X** y **L** son colineares y similarmente para el caso en que "*número de muestras*" << "*número de variables espectrales*". **X** y **L** serán casi singulares si el determinante de $\mathbf{X}^t\mathbf{X}$ o de $\mathbf{L}^t\mathbf{L}$ es cercano a cero, o bien, si el número de condición (*condition*

number) es significativamente grande, habrá un condicionamiento deficiente (Kalivas y Lang, 1994). El método RR expresado en la ecuación (13) está provisto de un mecanismo que fuerza a $\mathbf{X}^t\mathbf{X}$ para tener rango completo (*full rank*) a través de la adición de un número pequeño a la diagonal de $\mathbf{X}^t\mathbf{X}$, lo cual estabiliza la operación de inversión (Hoerl y Kennard, 1970). Siempre que λ no sea cero, las últimas *n* filas de la matriz aumentada en la ecuación (14) serán linealmente independientes, lo cual hará que la matriz aumentada tenga rango completo. Es decir, ya que **I** es una matriz diagonal de unos, al ser multiplicada por un número no nulo se estarán generando filas que serán todas distintas entre sí, puesto que sólo una variable será diferente de cero para cada fila. Cuanto mayor sea el valor de λ, mayor será el grado de ortogonalidad (o equivalentemente, de no-singularidad). Ya que $\mathbf{L}\neq\mathbf{I}$ para transferencia de Calibración, la estructura de **L** tiene un impacto significativo en la operación de inversión y si bien el problema seguirá siendo del tipo TR, ya no será en su forma estándar RR. Debido a este cambio en **L**, λ ya no será un regulador de norma, sino que será el ponderador del ajuste de las muestras en **L**. A nivel práctico, la inestabilidad en la inversión de matrices para distintos valores de λ en los diferentes modelos TR repercute de manera tal que amplios intervalos de valores puestos a prueba para λ son localizados en el mismo lugar de las gráficas de Pareto, como si fuesen modelos iguales aún cuando realmente no lo sean. En otras palabras, el efecto podría traducirse como una disminución de la sensibilidad al cambio de λ, plasmado en la superposición gráfica de modelos.

Por lo anterior y con el objetivo de estabilizar la operación de inversión, en estos casos será necesario contar con un meta-parámetro de regularización adicional. En el presente trabajo se presenta dicha adaptación a través de modificaciones en la expresión (8) y en las ecuaciones (3) y (9), las cuales se plantean, respectivamente, de la siguiente manera:

$$\min\left(\left\|\mathbf{Xb}-\mathbf{y}\right\|_2+\tau^2\left\|\mathbf{b}\right\|_2+\lambda^2\left\|\mathbf{Lb}-\mathbf{y_L}\right\|_2\right) \quad (16)$$

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\\ \lambda\mathbf{y_L}\end{pmatrix}=\begin{pmatrix}\mathbf{X}\\ \tau\mathbf{I}\\ \lambda\mathbf{L}\end{pmatrix}\mathbf{b} \quad (17)$$

$$\hat{\mathbf{b}}=\left(\mathbf{X}^t\mathbf{X}+\tau^2\mathbf{I}+\lambda^2\mathbf{L}^t\mathbf{L}\right)^{-1}\left(\mathbf{X}^t\mathbf{y}+\lambda^2\mathbf{L}^t\mathbf{y_L}\right) \quad (18)$$

donde τ hace referencia al nuevo meta-parámetro de estabilización, el cual mejorará el grado de no-singularidad para la matriz de covarianza en la operación de inversión, tal y como con RR. A su vez τ será el único encargado de la regulación de las normas vectoriales, mientras que λ estará dedicado únicamente a ponderar los aportes de la información secundaria de transferencia. Es

posible notar que el último término de la expresión (16) será proporcional al RMSEL, por lo que λ modificará la relevancia del error de predicción para las muestras en **L** respecto de las presentes en **X** (cuyo modificador será simplemente el escalar 1), lo cual dirigirá parcialmente la minimización planteada e indirectamente las tendencias finales a contemplar determinados tipos de información que tendrá el vector de regresión calculado en la ecuación (18).

Este nuevo planteo de TR es por tanto denominado Doble Regularización (DR) de Tikhonov, pues contendrá 2 meta-parámetros reguladores, y será en este estudio el mecanismo a través del cual se intentarán actualizar modelos primarios RR. Cabe destacar que en lugar de utilizar TR con este segundo meta-parámetro, PLS o PCR podrían ser aplicados directamente a las ecuaciones (3) y (11). Sin embargo, aún así se requeriría la determinación de 2 meta-parámetros (Variables Latentes o Componentes Principales, y λ).

Una inspección de las expresiones (8), (10), (15) y (16) proporciona un mayor entendimiento de las situaciones. Ya que **L** está compuesta de espectros (o diferencias derivadas), el significado físico del producto **Lb** en el último término de las expresiones (8) y (10) se corresponde con el de predicciones para **L** y, por lo tanto, no hay una regulación directa y explícita del tamaño del vector de regresión como en las expresiones (15) y (16). Por ende, la introducción del segundo meta-parámetro refuerza el control sobre la minimización.

Cuando $\mathbf{y}_L = \mathbf{0}$, la expresión (10) y las ecuaciones (11) y (12) serán ajustadas apropiadamente para tener en cuenta al segundo meta-parámetro. Sin embargo en esta situación una forma alternativa sería posible y el proceso consiste en transformar la forma general de TR en la expresión (10) a un formato tal como el de la expresión (15), es decir, llevar la resolución al campo de RR (Hansen, 1998). En este sentido, el vector de regresión puede ser obtenido usando un algoritmo de RR estándar (o PLS, PCR, etc.) y luego puede ser transformado a su forma general nuevamente (DiFoggio, 2005). El vector obtenido tras estos pasos estaría desensibilizado a los efectos presentados en **L** y podría ser utilizado para predecir nuevas muestras, aunque este mecanismo no será objeto de estudio.

En resumen, cuando **L** = **I** para RR, la operación de inversión es estabilizada con el uso de λ. Si en lugar de **I** se utilizan espectros en **L**, el segundo meta-parámetro τ será usualmente necesario para cumplir esa función. Mientras **X** y **L** se aproximen a ser matrices de rango completo, τ se hará cada vez menor y, en el límite donde $\mathbf{L}^t\mathbf{L}$ se aproxime a **I**, el valor de τ se aproximará a 0. Aún cuando con métodos como PLS, PCR o MLR no hay necesidad de regularización adicional por

medio de τ en las ecuaciones (3) y (11), otros meta-parámetros dependientes de cada método serán necesarios.

1.4.5 Generalización de la TR para transferencia de Calibración

Aunque en trabajos previos no han sido tenidos en cuenta los efectos pertinentes a la inestabilidad en la operación de inversión, sí han sido propuestas generalizaciones de TR para desensibilizar modelos respecto de múltiples efectos espectrales anticipados (Kalivas, 2004) (DiFoggio, 2007). El proceso consiste en usar un peso único para cada efecto presente en **L**. En ese caso, la ecuación (3) puede ser expresada de la siguiente manera:

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{\Lambda y_L} \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \mathbf{\Lambda L} \end{pmatrix} \mathbf{b} \quad (19)$$

donde $\mathbf{\Lambda}$ representa una matriz diagonal de tamaño l x l con los pesos λ_i correspondientes a cada efecto en **L**. Esta adaptación es aplicable a transferencia de Calibración, aunque como se ha dicho, no se han tenido en cuenta los problemas que pudieran surgir al intentar solucionar la ecuación (19). Cuando el término $\mathbf{\Lambda y_L} = \mathbf{0}$ y $\mathbf{L} = \mathbf{I}$, la ecuación (19) representa una generalización de RR (Hoerl y Kennard, 1970).

Si se requieren correcciones particulares para múltiples efectos, como desplazamientos, nuevos componentes químicos, o temperatura, entonces **L** debería estar compuesto de espectros que únicamente caractericen cada efecto y en tal caso la ecuación (19) podría ser usada. Al respecto, dos dificultades asociadas pueden identificarse. Una de ellas corresponde a obtener un protocolo para determinar cada λ_i en $\mathbf{\Lambda}$, mientras que la otra estará asociada a la necesidad de obtener espectros que caractericen solamente un efecto. Ante esto, debe considerarse el uso de un solo valor de λ, representando un compromiso o peso promedio para todos los efectos, que sería algo similar a la selección de Variables Latentes para PLS cuando la matriz **Y** contiene múltiples analitos.

1.4.6 Otros usos de la TR

El esquema general de TR expresado en la ecuación (3) ha sido útil en varias situaciones. Una matriz diagonal **L** con ruido espectral relativo a cada longitud de onda ha sido usada para remover variables irrelevantes del vector de regresión (DiFoggio, 2005; Stout y Kalivas, 2006). Ya que **L** era diagonal y por lo tanto de rango completo, el segundo meta-parámetro τ no fue necesario. TR también ha sido utilizado para suavizar **X**, así como también el vector de regresión. Una variante de

TR fue usada para problemas de resolución de curvas auto-modeladas (DiFoggio, 2005; Eilers, 2003; Gemperline y Cash, 2003). Reemplazando la norma euclidiana (norma-2) del vector de regresión con la norma-1 y ajustando **L** = **I**, se ha usado el procedimiento para selección de variables, lo cual se ha dado a conocer como Operador de Selección y Compresión Absoluta Mínima (LASSO) (Claerbout y Muir, 1973; Tibshirani, 1996; Stout y col., 2007). En un estudio limitado, una estimación de **b** fue ajustada al espectro del componente puro de un analito, con **L** = **I** en las ecuaciones (6) y (7) (Shih y col., 2007). Métodos relacionados a TR también han sido utilizados para clasificación de cáncer con datos de expresión de genes (Andries y col., 2007).

1.5 Materiales y Métodos

1.5.1 Software

Matlab (MATLAB *7.6.0*, 2008) fue utilizado como plataforma de cálculo y desarrollo. Como parte del presente trabajo fueron escritas las funciones para DR y RR, para diferentes estrategias de centrado y otras auxiliares para el análisis de resultados y la extracción de conclusiones. Funciones del denominado PLS Toolbox 3.52 (Wise y col., 2005) fueron utilizadas para el cálculo de modelos PLS y para el desarrollo de PDS durante algunas comparaciones. El método de Kennard-Stone (Kennard y Stone, 1969) para selección de muestras representativas fue aplicado a partir de funciones obtenidas del ChemoAC Toolbox 2.0 (Wu, 1998) y en cada caso donde fue necesario, se optó por obtener como primera muestra seleccionada a la más cercana al centroide de los datos disponibles.

1.5.2 Conjuntos de datos

Dos conjuntos de datos fueron utilizados en este trabajo, denominados Maíz y Temperatura. Con ambos conjuntos inicialmente se realizaron múltiples experiencias para evaluar los efectos provenientes del número de muestras presentes en **L** y del tipo de estrategia de centrado. Posteriormente se analizaron casos específicos con mayor detalle, con estrategias de centrado y número de muestras en **L** ya definidos.

En diferentes partes de este trabajo se utilizaron distintos subconjuntos de los datos (muestras de Calibración, transferencia y Validación), los cuales serán descriptos convenientemente cuando se describan las experiencias y se discutan los resultados. A continuación, una breve descripción

general de cada conjunto de datos.

1.5.2.1 Datos “Temperatura”

El conjunto está compuesto por 22 mezclas ternarias de agua, Etanol y 2-propanol. La figura 2, obtenida de (Wulfert y col., 1998) y posteriormente adaptada, detalla la composición de 19 muestras, todas ellas mezclas de al menos 2 de los componentes (las restantes 3 muestras corresponden a los analitos puros), en términos de fracciones molares porcentuales.

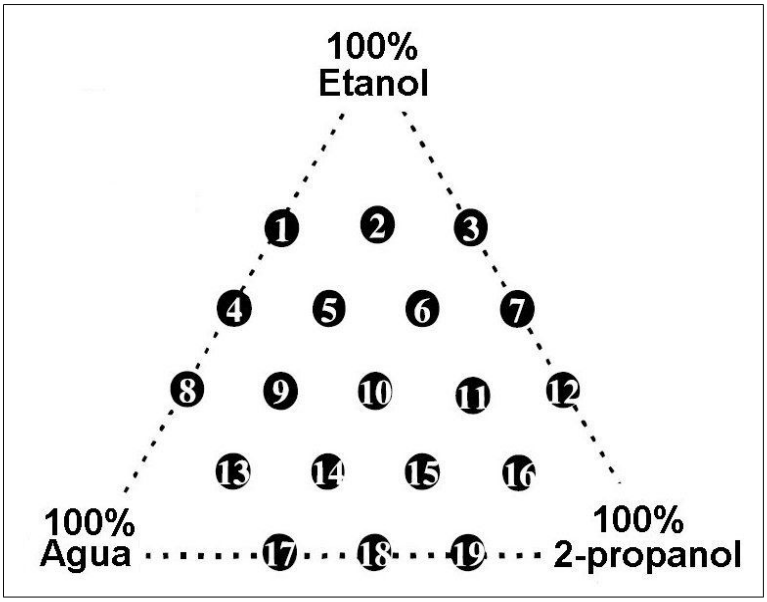


Figura 2: Diseño experimental para datos “Temperatura”. Fracciones molares porcentuales de Etanol, agua y 2-propanol.

Referencias: Las fracciones molares aproximadas de Etanol son de 0.66, 0.50, 0.33, 0.16 y 0 para las muestras 1-3, 4-7, 8-12, 13-16 y 17-19, respectivamente.

Las mediciones originales se realizaron en el intervalo 590-1091 nm, cada 1 nm, a las temperaturas de 30, 40, 50, 60 y 70ºC (Wulfert y col., 1998). En el presente trabajo sólo fueron usadas las variables entre 850 y 1049 nm, es decir, sólo se utilizaron 200 de las variables pertenecientes a los espectros NIR, al igual que en trabajos previos con este conjunto de datos, en los cuales también se encontró que la mayor sensibilidad de cambio con la temperatura ocurría fundamentalmente en longitudes de onda mayores a 950 nm (Wülfert y col., 2000a; Wülfert y col., 2000b; Marx y Eilers, 2002; Eilers y Marx, 2003). Ya que la absorción espectroscópica en IR depende de los modos vibracionales de las moléculas y dado que estos modos son afectados por fuerzas tales como los puentes de hidrógeno, los cuales son afectados por la temperatura, estos

espectros sirven como datos sensibles al cambio de temperatura y por ende son útiles en el estudio de transferencia de Calibración entre dos temperaturas. Específicamente, los datos obtenidos a 30ºC y 50ºC fueron considerados como primarios y secundarios, respectivamente. A su vez, en todas las pruebas realizadas el analito de interés fue Etanol. En la figura 3 se pueden observar los espectros de los componentes puros a distintas temperaturas en las longitudes de onda con mayor variabilidad, lo cual evidencia la influencia del cambio de temperatura en la absorción IR de los componentes y por ende en las mezclas.

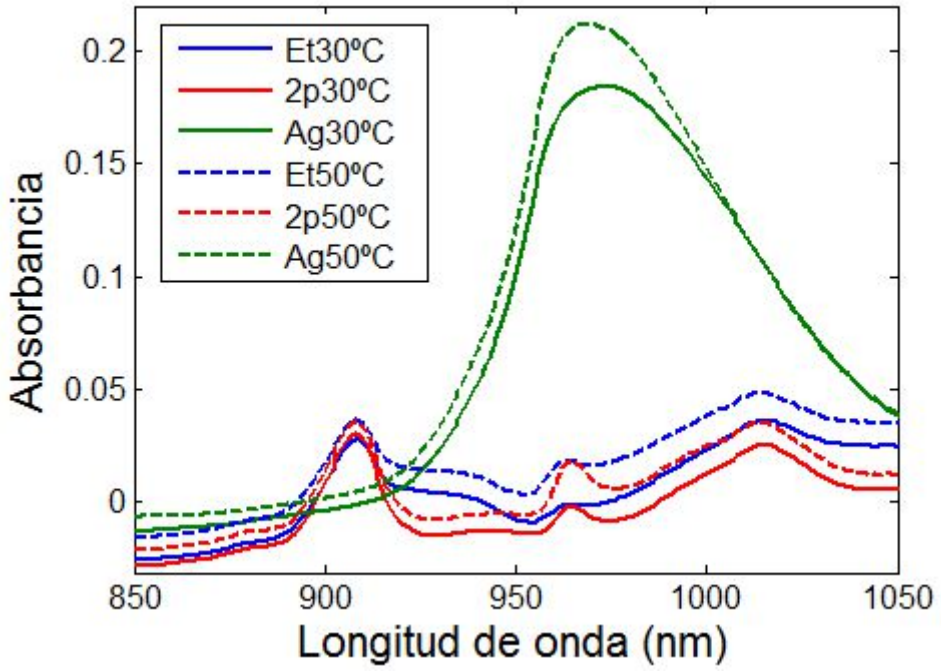


Figura 3: Espectros IR de los componentes puros para datos "Temperatura"

Referencias: Et: Etanol, 2p: 2-propanol, Ag: Agua

1.5.2.2 Datos "Maíz"

El conjunto está compuesto por 80 muestras de maíz medidas desde 1110 nm hasta 2498 nm en intervalos de 2 nm. Las mediciones se realizaron en 3 espectrofotómetros para IR cercano (NIR), denominados m5, mp5 y mp6 (Wise y col., 2005). Los datos están provistos con valores de referencia para contenido de aceites, proteínas, almidón y humedad. En el presente trabajo el contenido proteico, en un intervalo desde 7.654 hasta 9.711 % p/p, fue la propiedad de interés y, al igual que en otro anterior (Stout y Kalivas, 2006), sólo se utilizó una de cada dos variables disponibles, empezando por la segunda, dando un total de 350 variables. Respecto de los

instrumentos, m5 actuó como primario, mp5 como secundario, y mp6 no fue utilizado porque las señales obtenidas de este instrumento son similares a las de mp5. Adicionalmente, el conjunto posee mediciones de estándares de vidrio del NBS (*National Bureau of Standards* hasta 1988, actualmente conocido como *National Institute of Standards and Technology, NIST, Estados Unidos*). Específicamente, se cuenta con 3 repeticiones en m5 y 4 en mp5. En la figura 4 pueden observarse los espectros medios de las 80 muestras en ambos instrumentos, como así también los espectros medios de los estándares de vidrio.

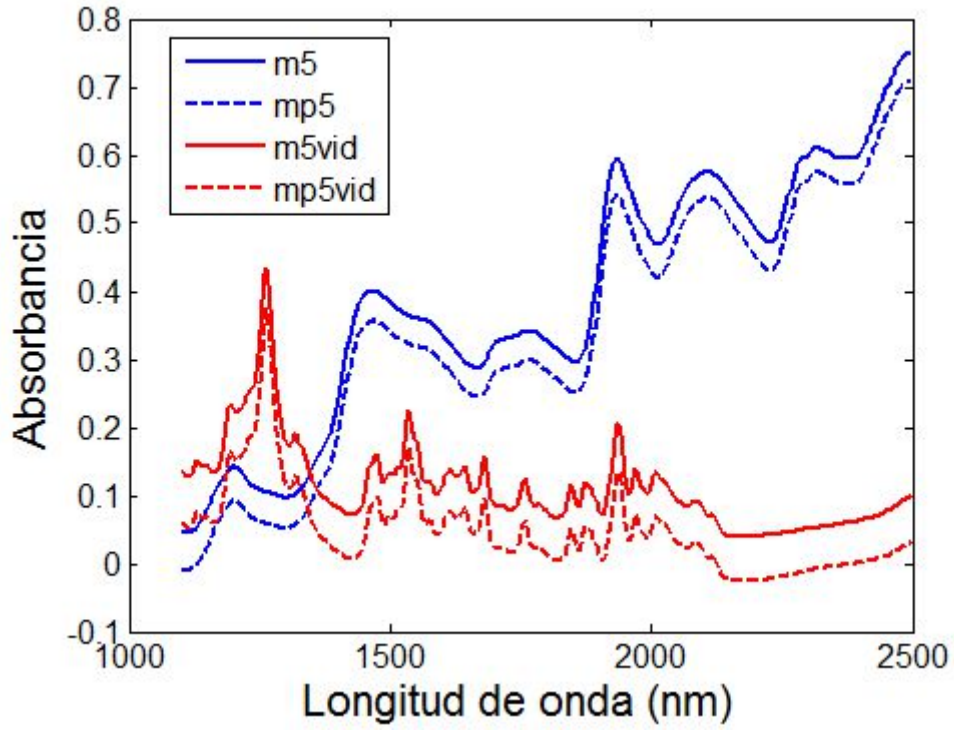


Figura 4: Espectros IR medios para las 80 muestras de datos "Maíz" y para los estándares de vidrio, en ambos instrumentos

Referencias: m5: Espectro medio de 80 muestras en el instrumento m5, mp5: Espectro medio de 80 muestras en el instrumento mp5, m5vid: Espectro medio de los estándares de vidrio en m5, mp5vid: Espectro medio de los estándares de vidrio en mp5.

En la figura 4 puede destacarse que existe fundamentalmente una deriva entre instrumentos. Dicha deriva no es constante al variar la longitud de onda de análisis (no mostrado). Cabe mencionar que este conjunto de datos puede obtenerse desde la siguiente dirección en Internet: http://software.eigenvector.com/Data/Corn/corn.mat.

1.5.3 Modos de transferencia con DR: SAC y DIFF

En el presente trabajo se estudiaron dos opciones para la introducción de información secundaria en **L**, con el fin de modificar dominios primarios previamente calibrados con **X** y así obtener nuevos vectores de regresión actualizados. Estas dos opciones han recibido las denominaciones SAC y DIFF. En ambos casos los datos son utilizados en la expresión (16) y en las ecuaciones (17) y (18). A continuación, una breve descripción de cada opción.

1.5.3.1 DR-SAC

Este esquema propone utilizar señales actuales o secundarias y sus valores de referencia para aumentar información primaria y, en ese sentido, no es más que un modelo mixto como alguno que podría obtenerse con PLS, PCR, u otro, aunque evidentemente esta mixtura tendrá características propias de DR, como ser la ponderación de la información en **X** y **L**, la restricción en la norma de los vectores, etc. La sigla SAC proviene de Espectros Y Concentraciones (del inglés *Spectra And Concentrations*), dado que en este trabajo las señales son espectros, aunque podrían ser otro tipo de señales. Vale destacar que el término "Concentraciones" hace referencia a analitos y no a propiedades de interés, aunque éstas últimas serán igualmente aplicables. Por lo tanto, para utilizar este esquema debe contarse tanto con las señales como con los valores de referencia para el A/PI en cuestión en las muestras de transferencia que conforman a **L**. Observando la expresión (16) puede notarse que λ es el meta-parámetro encargado de ponderar el error de predicción para las muestras en **L** por sobre el error de predicción de las muestras en **X**. Así por ejemplo, cuando λ sea 1, la minimización procederá sin ponderar especialmente el error de las muestras primarias ni el de las secundarias. Dado que será usual que el número de muestras en **X** sobrepase al correspondiente en **L**, valores de λ mayores que 1 intentarán darle mayor importancia al ajuste de la información de actualización para contrarrestar la disparidad numérica. Sin embargo, se verá posteriormente que este razonamiento, también válido para DIFF, no siempre será óptimo para predecir muestras incógnita del dominio secundario.

1.5.3.2 DR-DIFF

Este esquema propone utilizar diferencias (de allí el uso de DIFF, del inglés *Differences*) entre señales primarias (**L1**) y secundarias (**L2**) obtenidas de las mismas muestras en diferentes dominios para obtener a **L**, según **L** = **L2** - **L1**. Por ende, en este caso se asume una de dos opciones: o bien se

cuenta con muestras primarias estables, capaces de ser medidas en las nuevas situaciones secundarias, o bien se cuenta con muestras secundarias nuevas, capaces de ser medidas en la situación actual, pero también en la situación primaria (o al menos con la posibilidad de obtener la información previamente obtenida de esas muestras). En ambos casos podría o no contarse con el valor de referencia para el A/PI en las muestras que dieron origen a las diferencias. La principal ventaja del uso de diferencias entre señales es justamente la independencia respecto de los valores de referencia. Es decir, habiendo garantizado que las mismas muestras (o equivalentes desde algún punto de vista) hayan producido las señales primarias y secundarias, la diferencia entre señales, **L**, estará asociada a un valor de $\mathbf{y}_L$ que siempre será cero para cada diferencia en **L**. Por tal, contar con la información de los valores de referencia en dichas muestras no será de utilidad directa para encontrar los vectores de regresión por medio de DR-DIFF.

En este contexto, a diferencia del esquema en SAC, debe entenderse que la información que uno proporcionará a DR estará ligada a la fracción de las señales secundarias que no deberá ser tenida en cuenta en las predicciones futuras. Esto es, si se asume que la señal primaria es A, y que la secundaria es B = A + A', es decir, B es A modificada de alguna manera, el uso de diferencias estará haciendo hincapié en que A' proviene de un cambio en las señales, sea instrumental, físico, químico, u otro, y que no deberá ser tenido en cuenta pues proviene de información que no estará directamente relacionada al A/PI. Demás está decir que con el ejemplo anterior no debe entenderse a DR en su esquema DIFF como el uso de una simple resta, cual línea de base por ejemplo, sino como un protocolo que asume sus hipótesis partiendo de información diferencial y que, en base a ésta, realizará una manipulación de los datos un tanto más compleja que una resta. Claro está que el vector final de una DR tipo DIFF tendrá cierta potencialidad de distinguir, ante una señal secundaria, cuál es la información de interés y cuál no. Pero a menos que se pueda garantizar que el vector podrá distinguir entre información primaria y secundaria, el mismo vector tenderá a tener un defecto, relacionado a la información primaria y a encontrar en ésta algo para no tener en cuenta que, efectivamente, sí debería ser contemplado. Sin embargo, uno de los fines de DR es reutilizar la información en **X** aún a costa de aumentos razonables en el error de sus predicciones, pues para estas señales existe el modelo primario.

Durante el desarrollo posterior del trabajo, se verá que una estrategia de centrado apta para DIFF se denominará MC1. A su vez, con uno de los conjuntos de datos del trabajo ("Maíz") se tendrá el caso en el cual la transferencia tipo DIFF se realizará a partir de estándares de vidrio medidos en 2 instrumentos, sin utilizar muestras reales primarias y secundarias. Sólo por el caso de

los estándares de vidrio es que se utilizará MC1 para DIFF, pues esta estrategia no requiere los valores de referencia que, evidentemente, no tendrán los estándares de vidrio. El resto de las experiencias con DIFF se realizará con muestras reales. Por consiguiente, si se sabe que dichas muestras han sido medidas en 2 dominios distintos (instrumentos o temperaturas, según el conjunto de datos), es ilógico pensar que no se contaría con sus valores de referencia para sus A/PI. Por consiguiente, a pesar de que la principal ventaja de DIFF es su capacidad de no requerir valores de referencia, dichos valores podrán ser utilizados y esto será útil por 2 razones. La primera es que se podrá establecer una variante de MC1 utilizando esos valores. La segunda es que ante la necesidad de determinar un valor de λ óptimo, las muestras secundarias (**L2**), que junto a sus equivalentes primarias (**L1**) hayan dado origen a diferencias utilizadas en las transferencias (**L = L2 - L1**), podrán ser procesadas tal cual se haría para una muestra secundaria futura, es decir, centrándolas con lo que haya sido determinado como óptimo e introduciéndolas en el modelo que fue optimizado a partir de diferencias. De la operatoria anterior será posible obtener qué valor de λ resultaría apropiado para todo el conjunto de Validación secundario, asumiendo que las muestras en **L2** y sus valores de referencia serían lo suficientemente representativos de ese dominio. Se aclara que esto podrá realizarse porque la información en **L2** no participa directo de la elaboración de los modelos, sino a través de sus diferencias con **L1**. En cambio, en la estrategia SAC **L2** es en sí lo que actualizará a los modelos, por lo que a partir de su propio ajuste no será posible determinar qué valor de λ sería apropiado para todo el dominio secundario, ya que es de esperarse un ajuste mayor para **L2**.

1.5.3.3 Breve resumen comparativo entre SAC y DIFF

Operativamente, SAC y DIFF presentan ventajas y diferencias, mutuas y excluyentes. Comparten el problema de la representatividad de las muestras en **L** en relación a todo el dominio secundario, no sólo por el hecho de que no habrá manera de garantizar estrictamente que la información allí provista representará bien a cualquier información secundaria futura (un problema que acontecerá también al usar algoritmos de Calibración convencionales con la información en **X**, entre otros), sino también por su intrínseca naturaleza “ahorrativa”, con la cual muy pocas muestras deben representar a un todo a veces mucho mayor en cantidad. Ambos, claro está, encuentran una de sus ventajas exactamente en eso, pues si el proceso se realiza de manera exitosa, uno habrá precisamente ahorrado recursos. Teniendo en cuenta que podría ser normal esperar que datos obtenidos en el pasado (**X**, primarios) estén disponibles en los registros de un laboratorio para

realizar nuevos cálculos, el requerimiento experimental de SAC se limitará simplemente a obtener señales para muestras en las nuevas situaciones, actuales y por ende accesibles. Como desventaja, deberán invertirse recursos para hallar los valores de referencia de las muestras en **L**. Por otro lado, DIFF presenta la ventaja de no necesitar estas determinaciones. Sin embargo, ya que es necesario equiparar señales secundarias con primarias, habrá que tener acceso a las últimas. Una opción radicará en confiar en la estabilidad de las muestras originales (si es que se conservan aún) o en la inalterabilidad de estándares que pudieran ser medidos nuevamente en las condiciones actuales. Otra opción, en algunas ocasiones más propensa a lo fortuito y en otras limitada a casos de muestras y/o relaciones poco complejas, será obtener nuevas muestras secundarias que, por casualidad o por diseño, respectivamente, hayan estado presentes en conjuntos de señales primarias obtenidas con anterioridad. Se insiste en que si se desea reproducir una situación antigua, necesariamente deberá existir simpleza en las actividades requeridas. Por ejemplo, uno podría re-elaborar mezclas del conjunto de datos “Temperatura”, porque pertenecen a una mezcla ternaria de componentes simples (demás está decir que una real reproducción será imposible y que uno aceptará la igualdad dentro de ciertos límites experimentales), pero difícil sería reproducir mediante síntesis la composición de muestras del conjunto “Maíz”, dada la mayor complejidad de un vegetal. Al mismo tiempo, la relación modelada limitará la posibilidad de reproducción. Es decir, será más fácil hacer una solución y decir “esta solución, al igual que una obtenida en el pasado, contiene A% de agua, B% de Etanol y C% de 2-propanol y, por ende, puede equiparársela a una muestra original”, que decir “esta muestra de maíz, al igual que una obtenida en el pasado, tiene X% de humedad y, por ende, puede equiparársela a una muestra original”, siendo que múltiples muestras, muy diferentes entre sí en su composición, podrían tener ese % de humedad.

También en relación a los valores de referencia, se realizaron experiencias en las que se suponía estaban disponibles tanto para DIFF como para SAC (no fue el caso de los estándares de vidrio), por lo cual era posible plantear una estrategia mixta, presentando simultáneamente en **L** tanto a las señales como a sus diferencias (asociadas a sus concentraciones y a ceros, respectivamente). Los resultados no serán expuestos, pero se comenta que no fueron necesariamente mejores que los obtenidos con las estrategias por separado y que incluso su interpretación se tornó más dificultosa.

1.5.4 Cifras de mérito

Fundamentalmente se utilizaron 2 cifras de mérito para evaluar a todos los modelos calculados:

– RMSE (del inglés *Root Mean Square Error*): Son las Raíces cuadradas de los Errores Cuadráticos Promedio. Distintos RMSE fueron calculados específicamente para diferentes subconjuntos de muestras, siempre según la siguiente ecuación:

$$\mathrm{RMSE} = \sqrt{\frac{\sum (\mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i})^2}{n}} \quad (20)$$

donde $\hat{\mathbf{y}}_i$ representa las predicciones de determinadas muestras, $\mathbf{y}_i$ contiene a los valores de referencia de las mismas muestras y n es el número de muestras del subconjunto al cual pertenecían las muestras en cuestión.

- Los RMSE específicos más utilizados y los subconjuntos de muestras que les dieron origen son:
  - RMSEC (RMSE de Calibración): Relacionado a los errores en las predicciones de las muestras primarias (**X**) que participaron de cada modelo
  - RMSEL (RMSE de **L**): Relacionado a los errores en las predicciones de lo contenido en **L**, sean espectros secundarios (SAC) o diferencias espectrales (DIFF). Se hace hincapié en que en el último caso RMSEL se corresponderá con el ajuste de las diferencias a sus valores de referencia, los cuales siempre serán de 0.
    - En los casos donde se expongan errores de predicción para las muestras secundarias (**L2**) que dieron origen a diferencias utilizadas en **L**, los RMSE pertinentes serán explicitados oportunamente.
  - RMSEV (RMSE de Validación): Relacionado a los errores en las predicciones de las muestras secundarias de Validación (**V2**), las cuales serán definidas oportunamente según las experiencias reportadas.
- Otros RMSE particulares podrán ser definidos durante el desarrollo del escrito.

– nb: Corresponde a la norma-2 o Euclidiana de los vectores de regresión (**b**) de cada modelo:

$$\| \mathbf{b} \|_2 = \sqrt{\sum \mathrm{b_i^2}} \quad (21)$$

donde $\mathrm{b_i}$ representa cada coeficiente de regresión en **b**.

1.5.5 Meta-parámetros τ y λ

Sin importar en qué experiencias fueron utilizados los meta-parámetros, los resultados de DR provendrán de la ecuación (18), donde puede observarse que los meta-parámetros se encuentran elevados al cuadrado. Siempre que en cualquier experiencia se utilizaron valores de prueba, éstos fueron directamente provistos a las rutinas de cálculo como valores ya elevados al cuadrado. Por otro lado, la minimización en la expresión (16) es más sencilla o intuitiva para interpretar los roles de los meta-parámetros como ponderadores, por lo que dicha expresión resultará útil durante el análisis de resultados y será recurrente referirse a ésta. Por lo tanto, en todos los casos desde aquí en lo sucesivo, cada vez que se haga referencia a alguno de los valores ya elevados al cuadrado éstos serán denominados como "tau" (para τ^2) y "lam" (para λ^2).

En el presente trabajo se realizaron experiencias múltiples con componentes pseudo-azarosos y posteriormente se analizaron casos puntuales.

En las ejecuciones múltiples el objetivo fue evaluar efectos provenientes del número de muestras en **L** y del tipo de centrado. En este caso se pusieron a prueba varios valores de tau con el objeto de obtener vectores con valores distintos de norma para evaluar los efectos sin tener por objeto seleccionar un tau específico. Los valores de tau fueron obtenidos partiendo de la matriz de Calibración primaria centrada, de la cual se obtuvo la matriz de covarianza, y luego ésta fue sometida a una Descomposición en Valores Singulares (SVD). Los valores singulares resultantes fueron utilizados en una función que selecciona N valores de prueba de manera exponencial en decaimiento, incluyendo al valor cero. Aunque de esta manera los valores dependerán de los datos primarios, en general su máximo no superará la unidad y si se tienen varios valores de prueba entre cero y uno, será suficiente para encontrar una zona apta donde los nuevos modelos puedan ser evaluados. Vale destacar que esta forma de obtener los valores no es importante en sí, pues el objetivo consiste en obtener una secuencia apropiada para las pruebas, de manera tal que al variar tau existan cambios perceptibles y útiles en los modelos, y que la zona armónica esté representada. Los valores de lam puestos a prueba fueron obtenidos con una función que recibe un valor máximo, un mínimo y un número deseado de valores de prueba entre ambos. La función genera tantos valores como el número requerido, entre el máximo y el mínimo, también en decaimiento exponencial. Originalmente se utilizaron valores entre 10^9 y 0, aunque posteriormente se decidió reportar solo los resultados provenientes de una parte del intervalo anterior (más detalles serán dados oportunamente). Claro está, aquí también podría haberse utilizado otra lógica en la obtención

de los valores de lam.

Similarmente, durante el análisis de casos específicos se probaron varios valores de tau. Estos fueron coincidentes con algunos de los valores de las experiencias múltiples, sólo que entre éstos se adicionaron más valores intermedios para aumentar la resolución. Sin embargo, en estas experiencias específicas sí fue objetivo seleccionar un único valor de tau sobre el cual evaluar las optimizaciones. Este valor para tau quedará determinado por una elección previa del mismo meta-parámetro en una situación primaria hipotética. El criterio de elección de un tau definitivo estuvo basado en la armonía entre error de predicción RMSEC (u otro similar que se indique) y norma de los vectores de regresión, es decir, el tau seleccionado debe generar un modelo armónico. Por el lado de lam, los valores fueron obtenidos mediante una estrategia distinta a la aplicada en las ejecuciones múltiples, lo cual también será detallado oportunamente. A su vez, en determinados casos fue conveniente agregar valores de lam previamente no generados por la estrategia en cuestión. En dichos casos, los valores agregados serán informados mediante la notación "inicio:salto:fin" (por ejemplo, 4:-2:0 generaría los valores 4, 2 y 0).

Las salvedades anteriores provienen del hecho de que al utilizar en simultáneo muchos valores de tau y lam, el análisis de las gráficas se dificulta notablemente. Por lo tanto, en ocasiones es conveniente poner a prueba pocos valores, y posteriormente agregar y analizar más en ciertas zonas de interés.

Finalmente, es conveniente aclarar que la determinación de valores óptimos para los meta-parámetros podría realizarse en base a los resultados obtenidos de un subconjunto de muestras destinadas pura y exclusivamente a tal fin, que no hubieran participado directamente de las etapas de cálculo sino solamente de la evaluación (y posiblemente refinamiento) final de un modelo definido solamente por las muestras en **X** y **L**, y por sus respectivos valores de referencia. En dichos casos se supondrá que la información contenida en las muestras de transferencia sería lo suficientemente representativa de las nuevas situaciones a modelar. Sin embargo, en el contexto de este escrito una de las premisas está relacionada al ahorro de recursos necesarios para realizar las transformaciones. De aquí se deriva que en las situaciones supuestas no sería posible obtener demasiada información experimental nueva, y que esta información podría ser solamente utilizada en **L**, o bien parte de esta podría ser reservada para actuar fuera de **L** en la selección de meta-parámetros. Con el fin de suponer una situación de grandes limitaciones, en este trabajo se decidió que toda la información nueva disponible sería escasa y sólo utilizada en **L**. A su vez, en distintas experiencias fueron siendo conformados conjuntos de muestras secundarias denominadas

genéricamente de “Validación”. La evaluación de los resultados provenientes de la variación en los meta-parámetros se realizará priorizando las mejores predicciones para los datos de “Validación”, pero no se supondrá que éstos estarían realmente disponibles (si lo estuvieran, podría optarse por re-calibrar en las nuevas condiciones), sino que éstos representarían muestras incógnita futuras. Por lo tanto, se intentará evaluar si las evoluciones observadas en las curvas de muestras incógnita potenciales podrían ser predecibles en función de cómo estén variando los meta-parámetros, las normas de los vectores y los errores de ajuste para **X** y **L**, siendo que todos estos valores serían realmente conocidos. Quedará a voluntad de quienes decidan aplicar DR definir cuántas muestras realmente estarían destinadas a determinar valores óptimos para los meta-parámetros.

1.5.6 Estrategias de centrado

Aunque hubiese sido posible explorar el efecto de múltiples estrategias de pretratamiento para los datos, en el presente trabajo sólo se han evaluado cuatro estrategias de centrado en común para todas las aplicaciones de DR (en determinado estado de avance del desarrollo se derivará una más). No obstante, en su debido momento se hará evidente la importancia de estas estrategias. El objetivo de estas evaluaciones está relacionado a la obtención de resultados pseudo-óptimos a nivel general tanto para DIFF como para SAC, por lo cual los diferentes centrados fueron puestos a prueba en las ya mencionadas experiencias múltiples. Una vez que se pudo determinar cuáles eran los centrados más apropiados, sólo éstos fueron evaluados en el análisis de casos puntuales.

Debe notarse que, a excepción de los valores de los meta-parámetros que son impuestos, los centrados determinan qué valores son usados realmente en la expresión (16) y en las ecuaciones (17) y (18) para todas las matrices y vectores presentes, ya que hasta este punto del desarrollo no se habían asumido pretratamientos para **X**, **L** y sus respectivos valores de referencia.

A continuación, una breve descripción de cada estrategia de centrado.

1.5.6.1 MC1 (sin **L**)

Los datos en **L** e $\mathbf{y}_L$ no son modificados antes de ser introducidos en la etapa de cálculo. A su vez, en **L** sólo habrá diferencias de espectros e $\mathbf{y}_L$ será solamente un vector de ceros del tamaño apropiado, por lo que no son necesarios valores de referencia para $\mathbf{y}_L$. Por otro lado, **X** e **y** son centrados a sus respectivas medias y éstos valores son usados para centrar a los espectros incógnita

secundarios y para escalar a las predicciones de sus concentraciones una vez hecho el cálculo. Es decir, medias primarias son utilizadas para centrar datos secundarios, ya que éstos no pueden ser centrados con medias de su propio dominio porque se asume que no se tienen valores de referencia secundarios. Dado que **L** e $\mathbf{y}_L$ no sufren cambios, los valores allí presentes deberán ser compatibles en magnitud con los de **X** e **y** ya centrados. Por ende, este centrado debería ser fundamentalmente útil para DIFF.

1.5.6.2 MC2 (Clásico)

X e **y** son centrados a sus respectivas medias y éstos valores son usados, respectivamente, para centrar a espectros incógnita y a **L**, y para escalar a las predicciones de éstos luego del cálculo. Esta estrategia es similar a la convencionalmente utilizada en modelos de Calibración, donde las medias sólo se obtienen de ese conjunto y cualquier otro dato deberá ser referido a esos valores. Este centrado será el utilizado cuando se evalúen los modelos primarios que hipotéticamente hubiesen comenzado a fallar.

1.5.6.3 MC3 (Local)

X e **y** son centrados a sus respectivas medias, mientras que **L** e **yL** lo harán respecto de las suyas. Este centrado de tipo local ha mostrado mejorías de performance en otros trabajos (Du y col., 2005; Kalivas, 2008). Por otro lado, los espectros del dominio secundario son centrados con la media de **L** y sus predicciones escaladas con la media de $\mathbf{y}_L$ (ambos en relación con dicho dominio). Si hubiera espectros incógnita primarios, serían centrados con la media de **X** y sus predicciones escaladas con la media de **y**, aunque en este trabajo no se hará alusión a tales datos incógnita primarios.

A diferencia de las restantes estrategias de centrado, esta estrategia estará limitada al uso de más de una muestra en **L**. Esto puede entenderse observando la ecuación (18), en la cual se calculan los vectores de regresión, recordando que donde se indica **L** e $\mathbf{y}_L$ en realidad se estarán usando sus versiones centradas. Si se usara una única muestra de transferencia y ésta fuera centrada a su propia media, tanto **L** como $\mathbf{y}_L$ serían convertidas a valores de cero al ser centradas (para **L** sería un vector de ceros, mientras que para $\mathbf{y}_L$ sería el escalar cero). Lo anterior daría como resultado un caso puntual de la ecuación (18):

$$\hat{\mathbf{b}} = (\mathbf{X}^t\mathbf{X} + \tau^2\mathbf{I} + \lambda^2\mathbf{0}^t\mathbf{0})^{-1}(\mathbf{X}^t\mathbf{y} + \lambda^2\mathbf{0}^t 0) \quad (22)$$

o bien:

$$\hat{\mathbf{b}} = (\mathbf{X}^t\mathbf{X} + \tau^2\mathbf{I})^{-1}(\mathbf{X}^t\mathbf{y}) \quad (23)$$

La ecuación (23) revela información importante. En primer lugar que será inútil la variación de lam, en segundo lugar que **L** ni $\mathbf{y_L}$ aportan información en la etapa de cálculo (y por eso MC3 no se puede evaluar con una sola muestra en **L**) y en tercer lugar que los vectores obtenidos sólo dependerán de tau y de información primaria. Precisamente esa ecuación determinará cuáles hubiesen sido los modelos primarios del tipo RR que comenzaron a fallar, de los cuales solamente algunos para determinados tau hubiesen sido considerados armónicos y aplicables. El único aporte de información que provendría de **L** e $\mathbf{y_L}$ se daría en le etapa de centrado de espectros incógnita secundarios. Esto se realizaría a través del aporte de sus medias (que por ser muestras únicas serían sus mismos valores y probablemente no serían representativos de la media secundaria general) y luego en el escalado de las predicciones, pero los vectores en sí sólo provendrían de información primaria.

1.5.6.4 MC4 (Mixto)

X y **L** son reunidos y de éstos se extrae una media global. Lo mismo se hace con **y** e $\mathbf{y_L}$. Los espectros medios son utilizados para centrar a **X**, a **L** y a los espectros incógnita secundarios, mientras que las concentraciones medias son usadas para centrar a **y** e $\mathbf{y_L}$, y para escalar a las predicciones de las incógnitas.

1.6 Resultados y Discusión

A continuación se expondrán resultados de experiencias múltiples con muestras de transferencia al azar. El objetivo de éstas será evaluar efectos de la aplicación de distintas estrategias de centrado y efectos provenientes del cambio en el número de muestras en **L**.

Posteriormente se expondrán resultados de análisis de casos puntuales, y en estos casos el objetivo será apreciar detalles específicos para poder relacionarlos con las condiciones impuestas en los cálculos.

1.6.1 Experiencias con ejecuciones múltiples: muestras de transferencia al azar

Originalmente se pusieron a prueba intervalos con muchos valores para lam y tau, específicamente 50 valores por meta-parámetro. Sin embargo, el análisis de los primeros resultados obtenidos indicó que dichos intervalos no serían de utilidad en su totalidad, ya que creaban confusión durante su análisis y aún más cuando las experiencias analizadas eran múltiples. La figura 5, a modo de ejemplo, reporta resultados promedio de 30 ejecuciones de DR tipo SAC para datos "Temperatura" con 4 muestras en **L** y MC3, conservando todo el intervalo de lam y solamente 11 valores para tau. En cada ejecución las muestras para **L** fueron seleccionadas al azar, lo cual será descripto oportunamente fuera de este ejemplo.

En la figura 5 puede observarse cómo evolucionaron las cifras de RMSE y la norma de los vectores ante la variación de los meta-parámetros. Como puede apreciarse, sólo con los tau utilizados ya es suficiente para obtener curvas en zonas de escaso poder predictivo (normas bajas), en zonas armónicas (normas y errores medios) y en zonas de potenciales sobreajustes (normas altas). A su vez, el principal objetivo de la figura 5 es exponer el efecto de utilizar un gran intervalo para lam. En la gráfica de RMSEV se destacó una región en particular (zoom) y allí los rectángulos rojos han sido posicionados sobre modelos específicos, los cuales están asociados al número más cercano y corresponden a determinados valores de lam en términos ordinales. Por ejemplo, la curva negra tiene señalados 3 puntos en particular. El número 50 indica la menor de todas las lam, que en dicho caso fue de 3.552×10^{-6}. Se aprecia que el RMSEV es el más alto para esa curva, lo cual parece bastante lógico en este punto del desarrollo puesto que el valor de lam es muy bajo, por lo cual la información aportada por **L** básicamente no está contemplada y esto repercute en

predicciones con gran error medio para las muestras que conformaron a **V** (de forma análoga, RMSEL obtiene sus valores más altos en dichos valores de lam, cuyos modelos siempre serán los que, para cualquier tau, estén más hacia la derecha del gráfico respectivo).

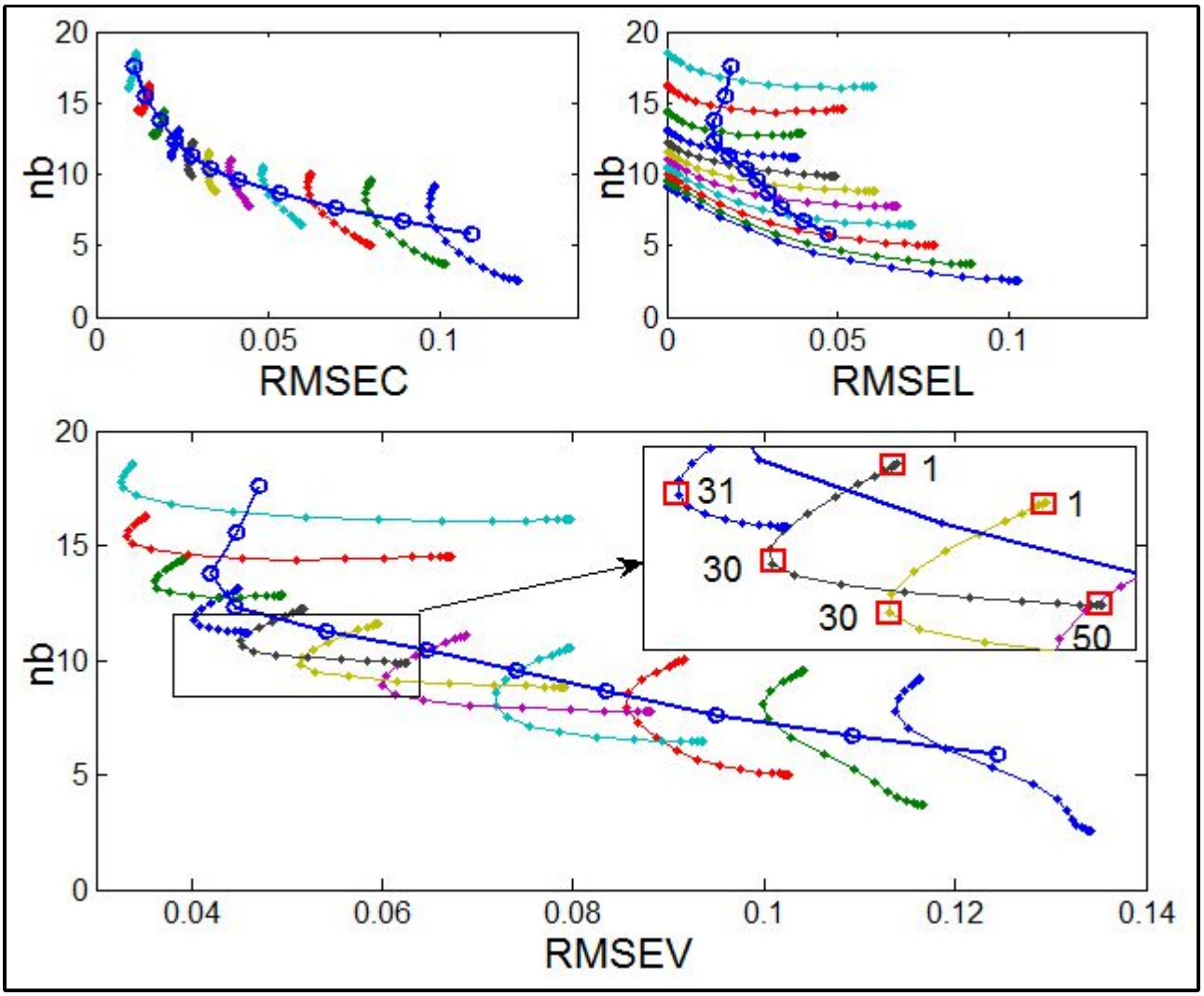


Figura 5: Promedio de RMSE (C, L y V) versus promedio de norma de los vectores de regresión (nb) para datos "Temperatura" en 30 ejecuciones DR-SAC-MC3 con 4 muestras de transferencia seleccionadas al azar en cada ejecución

Referencias: Cada color representa un tau (11 en total), cada punto en cada tau representa el promedio de las 30 ejecuciones en cada uno de los valores de lam (50 en total). La curva azul con círculos presente en todas las gráficas corresponde al dato medio para las 50 lam en cada tau en las 30 ejecuciones. Lo contenido dentro del zoom en la gráfica inferior es de interés para su análisis (ver texto).

Otro de los puntos destacados en la misma curva corresponde a lam ordinal 1, con un valor de $1x10^9$, el cual también fue destacado en la curva amarilla visible en el zoom.Este fue el máximo de los valores puestos a prueba y se supone que con dicho valor se obtendría una ponderación extrema en la predicción correcta de las muestras en **L**. Como resultado, la minimización estará dirigida fundamental y exageradamente por las muestras de transferencia (por lo cual los errores en las curvas de RMSEL llegan básicamente a cero, situándose siempre a la izquierda de esas gráficas), con lo cual también se pierde generalización y por lo tanto se obtienen cifras de RMSEV que indudablemente no serán las mínimas posible de obtener. Estos resultados mínimos han sido destacados en las 3 curvas del recuadro, señalados con los valores ordinales 30 y 31, que equivalen a 1.8626 y 0.9313, respectivamente. Como puede apreciarse, de la gran cantidad de valores de lam puestos a prueba, los mejores resultaron ser aquellos cercanos al valor de 1. Retomando la expresión (16), cuando lam toma el valor 1 la minimización pondrá el mismo énfasis entre las predicciones de **X** y las de **L** (sin tener en cuenta el desnivel en el número de muestras de cada una, que fueron de 10 y 4, respectivametne), ya que en ese caso ambos errores de predicción estarían siendo afectados por el mismo coeficiente, la unidad.

En relación a **X**, conviene destacar la diferencia de amplitud entre RMSEC máximos y mínimos para cada tau. En los tau de las zonas inferiores la amplitud es mayor y decrece al ir hacia las zonas medias, lo cual indica que el grado de ajuste a **X** cuando lam varía no siempre será el mismo, y a su vez que en las zonas medias (y muchas veces también en las zonas superiores) suele apreciarse que **X** se mantiene básicamente inalterado en cuanto a ajuste, más allá de la variación en lam. Esto último se traduce en que muchos puntos que deberían verse más separados se vean unidos en las curvas de los tau medios y superiores, y en que dichas curvas sean prácticamente verticales, es decir, básicamente lo que se modificará será la norma.

Respecto de la modificación en las normas, una tendencia en particular que puede ser observada en las gráficas anteriores y que se repitió en muchas otras experiencias (no en todas las realizadas), algunas de las cuales serán reportadas, es el ascenso de la norma vectorial con el aumento de los valores de lam. Si bien tau es el principal regulador de dicha norma, ésta tenderá a crecer si eso implica lo óptimo para la minimización de la expresión (16) y claro está que dicho óptimo dependerá también de lam. Al crecer el valor de lam, el error para las muestras de transferencia que posibilitará un mínimo de la expresión tenderá a ser cada vez menor. Para lograrlo será necesario que los coeficientes del vector de regresión sean modificados, de forma tal que cada vez sean apreciados con mayor relevancia los detalles espectrales de las muestras de transferencia y en

especial su relación con sus valores de referencia. Cuanto mayor sea el nivel de detalle y ajuste a los valores de referencia, más necesario será que los coeficientes sean adaptados. Cuando se observa cómo son modificados los gráficos de los vectores de regresión para un determinado tau a medida que cambia lam (no mostrado), se aprecia que los coeficientes relacionados a zonas de variables cercanas varían en conjunto, de lo cual se deduce que estas adaptaciones son realizadas para contemplar en mayor medida unos detalles por sobre otros. Así, al cambiar lam, algunas zonas decrecen en sus coeficientes (en valores absolutos) mientras que otras crecen, o viceversa, según como esté variando lam. La tendencia observada indica que al aumentar lam los coeficientes que crecen en valor absoluto lo hacen en mayor medida respecto de lo que disminuyen los que decrecen y como resultado las normas suelen elevarse al aumentar lam.

También vale destacar que donde fueron marcados los valores extremos y ordinales de lam 1 y 50 existen en realidad múltiples puntos casi superpuestos (en las curvas de RMSEC y RMSEL se aprecia lo mismo), aunque no de forma exacta ya que con mayores ampliaciones pueden observarse pequeñas diferencias (datos no mostrados). Estos puntos concentrados, visibles en cualquier tau, son mayoría y hacen que la distribución de resultados se aleje de la normalidad, por lo que los datos medios se proyectan en zonas donde no hay ningún punto individual (comparar la curva de cualquier tau con su respectivo dato medio en la curva azul con círculos). Ante esto, se hace necesario restringir el intervalo de lam para que los datos medios tengan sentido real.

Por lo expuesto, a pesar de que se realizaron experiencias con intervalos extensos para los meta-parámetros, los análisis posteriores contemplarán cantidades similares a las expuestas para tau, y valores de lam cercanos a 1. Todos éstos valores se especifican en la tabla 1 para ambos conjuntos de datos.

Vale destacar que las experiencias con matrices de transferencia al azar probablemente hayan dado origen a modelos en los cuales las muestras de transferencia no necesariamente representarían correctamente a la información proveniente del dominio secundario. Sin embargo, estas pruebas pretenden ver la robustez de las operaciones necesarias y de allí la aplicación de selecciones azarosas. Al mismo tiempo, las selecciones obtenidas fueron revisadas de forma tal que no contuvieran muestras repetidas en una misma matriz **L**, y que no existieran matrices **L** iguales.

Orden	tau T	lam T	tau M	lam M
1	9.64E-3	2.98E+1	3.43E-2	1.49E+1
2	4.82E-3	1.49E+1	1.72E-2	3.73E+0
3	2.41E-3	7.45E+0	8.58E-3	9.31E-1
4	1.21E-3	3.73E+0	4.29E-3	2.33E-1
5	6.03E-4	1.86E+0	2.15E-3	5.82E-2
6	3.01E-4	9.31E-1	1.07E-3	-
7	1.51E-4	4.66E-1	5.36E-4	-
8	7.53E-5	2.33E-1	2.68E-4	-
9	3.77E-5	1.16E-1	1.34E-4	-
10	1.88E-5	5.82E-2	6.70E-5	-
11	9.42E-6	-	3.35E-5	-
12	-	-	1.68E-5	-
13	-	-	8.38E-6	-

Tabla 1 : Valores de tau y lam para los modelos reportados en experiencias con muestras de transferencia al azar

Referencias: T: datos “Temperatura”, M: datos “Maíz”. E+n=x10^{n}.

1.6.1.1 Efecto del tipo de centrado en DR-SAC

Las siguientes experiencias estuvieron destinadas a obtener resultados basados en diferentes estrategias de centrado en DR-SAC. El número de muestras presentes en **L** fue siempre de 4.

1.6.1.1.1 Datos Maíz

Las 80 muestras disponibles en ambos instrumentos fueron divididas en 2 conjuntos utilizando el algoritmo de Kennard-Stone sobre los 80 espectros primarios. Las 30 primeras muestras de la selección fueron asignadas al conjunto de Calibración, y las restantes 50 para Validación. Estas asignaciones se corresponden tanto para los espectros primarios como para los secundarios, aunque durante las experiencias sólo se utilizaron los datos primarios de Calibración en **X** y los secundarios de Validación en **V**, además de sus respectivos valores de referencia.

La selección de 4 muestras para **L** se realizó 30 veces al azar y esas mismas tétradas fueron puestas a prueba con todas las estrategias de centrado. En todos los casos, dichas muestras solamente provinieron de las 30 de Calibración secundarias disponibles, que no estarían disponibles (las 30, aunque sí 4 de ellas) en la práctica real puesto que sino se realizaría una re-Calibración. El conjunto de Validación obtenido con el método de Kennard-Stone no aportó información primaria ni secundaria en las etapas de modelado y sólo la secundaria fue utilizada en las predicciones

definitivas. Los valores de tau y lam fueron los reportados en la tabla 1.

La figura 6 expone los resultados promedio obtenidos con las 30 matrices **L** al azar.

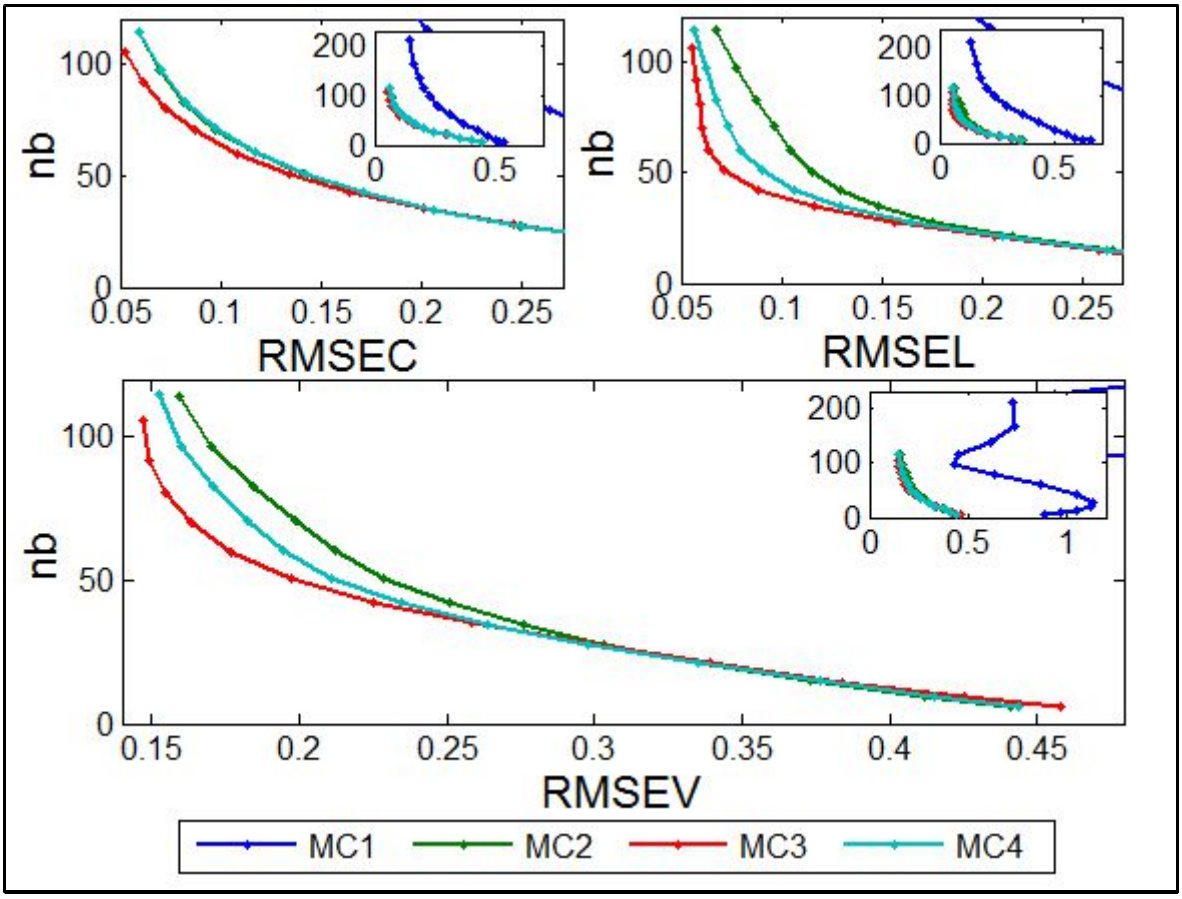


Figura 6: Promedio de RMSE (C, L y V) versus promedio de norma de los vectores de regresión (nb) para datos "Maíz" en 30 ejecuciones DR-SAC con 4 muestras de transferencia seleccionadas al azar en cada ejecución y diferentes estrategias de centrado

Referencias: Cada punto en cada curva representa el promedio para toda lam (5) en todas las ejecuciones (30), para cada tau individual (13). Los recuadros insertos en cada gráfica representan a su misma gráfica pero en una escala apta para poder apreciar también los resultados de MC1.

En la figura 6, las curvas de RMSEC dejan ver que MC3 produce los frentes óptimos en términos de Pareto básicamente para todo tau. Las estrategias sin centrado local puro MC2 y MC4 producen resultados muy similares, tanto que en la gráfica se encuentran básicamente solapados. Esto es debido a que en **X** existen 30 muestras y en **L** sólo 4, por lo cual el promedio de **X** y el promedio conjunto de **X** y **L** es similar (lo mismo se aplica a sus valores de referencia). Por su parte, el recuadro en otra escala deja ver que MC1 produce los frentes más lejanos al origen, lo cual

sugiere que para estos datos, si **L** no es centrada afectará muy negativamente al ajuste de **X**, lo cual es bastante lógico ya que además de que los datos de **X** ya centrados serán menores que los de **L** no centrados, se presume una deriva espectral entre ambos instrumentos. También puede observarse que las normas para MC1 fueron muy superiores a las obtenidas con las otras estrategias.

Las curvas de RMSEL son coherentes con las de RMSEC. El ajuste es mejor cuanto más local es el centrado de los datos de cada dominio, tanto espectros como valores de referencia. El recuadro permite observar que con MC1 los datos en **L** no serán bien ajustados, aunque haya sido suficiente para producir peores ajustes para **X**. Es decir, ni los datos primarios ni los secundarios que participan de las etapas de cálculo se ven beneficiados en el ajuste, por lo que no se podrían esperar resultados mejores para datos futuros que ni siquiera hubieran participado del modelado.

Finalmente, las curvas de RMSEV también confirman las tendencias de que lo mejor para **V** será un centrado proveniente de datos de su mismo dominio (MC3) o al menos con participación parcial (MC4) y no nula (MC2) en el cálculo de medias. El recuadro en MC1 permite observar que los errores para **V** serían inadmisibles.

1.6.1.1.2 Datos Temperatura

Se utilizaron 16 de las 19 muestras descriptas en la figura 2, todas con Etanol, quedando excluidas para estas experiencias las mezclas sin Etanol (muestras 17, 18 y 19) y las muestras con componentes puros (ausentes en la figura 2, pero disponibles en el conjunto total de datos).

El conjunto de Calibración primario, **X**, estuvo compuesto con 10 muestras (1, 2, 3, 4, 7, 8, 10, 12, 13 y 16 de la figura 2) cuyos espectros corresponden a la temperatura primaria de 30ºC. Este conjunto fue constante siempre, más allá de que variaron las muestras en **L** y en **V**.

Desde las muestras del dominio secundario en 50ºC se obtuvieron 30 combinaciones azarosas de 4 muestras para conformar **L**, partiendo de las 16 muestras en juego, y en cada caso las restantes 12 muestras no seleccionadas dieron origen al respectivo conjunto de Validación **V**, lo cual difiere de datos “Maíz”, donde **V** fue siempre un conjunto invariable de 50 muestras secundarias. Esto puede ser cuestionable, porque algunas de las muestras secundarias en **V** también estarán contenidas a través de sus primarias respectivas en **X**, participando del modelo DR. Esta decisión se tomó debido al escaso número de muestras disponibles en el conjunto de datos “Temperatura”. En relación a lo último, también debe notarse que el número de muestras primarias de Calibración fue 3 veces inferior al utilizado con datos “Maíz”. Los valores de lam y tau fueron los reportados en la

tabla 1. A su vez, los valores de tau fueron los mismos utilizados en la figura 5.

La figura 7 expone los resultados promedio obtenidos con las 30 matrices **L** al azar.

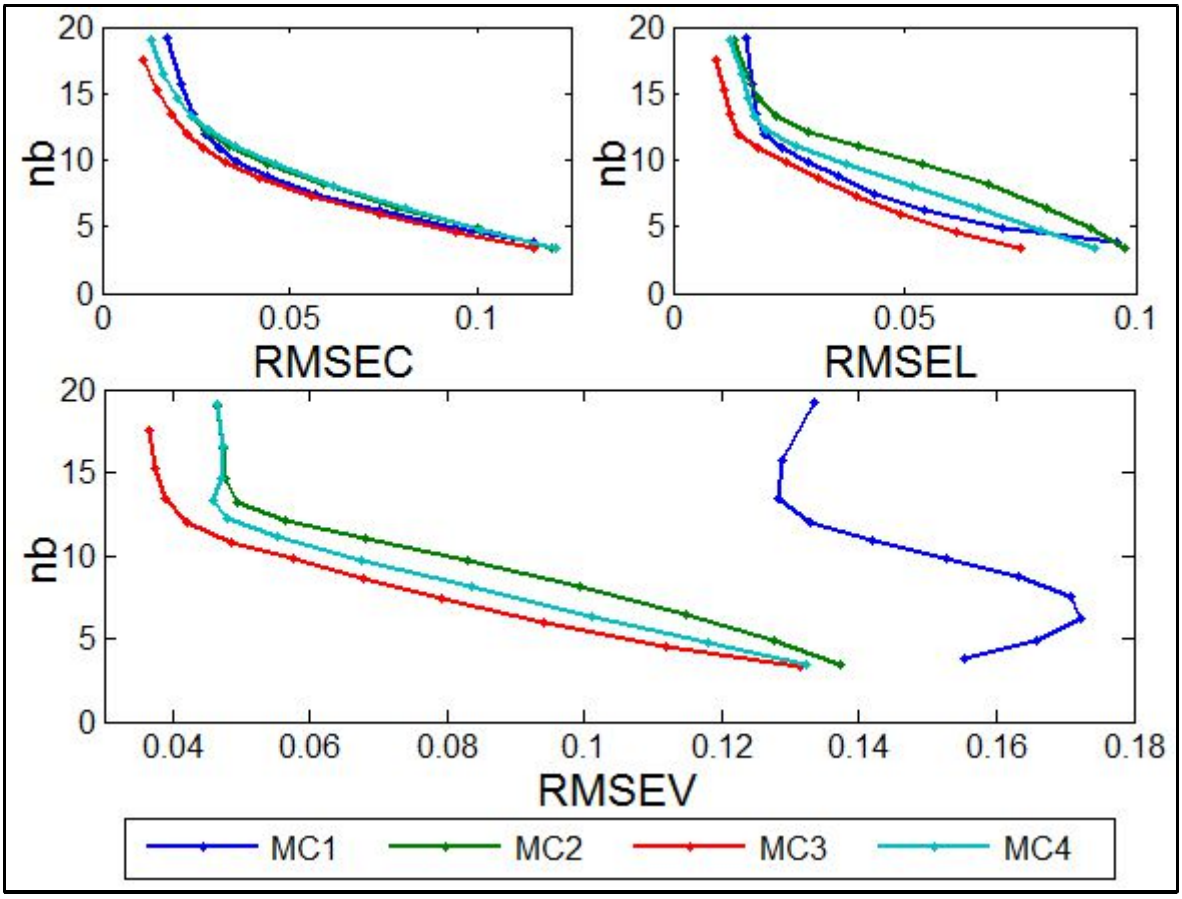


Figura 7: Promedio de RMSE (C, L y V) versus promedio de norma de los vectores de regresión (nb) para datos "Temperatura" en 30 ejecuciones DR-SAC con 4 muestras de transferencia seleccionadas al azar en cada ejecución y diferentes estrategias de centrado

Referencias: Cada punto en cada curva representa el promedio para toda lam (10) en todas las ejecuciones (30), para cada tau individual (11).

En la figura 7, las curvas para RMSEC dejan ver que MC3 resulta en los menores errores y normas vectoriales para todo tau, es decir que nuevamente las soluciones obtenidas con MC3 pueden considerarse Pareto superiores. Los centrados MC2 y MC4 muestran resultados similares, lo cual indica que el auto-centrado de **X** y de **y** (en MC2) no se ve demasiado afectado cuando el centrado se produce en base a datos mixtos provenientes de **X**, **y**, **L** e $\mathbf{y}_L$ (MC4), aunque tanto los resultados de MC2 como de MC4 se vean afectados por el centrado no local de **L**, característico solamente de MC3. Por su parte, MC1 se comporta similar a MC3 en normas bajas, pero diverge en

las altas, aunque MC1 resultará nuevamente de poco interés para la estrategia DR-SAC una vez analizados los resultados de RMSEV. No obstante, vale destacar que en datos "Maíz" MC1 obtenía resultados mucho peores que el resto de los centrados tanto para **X** como para **L** (y previsiblemente para **V**), mientras que en el caso de datos "Temperatura" el uso de espectros secundarios no centrados casi no perjudica a **X** e incluso se obtienen mejores resultados que con MC2 y MC4 para **L**. También se aprecia que no existen diferencias groseras de norma entre estrategias de centrado.

Las curvas de RMSEL sugieren que MC3 también es la mejor estrategia para todo tau, lo cual proviene de la forma local en que son centradas **L** e $\mathbf{y}_L$. Para este conjunto de datos eso explica también por qué MC2 es el centrado menos beneficioso para RMSEL, ya que en ese caso los datos de transferencia son centrados con los promedios de Calibración primarios y en el resto de los centrados los primeros son centrados con información total (MC3) o al menos parcial (MC4) de su propio dominio. A su vez. el caso intermedio de MC4 se presenta previsiblemente entre MC2 y MC3.

Las curvas de RMSEV confirman que MC3 es el centrado óptimo. A medida que el carácter local de centrado para los datos secundarios decae desde MC3, pasando por MC4, hasta MC2, se observan resultados menos óptimos, lo cual indica la importancia de centrar a los datos secundarios de **V** con información de su mismo dominio si es posible. El caso de MC1 era previsible y muestra que aunque **X** y **L** pudieron ajustarse relativamente bien a sus respectivos valores de referencia, estos ajustes no representarán a muestras que no participen del modelado.

1.6.1.2 Efecto del tipo de centrado en DR-DIFF

Con lo visto en las experiencias SAC para distintos centrados, es lógico suponer que para DIFF los resultados no serán los mismos. Específicamente, se espera que MC1 sea esta vez la estrategia más adecuada para **V**, pues al insertar diferencias en **L** y valores de referencia de 0 en $\mathbf{y}_L$, y no espectros con valores de referencia distintos como en SAC, cualquier centrado local o mixto de los datos secundarios de **V** basado en las diferencias y ceros no tendría demasiado sentido. Por lo tanto, en las siguientes experiencias se reportan sólo 5 ejecuciones al azar por centrado, las cuales serán suficientes para mostrar lo que se espera.

Para datos "Maíz" los valores de tau y lam fueron equivalentes a los de la figura 6, solo que se omitieron algunos tau y sólo se conservaron 7 de ellos, los más apropiados para evaluar modelos en la zona armónica (ni normas muy bajas ni muy altas). Para datos "Temperatura", los valores para los meta-parámetros fueron los mismos que dieron origen a la figura 7.

En todos los casos, los índices de muestras que en SAC dieron origen a las 5 primeras matrices **L** de las 30 totales fueron utilizados para obtener los espectros de las respectivas muestras primarias y luego se obtuvieron las 5 matrices de diferencias del tipo **L2-L1** para conformar las matrices **L** de DIFF. Todos los valores en $\mathbf{y}_L$ fueron de 0. Los conjuntos de Calibración y Validación fueron construidos como se describió en las experiencias con SAC.

1.6.1.2.1 Datos Maíz

La figura 8 corresponde a los resultados promedio obtenidos con las 5 matrices **L** al azar.

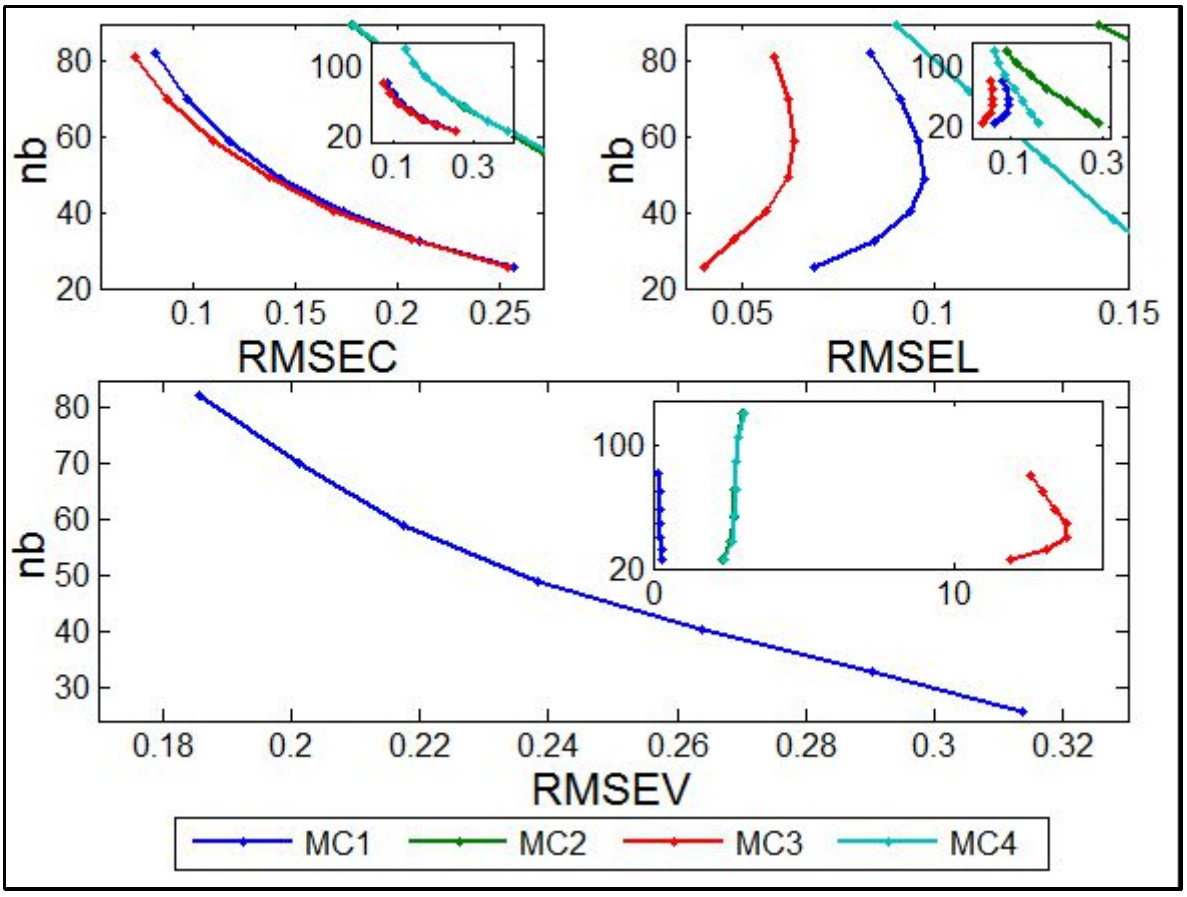


Figura 8: Promedio de RMSE (C, L y V) versus promedio de norma de los vectores de regresión (nb) para datos "Maíz" en 5 ejecuciones DR-DIFF con 4 diferencias de espectros de transferencia seleccionados al azar en cada ejecución y diferentes estrategias de centrado

Referencias: Cada punto en cada curva representa el promedio para toda lam (10) en todas las ejecuciones (5), para cada tau individual (7). Los recuadros insertos en cada gráfica representan a su misma gráfica pero en una escala apta para ver los resultados de todas las estrategias de centrado.

En la figura 8, las curvas de RMSEC dejan ver que que el ajuste a **X** fue similar con MC1 y MC3, es decir que las diferencias no interfirieron de manera negativa cuando no fueron centradas en MC1 y tampoco cuando lo fueron con sus propias medias en MC3. A su vez en el último caso la media de $\mathbf{y}_L$ es en sí 0 y por tal MC3 no tiene un efecto diferente al de MC1 sobre los valores de referencia. Por otro lado en el recuadro inserto se observa que MC2 y MC4 obtuvieron resultados peores y similares, sugiriendo que los centrados con información primaria total o parcial, respectivamente, no son lo apropiado para la introducción de diferencias en **L**. También debe recordarse que dada la relación entre las cantidad de muestras en **X** y **L**, la influencia de las últimas es escasa en el cálculo de medias de MC4 y por ende los datos medios de MC2 y MC4 son muy similares. Otra cuestión notable que es con MC2 y MC4 las normas de los vectores también resultaron mayores.

Las curvas de RMSEL permiten apreciar que el centrado local de MC3 sigue siendo lo mejor para la introducción de información en **L**, al menos para su propio ajuste. Seguido en orden de mérito se ubica MC1, aunque la diferencia no es menor. A su vez, puede verse que los valores de RMSEL con MC1 y MC3 para todo tau son bajos e incluso esos mismos errores son logrados en RMSEC para unos pocos valores de tau (los de normas superiores). Para el caso de RMSEL los resultados de MC2 no se encuentran solapados con los de MC4 (ver recuadro inserto) y el último es mejor en todos los puntos. Esto sugiere que el centrado mixto de MC4, aunque no haya influenciado de manera radicalmente distinta a **X** respecto de MC2, efectivamente proporciona cierta ventaja para el ajuste de **L** a sus valores de referencia.

Las tendencias esperadas se confirman en las curvas de RMSEV. Se aprecia que las curvas para MC2, MC3 y MC4 sólo son visibles en el recuadro pertinente. MC3 produce los mayores errores y esto es debido a que los datos en **V** son centrados con la media de las diferencias, por lo cual a los espectros en **V** sólo se les substrae un vector de valores muy pequeños antes de ser predichos, algo impropio para el caso de derivas apreciables. A su vez, como la media de $\mathbf{y}_L$ es 0, luego de las predicciones no se produce ningún tipo de re-escalado. Por el lado de MC2 y MC4, como los datos medios serán similares, los resultados para **V** centrado también lo serán. A su vez, ya las curvas de RMSEC y RMSEL no presentaban buenos resultados, por lo cual los obtenidos para **V** eran esperables. Si bien éstos no valen la pena, vale destacar que funcionaron mejor que MC3 porque al menos los espectros de **V** fueron pre-colocados en el hiperespacio de variables y sus predicciones fueron re-escaladas con la información primaria de escala un poco más adecuada. Esto último también se dio con MC1, que a su vez produjo modelos donde **X** y **L** fueron ajustados de manera

aceptable, por lo que sus resultados para **V** resultaron ser los mejores, tal y como se esperaba.

1.6.1.2.2 Datos Temperatura

La figura 9 corresponde a los resultados promedio obtenidos con las 5 matrices **L** al azar.

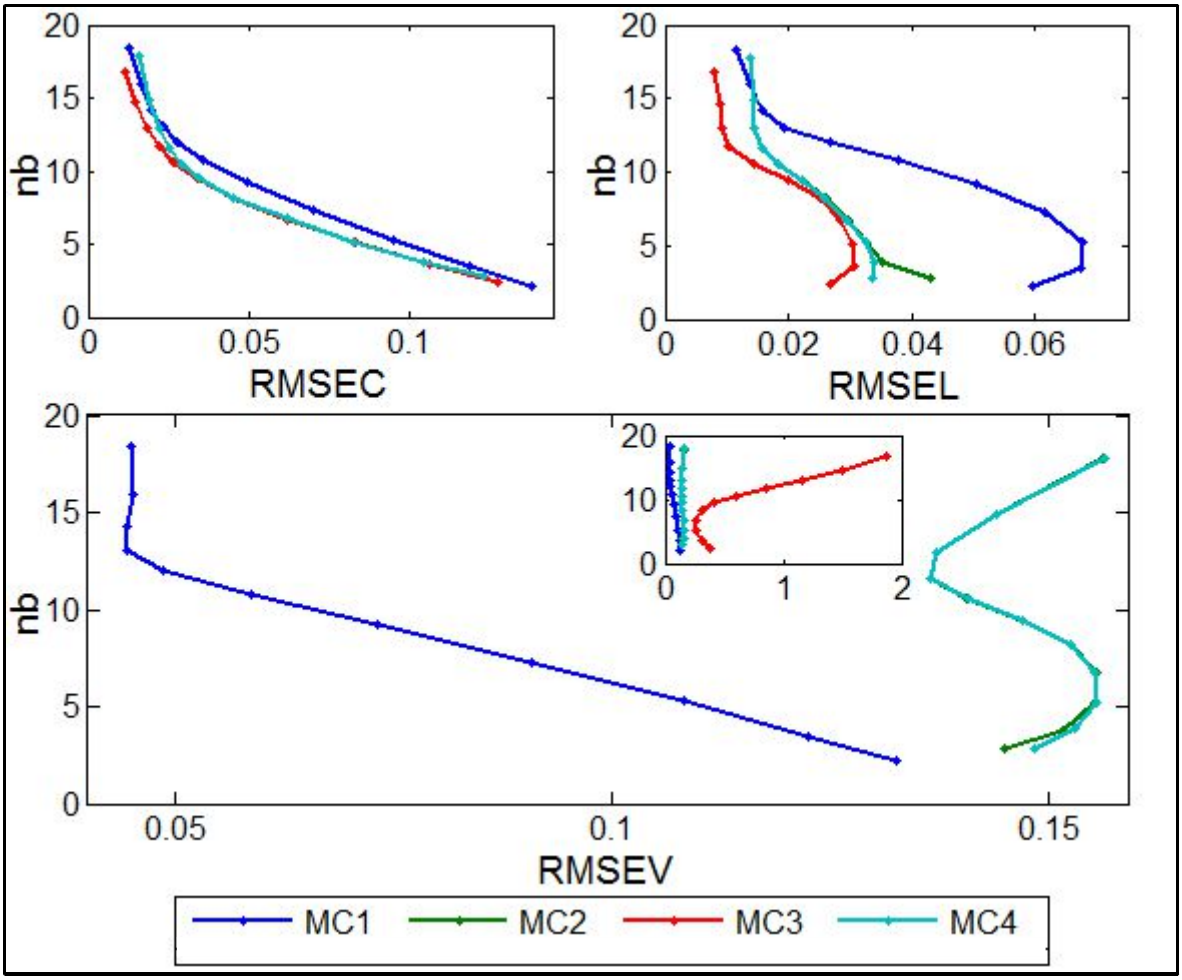


Figura 9: Promedio de RMSE (C, L y V) versus promedio de norma de los vectores de regresión (nb) para datos "Temperatura" en 5 ejecuciones DR-DIFF con 4 diferencias de espectros de transferencia seleccionados al azar en cada ejecución y diferentes estrategias de centrado

Referencias: Cada punto en cada curva representa el promedio para toda lam (10) en todas las ejecuciones (5), para cada tau individual (11). El recuadro inserto en la gráfica de RMSEV representa a su misma gráfica pero en una escala apta para ver también los resultados de MC3.

En las curvas de RMSEC de la figura 9 puede apreciarse que el ajuste a **X** se vio poco influenciado en general, independientemente del centrado aplicado, aunque MC1 presente

resultados levemente peores. Similarmente, en las curvas de RMSEL se observa que aunque puede establecerse un orden de méritos con MC1 en último lugar, los resultados de todas las curvas son bajos si se los compara con los errores en RMSEC. Otra observación que puede realizarse es que las normas se mantuvieron similares para las 4 estrategias de centrado. Se percibe que las conclusiones que de aquí pueden extraerse no son exactamente las mismas que en el caso de datos "Maíz", aunque no hay una razón estricta que debería determinar parecidos mayores entre datos diferentes. En relación, se recuerda que en los casos de SAC ambos conjuntos de datos se comportaban de forma más similar respecto de los diferentes centrados.

Lo más importante para destacar son los resultados de RMSEV, donde sí existe concordancia entre los conjuntos de datos, ya que el orden de mérito es muy similar al visto para datos "Maíz", con MC1 como lo más apropiado, con MC2 y MC4 como segundas opciones, y finalmente con MC3 exponiendo los mayores errores (sólo visibles en el recuadro inserto). Por consiguiente, de estas experiencias también se concluye que para obtener mejores resultados para muestras secundarias futuras, MC1 debería ser la estrategia de centrado si han de utilizarse diferencias en **L**.

1.6.1.3 Efecto del número de muestras en L

Habiendo definido en las experiencias anteriores que el centrado óptimo para DR-SAC es MC3 y que para DR-DIFF es MC1, a continuación se evalúa el efecto del número de muestras en **L** con dichos centrados en ambos conjuntos de datos, para los cuales los valores de lam y tau fueron los expuestos en la tabla 1.

Similarmente al modo en que fueron divididos los juegos de datos en las 30 ejecuciones con 4 muestras de transferencia al azar, se realizaron selecciones azarosas de 30 ternas y de 30 duplas de muestras para **L** (de uso directo en SAC, o bien a través de **L2**-**L1** en DIFF), se conformaron las respectivas matrices **X** y **V**, y junto a los valores de referencia respectivos se obtuvieron modelos SAC y DIFF. Cuando las transferencias fueron realizadas con muestras únicas (vectores **L**), en el caso de datos "Maíz" cada una de las 30 muestras de Calibración secundarias dio origen a una matriz **L** diferente , mientras que para datos "Temperatura" cada una de las 16 muestras utilizadas en estas experiencias aportó los datos para las transferencias, por lo cual sólo en este caso los resultados promedio que serán expuestos provendrán de 16 ejecuciones y no de 30.

Sin ahondar en detalles sobre cómo varían RMSEC y RMSEL, a continuación se exponen los resultados obtenidos con ambos conjuntos de datos en ejecuciones múltiples para los conjuntos de Validación respectivos, con 1, 2, 3 y 4 muestras de transferencia tanto para SAC como para DIFF.

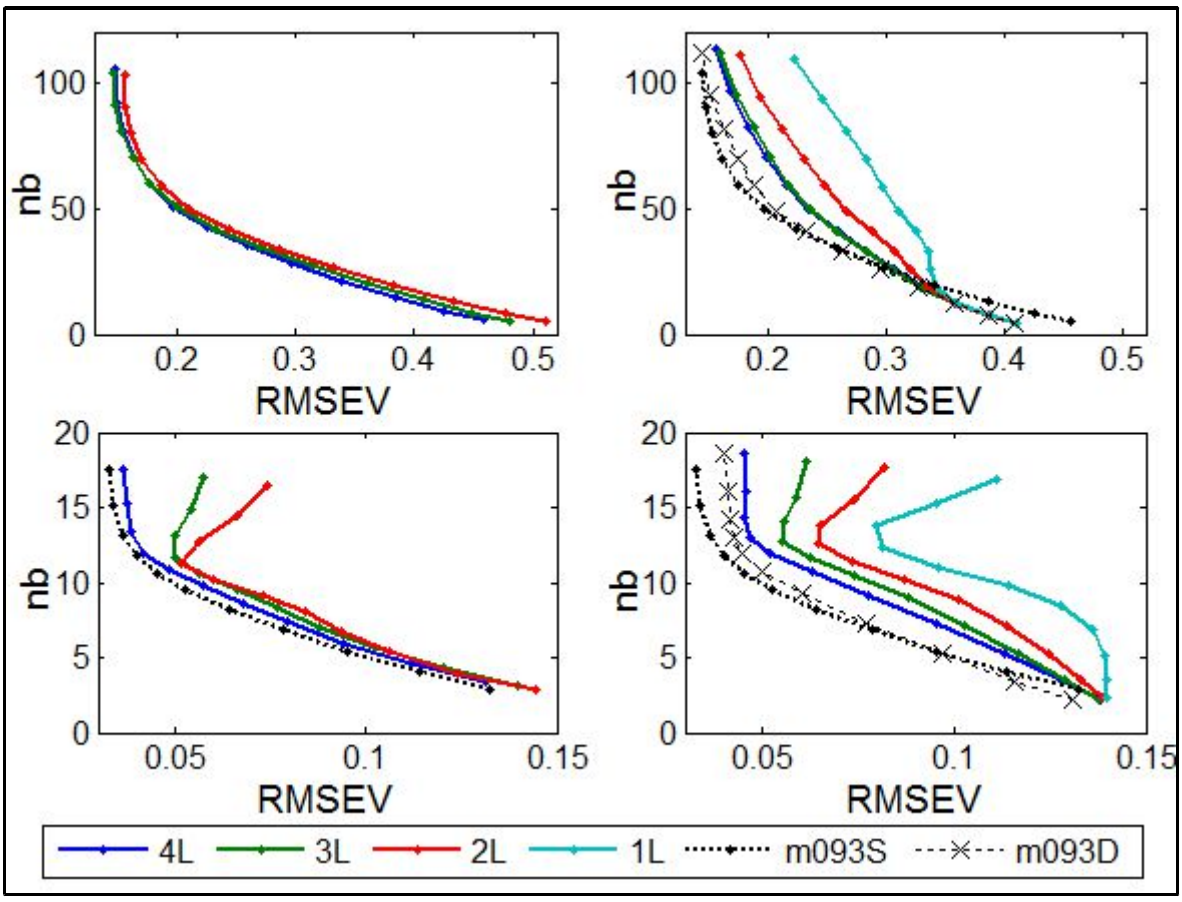


*Figura 10: Promedio de RMSEV versus promedio de norma de los vectores de regresión (nb) para datos "Maíz" (arriba) y "Temperatura" (abajo), en 30 ejecuciones DR-SAC-MC3 (izquierda) y DR-DIFF-MC1 (derecha), con 1 a 4 muestras en **L** seleccionadas al azar en cada ejecución*

Referencias: nL: Número de muestras en **L**, m093S: curva media de 30 ejecuciones DR-SAC-MC3 con 4L en lam=0.9313, m093D: curva media de 30 ejecuciones DR-DIFF-MC1 con 4L en lam=0.9313. Para el resto de las curvas, cada punto representa el promedio para toda lam en todas las ejecuciones, para cada tau individual. Las ejecuciones para datos "Temperatura" con 1 muestra en **L** no fueron 30 sino 16 (total de muestras seleccionadas para estas experiencias). Para DR-SAC-MC3 no se realizaron experiencias con 1L.

En la figura 10, las gráficas para datos "Maíz" en SAC dejan ver que los resultados son muy similares sin importar el número de muestras de transferencia. La curva de 2 muestras se ve levemente separada de las otras, las cuales se encuentran solapadas en gran parte del intervalo de normas. En el caso de DIFF, se aprecia que al pasar desde 1 hacia 2 muestras existe mejoría, y desde 2 hacia 3 también, aunque menor. El aumento posterior de una muestra no parece ser relevante para RMSEV y por eso nuevamente las transferencias con 3 y 4 muestras se encuentran

muy solapadas. De ambas gráficas puede deducirse que para este conjunto de datos 3 muestras de transferencia serían suficientes, o dicho en otras palabras, que una cuarta muestra casi no redundará en mejorías.

Si se comparan SAC y DIFF, el primero resulta mejor a nivel medio en el intervalo de lam reportado. Sin embargo, en la gráfica para DIFF pueden también observarse los datos medios de las 30 ejecuciones para cada tau sólo en lam=0.9313, la más cercana a la unidad, donde se aprecia que SAC4 y DIFF4 logran resultados muy similares, aunque SAC4 obtiene mejorías leves en la zona armónica y en las superiores. Dicha curva para SAC no fue inserta en su propio gráfico debido a que casi se solapaba con la media a toda lam. En ese sentido, el intervalo de lam reportado afectó de forma similar a los modelos en términos medios que en el caso específico de la media a lam=0.9313. En el caso de DIFF, el comportamiento medio en el intervalo entero produjo resultados de calidad inferior si se los compara con los obtenidos solo en lam=0.9313.

Las gráficas inferiores para datos “Temperatura” permiten ver que el aumento del número de muestras de transferencia produce un disminución gradual de RMSEV. En el caso de SAC esta disminución no es demasiado relevante y por eso las curvas se observan muy cercanas entre sí. En las normas fundamentalmente superiores (arriba de 10) se produce una diferencia clara y a medida que disminuyen las muestras en **L**, el dominio secundario comienza a ser relegado y por lo tanto aumentan los RMSEV. Al notar que con 4 muestras no se produce el mismo fenómeno (aumento de RMSEV), queda indicado que para este conjunto de datos ese número de muestras de transferencia parece ser apropiado. También para SAC se agregó la curva m093S respectiva, de lo cual se puede apreciar que los resultados medios a toda lam fueron de calidad levemente inferior. En el caso de DIFF las diferencias de RMSEV respecto del número de muestras en **L** son más notorias y también se observa que en las normas superiores sólo con 4 muestras no se producen aumentos del error, lo cual vuelve a indicar que ese número de muestras parece apropiado. Al respecto, DIFF4 puede ser comparado entre sus resultados a toda lam y sólo en lam=0.9313, siendo los últimos Pareto superiores, aunque no respecto de SAC4 en la misma lam.

Para ambos conjuntos de datos en las gráficas anteriores puede no apreciarse un detalle dada la escala utilizada, pero casi sin excepciones y sin importar si se utilizó SAC o DIFF, al aumentar el número de muestras de transferencia aumenta también la norma de los vectores respectivos y esto ocurre básicamente para todo tau. Para interpretar lo anterior es conveniente recordar las curvas de la figura 1 para RR y PLS. En ambos casos puede apreciarse que los datos que dan origen a los modelos (en dicho ejemplo eran solamente datos de Calibración sin transferencia) serán cada vez

mejor ajustados a media que se otorgue libertad al cálculo, sea a través de restricciones menores a la norma en RR o del agregado de Variables Latentes en PLS. Así, este aumento de libertad permitirá modelar más profundamente la varianza presente, pero para lograrlo los vectores de regresión deberán establecer relaciones cada vez más estrictas o específicas entre los datos modelados. Esta especificidad provendrá de que los coeficientes trascendentales para modelar más varianza tomarán valores cada vez más lejanos de 0, pero como producto de eso pequeñas variaciones en los datos modelados repercutirán con gran impacto en el ajuste. El alejamiento de 0, sea hacia coeficientes más positivos o más negativos, conllevará valores absolutos mayores y por ende aumentarán las normas vectoriales. Retomando la observación sobre el crecimiento de las normas para el caso de las transferencias con aumento de muestras en **L**, el agregado de información secundaria a la primaria pre-existente representará también más varianza para modelar, y tanto más cuanto más sean las muestras de transferencia (en general), por lo cual el ascenso de las normas vectoriales podría ser un resultado esperado. Debe entenderse que estos efectos sobre las normas provendrían del aumento de información que debe ser modelada y no de cómo la nueva información es ponderada, ya que lo último tendría más relación con lo dicho durante el análisis de la figura 5 (50 valores de lam por cada tau), donde se destacó la tendencia de la elevación de las normas con el aumento de los valores en lam.

1.6.2 Experiencias con muestras de transferencia específicas

Con los resultados de las experiencias anteriores se definieron los centrados óptimos para SAC y DIFF, y se pudo apreciar que para ambos conjuntos de datos las transferencias con 4 muestras serían apropiadas (en datos “Maíz” con 3 sería suficiente). Estas condiciones fueron reproducidas en las siguientes experiencias, donde el foco de análisis estará en los detalles provenientes de transferencias específicas, sin promediar ejecuciones.

También en las experiencias anteriores pudo observarse que los valores de lam cercanos a la unidad (0.9313) dieron modelos aceptables en calidad. En relación a estos valores, si se recuerda la figura 5 donde los resultados expuestos para datos “Temperatura” correspondieron a intervalos de lam de 50 valores, se vio que valores mucho mayores a la unidad producirían resultados no óptimos, lo cual sugiere que el aumento de lam más allá de la unidad debería realizarse siempre con precaución. Por lo tanto, para cada conjunto de datos inicialmente se realizaron experiencias comparativas cuyos resultados serán expuestos en lam=1, y finalmente se expondrán resultados de

experiencias específicas (incluidas en las comparaciones) utilizando un intervalo de valores para lam. Se decidió que dichos valores provendrían de una metodología común para ambos conjuntos de datos, de forma tal que cada transferencia determine el máximo de lam y algunos otros de los valores del intervalo puestos a prueba. La estrategia que generó 10 valores de lam para estas transferencias fue la siguiente:

- Máxima lam (Maxilam): proviene de relacionar el número de muestras primarias de Calibración (nX) con el de muestras secundarias de transferencia (nL). Así pues: Maxilam = nX/nL De esta forma en Maxilam la relevancia numérica de las muestras en **L** sería igual a la de las muestras en **X**.
  - Por ejemplo, para datos "Maíz" y SAC4 o DIFF4: nX=30, nL=4 y Maxilam=7,5. Entonces 1x30=7.5x4, los aportes serán equivalentes durante la minimización respectiva.
- 4 valores equidistantes entre Maxilam y 1 (p1, p2, p3 y p4)
- 5 valores fijos para cualquier transferencia: 1.0687, 1, 0.9313, 0.75 y 0.5
  - El valor 1 es de interés por lo visto cerca de la unidad
  - Los valores 0.75 y 0.5 servirán para evaluar valores pequeños (baja ponderación para **L**)
  - El valor 0.9313 se reutiliza porque en las ejecuciones al azar estuvo presente y esto permitirá realizar comparaciones con aquellas experiencias
  - El valor 1.0687 se aleja tanto de la unidad como 0.9313, solo que siendo mayor a esta. Con este valor se pretendió evaluar un efecto de "ponderación simétrica"
- La conjunción de todos los valores ordenados otorgará 10 lam normalmente
  - Maxilam, p1, p2, p3, p4, 1.0687, 1, 0.9313, 0.75 y 0.5

También vale destacar que en ocasiones fue apropiado agregar y/o suplantar valores de lam por algunos de interés para su análisis. Uno de los valores que fue agregado selectivamente fue lam=0, el cual no se tuvo en cuenta en la estrategia de generación común y se obvió en el reporte de algunas experiencias puesto que a veces genera distorsiones sin sentido útil en ciertas gráficas (en especial las de RMSEL). Sin importar cuáles fueron los cambios en el vector común de 10 lam, éstos serán debidamente explicitados.

1.6.2.1 Datos Maíz

1.6.2.1.1 Valores de tau

Para no exponer figuras con escalas desmedidas que harían perder ciertos detalles, se decidió realizar las siguientes experiencias con un subconjunto de los tau utilizados en las experiencias con 30 ejecuciones al azar, de forma tal de conservar aquellos que dieran modelos en zonas de norma intermedia (ni muy deficientes, ni muy sobreajustados). Específicamente fueron seleccionados los tau 6, 7, 8, 9, 10 y 11. Luego, para aumentar la resolución intermedia, entre cada tau se agregaron 2 más (uno de ellos a 1/3 de la distancia entre ambos, el otro a 2/3). Los valores son expuestos en la tabla 2.

Orden	Valor
1 (6)	1.07E-3
2	8.94E-4
3	7.15E-4
4 (7)	5.36E-4
5	4.47E-4
6	3.58E-4
7 (8)	2.68E-4
8	2.23E-4
9	1.79E-4
10 (9)	1.34E-4
11	1.12E-4
12	8.94E-5
13 (10)	6.70E-5
14	5.59E-5
15	4.47E-5
16 (11)	3.35E-5

Tabla 2: Valores de tau para datos "Maíz" en ejecuciones únicas

Referencias: Los valores entre paréntesis indican el orden del mismo tau en las experiencias con ejecuciones replicadas al azar. E+n=x10^{n}.

A su vez, en algunas experiencias se tuvieron que utilizar los mismos 13 tau que en las ejecuciones al azar, pero en dichos casos se hará explícito el cambio.

1.6.2.1.2 Conjuntos de Transferencia, Calibración y Validación

La selección de 4 muestras para las transferencias se realizó a partir las 30 muestras secundarias de Calibración disponibles para estos estudios, aplicando el algoritmo de Kennard-Stone sobre la

información espectral. En el caso de experiencias con 3 muestras, éstas fueron las primeras 3 seleccionadas por el mencionado algoritmo.

Los conjuntos de Calibración y Validación fueron los mismos que en la experiencias con ejecuciones múltiples (30 muestras primarias y 50 secundarias, respectivamente).

1.6.2.1.3 Experiencias, resultados y análisis

Lo primero que debe ser determinado es cuáles hubiesen sido las condiciones primarias que dejaron de ser aplicables por alguna razón, que para este conjunto de datos sería un cambio de instrumento. La figura 11 expone resultados en relación a lo anterior.

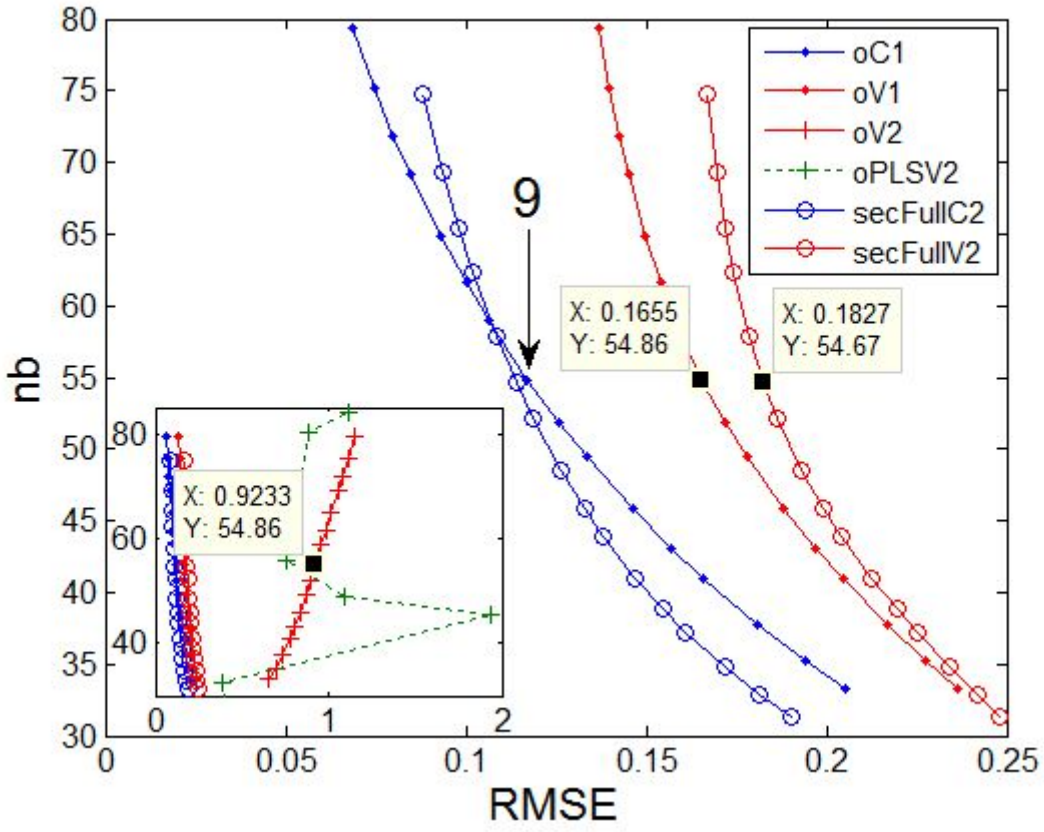


Figura 11: RMSE (C y V) versus norma de los vectores de regresión (nb) para modelos primarios sin transferencias y Re-Calibraciones Completas, datos "Maíz"

Referencias: oC1: Original RMSEC1, oV1: Original RMSEV1, oV2: Original RMSEV2, oPLSV2: RMSEV2 de PLS en dominio original, secFullC2: Secundario Re-Calibración Total (Full) RMSEC, secFullV2: Secundario Re-Calibración Total (Full) RMSEV. El recuadro inserto en la gráfica representa a su misma gráfica pero en una escala apta para ver otros. Los recuadros amarillos representan puntos de interés. X es RMSE e Y es nb (ver texto)

Antes de comenzar con el análisis de la figura 11, se quiere aclarar que en ésta y en otras figuras se señalarán algunos puntos de interés mediante el agregado de recuadros amarillos con valores de X y Y. Los valores mostrados son automáticamente formateados por Matlab y ésto no tiene en cuenta cuestiones relativas a las cifras significativas, como por ejemplo qué cantidad de valores deben mostrarse luego del punto decimal. Por lo tanto, estas cifras deberán tomarse como aproximaciones con rigor indefinido en el tratamiento de los datos.

La figura 11 presenta en las curvas oC1, oV1, oV2 y oPLSV2 resultados de modelos obtenidos en el dominio original (primario ó 1) que habrían perdido su validez y que requerirían transferencia de Calibración. En el contexto de las ecuaciones de DR, los modelos sin transferencia provendrían de la minimización de la expresión (16) sin tener en cuenta al último término, es decir, modelos con información de Calibración y con regulación de norma, pero sin información de transferencia (sin λ, $\mathbf{L}$ ni $\mathbf{y}_L$), tal y como en RR o como se explicitó en la ecuación (23). La curva oC1 señala los resultados de los modelos RR primarios para las 30 muestras de Calibración primarias (C1, o $\mathbf{X}$ en el contexto de DR) y lo mismo hacen oV1 y oV2 con las muestras de Validación primarias (V1) y secundarias (V2, o en este contexto simplemente $\mathbf{V}$), respectivamente. Similarmente, oPLSV2 indica las cifras para modelos PLS con distinto número de Variables Latentes en la predicción de V2. Los espectros en V1 y V2 fueron centrados con la media de C1, como si se utilizara la clásica estrategia de centrado que en este trabajo se denominó MC2, es decir, todo lo que debe ser predicho sería previamente centrado con dicha media. La decisión de centrar con MC2 proviene de que en este momento del desarrollo, no se supone que se tenga conocimiento de la desactualización de los modelos y por ende, al creer que el modelo aun funcionaría bien, lógicamente el centrado provendría de los datos de Calibración como sería usual. Es justamente a partir de este momento del desarrollo en que, al evaluar los resultados de las nuevas muestras y tratando a éstas como si la situación fuera normal, se obtendrían resultados inesperados y con esto la necesidad de correcciones. Tanto para RR como para PLS se observa que los errores de predicción de V2 son excesivos, de lo cual se deriva la necesidad de actualizar los modelos, o bien de realizar una re-Calibración del dominio secundario. Este último procedimiento, que se quiere evitar en este trabajo, deja ver sus resultados en las curvas secFullC2 y secFullV2, que provienen de haber calibrado sólo con C2 y validado con V2. Como puede notarse, los errores en secFullV2 serían mucho menores que los de oV2. También se aprecia que cada dominio produciría errores de Calibración similares y que el dominio primario obtendría errores menores en su conjunto de Validación, aunque las diferencias tampoco serían groseras. Más aun, el conjunto de datos “Maíz” fue dividido aplicando el

algoritmo de Kennard-Stone sobre las 80 muestras primarias, por lo que las 30 seleccionadas para calibrar bien podrían ser más representativas de su conjunto de Validación que las mismas 30 muestras pero en el dominio secundario con sus propias muestras de Validación.

Volviendo al contexto de la necesidad de transferencia, se requiere ahora la selección de un modelo de los presentes en oC1. Cualquiera de los puntos en dicha curva sería un modelo original posible de haber sido utilizado y eso dependería del criterio de selección de modelos. En este trabajo dicho criterio se basa en la armonía entre error en C1 y norma del vector respectivo. El número 9 indica un tau que otorgó un modelo armónico en términos de equilibrio entre RMSEC y nb, y aunque otros cercanos también podrían haberse escogido, se supondrá que ese era el modelo primario original que comenzó a fallar. De aquí puede obtenerse información importante para lo que resta del análisis:

- Los modelos futuros que han de evaluarse aplicando DR deberían tener cifras de nb cercanas y, según algunas tendencias observadas, levemente superiores a la del modelo original en tau 9. Se supone que el agregado de información secundaria de transferencia a la información primaria de Calibración podría conducir a un aumento de norma en el vector de regresión. Esto no es estricto, pero se observó ampliamente durante muchas experiencias con los valores de los meta-parámetros puestos a prueba. También sería posible la elección de un tau distinto si se considera apropiado (menor en valor, de forma tal que otorgaría un vector de regresión con norma mayor), pero en dicho caso no se estaría actualizando el modelo primario estrictamente, porque ese modelo ya habría sido definido con un valor de tau. No obstante, elegir valores de tau diferentes para obtener modelos en normas superiores sería similar a agregar Variables Latentes a un modelo PLS primario con el objeto de contemplar información secundaria. Esto no sería incorrecto en sí, pero debería tenerse información de que al aumentar las normas el riesgo de sobreajuste no afectaría a los datos que se pretenden predecir en el futuro, más allá de que también se hubiese sacrificado el criterio de armonía original.
- Se supondrá que el RMSE para V1 (RMSEV1) era una cifra aceptable mientras el modelo primario funcionaba correctamente. Dicha cifra es de 0.1655 (resaltada en oV1) y en el mejor de los casos una transferencia debería acercarse o ser menor a ese valor para la predicción de V2, es decir, se buscará tener la misma calidad en los resultados que la que se tenía originalmente y no la que se tendría si hubiera re-Calibración. También se resaltó RMSEV2 en secFullV2 a una norma similar, aunque el tau no sea el mismo para ambas

curvas.

- Para el tau 9 en cuestión, el modelo primario tiene un error en V2 de 0.9233 (resaltado en el cuadro inserto), aproximadamente 5.6 veces mayor a lo que debería.

En el contexto de transferencia, se supone que uno reutilizará C1 (porque C2 no existiría) y mediante DR y muestras de transferencia obtendrá una actualización para predecir muestras futuras. Dichas muestras en este escrito son las de V2. Por lo tanto, cuando se haga referencia simplemente a RMSEV, se lo estará haciendo siempre con V2. A su vez C1 será normalmente representado por **X**, para mantener coherencia con las ecuaciones, aunque su error será siempre nombrado como RMSEC.

En la figura 12 se exponen resultados de distintas aplicaciones de DR-SAC y variantes. El análisis de las curvas presentes en dicha figura es el siguiente:

SAC3, SAC4 y m093S: En términos generales y más allá de que SAC3 presenta resultados levemente mejores a partir del tau cercano a nb=65 (aproximadamente), la transferencia con 4 muestras resultó en cifras de RMSEV levemente inferiores que con 3, aunque los resultados son muy similares, tal y como se había observado en las experiencias con 30 ejecuciones al azar para SAC3 y SAC4. Al respecto, se observa la curva para dichas ejecuciones con SAC4 en su lam=0.9313 (el valor más cercano a 1 de aquellas ejecuciones), por lo cual se aprecia que el conjunto seleccionado para transferir obtuvo resultados mejores en términos de RMSEV y muy similares en términos de norma. De hecho, en la curva m093S uno de sus puntos fue resaltado dentro de un óvalo junto a 2 puntos de SAC3 y SAC4. Lo que se quiso hacer notar allí es que las normas de SAC4 en ese tau (su propio 10) y las de las ejecuciones múltiples en su su propio tau 9 (coincidente en valor con el 10 de SAC4, tal como puede apreciarse en la tabla pertinente) son casi coincidentes. En ese sentido, la norma de SAC4 en su tau 10 fue similar a la obtenida en ejecuciones al azar que también implicaron “30+4” (nX+nL) muestras finales dando origen a los modelos, aunque en el caso de SAC4 lam es 1 y en el otro caso sólo se aproxima (0.9313). Para SAC3 y SAC4 también debe apreciarse que en el tau 9 (sólo señalado para SAC4 con un recuadro amarillo y los valores X=0.1695 e Y=55.27) los vectores tienen normas levemente mayores en relación a la del modelo primario (54.86), y este aumento de norma era previsible teniendo en cuenta tendencias ya mencionadas. En relación a lo anterior, también se verifica que en el tau 9 SAC4 posee una norma levemente mayor a la de SAC3 (es el punto verde a la derecha del señalado

en SAC4). En ambos casos se obtienen cifras de RMSEV cercanas al objetivo de 0.1655, lo cual confirma que estas transferencias podrían ser realizadas de forma aceptable con 3 ó 4 muestras en **L**.

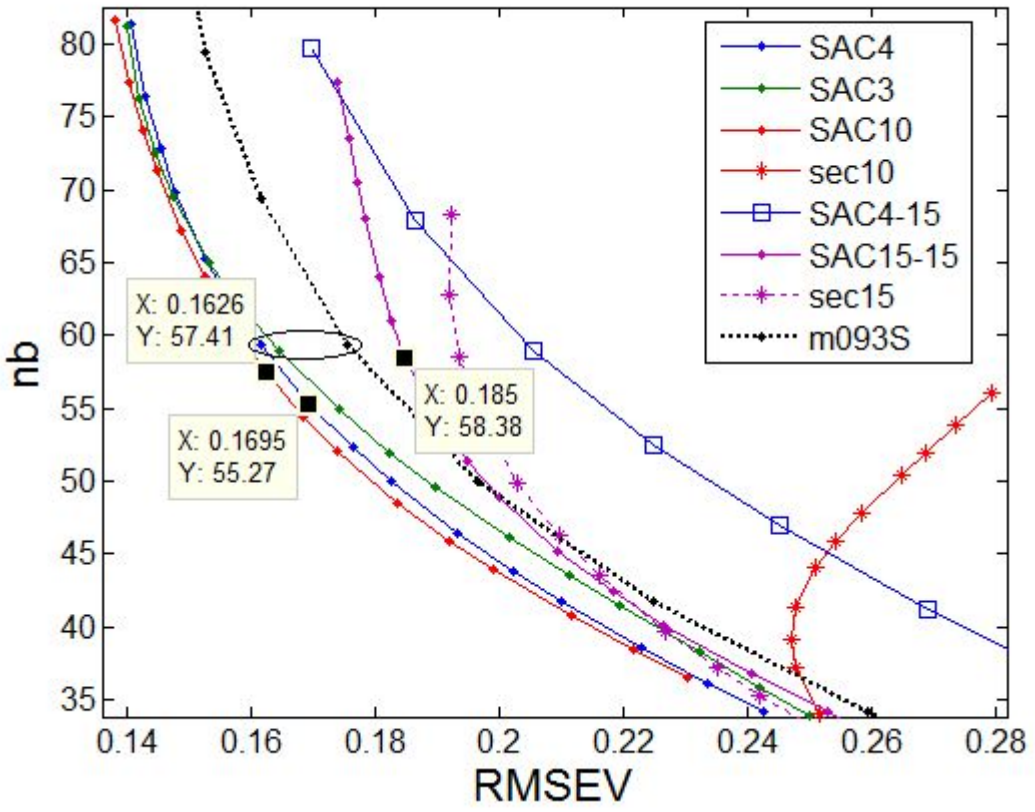


Figura 12: RMSEV versus norma de los vectores de regresión (nb) para DR-SAC y variantes en lam=1, datos "Maíz"

Referencias: SACn: DR-SAC con n muestras de transferencia y 30 calibradores primarios, secN: modelo RR secundario con N muestras secundarias usadas para calibrar, SACn-m: DR-SAC con n muestras de transferencia y m muestras de Calibración primarias, m093S: media de 30 ejecuciones al azar para DR-SAC4 en lam=0.9313. Los recuadros amarillos representan puntos de interés. X es RMSEV e Y es nb (ver texto)

SAC10: Se tomaron 10 espectros para la transferencia. Dichos espectros se seleccionaron desde los 30 de Calibración (secundarios) disponibles, desde el segundo de éstos, omitiendo 3, seleccionando uno, omitiendo 3 y así sucesivamente (en notación, 2:3:30). No tiene mucho sentido en este contexto tomar 10 muestras sólo para transferir, pues se supone que la transferencia es una estrategia ahorrativa. Sin embargo, se quiso ver cuál sería el efecto de elevar abruptamente la cantidad de información secundaria sin alterar la cantidad de información primaria. Se aprecia que

en relación a SAC3 y SAC4, para un mismo valor de tau, se obtienen RMSEV inferiores, pero la norma siempre es mayor también, puesto que hay mucha más información para modelar. Este aumento claro de la norma se puede observar a partir de las cifras Y en los recuadros amarillos para SAC4 y SAC10, los cuales pertenecen al mismo tau (9). Para SAC10 también puede apreciarse que en dicho tau la cifra de RMSEV es incluso menor que la obtenida en el modelo primario para los datos de Validación primarios (0.1655). Esto último indica que la transferencia con 10 muestras sería lo suficientemente aceptable como para recuperar la calidad de los resultados originales.

A la vista de los resultados en las normas de SAC3, SAC4 y SAC10, es oportuno aquí hacer una aclaración. Siendo que nb es uno de los criterios de armonía, dejar que sea mayor no parece lo apropiado, y esto se asemeja al caso de haber elegido un modelo primario en un tau con norma superior, menos armónico que el seleccionado. Sin embargo, a tau fijo se observa que usualmente se elevará la norma por tener que contemplar también a la información de transferencia y no solo a la original primaria. Por ende, no debe considerarse inapropiado que uno de los criterios de armonía se eleve (no excesivamente) si el regulador de la norma sigue siendo el mismo que el original.

Con lo anterior, también se aprecia que no es necesario un mecanismo de selección de un nuevo tau, puesto que si el tau original se mantiene, los modelos actualizados predicen con cifras de error aceptables. Si bien a la vista de las curvas de RMSEV es sencillo notar que se obtendrían menores errores con valores de tau en normas superiores, debe entenderse que eso sí sería violar el criterio de armonía original, pues ciertamente al cambiar el tau se estaría optimizando otro modelo, no el mismo primario. No obstante, lo anterior también podría ser válido si se interpreta que la selección de un tau distinto al original y menor que éste (con posibilidad de norma mayor) estaría basada en la decisión de contemplar en simultáneo más fuentes de varianza que las originalmente tenidas en cuenta. En este estudio, la relevancia en los análisis estará sentada en el mantenimiento del tau original, aunque podrá hacerse referencia a resultados de otros valores de tau.

sec10: La curva expuesta representa modelos RR realizados directamente en el dominio secundario, sólo que las muestras usadas para transferir en SAC10 serían usadas para calibrar directo en su propio dominio y serían las únicas aportando información al modelo. Esto sería con lam=0 (no existe una situación secundaria qué ponderar puesto que se está en la situación secundaria en sí) y con $\mathbf{L}$ e $\mathbf{y}_L$ ocupando el lugar de $\mathbf{X}$ y de $\mathbf{y}$ en la ecuación (16). En principio, conviene notar que como dicho modelo sólo contiene 10 muestras que dieron origen al vector de regresión y no “30 + nL” como fue usual hasta aquí, la norma es mucho menor y eso está de

acuerdo con la idea general de “más muestras, norma mayor”. Al mismo tiempo se observa que el RMSEV comienza a aumentar a partir de cierto tau hacia normas mayores, lo cual indica pérdida de generalización, algo esperable dado que sólo se ha modelado con 10 muestras. Por lo tanto, en esa curva a partir de cierto tau los ajustes sólo estarán destinados a los detalles de las 10 muestras de Calibración, lo cual ejemplifica un caso en el cual el aumento no controlado de la norma vectorial conduciría a errores mayores y no óptimos para muestras futuras. A su vez, lo que reviste mayor importancia es que al no utilizarse la información primaria, las cifras de RMSEV son mucho mayores. Es decir, se obtienen mejores resultados utilizando las 10 muestras secundarias para transferir un modelo que también reutiliza información primaria, en relación a su sólo uso para calibrar. En partes anteriores de este texto, se ha sentado que la importancia del proceso de transferencia radica en el ahorro de recursos, pues se confía en que los datos primarios reutilizados brindarán una “estructura” sobre la cual podrán reposar los modelos actualizados con información secundaria. Los resultados en la curva sec10 muestran que esa no es la única ventaja, sino que una transferencia puede lograr lo que quizá sería sencillamente inviable sólo con la información secundaria. Igualmente bien podría decidirse utilizar un modelo secundario obtenido a partir de pocas muestras sólo por poner énfasis en la certeza de que las muestras futuras estarían siendo unicamente representadas por otras de su mismo dominio, pero esa decisión deberá tomarse a sabiendas de que la calidad de las predicciones estaría en condiciones de ser mayor a la que se obtendrá. Finalmente, vale también destacar que en sec10 se utilizó el centrado clásico MC2, es decir, todo se centró a la media de los 10 calibradores secundarios, pues en estas condiciones no existirían datos más apropiados con los cuales centrar al resto de las muestras secundarias.

SAC4-15 En esta curva se puede observar el comportamiento de la transferencia SAC4 reutilizando sólo la mitad de la información primaria de Calibración, es decir, 15 muestras seleccionadas desde las 30 originales (en notación, 1:2:30), lo cual podría interpretarse como un intento de aumentar la trascendencia secundaria sobre la primaria. Puede apreciarse que esta metodología no es favorable, ya que los RMSEV aumentan considerablemente respecto de otros con 30 muestras primarias. Debe por tal notarse que a pesar de haber enriquecido la proporción de muestras de transferencia secundarias respecto de las primarias (como en SAC10), las muestras de Validación no se ven mejor representadas y la información enteramente reutilizada es ciertamente útil. A su vez, por contener menos información para modelar, las normas de estos vectores resultaron muy bajas, por lo que para la obtención de esta curva fue necesario reutilizar los mismos

13 valores de tau que en las 30 ejecuciones al azar con SAC4.

SAC15-15: De forma similar a la anterior, se utilizaron las mismas 15 muestras primarias, pero en la transferencia se utilizaron 15 muestras secundarias (correspondientes a las 15 primarias que no fueron usadas). A diferencia del caso anterior, no fue necesario reutilizar los 13 tau de m093S, ya que estos modelos contuvieron 30 muestras totales y sus normas estuvieron a la altura de la zona analizada. Si bien no tiene sentido ahorrativo una transferencia con tantas muestras, se pretendió evaluar el comportamiento con aportes iguales (en número de muestras) de la información de ambos dominios. Se observa que los RMSEV mejoran respecto de SAC4-15, pero no llegan al nivel de los frentes vistos a la izquierda de la figura. Nuevamente, se deduce que la reutilización de la información primaria es trascendental.

sec15: Las mismas 15 muestras secundarias de transferencia en SAC15-15 fueron utilizadas para elaborar modelos RR secundarios directamente, sin intervención de información primaria (como se explicó en sec10). Puede observarse una mejoría respecto de sec10 y a su vez que los RMSEV ya estarían en valores cercanos al objetivo de 0.1655, pero aún así no se obtendrían resultados mejores que los obtenidos con transferencias y reutilización de toda la información primaria. Por ser modelos simplemente secundarios, se centró con MC2.

Hasta aquí, se evaluaron comportamientos sin haber buscado valores óptimos para lam. Se pudo observar que la información primaria reutilizada es conveniente a pesar de que por sí sola ya no predeciría bien, y similarmente se observó que el sólo uso de la información secundaria disponible no es mejor que su combinación con información primaria. A su vez, se apreció que con 3 ó 4 muestras básicamente se obtendrían resultados aceptables. En este punto del desarrollo, todo esto habla a favor del proceso de transferencia.

Habiendo evaluado el comportamiento de DR-SAC, se prosigue con el de DR-DIFF y variantes, cuyos resultados se pueden apreciar en la figura 13.

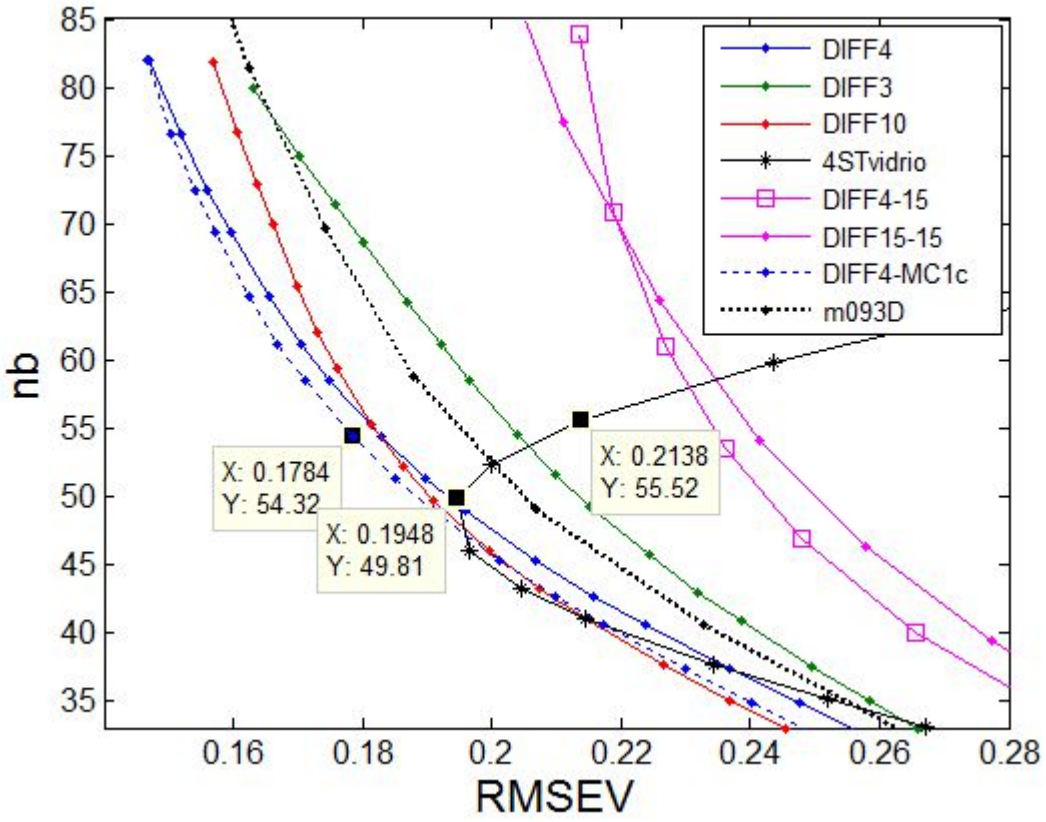


Figura 13: RMSEV versus norma de los vectores de regresión (nb) para DR-DIFF y variantes en lam=1, datos "Maíz"

Referencias: DIFFn: DR-DIFF con n muestras de transferencia y 30 calibradores primarios, 4STvidrio: DR-DIFF4 a partir de estándares de vidrio (ver texto), DIFFn-m: DR-DIFF con n muestras de transferencia y m muestras de Calibración primarias, DIFF4-MC1c: DR-DIFF4 utilizando una variante de centrado de datos (ver texto), m093D: media de 30 ejecuciones al azar para DR-DIFF4 en lam=0.9313. Los recuadros amarillos representan puntos de interés. X es RMSEV e Y es nb (ver texto)

El análisis de las curvas de la figura 13 es el siguiente:

DIFF4, DIFF3 y m093D: La transferencia con 4 muestras resultó apreciablemente mejor que con 3 para RMSEV, lo cual era previsible si se recuerdan las experiencias con 30 ejecuciones al azar. Al respecto, la curva media de éstas últimas en su lam=0.9313 permite afirmar que el conjunto de 4 muestras de transferencia utilizado generó resultados superiores en calidad, fundamentalmente en RMSEV. Lo más importante que hay que destacar en este punto del desarrollo es que DIFF4 y DIFF3 obtuvieron en tau 9 normas levemente menores a las del modelo original (no remarcado), lo cual no estaría de acuerdo con las tendencias observadas en la elevación de la norma con el aumento de la información a modelar. Si se quisieran respetar las tendencias vistas, se podría optar

por el siguiente tau en la secuencia (10), tal que la norma superara a la del vector original, o bien podría mantenerse el valor de tau y luego podría buscarse que a través del cambio de lam se recuperaran las tendencias. Obviamente, también podría ignorarse lo visto anteriormente. Experiencias posteriormente reportadas darán cuenta de estos asuntos.

DIFF10: La elección de muestras para obtener diferencias se hizo de manera equivalente a lo explicado en SAC10. La transferencia con 10 diferencias no resultó tan provechosa como en aquel caso y de hecho solo se superaron los resultados de DIFF4 en normas bajas. No obstante, estos son resultados en lam=1 y en otros valores podrían esperarse mejorías, aunque igualmente no sería suficiente para justificar el uso de 10 muestras para el procedimiento si con 4 sería suficiente para el caso. Más aun, la curva Pareto-óptima en la figura 13 se obtuvo con una variante del centrado MC1 (posteriormente será explicada). Si a DIFF10 se lo modifica con dicha variante, los resultados básicamente no se modifican.

4STvidrio: Los estándares de vidrio produjeron resultados más que aceptables si se tiene en cuenta que no portan información de composición en sí, a la vez que en la práctica serían más fáciles de conservar que muestras de transferencia en ambos dominios. Para realizar el cálculo de modelos tuvieron que hacerse algunas adaptaciones. Esto fue debido a que dichos estándares poseen 3 mediciones en el instrumento primario y 4 en el secundario, por lo cual se necesitó optar por una medición primaria para repetirla y así poder generar las 4 diferencias. Evaluando gráficamente los 7 espectros, se observó que las correlaciones entre mediciones primarias y secundarias no eran constantes entre muestras, es decir, según qué picos se observaban había determinadas correlaciones. Por lo tanto, la obtención de la cuarta muestra primaria necesaria se realizó al azar, resultando en la primera de ellas (vale destacar que otra forma de resolver la disparidad consistió en promediar los espectros de cada dominio y realizar las transferencias a través de sus únicas diferencias, pero los resultados no fueron mejores). Una vez que esto fue realizado, la transferencia procedió normalmente con las diferencias entre espectros. En la curva graficada se remarcaron 2 puntos. El inferior (X=0.1928, Y=49.75) corresponde a tau 7 y presentó el RMSEV más bajo. El superior (X=0.2085, Y=55.46) corresponde a tau 9 y presentó una norma mayor a la del modelo RR original (54.86). En ambos casos los resultados no fueron tan buenos como con diferencias de muestras reales, pero como ya se dijo, si se aceptaran errores en RMSEV como los presentados, la comodidad práctica sería inconmensurable, pues estos mismos estándares podrían ser usados una y

otra vez si es que el instrumental volviera a presentar cambios no modelados. Vale destacar que de usarse este tipo de estándares ya debería existir conocimiento previo de su aplicabilidad, es decir, deberían haber sido previamente validados de alguna manera.

Más allá de que en el contexto de este escrito es sabido que estas transferencias son realizadas entre instrumentos distintos, los resultados con los estándares de vidrio sugieren que la diferencia fundamental entre la situación primaria y la secundaria es instrumental. Si simultáneamente existieran otros efectos del tipo químico, físico, ambiental, etc., entonces el uso de estos estándares no sería suficiente para caracterizar de forma aceptable a las nuevas condiciones con el fin de actualizar los modelos y a su vez se esperaría que fueran necesarios ponderadores individuales de cada efecto (y representaciones de cada efecto para **L**) como en la ecuación (19).

DIFF4-15 y DIFF15-15: Debe recordarse que en el primero de estos casos aumenta la proporción de información de transferencia por disminución en el uso de información primaria mientras que en el segundo ocurre lo mismo sólo que a su vez existe un aumento de información secundaria. Los resultados se muestran con calidad muy inferior a los expuestos en otras curvas de la misma gráfica y también en comparación con las respectivas experiencias de SAC. Nuevamente, lo importante aquí es notar la importancia de reutilizar la información primaria completa. También en el caso de DIFF4-15 la norma de los vectores resultó más baja de lo normal (menos información para modelar), por lo que debieron reutilizarse los mismos 13 tau que en las 30 ejecuciones al azar.

Vale destacar que a diferencia de DR-SAC donde se probaron modelos del tipo secN, en DR-DIFF no tiene sentido elaborar modelos secundarios puros a partir de las diferencias entre muestras, puesto que todas las concentraciones calibradas serían 0.

DIFF4-MC1c: Para realizar esta experiencia se realizó una modificación del centrado MC1 que había sido adoptado. Esta modificación utiliza los valores de referencia de las muestras secundarias de transferencia, así como sus espectros, para realizar centrados y escalados para **V**. En general los valores de referencia serán concentraciones y de allí el uso de la letra “c” detrás de MC1, más allá de que en el caso de datos “Maíz” no sean concentraciones sino contenido proteico. La aplicación de esta variante no sería posible en cualquier caso. Por ejemplo con los estándares de vidrio no se podría aplicar, dado que éstos no contienen valores de referencia. Tampoco se podría aplicar en el hipotético caso en que sólo se contara con la información de las diferencias espectrales sin saber qué valores de referencia tendrían las muestras que les dieron origen, aunque estos casos no serían

los más comunes, pues normalmente se sabrá de donde proviene la información.

Como en MC1, **X** es centrada en sí misma, **L** no es centrada ni tampoco lo es $\mathbf{y}_L$. Sin embargo, la modificación consiste en que los espectros de muestras de Validación secundarias serían centrados con la media de los espectros secundarios que dieron origen a las diferencias presentes en **L** (es decir **L2**, o bien la **L** que se utiliza en SAC). Esto es similar a MC3 (sólo que **L** sigue sin ser centrada y su uso sigue siendo directo en la ecuación (16)), donde los espectros son centrados con la media de los espectros de su propio dominio, es decir localmente. Finalmente, hechas las predicciones de los espectros secundarios de Validación centrados, éstas serán re-escaladas utilizando la información de los valores de referencia en $\mathbf{y}_L$ (no ceros, sino como en $\mathbf{y}_L$ para SAC) pertenecientes a su mismo dominio. Como se aprecia, esto otorgó los mejores resultados en términos de RMSEV. La curva marca el frente donde las soluciones son óptimas para esa cifra, al menos en buena parte del recorrido de normas y fundamentalmente en la zona de interés. Las normas de los vectores con con 4 muestras en **L** serán siempre iguales más allá de si se utiliza MC1 o MC1c, ya que la diferencia entre éstos no produce un efecto durante la etapa de obtención de modelos (donde se determina la norma) sino sólo en las predicciones de las muestras de Validación. A su vez, en la curva para MC1c se resaltó un punto (X=0.1784, Y=54.32), correspondiente al tau 9, donde se aprecia que la norma resultó nuevamente menor que la del modelo primario original. Posteriormente se evaluarán cuestiones relacionadas a lo último.

Se recalca que el uso de una estrategia de centrado del tipo local como MC1c se postergó hasta aquí puesto que no hubiese sido aplicable para el caso de los estándares de vidrio.

Habiendo analizado variantes de SAC y DIFF, corresponde ahora comparar los mejores resultados con los obtenidos mediante otras estrategias. Vale aclarar que para el caso de SAC no se expone como el mejor resultado al de SAC10, puesto que éste fue obtenido con fines analíticos pero no representa una transferencia ahorrativa en sí. En la figura 14 fueron graficados los resultados obtenidos con SAC4 y DIFF4-MC1c, representando a las mejores variantes obtenidas hasta ahora. Se aprecia que cualquiera produciría mejores soluciones respecto de las de su modelo PLS equivalente aumentado (a4PLS para SAC4, a4PLS-MC1c para DIFF4-MC1c), lo cual se observa a cualquier valor de nb y RMSEV. Por lo tanto las soluciones DR propuestas son Pareto superiores a las de PLS, a pesar de haber utilizado la misma información y metodologías de centrado.

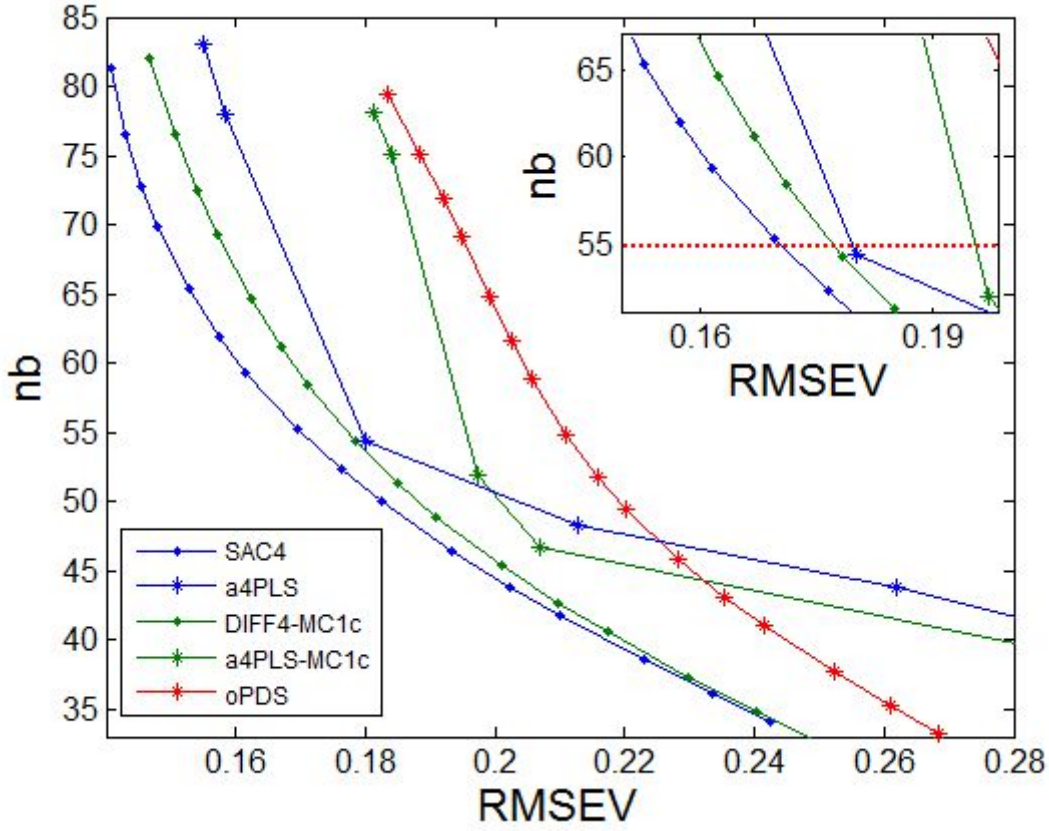


Figura 14: RMSEV versus norma de los vectores de regresión (nb) para DR-SAC y DF-DIFF en sus mejores variantes con lam=1, para modelos PLS aumentados y para estandarizaciones con PDS, datos "Maíz"

Referencias: a4PLS: Modelos PLS aumentados con las mismas 4 muestras de transferencia que en SAC4, a4PLS-MC1c: Modelos PLS aumentados con las mismas 4 diferencias de muestras de transferencia que en DIFF4-MC1c, oPDS: Predicciones realizadas con los modelos primarios originales para espectros secundarios estandarizados con PDS a partir de las 4 muestras de transferencia primarias y secundarias. La línea de puntos roja en el recuadro inserto señala el valor de nb para el modelo primario original.

A su vez, la comparación entre SAC4 y DIFF4-MC1c deja ver mejores resultados para la primera, y esto mismo se cumple si se comparan otros resultados ya obtenidos. Por ejemplo, ya se ha visto que con SAC10 se obtenían mejorías que no podrían traducirse de forma equivalente para DIFF10 o DIFF10-MC1c. Por ende, todos estos resultados sugieren que para este conjunto de datos, la desensibilización de los modelos proveniente de la ortogonalización propuesta por DIFF no utiliza la información disponible apropiadamente y no resulta tan eficiente como el simple agregado de espectros y concentraciones secundarias en SAC.

Las curvas de la figura 14 no relacionadas a DR se analizan por separado:

a4PLS: Estos modelos se obtuvieron poniendo a prueba 8 Variables Latentes de PLS. La misma **L** proveniente de SAC4 se utilizó para aumentar el conjunto **X** de datos primarios, con un valor de lam implícito igual a 1, es decir, simplemente se agregaron muestras a **X** y no fueron ponderadas ni unas ni otras. A su vez, la estrategia de centrado se corresponde con MC3 de centrados locales. Finalmente las muestras de Validación secundarias fueron predichas y éstas predicciones fueron re-escaladas como en MC3. En la curva expuesta en el recuadro inserto se observa el modelo con 6 Variables Latentes. Dicho modelo explica 99.99% de la varianza espectral y 93.06% de la varianza en las concentraciones. Se observa que el resultado es apenas mayor en RMSEV comparado a los diferentes modelos DR, y siendo que PLS es un algoritmo confiable y de uso extendido desde hace muchos años en el mundo de la Quimiometría, los resultados obtenidos con las estrategias de DR son alentadores. También se destaca que se optó por 6 Variables Latentes aún a pesar de que la norma de dicho vector es levemente inferior a la del modelo primario original, puesto que de optar por 7 esa norma ascendería a 77 aproximadamente, con lo cual las comparaciones ya no serían igualmente válidas dado que sería una zona completamente diferente y no armónica (aun así, en dicha zona los modelos DR se muestran siempre Pareto superiores). También vale destacar que en la curva de a4PLS se observa de forma muy clara que en 6 Variables Latentes se produce un cambio en el ritmo de crecimiento de nb y RMSEV, por lo que queda explícita la zona armónica. Este hecho sugiere que la elección de tau 9 para DR también fue apropiada.

a4PLS-MC1c: Similarmente, en este modelo se aumentó **X** con la misma **L** utilizada en DIFF4-MC1c, el valor de lam implícito también fue de 1 y tal como indica su nombre, la estrategia de centrado fue la más apropiada que se puso a prueba en las experiencias anteriores. En resumen, se aplicó PLS aumentado manteniendo coherencia con DIFF4-MC1c y para el mismo número de Variables Latentes que en a4PLS, el modelo explica 99.99% de la varianza espectral y 91.47% de las concentraciones, siendo éstas cifras similares en ambos casos. Básicamente se observa que los resultados no fueron tan buenos como en a4PLS. También se observa que los modelos PLS de 6 Variables Latentes, ambos visibles en el recuadro debajo de la línea roja de puntos, poseen normas distintas además de RMSEV distintos. Esta misma relación se verifica para todos los modelos con igual cantidad de Variables Latentes, a excepción de con 1 ó 2 (no visibles en los gráficos). Por lo anterior y de forma análoga a algunas tendencias observadas en DR, para PLS puede pensarse que

la incorporación de espectros y valores de referencia (tipo SAC) en lugar de diferencias y ceros (tipo DIFF) conlleva un aumento mayor en la norma, en la zona armónica de análisis y sobre esta también. Otro punto de acuerdo radica en que de las diferencias no parecen sacarse las mismas ventajas que de los espectros.

oPDS: Para su realización se utilizaron las 4 muestras de transferencia en juego, tanto primarias como secundarias. La matriz de transformación para estandarizar a los espectros secundarios de Validación hacia el dominio primario se obtuvo con ventana de 3 elementos y tolerancia de los modelos locales con un valor de 1%. Finalmente los espectros secundarios estandarizados fueron predichos por los modelos primarios originales, los cuales aportaron también los valores de norma de esa curva. Como se aprecia, los resultados obtenidos fueron básicamente los de menor calidad y la única ventaja que podría representar PDS en este caso es que los valores de referencia de las muestras de transferencia no serían necesarios si se tiene certeza de que estas muestras pueden considerarse equivalentes entre ambos dominios. Estos valores de referencia tampoco fueron utilizados en oPDS para centrar, pues se supone que PDS convierte directamente los espectros secundarios en primarios y luego los últimos son predichos, pero como éstos tendrían características primarias adquiridas, no deberían ser centrados con información secundaria de transferencia, sino con información primaria proveniente solamente de **X**.

Vale destacar que las variantes de PLS utilizadas podrían haber sido puestas a prueba con valores de lam distintos de 1. De la misma forma, las estandarizaciones espectrales provenientes de PDS pudieron haber sido profundizadas evaluando distintos tamaños de ventanas y tolerancias. No obstante, no es objetivo de esta parte del trabajo ahondar en estos detalles y sólo se utilizaron esas estrategias en sus formas más convencionales para obtener comparaciones con los modelos DR en estudio.

Las figuras anteriores permitieron comparar de forma genérica a SAC y DIFF en algunas variantes, así como también su rendimiento en relación a otras metodologías como PDS y PLS. A continuación se exponen resultados más detallistas sobre las curvas de SAC4 y posteriormente se realizará algo similar para las correspondientes a DIFF4-MC1c.

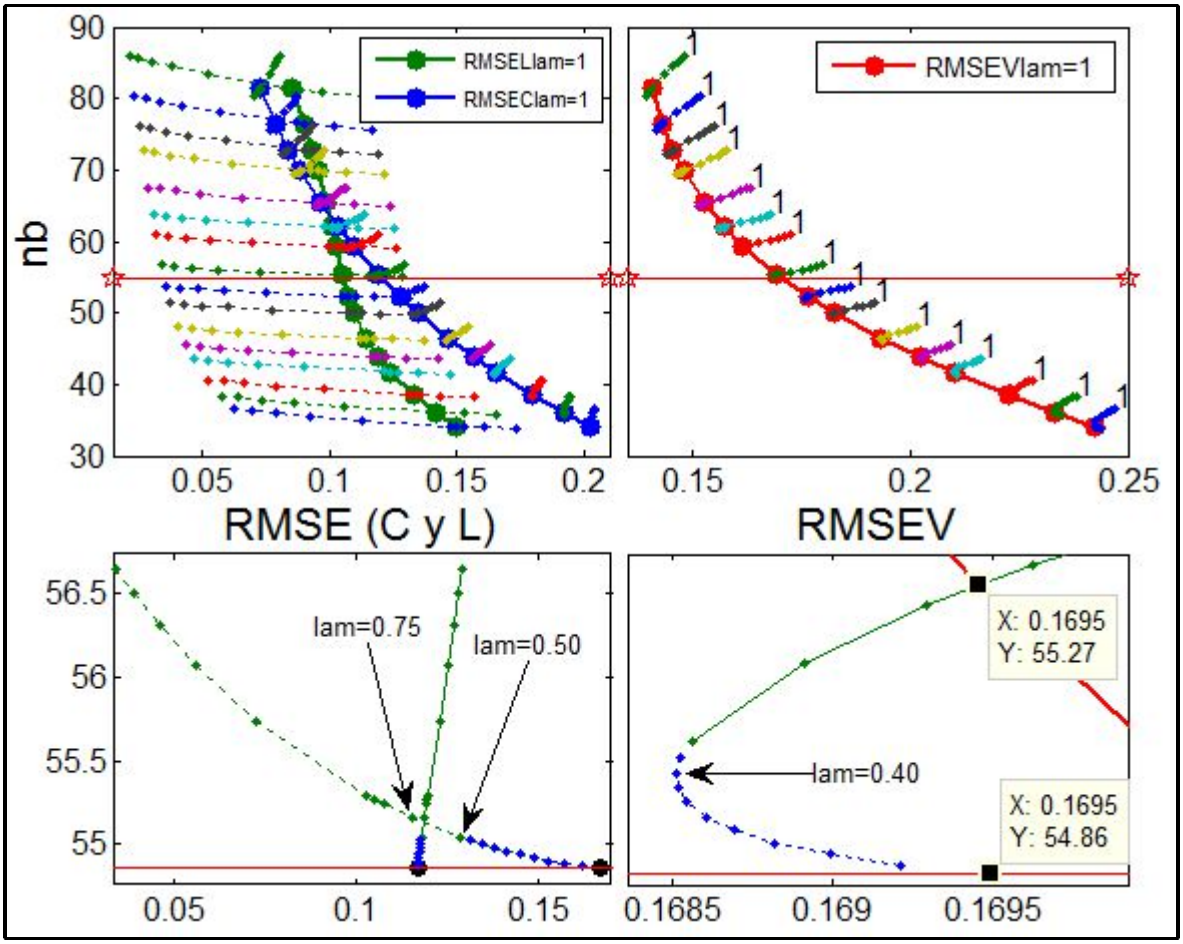


Figura 15: RMSEC y RMSEL (izquierda arriba), y RMSEV (derecha arriba), versus norma de los vectores de regresión (nb) para SAC4, datos "Maíz"

Referencias: La línea entre estrellas rojas señala la norma del modelo original. Los zoom inferiores contienen puntos agregados no presentes en las figuras de origen superiores respectivas (ver texto). Los recuadros amarillos representan puntos de interés. X es RMSE e Y es nb (ver texto)

Para obtener todos los modelos graficados en la parte superior de la figura 15 se utilizaron los valores de tau expuestos en la tabla 2 y los 10 valores de lam comunes entre Maxilam y 0.5 según la estrategia de generación de valores descripta.

El análisis de las curvas representadas en la figura 15 comienza por los resultados de RMSEC y RMSEL. A partir de las curvas en lam=1 respectivas puede observarse la relación de ajustes cuando el coeficiente de ponderación para **X** y para **L** es 1 (recordar que para **X** siempre es 1). En un amplio intervalo de normas inferiores y medias el ajuste favorece a **L** y en las normas superiores esto se invierte. A su vez las tendencias son distintas, pues a medida que decrece la restricción sobre las normas vectoriales el ajuste a **X** es cada vez mayor con aumentos de norma apreciables pero no

exagerados, mientras que el ajuste a **L** llega hasta cierto error y luego de eso el ritmo de decaimiento de estos errores es básicamente menor. No obstante, lo visto no será una constante que uno podría evidenciar en cualquiera de estos gráficos, ya que esta relación dependerá fundamentalmente de las características de las muestras en **L** y variará entre transferencias con muestras distintas.

En cuanto a las curvas para cada tau (cada uno en un color), se aprecia que las de RMSEC (líneas continuas) están bien comprimidas, es decir que en el intervalo de lam puesto a prueba, para todo tau se observa que el conjunto original de Calibración será levemente modificado en esa cifra, lo cual tiene sentido si se tiene en cuenta que son 30 muestras siendo afectadas sólo por 4. Por el lado de las curvas para RMSEL (líneas de puntos), se observa una variación mucho más significativa con el cambio de lam. Siendo que en **L** hay sólo 4 muestras, la sensibilidad es mayor ante cualquier mejoramiento en el ajuste. En los tau inferiores puede verse que aún con el valor más pequeño de lam puesto a prueba (0.5) **L** obtiene errores menores que **X**, pero a medida que los valores de tau decrecen **X** es mejor contemplada y para que **L** obtenga errores similares deberán utilizarse valores de lam mayores. Al respecto, dada la relación entre cantidad de muestras en **L** y **X**, Maxilam tomó el valor de 7.5 y los modelos provenientes de este valor se observan a la izquierda de las curvas para todo tau con el menor de los errores para **L**, siendo estos siempre mucho menores que los respectivos RMSEC. También puede observarse que para todo tau se obtienen normas mayores a medida que lam crece, lo cual está de acuerdo con tendencias previamente comentadas.

En la gráfica de RMSEV se insertaron múltiples “1”, uno por cada tau, señalando el orden de lam (no su valor). Puede apreciarse que a medida que crece el valor de lam hacia el ordinal 1, el cual corresponde a Maxilam, se producen detrimentos cada vez mayores tanto en nb como en RMSEV, lo cual podría ser genéricamente denominado “efecto de expulsión del origen”. En primera instancia esto no parece tener lógica, puesto que lam pondera a **L**, la cual representa a **V** por ser del mismo dominio, y sin embargo RMSEV nunca mejora aun cuando sube nb. Esta misma tendencia, aunque no fue graficada, también se dio en la evolución de las 30 ejecuciones al azar, por lo que no es un problema del conjunto de transferencia, sino algo genérico. En este punto, conviene observar el zoom debajo de RMSEV, correspondiente al tau 9 (en verde, sobre la línea roja de la norma original). La curva verde representa a los modelos para dicho tau y el modelo observado más abajo que el resto corresponde a lam=0.5, el menor de los valores de lam probados normalmente. La curva azul (línea de puntos) fue agregada y pertenece a modelos calculados con valores de lam extra, entre 0.45 y 0.05, decayendo en 0.05 unidades (en notación: 0.45:-0.05:0.05). El punto

señalado con una flecha corresponde a lam=0.40 y estrictamente hablando es el óptimo en RMSEV. El punto señalado con un recuadro amarillo (X=0.1695, Y=54.86) sobre la recta roja es el modelo con lam=0, en cuyo caso la minimización no tiene en cuenta a la información de transferencia. A su vez, este punto es en sí el modelo original prediciendo al conjunto secundario de Validación, sólo que éste ha sido previamente centrado con la media de **L** como indica MC3, y no con la media de **X**, como si fuera la predicción de cualquier muestra con el modelo primario original. Por otro lado, **X** se centra a su propia media, entonces la minimización no sólo que no depende de **L**, sino que a su vez la información de centrado de los términos de la minimización también depende solamente de **X**, o dicho de otra forma, básicamente no existe ninguna influencia de la información de transferencia más allá de su efecto de centrado. Por consiguiente, la corrección de la deriva espectral de estos datos proviene casi exclusivamente de la estrategia de centrado y no de DR-SAC. El único aporte extra de DR-SAC en este caso es la diferencia de RMSEV entre el punto en lam=0 y el producido con lam=0.40, el cual puede considerarse despreciable si se observa que la variación de RMSEV se da en el tercer decimal. A su vez, entre estos 2 puntos se produce una evolución con lam que eleva nb pero reduce RMSEV, mientras que luego de lam=0.4 se observa verdaderamente el efecto de expulsión del origen. Otros estudios sobre transferencia de Calibración habían sugerido que en ocasiones, el solo hecho de centrar a las muestras de Validación secundarias con la media del conjunto de transferencia obtenida también en el dominio secundario sería suficiente pre-procesamiento para permitir la reutilización de un modelo primario (Anderson y Kalivas, 1999; Swieranga y col., 1998). Dicho de otra forma para el presente caso, el centrado local de los datos con las respectivas medias del dominio del cual provenían colocó a los espectros secundarios muy apropiadamente en el espacio multivariado ocupado por las muestras de Calibración primarias ya auto-centradas. Estas observaciones están de acuerdo con que la diferencia más importante entre dominios es instrumental (deriva) y no química o física.

También vale aclarar que es casual la igualdad de valores de RMSEV cuando lam=0 y cuando lam=1, lo cual puede observarse con los valores resaltados en los cuadros amarillos dentro del zoom de RMSEV. Si se analizaran otros valores de tau, se observaría que los valores de RMSEV diferirían además de los de nb.

Observando el zoom de la izquierda, puede verse la relación entre RMSEC y RMSEL. Los puntos señalados marcan entre qué valores de lam un error comienza a ser mayor que el otro. Se aprecia que ninguno es el óptimo de 0.40 y a nivel algorítmico sería bueno poder deducir desde RMSEC y RMSEL cuál sería el óptimo para RMSEV. Quien escribe puso a prueba todo tipo de

comparaciones, utilizando transformaciones logarítmicas, derivadas de orden 1 y 2, escalados, entre otras estrategias. Nunca se puso en evidencia un mecanismo de selección que pudiera ser aplicado a la curva de cualquier tau evaluado en cualquier intervalo de lam y para cualquier conjunto de transferencia, pero sí se puede decir que un método de selección de lam basado en la igualdad de errores entre **X** y **L** no tiene por qué conducir a soluciones pseudo-óptimas. Más aun, en ocasiones las curvas de RMSEC y RMSEL no se cruzan en ningún punto, por lo cual nunca llegan a presentar errores iguales.

Vale destacar que en los tau de normas inferiores el aporte de DR-SAC sobre su sólo efecto de centrado fue levemente más apreciable que el recientemente analizado, pero en esa zona no existe armonía y los errores son inadmisibles. También se observaron mejorías apreciables con otros conjuntos de muestras de transferencia, pero similarmente éstas solo tenían sentido porque se encontraban en una zona mejorable más allá del centrado, sin llegar a los mínimos de RMSEV vistos aquí. También vale comentar que entre estos conjuntos uno fue analizado en detalle, seleccionado a partir de los valores de referencia para contenido proteico en el conjunto de Calibración, de forma tal de abarcar aceptablemente bien dicho intervalo con 4 muestras, a pesar de que dada la complejidad de una matriz vegetal como “Maíz” se puedan encontrar muestras con contenidos similares pero de naturalezas muy diferentes y con las consiguientes diferencias espectrales, por lo que no es posible evaluar una correlación directa entre espectros y valores de referencia (distinto será el caso de datos “Temperatura”). No obstante, los resultados no fueron mejores a los expuestos, pero esto último sólo se comenta anticipando algo relacionado que será visto durante el análisis de datos “Temperatura”, donde el método de selección de Kennard-Stone se mostró deficiente. Sin embargo, en el presente conjunto de datos dicho método funcionó muy bien, ya que proveyó de muestras en **L** con medias espectrales y de valores de referencia lo suficientemente representativas como para que con sólo centrar a las muestras de Validación básicamente fuera suficiente para reutilizar el modelo primario. Esta observación pone en evidencia la importancia del conjunto de transferencia, no sólo para abarcar apropiadamente a las condiciones que debe representar, sino también por el hecho de capturar una buena aproximación de los datos medios.

Adicionalmente, si en lugar de realizar SAC4 se utilizan las mismas 4 muestras de transferencia, pero utilizando sólo 3 para el cálculo de modelos y la restante para evaluar dichos modelos (cual muestra de Validación única), y esto se hace en todas las combinaciones posibles (4 en total), pueden luego evaluarse los distintos comportamientos. En el caso del conjunto de

Kennard-Stone, 3 de las 4 combinaciones mostraron el efecto de expulsión del origen al aumentar el valor de lam. Sólo una de las combinaciones presentó una tendencia inversa (aunque sus resultados eran peores que el resto) y el comportamiento medio de las 4 mostró el efecto de expulsión. Esto se comenta pues a falta de más muestras cuantificadas para determinar si lam debe aumentarse o disminuirse, el procedimiento con las combinaciones podría ya dar una pista de que lam no debería aumentar en este caso.

Finalmente cabe analizar la opción de no mantener el mismo tau para los modelos actualizados, utilizando algún otro criterio, pero partiendo siempre de modelos originales que podrían considerarse armónicos. De entre muchos criterios posibles y aún respetando los centrados MC3 y MC1c para SAC y DIFF, respectivamente, uno de estos podría ser encontrar (gráfica o analíticamente) un valor de tau tal que el error en **X** con lam=1 fuera igual al error en lam=0 para el tau original. Dicho en otras palabras, intentar recuperar el mismo nivel de error original para las muestras de Calibración sin ahondar en la optimización de lam y suponiendo un simple agregado (la ponderación será sencillamente 1) de información actualizada proveniente de muestras secundarias de transferencia, a costa de un posible incremento en la norma vectorial que no sería asumido como un sacrificio en una de las cifras de mérito, sino a lo sumo como una necesidad producto de dicho agregado. Si se observa la evolución de la curva de RMSEC para lam=1, puede deducirse que dicho tau estaría entre el original y el siguiente de la gráfica (en rojo), mucho más cercano al primero, por lo cual todas las cifras de mérito serían básicamente las mismas que en el tau 9 original. Este criterio dará otros resultados en experiencias descriptas posteriormente, y se lo destaca por su simpleza y porque el ascenso de la norma vectorial no se deja librado al azar, sino solamente lo necesario para obtener resultados de calidad similar (ni peores ni mejores) a la que se tenía originalmente para los calibradores primarios.

La figura 16 expone detalles para las curvas obtenidas mediante DIFF4-MC1c. En dicha figura, obtenida con valores de tau como los de la tabla 2, pueden observarse varias cosas sobre DIFF4-MC1c. En primer lugar se han agregado valores de lam, ya que además de los 10 generados con la estrategia descripta cuyo mínimo era de 0.50, se agregaron en secuencia desde 0.45 hasta 0, en decrementos de 0.005 unidades (en notación: 0.45:-0.005:0). Esto dio origen a un total de 101 valores de lam y esa es la razón por la cual algunas curvas se ven muy densas en cuanto a cantidad de puntos.

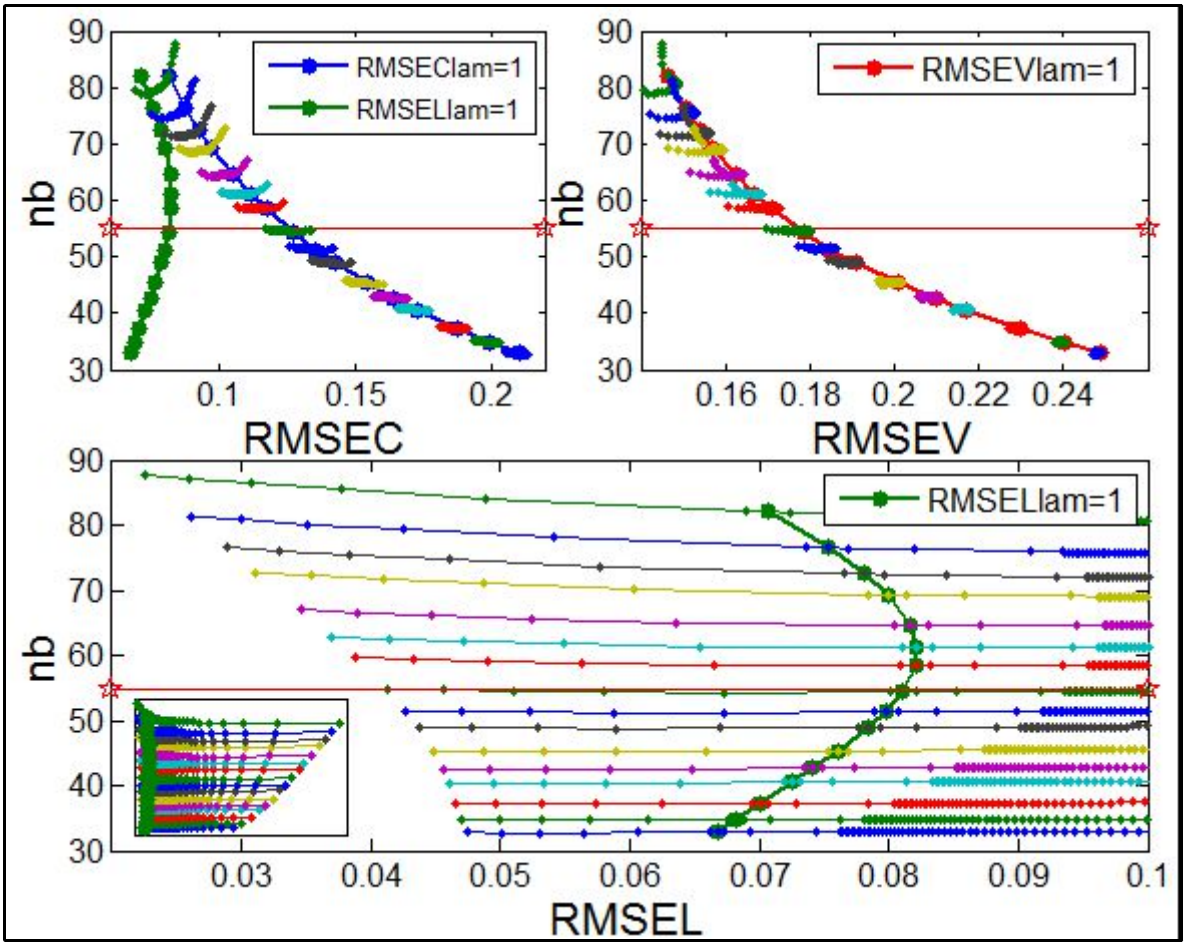


Figura 16: RMSEC y RMSEV (arriba), RMSEL (abajo), versus norma de los vectores de regresión para DIFF4-MC1c, datos "Maíz"

Referencias: La línea entre estrellas rojas señala la norma del modelo original. El recuadro inserto en la gráfica de RMSEL permite ver resultados no visibles en la misma.

De la gráfica de RMSEC puede obtenerse una comparación entre el ajuste de **X** en lam=1 con el respectivo de **L**. Se aprecia que para todo tau **L** se ajusta mejor que **X**, lo cual resulta en un comportamiento diferente al visto en SAC. De esto se deduce que en lam=1, las diferencias insertas influencian realmente los cómputos en favor de su propia predicción. Sin embargo, al observar la curva de RMSEV en lam=1, se aprecia que existen modelos Pareto superiores. En efecto, y sin tener en cuenta a los modelos en normas inferiores donde la modificación con lam es escasa, para los modelos de las zonas media y superior los RMSEV más bajos se obtienen en lam=0 (la evolución desde lam=0 hacia Maxilam será examinada posteriormente). Desde ya se deduce que para este conjunto de datos, el incremento de ajuste de las diferencias con lam no beneficiará directamente al conjunto de Validación secundario. También para RMSEV en lam=1 puede apreciarse una curva de

forma similar a la de RMSEC, lo cual indica la importancia de la información primaria para predecir a la secundaria. En relación a lo anterior, si se utiliza la estrategia de cambiar el tau hacia otro tal que el RMSEC en lam=1 sea igual al RMSEC en lam=0 para el tau original, se obtendrían buenos resultados para RMSEV, teniendo en cuenta que no habría una optimización posterior de lam. Por ejemplo, en la gráfica de RMSEC, el tau siguiente (rojo, 10) en sentido ascendente respecto del original (verde, 9) posee un error de 0.1171 en lam=1. Si se mantiene el tau original, su error en lam=0 es de 0.1170, por lo cual el tau 10 es muy cercano en valor al tau buscado donde los RMSEC se igualan. Si se comparan los RMSEV respectivos, el tau 9 tiene su óptimo en lam=0 (la figura siguiente mostrará detalles) con un valor de 0.1695, mientras que en el tau 10 con lam=1 dicho error es de 0.1711, por lo cual no existen grandes diferencias. Vale destacar que en el tau 10 existen mejores resultados de RMSEV (de hecho, su óptimo estará en lam=0 también), pero el objetivo de lo propuesto como alternativa es no tener que lidiar con la optimización de lam.

Aun cuando los modelos de las zonas bajas no han sido de interés en este desarrollo por estar fuera de la zona armónica pero a su vez por ser siempre los de peores características predictivas, conviene destacar un detalle claramente manifiesto en las curvas de RMSEL y RMSEC con lam=1. Si estas curvas son recorridas desde arriba hacia abajo, a partir de normas aproximadamente de 60 se observa que el ajuste será cada vez mejor para **L** y lo contrario para **X**. A su vez, intuitivamente puede percibirse que si se continuara bajando la norma, el error en **L** eventualmente llegaría a 0 o a un valor muy pequeño (sin siquiera aumentar lam más allá de 1), mientras que el error en **X** tendería a crecer. Aunque esto no es estrictamente así, experiencias no reportadas indicaron que hasta valores de norma cercanos a cifras tan bajas como 1 se verifica lo intuido (no visible en las gráficas). Es decir, cuando el valor de tau es elevado (en relación a los usados normalmente) la restricción sobre la norma vectorial es mayor, por lo cual la norma resultante de la minimización (16) resulta en un valor bajo. Un vector de regresión con baja norma será un vector cuyos coeficientes individuales deberán ser necesariamente escasos en valor absoluto. Por lo tanto aquí vale plantear qué sucedería en el caso extremo (y absurdo ciertamente) de obtener un vector de regresión cuya norma fuera 0, es decir, con la totalidad de sus coeficientes valiendo 0. Con lo visto hasta aquí, se podría pensar que por estar en una zona tan baja la capacidad predictiva sería escasa, pero también debería tenerse en cuenta que un vector de regresión obtenido en condiciones no óptimas (desde cualquier punto de vista) debería predecir todo de forma deficiente, cuando lo que se observa realmente es que las deficiencias sólo se dan en el ajuste de **X** y no en el de **L**. En este momento conviene tener en

cuenta que los valores de referencia para $\mathbf{y}_L$ son impuestos en 0, por lo cual un vector de regresión con norma nula, paradójicamente realizaría sus predicciones para las diferencias en **L** con absoluta exactitud. Por lo tanto en la minimización (16) los términos relativos a la norma vectorial y al error en el ajuste de **L** estarían en valores de 0 y solamente el error para **X** definiría el resultado de la optimización. Este planteo sólo fue realizado para explicar algunas tendencias vistas, pero sirve para sugerir que la introducción de diferencias asociadas a valores de referencia de 0 podría influir a las minimizaciones de una forma inteligible desde lo estrictamente matemático pero no relacionada a solucionar el problema instrumental en cuestión, con lo cual sería esperable que se produjeran efectos extraños en los modelos actualizados con diferencias.

Volviendo a la figura 16, resta analizar la gráfica inferior de RMSEL. Fundamentalmente se hace notar que el zoom inserto en dicha gráfica deja ver hasta dónde crecerán los errores para **L** cuando lam sea 0 o tendiente a 0, es decir, lo mal que serían predichas estas diferencias por los modelos originales. La gráfica de RMSEL deja ver más claramente algo que en las otras no se aprecia por la densidad de puntos: en el tau original la norma de los vectores nunca supera a la norma original, al menos en el intervalo de lam utilizado. El hecho de que el valor 1 esté en dicho intervalo explica por qué en figuras anteriores donde los resultados se exponían sólo en lam=1 las normas resultaban menores a la original para DIFF cuando lo mismo no sucedía en SAC, y esto se tornaba contrario a la tendencia de "más información para incorporar, mayor norma". La figura 17 expone resultados al respecto en varias gráficas. La externa que contiene al resto es la de RMSEL y fue elegida para ocupar ese espacio por la gran dispersión que presentan los puntos en DIFF, ubicados debajo de la norma original. Sobre dicha norma se insertó la curva para SAC en el mismo tau en línea de puntos, y este mismo formato fue aplicado para RMSEC y RMSEV. En la última, a su vez se insertó la curva pertinente al RMSE de las muestras secundarias que dieron origen a **L** (**L2**), y esta inserción se realizó en la gráfica de RMSEV puesto que las muestras en **L2** fueron tratadas como cualquier muestra de Validación secundaria. Esto es posible ya que esas muestras no participan directamente en la elaboración de modelos, sino a través de sus diferencias con las respectivas primarias (**L1**). También en la curva de RMSEV para SAC a propósito se dejaron en color azul los puntos para valores de lam que habían sido oportunamente agregados (en notación: 0.45:-0.05:0.05) y que habían permitido determinar que el resultado óptimo de SAC4 se encontraba en lam=0.40, muy cercano al del modelo primario una vez centradas las muestras de Validación. Esos mismos valores también se agregaron para las curvas de RMSEC y RMSEL de SAC4.

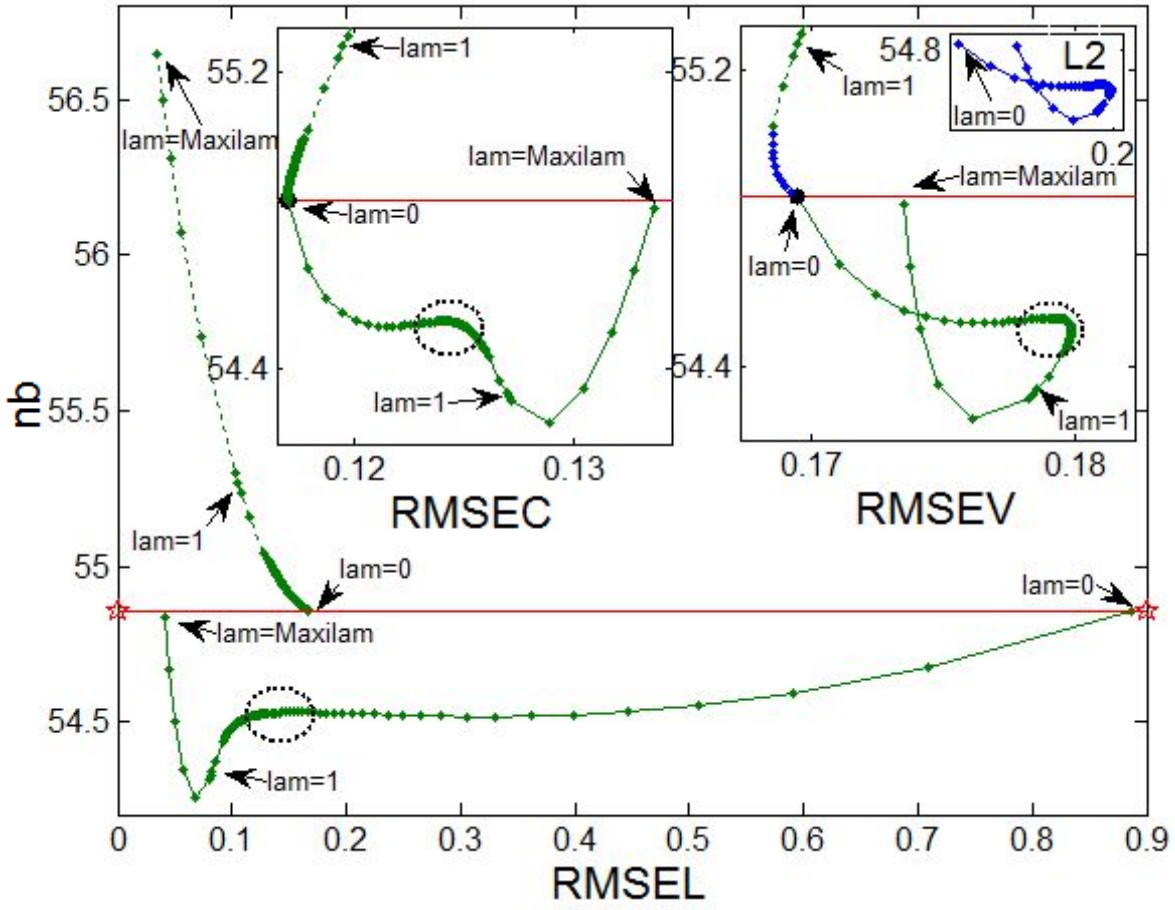


Figura 17: RMSEC y RMSEV (insertas arriba), y RMSEL (toda la figura), versus norma de los vectores de regresión (nb), para SAC4 y DIFF4-MC1c en el tau 9, datos "Maíz"

Referencias: Las líneas rojas señalan la norma del modelo original. Los valores de lam señalados, el óvalo negro en línea de puntos y la gráfica L2 son de interés para su análisis (ver texto).

En primer lugar se observarán las curvas de RMSEC. Como podría preverse, en lam=0 se aprecia una convergencia de SAC y DIFF-MC1c con el modelo primario. Aunque un aumento de lam conlleve en ambos casos detrimentos de RMSEC, lo cual parece bastante razonable, la evolución es muy diferente. A diferencia de lo analizado en SAC4, se aprecia que en DIFF-MC1c disminuye la norma y lo hará incluso más allá de lam=1, hasta que finalmente comenzará a crecer nuevamente con su máximo en Maxilam. Vale aclarar que la norma en Maxilam es cercana pero no igual a la original y esto es sólo casualidad, ya que en otros tau no se verificaría lo mismo. Claro está que si Maxilam hubiese sido mayor, entonces el crecimiento en la norma continuaría, por lo que puede decirse que a partir del valor de lam que produce la menor norma, el agregado de información en DIFF-MC1c también estará de acuerdo con la tendencia en la elevación de las

normas que se observaba fundamentalmente en SAC.

Las curvas de RMSEL para SAC y DIFF-MC1c no se muestran convergentes en lam=0 y de hecho obtienen resultados muy diferentes en favor de SAC. Esto es así dado que en SAC los espectros de **L** ya centrados con MC3, aunque no sean tenidos en cuenta en la minimización, al menos se encuentran bastante bien ubicados en el hiperespacio abarcado por los espectros de **X** auto-centrados. En cambio en DIFF no hay espectros sino diferencias, y éstas estarán en niveles cercanos a cero asociadas a valores de referencia ($\mathbf{y}_L$) exactamente de cero. Sin embargo, para el modelo en lam=0 y debido al auto-centrado de **X**, las señales que estarán cerca de cero serán aquellas cercanas a la media original de **X** y los valores de referencia asociados (**y**) rondarán el valor cero, pero no serán exactamente cero como en el caso de $\mathbf{y}_L$. Esta incompatibilidad entre señales y valores de referencia, mientras lam sea 0, claramente beneficiará a **X** (pues no habrá aportes de la información en **L**) y por lo tanto RMSEL será máximo. Luego, cuando lam tome otros valores y aunque sean muy pequeños, existirá influencia del ajuste de **L** y las mejorías de RMSEL serán permanentes a medida que lam crezca, aunque no serán constantes.

Respecto de la gráfica para RMSEV, todos las cifras de error son superiores a las obtenidas como óptimas en SAC, por lo que se deduce que SAC fue mejor que DIFF-MC1c para este conjunto de datos. Específicamente se aprecia que a medida que lam aumenta hasta una primera elevación de norma (óvalo negro), RMSEV también aumenta y luego comienza a disminuir. Por lo tanto, sólo a partir de la zona remarcada con el óvalo podría pensarse que los modelos introducen información secundaria útil desde **L**, siendo que antes de dicha zona esa información sólo serviría para representarse a sí misma y no a muestras del mismo dominio; o más bien que a partir de la zona (y a pesar de que RMSEC siga aumentando siempre) se recuperen características más del tipo primarias que en este caso resultan más convenientes para predecir a las muestras de Validación. Esto último estaría de acuerdo con el hecho de que el óptimo de RMSEV para DIFF-MC1c se da en lam=0, es decir, en el modelo construido solamente con información primaria, aunque la colocación previa de los espectros de Validación en el hiperespacio de variables implique información secundaria.

Las observaciones realizadas sobre RMSEV se deducirían de la curva L2, con lo cual puede pensarse que las muestras secundarias de transferencia representan muy bien a las de Validación. Como se aprecia, el parecido de formas es notable y queda perfectamente evidenciado que lo mejor para las muestras de Validación sería no incrementar lam más allá de 0. También se destaca que

aunque no fueron graficados en RMSEV, se pusieron a prueba valores en DIFF superiores a Maxilam y nuevamente se observó el efecto de expulsión del origen de coordenadas.

El hecho de haber observado decrementos de norma merece ciertas reflexiones. En principio, que la norma se encuentre bajando no es un indicador de que se obtendrán siempre peores resultados para muestras futuras. De hecho, en la curva de RMSEV se aprecia que en un intervalo de lam hay bajada de norma y RMSEV mejora respecto de donde se encontraba. Segundo, no tiene por qué darse siempre para DIFF, sino que eso dependerá del equilibrio de términos de la minimización (16) y por ende de los datos analizados que aporten la información. Habiendo advertido lo anterior, se quiere comentar que las bajadas de norma podrían estar relacionadas a lo que se ejemplificó con el hipotético vector de regresión con norma=0. Por ejemplo, RMSEL disminuye desde 0.9 hasta 0.7, ambos aproximados, con sólo ir desde lam=0 hacia lam=0.005 (el primer punto luego de lam=0). Es decir, aunque el ponderador del error de **L** sea casi nulo, se obtiene una mejoría de 0.2 unidades, que en relación a los errores expuestos en las gráficas representan un avance muy relevante. Esto significa que en términos de la minimización (16) predecir muy mal a **L** se torna inviable o lejos de ser lo óptimo, y quizá una posibilidad para lograr un valor menor (para toda la minimización, no exclusivamente para el error en **L**) radique en re-adaptar al vector de regresión para predecir mejor a muestras con valores de referencia iguales a 0, lo cual podría tener relación con la bajada en la norma, a costa de aumentar el error en otras muestras, las de **X**. A medida que crezca lam en incrementos constantes, el ajuste de **X** será cada vez menos sacrificado y el de **L** será cada vez menos beneficiado, aunque siempre lo serán mínimamente. En lam ordinal 5 (o en algún valor de lam cercano pero no evaluado) se llegará al mínimo de norma y luego ésta comenzará a elevarse. Vale también destacar que entre lam=0 y lam ordinal 5 la bajada de norma no es absoluta, ya que existe la zona remarcada con el óvalo negro en la cual existe un pequeño incremento. Éste podría señalar algún cambio relevante, pues a pesar de que luego se retome la bajada en las normas, es allí a partir de donde los RMSEV comienzan a mejorar.

Por todo lo expuesto, se concluye que para el caso de instrumentos con deriva se obtendrán buenos resultados a través de centrados locales y cualquier esfuerzo por mejorar esa situación mediante estrategias SAC o DIFF será básicamente inútil. A su vez, la estrategia DIFF implica tener las mismas muestras en ambos dominios, lo cual no siempre será posible en la práctica.

En ocasiones en las cuales no sea posible aplicar centrados locales, como las vistas con los

estándares de vidrio, DIFF-MC1 puede resultar útil (no así MC1c pues los estándares de vidrio claro está que no contendrán proteínas). No obstante, deberá contarse con información previa que garantice que la aplicación de estándares o símiles producirá resultados medianamente aceptables. Entre los aquí analizados, la información proveniente de muestras reales (o de sus diferencias) resultó mejor, pero la elección radicará en las preferencias de los analistas y en el error que uno se permitiría cometer en la cuantificación del contenido proteico. El mejor resultado reportado de RMSEV con estándares de vidrio fue de 0.1928 y el contenido proteico medio de Validación fue de 8.6586, con lo cual se cometería un error aproximado de 2.23%, lo cual no parece nada mal teniendo en cuenta la simpleza de las operaciones con estándares de vidrio en lugar de con muestras reales.

A su vez el conjunto de datos “Maíz” podría representar una situación real en la cual por un determinado proceso se tuvieran que analizar muchas muestras en simultáneo. Por más confianza que uno tuviera, no sería razonable pensar que uno no cuantificaría absolutamente ninguna muestra y simplemente confiaría en un modelo primario hecho en otro tiempo o instrumento. Ante esto, es probable que algunas muestras sean cuantificadas para labores de control mínimas, y por ende de éstas muestras podría obtenerse la información para las transferencias con SAC o bien directamente con un centrado local. El caso de DIFF-MC1c, además de no parecer tan sencillo en su interpretación como de alguna forma sí lo fue SAC, exigiría que las nuevas muestras también pudieran ser obtenidas en el dominio primario y esto puede no ser aplicable siempre. Finalmente, si es posible hacer DIFF-MC1c, también será posible hacer SAC.

1.6.2.2 Datos Temperatura

Para este conjunto de datos no se supone solamente una deriva instrumental. Por lo tanto algunos resultados podrán ser comparados con los vistos con datos “Maíz”, mientras que otros provendrán de las particularidades de esta nueva situación bajo análisis.

1.6.2.2.1 Valores de tau

Los valores de tau seleccionados corresponden a un subconjunto de los utilizados durante las experiencias múltiples y a su vez se aumentó la resolución entre éstos intercalando valores. A excepción de casos en los que fueron reutilizados los mismos valores de tau que en las ejecuciones múltiples (será notificado), la tabla 3 reporta los valores seleccionados.

TAU reportados: Datos “Temperatura”	
Orden	Valor
1 (2)	4.82E-3
2	4.02E-3
3	3.21E-3
4 (3)	2.41E-3
5	2.01E-3
6	1.61E-3
7 (4)	1.21E-3
8	1.00E-3
9	8.03E-4
10 (5)	6.03E-4
11	5.02E-4
12	4.02E-4
13 (6)	3.01E-4
14	2.51E-4
15	2.01E-4
16 (7)	1.51E-4
17	1.26E-4
18	1.00E-4
19 (8)	7.53E-5
20	6.28E-5
21	5.02E-5

Tabla 3: Valores de tau para datos “Temperatura” en ejecuciones únicas

Referencias: Los valores entre paréntesis indican el orden del mismo tau en las experiencias con ejecuciones replicadas al azar. E+n=x10^{n}.

1.6.2.2.2 Conjuntos de Transferencia, Calibración y Validación

Mientras que en las ejecuciones múltiples las muestras primarias (30ºC) de Calibración fueron 10, en las siguientes experiencias se optó por utilizar 13, a través de la inclusión de las muestras 17, 18 y 19 del diseño experimental de la figura 2, todas sin Etanol.

Las muestras de Validación se correspondieron solamente con las secundarias (50ºC) 5, 6, 9, 11, 14 y 15. Este conjunto fue fijo (no variable como en las ejecuciones múltiples) y se puede apreciar que todas estas muestras poseen concentraciones distintas de 0 para los 3 componentes del diseño, aunque sólo se cuantificará Etanol.

Notando que los conjuntos de Calibración y Validación no fueron los mismos que en las experiencias múltiples, deberá tenerse en cuenta que las comparaciones no deberían realizarse de forma directa.

En cuanto a las muestras seleccionadas para realizar las transferencias, se utilizaron 2 conjuntos diferentes:

- Muestras “KS”: El algoritmo de Kennard-Stone (de allí KS) fue aplicado sobre los espectros de las 13 muestras de Calibración equivalentes en 50ºC para obtener 4 muestras de transferencia, que fueron las numeradas 1, 8, 16 y 18 en el diseño. Debe recordarse que este tipo de selección no toma en cuenta a las concentraciones en las muestras.

- Muestras “noKS”: La selección de 4 muestras fue realizada a partir de las mismas 13 de Calibración en 50ºC, pero no con el algoritmo KS (de allí noKS). En este caso, se optó por seleccionar muestras que estuvieran esparcidas en el espacio mixto de concentraciones y componentes del diseño, sin poner énfasis en las características espectrales. Las muestras 2, 8, 12 y 18 fueron las seleccionadas. En el diseño puede apreciarse que éstas muestras “rodean” a las de Validación.

Aprovechando que en el caso de datos “Temperatura” la información proviene de un diseño y por lo tanto la relación entre concentraciones se torna más evidente que con datos “Maíz”, vale destacar algunos aspectos sobre las selecciones KS y noKS.

En las dos selecciones fue incluida la muestra secundaria 18, sin Etanol. Cuando el conjunto de transferencia se compone con muestras conteniendo al analito de interés, se supone que puede obtenerse una mejor representación del espectro del analito afectado por su matriz y por las nuevas condiciones. Distintos niveles de efectos provenientes de la matriz podrán ser representados en función del nivel de similitud entre las muestras de transferencia y las futuras. También puede suponerse que cuanto más sean los aspectos relativos a la matriz que no sean ejemplificados a través de las muestras de transferencia, mayor será la ineficiencia del procedimiento, tal y como se vio con los estándares de vidrio (datos “Maíz”). Todo lo anterior debería ser válido tanto para espectros como para diferencias. El caso específico de la inclusión de la muestra 18 resulta interesante puesto que se estará representando uno de los niveles de concentración realmente calibrados (0, a través de las muestras primarias 17, 18 y 19), pero al mismo tiempo se estarán exponiendo de forma balanceada (50% y 50%) los aportes de agua y 2-propanol a 50ºC, y éstos representan a la matriz también afectada por el cambio de temperatura en los modelos bajo análisis.

Por otro lado, el conjunto KS representa 4 de los 5 niveles de concentración calibrados, mientras que noKS sólo lo hace con 3 (superior, medio e inferior). No obstante, noKS contiene a la muestra 12 en balance con la 8, de forma tal que ambas muestran la misma concentración de Etanol

pero sólo con uno de los otros dos componentes en cada caso. Más detalles relacionados a ambos conjuntos serán expuestos oportunamente.

1.6.2.2.3 Experiencias, resultados y análisis

En primer lugar, se analizan los modelos RR primarios y lo relativo al cambio de temperatura desde 30ºC hacia 50ºC. Los resultados obtenidos se muestran en la figura 18.

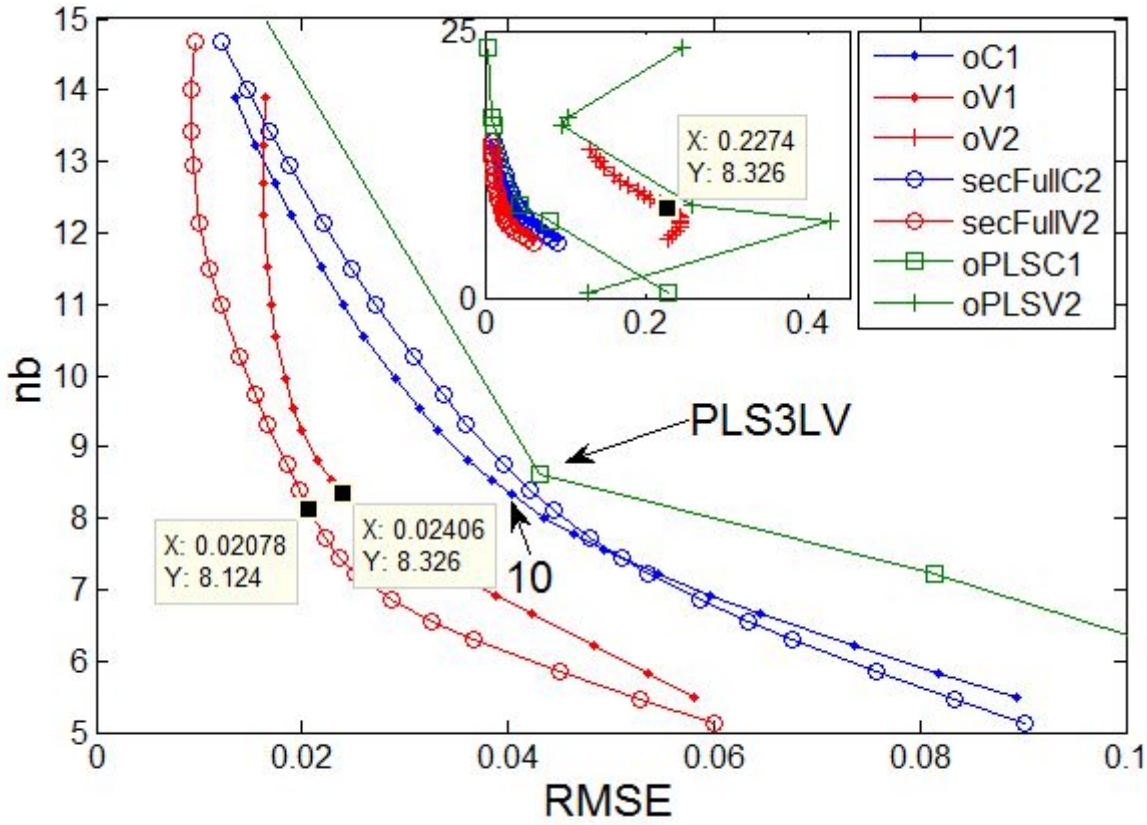


Figura 18: RMSE (C y V) versus norma de los vectores de regresión (nb) para modelos primarios sin transferencias y Re-Calibraciones Completas, datos "Temperatura"

Referencias: oC1: Original RMSEC1, oV1: Original RMSEV1, oV2: Original RMSEV2, secFullC2: Secundario Re-Calibración Total (Full) RMSEC, secFullV2: Secundario Re-Calibración Total (Full) RMSEV, oPLSC1: RMSEC de PLS en dominio original, oPLSV2: RMSEV2 de PLS en dominio original. El recuadro inserto permite ver a todas las curvas de la gráfica. Los recuadros amarillos representan puntos de interés. X es RMSE e Y es nb (ver texto).

En la figura 18 se aprecia de manera poco usual que las curvas oV1 y secFullV2 presentan errores menores que oC1 y secFullC2 (todas provenientes de modelos RR), y esto se cumple

básicamente para todo tau. Es decir, las predicciones promedio serían mejores para las muestras de Validación que para las muestras de Calibración que dieron origen a los modelos en cada dominio. Sin embargo, esto puede entenderse observando el diseño experimental y contemplando que las muestras de Calibración se encuentran en la "periferia" (a excepción de la 10 que es la central). Si se supone que todas tuvieron aproximadamente la misma influencia, entonces todas debieron haber influenciado al resto, por lo cual ninguna en sí lograría una predicción óptima. Por su parte, las muestras de Validación se encuentran "a mitad de camino" entre el centro y la periferia, y por eso obtienen errores promedio menores.

En oC1 fue señalado el tau 10, el cual corresponde a un modelo RR primario con potencial de ser elegido en términos de armonía (como otros cercanos) y que representará al modelo primario original en lo que resta del trabajo. Los valores resaltados en oV1 y secFullV2 corresponden al mismo tau. En este sentido, el RMSEV objetivo de las experiencias con transferencias de Calibración sería 0.02406. La necesidad de transferencia queda plasmada en la curva oV2, donde se ha resaltado el RMSEV que obtendría el modelo primario elegido (0.2274). Similarmente en oPLSV2 se observa que los modelos PLS primarios predecirían a las muestras de Validación secundarias con errores muy apreciables. A su vez, en la curva oPLSC1 se observan las predicciones de Calibración para PLS con distinto número de Variables Latentes. Puede apreciarse que el modelo más cercano al RR en tau 10 de oC1 corresponde a 3 Variables Latentes (señalado con PLS3LV), lo cual es bastante lógico puesto que en el sistema modelado hay 3 componentes reales. Por lo tanto, esto último también es un buen indicador de que el tau 10 seleccionado a partir de criterios armónicos hubiese sido apropiado. A su vez, se aprecia los modelos RR poseen valores de RMSEC levemente menores que los de PLS.

En la figura 19 se exponen resultados provenientes de las transferencias con DR-SAC y aspectos relacionados. En primer lugar, se repitió el gráfico de oC1 donde se aprecia señalado el modelo RR primario en el tau 10 que había sido elegido según criterios armónicos. Con lo visto en datos "Maíz", es conveniente resaltar qué fracción de las mejorías proviene solamente del efecto del centrado local, es decir, cuando lam=0. Para esto, en la curva MC3noKS se resaltó el punto correspondiente al tau 10, donde el RMSEV es de 0.08442. Aunque no fue resaltado en MC3KS, su valor en el mismo tau es de 0.07441. Obviamente, ambos modelos presentan la norma original, pues con lam=0 sólo los datos primarios determinarán a **b** y a su norma. Siendo que el RMSEV para las muestras secundarias proveniente del modelo primario sin transferencia ni centrado es de 0.2274

(ver figura 18), los valores que se obtendrían con sólo centrar localmente serían aproximadamente 67% y 63% menores para los conjuntos KS y noKS, respectivamente. Por lo tanto, el aporte del centrado local es importante, pero no permitirá llegar al RMSEV objetivo de 0.02406. También vale destacar que los conjuntos no realizan su aporte de igual manera en todo tau, y que para ambos se observa expulsión del origen en normas superiores.

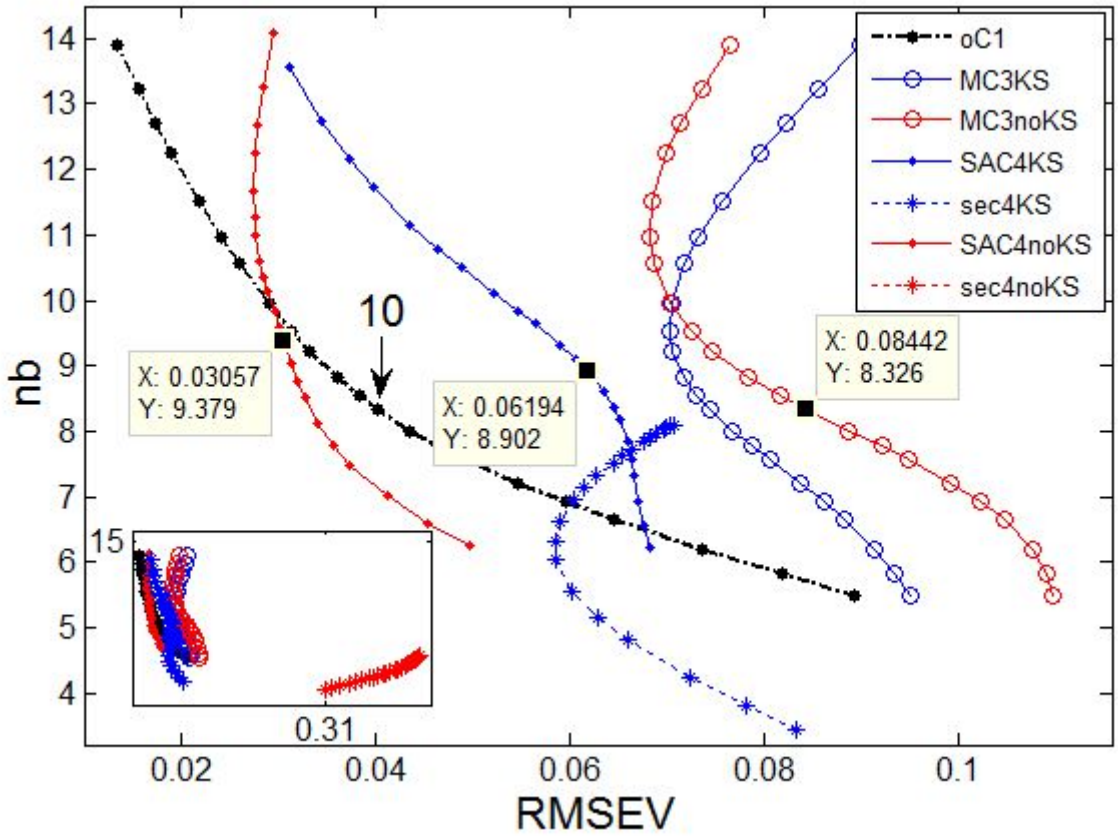


Figura 19: RMSEV versus norma de los vectores de regresión (nb) para DR-SAC en lam=1 y lam=0, y para modelos netamente secundarios, datos "Temperatura"

SACn: DR-SAC con n muestras de transferencia y 13 calibradores primarios, secN: modelo RR secundario con N muestras secundarias usadas para calibrar, KS: conjunto de transferencia con Kennard-Stone, noKS: conjunto de transferencia sin Kennard-Stone, MC3: Aplicación de centrado local con lam=0. El recuadro inserto permite ver a la curva sec4noKS en relación al resto. Los recuadros amarillos representan puntos de interés. X es RMSEV e Y es nb (ver texto)

Antes de evaluar las variantes de SAC4, conviene observar el desempeño de los conjuntos KS y noKS de 4 muestras cada uno para calibrar el dominio secundario. En este sentido, en la gráfica solamente es visible sec4KS, donde se observan normas menores (menos muestras calibrando) y

efecto de expulsión del origen a partir de cierto tau. Su óptimo de RMSEV (no señalado) es levemente menor a 0.06, por lo que se deduce que con el conjunto KS sería posible re-calibrar y obtener errores menores (y también normas) que con sólo centrar localmente. Estos resultados no serían óptimos como algunos vistos ni tampoco optimizables, puesto que al ser el modelo netamente secundario, no existe una posibilidad posterior de optimización como por ejemplo podría ser variando lam. En relación a la comparación con los resultados provenientes de sólo centrar localmente, también se recuerda que en las curvas del tipo "sec" los centrados para las muestras de Validación se realizan sólo con la información de las 4 muestras utilizadas para generar los modelos, pero esta colocación en el hiperespacio antes de la predicción es equivalente a la de las curvas del tipo "MC3". Por consiguiente, la diferencia no proviene de centrados distintos, sino de qué espectros y qué concentraciones aportan información al generar los modelos, siendo todas las muestras primarias de Calibración en las curvas "MC3", y sólo 4 muestras secundarias en las curvas "sec". Por otro lado, la curva sec4noKS sólo es visible en el recuadro inserto y se aprecia que el menor RMSEV que podría obtenerse será cercano a 0.31 en el tau de norma inferior, pues luego existiría efecto de expulsión. A su vez, los resultados serían de calidad muy inferior en relación a los de su respectiva curva "MC3". El hecho de que el conjunto de Validación posea muestras de 3 niveles de concentración podría explicar la diferencia entre la performance de KS y noKS en los modelos "sec", ya que el primero presenta muestras coincidentes en concentración con 4 muestras de Validación (8 con 9 y 11, 16 con 14 y 15) mientras que el segundo sólo lo hará con 2 (8 con 9 y 11). Por ende, si las muestras de transferencia no se usan como tales sino para calibrar directamente el dominio secundario, entonces es probable que lo mejor sea representar tantos más niveles como se espere tener luego en muestras futuras.

Resta analizar los resultados de SAC4 en lam=1. En principio, vale resaltar que en ambos casos las normas resultaron mayores que la original para el tau en cuestión, por lo que se cumplió la tendencia de la elevación de la norma con el agregado de información (noKS lo cumple en todo tau, KS no). La curva MC3KS, más allá de tener una forma diferente a la comúnmente vista para frentes de este tipo, posee señalado el punto correspondiente al tau 10, con un RMSEV de 0.06194 que es incluso superior al óptimo de sec4KS, por lo que el procedimiento SAC con KS no reporta ventajas. Muy distinto resulta el caso de noKS, que obtiene el mejor de los frentes, con un valor de RMSEV en tau 10 de 0.03057. La calidad de los resultados no es coherente con lo apreciado en las curvas "sec" donde KS resultaba mucho mejor que noKS. Sin embargo, en esas experiencias los espectros y concentraciones de las 4 muestras eran los únicos responsables del modelado, mientras que en

SAC4 existe una estructura primaria modelando todos los niveles de concentración posibles y las muestras de transferencia más bien aportan información actualizada. Por ende, en este caso la cantidad de niveles representados no parece tan importante, lo cual se supone era la ventaja de KS sobre noKS en las curvas "sec". Al actualizar los modelos con SAC4, parece más relevante en sí la relación de las concentraciones de transferencia con los espectros, siendo que las primeras serán sólo de Etanol y los segundos provendrán de todos los componentes en las mezclas. Al respecto y como ya se ha hecho notar, noKS presenta un mismo nivel de Etanol en 2 de sus muestras, pero lo hace acompañado en un caso solamente con agua (8) y en el otro con 2-propanol (12). También presenta otros 2 niveles de concentración de Etanol, el superior con la muestra 2 (fracción molar de 0.66) y el inferior con la muestra 18 (fracción molar de 0), y en ambos casos el aporte de agua y 2-propanol también se encuentra balanceado (fracciones molares de 0.16 y 0.50 para cada uno, respectivamente). Otra posible ventaja de noKS puede ser la forma en que rodea al conjunto de Validación, de forma que cada muestra del último será "vecina" directa (en términos de la gráfica del diseño) de una muestra de transferencia. Esto último no se dará en KS (donde la 1 toma el lugar de la 2), donde por ejemplo la muestra 6 no tendrá a ninguna muestra de transferencia tan cercana en composición. Esta última palabra no es trivial, pues los cambios se manifiestan en todos los componentes y no sólo en el Etanol. Más aún, en KS no hay ninguna muestra que sólo contenga Etanol y 2-propanol, ya que la 16 (que toma el lugar de la 12) posee 0.16 de agua. Sumado a ésto, las muestras 1 y 8 de KS también contendrán agua (0.66 y 0.33, respectivamente) y no 2-propanol, por lo que KS es un conjunto fundamentalmente enriquecido en agua que no expondrá tan claramente la interacción entre Etanol y 2-propanol. Si se observa la figura 3, puede apreciarse que el Etanol a nivel espectral es más parecido al 2-propanol que al agua, por lo que el enriquecimiento con la última parece aún menos apropiado. En resumen, KS puede tener representatividad espectral (de hecho el algoritmo de Kennard-Stone se basa en eso) y en este caso también en distintos niveles de concentración, pero carece de una buena relación entre dichas concentraciones y los respectivos espectros.

Habiendo analizado SAC y variantes, en la figura 20 se presentan los resultados obtenidos con DIFF. Como podía anticiparse, la variante de centrado MC1c mejoró los RMSEV para ambos conjuntos de transferencia. En el caso de noKS, se señalaron 2 puntos correspondientes al tau 10. El modelo que produce ambos puntos es el mismo en sí y la diferencia de RMSEV, aproximadamente de 0.027 unidades, proviene del centrado MC1 o MC1c. Como el modelo es el mismo, la norma

también, y se aprecia que es mayor que la norma original (8.326). Respecto de las diferencias entre MC1 y MC1c para KS, los RMSEV en tau 10 (aunque sólo se señaló un dato en la curva de MC1c) difieren en 0.013 unidades aproximadamente, también en favor de MC1c, y además se cumple que las normas son mayores a la original. Por ende, a diferencia de lo observado con datos "Maíz", tanto con KS como con noKS las normas superan a la original primaria y ésto concuerda con el agregado de información, tal como en tendencias ya comentadas. Posteriormente podrán ser evaluadas curvas específicas con lam variables para verificar si existen bajadas de norma no sugeridas en lam=1.

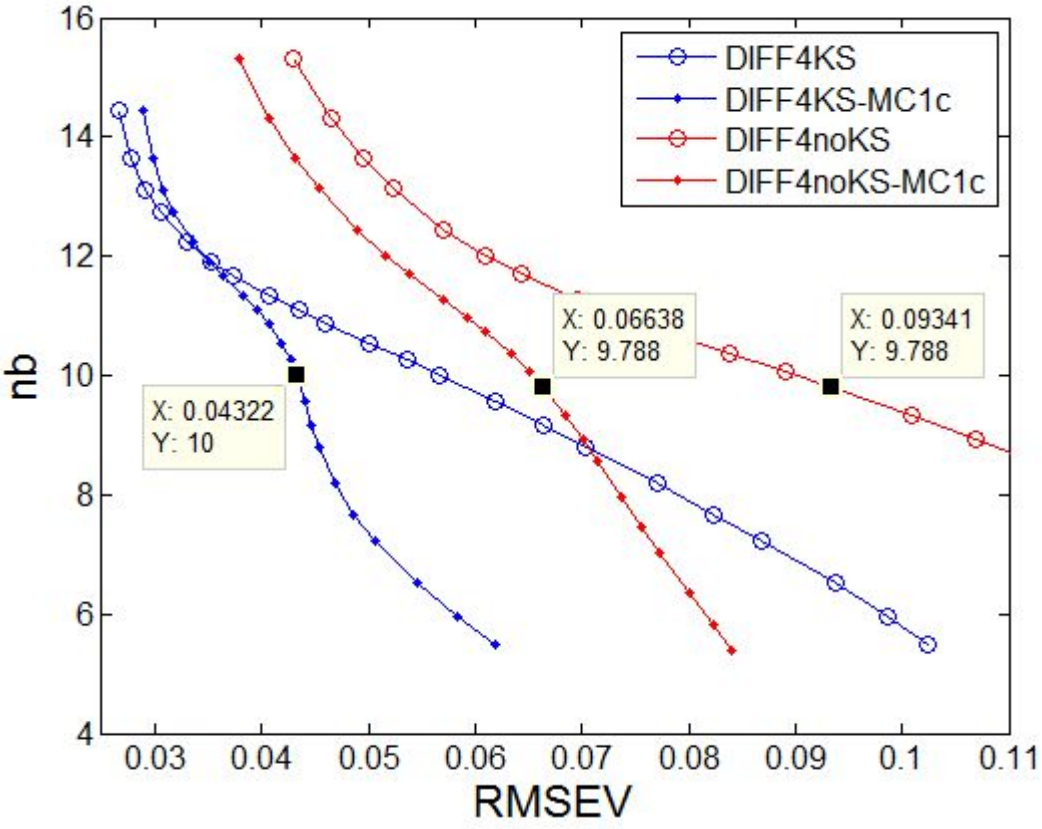


Figura 20: RMSEV versus norma de los vectores de regresión (nb) para DR-DIFF y variantes en lam=1, datos "Temperatura"

Referencias: DIFF4: DR-DIFF4 centrado según MC1, DIFF4-MC1c: DR-DIFF4 centrado según MC1c, KS: conjunto de transferencia con Kennard-Stone, noKS: conjunto de transferencia sin Kennard-Stone. Los recuadros amarillos representan puntos de interés. X es RMSEV e Y es nb (ver texto)

Retomando lo observado para MC1c, vale apreciar que la ventaja que pueda representar sobre MC1 no es la misma para todo tau y se hace más importante en la zona de normas bajas, donde se pueden verificar diferencias grandes entre curvas del mismo color evaluadas en el mismo tau. A medida que las normas suben y supuestamente los vectores de regresión adquieren mayores

capacidades para realizar las predicciones, el efecto de la diferencia de centrados para el conjunto de Validación se hace menor, y en el caso de KS, DIFF4 en normas altas incluso obtiene mejores RMSEV que DIFF4-MC1c. Por lo tanto y teniendo en cuenta que si el conjunto que aporta las medias (KS o noKS) es fijo entonces lo que se le resta a los espectros de Validación y lo que se le suma a sus predicciones será igual para todo tau, el efecto del centrado no debe ser interpretado como un simple movimiento que determina la posición a partir de la cual serán mejor o peor predichas las muestras, sino que a su vez determinados modelos contemplarán de mejor manera a muestras en determinadas posiciones del hiperespacio.

También vale recordar que la media de concentración en noKS es igual a la media de concentraciones de Calibración, y entonces el escalado final de las predicciones para el juego de Validación será el mismo con MC1 o con MC1c, por lo que las diferencias observadas entre ambos solamente pueden provenir del centrado de los espectros de Validación con la media espectral de **X** o con la de noKS, respectivamente. En el caso de KS no puede decirse lo mismo, ya que ni las medias espectrales ni las de concentración son las mismas.

Si se comparan KS y noKS, se puede decir que el primero resultó mejor que el segundo, lo cual no está de acuerdo con lo obtenido en SAC. No obstante, en figuras posteriores se podrá apreciar que esta observación es válida para un amplio intervalo de lam, el cual también evidentemente incluye a lam=1, pero se verá también que si lam puede ser mayor y es optimizada, entonces KS y noKS obtendrán resultados similares. De lo anterior puede deducirse que noKS presentará otra sensibilidad al cambio de lam, y que para lograr resultados óptimos serán necesarios valores de lam mayores a la unidad, es decir, las diferencias deberán ser más ponderadas que la información en **X**. Esta mayor necesidad de optimización para sacar mejor provecho de la información con noKS, así como el hecho de que en DIFF (MC1 o MC1c) con lam=1 KS funciona mejor que noKS, están de acuerdo con que la principal ventaja de noKS sobre KS es la relación entre concentraciones y espectros. A su vez, en KS no es tan perjudicial lo relativo a lo espectral, que de hecho se supone es una de las ventajas del algoritmo de Kennard-Stone para seleccionar muestras a partir de espectros, como lo es su asociación con las concentraciones respectivas. Es decir, cuando sólo se utiliza la información espectral para modelar, como en DIFF, y no la relación de esta información con sus valores de referencia (como en SAC), KS obtiene mejores resultados y en noKS sucede lo contrario, lo cual también podrá ser evaluado en la siguiente figura.

En la figura 21 se evalúan las mejores curvas obtenidas con SAC y DIFF tanto para KS como

para noKS, en relación con los resultados obtenidos por modelos PLS equivalentes y estandarizaciones espectrales provenientes de PDS.

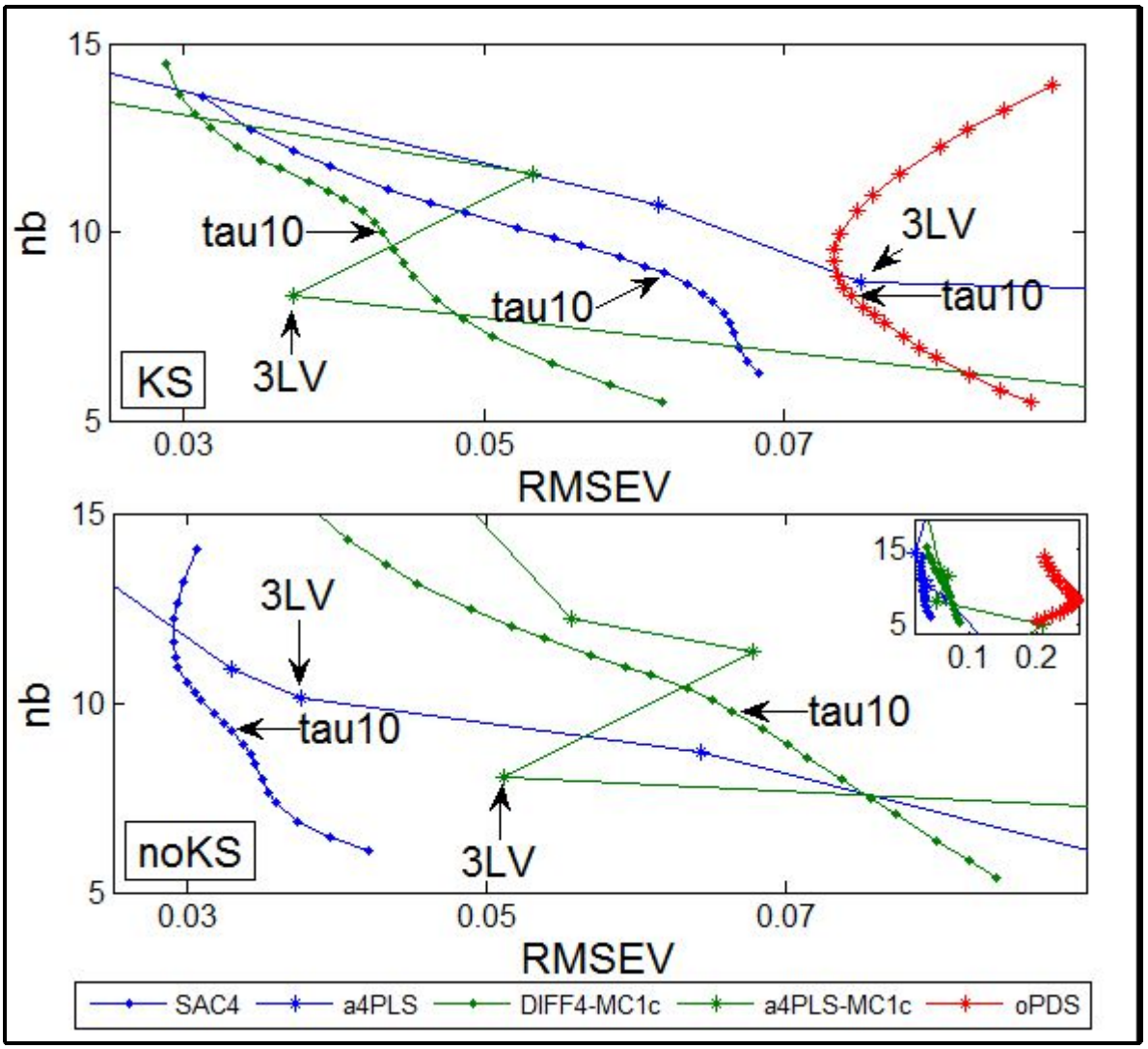


Figura 21: RMSEV versus norma de los vectores de regresión (nb) obtenidos con las muestras de transferencia KS (arriba) y noKS (abajo), para SAC4 y DIFF4-MC1c con lam=1, para modelos PLS aumentados y estandarizaciones con PDS, datos "Temperatura"

Referencias: a4PLS: Modelos PLS aumentados con las mismas 4 muestras de transferencia que en SAC4, a4PLS-MC1c: Modelos PLS aumentados con las mismas 4 diferencias de muestras de transferencia que en DIFF4-MC1c, oPDS: Predicciones realizadas con los modelos primarios originales para espectros secundarios estandarizados con PDS a partir de las 4 muestras de transferencia primarias y secundarias, LV: Variables Latentes para PLS. El recuadro inserto en noKS permite observar también a la curva oPDS respectiva. Los puntos señalados son de interés para su análisis (ver texto).

Las gráficas de la figura 21 serán analizadas de forma individual pero también comparativamente. En primer lugar conviene analizar las curvas SAC4 y DIFF4-MC1c. Así como se vio que para ambos conjuntos de transferencia los modelos DIFF obtenían normas mayores que la original para el mismo tau, también conviene notar que éstas son incluso mayores que las de los respectivos modelos SAC, lo cual es más evidente para KS (ver puntos marcados con tau10). Por lo tanto, el efecto de la inserción de diferencias espectrales y valores de referencia de cero sobre la norma de los modelos es muy diferente al observado durante el análisis de datos "Maíz". Esta diferencia podría sustentarse en algunos hechos. Uno de ellos es que el problema en cuestión para datos "Temperatura" es de otra naturaleza y va más allá de una deriva instrumental. Otro es que el número de muestras en **X** es bajo (13, y 30 para datos "Maíz") y por ende los desajustes de unas pocas muestras primarias de Calibración podrían ya ser suficientes para resultar en errores muy altos en la minimización (16), con lo cual podría ser más probable la obtención de vectores con norma mayor (y no descendente) pero capaces de modelar mejor en simultáneo a las muestras en **X** y a las diferencias de **L**. Un tercer hecho radica en que en datos "Temperatura" existen verdaderamente muestras modeladas cuya concentración es cero.

Otra comparación entre SAC4 y DIFF4-MC1c proviene de sus RMSEV, para lo cual ayuda que ambas gráficas tienen los mismos valores de RMSEV en sus ejes. Viendo que en general para todas las curvas en todo tau, pero en especial en el 10 de interés y con lam=1, se cumple que noKS obtiene el mejor resultado en SAC4 y el peor en DIFF4-MC1c, y que KS ubica a sus resultados entre los dos anteriores, siendo mejor DIFF4-MC1c, queda sugerida aún más la relevancia de la relación entre espectros y concentraciones de cada conjunto.

Las curvas oPDS dejan ver que KS obtiene un resultado aceptable y mucho mejor que el de noKS. Esto también habla en favor de que la información espectral de KS no es problemática en sí (sólo cuando se asocia a sus concentraciones), ya que PDS sólo utiliza espectros y el algoritmo de Kennard-Stone trabaja exclusivamente con éstos. A su vez, los resultados de noKS también están de acuerdo en que si se pierde la relación con las concentraciones de Etanol, como sucede en PDS, se pierde la gran ventaja del conjunto. Debe tenerse en cuenta que las muestras de noKS no fueron seleccionadas contemplando lejanía entre espectros y con esto fuentes diversas de varianza espectral (como sí sucede con KS), algo que efectivamente debe ejemplificar un conjunto de estandarización espectral que se precie de representar varianza futura.

Respecto de las curvas para PLS, vale mencionar ante todo que resultaron muy irregulares en general, incluso para normas no vistas en las gráficas. Esto se pone de manifiesto sobre todo en

a4PLS-MC1c, y se aprecia que para ambos conjuntos en 3 Variables Latentes (LV) se obtienen modelos que no siguen la tendencia del resto en sus respectivas curvas, pero que resultan ser Pareto superiores. No es objetivo de estos análisis profundizar en PLS, pero resulta interesante notar que estos óptimos son obtenidos en el número de LV que coincide con el de componentes afectados, y que el agregado de otra LV en un intento de modelar con una LV más la nueva información, no obtiene mejores resultados. Más aún, el análisis de los % de varianza explicada en cada LV (no mostrado) indica que todos los modelos con 3LV explican aproximadamente el 98% de la varianza espectral, pero a nivel de concentraciones a4PLS y a4PLS-MC1c explican aproximadamente 97% y 85%, respectivamente. Si en los últimos se aumenta el modelo con una LV, los porcentajes ascienden cerca de 98%, pero ésto último no resultará óptimo para las muestras de Validación. Distinto es el caso para aPLS, donde el agregado de LV en la zona de análisis disminuye progresivamente los errores de predicción y aumenta las normas. También vale destacar que en esta zona PLS sugiere las mismas ventajas y desventajas de KS y noKS. Para KS, a4PLS-MC1c obtiene mejores resultados que a4PLS para 3 y 4 LV, y lo inverso ocurre para noKS. Es interesante que los modelos SAC y DIFF puedan interpretarse de forma similar a los PLS (a4PLS y a4PLS-MC1c, respectivamente) en cuanto a qué sucederá con los modelos si el tratamiento de la información es similar para todos los algoritmos, o bien en cómo las relaciones de información (espectral y de concentraciones) afectarán los cálculos. Finalmente, se destaca que viendo los óptimos en todas las curvas el conjunto noKS con SAC4 aporta los frentes que en general podrían considerarse Pareto superiores al resto, y también que los modelos SAC para ambos conjuntos de transferencia resultaron superiores que sus equivalentes a4PLS, al menos en la zona analizada.

Habiendo evaluado y comparado las variantes de DR fundamentalmente en lam=1, cabe ahora analizar detalles de las curvas en intervalos de lam, y en especial en el tau 10. Los 10 valores de la estrategia de generación común fueron utilizados, y para estos casos (13 muestras en **X**, 4 en **L**) Maxilam toma el valor de 3.25. A su vez, en ocasiones fueron agregados valores de lam superiores (en notación: 50:-5:5), y aunque 3.25 ya no sería el máximo, igualmente conservará su denominación Maxilam. Similarmente y según fue o no necesario, se agregaron valores inferiores al menor común de 0.5 (en notación: 0.45:-0.05:0.05). También lam=0 fue agregada para análisis específicos. La inclusión selectiva de unos valores y no de otros sólo se realizó para mejorar la claridad con la que se exponen los resultados, ya que especialmente en las gráficas donde se exponen múltiples curvas se genera mucha confusión. Todos estos cambios en lam serán

debidamente notificados en el texto o señalados en las figuras posteriores.

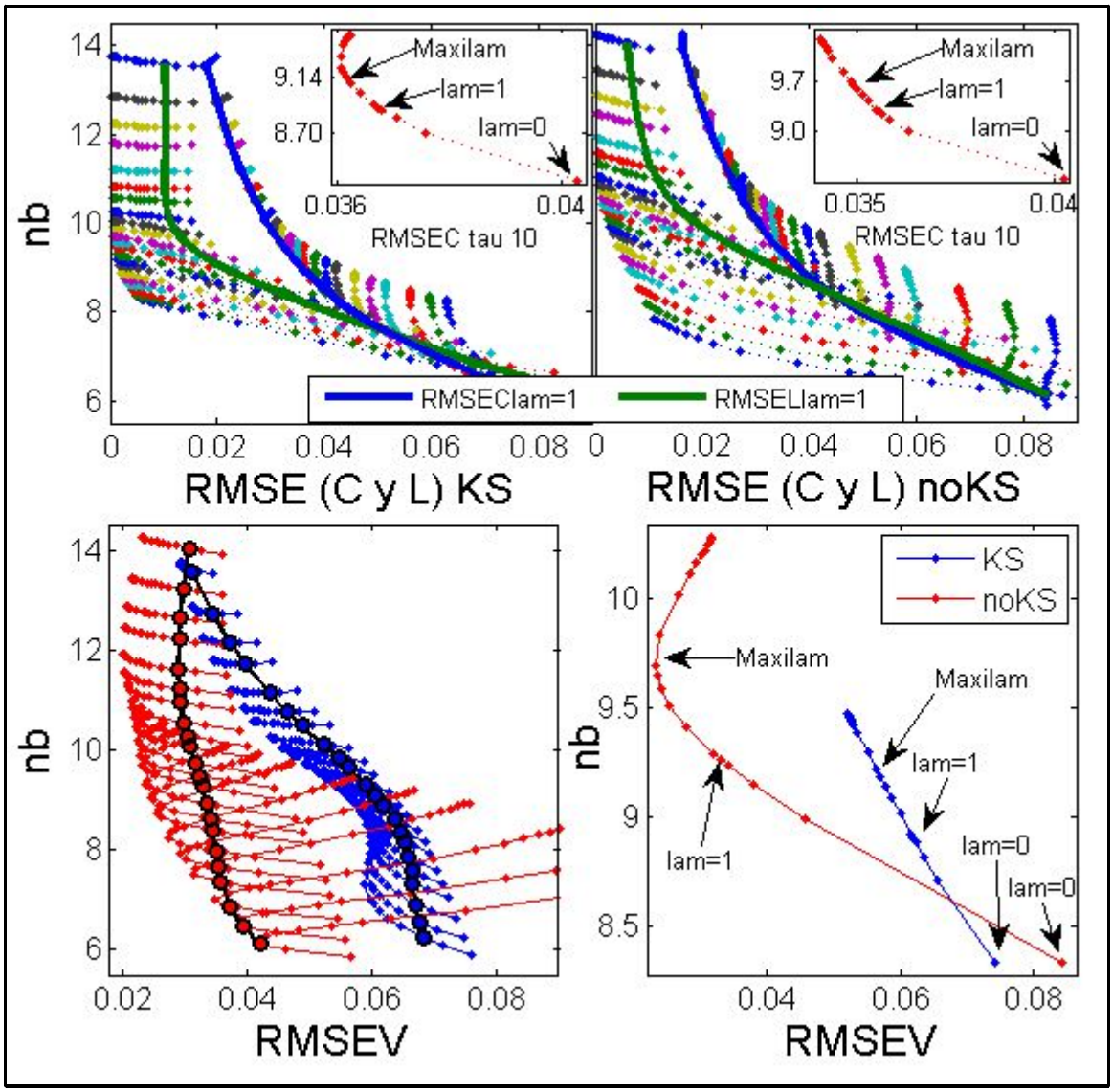


Figura 22: RMSE (C, L y V) versus norma de los vectores de regresión (nb) para SAC4 con las muestras de transferencia KS y noKS, en intervalos de lam y tau, datos "Temperatura"

Referencias: En las gráficas superiores cada color representa un tau generador de respectivas curvas de RMSEC y RMSEL del mismo color. En la gráfica de RMSEV izquierda las curvas negras corresponden a lam=1. Los valores de lam señalados son de interés para su análisis (ver texto).

Las gráficas superiores de la figura 22 fueron obtenidas con los 10 valores de lam comunes y con el agregado de valores mayores (50:-5:5), por lo que el valor mínimo presente fue de 0.5. Tanto para KS como para noKS se observa que en la zona de normas inferiores y medias existen curvas

para determinados valores de tau donde RMSEC puede disminuir (levemente) con el aumento de lam, es decir, el ajuste a **X** puede mejorar incluso cuando se esté ponderando cada vez más al de **L**. En principio esto parece positivo y deberá ser tenido en cuenta, más allá de que en la zona de normas superiores esto no se da (se dificulta verlo porque los puntos se encuentran muy comprimidos) y como se ha visto en otros casos RMSEC desmejore con el aumento de lam.

También puede apreciarse que las curvas en lam=1 para RMSEC son similares, y levemente diferentes para RMSEL. Esto indica que con ambos conjuntos de transferencia los ajustes a las muestras que dan origen a los modelos (**X** y **L**) serán similares, o dicho de otra forma que la información secundaria adicionada podrá compatibilizarse con la primaria pre-existente sin importar qué conjunto de transferencia haya sido usado, más allá de que viendo posteriormente los resultados de RMSEV quede indicado que no será lo mismo para muestras secundarias futuras.

En las ampliaciones insertas de RMSEC en tau 10 (rojo en la serie) también se incluyó lam=0. Allí se puede evaluar el RMSEC del modelo original y observar cómo éste disminuye en ambos casos con aumentos de lam en gran parte del intervalo. En otras partes de este escrito se propuso que una posibilidad para seleccionar un tau final distinto al original y sin optimizar a lam podría estar basada en recuperar el nivel de RMSEC que se tenía originalmente, para lo cual el nuevo tau en lam=1 debería tener el mismo RMSEC que el tau original en lam=0. Debido a las mejorías observadas en RMSEC con el aumento de lam, la estrategia de cambio de tau no tendría sentido, pues con las tendencias observadas dicho tau probablemente se presentaría en normas inferiores a la original, donde se supone que la capacidad predictiva de los modelos es aún menor. Suponiendo que por las razones anteriores se conservaría el tau original, una posibilidad para seleccionar un valor de lam final podría provenir de la relación entre el aumento de lam y la disminución de RMSEC, por lo cual se señalaron los puntos para Maxilam. En el caso de KS, Maxilam produce un RMSEC muy cercano al de un valor de lam superior a partir del cual ya comienza a aumentar RMSEC y en ese sentido Maxilam (o el valor que produce el mínimo RMSEC) podría indicar un tope para el ascenso de lam. En el caso de noKS Maxilam produce un RMSEC que no es especial en sí, ya que valores de lam superiores continúan con las disminución de RMSEC y entonces lo indicado sería proseguir con el aumento de lam. Posteriormente se evaluará si éstos criterios serían útiles para las muestras de Validación. Finalmente, de las ampliaciones en tau 10 también cabe notar que con KS las variaciones de RMSEC y norma son menores que con noKS (ver valores en los ejes).

En las gráficas inferiores dedicadas a RMSEV los colores azul y rojo representan a KS y noKS, respectivamente, aunque los valores de tau sean equivalentes a los de las gráficas superiores. En la

gráfica inferior izquierda (donde no graficaron los modelos para lam=0), se aprecia que noKS obtiene modelos Pareto superiores básicamente para todo tau y lam. Sólo en los modelos de normas inferiores y medias se da el caso de que al aumentar lam, con noKS se producen expulsiones del origen y esto hace que algunos modelos de noKS se superpongan con los de KS, pero en general noKS es mejor que KS, lo cual también queda indicado con las curvas para lam=1 en negro (las cuales ya habían sido expuestas en figuras anteriores). También es llamativo que las expulsiones se dan hasta zonas donde los RMSEC podían mejorar con el aumento de lam, y que no existen expulsiones (sino mejorías directas de RMSEV al aumentar lam) en la zona donde los RMSEC aumentan si aumenta lam. Otro detalle que puede apreciarse es que en general los RMSEV y las normas de KS varían menos que con noKS, lo cual también fue observado para las ampliaciones de RMSEC. Entonces se observa que en SAC4, la información que brinda KS en favor de las muestras de Validación no sólo que no resulta tan beneficiosa como la de noKS, sino que además no se deberían esperar grandes mejorías optimizando posteriormente a lam. Esto mismo puede verificarse en la gráfica inferior derecha, donde se aprecia que en términos de RMSEV todo el intervalo de lam para KS se encuentra contenido entre los modelos de noKS con lam=0 y lam=0.5. En esta misma gráfica puede verse que la selección de lam basada en la evolución de RMSEC no sería útil y que de hecho daría indicaciones de alguna manera opuestas para hallar los óptimos de RMSEV. Es decir, mientras que con KS en Maxilam el RMSEC podría indicar un tope de lam, para RMSEV se obtendrían mejorías incluso aumentando lam. Por el contrario, en noKS el óptimo para RMSEV se encuentra en Maxilam y no con valores mayores, mientras que en RMSEC las mejorías podían continuar aun con aumentos de lam. Por ende para optimizar lam deberían asignarse muestras a tal fin o utilizar criterios diferentes a los expuestos, o bien confiar en que en ambos casos con Maxilam se obtendrían errores óptimos o cercanos a éstos.

Habiendo evaluado las cifras de mérito para SAC4, se analizan resultados para DIFF-MC1c con ambos conjuntos de transferencia, los cuales pueden ser apreciados en la figura 23. En las gráficas superiores de la mencionada figura, para RMSEC y RMSEL (con 10 lam de generación común y habiendo agregado 50:-5:5) en principio conviene destacar que para los tau de normas inferiores, RMSEL en lam=1 no obtiene valores cercanos a cero como en el caso de datos “Maíz”, y en cambio la evolución de RMSEC y RMSEL con tau es similar si lam=1. Sólo con el aumento de lam se lograrán valores de RMSEL cercanos a cero, pero esto ocurrió siempre. Por lo tanto, de aquí también se aprecia que la inclusión de diferencias espectrales junto con valores de referencia de 0 se

manifestó de forma diferente en datos “Maíz” y en datos “Temperatura”, y a ésto también debe sumarse que para los últimos tampoco se observaron caídas de norma con el aumento de lam.

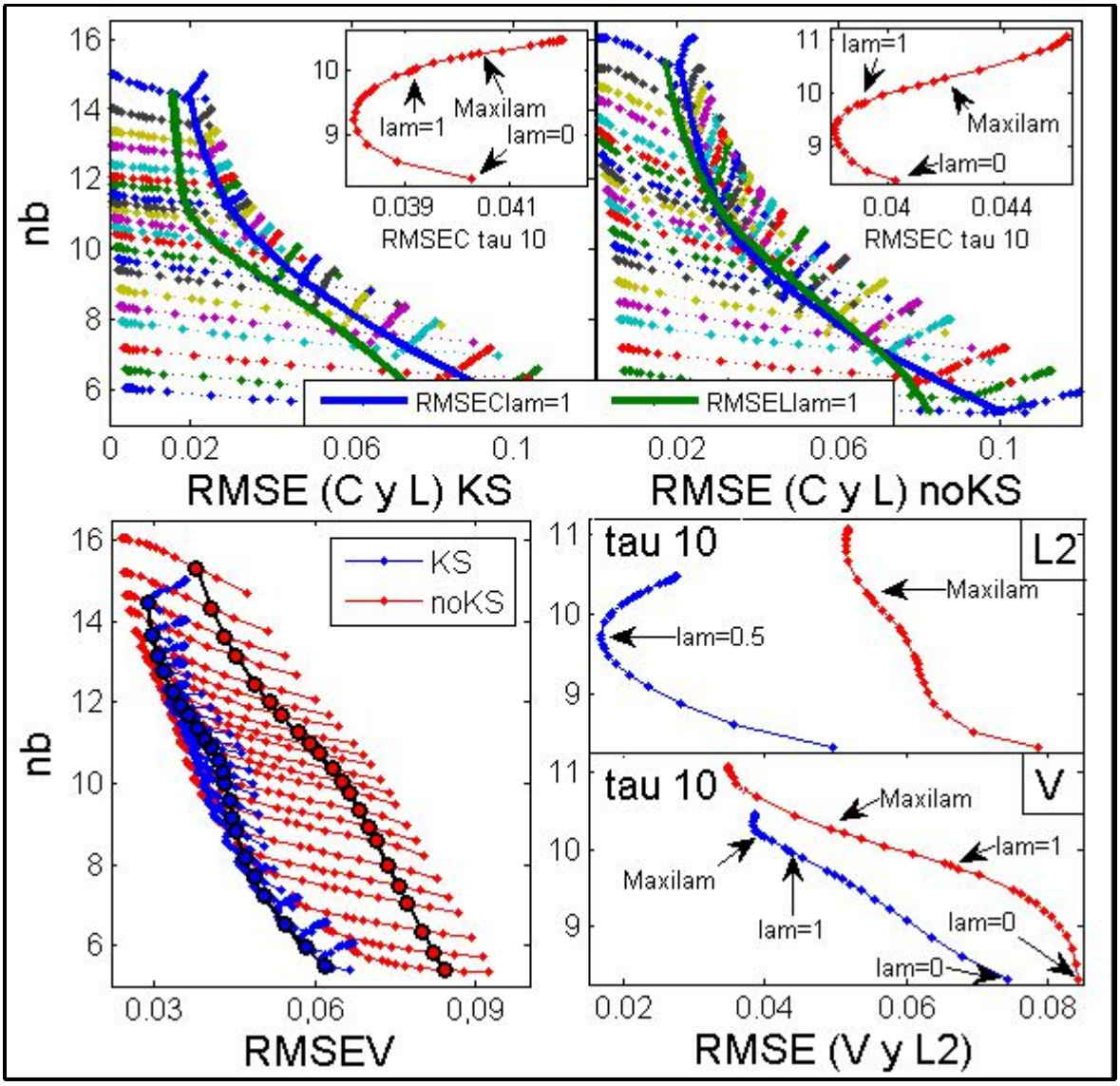


Figura 23: RMSE (C, L, L2 y V) versus norma de los vectores de regresión (nb) para DIFF4-MC1c con las muestras de transferencia KS y noKS, en intervalos de lam y tau, datos “Temperatura”

Referencias: En las gráficas superiores cada color representa un tau generador de respectivas curvas de RMSEC y RMSEL del mismo color. En la gráfica de RMSEV izquierda las curvas negras corresponden a lam=1. Los valores de lam señalados son de interés para su análisis (ver texto).

También en las gráficas superiores puede apreciarse que al crecer lam los RMSEC y las normas crecieron, pero ésto sólo es así porque el valor mínimo de lam fue de 0.5. En cambio, en las

ampliaciones insertas para RMSEC en tau 10, donde se agregaron valores menores de lam (en notación: 0.45:-0.05:0.05 y 0), puede verificarse que en realidad existe un intervalo de valores de lam para los cuales RMSEC mejora antes de comenzar a desmejorar. Como se evaluó anteriormente, el tope para el asenso de lam podría provenir del tope en la mejoría de RMSEC. Aunque posteriormente se verá que esto no será conveniente para las muestras de Validación, se podría evaluar en tau 10 qué valor de lam (o uno cercano) obtiene el mismo RMSEC que en el caso de lam=0, ya que en ambas curvas se aprecian RMSEC mayores al de lam=0 una vez que ambas curvas se encuentran en la zona de aumento permanente del RMSEC con lam (en las gráficas de SAC4 esto no se manifestó). En el caso de KS, esta igualdad se da aproximadamente en lam=2.8, que representa el punto siguiente a Maxilam. Por el lado de noKS, el valor respectivo es lam=1.9, entre Maxilam y lam=1, más cerca del último.

Antes de evaluar si los valores de lam recomendados según RMSEC serían o no aptos para las muestras de Validación, conviene analizar brevemente la gráfica inferior izquierda, obtenida con los mismos valores de lam que las superiores. En esta gráfica se aprecia que los resultados provenientes de noKS conforman el frente de modelos superiores en términos de Pareto, lo cual queda evidenciado con que los modelos son rojos para normas desde medias hasta altas, bien a la izquierda de la gráfica. Sin embargo, la diferencia de éstos óptimos con los de KS es realmente escasa. A su vez, para el último los RMSEV se encuentran muy comprimidos y son en general resultados aceptables, hasta el punto en que en lam=0.5 (menor valor de lam) predicen mejor que noKS en lam=1 e incluso mejor que en valores de lam superiores. En relación a la baja sensibilidad al cambio de RMSEV respecto de lam, esto puede ser evaluado de dos formas opuestas de alguna forma, relacionadas a la la necesidad de seleccionar un valor de lam final. Por un lado, no mostrar grandes cambios pero aún así obtener resultados aceptables como KS resulta positivo, sumado a que no sería necesario preocuparse por un método para la optimización de lam porque el resultado no variaría demasiado. Pero a su vez, no contar con la posibilidad de optimización para lam indicaría que otros conjuntos de transferencia (no KS ni noKS), quizá obtenidos sin la opción de evaluar la representatividad (espectral y/o de valores de referencia) por condicionamientos experimentales, resultarían en errores fijos (lam tomaría algún valor fijo) y no modificables, siendo que quizá cabría la posibilidad de obtener mejores resultados, como es el caso de noKS, el cual requiere más optimización en lam. Contar con una metodología de selección de lam representa ciertamente una ventaja, porque de otra manera o bien se confiaría a ciegas en las muestras de transferencia y en un valor de lam fijo obtenido con alguna otra metodología, o bien se tendrían que invertir esfuerzos

para obtener muestras extra destinadas sólo a ese fin, algo que en el contexto ahorrativo de transferencia querría evitarse.

Ya observando la gráfica inferior derecha de RMSEV (en la cual se agregaron los valores de lam 0.45:-0.05:0, lo cual también se hizo en la de **L2**) en tau 10, para el caso en cuestión se pueden comparar dos metodologías de selección de lam, la basada en los resultados de RMSEC y la basada en los de **L2** (o RMSEL2). Para KS, la primera indicaría un valor cercano a Maxilam, lo cual resultaría muy cercano al óptimo de RMSEV, mientras que para noKS resultaría en un RMSEV aproximado de 0.06 que aun podría ser optimizado hasta aproximadamente 0.035 (óptimo de RMSEV para noKS). En cuanto a las curvas **L2**, para KS la indicación sería no aumentar más allá de lam=0.5 con lo que el RMSEV no sería óptimo. En cambio para noKS, se aprecia que incluso valores mayores a Maxilam obtienen errores menores y si esto se observa realizando un zoom sobre la curva (no mostrado) el mínimo se presenta en lam=20, dando un valor de 0.036 casi óptimo para RMSEV. Por lo tanto, ninguna de las dos metodologías resulta apropiada para determinar resultados pseudo-óptimos con ambos conjuntos de transferencia.

Vale también destacar que las gráficas de RMSEV y RMSEL2 están en las mismas escalas. Por lo tanto, de modo muy simplista, puede pensarse que la mayoría de los puntos de **L2** para KS se encuentran en valores de error bajos que en general tenderán a crecer con lam, y lo inverso ocurre para noKS (errores altos decrecientes con lam). En una zona intermedia entre los anteriores, se puede encontrar a la mayoría de los puntos de ambas curvas de RMSEV y en especial a los óptimos. Lo que se quiere hacer notar es que las indicaciones provenientes de las curvas de RMSEL2 no deberían interpretarse igualmente para ambas si el objetivo de dichas indicaciones es obtener valores de lam óptimos para futuras. Indicar que el error descenderá al aumentar lam pero hacerlo desde errores altos, como con noKS, será más representativo de la evolución que tendrían muestras futuras (como las de Validación) que tampoco hubiesen participado de la elaboración de modelos. Por el contrario, indicar que lam no debería ascender demasiado pero hacerlo desde errores más bajos que los que en promedio se obtendrían para muestras futuras, como con KS, no resulta representativo. En este último caso además conviene realizar otra reflexión. Si se recuerda que en DIFF (en general) las diferencias espectrales y los valores de referencia de 0 aportan el mecanismo a través del cual se pretende mejorar las predicciones de Etanol en muestras secundarias que no aportarán información durante los cómputos, las cuales entre otras incluyen a las de **L2**, es esperable que al comprobar el ajuste obtenido para dichas muestras sus valores de referencia para Etanol tomen un rol importante. Como ya se ha dicho, en KS los espectros y sus valores de

referencia específicamente para Etanol de alguna forma no son muy compatibles, por lo tanto puede esperarse que un procedimiento de transferencia dedicado a refinar las predicciones de Etanol tampoco resulte compatible con las muestras de KS. Con lo anterior, observar que ya desde un valor pequeño y superior a lam=0.5 se obtendrán desmejoras de RMSEL2 es entendible, siendo que se supone que el ascenso de lam mayoritariamente aumentaría la capacidad de predecir Etanol a partir de espectros secundarios que resultaran compatibles con sus valores de referencia.

Finalmente, en la figura 24 se presentan resultados útiles para comparar SAC4 con DIFF4-MC1c en términos de RMSEV para ambos conjuntos de transferencia a través de curvas ya expuestas, aunque no simultáneamente en la misma gráfica.

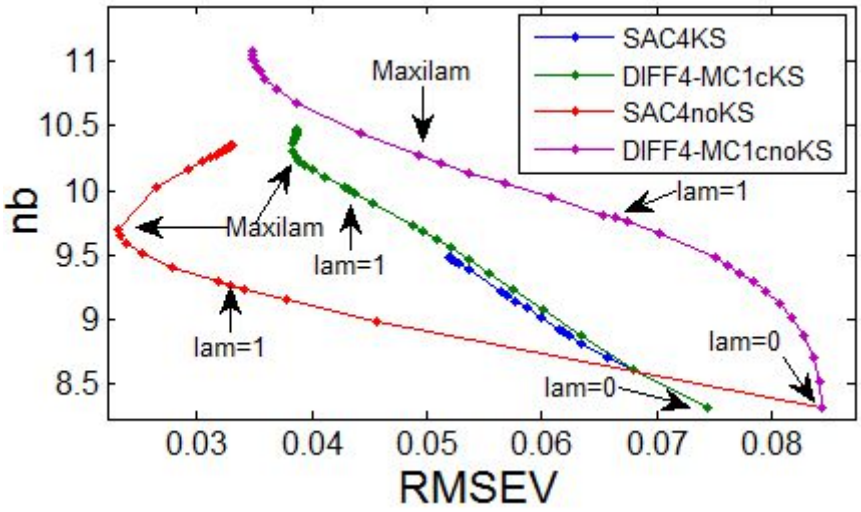


Figura 24: RMSEV versus norma de los vectores de regresión (nb) en tau 10, para SAC4 y DIFF4-MC1c, con los conjuntos de transferencia KS y noKS, datos "Temperatura"

Referencias: Los valores de lam señalados son de interés para su análisis (ver texto)

Antes de comparar definitivamente a las distintas variantes, cabe analizar la figura 24 para hacer algunas reflexiones adicionales. En principio, deben observarse los modelos en lam=0 y recordar que en éstos no existe un procedimiento de transferencia en sí, ya que el espacio multivariado calibrado sólo habrá dependido de las muestras primarias y por lo tanto no habrá existido transferencia de esa información a ningún otro espacio. Por ende, ya que se aprecia que en tau 10 los modelos KS resultan en RMSEV menores que los de noKS, esta diferencia puede atribuirse al uso de la información secundaria sólo para realizar centrados locales. Con lo anterior en mente, se puede pensar que la relación entre espectros y valores de referencia de KS no es tan conflictiva o

incompatible como se ha propuesto durante el desarrollo si dicha relación es simple como la implícita en un centrado. Es decir, restar la media de ciertos espectros conocidos a espectros desconocidos, predecir a los últimos y sumarles promedios de valores de referencia provenientes de los primeros, asumiendo con esto una relación simplista a través de la cual se supone que los valores de referencia de alguna forma serían linealmente proporcionales a los espectros y viceversa. Sin embargo, una vez que lam comienza a aumentar, el efecto no será solamente de centrado, los modelos provendrán de transferencias reales, y la relación entre espectros y valores de referencia tendrá influencia en los resultados de la minimización (16). Esta última relación, de mayor complejidad multivariada comparada a la de un centrado, y que a su vez será influenciada mediante un equilibrio con la norma vectorial y con las predicciones primarias, es la que se supone conflictiva para KS. Ante lo anterior, al crecer lam la variante de DR que obtiene los modelos sólo a partir de información espectral (DIFF-MC1c) resulta mejor que la que además utiliza los valores de referencia (SAC). Más aún, en el último caso y aún habiendo utilizado valores de lam tan altos como 50 (que ya de por sí resultan carentes de mucho sentido como ponderadores), el modelo óptimo obtendrá un RMSEV que estará lejos de ser el óptimo de DIFF-MC1c. Por lo tanto y como ya se ha expuesto, aunque para las muestras de Validación los espectros de KS puedan considerarse representativos, cuando éstos sean asociados a sus valores de referencia y en especial cuando esta asociación sea ponderada, dicha representatividad no podrá repercutir de manera tan ventajosa para el procedimiento. Vale también destacar que en términos de norma el ascenso de ésta a medida que disminuye el RMSEV es similar para ambos casos con KS, es decir, si las curvas se aproximan a rectas, éstas tienen pendientes muy similares. Por lo tanto, la diferencia fundamental estará en los RMSEV óptimos que cada variante pueda lograr, y ante esto se verifica que KS funciona mejor con DIFF-MC1c que con SAC.

Por el lado de noKS, se aprecia que el mejor de todos los resultados será obtenido con SAC4. Este RMSEV será de 0.02311 aproximadamente, y a su vez será menor que el RMSEV planteado como objetivo (0.02406), por lo cual esta transferencia habrá cumplido su cometido. Cuando en DIFF4-MC1c se utilicen sólo los espectros y no su ventajosa relación con sus valores de referencia, se obtendrán RMSEV bastante mayores que podrán ser mejorados sólo a través de la optimización de lam, pero que no lograrán los óptimos de SAC4.

Por todo lo expuesto, se concluye que en el caso de datos “Temperatura” la necesidad de transferencia es incluso mayor que para datos “Maíz”, pues no alcanzaría simplemente con realizar

centrados locales para alcanzar resultados óptimos. Es lógico que datos “Temperatura” deje ver estos resultados, pues el fenómeno involucrado no es una simple deriva entre instrumentos sino un cambio de temperatura que afecta de forma selectiva a los espectros de cada uno de los 3 componentes de las mezclas analizadas, más allá de que sólo se haya modelado Etanol.

El hecho de que una variante SAC haya otorgado los mejores resultados para las muestras de Validación concuerda con datos “Maíz”. Sin embargo, el análisis simultáneo con los conjuntos noKS y KS mostró que en casos como el último SAC no será mejor que DIFF-MC1c. Los resultados de varias experiencias sugieren que la supremacía de una variante por sobre la otra dependerá de características relacionadas a la representatividad espectral y de valores de referencia, pero por sobre todo de la relación entre ambas. Esto deberá ser tenido en cuenta a la hora de seleccionar muestras de transferencia, siempre que dicha selección sea posible.

1.7 Conclusiones

- Los resultados obtenidos sugieren que la derivación de la TR denominada DR en sus variantes SAC y DIFF, vistas como métodos de transferencia de modelos CMV de orden 1, proveen la flexibilidad necesaria para desensibilizar a un modelo primario armónico en términos de exactitud y varianza, con el objetivo de obtener mejores predicciones en nuevas condiciones experimentales, una vez que se ha determinado que nuevas muestras no serían compatibles con la situación original modelada.
- El problema de transferencia de calibración pudo ser encuadrado dentro del marco de expresiones y ecuaciones que encuentran su base en la TR. A través de planteos extra se derivó la DR, con la cual se solucionaron problemas relacionados a la inversión de matrices. A partir de la DR se apreciaron diversos efectos relacionados al cambio en los meta-parámetros y con esto los efectos relativos a la ponderación parcial de la información.
- Se pudo evaluar el efecto de los centrados, siendo los del tipo "local" como MC3 y MC1c los que obtuvieron los mejores resultados en general. Los análisis que involucraron cantidades variables de muestras de transferencia permiten decir que el procedimiento pudo realizarse con relativo éxito utilizando 3 ó 4 muestras. También, especialmente en el caso de datos "Temperatura", se pudieron evaluar efectos de la representatividad o de la calidad estructural (espectros, valores de referencia y relaciones) de las muestras de transferencia en relación a muestras futuras, lo cual sugiere que esos efectos deben ser tenidos en cuenta si es posible seleccionar las muestras de transferencia, en contraposición a una selección al azar o impuesta por condiciones experimentales. Todo lo anterior representa efectos que son reconocidos por condicionar usualmente a otros algoritmos de transferencia o a los de estandarización espectral como PDS, también íntimamente relacionados a resolver problemas cuando los modelos dejan de tener validez.
- Se observó que la información primaria reutilizada es conveniente a pesar de que por sí sola ya no predeciría bien, y similarmente se observó que el sólo uso de la información secundaria disponible no es mejor que su combinación con información primaria.
- Ni SAC ni DIFF resultaron mejores en todos los casos, aunque en experiencias de contraste con otros algoritmos (PLS y PDS) se hayan mostrado superiores en general. La principal ventaja de DIFF-MC1 radica en que no son necesarios los valores de referencia, y por ende

se ahorran determinaciones, aunque deberá tenerse información espectral primaria y secundaria equivalente al mismo tiempo. Si se utiliza DIFF-MC1c, entonces el grado de ahorro será menor, pero en general se vio que se obtendrían mejores resultados. Por el lado de SAC, lo más ventajoso a nivel experimental radica en que no serán necesarias muestras primarias equivalentes a las de transferencia secundarias. A nivel de determinaciones necesarias, SAC insumiría las mismas que requiere la aplicación del centrado MC1c.

- En ocasiones, las mejorías obtenidas serán debidas casi totalmente a la estrategia de centrado utilizada, como en el análisis de instrumental con deriva. En casos donde los cambios espectrales sean más complejos por deberse a modificaciones químicas o físicas, como en el caso de datos "Temperatura",más necesario será modificar las relaciones modeladas entre valores de referencia y variables espectrales. En estos casos el proceso de transferencia en sí representará disminuciones de error en las predicciones, más allá de que el centrado también será relevante.
- A pesar de que se realizaron experiencias y análisis con mucho detalle, y aunque se encontraron tendencias en relación al cambio en los meta-parámetros, los resultados analizados no otorgan la suficiente información como para elaborar un algoritmo que lleve a cabo automáticamente la tarea de encontrar resultados óptimos para las transferencias. No obstante, sí puede decirse que al menos en la zona de modelos armónicos podrían obtenerse resultados aceptables sin necesidad de cambiar el tau original. En cuanto a lam, deberán realizarse algunas reflexiones previas acerca de la información con la que se cuenta para determinar si es o no conveniente aumentar su valor. En los casos en los que lo fue, se apreció en general que lam podría ascender hasta el valor que se denominó Maxilam, es decir, hasta la relación entre el número de muestras primarias de Calibración y el de secundarias de transferencia. Con este valor los modelos obtuvieron buenos resultados, más allá de que en algunos casos los óptimos se encontraron en valores mayores o menores, pero cercanos.

CAPÍTULO 2: Estudio metabonómico para la detección de efectos de *stress* en frutos de tomate luego de tratamiento con Carbofurano, a partir de datos de Cromatografía Líquida-Espectrometría de Masa (LC-MS). Utilización de técnicas quimiométricas para resolución y clasificación de muestras.

2.1 Resumen

Se aplicó una estrategia quimiométrica basada en Resolución Multivariada de Curvas mediante Mínimos Cuadrados Alternantes (MCR-ALS) a datos de Cromatografía Líquida acoplada a Espectrometría de Masa (LC-MS). Dichos datos provinieron de frutos de tomate, específicamente del cultivar Rambo, luego de que algunas muestras fueran tratadas con el pesticida Carbofurano. El objetivo de estas experiencias fue obtener perfiles de concentración para una gran cantidad de potenciales metabolitos y a través de estos perfiles poder detectar la presencia de efectos de stress debidos al tratamiento con el mencionado pesticida. El análisis de los resultados provenientes de muestras colectadas en distintos días tras la aplicación de Carbofurano sugiere que algunos componentes tienen cinéticas específicas y claramente diferenciables entre muestras tratadas y no tratadas, más allá de que otras resultaran similares. Por lo tanto, esta metodología se mostró adecuada para verificar que la presencia de pesticida es causal de cambios en el tiempo sobre el comportamiento de algunos metabolitos endógenos del tomate, como resultado de un stress fisiológico.

De manera similar, la estrategia se utilizó para resolver muestras tratadas y no tratadas, pero esta vez en conjunto con muestras de una variedad de tomate llamada RAF, además de la ya utilizada Rambo. Ya que tanto las muestras tratadas como las que actuaron de blancos para ambas variedades fueron recolectadas con el mismo protocolo de muestreo e hipotéticamente bajo condiciones similares de maduración, algunas de las diferencias de comportamiento entre perfiles de concentración podrían ser interpretadas como parte de efectos debidos a las aplicación de pesticida. Basados en la hipótesis anterior, se realizaron algunos modelos de clasificación con Mínimos Cuadrados Parciales-Análisis Discriminante (PLS-DA) con el objetivo de comprobar si sería posible clasificar a las muestras usando los datos metabonómicos provenientes de los perfiles de concentración tras la aplicación de MCR-ALS. Los resultados mostraron que los modelos PLS-DA podrían ser útiles para detectar los efectos de stress debidos al tratamiento, el tipo de cultivar y, en menor medida, para distinguir simultáneamente a los 4 tipos de muestras estudiadas.

2.2 Introducción

En los últimos años, las nuevas disciplinas "ómicas" (metabolómica y metabonómica) han sido aplicadas de manera creciente en diversos campos, como ser genómica funcional, toxicología, farmacología, diagnóstico de enfermedades, ciencias de la alimentación y de la nutrición, ciencias ambientales, entre otras, a la vez que el número de reportes de estos estudios en revistas revisadas por pares ha ido creciendo año tras año. La metabonómica ha sido definida como "la medición cuantitativa de respuestas multiparamétricas y dependientes del tiempo en sistemas multicelulares ante estímulos patofisiológicos o modificaciones genéticas". Por su parte, la metabolómica encuentra su definición en "la medición de concentraciones y flujos de metabolitos en sistemas celulares aislados (y usualmente idénticos) o en complejos celulares" (Plumb y col., 2002). Más allá de estas dos definiciones, algunos autores usan estos términos de manera indistinta (Viant, 2008) y algunas estrategias (perfiles metabólicos, huellas metabólicas, análisis de metabolitos objetivo) han sido propuestas como análisis metabolómicos (Dunn y Ellis, 2005; Schauer y Fernie, 2006). Obviamente cualquiera de las opciones anteriormente mencionadas requiere medir un conjunto de compuestos de gran variedad en cuanto a sus propiedades fisicoquímicas (peso molecular, polaridad, solubilidad, volatilidad, etc.) y en un amplio rango de concentraciones, desde picomolares hasta milimolares. Estos compuestos son los denominados metabolitos, representados por ácidos orgánicos, azúcares, aminoácidos, vitaminas, péptidos pequeños, entre otros, y están involucrados en procesos metabólicos, siendo productos de éstos, o bien partes funcionales e intermedias en muchos procesos biológicos.

Como en el caso específico del presente trabajo, la metabonómica depende de la posibilidad en la determinación de cambios en los metabolitos. Entre las estrategias analíticas utilizadas en metabonómica, pueden nombrarse la espectroscopia de Resonancia Magnética Nuclear de protón de campo alto (1H NMR), la inyección directa en espectrómetro de Masa, la espectroscopia Infrarroja con Transformada de Fourier (FT-IR), y técnicas separativas tales como Cromatografía Líquida o Gaseosa y Electroforesis capilar con detección mediante espectrometría de Masa (LC-MS, GC-MS y CE-MS) (Lenz y Wilson, 2007). Actualmente, desarrollos emergentes en las tecnologías analíticas, tales como los sistemas separativos rápidos de alta resolución (por ejemplo Cromatografía Líquida de Ultra Rendimiento, UPLC) o como los instrumentos con amplios rangos dinámicos y alta exactitud de Masas, entre los que se pueden nombrar "Tiempo De Vuelo-Masa"

(TOF-MS), “Cuadrupolo-Tiempo De Vuelo-Masa” (Q-TOF-MS), “Resonancia de Ciclotrón con Transformada de Fourier-Masa” (FT-ICR-MS), espectrometría de Masa con Impacto de Electrones (EI-MS, con o sin pirólisis) y “Orbitrap con Transformada de Fourier-Masa” (FT-Orbitrap-MS), pueden proveer más información a partir de los datos experimentales generados, dando lugar a una mejor asignación de los metabolitos (Moco y col., 2007; Ducruix y col., 2008). Ante lo expuesto, es evidente que no existe un método universal para detectar todo tipo de componente.

Lograr una separación cromatográfica completa de todos los componentes de una muestra biológica compleja es a menudo algo muy difícil de alcanzar. Esto fundamentalmente es debido a la existencia de picos solapados, incluso en condiciones óptimas para la separación, siempre que se utilicen sistemas separativos convencionales. Entre las cromatografías, la líquida es probablemente la más versátil, ya que permite la separación de componentes en un amplio rango de polaridades y de estabilidades térmicas. Actualmente, el uso conjunto de técnicas cromatográficas y espectroscópicas con detección de Masa permite obtener datos de segundo orden que combinan señales instrumentales provenientes de los dominios de tiempo y espectrales, a través de los cuales puede ser determinada una amplia variedad de compuestos con concentraciones menores a las que pueden ser detectadas, por ejemplo, mediante NMR. Estas respuestas pueden manipularse como datos matriciales, donde cada columna corresponde a una relación “masa/carga” (m/z) y donde cada fila representa un tiempo específico de la separación (o viceversa si se desea). El uso de este tipo de datos junto con métodos multivariados de resolución permite obtener perfiles de Masa y concentración (espectros de Masa y cromatogramas, respectivamente) para distintos componentes de una muestra.

Aun con lo anterior, debe entenderse que los estudios metabonómicos destinados a analizar datos provenientes de técnicas mixtas como GC-MS o LC-MS tienen algunos inconvenientes, como por ejemplo que los datos presentan alto grado de ruido de fondo, o que las matrices obtenidas contienen información que, para ser procesada, requiere recursos de cálculo considerables, aunque estos inconvenientes no son exclusivos de los tipos de señales nombradas. Este tipo de problemas y otros pueden ser evitados mediante el uso de técnicas quimiométricas para eliminación de ruido y compresión de datos que, utilizados apropiadamente, no conducirán a pérdida de información relevante. Existen también otros inconvenientes más relacionados a la cromatografía en sí, como cuando existe redundancia en el sentido de que las señales para algunas relaciones de m/z pueden provenir de analitos solapados. Más aun, el ruido de fondo puede variar con los tiempos de retención, a la vez que la forma de algunos picos y los tiempos de retención pueden variar en

análisis sucesivos para muestras muy complejas. Sobre todas las cosas, la resolución cromatográfica completa no puede ser posible para todos los analitos si éstos existen en gran número y la dimensión de separación es una sola, lo cual conducirá indefectiblemente a solapamiento de señales.

Más allá de lo detallado, con este tipo de datos pueden obtenerse modelos capaces de utilizar la llamada ventaja de segundo orden, por la cual es posible calibrar con patrones puros o mezclas de éstos sin tener en cuenta a los potenciales interferentes, conduciendo a la solución de problemas como el corrimiento de picos y el solapamiento de señales, entre otros. Los modelos con resolución de curvas se aplican principalmente para alcanzar estos objetivos (Halket y col., 1999; Shen y col., 2001; Jonsson y col., 2004; Jonsson y col., 2005; Jonsson y col., 2006).

Los métodos quimiométricos incluyen métodos iterativos, así como también otros basados en la selección de variables más puras. Dentro de las opciones quimiométricas no iterativas, basadas en la evolución natural de los datos, se pueden nombrar los análisis de factores EFA (del inglés *Evolving Factor Analysis*) (Maeder, 1987; Maeder y Zilian, 1988), WFA (del inglés *Window Factor Analysis)* (Malinowski, 1992; Den y Malinowski, 1993) y SFA (del inglés *Subwindow Factor Analysis*) (Manne y col., 1999; Shen y col., 1999), así como también los métodos con proyecciones tales como HELP (del inglés *Heuristic Evolving Latent Projections*) (Kvalheim y Liang, 1992*;* Liang y col., 1992), OPR (del inglés *Orthogonal Proyection Resolution*) (Liang y Kvalheim, 1994) y EWOP (del inglés *Evolving Window Orthogonal Proyections*) (Xu y col., 1999). Por otra parte, dentro de las opciones iterativas se puede encontrar a ITTFA (del inglés *Iterative Target Transformation Factor Analysis*) (Vandeginste y col., 1985) y MCR-ALS (del inglés *Multivariate Curve Resolution-Alternating Least Squares*) (Navea y col., 2001), entre otros. Por su lado, los métodos de selección de variables puras son los más simples para usar e incluyen a SIMPLISMA (del inglés *Simple to use Interactive Self-modeling Mixture Analysis*) (Windig y Guilment, 1991), OPA (del inglés *Orthogonal Proyection Approach*) (Cuesta Sánchez y col., 1994; Sánchez y col., 1996), IKSFA (del inglés *Iterative Key Set Factor Analysis*) (Malinowski, 1982) y SBM (del inglés *Simplified Borgen Method*) (Grande y Manne, 2000).

Algunos de los métodos mencionados fueron utilizados con datos provenientes de vegetales, como en nuestro caso. El método HELP fue utilizado sobre datos proveniente de GC-MS para determinar compuestos químicos de aceites esenciales en *Cortex cinnamoni* desde cuatro áreas diferentes de producción, lo cual resultó en la separación de 88-93 componentes y en la determinación de 58-64 de ellos, lo que representa aproximadamente el 90% del contenido total

(Gong y col., 2001b). El mismo algoritmo fue aplicado en combinación con datos de GC-MS con el objeto de analizar componentes volátiles en preparaciones tradicionales de la medicina en China. En esta experiencia, 93 componentes fueron separados y 65 de éstos fueron analizados cualitativa y cuantitativamente, representando aproximadamente el 90% del contenido total (Gong y col., 2001a). Más aun, en una caracterización de componentes en aceites esenciales de geranios iraníes, un total de 61 componentes fueron identificados usando búsqueda de similitudes entre los espectros de Masa obtenidos y una base de datos MS. Posteriormente, tras aplicar algoritmos para determinar el número de componentes, las variables más puras y las regiones más selectivas, se utilizó HELP y el número de componentes ascendió a 85, ya que pudieron resolverse algunos grupos de picos solapados (Jalali-Heravi y col., 2006). El análisis de datos de GC-MS mediante OPR y DS-MCR-ALS (*Distance Selection* MCR-ALS) también fue utilizado para analizar componentes de otros aceites esenciales de Irán, en ese caso de comino y alcaravea. Sin los métodos quimiométricos, se habían identificado 19 y 39 compuestos por búsqueda directa de similitudes, mientras que luego de su aplicación en zonas con picos solapados, los números ascendieron a 49 y 98, respectivamente (Jalali-Heravi y col., 2007).

La utilización de pesticidas, más allá del riesgo en su ingesta, implica inconvenientes en el campo medioambiental. En este mismo campo, MCR-ALS se utilizó con datos de LC-ESI-MS (LC-*Electro Spray Ionization*-MS) para investigar las fuentes principales de disruptores endócrinos en aguas costeras y de puertos (Peré-Trepat y col., 2004) y luego para analizar aguas residuales y sedimentos fusionando datos de LC-DAD (LC-*Diode Array Detection*) y de LC-ESI-MS (Peré-Trepat y Tauler, 2006). En la última referencia, cabe destacar que utilizaron la Transformada Wavelet (WT) para reducir las dimensiones de los datos de MS previo a realizar cálculos junto a los datos de LC-DAD, y que al haberlo hecho no sólo redujeron el tiempo de cómputo que conlleva el análisis, sino también que la resolución y cuantificación de los componentes que co-eluían fue más sencilla. Por lo anterior, y debido a que en el presente trabajo los datos también tuvieron origen a partir de análisis de muestras mediante LC-ESI-MS, la estrategia utilizada en (Peré-Trepat y Tauler, 2006) fue la base del tratamiento de los datos con WT.

El tomate es un fruto con orígenes en sitios elevados de Perú y Ecuador. Pertenece a la familia de las solanáceas (*Solanaceae*) y representa a unos de los vegetales más distribuidos alrededor del mundo, con un alto valor económico. Tradicionalmente ha sido obtenido en campos, pero su cultivo en condiciones de protección ha permitido expandir su ciclo y disponibilidad a lo largo de todo el año. Los tomates utilizados pertenecen a dos cultivares llamados Rambo y RAF. Los primeros

tienen frutos de buen tamaño (tamaño G-GG), firmes y esféricos. La piel es relativamente delgada y tiene un atractivo color rojo con rayas verdes. Ya que posee buen sabor tanto en tempranos como en avanzados estados de maduración, puede ser incorporado con facilidad en la dieta diaria. A su vez, es muy resistente al virus del mosaico de tomate, a *Fusarium* 1 y 2, a *Verticillium* y a *Stemphylium radicis*, aunque tiene una moderada resistencia a nemátodos (Vademécum de Variedades Hortícolas. Portagranos 2005-2006, 2005). Por otro lado, la resistencia a *Fusarium* es la que da el nombre a los tomates RAF (Resistente Al *Fusarium*). Este cultivar ha sido obtenido a partir de la cruza, mediante selección artificial, de variedades tradicionales de la zona de Almería, España. Además de su sabor dulce, color (verdes con manchas quasi negras) y textura, se destaca por una alta resistencia a aguas muy salinas, llegando a soportar concentraciones salinas hasta 10 veces mayores que la habitual. Otra característica distintiva de esta variedad es su morfología, sumamente irregular y diferenciable a simple vista de los frutos del cultivar Rambo. Los tomates RAF son muy codiciados y su precio suele ser elevado si se lo compara con otros, razón por la cual existe interés para determinar adulteraciones de distintos tipos en productos derivados (Vademécum de Variedades Hortícolas. Portagranos 2005-2006, 2005).

El Carbofurano es un insecticida sistémico y representativo de los carbamatos, con aplicación contra nemátodos, insectos y ácaros, el cual actúa por contacto a nivel superficial y a través de la ingestión, interfiriendo con la transmisión de impulsos nerviosos, ya que inhibe a la colinesterasa y provoca una acumulación de acetilcolina (Vademécum de Variedades Hortícolas. Portagranos 2005-2006, 2005). Es altamente tóxico para humanos y otros animales a través de su ingestión oral o nasal, siendo su LD_{50} oral en ratas de 8 mg/Kg (Abad y col., 1997). La figura 1 expone la estructura molecular del compuesto.

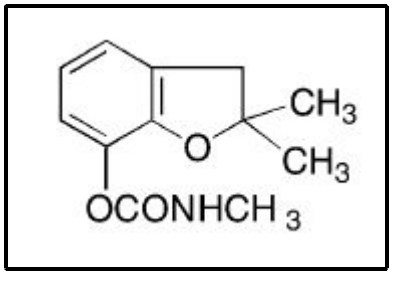


Figura 1: Estructura molecular del Carbofurano (2,2-dimetil, 2-3-dihidro-7-benzofuranil-N-metilcarbamato)

En este trabajo, mediante el uso de la WT, MCR-ALS y PLS-DA se estudiaron los datos obtenidos de muestras de tomates (*Lycopersicon esculentum*) Rambo y RAF, y se pretendió detectar

cambios en la concentración y/o evolución de sus metabolitos como resultado de un *stress* luego de la aplicación del mencionado insecticida.

2.3 Objetivos

- Diseñar y poner a prueba una estrategia quimiométrica para resolver y poder interpretar datos de LC-MS a través de los cuales pueda realizarse un estudio metabonómico para la detección de efectos de *stress* en frutos de tomate debidos a tratamiento con Carbofurano.
- Establecer metodologías de clasificación para las muestras analizadas.

2.4 Teoría

2.4.1 Pretratamiento de datos: uso de la Transformada Wavelet (WT) para la eliminación de ruido y compresión de matrices de datos.

Desde 1996, el número de artículos relacionados con la WT y tratamientos derivados para comprimir datos y eliminar ruido de señales ha crecido considerablemente. Algunos de estos trabajos están relacionados con datos de LC-DAD (Collantes y col., 1997; Shao y col., 1997; Chau y col., 2004), otros han usado detectores de Masa (LC-MS y GC-MS) (Peré-Trepat y col., 2007; Cocchi y col., 2004). A su vez, la WT ha sido aplicada como método para pretratamiento de señales previo al uso de técnicas quimiométricas como las de resolución de curvas (Chau y col., 2004) y SIMPLISMA (Chen y Harrington, 2003), así como también para preprocesar espectros IR (Bos y Vrielink, 1994), entre muchas opciones.

Las denominadas Wavelets son un conjunto de funciones madre, algunas de las cuales (Haar, Daubechies, Symmlets, Coiflets, entre otras) poseen dominio compacto, es decir, toman valores diferentes de cero sólo en un subdominio limitado, a la vez que su valor medio es cero. Los miembros de este conjunto son las bases que, a través de modificaciones, dan origen a familias de Wavelets. El hecho de tener dominios reducidos, así como otras propiedades tales como la asimetría de algunas de éstas familias, las hacen especialmente apropiadas para representar características de ciertas señales, como ser cambios bruscos, ruido y discontinuidades (Perrin y col., 2001; Walczak, 2000).

Pueden encontrarse algunas analogías entre la WT y la transformada de Fourier, ya que ambas miden similitud entre una señal y funciones de análisis a través del cómputo del producto interno,

dando como resultado representaciones de las señales originales a través de partes constituyentes, facilitando a su vez la interpretación. Mientras que en la WT se utilizan versiones escaladas (expansiones y compresiones) y trasladadas de funciones madre, en la transformada de Fourier se utilizan exponenciales complejas de las cuales pueden derivarse funciones del tipo seno con distintas frecuencias y sin restricciones de dominio. A su vez, las últimas tienen formas suaves y predecibles, mientras que las funciones Wavelet madre suelen ser irregulares y asimétricas. Debido a esto último, estas funciones suelen ser más apropiadas para analizar señales con apariciones esporádicas de cambios bruscos en su forma.

Dado que la WT extrae información a distintas escalas, podrán ponerse de manifiesto características de las señales analizadas que, para ser notadas, dependen de las escalas utilizadas. Dicho de otra forma, algunos fenómenos sólo serán perceptibles a través de largos períodos de tiempo o espacio, y otros lo serán a escalas mucho menores. Obviamente, también podrán ponerse de manifiesto fenómenos con invariancia de escala. La figura 2, adaptada desde la referencia (MATLAB *7.6.0*, 2008), ejemplifica lo anterior:

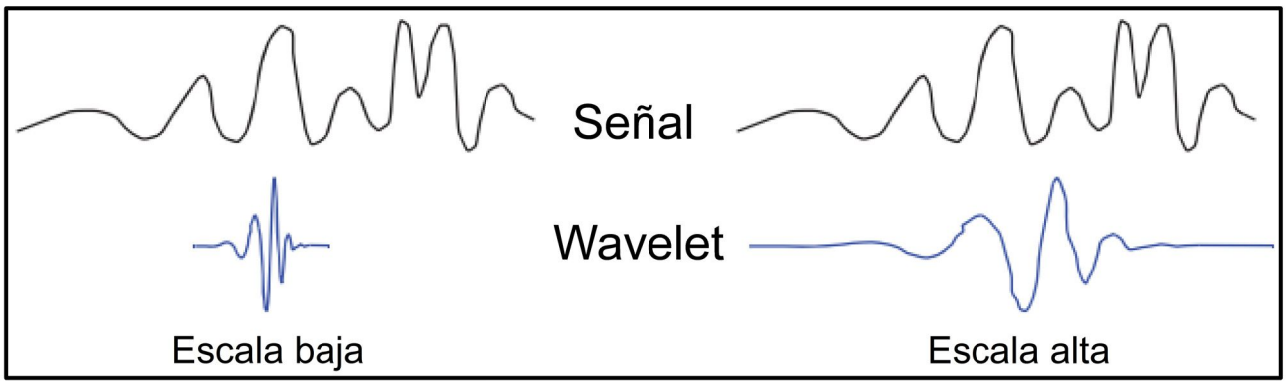


Figura 2: Semejanza entre una señal genérica y Wavelets en distintas escalas

En la figura 2 puede apreciarse que la wavelet en escala alta se asemeja más a la señal que su versión en escala baja. De esta manera, en la escala alta se obtendrá un coeficiente más significativo que en la escala baja. No obstante, la señal ejemplificada no posee variaciones de alta frecuencia, en cuyo caso también se obtendría un coeficiente significativo a baja escala. En general, los coeficientes obtenidos de las escalas altas se corresponderán con Wavelets expandidas y brindarán información más global y de menores cambios en la señal, mientras que los obtenidos de las escalas bajas estarán relacionados a Wavelets comprimidas y darán información sobre cambios de alta frecuencia.

Los efectos conjuntos de traslación y escalado en la WT pueden observarse en la figura 3, adaptada desde la referencia (MATLAB *7.6.0*, 2008):

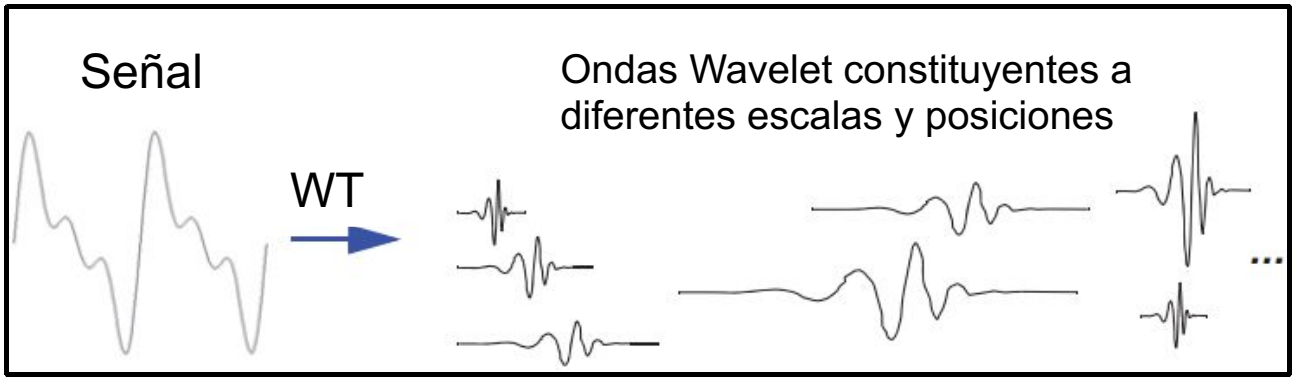


Figura 3: Efectos conjuntos de traslación y escalado en WT

Como puede apreciarse en la figura 3, tras la WT la señal estará representada por distintas versiones de la Wavelet madre utilizada. Aquellas que, por escala y posición, tengan similitud con la señal analizada (o una porción de ésta), obtendrán coeficientes cuyo valor absoluto será mayor que aquellas que no sean similares.

Las técnicas que involucran a la WT pueden encuadrarse en 2 categorías: la WT Continua (CWT) y la WT Discreta (DWT).

Para una función f(t) y una función wavelet madre Ψ(t), donde t representa al tiempo, la CWT, con un parámetro a>0 de escala y con otro b de posición, puede definirse de la siguiente manera:

$$C(a,b,f(t),\psi(t))=\int_{-\infty}^{+\infty} f(t)\,a^{-1/2}\Psi^{*}\left(\frac{t-b}{a}\right)dt \quad (1)$$

donde * representa al conjugado complejo. De la ecuación anterior, se deduce que los coeficientes serán función de los parámetros de escala y posición, así como también del producto interno entre la función o señal analizada y la función Wavelet elegida. El cómputo de estos productos internos es el responsable de evaluar las similitudes.

Aunque la CWT puede utilizarse, el cálculo de coeficientes Wavelet a cada escala y posición posibles implica muchos recursos de cómputo y la generación de una gran cantidad de información parcialmente redundante y de difícil interpretación. De lo anterior sobreviene la posibilidad de seleccionar algunas escalas y posiciones para realizar los cálculos. Específicamente, si las escalas y posiciones están basadas en potencias de 2, el análisis puede ser eficiente y exacto. Esto da origen a la DWT, que puede implementarse en base al algoritmo de descomposición rápida propuesto por

Mallat (Mallat, 1989). La DWT es capaz de retener suficiente información de la señal, puede ser implementada con menores recursos y más rápido que la CWT, a la vez que la descomposición de la señal es única (Walczak y Massart, 1997a).

La DWT, que no es otra cosa que una discretización de la CWT, es una técnica que envuelve cálculos a través de ventanas, es decir, las comparaciones entre las Wavelet y la señal se realizan por partes. Estas ventanas, a su vez, serán móviles por traslación, e irán variando su tamaño al cambiar la escala de análisis. Las ventanas más amplias darán información referente a componentes de baja frecuencia en la señal, mientras que las más angostas lo harán para los de alta frecuencia. Vale destacar que la CWT, al ser aplicada en una computadora común, no es en realidad continua sino discreta y por ende merecería también ser considerada DWT. Sin embargo, se la considera continua porque admite cualquier tipo de escalas y, en relación a las ventanas móviles, éstas son trasladadas suavemente a través de la señal en todo su dominio, solapándose entre una posición y la siguiente. De lo último surge que la información en la CWT es parcialmente redundante.

Habiendo elegido escalas y posiciones, la Wavelet madre Ψ(t) dará origen a múltiples Wavelets para el análisis, según:

$$\psi_{a,b} = a^{-1/2} \Psi\left(\frac{t-b}{a}\right) \quad (2)$$

donde a, b ∈ R y a ≠ 0. En la ecuación (2), "a" representa a una variable de escalado, "b" a una variable de traslación y "t" a una variable de tiempo. En el caso de la DWT, los valores de a y b irán variando en potencias de 2. Cuando se usan factores mayores que 2, se suele hablar de Wavelets de alta multiplicidad (Walczak, 2000).

La DWT, que es la que se utilizó en el presente trabajo, también puede ser representada mediante notación vectorial/matricial:

$$\mathbf{w} = \mathbf{W} f \quad (3)$$

donde "f" representa a la señal de interés, "**w**" es el vector de coeficientes de la WT y "**W**" es una matriz (ortogonal para algunas familias de Wavelets) compuesta de las funciones Wavelet base. Suponiendo que la señal analizada tiene una cantidad de elementos N coincidente con una potencia de 2 (caso contrario, los algoritmos de Matlab utilizados en este trabajo están adaptados para poder realizar igualmente los cálculos), la matriz **W** será cuadrada, con dimensiones N × N. Esta matriz puede ser vista como un apilamiento de 2 submatrices de dimensiones N/2 × N. La submatriz superior, al ser multiplicada por la señal, producirá N/2 promedios ponderados de los componentes

de la señal tomados de a 2, denominados coeficientes de aproximación, los cuales pueden ser vistos como una versión suavizada de la señal original a la mitad de su resolución. A su vez, esta banda superior que suele denominarse filtro de paso bajo, estará compuesta de tal manera que la primera fila contendrá una serie de valores propios de cada familia de Wavelets, la segunda fila contendrá a los mismos valores, pero trasladados 2 lugares hacia la derecha, y así sucesivamente con el resto de las filas. Similarmente, la parte inferior producirá N/2 diferencias de elementos constituyentes tomados de a 2, denominadas coeficientes de detalle, más bien útiles para obtener información sobre cómo se producen cambios entre elementos consecutivos de la señal, algunos propios de cada escala (Walczak y Massart, 1997b). La parte inferior de **W**, comúnmente denominada filtro de paso alto, es construida de igual manera que la superior, sólo que con una serie de valores distintos y que, a su vez, deben sumar cero. En algunas familias de Wavelets, existe una relación entre una serie de valores y la otra, es decir, una serie puede ser construida a partir de la otra.

Respecto de **W**, si se pretende aplicar la WT y luego tener la posibilidad de retornar al dominio original en un paso inverso, también hay que destacar que la inversa de la matriz Wavelet debe ser la transpuesta de una matriz Wavelet, pudiendo o no ser coincidentes. En algunos casos la matriz para WT se construye con valores que la hacen ortogonal, por lo que la inversa de la matriz de transformación es su transpuesta, lo cual es beneficioso en términos computacionales, ya que por simple transposición se evitan cálculos complejos cuando se requiere la inversión de matrices. Sin embargo, existen otros casos en los cuales la matriz de transformación se genera con valores que no permiten la ortogonalidad. En estos casos, la inversa de la matriz se obtendrá a partir de la transposición de otra matriz distinta, compuesta de otros valores. Obviamente, para trabajar con estas familias de Wavelets será necesario contar con ambas matrices.

Para ejemplificar una transformación, la figura 4 esquematiza la reducción de una señal de 4 variables. En dicha figura, las matrices Wavelet corresponden a las de Haar, de las cuales se hablará posteriormente. Puede verificarse para ambas matrices **W** que sus inversas son iguales a sus transpuestas. También puede apreciarse que una vez que la señal es descompuesta en una escala, los coeficientes de aproximación obtenidos tomarán el lugar de la señal en el siguiente nivel de descomposición y por lo tanto darán aproximaciones y detalles si son descompuestos en una escala posterior. Es con la lógica anterior como se realiza el análisis multiresolución. En la primera matriz Wavelet definida puede observarse que tanto las aproximaciones como los detalles son calculados sin solapamiento. Es decir, dada la disposición de los elementos en la matriz, la primera

aproximación y el primer detalle sólo provendrán de los valores s_1 y s_2, sin que s_3 y s_4 participen en dichos cálculos. Esto es propio de las matrices de Haar, ya que otras familias de Wavelets sí contemplan solapamientos.

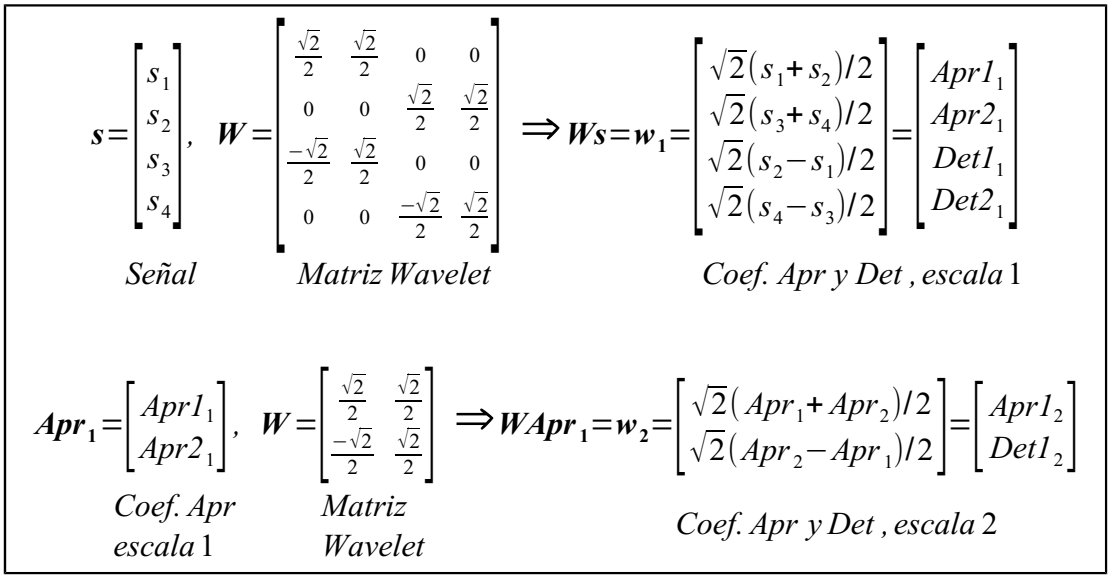


Figura 4: Reducción unidimensional mediante WT de una señal genérica de 4 variables en 2 escalas

Referencias: $\mathbf{w_n}$: coeficientes Wavelet para la escala n, $AprK_n$: K-ésimo coeficiente de Aproximación para la escala n, $DetK_n$: K-ésimo coeficiente de Detalle para la escala n. En todos los casos, K toma valores enteros desde 1 hasta la mitad del número de filas o columnas de la matriz **W** utilizada en la WT que dio origen a los coeficientes.

Debe notarse el carácter local de la WT, en el sentido de que los cómputos son realizados entre valores cercanos y sólo unos pocos valores son afectados en cada tramo de la transformación. Lo anterior contrasta por ejemplo con la Transformada Rápida de Fourier (FFT) en la cual, si existe alguna perturbación en la señal analizada, el resultado entero de la FFT se verá afectado. Esta habilidad de las Wavelets para localizar puntualmente cambios es una de las grandes ventajas de esta transformación. No obstante, la localización de algunos cambios depende de la cantidad de escalas utilizadas. Por ejemplo, suponer que se tiene una señal vectorial de 8 variables, donde el valor de todas es similar a excepción de la quinta y la sexta, con valores mayores pero similares entre sí, lo cual representaría un salto abrupto en la señal. Si la WT se ejecuta en una única escala y con la matriz de Haar, el cambio brusco entre la cuarta y la quinta variable no será percibido en los coeficientes de detalle. Esto ocurriría porque la cuarta variable sería restada con la tercera, y la sexta con la quinta, obteniendo en ambos casos coeficientes de detalle cercanos a cero, ya que los pares

comparados tienen valores similares. En cambio, si la WT con los filtros de Haar fuera nuevamente aplicada en escalas posteriores, entonces sí habría detección del salto brusco. Este defecto que se ha remarcado es propio de la matriz de Haar, ya que en el cálculo de coeficientes no existen solapamientos. Las matrices Wavelet de otras familias que sí utilicen solapamiento, por ejemplo las de Daubechies, detectarían los cambios bruscos desde la primera escala, aunque a nivel de cómputo requieran más recursos.

De acuerdo al algoritmo de descomposición rápida propuesto por Mallat (Mallat, 1989), el vector entero de la señal a ser descompuesta es filtrado a través de los filtros de paso bajo y alto, y sus salidas son separadas en lo que se conoce como coeficientes de aproximaciones y de detalles, respectivamente. Este procedimiento puede continuar con la transformación de las aproximaciones hasta que ya no sea posible proseguir con las reducciones, es decir, hasta que quede solo un coeficiente de aproximación y sólo uno de detalle, pero usualmente la ejecución del algoritmo es detenida una vez que el nivel de descomposición puede considerarse óptimo, lo cual dependerá de los objetivos de cada caso en particular. El análisis en múltiples escalas se esquematiza en la figura 5, obtenida de la referencia (MATLAB *7.6.0*, 2008):

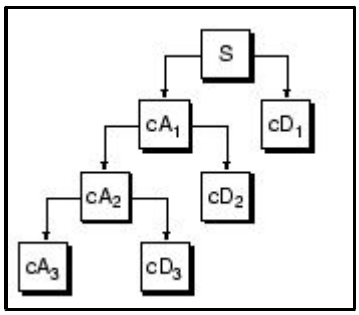


Figura 5: Esquema de WT en 3 escalas

Referencias: S: señal analizada, cA_n: coeficientes de aproximación para la escala n, cD_n: coeficientes de detalle para la escala n

Luego de lo anterior, la recuperación de la señal original puede llevarse a cabo mediante la operación inversa, denominada WT Inversa (IWT), mediante el uso de filtros de paso bajo y alto de reconstrucción. Vale destacar que en Matlab, el algoritmo propuesto por Mallat para la WT es ejecutado de manera tal que luego de filtrar la señal con los filtros de paso bajo y alto, es necesario remover 1 de cada 2 elementos en los coeficientes resultantes, procedimiento que en inglés se denomina *downsampling* y que no envuelve pérdida de información. Similarmente, en la IWT se

requiere que entre cada coeficiente se intercale un 0 antes de la reconstrucción final, lo cual en inglés se denomina *upsampling*. A su vez, los filtros involucrados en la WT y sus respectivos de la IWT forman sistemas conocidos como Filtros Espejo en Cuadratura (QMF, del inglés *Quadrature Mirror Filters*). La figura 6, adaptada desde la referencia (MATLAB *7.6.0*, 2008), esquematiza estos conceptos:

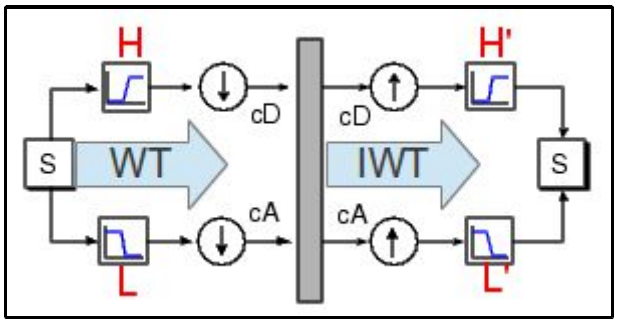


Figura 6: Esquema de descomposición mediante WT y reconstrucción a través de IWT de una señal genérica utilizando el algoritmo de Mallat

Referencias: S: señal analizada, H: filtro de paso alto en WT, L: filtro de paso bajo en WT, H': filtro de paso alto en IWT, L': filtro de paso bajo en IWT, cD: coeficientes de detalle, cA: coeficientes de aproximación, flecha hacia abajo: *downsampling*, flecha hacia arriba: *upsampling*. Los filtros conforman un QMF.

La elección de los filtros no debería ser trivial, ya que determina si la reconstrucción de señales se puede realizar perfectamente o no, además de que son los filtros los que determinan las formas de las Wavelets que se usan en los análisis (MATLAB *7.6.0*, 2008). En efecto, dado un par de filtros de reconstrucción, L' y H', la realización de *upsampling* sobre el último, la convolución del resultado con L' y la repetición iterativa de los 2 últimos pasos, se obtendrá una función denominada propiamente Wavelet (Ψ, *psi*), cuya forma estará totalmente determinada por los coeficientes de los filtros de reconstrucción. Para algunas Wavelets (aunque no para todas) existe otra función, denominada función de Escalado (Φ, *fi*), la cual puede obtenerse de forma similar a la descripta, solo que la dependencia residirá en los coeficientes del filtro de paso bajo.

Todo lo discutido hasta aquí con señales de primer orden puede aplicarse también para el tratamiento de datos matriciales. En este caso, la reducción de señales puede realizarse mediante WT Bidimensional (WT2). Existen dos formas de generalizar la WT clásica hacia dos dimensiones: la estándar y la no estándar (Walczak y Massart, 1997b). La primera es atractiva por su simplicidad, ya que solo requiere realizar una WT unidimensional en cada vector fila de una matriz,

prosiguiendo de la misma manera con los vectores columna. La segunda alterna las operaciones entre filas y columnas. La elección de una u otra depende de la aplicación que debe llevarse a cabo. En cada nivel, la WT2 producirá coeficientes de aproximación, y 3 tipos de coeficientes de detalle: horizontales, verticales y diagonales (específicamente cambios a 45º).

Independientemente de la naturaleza vectorial o matricial de los datos, así como también de la Wavelet elegida, existe un problema común asociado a los bordes de las señales cuando se aplica la WT. El algoritmo de la DWT no está limitado a ser aplicado en señales con un número de variables coincidente con una potencia de 2 y básicamente tiene 2 pasos: convolución con los filtros y *downsampling*. El hecho de realizar convoluciones en señales finitas trae aparejadas distorsiones en los bordes. De lo anterior se deduce que el tratamiento en los bordes debe ser diferente al del resto de las señales analizadas. Específicamente, la señal deberá ser extendida en los bordes con el objeto de alcanzar una longitud apta para la WT, para lo cual existen diversas alternativas. Estas extensiones conllevan el cálculo de coeficientes extra en cada escala de descomposición, lo cual permitirá la reconstrucción exacta de las señales reducidas. Dentro de las opciones para extensiones, se pueden nombrar el rellenado con ceros, la simetrización en los bordes (modo por defecto en Matlab), la periodización en los bordes, extrapolaciones de distintos tipos y combinaciones de las anteriores (MATLAB *7.6.0*, 2008).

Entre los distintos tipos de Wavelets, la más simple es la de Haar (Haar, 1910), la cual también es el primer miembro de la familia Daubechies (Daubechies, 1992) de Wavelets ortogonales, caracterizada por 2 coeficientes, c_0 y c_1 (Trygg y Wold, 1998). En Matlab, por cuestiones relativas a la ortogonalidad de matrices y a sus inversiones, el valor absoluto de ambos coeficientes es $(2^{1/2})/2$, pero podrían ser otros. El filtro de paso bajo se define como (c_0 , c_1), y el de paso alto como ($-c_0$, c_1). La wavelet Haar es la única que presenta simetría, dominio compacto y ortogonalidad al mismo tiempo (Walczak, 2000). También, dado a que en su filtro de paso bajo ambos coeficientes son positivos, garantiza la propiedad de no-negatividad en las aproximaciones de la señal, lo cual permite la aplicación de métodos comunes para resolución multivariada de curvas con restricciones de no-negatividad (Peré-Trepat y Tauler, 2006), como en el presente trabajo. Sin embargo, debe aclararse que esto también depende de que no haya datos negativos en los datos originales, pues en ese caso sí podrían obtenerse coeficientes de aproximación negativos. No obstante, otras familias de Wavelets no garantizan la propiedad de no-negatividad aun cuando se garantice que los datos a ser reducidos sean todos mayores que cero. Por ejemplo, el filtro de paso bajo de Daubechies-4 contiene un coeficiente negativo, específicamente el cuarto, de menor valor absoluto que el resto y

por ende con menor efecto en la ponderación del promedio. Por lo tanto, es posible que al multiplicar 4 valores de la señal con la Wavelet en cuestión se obtenga un promedio ponderado negativo, si es que el cuarto valor de la señal es mucho mayor en relación al resto. A su vez, la wavelet Haar no es suave ni continua y, por ende, no es diferenciable, aunque su forma puede resultar ventajosa en el análisis de señales con transiciones abruptas (siempre que se garantice el nivel de escalas necesario), como por ejemplo las provenientes de espectrometría de Masa aquí utilizadas.

La figura 7 esquematiza las funciones asociadas a la Wavelet de Haar, así como también su QMF y los valores asociados de los coeficientes:

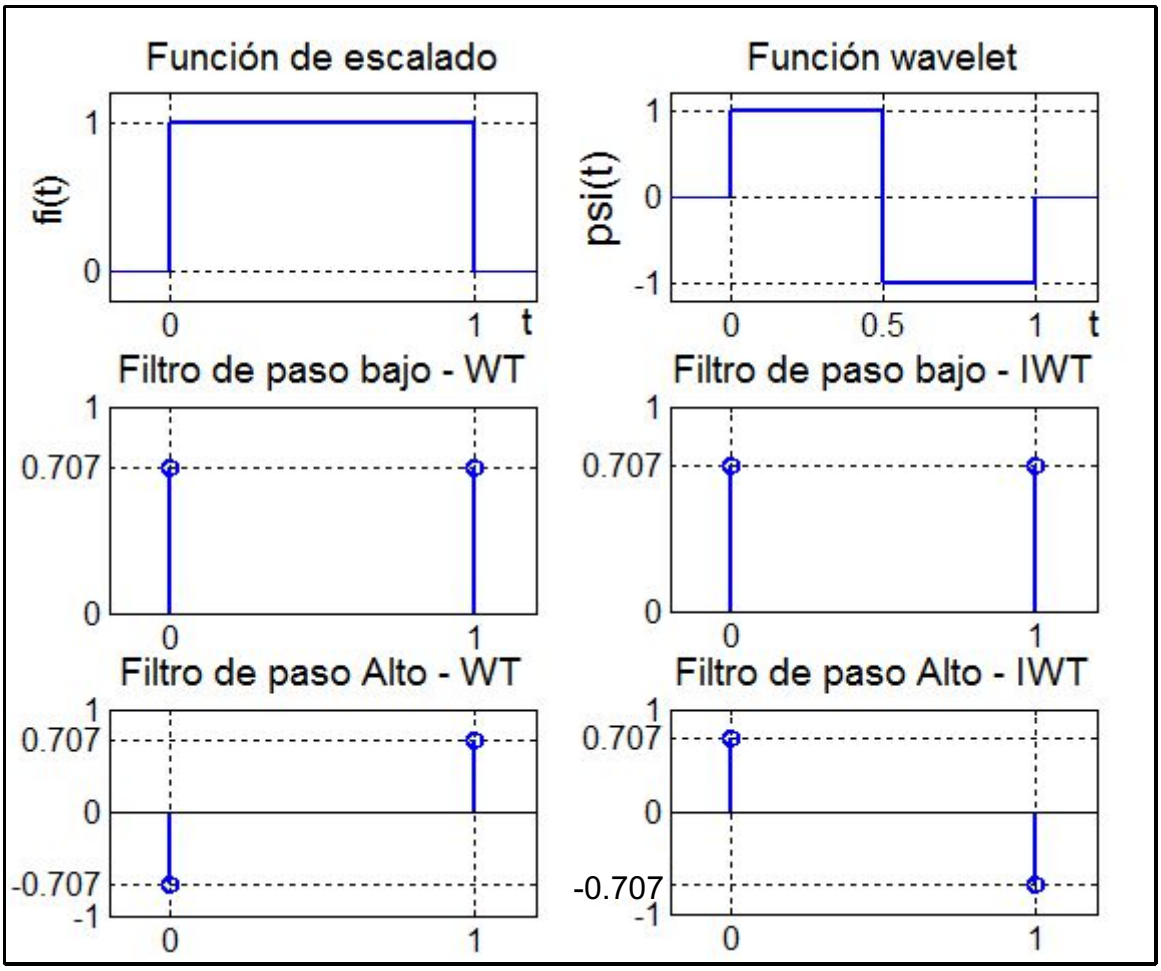


Figura 7: Función de Escalado, Wavelet y QMF de Haar

Referencias: t: tiempo. En los ejes horizontales de los filtros se utilizaron los valores 0 y 1 solamente para nominar a cada valor.

La función de escalado de Haar se define como:

$$\Phi(t)=\begin{matrix}1 & 0\le t<1\\ 0 & cualquier\ otro\ caso\end{matrix} \quad (4)$$

Por su parte, la función wavelet de Haar se define como:

$$\Psi(t)=\begin{matrix}1 & 0\le t<1/2\\ -1 & 1/2\le t<1\\ 0 & cualquier\ otro\ caso\end{matrix} \quad (5)$$

2.4.2 Resolución Multivariada de Curvas mediante Mínimos Cuadrados Alternantes (MCR-ALS).

Los denominados instrumentos de segundo orden pueden generar datos bilineales o no-bilineales, lo cual determina qué tipo de algoritmo puede ser usado para el tratamiento. En el modelo bilineal, se asume que los valores presentes en una matriz **D** son dependientes a nivel lineal respecto de 2 tipos de componentes. Por ende **D** puede ser entendida y descompuesta en el producto de dos matrices (una para cada tipo de componentes) **C** y $\mathbf{S}^T$ (T representa la transposición de la matriz **S**), las cuales contendrán los perfiles de concentraciones y los espectros puros de los componentes en **D**, respectivamente:

$$\mathbf{D}=\mathbf{C}\mathbf{S}^{\mathbf{T}}+\mathbf{E} \quad (6)$$

donde **D** (J × K) es la matriz de datos originales, **C** (J × N) y $\mathbf{S}^T$ (N × K) son las matrices que contendrán los perfiles de concentración (en las columnas de **C**) y los espectros puros (en las filas de $\mathbf{S}^T$) de cada componente, respectivamente, y **E** es la matriz de error o residuos no ajustados, la cual representará a las variaciones residuales del conjunto de datos que, en el mejor de los casos, no estarán relacionadas con ninguna contribución química. Los parámetros J y K representan al número de filas y columnas, respectivamente, de la matriz **D**, que en nuestro caso se corresponden con tiempos cromatrográficos y variables de m/z, respectivamente. Por su parte, N es el número de componentes químicos o principales de la mezcla o proceso. Por lo tanto, para una mezcla de N componentes sometidos a HPLC y detectados con MS, **D** será un espectrocromatograma bidimensional compuesto de espectros de MS (horizontales) en función de su tiempo de elución cromatográfico (en la vertical) y de su descomposición podrán obtenerse N perfiles de concentración en **C** y N perfiles espectrales en **S**.

La función principal de los métodos de resolución es la descomposición matemática de señales

instrumentales globales y con componentes mixtos en sus contribuciones puras debidas a cada componente del sistema. MCR-ALS es una herramienta popular en el mundo de la quimiometría, la cual ha sido utilizada satisfactoriamente para resolver respuestas de componentes múltiples a partir de muestras desconocidas (Tauler y col., 1993b; Zachariassen y col., 2006; Jaumot y col., 2005). Así pues, esta técnica se ha mostrado útil en el tratamiento de arreglos de datos de 2 y 3 vías, siendo una de sus principales ventajas la adaptación a conjuntos de datos de complejidad y estructura diferentes, dando soluciones óptimas en términos de mínimos cuadrados (Tauler y col., 1998). A su vez, su versatilidad permite su aplicación a cualquier sistema multicomponente, a través de datos matriciales que puedan ser descriptos con un modelo bilineal, provenientes de reacciones químicas, eluciones cromatográficas, datos ambientales y otros, monitoreados a través de respuestas multivariadas, como medidas espectroscópicas, señales electroquímicas, perfiles de composición, entre otras (Walczak, 2000).

MCR-ALS se corresponde con un algoritmo de resolución iterativo que es utilizado para recuperar las contribuciones que dan origen a los datos, pudiendo éstas ser expresadas normalmente como perfiles de concentración y como espectros puros para cada componente involucrado (de Juan y col., 2000; Muñoz y de Juan, 2007). Por ende, el objetivo de aplicar MCR-ALS a nuestros datos bilineales consistió en solucionar la ecuación (6), obteniendo en **C** y en **S** todas las señales individuales de cada componente puro en los 2 órdenes o vías de medición.

La optimización se realiza postulando un número de componentes que hipotéticamente generan las variaciones experimentales (la metodología para calcular este número será discutida posteriormente). Para cada componente, en principio debe obtenerse una estimación inicial que puede corresponderse con su perfil espectral o de concentración. Cabe mencionar que dentro de lo posible, las estimaciones iniciales deberían ser parecidas a los perfiles finalmente resueltos, ya que de esta manera se verá favorecida la resolución, tanto en la velocidad de los cómputos como en la minimización de efectos no deseados que serán discutidos posteriormente. Si el sistema analizado es conocido, es posible estimar los resultados finales y usarlos para iniciar la optimización. También, si cabe la posibilidad, se pueden utilizar espectros puros de los componentes, obtenidos de estándares apropiados o de bases de datos. Cuando el sistema es realmente desconocido, las aproximaciones iniciales pueden obtenerse a través de SIMPLISMA (o métodos similares y derivados) (Windig y Guilment, 1991; Garrido y col., 2004), el cual puede interpretarse en este contexto como un algoritmo de resolución multivariada de curvas que extrae las variables más puras y, a su través, estimaciones de los espectros puros de los componentes a partir de una mezcla de

espectros de composición variada. El objetivo de asociar a los componentes con ciertas variables y no con otras se relaciona con el aumento de selectividad en la resolución final. En esta tesis, exceptuando situaciones oportunamente identificadas, todas las estimaciones fueron obtenidas con SIMPLISMA. Suponiendo que las aproximaciones se corresponden con información espectral, la optimización puede comenzar hallando una primera estimación de los perfiles de concentración mediante mínimos cuadrados, según:

$$\hat{\mathbf{C}} = \mathbf{D}(\mathbf{S}^{\mathrm{T}})^{+} \quad (7)$$

donde $\hat{\mathbf{C}}$ representa la estimación de perfiles de concentración, el signo + simboliza inversa generalizada y como es usual, T representa transposición. Habiendo obtenido a $\hat{\mathbf{C}}$, es posible alternar la información y realizar un nuevo paso de ajuste por cuadrados mínimos, obteniéndose una nueva aproximación de los perfiles espectrales, según:

$$\hat{\mathbf{S}} = ((\hat{\mathbf{C}})^{+}\mathbf{D})^{\mathrm{T}} \quad (8)$$

donde $\hat{\mathbf{S}}$ representa la estimación de los perfiles espectrales. A continuación es posible calcular la matriz de residuos **E**, utilizando a **D**, $\hat{\mathbf{C}}$ y a $\hat{\mathbf{S}}$ en la ecuación (6). Los pasos descriptos serán repetidos a lo largo de las iteraciones, hasta que se alcance un determinado criterio de convergencia o alguna condición adicional de parada. Así pues, excluyendo restricciones adicionales, la fuerza directriz en la resolución de los datos radicará en la reducción de la norma de la matriz de residuos **E** mediante ALS.

Debe notarse que la solución a la ecuación (6) no es única, ya que muchos pares de matrices **C** y **S** podrían multiplicarse y satisfacerla con el mismo grado de error. Desde el punto de vista cualitativo esto no es necesariamente problemático, pero desde lo cuantitativo sí, ya que distintos perfiles en **C** representarían concentraciones distintas. En la práctica, cuando se desean cuantificar los componentes con métodos como HPLC-MS, deberán utilizarse patrones obtenidos en las mismas condiciones cromatográficas que las muestras incógnita. De esta manera, no importará qué proporción haya tomado **C** sobre **S**, ya que la cuantificación podrá realizarse relacionando las áreas bajo los perfiles de los componentes con las de sus respectivos patrones, aunque estos no posean a potenciales interferentes modelados. Lo anterior representa la base de la denominada ventaja de segundo orden (Booksh y Kowalski, 1994). A su vez, si se cuenta con patrones, podrán establecerse restricciones a los perfiles a resolver, de manera tal que los perfiles de los patrones sean respetados sin ser modificados o con leves modificaciones y esto producirá un aumento de la selectividad en la resolución de las incógnitas.

El tipo de ambigüedad descripto en la intensidad de las señales no es el único presente. Existe también la denominada ambigüedad rotacional, que ocurre cuando 2 o más componentes del sistema poseen uno o ambos perfiles, sean de concentraciones o espectrales, con alta correlación. De esto sobreviene una pérdida de selectividad y los perfiles resultantes suelen no ser propios de ninguna especie, sino una combinación que no se corresponderá con ninguno exactamente.

Los problemas relativos a la ambigüedad de las posibles resoluciones pueden ser minimizados con algunas estrategias, de forma tal que las resoluciones obtenidas no sólo sean aceptables desde un punto de vista lógico-matemático, sino también desde lo experimental. Entre ellas, vale nombrar la estrategia de aumentar los datos mediante el procesamiento conjunto de muestras y la aplicación de restricciones matemáticas durante los procesos iterativos en la ejecución de MCR-ALS, las cuales serán discutidas posteriormente.

2.4.2.1 Cifras de mérito para MCR-ALS

La calidad de un modelo MCR-ALS queda indicada con diferentes cifras de mérito relacionadas a la reproducción correcta de los datos originales a partir de los perfiles resueltos. En el presente trabajo se han utilizado las siguientes:

- Falta de ajuste experimental porcentual (%LOF o %LOF EXP, del inglés *lack of fit*):

$$\%\,\mathrm{LOF} = \sqrt{\frac{\sum (d_{ij}^{*} - d_{ij})^2}{\sum d_{ij}^2}} \times 100 \quad (9)$$

donde d_{ij} es un elemento de la matriz experimental **D**, mientras que d^{*}_{ij} es un elemento de la matriz reproducida a partir de los perfiles resueltos ($\mathbf{D}^{*}=\mathbf{CS}^{T}$).

- Porcentaje de varianza explicada ($\%R^2$)

$$\%\,R^2 = \frac{\sum (d_{ij}^{*})^2}{\sum d_{ij}^2} \times 100 \quad (10)$$

2.4.2.2 Resolución conjunta de múltiples muestras mediante apilamiento

MCR-ALS puede ser aplicado a matrices individuales interpretadas simplemente como datos de 2 vías, o también a datos (o arreglos) de 3 vías. La tercera vía puede lograrse mediante el agrupamiento de 2 o más muestras, colocando cada matriz al lado de otra, o bien una debajo de otra, lo cual dependerá de la naturaleza de las experiencias que dan origen al análisis. En ambos casos,

los datos son considerados aumentados, y estos arreglos de datos de experimentos múltiples también siguen el modelo bilineal, tal y como si fueran una única matriz (de Juan y Tauler, 2003). En el presente trabajo se utilizaron apilamientos verticales de matrices porque todas las muestras fueron cromatografiadas y reveladas solamente con espectrometría de Masa, lo cual es esquematizado en la figura 8 para 3 matrices con 3 componentes hipotéticos.

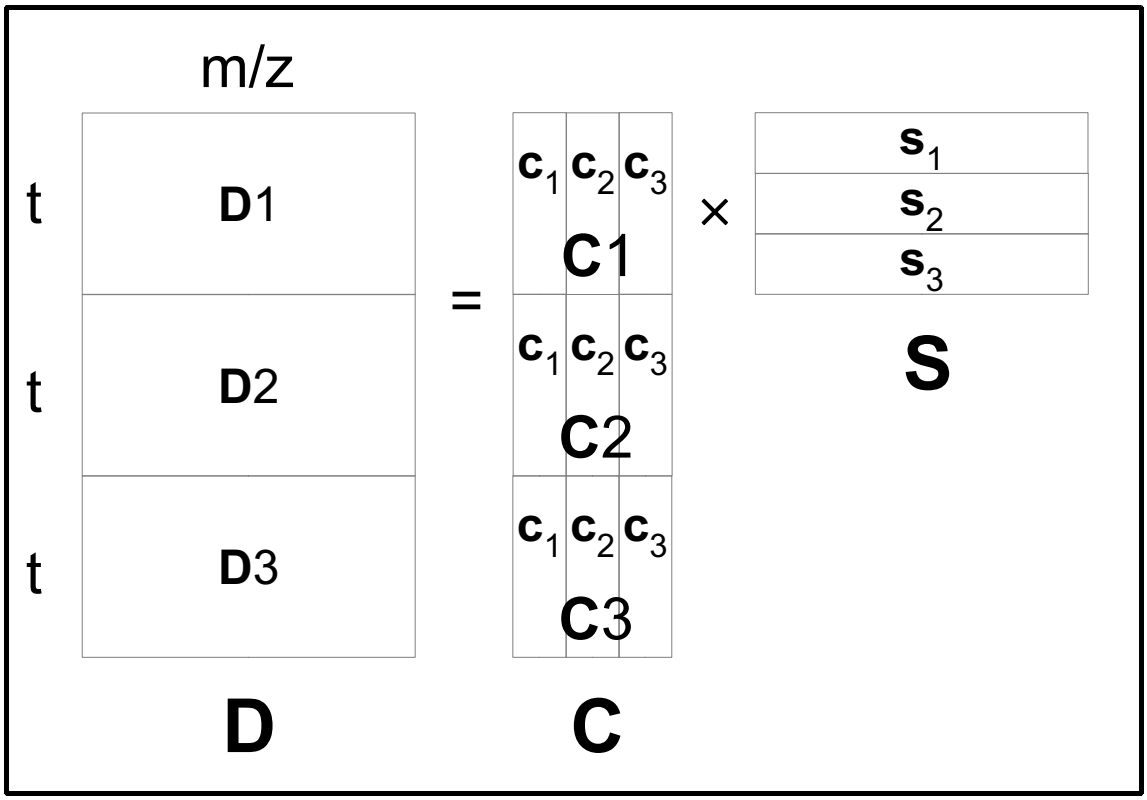


Figura 8: Esquema de apilamiento vertical de matrices para MCR-ALS

Referencias: t: tiempo, m/z: relación masa/carga, **D**: matriz apilada, **D**N: Matrices de muestras individuales, **C**: perfil de concentraciones de **D**, **C**N: perfiles de concentraciones para la matriz **D**N, $\mathbf{c}_k$: perfil de concentración del componente k, **S**: perfil espectral común a todas las **D**N. N va desde 1 hasta 3 y su objetivo es identificar a cada muestra individual. k va desde 1 hasta 3 y su objetivo es identificar a cada componente en la resolución.

En el esquema de la figura 8 se ha obviado la matriz de residuos no ajustados **E**, aunque igualmente es su norma lo que se minimiza a lo largo de las iteraciones de MCR-ALS. No obstante, puede apreciarse que el apilamiento vertical conduce a que las columnas de cada matriz (información espectral) serán las mismas que las del apilamiento final. Por ende, pueden existir

variaciones de los perfiles de concentración a nivel de componentes individuales en cada muestra apilada, pero la información espectral deberá ser la misma para toda la resolución. Esta imposición conduce a la minimización del efecto de ambigüedades y deficiencias de rango. En la referencia (Peré-Trepat y Tauler, 2006) se hizo notar que en la resolución de muestras ambientales los resultados obtenidos por apilamiento fueron mejores que los logrados habiendo analizado cada muestra por separado. Más aun, en dicho trabajo se menciona que el uso adicional de estándares puros o mezclas de estándares y la resolución de éstos en conjunto con las muestras incógnita mediante apilamiento, aportaron información valiosa relacionada con aumentos de selectividad, lo cual permitió la resolución de componentes cuyos perfiles de elución estaban fuertemente solapados. Vale destacar que si existe información fehaciente, las restricciones en los cómputos (que serán discutidas a la brevedad) pueden ser aplicadas al conjunto de los datos aumentados, a muestras individuales e incluso a ciertos componentes en particular.

2.4.2.3 Aplicación de restricciones en MCR-ALS

Tal como se explicó anteriormente, la resolución realizada por MCR-ALS comienza con una estimación inicial de algunos perfiles (**C** o **S**) que, a través de un proceso por iteraciones, optimiza los perfiles en cuestión, introduciendo la información disponible sobre el sistema en estudio mediante la implementación de restricciones (Leger y Wentzell, 2002). Aunque no sea estrictamente necesaria para la resolución de curvas, la imposición de restricciones en los cómputos puede mejorar los resultados, fundamentalmente para que éstos sean coherentes con la naturaleza de las muestras en cuestión, más allá de la lógica estrictamente matemática. Las restricciones pueden definirse como propiedades matemáticas y/o químicas (en este contexto) que deban ser cumplidas sistemáticamente por el sistema entero en estudio o por alguna de sus contribuciones puras. Así pues, las restricciones son traducidas a lenguaje matemático y pueden forzar el proceso iterativo de optimización de manera tal de modelar los perfiles respetando las condiciones establecidas, minimizando el efecto de ambigüedades de intensidad y rotacionales (de Juan y col., 1997). A su vez, durante la optimización las restricciones afectan el cálculo de pseudoinversas, por lo cual las soluciones halladas no serán estrictamente soluciones en términos de mínimos cuadrados. Vale destacar que la aplicación de las restricciones debe ser cautelosa y que lo recomendado es tener certeza de su validez. Aun cuando existan restricciones potencialmente aplicables, estas pueden resultar en un detrimento de los resultados si, por ejemplo, existe ruido experimental o problemas instrumentales que puedan distorsionar los perfiles de algunos compuestos hasta el punto tal que ya

no se cumplan las condiciones asumidas. También cabe mencionar que debe existir una forma de indicar cuánta tolerancia ha de tenerse para considerar que una restricción tiene o no que ser impuesta, en especial cuando los datos son reales y contienen problemas de ruido, para lo cual deberán permitirse leves desviaciones del comportamiento ideal (de Juan y col., 1997).

En el presente trabajo se utilizaron las siguientes restricciones, aunque existen otras disponibles:

- No negatividad (Lawton y Sylvestre, 1971): Previene la presencia de valores negativos en los perfiles. En su forma general se realiza convirtiendo los valores negativos a cero en los perfiles optimizados antes de continuar con las iteraciones siguientes. Se aplicó tanto a los perfiles de concentraciones como a los espectrales, ya que ambos no pueden tener valores negativos. En el caso espectral sí es posible ya que algunas espectrometrías otorgan intensidades negativas, pero no para los datos de espectrometría de Masa aquí trabajados. En particular, se utilizó el algoritmo descripto en la referencia (Bro y De Jong, 1997).

  - Tolerancia a la no negatividad: el algoritmo impone una por defecto y así fue utilizado.

- Unimodalidad (Tauler y col., 1993a): Fuerza los cómputos y garantiza la presencia de un único máximo por perfil calculado. En nuestro caso fue aplicado a los perfiles de elución cromatográficos y, de las opciones disponibles, se utilizó el modo promedio, en el cual los máximos secundarios son corregidos mediante el cálculo de promedios, tal y como en los algoritmos de unimodalidad con mínimos cuadrados (Bro y Sidiropoulos, 1998). En general esta restricción puede aplicarse a perfiles de concentración que muestran una forma de aparición-decaimiento, clásicos en HPLC y otros métodos.

  - Tolerancia a la unimodalidad: la interfaz de MCR-ALS permite indicar mediante un valor en qué proporción pueden desviarse los perfiles calculados de la idealidad. En la referencia (Jaumot y col., 2005) se especifica que si el valor es 1.5, esto significa que puede existir un 50% de desvío o, en otras palabras, que en la bajada desde el pico máximo un punto en particular puede incrementarse en un máximo del 50% respecto del valor previo, antes de considerar que la restricción de unimodalidad debe ser aplicada. Se recomiendan valores entre 1.0 (no existe posibilidad de desvíos de la idealidad) y 1.1 para sistemas de bajo o mediano nivel de ruido. En este trabajo se utilizó el valor 1, es decir, sin tolerancia.

- Criterio de convergencia: existe la posibilidad de establecer una condición de parada de las iteraciones, basada en la diferencia de la desviación estándar de los residuos entre

iteraciones consecutivas. Utilizamos el valor por defecto, que es del 0.1%.

- Cantidad de iteraciones: esto simplemente evita que el cálculo, de no hallarse convergente, continúe de manera indefinida. Se utilizó un máximo de 300.
- Normalización de los perfiles espectrales: Como ya se explicó, varios pares de matrices **C** y **S** podrían producir resultados igualmente válidos. Por lo anterior y para evitar problemas con las escalas de las soluciones propuestas, se optó por normalizar los perfiles espectrales, opción incluida en la interfaz.

    - Esta restricción en las escalas de los perfiles espectrales conlleva indirectamente un escalado de los perfiles de concentración que, como ya se ha discutido, implicaría el uso de estándares si se requirieran cuantificaciones. En esos casos, el uso de patrones no sólo aumentaría la selectividad, sino que a su vez tiene relación con una restricción no utilizada en el presente trabajo, pero que vale la pena mencionar en este contexto. La restricción en cuestión trata sobre el establecimiento certero de la presencia y/o ausencia de ciertos componentes en las muestras analizadas. Como el par presencia/ausencia brinda información netamente cualitativa, esta restricción suele informarse con una matriz binaria dedicada a tal fin (Ruckebusch y col., 2006).

2.4.2.4 Descomposición en Valores Singulares (SVD) para estimar el número de componentes generadores de varianza.

MCR-ALS soluciona iterativamente la ecuación (6) para un número propuesto de componentes que, hipotéticamente, generan toda o la mayoría de las variaciones en los datos. Si se tiene un conocimiento suficiente, este número puede ser determinado directamente, pero en el caso en que se desconozca, su estimación debe ser el primer paso a realizar. Los métodos para realizar las estimaciones tienen distintos fundamentos, pero en definitiva han de realizar alguna suposición sobre los datos, su distribución, su variabilidad, entre otras. Sin embargo, estos esquemas están basados en principios matemáticos y/o estadísticos, por lo cual la información que otorgarán sobre los datos no necesariamente estará relacionada de manera directa, aunque sí parcialmente, a la naturaleza química de éstos. En relación a lo anterior, es común en el tratamiento de datos complejos que se reserve un número usualmente pequeño de componentes hipotéticos útiles para modelar ruido o variaciones sistemáticas en los datos, aun cuando en ninguno de los casos anteriores se puede hablar de naturaleza química. Aunque la reserva de componentes no se realice,

es posible que la estimación matemática sea mayor que el verdadero número de componentes variables a nivel químico, por lo cual las estimaciones pueden imponer un máximo de identidades diferentes capaces de resolver el sistema, más que el número exacto.

Algunos de los métodos utilizados asumen que las variaciones son propias de cada componente, es decir, sin interacciones. Bajo ese punto de vista, los métodos descomponen los datos con ciertos criterios que permiten sistematizar la distribución de la varianza. Al suponer que no existen interacciones se estará postulando la independencia de los componentes y de esto, si los datos son matriciales, sobreviene el concepto de rango matricial, el cual indica la cantidad de componentes linealmente independientes que generan las variaciones. Este concepto da origen a la llamada deficiencia de rango, propiedad de algunos datos matriciales, lo cual significa que ante ciertas condiciones, como por ejemplo cuando los perfiles de 2 o más componentes son muy similares, se obtendrá una representación matemática de menor valor que la real cantidad de especies involucradas (Ruckebusch y col., 2006).

Las estimaciones suelen realizarse con Análisis de Componentes Principales (PCA), una de las herramientas quimiométricas más básicas y de uso extendido dedicada a obtener el número y la dirección de las fuentes relevantes de variación en datos bilineales (Malinowski, 2002). En este trabajo, se ha utilizado la Descomposición en Valores Singulares (SVD) (Lorenz, 1956), íntimamente relacionada con la anterior.

Un valor singular y sus correspondientes vectores singulares para una matriz rectangular **A** son, respectivamente, un escalar σ y un par de vectores **u** y **v**, tales que:

$$\mathbf{A}\mathbf{v} = \sigma\,\mathbf{u} \quad \text{y} \quad \mathbf{A}^T\mathbf{u} = \sigma\,\mathbf{v} \tag{11}$$

Si los valores singulares son agrupados en una matriz diagonal **Σ**, y los vectores singulares lo son en las columnas de 2 matrices distintas, **U** y **V**, ambas ortogonales, se tiene que:

$$\mathbf{A}\mathbf{V} = \mathbf{U}\boldsymbol{\Sigma} \quad \text{y} \quad \mathbf{A}^T\mathbf{U} = \mathbf{V}\boldsymbol{\Sigma} \tag{12}$$

Dado que **U** y **V** son ortogonales, se puede llegar a la siguiente expresión para la SVD:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \tag{13}$$

Si la matriz **A** tiene dimensiones m × n, la SVD descompondrá los datos matriciales en **U** (m × m), **Σ** (m × n) y **V** (n × n). Si **A** tiene más filas que columnas, **U** contendrá columnas que serán multiplicadas con ceros en **Σ**, por lo cual existen variantes algorítmicas de la SVD, a veces denominadas “económicas”, que solamente calculan el número máximo de columnas posibles para

U, reduciendo sus dimensiones a m × n, cambiando las de **Σ** a n × n y dejando **V** intacta. Algo similar ocurre si **A** tiene más columnas que filas, sólo que **U** permanecerá intacta y los cambios afectarán las dimensiones de **Σ** y de **V**. A su vez, una matriz de dimensiones m × n contendrá, como máximo, una cantidad de valores singulares distintos equivalente al mínimo entre m y n, y este máximo posible es el utilizado en las versiones "económicas". Ya que la matriz **Σ** contiene en su diagonal a estos valores, y el resto de la matriz son ceros, la cantidad de elementos positivos en la diagonal coincide con el rango de la matriz analizada. De aquí que la presencia de estos elementos puede utilizarse como indicador de la cantidad de fuentes de varianza presentes y linealmente independientes. A su vez, es común que el algoritmo sea aplicado de forma tal que los elementos de la diagonal en **Σ** estén ordenados de manera decreciente. Así pues, el primer valor singular indicará que el componente asociado es el que genera mayor variabilidad, el segundo lo hará con el siguiente y así sucesivamente. Este orden es estrictamente matemático y por lo tanto no debe pensarse que a nivel químico existe el mismo orden de generación en la varianza.

Una de las variantes de la SVD contempla que las matrices **U** y **V** sean unitarias, y siendo éstas ortogonales, serán también ortonormales. En estos casos, **Σ** puede ser claramente vista como una matriz de escalado. Sean o no unitarias, los vectores columna de **U** y **V** son llamados vectores singulares izquierdos y derechos, respectivamente.

En contraste con otros métodos, una de las características de la SVD es que si la matriz analizada es real, la descomposición también lo será. Por lo tanto, el método se torna apto para datos como los aquí estudiados. En efecto, las matrices a ser analizadas fueron sometidas a SVD y luego de esto, la composición de la matriz **Σ** emergente fue interpretada de manera tal que se pueda estimar un número apropiado de componentes variantes. Específicamente, se estableció una fracción de la varianza total a explicar (en todos los casos fue de 0.90) y se halló la mínima cantidad de valores singulares ordenados que acumularían una suma, relativa a la suma total de los valores singulares, igual o levemente superior al umbral establecido.

2.4.3 Análisis Discriminante - Mínimos Cuadrados Parciales (PLS-DA)

El algoritmo de PLS (Wold y col., 1984; Geladi y Kowalski, 1986; Haaland y Thomas, 1988) inicialmente fue obtenido para análisis cuantitativos, aunque también se puede aplicar a clasificación de muestras, como en el presente trabajo. En su forma genérica (análisis cuantitativos), este análisis de tipo supervisado está basado en la relación entre intensidades espectrales

normalmente descriptas por una matriz **X** (n observaciones × m variables) y características de las muestras comúnmente asociadas a su composición y representadas por el vector **y**. De hecho, los factores o, más apropiadamente, Variables Latentes (LV) de PLS se construyen teniendo en cuenta un compromiso entre 2 propósitos: describir el juego de variables que explican el sistema y predecir las respuestas (Galtier y col., 2011). Comparado con otros algoritmos, PLS se encuentra en un punto que podría denominarse intermedio entre PCR (Regresión en Componentes Principales) y MLR (Regresión Lineal Múltiple). PCR calcula sus factores intentando maximizar la varianza capturada respecto de la presente en las variables de predicción (espectros por ejemplo), mientras que MLR intenta correlacionar de la mejor manera posible a las variables de predicción con las variables a predecir. En este sentido, PLS intenta hacer ambas cosas (capturar varianza y encontrar correlaciones) y por eso comúnmente se dice que PLS intenta maximizar la covarianza (Wise y col., 2005). A su vez, si el espacio de variables a predecir es unidimensional (un sólo analito o propiedad a predecir), entonces se habla de PLS1, siendo PLS2 el correspondiente modelo global para espacios multidimensionales, en cuyo caso **y** encuentra su equivalente en la matriz **Y**. Según los autores de la referencia (Haaland y Thomas, 1988), en análisis de muestras reales se obtuvieron mejores resultados predictivos aplicando PLS1 para más de un analito de manera separada que aplicando PLS2, aunque el primero sea un caso especial del último. Lo anterior suele deberse a cómo se realiza la selección de variables latentes, que son las variables emergentes del modelo. Mientras que en PLS1 esta selección (sin importar el método) se realiza de manera dedicada a un único analito o propiedad, en PLS2 se conjugan selecciones usualmente incompatibles y la decisión final suele no ser óptima para todos los analitos modelados al mismo tiempo, a la vez que estos modelos mixtos son más difíciles de interpretar. En general, PLS es apto para modelar datos cuando hay muchas más variables de predicción que muestras independientes, y cuando entre las variables nombradas existe un alto grado de colinearidad.

A nivel algorítmico existen distintas maneras de calcular los parámetros para modelos PLS, por ejemplo en la referencia (de Jong, 1993) se detalla el algortimo SIMPLS. No obstante, uno de los métodos más intuitivos se conoce como NIPALS (*Nonlinear Iterative Partial Least Squares*) (Wold, 1966), a través del cual pueden ser obtenidas secuencialmente todas las matrices necesarias para el modelado. Según los autores de la referencia (Wise y col., 2005), la descripción de NIPALS es la siguiente:

- La descomposición comienza seleccionando una columna de **Y**, usualmente la de mayor

variabilidad, como primera estimación de $\mathbf{u}_1$, que es el vector de *scores* de concentración. Si se tiene solamente a **y**, entonces $\mathbf{u}_1$=**y**. Comenzando con la matriz de espectros **X**:

$$\mathbf{w}_1 = \frac{\mathbf{X}^T \mathbf{u}_1}{\left\| \mathbf{X}^T \mathbf{u}_1 \right\|} \quad (14)$$

donde las barras dobles significan el cálculo de la norma, y $\mathbf{w}_1$ es el primer vector de coeficientes de peso. Estos coeficientes son necesarios para mantener la ortogonalidad de los *scores*, aunque se aclara que la ortogonalidad absoluta entre factores puede no lograrse finalmente, ya que PLS puede sacrificarla en beneficio de la capacidad predictiva del modelo. Luego:

$$\mathbf{t}_1 = \mathbf{X}\,\mathbf{w}_1 \quad (15)$$

donde $\mathbf{t}_1$ es el primer vector de *scores* para **X**.

- Posteriormente, para los datos de composición, si éstos son multivariados:

$$\mathbf{q}_1 = \frac{\mathbf{Y}^T \mathbf{t}_1}{\left\| \mathbf{Y}^T \mathbf{t}_1 \right\|} \quad (16)$$

donde $\mathbf{q}_1$ representa el primer vector de *loadings* de composición. Luego:

$$\mathbf{u}_1 = \mathbf{Y}\,\mathbf{q}_1 \quad (17)$$

donde $\mathbf{u}_1$ representa el primer vector de *scores* de composición.

- En este punto es necesario realizar una comparación entre $\mathbf{t}_1$ y su correspondiente a la iteración anterior (si es que existe). El objetivo es evaluar si existe convergencia dentro de ciertos límites de error numérico. Si de la comparación se puede deducir que ambos vectores son iguales, entonces es necesario seguir con el cómputo en (19). Caso contrario, se debe regresar a (15) y usar a $\mathbf{u}_1$ proveniente de (18). Si los datos de composición son univariados, las ecuaciones (17) y (18) pueden ser omitidas, $\mathbf{q}_1$ toma el valor 1 y no se requerirán iteraciones.
- A continuación debe calcularse el primer vector de *loadings* ($\mathbf{p}_1$) para los datos en **X** y luego deben reescalarse, haciendo lo mismo con los *scores* y coeficientes de peso ya calculados (en las ecuaciones de escalado (20), (21) y (22), si una misma variable aparece a ambos

lados del signo igual, debe entenderse que primero se realiza el cálculo a la derecha con los valores actuales, y luego se almacenan los resultados en la variable presente a la izquierda, la cual adoptará su valor/es definitivo/s):

$$\mathbf{p}_1 = \frac{\mathbf{X}^T \mathbf{t}_1}{\left\| \mathbf{t}_1^T \mathbf{t}_1 \right\|} \quad (18)$$

$$\mathbf{p}_1 = \frac{\mathbf{p}_1}{\left\| \mathbf{p}_1 \right\|} \quad (19)$$

$$\mathbf{t}_1 = \mathbf{t}_1 \left\| \mathbf{p}_1 \right\| \quad (20)$$

$$\mathbf{w}_1 = \mathbf{w}_1 \left\| \mathbf{p}_1 \right\| \quad (21)$$

Lo siguiente consiste en hallar un coeficiente para la relación interna entre ambos tipos de *scores*:

$$b_1 = \frac{\mathbf{u}_1^T \mathbf{t}_1}{\mathbf{t}_1^T \mathbf{t}_1} \quad (22)$$

- Una vez realizados los cómputos para la primera LV, pueden calcularse los residuos para **X** e **Y**:

$$\mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1 \mathbf{p}_1^T \quad (23)$$

$$\mathbf{F}_1 = \mathbf{Y} - b_1 \mathbf{t}_1 \mathbf{q}_1^T \quad (24)$$

- Si se requiere el cálculo de más LV, se repetirá el procedimiento entero, comenzando desde (15). En este caso, **X** e **Y** serán reemplazadas con sus matrices de residuos $\mathbf{E}_1$ y $\mathbf{F}_1$, respectivamente, y todos los subíndices serán aumentados en una unidad. A su vez, los vectores calculados irán siendo ensamblados en sendas matrices, las cuales serán finalmente utilizadas para obtener el modelo de regresión.

Dado que en este contexto PLS sería utilizado para hallar un vector de regresión **b** para satisfacer **y**=**Xb**, vale destacar que la inversa de **X** según PLS se obtendría de la siguiente manera:

$$\mathbf{X}^+ = \mathbf{W} \left(\mathbf{P}^T \mathbf{W} \right)^{-1} \left(\mathbf{T}^T \boldsymbol{T} \right)^{-1} \mathbf{T}^T \quad (25)$$

También vale mencionar que los *scores* y *loadings* calculados pueden ser pensados como aquellos utilizados en PCA, solamente que con determinada rotación para ser más relevantes en la

predicción de **y**. Al igual que PCR, PLS converge a MLR si todas las LV son incluidas (Wise y col., 2005).

Más allá de lo descripto a nivel algorítmico, en este trabajo PLS ha sido utilizado con fines de clasificación. En este contexto, algunas de las técnicas de reconocimiento de patrones puros están orientadas a la discriminación entre diferentes grupos de muestras y operan dividiendo el hiperespacio en tantas regiones como grupos existentes. Así, si una muestra se encuentra representada por una región del hiperespacio correspondiente a una categoría particular, entonces se considera que la muestra pertenece a dicha categoría. En estos casos, las muestras son asignadas solamente a un único grupo (Galtier y col., 2011). Si bien un discriminante que separa 2 clases puede ser calculado mediante regresión lineal, por otro lado, cuando las clases están descriptas por objetos multivariados y hay más variables que objetos, una regresión obtenida con PLS puede ser usada para modelar discriminantes para clasificación (Barker y Rayens, 2003; Ni y col., 2009). De esta forma, PLS utilizado para clasificación toma el nombre de PLS-DA. Esta herramienta de clasificación se diferencia de otras, como por ejemplo de SIMCA, en la cual el objetivo es encontrar variaciones internas de cada clase sin tener por meta encontrar direcciones en el hiperespacio que diferencien directamente a las clases. PLS-DA realiza lo anterior y en este sentido es similar a otra técnica, LDA (Análisis de Discriminantes Lineales), hasta el punto que se ha sugerido en (Barker y Rayens, 2003) que PLS-DA es esencialmente equivalente a LDA pero la solución es hallada mediante mínimos cuadrados inversos, por lo cual produce los mismos resultados pero con las ventajas de seleccionar variables y de reducir el efecto del ruido en las señales (Wise y col., 2005).

Si la clasificación es binaria, las membresía de clase para cada muestra de calibración es codificada en el vector **y**, con una cantidad de componentes igual al número de muestras de calibración, y con un valor de 1 para el caso de las muestras que pertenecen a la clase en cuestión y de 0 para las no pertenecientes. Estos valores suelen ser normalmente utilizados, aunque podrían cambiarse por otros, ya que simplemente cumplen fines nominales. Así pues, en la etapa de calibración se estima la relación entre la información espectral y los códigos de clase. Más allá de cuáles sean los valores codificantes, mediante una regresión del tipo PLS1 es posible obtener un modelo de discriminante simple y un valor umbral, el cual deberá ser superado en la predicción de una muestra para que ésta pueda ser considerada perteneciente a una clase (Ni y col., 2009; Arancibia y col., 2008). El valor umbral debe ser calculado debido a que sería muy poco probable que un modelo codificado con clases binarias predijera exactamente 1 ó 0 para muestras incógnita (incluso las muestras de calibración pueden obtener valores cercanos a 1 ó 0). El cálculo del umbral

por clase según la versión de PLS-DA utilizada se realiza mediante el teorema de Bayes y contemplando la información disponible con la finalidad de minimizar los errores de predicción. El umbral Bayesiano se calcula asumiendo que los valores predichos para las muestras de calibración siguen una distribución similar a la que tendrían las muestras incógnita posteriormente clasificadas. Utilizando las distribuciones estimadas, el umbral se selecciona en el punto donde las dos distribuciones se cruzan, y este es el valor para el cual el número de falsos positivos y negativos debería ser minimizado para predicciones futuras. A su vez, el cálculo asume que la distribución de las predicciones para cada clase es aproximadamente normal. Si existe un número pequeño de muestras representando a cualquiera de las clases (las pertenecientes o las no pertenecientes) el cálculo del umbral puede ser sesgado. A su vez, es posible obtener un umbral utilizando muestras extra destinadas a tal fin. Ya que el umbral determinará la sensibilidad y la especificidad de las clasificaciones, también es posible obtener su valor gráficamente a través del análisis de curvas tipo ROC (Curvas Receptor Operador), confrontando sensibilidad y especificidad en un mismo gráfico para distintos valores del umbral (Wise y col., 2005).

Cuando hay más de 2 clases presentes, la información de clase puede ser tratada igualmente como binaria y mediante PLS1 se puede computar un conjunto de discriminantes entre cada clase y todas las restantes. También es posible codificar el conjunto de identidades de clase en una matriz **Y**, donde cada columna representa la membresía de clase en términos binarios para las muestras representadas por cada fila de **Y**, y posteriormente utilizar PLS2 para producir un conjunto de discriminantes separando cada clase del resto (Ni y col., 2009). La codificación binaria para el caso de más de 2 clases suele utilizarse por sobre la simple asignación de un valor por clase debido a que esta última opción presenta algunos inconvenientes. Por ejemplo, si 3 clases fueran codificadas con los números 1, 2 y 3, esto implicaría indirectamente que la clase 2 se encontraría entre la 1 y la 3, lo cual rara vez estará relacionado con la realidad experimental (Wise y col., 2005).

2.4.3.1 Cifras de mérito para PLS-DA

Para seleccionar el número de LV para los modelos PLS-DA se utilizó Validación Cruzada (CV) con las muestras de calibración, dejando una muestra de lado en cada iteración (lo que comúnmente se denomina LOOCV, del inglés L*eave One Out Cross Validation*). Así, para cada número de LV, se realizó LOOCV y se obtuvieron las Raíces de los Errores Cuadráticos Medios para Validación Cruzada (RMSECV, del inglés *Root Mean Squares Error of Cross Validation*), cuya definición es:

$$RMSECV = \sqrt{\frac{\sum (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}{n}} \quad (26)$$

donde $\hat{\mathbf{y}}_i$ representa a las predicciones de las muestras dejadas de lado, $\mathbf{y}_i$ contiene a los valores de referencia o nominales de las mismas muestras y n es el número de muestras de calibración.

Por otro lado, una vez que los modelos fueron definidos y utilizados, la calidad de sus predicciones simplemente se evaluó relacionando el número de clasificaciones correctas para un determinado conjunto de muestras (calibración o validación) respecto del número total de muestras en dicho conjunto.

2.5 Materiales y Métodos

El autor del presente escrito participó solamente de las etapas posteriores a la obtención de los datos mediante HPLC-MS. No obstante, se exponen los detalles experimentales que dieron origen a todos los datos.

2.5.1 Reactivos y Solventes

Acetonitrilo (ACN) grado HPLC fue obtenido de J.T.Baker (Holanda). Ácido acético glacial (AcOH, 99.7%) fue obtenido de Panreac (España). Sulfato de Magnesio anhidro ($MgSO_4$) y Acetato de Sodio Tri-hidratado ($NaAc\cdot 3H_2O$) fueron obtenidos de Merck (Alemania). El agua ultrapurificada fue obtenida de un sistema de purificación de agua Milli-Q marca Millipore (Bedford, MA). Las fases móviles fueron sometidas a filtración con 0.45 µm de Acetato de Celulosa en el caso del agua y Politetrafluoretileno (PTFE) para los solventes orgánicos, a la vez que fueron desgasificadas con Helio antes de su uso. Todos los extractos fueron filtrados a través de una membrana Millipore de Acetato de Celulosa de 0.45 µm antes de ser bombeados en el sistema cromatográfico. Finalmente una suspensión concentrada de Botrán 20 (Carbofurano aproximadamente 20% p/v) fue obtenida de Tragusa (Sevilla, España) y utilizada para aplicar el tratamiento a los frutos.

2.5.2 Instrumentos y Programas

La separación por HPLC fue llevada a cabo con un sistema Hewlett-Packard (HP) Serie 1100 (Wilmington, DE) y operada mediante el software HP ChemStation con control de MS y

procesamiento espectral. El HPLC contó con una bomba modelo G 1311 y con una válvula de inyección de 6 puertos modelo 7725i con 20 µl de loop. La separación analítica se llevó a cabo con una columna (150 mm × 4.6 mm) Agilent Zorbax EclipseXDB C_8, para un tamaño de partícula de 5 µm. Para la detección de los componentes objetivo a partir de la separación cromatográfica se utilizó una plataforma HP G1948A para espectrometría de Masa con simple cuadrupolo y con interfaz ESI.

También se utilizaron una trituradora Sammic S.L. (Azpeitia, España) de 0.5 kW de potencia máxima, un politrón PT1035 de Kinematica AG (Suiza) y un evaporador rotatorio (R-114) con un baño termostatizado de agua (B-480) obtenido en Buchi (Flawil, Suiza). Durante los pasos de extracción se usó una centrífuga Sigma 4-15 con un rotor incorporado Sigma 11150. Las muestras trituradas y homogeneizadas fueron almacenadas en un freezer de baja temperatura (-84 ºC).

Matlab (MATLAB *7.6.0*, 2008) fue utilizado como plataforma de desarrollo, proveyendo conjuntos de funciones (*toolboxes*) apropiadas para el trabajo con Wavelets (*Wavelet Toolbox*) y para estadística en general *(Statistics Toolbox)*. Las clasificaciones con PLS-DA fueron hechas con PLS *Toolbox* 3.52 (Wise y col., 2005) para Matlab. MCR-ALS se utilizó a partir de una interfaz gráfica que adicionalmente provee información detallada acerca de la implementación de este algoritmo (Jaumot y col., 2005), aunque para la realización de algunos cálculos se utilizaron las funciones de ALS directamente en línea de comando o mediante rutinas escritas en Matlab.

2.5.3 Plantación y tratamiento con pesticida

Los cultivares de tomate Rambo, RAF y Zayno fueron plantados en un invernadero de 1 Ha dentro de la granja experimental “Universidad de Almería-ANECOOP”.

Los cultivares fueron sembrados con un diseño tendiente a la uniformidad y a la simetría con el objetivo de minimizar posibles errores debidos a la distribución espacial de las plantas. Un esquema de la plantación puede observarse en la figura 9. En dicho esquema puede observarse que cada color (representando un tipo de cultivar) abarca 2 líneas paralelas verticales, lo cual alude a que las filas de plantas fueron realizadas de a pares. A su vez, el corredor central y las áreas de seguridad (ambas con fondo blanco) cumplieron la función de evitar que las muestras blanco fueran contaminadas con pesticida cuando este era utilizado para tratar a las plantas objetivo. Las zonas con fondo gris, para blancos y muestras tratadas, esquematizan la procedencia real de todas las muestras utilizadas. Las condiciones promedio del invernadero fueron de 20.1 ºC para temperatura y de 74.4% para humedad. Todas las plantas recibieron un tratamiento horticultural de rutina. Las plantas tratadas

fueron sometidas a dosis recomendadas (4 litros/ha) de Botrán 20, incorporando al insecticida directamente en el agua de irrigación. En este trabajo sólo se reportan resultados obtenidos con las variedades Rambo y RAF.

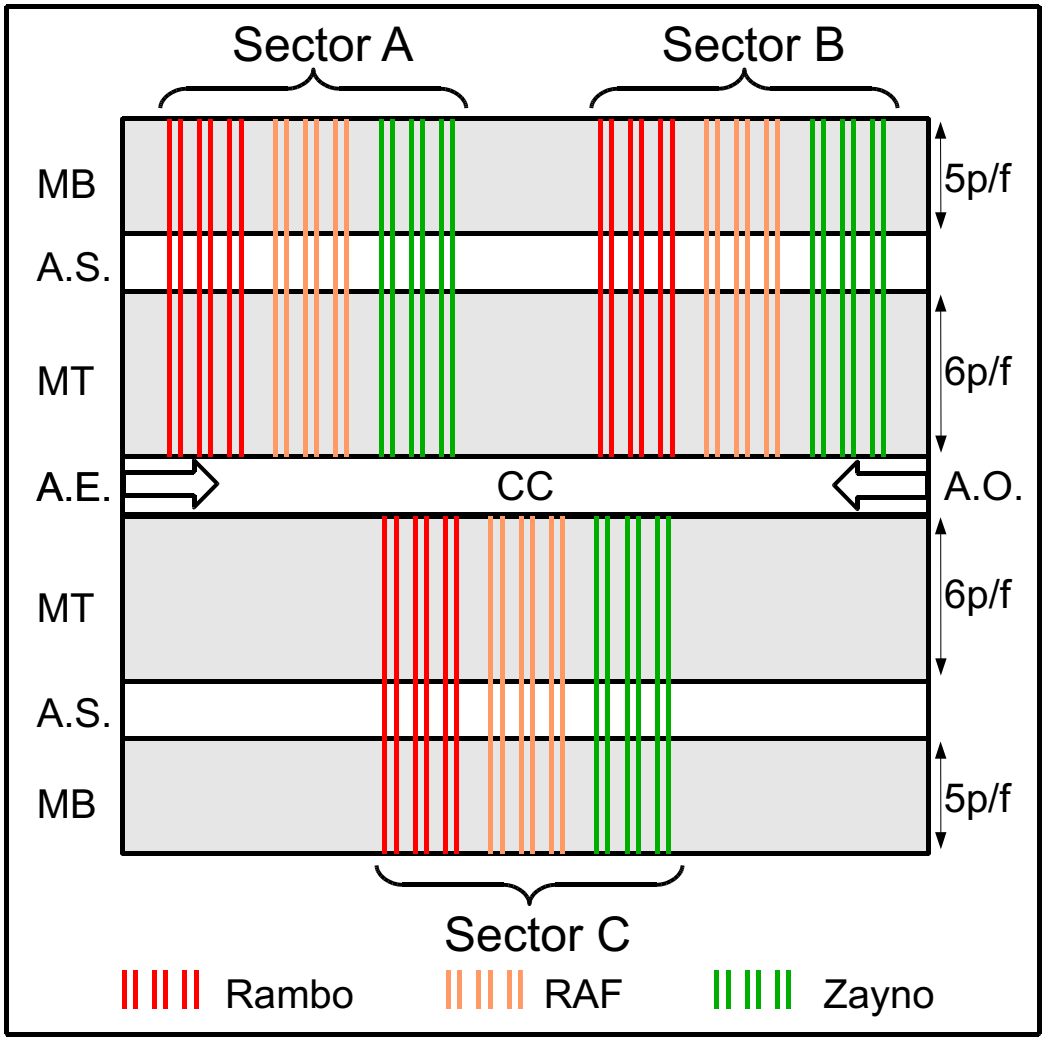


Figura 9: Esquema de la plantación desde la cual se obtuvieron los frutos

Referencias: MB: plantas sin tratamiento con pesticida, MT: plantas con tratamiento con pesticida, A.S.: Área de Seguridad, A.E.: Acceso Este, A.O.: Acceso Oeste, C.C.: Corredor Central, p/f: plantas/fila

2.5.4 Procedimiento de muestreo y almacenamiento

El muestreo fue realizado 8 veces, a través de días no consecutivos, durante los 21 días posteriores al tratamiento con Carbofurano. Independientemente del cultivar, del sector y del tratamiento, el procedimiento de muestreo implicó la selección de una planta y la recolección de 3 frutos (parte inferior, media y superior), los cuales fueron guardados en bolsas de polietileno con su

correspondiente rótulo y trasportados lo más rápido posible al laboratorio donde se realizaron los análisis. Debe entenderse que por cada muestra tratada se obtuvo simultáneamente una muestra equivalente del mismo cultivar y sector pero sin tratamiento.

Cada muestra fue procesada para dar una muestra en términos analíticos. Para esto, los 3 frutos de cada planta fueron cortados, triturados, mezclados vigorosamente y homogeneizados en un politrón. Finalmente, 200 g de cada mezcla fueron seleccionados como partes representativas que luego fueron conservadas en freezer a -84 ºC hasta el momento de ser analizadas.

Para diferenciar a las muestras, se adoptó una nomenclatura del tipo Cultivar-Sector-Día de Muestro-Tratamiento. Cultivar fue reemplazado con R para Rambo y F para RAF. Para Sector se utilizaron las letras A, B y C, mientras que para día de muestreo se usaron enteros del 1 al 8. Para tratamiento se agregó una “b” para simbolizar “Blanco” y nada en caso de que la muestra hubiese sido tratada con pesticida. Dado que se utilizaron 2 cultivares en 3 sectores durante 8 recolecciones tanto para blancos como para muestras tratadas, el conjunto final de muestras fue de 96. La tabla 1 ejemplifica los tiempos de recolección y la nomenclatura derivada de las muestras Rambo obtenidas durante el período de muestreo desde el sector A.

O.M.	D.P.T.	MT	MB
1	1	RA1	RA1b
2	3	RA2	RA2b
3	7	RA3	RA1b
4	9	RA4	RA2b
5	11	RA5	RA1b
6	14	RA6	RA2b
7	18	RA7	RA1b
8	21	RA8	RA2b

Tabla 1: Tiempos de recolección de frutos Rambo desde el sector A y nomenclatura derivada

Referencias: O.M.: Orden de Muestreo, D.P.T.: Días Post Tratamiento, MT: Muestra Tratada, MB: Muestra Blanco

La tabla 1 a su vez deja ver cuáles fueron las muestras utilizadas en una primera parte de este trabajo, donde sólo se obtuvieron datos de 16 de las 96 muestras disponibles. El total en sí fue utilizado en la segunda parte del trabajo, lo cual será detallado oportunamente.

2.5.5 Extracciones y preparación de las muestras para análisis

Los extractos fueron preparados usando el método QuEChERS (del inglés *Quick Easy Cheap Effective Rugged Safe*, página web oficial en [http://quechers.cvua-stuttgart.de/]) (Anastassiades y col., 2003; Lehotay y col., 2005). Los pasos de extracción fueron los siguientes:

1- Pesar 15 g de una muestra homogeneizada en un tubo de Teflón de 50 mL para centrífuga

2- Adicionar 15 mL de ACN acidificado con AcOH 1%

3- Adicionar 6 g de $MgSO_4$ y 2.5 g de $NaAc \cdot 3H_2O$

4- Agitar vigorosa y manualmente durante 3 minutos

5- Centrifugar el tubo a 3700 rpm por 5 minutos

Por otro lado, se llevó a cabo un paso de pre-concentración evaporando a sequedad alícuotas de 10 mL de sobrenadante en un evaporador rotatorio, las cuales fueron reconstituidas con 1 mL de ACN. Finalmente, los extractos fueron filtrados a través de las membranas Millipore ya descriptas antes de ser inyectados en el sistema cromatográfico.

En el método QuEChERS original existe un paso de extracción en fase sólida (SPE). Esto no fue realizado luego de la extracción con ACN debido a que esta última extracción resultó ser la mejor opción para el caso.

2.5.6 Análisis LC-ESI-MS

Durante los pasos de separación cromatográfica, además de la columna de separación (EclipseXDB C_8) se utilizó otra pre-columna Phenomenex C_8. Los análisis fueron realizados con un gradiente de solvente dado por el solvente A (ACN) y por el solvente B (Formiato de Amonio 50 mM acidificado con Acido Fórmico hasta pH 3.5). El programa de gradiente se inició con 3 minutos con 75% B y continuó con 15 minutos de gradiente lineal con 40% B, 7 minutos de gradiente lineal con 100% A, 1 minuto con 100% A, y finalmente 4 minutos con un gradiente lineal hacia las condiciones iniciales (75% B), más 1 minuto con 75% B. La fase móvil fue ajustada para tener un flujo de 1 mL/min. La temperatura de la columna fue fijada en 25 ºC y el volumen de inyección fue de 20 µL. La desolvatación fue optimizada para obtener la respuesta analítica más alta para Carbofurano. La fuente de temperatura durante la desolvatación se fijó en 325 ºC y los fragmentos iónicos fueron generados usando Nitrógeno de alta pureza como gas de secado a un flujo de 9 L/min. El voltaje de fragmentación fue de 60V. Los espectros fueron obtenidos en modo de adquisición *full scan* en un rango de m/z de 50-750 amu.

2.5.7 Datos obtenidos: tratamientos generales

Cada muestra corrida en los análisis LC-ESI-MS dio como resultado una matriz de dimensiones 507 × 2951, donde el primer valor indica tiempos de retención para los cromatogramas y el segundo se relaciona con la cantidad de variables de m/z en cada espectro. Los datos fueron obtenidos desde el software HP ChemStation en formato cdf y luego fueron convertidos a formato ASCII para ser procesados en Matlab. Habiendo notado una importante presencia de ceros en las matrices, cada matriz fue reducida eliminando todas las columnas entre la 711 y la 2951 (incluyendo a ambas). Por ende, cada muestra quedó representada por una matriz cuyas dimensiones fueron 507 × 710. Vale destacar que viendo que muchas variables de m/z registradas no representaron información relevante y fueron eliminadas, es criticable el hecho de no haber reducido el intervalo durante la adquisición de datos. Si se hubiera hecho, seguramente el detector hubiese brindado información más valiosa sobre los metabolitos.

A excepción de una parte del trabajo en la cual se estudió el efecto de la WT y donde se utilizaron las matrices de 507 × 710, en el resto de las experiencias se trabajó con las matrices reducidas mediante DWT bidimensional (DWT2) en 2 niveles con la Wavelet de Haar, obteniendo matrices con un tamaño aproximado a ¼ del tamaño original, específicamente de 127 × 178, bajo el supuesto de que la información más relevante no se perdería en la reducción. Esto fue realizado con el objetivo de reducir las demandas computacionales, disminuyendo el tiempo de los cómputos y facilitando la aplicación de las herramientas quimiométricas utilizadas. A su vez, la compresión aquí practicada está inherentemente asociada con una reducción del ruido en las señales, ya que se asume que los coeficientes de detalle representan, entre otras cosas, a la componente de ruido en general (Walczak y Massart, 1997b). Por otro lado, aunque también con el objetivo de facilitar los cómputos y superar limitaciones de *hardware*, las matrices reducidas fueron divididas en 4 regiones (A, B, C y D), conservando todas las columnas y dividiendo las filas según el esquema de la figura 10, el cual ejemplifica la primera muestra tomada de los tomates Rambo tratados del sector A (RA1) reducida mediante DWT2. Para conservar la nomenclatura propuesta, en cada caso en que fuera necesario señalar una determinada región de una matriz ya nomenclada, se agregó el sufijo correspondiente a la región en cuestión. De esta forma, cada vez que se ejecutó MCR-ALS con matrices aumentadas, los apilamientos se produjeron para cada región de manera independiente y por ende, se obtuvo una resolución por región.

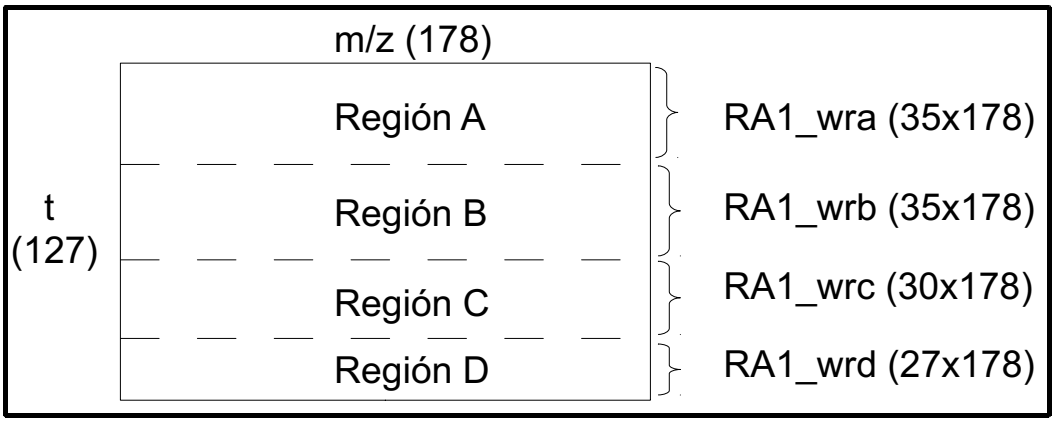


Figura 10: Esquema de la división en regiones de la matriz RA1 reducida con DWT2

Referencias: t: tiempo, m/z: relación masa/carga, wrx: versión de una matriz reducida mediante DWT2 para la región x (a,b,c y d)

Antes de aplicar MCR-ALS, el número de componentes significativos y responsables de la generación de varianza fue estimado mediante SVD aplicado a las matrices aumentadas obtenidas por apilamiento de regiones equivalentes de las matrices reducidas (la identidad de estas matrices depende de cada experiencia y será revelada posteriormente). Finalmente se ejecutó MCR-ALS con las restricciones ya enunciadas sobre las matrices aumentadas, de lo cual se obtuvieron los perfiles espectrales y de concentraciones de los componentes individuales presentes en cada región.

2.5.8 Datos obtenidos: separación del estudio en partes

La descripción anterior corresponde a una estrategia general de la cual se obtuvieron resultados a partir de ciertos grupos de muestras y bajo determinados tratamientos, con lo que puede obtenerse una división del trabajo en partes. Antes de dar detalles sobre cada una, en la figura 11 se esquematiza la forma en que fueron procesadas las muestras, así como también los tratamientos generales (reducciones, divisiones, apilamientos, entre otros) aplicados a los datos, los distintos algoritmos ejecutados y cuáles de todos los datos disponibles dieron origen a cada una de las partes mencionadas:

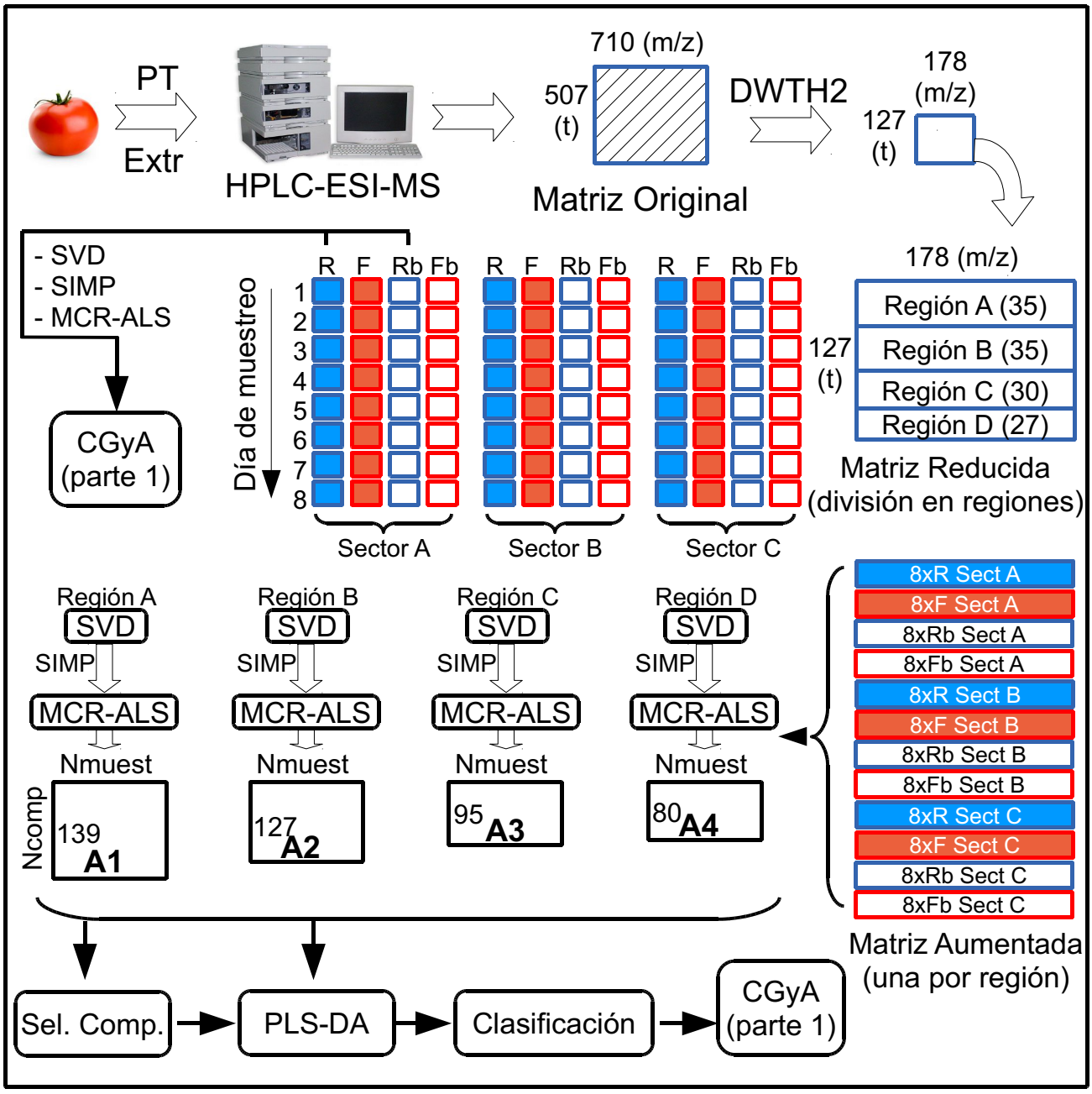


Figura 11: Esquema de trabajo

Referencias: PT: PreTratamientos, Extr: Extracción, DWTH2: Transformada Wavelet Discreta Bidimensional de Haar, SIMP: SIMPLISMA, Nmuest: Número de Muestras (96), Ncomp: Número de componentes resueltos, An: Áreas basadas en los perfiles de concentración (n desde 1 hasta 4), Sel. Comp.: Selección de Componentes, CGyA: Comparación Gráfica y Analítica.

El detalle de las partes en que fue dividido el trabajo es el siguiente:

Parte 1: En una primera instancia, se utilizaron solamente las muestras de tomates Rambo, tratadas y no tratadas durante los 8 días de muestreo, correspondientes sólo al sector A, contabilizando 16 muestras, todas ellas reducidas mediante DWT2. La aplicación de MCR-ALS a matrices aumentadas obtenidas apilando regiones equivalentes de las 16 matrices reducidas condujo a la obtención de los perfiles de concentración que, en conjunto con la información de los días de muestreo, permitieron realizar una comparación gráfica y analítica de las cinéticas observadas para los componentes modelados. Para este análisis se elaboraron rutinas para el cálculo de coeficientes de correlación de Pearson, con lo cual las cinéticas para determinados componentes a través de los 8 días de muestreo fueron cotejadas entre muestras tratadas y blancos. Estas comparaciones también se realizaron solapando parcialmente los días de muestreo a través de movimientos relativos de los perfiles evolutivos comparados, en busca de altas correlaciones, con el objeto de verificar si algunas vías metabólicas habían sido retardadas o aceleradas debido al tratamiento con pesticida.

Parte 2: En segundo lugar se utilizaron las muestras tratadas y no tratadas de los cultivares Rambo y RAF, correspondientes a todos los sectores, contabilizando 96 muestras representadas por sus respectivas matrices reducidas con DWT2. Las regiones equivalentes de todas las muestras fueron apiladas, conformando 4 matrices aumentadas. Se estimaron los componentes mediante SVD y se obtuvieron los perfiles resueltos con MCR-ALS. Luego se utilizaron los perfiles de concentración en modelos de clasificación PLS-DA, con el objetivo de diferenciar cultivares y presencia/ausencia de tratamiento. Para esto, de cada perfil de concentración se tomó el área resuelta bajo el perfil y estas áreas para los perfiles de todos los componentes y de todas las muestras fueron tabuladas en una matriz común de dimensiones Ncomp × Nmuest, donde Nmuest representa la cantidad de muestras disponibles (96) y Ncomp equivale a la suma de los componentes estimados por SVD y resueltos entre todas las regiones. En este sentido, cada matriz de datos quedó representada por un vector vertical de áreas resueltas.

Debe tenerse en cuenta que la clasificación en grupos de muestras basada en los perfiles metabólicos podría no ser una tarea sencilla debido al bajo número de muestras en comparación con el alto número de variables (componentes modelados). El gran número de picos en estas muestras representa desafíos de modelado y validación, ya que todos los componentes resueltos son potenciales biomarcadores. A su vez, el número de muestras necesarias para describir con exactitud este tipo de problemas clasificatorios se incrementa exponencialmente con el número de variables modeladas y, en general, el número de muestras para este tipo de aplicaciones suele ser mucho

menor que el de variables. Aunque PLS-DA es uno de los métodos de análisis usados en este tipo de casos, desafortunadamente los modelos suelen sobreajustarse a los datos de calibración (Westerhuis y col., 2008). Con lo anterior en mente, se puso a prueba un procedimiento simple a través del cual es posible seleccionar componentes en un intento de reducir el número de variables activas durante el modelado. Dadas las áreas resueltas en MCR-ALS, se selecciona un grupo de muestras que actuará como conjunto de calibración, por lo cual sus clases deben ser conocidas, dejando al resto de las muestras como validadores. Luego se evalúa si para cada componente (variable) existe igualdad, en términos estadísticos, de la media de las áreas entre muestras de clases distintas. Para esto, se aplica el test de Levene (más apropiado que el de Bartlett cuando la distribución de las muestras no es normal y cuando pueden existir datos anómalos (MATLAB *7.6.0*, 2008)) y se verifica si las varianzas de los grupos comparados pueden ser consideradas equivalentes o no. Posteriormente se realiza un test t-Student tomando en cuenta lo anterior, es decir, con o sin igualdad de varianzas, para un nivel de confianza que en este trabajo fue 95%. Finalmente, todos los componentes para los cuales la hipótesis de igualdad en la media de las áreas no puede ser rechazada son descartados, siendo el resto conservados. Ya que en esta parte del trabajo existieron 4 clases distintas (R, F, Rb y Fb), para la selección de variables se realizaron todas las comparaciones posibles entre 2 clases (6 combinaciones), es decir, uno contra uno. Hecho lo anterior, todos los componentes seleccionados en todas las evaluaciones fueron agrupados y, en el caso de componentes repetidos, sólo una copia fue introducida en la selección definitiva.

En resumen, los modelos PLS-DA fueron hechos a partir de las áreas de los componentes resueltos con y sin selección de estos. En ambos casos las matrices **X** e **y** fueron centradas a sus respectivas medias, siendo este paso el único preprocesamiento aplicado a los datos antes del modelado. Más allá del hecho de que algunos problemas pueden surgir durante la optimización de modelos usando Validación Cruzada (Westerhuis y col., 2008; Anderssen y col., 2006) y teniendo en cuenta que los objetivos principales de este trabajo no están relacionados a un estudio exhaustivo de estos temas, el número de variables latentes seleccionadas para cada modelo fue obtenido mediante LOOCV partiendo del conjunto de calibración. En general, este número coincidió con el primer mínimo encontrado al graficar RMSECV vs LV. Debe notarse que al aplicar LOOCV a modelos hechos para clasificaciones no binarias (usando el algoritmo PLS2, en los cuales una única cantidad de LV debe ser seleccionada para todas las clases modeladas) puede existir desacuerdo en términos de qué número de LV debería ser considerado óptimo, ya que el mínimo error de CV para cada clase puede no ser el mismo para todas. En estos casos, la cantidad de LV seleccionadas fue

coincidente con aquella que produjera el menor error para la clase peor predicha, aunque para el resto de las clases este número no hubiera sido la mejor opción. Se decidió lo anterior en un intento de mantener a todas las clases más o menos contempladas por los modelos.

Finalmente, dado que los modelos PLS-DA fueron hechos con PLS Toolbox 3.52, cada muestra obtuvo una predicción para todas las clases puestas en juego en cada modelo, así como también una probabilidad asociada a dicha predicción. Para establecer definitivamente la clase de una muestra, esta fue asignada a la clase para la cual su predicción superó el umbral de dicha clase. Cuando las predicciones superaron más de un umbral de clase, las muestras fueron asignadas a la clase con mayor probabilidad asociada. Los umbrales y probabilidades se estiman usando el teorema de Bayes y la información de calibración disponible. Más información acerca de cómo son calculados estos valores puede ser encontrada en (Wise y col., 2005). Los resultados de todos los modelos fueron analizados a través de cifras de mérito y de gráficas. Como en la primera parte, también se realizaron comparaciones gráficas y analíticas de los perfiles evolutivos de los metabolitos, focalizando el análisis en cinéticas de componentes que podrían considerarse biomarcadores según los resultados de los modelos de clasificación.

2.6 Resultados y Discusión

Antes de evaluar los resultados de las distintas experiencias realizadas, es necesario realizar una aclaración. El tratamiento con MCR-ALS que se le dio a los datos permitió descomponerlos en sus contribuciones puras o en aproximaciones de éstas. Con lo anterior se hace posible representar la evolución de los metabolitos endógenos de tomate (y de los derivados por el tratamiento con Carbofurano) a través del período de muestreo y, con esta información, pueden postularse modelos clasificatorios que indiquen si una muestra fue o no tratada, por citar un ejemplo. De estas mismas resoluciones podría obtenerse información del tipo espectral, lo cual conduciría a la identificación de los componentes en cuestión siempre que sea posible realizar una comparación con bases de datos para metabolitos de tomate con información obtenida en las mismas condiciones experimentales que las aquí utilizadas, en especial en lo referente a la variante de espectrometría de Masa practicada. Lo que se quiere aclarar es que esto último no ha sido posible, ya que al momento de realizar estas labores no se encontraron las mencionadas bases de datos, y las que fueron encontradas contenían información de otros tipos de espectrometría de Masa.

2.6.1 Muestreo, extracciones y pre-concentraciones

La dosis recomendada para utilizar Carbofurano en frutos de tomate es de 4 L/ha con un intervalo previo a la cosecha de 45 días (Ministerio de Agricultura, Alimentación y Medio Ambiente de España, 2008). Lo anterior es un parámetro agronómico relacionado al tiempo mínimo que debe transcurrir entre la aplicación del pesticida y la cosecha. Luego de ese tiempo, se presume que la concentración de pesticida en los frutos será menor a su Límite Máximo de Residuos (MRL), el cual ha sido definido por la Comisión del Codex Alimentario de España (Food and Agriculture Organization of the United Nations, 2008) como la máxima concentración de residuos de pesticida, expresada en mg/kg, la cual es legalmente permisible para su uso en la superficie o dentro de alimentos para consumo humano y animal. Los MRL están basados en datos de Buenas Prácticas de Agricultura (GAPs) y su objetivo es asegurar que la comida derivada de productos de uso común, que cumplan con sus respectivos MRLs, sea aceptable desde un punto de vista toxicológico (Food and Agriculture Organization of the United Nations, 2008). Por otro lado, el muestreo fue llevado a cabo de acuerdo a un protocolo propuesto por la Unión Europea (Dirección General de Agricultura-Comisión de la Comunidad Europea, 2008). Si bien se muestreó en un período total de 21 días y no de 45, se supone que durante los primeros días post tratamiento debería ocurrir gran parte de las potenciales modificaciones metabólicas asociables a un stress fisiológico de este tipo.

La extracción de los componentes fue llevada a cabo con el método QuEChERS, el cual es ampliamente utilizado en la determinación de residuos de pesticidas a partir de muestras alimenticias. Este método, más allá de las ventajas que representan las características que componen su nombre (rápido, fácil, económico, efectivo, resistente y seguro), tiene un amplio espectro de capacidad para la extracción de compuestos de diferentes polaridades. Como ya se explicó, el método sugiere un paso de extracción en fase sólida que fue obviado en este trabajo. No obstante, el paso de pre-concentración llevado a cabo con la pre-columna Phenomenex C_8 permitió incrementar la sensibilidad, compensando parcialmente la pérdida asociada al modo de adquisición utilizado.

2.6.2 Tratamiento de datos: Reducción mediante DWT2 con Wavelet de Haar

A menudo, los datos obtenidos mediante LC-MS en modo *full scan* son de procesamiento dificultoso debido a la gran cantidad de información colectada. Por esta razón la DWT2 se muestra como una técnica potencial para la compresión de los datos. La elección de esta técnica y del filtro de Haar se basó a su vez en las ventajas expuestas en la sección de Teoría.

Inicialmente, se realizaron algunas experiencias con el objetivo de determinar si la compresión de datos mediante WT bidimensional con la wavelet de Haar (WTH2) era apropiada para este estudio. Para esto, se decidió trabajar únicamente con la muestra 1 del Sector A de tomates Rambo tratados con Carbfurano, denominada RA1, con dimensiones 507 × 701 (tiempos × m/z). Esta muestra, mediante WTH2 a 2 niveles de escalado, dio origen a RA1_wr, con dimensiones 127 × 178. El nivel de escalado elegido implicó un compromiso entre compresión y resolución, de forma que los cómputos pudieran realizarse a una velocidad admisible en términos prácticos sin que por esto se perdiera demasiada información. No obstante, la reducción del tamaño no se mostró suficiente para superar limitaciones relacionadas a las capacidades de cómputo disponibles, por lo cual se aplicó la estrategia de dividir el dominio de tiempo en 4 regiones. Para esto, la matriz original y la reducida fueron subdivididas en 4 regiones, denominadas A, B, C y D, conservando el número de columnas y restringiendo el número de filas a 140 (RA1a), 140 (RA1b), 120 (RA1c) y 107 (RA1d) para los datos originales, y a 35 (RA1_wra), 35 (RA1_wrb), 30 (RA1_wrc) y 27 (RA1_wrd) para los datos reducidos. Esta misma división por regiones también se utilizó en experiencias cuyos resultados serán presentados posteriormente.

Como puede apreciarse, la reducción de dimensiones con WTH2 se acerca al 25%, correspondiéndose con el nivel de escalado 2, y las pequeñas diferencias provienen de la forma en que el algoritmo de DWT es implementado cuando los datos de partida no tienen dimensiones que se correspondan con potencias de 2. Esto último no es indispensable en realidad, puesto que si al menos se parte de dimensiones pares, contrario al caso presentado, no se llegará a una situación de desproporcionalidad entre datos originales y reducidos, siempre y cuando en los sucesivos escalados no se produzcan dimensiones impares.

Las regiones de ambas matrices fueron sometidas al mismo proceso, a saber:

- Cálculo del número de componentes (ncomp) que podrían explicar un 90% de la varianza observada, mediante SVD.
- Obtención de estimaciones espectrales iniciales para los ncomp componentes posibles, mediante SIMPLISMA.
- Resolución mediante MCR-ALS con restricciones:
  - No-negatividad en espectros y concentraciones de todos los componentes
  - Unimodalidad para concentraciones de todos los componentes
  - Normalización de los espectros puros

- Criterio de convergencia del 0,1% en la diferencia de la desviación estándar de los residuos entre iteraciones sucesivas
- Cantidad de iteraciones en 300 permitidas como máximo.

Otros procedimientos fueron realizados con estas matrices y sus derivados. Todos los pasos realizados se esquematizan en la figura 12 para la región "A" de la matriz RA1, aunque el tratamiento fue el mismo para las otras regiones.

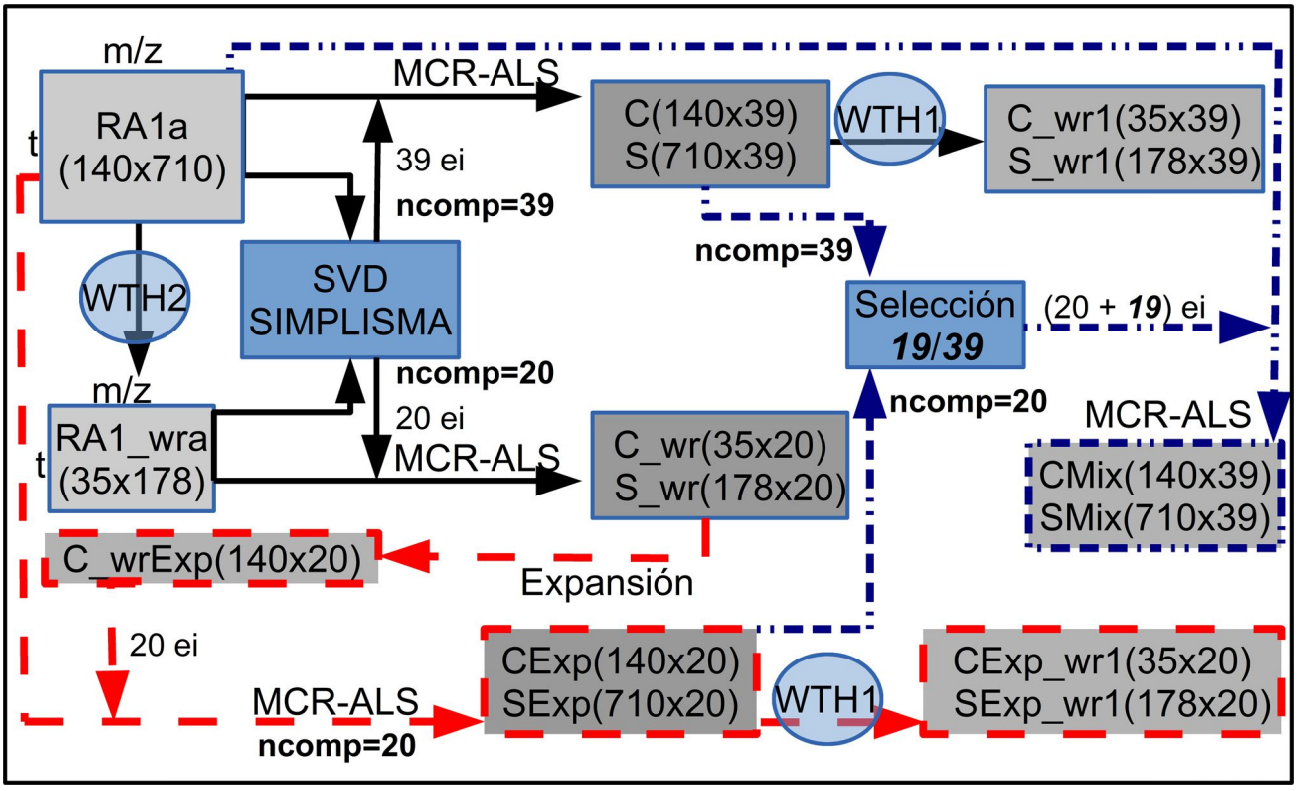


Figura 12: Esquema de experiencias realizadas para evaluar el desempeño de la WT

Referencias: WTH2: WTH bidimensional, WTH1: WTH unidimensional por columnas, ncomp: número de componentes, ei: Estimaciones Iniciales, Expansión: Expansión de columnas de C_wr mediante interpolación cúbica por partes, Selección 19/39: Selección de 19 componentes de los 39 disponibles en base a su comparación con CExp

La resolución multivariada de **RA1a** mediante MCR-ALS dio como resultado a la matriz de perfiles cromatográficos **C** y a la matriz de perfiles espectrales **S**, ambas para 39 componentes, y de manera similar, a partir de **RA1_wra** se obtuvieron **C_wr** y **S_wr** para 20 componentes. Lo mismo fue realizado para las matrices de las regiones B, C y D, cada una con su respectivo número de

componentes.

Con el objeto de examinar el efecto de otras aproximaciones iniciales en la resolución de **RA1a** y teniendo resuelta a **C_wr**, las columnas de la última matriz fueron expandidas hasta alcanzar un número de dimensiones compatible con **RA1a**, es decir, desde 35 hacia 140 tiempos. Para realizar estas expansiones, se utilizó interpolación cúbica por partes a través de la función "pchip" de Matlab. Vale destacar que este procedimiento podría haberse realizado con otros algoritmos, y en este contexto, la IWT merecería atención. No obstante, la IWT no fue utilizada por algunas razones:

- La WTH2 que dio origen a **RA1_wra** fue aplicada en 2 dimensiones. Esto significa que los coeficientes calculados provienen de valores puntuales y de valores en la vecindad, tanto a nivel de columnas como de filas. En cambio, las columnas de **C_wr** representan perfiles cromatográficos que son específicos para cada componente y no existe relación entre una columna y otra.
- MCR-ALS fue aplicado con restricciones y entre éstas los espectros fueron normalizados. Esto significa que los valores de los perfiles cromatográficos han sido escalados para cumplir con $\mathbf{D}=\mathbf{C}\mathbf{S}^{T}$. Por lo tanto, aunque estos perfiles derivan de coeficientes de aproximación de la WT, la escala de los coeficientes y de los perfiles resueltos no tiene por qué ser la misma. Por ende, si se aplicase IWT existirían inconvenientes de escalado, aunque serían salvables si éstos perfiles sólo fueran utilizados como aproximaciones iniciales.
- La aplicación de IWT para la región D resultaría en 108 tiempos y no en los 107 originales. Esto se debe a que al comprimir 107 tiempos mediante WTH2, en el primer nivel de escalado se producirían 54 tiempos, ante la imposibilidad de generar 53,5 tiempos. Esta es una limitación algorítmica y se produce para otro tipo de Wavelets además de la de Haar. Como resultado, de la IWT pueden regenerarse los datos transformados con mayor cantidad de dimensiones que las originales. La solución normalmente consiste en verificar el mayor nivel de alineamiento de un resultado respecto de algún patrón y seleccionar sólo las variables más alineadas.

Por las razones anteriores, se utilizó interpolación cúbica por partes para las expansiones. Debe entenderse, además, que el número de estimaciones iniciales proviene de la resolución de **RA1_wra**, y que este número (20) es menor al calculado para la matriz **RA1a** (39). En definitiva, MCR-ALS fue aplicado a **RA1a** con 20 perfiles de concentración como aproximaciones iniciales,

dando origen a **CExp** y **SExp**.

Ya que la experiencia anterior contemplaba para **RA1a** un número de componentes menor al obtenido por SVD, se realizó una nueva experiencia. En este caso, los 20 perfiles de concentración obtenidos en **CExp** fueron agrupados con 19 de los 39 obtenidos en **C**. Para seleccionar a los 19 complementarios, los 20 perfiles de **CExp** fueron comparados con los 39 de **C**, obteniendo un coeficiente de correlación de Pearson en cada comparación. Una vez que para cada perfil de **CExp** se hubiera seleccionado uno de **C** como el más correlacionado, estos últimos serían obviados para formar el grupo complementario. Como resultado, los 19 seleccionados fueron los perfiles con menores correlaciones. Con el grupo de 39 aproximaciones iniciales conformado, 20 de ellos provenientes de **CExp** y 19 desde **C**, se ejecutó nuevamente MCR-ALS, obteniendo las matrices **CMix** y **Smix**. Vale destacar que si durante la evaluación de correlaciones 2 ó más componentes de **CExp** hubieran seleccionado al mismo componente en **C** como el más apropiado, aquel que tuviera el menor valor de correlación sería agrupado con otro componente de **C**, el siguiente en orden de correlación.

También se aplicó WTH1 para reducir las dimensiones de las columnas de **C** y de **S**, dando origen a **C_wr1** y a **S_wr1**. Similarmente, se comprimieron las columnas de **CExp** y **SExp**, con lo cual se obtuvieron **CExp_wr1** y **SExp_wr1**. Estas reducciones se realizaron para poder compararlas con las columnas de **C_wr** y de **S_wr**. Si bien se explicó anteriormente por qué no usar la IWT para expandir los resultados, entre otras cuestiones por posibles disparidades en las escalas, el uso de la WTH1 para comprimir los vectores resueltos no presenta inconvenientes, ya que aunque las escalas sean diferentes, el cálculo de correlaciones no debería verse afectado.

Resultados provenientes de las experiencias anteriormente descriptas para la región A, así como también algunos valores derivados, fueron además obtenidos para el resto de las regiones y se resumen en la tabla 2. En ésta se puede observar, para todas las regiones, que el número de componentes estimados en la matriz original decae aproximadamente a la mitad luego de reducida la matriz con WTH2. Intuitivamente, lo anterior podría esperarse de una reducción a la mitad en las dimensiones, pero sin embargo dicha reducción ha sido a un cuarto de las dimensiones originales. Por lo tanto, podría decirse que la reducción de dimensiones a través de la WTH2 es mayor que la pérdida de información que podría esperarse. En relación a esto, no debe olvidarse que la WT también tiene funciones de eliminación de ruido en las señales y, por ende, algunos componentes en las matrices originales pudieron haber estado modelando este tipo de interferencias. De todas formas, la estimación del número de componentes mediante SVD corresponde a una simplificación

matemática que, por ejemplo, no contempla que las fuentes de varianza puedan ser no lineales. Por lo tanto, no debe pensarse que cuando se produce una reducción en el número estimado de componentes, éstos mismos (junto a otros más) hubiesen pertenecido al grupo mayor del cual se obtuvo la reducción.

	Región A				**Región B**			
Matriz	RA1	RA1_wr	RA1-Exp	RA1-Mix	RA1	RA1_wr	RA1-Exp	RA1-Mix
ncomp	39	20	20	39	32	16	16	32
%LOF EXP	9.132	15.665	19.673	8.976	21.107	20.900	24.066	20.785
$\%R^2$	99.166	97.546	96.130	99.194	95.545	95.632	94.208	95.680
Dato Medio (A)	51.249	204.421	51.249	51.249	72.039	287.346	72.039	72.039
Media |res| (B)	14.314	50.730	24.789	14.241	19.742	70.298	29.810	19.854
B/A	0.279	0.248	0.484	0.278	0.274	0.245	0.414	0.276
iter	44	5	51	17	14	5	5	7
	Región C				**Región D**			
Matriz	RA1	RA1_wr	RA1-Exp	RA1-Mix	RA1	RA1_wr	RA1-Exp	RA1-Mix
ncomp	26	14	14	26	25	11	11	25
%LOF EXP	17.638	21.214	26.504	16.752	11.705	22.255	19.306	11.133
$\%R^2$	96.889	95.500	92.976	97.194	98.630	95.047	96.273	98.761
Dato Medio (A)	88.021	351.096	88.021	88.021	140.664	555.879	140.664	140.664
Media |res| (B)	23.266	100.178	38.767	23.027	30.552	127.558	54.223	30.489
B/A	0.264	0.285	0.440	0.262	0.217	0.229	0.385	0.217
iter	8	6	5	4	42	2	3	3

Tabla 2: Detalles y cifras de mérito en la resolución mediante MCR-ALS de la muestra RA1 y de matrices derivadas (RA1_wr, RA1-Exp y RA1-Mix)

Referencias: ncomp: número de componentes en MCR-ALS, %LOF EXP: Porcentaje de Falta de Ajuste Experimental, $\%R^2$: Porcentaje de Varianza Explicada, Dato Medio (A): Valor medio de la matriz analizada, Media |res| (B): Valor medio de los valores absolutos de los residuos, B/A: cociente entre Dato Medio y Media |res|, iter: cantidad de iteraciones

El decaimiento de componentes significativos puede aclararse más con la figura 13, la cual permite comparar una sección de la matriz original con su respectiva reducción. Varios aspectos de esta figura serán discutidos en su momento, pero ahora debe notarse que la WT tiende a eliminar componentes minoritarios. Por ejemplo, en la gráfica superior existen pequeños conglomerados de baja intensidad pero diferenciables del nivel 0, como algunos observables entre los tiempos 1 y 16 (aproximadamente) para valores de m/z entre 64 y 128. Si se observa la gráfica inferior entre los

tiempos 1 y 4 para valores de m/z entre 16 y 32, estas zonas reducidas y equiparables a las anteriores ya no muestran con claridad a los conglomerados. Aunque los últimos sean levemente detectables a simple vista, es probable que en la ejecución de la SVD no sean considerados relevantes y por consiguiente el número de componentes estimados podrá disminuir.

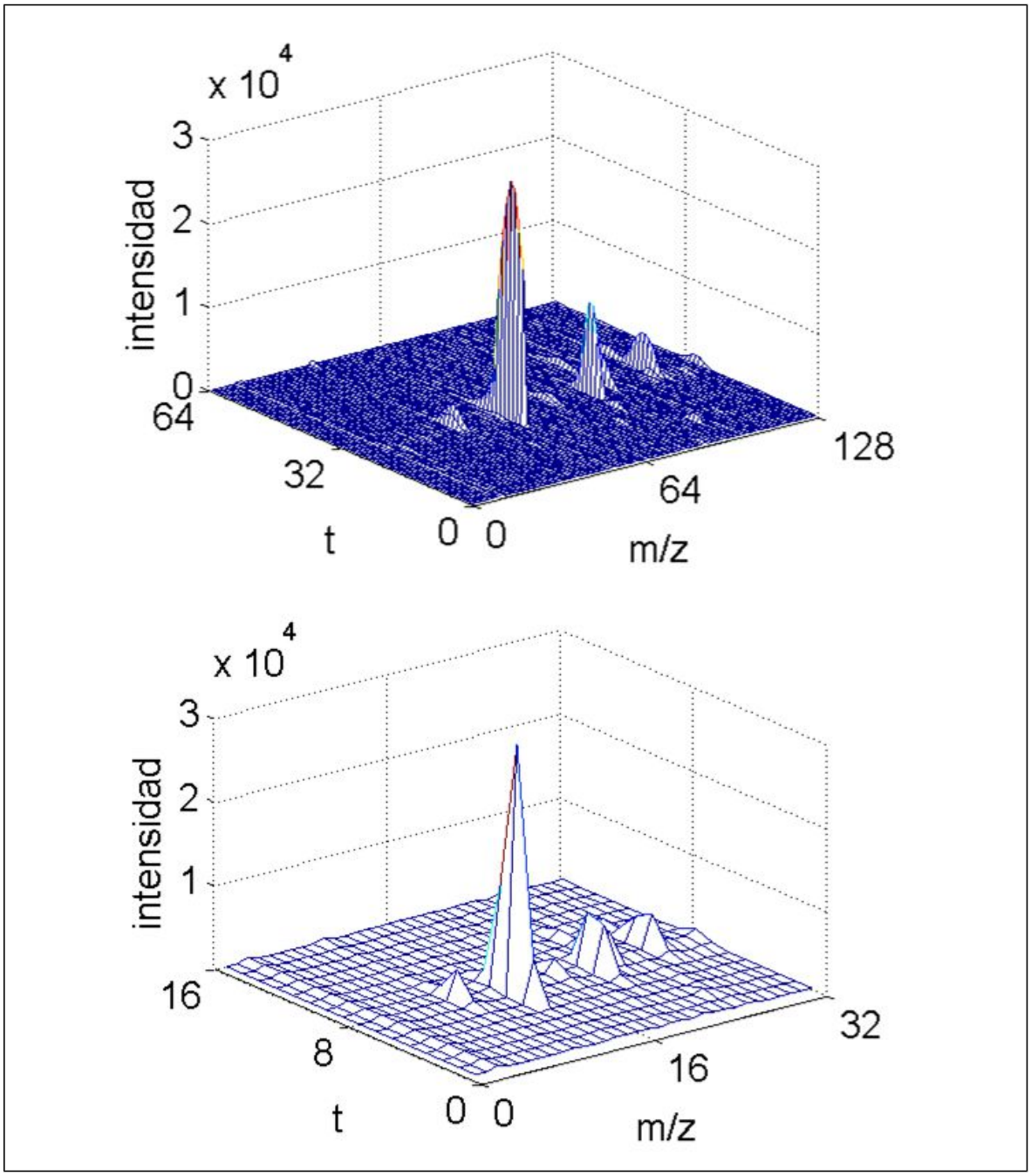


Figura 13: Sección de una matriz original con su correspondiente reducción en 2 escalas mediante WTH2

Referencias: Arriba: sección de RA1a, tiempos desde 1 hasta 64, m/z desde 1 hasta 128. Abajo: sección de RA1_wra, tiempos desde 1 hasta 16, m/z desde 1 hasta 32. Los valores de t y m/z son ordinales.

Otro aspecto a tener en cuenta desde la figura 13 es que el proceso de compresión, más allá de si los componentes eran o no minoritarios, produce un efecto de acercamiento entre los conglomerados. Así, componentes distintos pero cercanos en su tiempo de elución o con espectros con escasas diferencias serán condensados en pseudo-componentes al calcular los coeficientes de aproximación. Un ejemplo de esto puede observarse en la gráfica inferior, cerca del tiempo 8 y en valores de m/z de 20-24 aproximadamente. La altura máxima alcanzada allí es menor que la equivalente en la gráfica superior y esto es producto de la fusión con datos de menor valor en las cercanías. Todo esto también repercutirá en una disminución del número de componentes durante la SVD.

El hecho de que existan razones para pensar que la WT reduce la cantidad de información hasta el punto tal que definitivamente existirán algunos componentes no modelados o bien modelados de forma incorrecta, no es suficiente para descartar al método, pues aun en estos casos persistirán los componentes mayoritarios y en base a éstos es probable que puedan evaluarse los efectos de *stress* provocados con el Carbofurano.

Regresando al análisis de los resultados tabulados, los resultados de falta de ajuste y de varianza explicada mantienen siempre una relación. Al respecto, las regiones A y D, y en menor medida la C, muestran un aumento de falta de ajuste y una disminución de la varianza explicada al aplicar WT. La excepción puede verse en la región B, donde por ejemplo %LOF EXP es levemente menor para la matriz reducida. No obstante, la región B presenta los peores valores de %LOF EXP y de $\%R^2$ para la matriz sin reducir. De lo anterior, no puede deducirse con claridad el efecto de la WT sobre las cifras de mérito en cuestión, aunque de todas formas en el peor de los casos el porcentaje de varianza que no ha podido ser explicado no supera el 5%, cuando podría haberse elevado hasta 10% según el criterio utilizado en la SVD. Una explicación posible para interpretar las cifras obtenidas con las matrices reducidas, a excepción de las de la región B, podría estar relacionada con la subestimación de componentes producto de la aplicación de la WT. Se observó con anterioridad que pequeños conglomerados en las matrices originales eran prácticamente eliminados después de la reducción. Aunque a nivel visual eran levemente notables, se postuló que probablemente no serían tomados en cuenta como componentes relevantes en la SVD. No obstante, aunque mínimos y quizá no modelados, estos valores persistieron y por ende representaron parte de la varianza total a modelar. Habiendo menos componentes para el modelado, es probable que se hayan destinado a zonas de las matriz con intensidades bien apreciables. Por consiguiente, las zonas donde pudieron encontrarse los valores mínimos persistentes quizá no recibieron el soporte de ninguno de los

componentes modelados. Dicho en otras palabras, desde el punto de vista del modelo esas zonas deberían haberse resuelto en niveles de intensidad nulos por no tener componentes asociados, por lo que al comparar lo resuelto con los datos experimentales, los mínimos persistentes, al no valer 0, representarían residuos. Esto conduciría a un aumento en la falta de ajuste. A su vez, si estos aportes mínimamente variantes no fueron modelados, también se puede esperar una disminución en el % de varianza explicada. Comparando a las matrices originales con las reducidas por WT para todas las regiones, también se verifican caídas en el número de iteraciones necesarias, lo cual se produce de manera más notoria en las regiones A y D. Esto parece ser coherente con las observaciones realizadas. Las regiones A y D fueron las que sufrieron los peores detrimentos de %LOF EXP con las compresiones, tanto en términos absolutos como relativos a los valores que habían obtenido en el dominio original, y esto pudo haberse debido a interrupciones tempranas del algortimo. Si se recuerda que en MCR-ALS se aplicó un criterio de parada de las iteraciones en caso de que la diferencia de la desviación estándar de los residuos entre iteraciones sucesivas fuera menor al 0.1%, es posible que los modelos con componentes subestimados hayan tenido menos opciones de variación en pro de un mejor ajuste y por lo tanto en iteraciones sucesivas hayan obtenido residuos relevantes pero fundamentalmente similares, lo cual pudo haber detenido tempranamente al proceso de ajuste. De todas formas, un mayor número de iteraciones no siempre conlleva mejores resultados en la resolución, y dicho número no debe ser visto como una cifra de mérito en sí.

Antes de proseguir con el análisis, es necesario realizar una aclaración. Las matrices originales, las reducidas mediante WT y todas las matrices de residuos que pudieron calcularse (incluyendo a las experiencias “Exp” y “Mix”) fueron vectorizadas y a estos vectores se los sometió a un test de normalidad, específicamente el de Lilliefors (Lilliefors, 1967) con un 95% de confianza, mediante la función “lillietest” de Matlab. En todos los casos, la hipótesis nula de que los datos poseían distribución normal fue descartada. La aclaración proviene del hecho de que en la tabla de resultados analizada, existen cifras de mérito que deberían ser obtenidas de datos normalmente distribuidos, contrario a lo obtenido mediante el test de Lilliefors. Más aun, como en este caso los datos provienen de espectrometría de Masa, dadas las características típicas de estos espectros (valores nulos en gran parte de las variables, con valores medios o altos en variables aisladas), es lógico pensar que estos datos no podrían catalogarse como normalmente distribuidos. Sin embargo, cifras como el dato medio se han utilizado como indicadores popularmente conocidos y para realizar comparaciones básicas. Por ejemplo, si se comparan los datos medios de cualquier matriz original con los de su respectiva reducción por WTH2, puede apreciarse que los últimos son

siempre mayores y que el cociente entre ambos valores se acerca a 4. Esto corresponde al efecto de la WTH2 aplicada en 2 escalas y en 2 vías (tiempo y m/z) a través del algoritmo DWT de Matlab y deja entrever que las aproximaciones no son simples promedios de valores consecutivos, sino que son promedios multiplicados por valores constantes, dado que en la matriz de Haar utilizada los valores para el filtro de paso bajo son $(2^{1/2})/2$ en ambos casos, mientras que para el de paso alto son $-(2^{1/2})/2$ y $(2^{1/2})/2$. Específicamente, la reducción con WTH de una señal vectorial en una primera escala hará que el promedio de 2 valores consecutivos sea multiplicado por $2^{1/2}$, y en la siguiente escala ocurrirá lo mismo, contabilizando un total de 2 en la ampliación de los promedios. Al aplicar el algoritmo en 2 vías y con 2 escalas, los promedios serán multiplicados por 4. El cociente de datos medios de las matrices aquí obtenidas tiende al valor de 4 y no llega a éste debido a que las matrices originales no poseen dimensiones que se correspondan exactamente con potencias de 2, lo cual trae aparejados inconvenientes en los extremos de las señales y con esto leves distorsiones de la idealidad. Si de estas mismas matrices originales se seleccionan, por ejemplo, 256 tiempos y 256 m/z, y a estas selecciones se les aplica la WTH en 2 vías y 2 escalas, el cociente de medias entre los datos originales y los reducidos será exactamente 4. La relación de medias de distintas escalas tiene las particularidades descriptas porque en el proceso en sí de la reducción con WTH existe el cálculo de medias para los coeficientes de aproximación. Lo dicho acerca de la relación entre medias no se aplica a otro tipo de estadísticos similares, como podrían ser la mediana y la moda. Tampoco a otros como máximos y mínimos, de lo cual por ejemplo no se puede esperar que el máximo de los datos reducidos en 2 escalas sea 4 veces mayor que el máximo de los datos originales (de hecho, resultaron bastante similares).

Vale aclarar que la WTH podría aplicarse con filtros de paso bajo y alto cuyos valores fueran distintos a los expuestos, siempre que se respete que para el filtro de paso bajo ambos valores sean positivos e iguales (por ejemplo ambos con valor de 1) y que para el filtro de paso alto los valores absolutos sean los mismos que para los de paso bajo, pero con signo negativo en el primero de ellos (continuando con el ejemplo, serían -1 y 1). Estos valores generarían otro tipo de promedios, pero cualitativamente el efecto sería el mismo que con la matriz realmente utilizada. Sin embargo, los filtros de paso bajo y alto poseen los valores descriptos porque de esta manera se simplifica, a nivel algorítmico, el cálculo de inversas matriciales al aplicar IWT, ya que con esos valores, la inversa de la matriz de Haar es igual a su transpuesta (es decir, hay ortogonalidad).

Como pudo apreciarse en la figura 13 los valores de las señales relevantes no difieren demasiado, al menos a simple vista y contemplando que el efecto de la reducción conlleva curvas

menos suaves que en los datos originales. Sin embargo, de la diferencia de valores promedio entre matrices originales y reducidas, y sumando el hecho de que la cantidad de componentes estimados para resolver ambos sistemas fue diferente, pueden esperarse también diferencias en los valores de los residuos luego de cada ejecución de MCR-ALS. Esto puede observarse en las correspondientes cifras de Media |res|, cuyos valores son mayores para las matrices reducidas que para las originales. No obstante, la relación de éstos valores respecto de los datos que les dieron origen también debe ser contemplada. Esto puede observarse en los valores de B/A, que indicarían mejores relaciones para las matrices reducidas en las regiones A y B, mientras que para las regiones C y D los datos originales muestran menores cocientes. Aún así, todos los valores son cercanos, de manera tal que el promedio de B/A para todas las matrices originales es 0.259, siendo de 0.252 para las reducidas. Viendo las similitudes se puede inferir que, sólo en este sentido, MCR-ALS obtendría resultados cuya calidad sería similar en ambos casos, pero con la ventaja de que al aplicar WTH, además del filtrado de ruido, los cálculos podrían realizarse con menores recursos de cómputo (o en igualdad de recursos a mayor velocidad).

Las experiencias “Exp”, con expansiones de perfiles de concentración reducidos actuando como aproximaciones iniciales, tienen algunas particularidades que deben ser analizadas.

En primer lugar, para las matrices originales y para las reducidas con WTH2, las aproximaciones iniciales para MCR-ALS se obtuvieron con SIMPLISMA, de lo cual no podría esperarse que todas esas aproximaciones, obtenidas de datos complejos como los aquí analizados, tuvieran forma de perfil espectral. En cambio, en las experiencias “Exp”, las formas de todas las aproximaciones iniciales sí serían las correspondientes a perfiles de concentraciones, o al menos similares, además de que provienen también de una optimización y deberían contener información útil para la localización de los componentes. Desde estos puntos de vista el ajuste en “Exp” contaría con cierta ventaja.

En segundo lugar, el número de componentes en estas experiencias se heredó de los calculados para las matrices reducidas y, teniendo en cuenta que para las correspondientes matrices originales estos valores son mayores, es de esperarse que exista una subestimación en el número de componentes. Es decir, en las experiencias “Exp” se debió resolver la misma información que con las matrices originales, pero restringiendo la cantidad de componentes que podrían explicar las variaciones. Esto ha repercutido con bastante claridad en varias cifras de mérito. En todas las regiones los valores de %LOF EXP y de $\%R^2$ han desmejorado respecto de la resolución de las curvas originales con más componentes, siendo la región C la de peor ajuste, con poco más que el

7% de la varianza no explicada. De manera similar, los valores para Media |res| aumentaron en las experiencias “Exp”, con los consiguientes aumentos de B/A, ya que A conservó su valor (los datos fueron los originales nuevamente). Observando que en general el número de iteraciones disminuyó, puede deducirse que no existieron cambios relevantes entre los residuos de iteraciones consecutivas y esto puede nuevamente relacionarse con la carencia de componentes para el modelado aunque de forma más evidente, pues las cifras de mérito desmejoraron claramente. La región A utilizó más iteraciones que antes pero no obtuvo mejores resultados, sino todo lo contrario, ya que por ejemplo obtuvo el valor más alto de B/A. Concluyendo, el efecto que podría provenir de supuestas mejores estimaciones iniciales no fue superior al efecto de la subestimación de componentes.

Finalmente, resta analizar las experiencias “Mix”, en las cuales el número de componentes que debían ser resueltos se heredó del análisis de las matrices originales. Como ya se explicó, algunas de estas estimaciones iniciales provenían de perfiles de concentración obtenidos en las experiencias “Exp”. Al comparar estos resultados con los obtenidos originalmente, para todas las regiones puede notarse una leve mejoría en todas las cifras de mérito relativas a la calidad del ajuste (%LOF EXP, $\%R^2$, Media |res| y B/A), a excepción de Media |res| para la región B, por lo cual B/A aumenta mínimamente. A su vez el número de iteraciones disminuyó en todos los casos, muy probablemente porque las estimaciones iniciales, esta vez en cantidad más apropiada que en “Exp”, tenían mejores características que las originales obtenidas con SIMPLISMA. La disminución en la cantidad de iteraciones fue aproximadamente de 61%, 50%, 50% y 93%, para las regiones A, B, C y D, respectivamente. Si bien la mejoría de las cifras de mérito no fue extrema, se obtuvieron resultados de calidad similar con menos iteraciones, lo cual puede representar tiempo ganado. Con estos valores, puede pensarse en una estrategia para la resolución de este tipo de matrices. En principio, los datos originales deberían ser reducidos con WTH2, utilizando varios niveles de escalado. Una vez que se hubiera definido un nivel mínimo, que para este ejemplo será 4, se estimaría la cantidad de componentes y mediante SIMPLISMA se obtendrían aproximaciones iniciales, para luego ejecutar MCR-ALS. Los perfiles de concentración que se hubieran obtenido deberían expandirse para tener dimensiones concordantes con las matrices reducidas hasta un nivel anterior al utilizado (para el ejemplo, 3), es decir, del doble de tamaño en dimensiones. Hechas las expansiones, estos perfiles serían utilizados como aproximaciones iniciales para el nivel actual, aumentando con otras aproximaciones (obtenidas de alguna forma) para completar el número de componentes recomendado para dicho nivel. Hecho lo anterior, se procedería a ejecutar MCR-ALS. Esta secuencia de pasos sería repetida hasta alcanzar las dimensiones de los datos originales o, dicho de

otra manera, en el nivel de escala 0. Aunque con este procedimiento no habría garantías de que el número total de iteraciones sería menor que el necesario al aplicar MCR-ALS directo sobre las matrices originales, pueden darse casos en los cuales esto suceda. Al mismo tiempo, podría obtenerse información útil al ir desde una escala hacia otra. Por ejemplo, evaluando los perfiles resueltos en una escala y comparándolos con los de la siguiente, en la cual habría más componentes recomendados para explicar el sistema desde un punto de vista matemático, pero quizá sin sentido físico y/o químico. También podría obtenerse información acerca de qué escala sería necesaria para que cada componente, o al menos los que fueran de interés, pudiera comenzar a ser percibido, o qué componentes se expresan como relevantes a todas las escalas.

También fueron realizadas algunas comparaciones entre los perfiles resueltos con las distintas estrategias. Estas comparaciones se realizaron a través del cálculo de coeficientes de correlación de Pearson.

En primer lugar, se compararon los perfiles espectrales derivados de las matrices originales (**S**) con los obtenidos luego de la expansiones (**SExp**). Ya que los últimos eran menos que los primeros, cada perfil en **SExp** fue comparado contra todos los posibles en **S**, rescatando de esta comparación al componente original con mayor correlación para cada caso. Este planteo (y otros similares que se expondrán a continuación) implican la posibilidad de que más de un perfil expandido optara por el mismo perfil original al mismo tiempo, pero esto se ha dado en escasas ocasiones. Similarmente, fueron comparados los perfiles de concentración **C** y **CExp** provenientes de las mismas experiencias. Adicionalmente, se cotejaron los perfiles espectrales **SExp_wr1** y **S_wr** por un lado, y **S_wr1** y **S_wr** por el otro. Los resultados pueden ser observados en la tabla 3. La primera de las comparaciones (**S** vs **SExp**) muestra valores de correlación aceptables, siendo mejores los de la región C, incluyendo una menor desviación estándar. A su vez, en dicha región se presentó una repetición en la selección de componentes, y ninguna en el resto de las regiones. Por otro lado, si se analizan individualmente los valores que dieron origen a los promedios, se puede apreciar que muchos de ellos son iguales o superiores a 0.98, específicamente 12 de 20, 9 de 16, 8 de 14 y 8 de 11 para las regiones A, B, C y D, respectivamente. Estos resultados muestran que existió una convergencia notable en varios de los espectros resueltos, y debe tenerse en cuenta que los presentes en **S** partieron de estimaciones hechas con SIMPLISMA, mientras que los correspondientes a **SExp** provienen de una serie de pasos distinta (compresión WTH2 de los datos originales, SVD y SIMPLISMA, MCR-ALS, Expansión polinómica de perfiles de concentración y resolución con MCR-ALS nuevamente). A su vez, los vectores comparados tenían 710 variables cada uno, de lo

cual no parece intuitivo pensar que por azar se obtendrían correlaciones altas, aunque debe recalcarse que dadas las características de los espectros de Masa (señales relevantes en pocas y aisladas variables), sumado a las restricciones de no-negatividad impuestas en MCR-ALS, muchos de los valores de éstas variables deberían ser 0 en ambos tipos de perfiles, con lo cual el cálculo de correlaciones tendería a entregar coeficientes altos, aun cuando lo más importante para evaluar serían las pocas variables con valores distintos de 0.

	Región A	**Región B**	**Región C**	**Región D**
	S_{39} vs $SExp_{20}$	S_{32} vs $SExp_{16}$	S_{26} vs $SExp_{14}$	S_{25} vs $SExp_{11}$
r^2 medio	0.935	0.934	0.961	0.941
std r^2	0.114	0.113	0.050	0.103
	C_{39} vs $CExp_{20}$	C_{32} vs $CExp_{16}$	C_{26} vs $CExp_{14}$	C_{25} vs $CExp_{11}$
r^2 medio	0.966	0.947	0.964	0.956
std r^2	0.055	0.059	0.037	0.056
	$SExp\_wr1_{20}$ vs S_wr_{20}	$SExp\_wr1_{16}$ vs S_wr_{16}	$SExp\_wr1_{14}$ vs S_wr_{14}	$SExp\_wr1_{11}$ vs S_wr_{11}
r^2 medio	0.917	0.960	0.973	0.943
std r^2	0.127	0.083	0.030	0.094
	S_wr1_{39} vs S_wr_{20}	S_wr1_{32} vs S_wr_{16}	S_wr1_{26} vs S_wr_{14}	S_wr1_{25} vs S_wr_{11}
r^2 medio	0.909	0.960	0.950	0.951
std r^2	0.101	0.048	0.063	0.049

Tabla 3: Comparación de perfiles de concentración y espectrales resueltos mediante MCR-ALS utilizando distintas estrategias con WT y derivados.

Referencias: r^2 medio: Coeficiente de correlación de Pearson promedio, std r^2: desviación estándar de los coeficientes de correlación de Pearson. Los subíndices indican la cantidad de componentes comparados en cada grupo.

Las comparaciones siguientes (**C** vs **CExp**) muestran incluso mejores valores en los promedios y en sus desviaciones estándar. Sólo se encontró una selección repetida en la región A. En relación al número de coeficientes que igualaros o superaron el valor 0.98, éstos fueron 10 de 20 en A, 7 de 16 en B, 7 de 14 en C y 7 de 11 en D. Estas buenas correlaciones también pueden verse favorecidas por la restricción de no-negatividad tal y como los perfiles espectrales, aunque es meritoria la localización de los picos en el tiempo. Vale destacar que, con escasas excepciones, los componentes originales seleccionados mediante perfiles espectrales y mediante perfiles de concentraciones fueron exactamente los mismos en términos de identidad.

Las últimas dos comparaciones analizadas involucran a las experiencias “Exp” y si se recuerda el análisis de la tabla de cifras de mérito, dicha experiencia presentaba varias deficiencias probablemente debidas a un número insuficiente de aproximaciones para el modelado. No obstante, debe notarse que el ajuste de varios componentes concluyó en resultados similares respecto de la situación original (donde la deficiencia podría existir pero sería menor), lo cual queda plasmado con los altos coeficientes de correlación encontrados. Por consiguiente, puede pensarse que la subestimación de componentes no se traduce en un reparto equitativo del error entre los componentes numéricamente estimados, sino que algunos componentes son dominantes y obtienen resultados de ajuste mejores y similares, haya o no deficiencias. Las zonas de la matriz no relacionadas a estos componentes presentarán grandes errores cuando haya subestimación porque no habrá componentes para modelar dichas zonas y esto traerá aparejado el detrimento de las cifras de mérito generales, tal como ya se observó para “Exp”. Otra conclusión puede extrapolarse en relación a la WT como técnica de compresión, en el sentido de que parece haber conservado la suficiente cantidad de información como para que algunos componentes se manifiesten y se resuelvan de manera similar con o sin reducción. Esto último es positivo, pues entre los objetivos de usar la WT está el que no se pierda la información relevante de los componentes siempre que sea posible.

La siguiente comparación contempló a los perfiles espectrales resueltos a partir de las matrices reducidas (**S_wr**) con aquellos obtenidos utilizando las expansiones de los perfiles de concentración como estimaciones iniciales para resolver las matrices originales, con posterior compresión mediante WTH1(**SExp_wr1**). A diferencia de las comparaciones anteriores, en este caso la cantidad de elementos comparados fue la misma (20 vs 20), ya que los perfiles espectrales en **SExp_wr1** provienen indirectamente de los de **S_wr** y por esto es posible también comparar las identidades de estos perfiles. Los resultados mostraron buenas correlaciones nuevamente. En cuanto a la cantidad de componentes que mostraron correlaciones mínimamente de 0.98, estos fueron 10 de 20 para la región A, 12 de 16 para la región B, 7 de 14 para la región C y 6 de 11 para la región D. En todas las regiones se presentó una selección repetida y, en particular la región A, la cual tuvo la menor de las correlaciones, presentó además 3 selecciones permutadas. En otras palabras, de los 20 componentes comparados, 16 de ellos presentaron la misma identidad, 1 de ellos fue seleccionado 2 veces, y los restantes 3 fueron asociados a otros 3 que no habían sido previamente seleccionados, pero sin coincidir en su identidad (de allí lo de “selecciones permutadas”). Específicamente, los perfiles espectrales de los componentes 17, 18 y 19 de **S_wr** encontraron su mayor correlación con los

perfiles espectrales de los componentes 20, 19 y 18, respectivamente, en **SExp_wr1**. Cabe también recordar que sólo en la región A se necesitaron muchas iteraciones (51) en las experiencias "Exp", por lo cual se puede pensar que durante los ajustes de MCR-ALS, la mayor parte de las modificaciones fueron realizadas sobre los componentes permutados.

Las siguientes comparaciones relacionan a los perfiles espectrales resueltos desde las matrices reducidas (**S_wr**) con aquellos obtenidos desde las matrices originales directamente y con posterior compresión mediante WTH1 (**S_wr1**). Las correlaciones promedio y sus desviaciones estándar son aceptables y similares a las anteriores para las regiones B, D y C, presentándose en la última la única selección repetida. La región A presenta los menores valores de correlación promedio en relación a las comparaciones anteriores, pero de todas formas estas cifras no son bajas. Más específicamente, 7 de 20 componentes de la región A mostraron correlaciones mínimas de 0.98, 9 de 16 en la B, 6 de 14 en la C y 5 de 11 en la D.

A su vez, puede realizarse otra comparación con estos perfiles. En la última comparación analizada, N perfiles espectrales de las matrices originales reducidos mediante WTH1 fueron cotejados contra n perfiles con procedencia directa desde las matrices reducidas, siendo N>n, y dependiendo ambos valores de la región en cuestión (por ejemplo, para la región A los valores son 39 para N y 20 para n). Similarmente, en la primera comparación vista (**S** vs **SExp**), N perfiles obtenidos directamente desde las matrices originales fueron evaluados en su correlación con n perfiles que también fueron obtenidos de manera directa desde los datos originales, pero que a su vez contaban con aproximaciones iniciales que, indirectamente, provenían de lo resuelto con las matrices reducidas. Si se evalúa cada región por separado en estas dos experiencias, se pueden evaluar también las identidades de los componentes seleccionados. En efecto, al realizarlo se observa que en ambos casos y para todas las regiones, varias de las identidades seleccionadas son las mismas. En la región A 14 de 20 componentes coincidieron, en la B lo hicieron 14 de 16, en la C fueron 12 de 14 y 8 de 11 en la D.

Las cifras de mérito evaluadas y diferenciadas según cada situación, así como las altas correlaciones encontradas para los perfiles resueltos, tanto de concentraciones como espectrales, mediante mecanismos diferentes, sugieren que existió una equivalencia aceptable entre distintas estrategias, de lo cual puede pensarse que la resolución de las matrices reducidas con WTH2 podría representar en cierto grado a la resolución de la información original, conservando características importantes en las señales y disminuyendo los recursos de cómputo necesarios, aun cuando la

calidad de los ajustes no sea la misma.

Los análisis recientes fueron derivados de la resolución mediante MCR-ALS de una única muestra, lo cual ha resultado suficiente para verificar que la WT puede ser utilizada como método de pretratamiento de los datos en lo que resta del trabajo. En los análisis posteriores se implementará MCR-ALS con matrices apiladas, lo cual conlleva la ventaja de contener mayor cantidad de información útil para resolver ambigüedades y determinar, para especies comunes a diversas matrices, los mismos perfiles espectrales. Por lo tanto, no es de esperarse que los perfiles resueltos para la matriz RA1 serían exactamente los mismos que los ya obtenidos.

Si bien la estrategia de compresión con WT estuvo basada según (Peré-Trepat y Tauler, 2006), una crítica importante a cómo fue aplicada la WT en ese y en el presente trabajo proviene del hecho de que al no haber usado explícitamente los coeficientes de detalle en absoluto, y al no haber realizado IWT en general, el uso que se le dio a la WT podría haber sido reemplazado por simple promediación de valores, que es lo que realiza la Wavelet de Haar con las aproximaciones (aunque luego afecte al promedio con una constante). Más aun, si se hubiese elegido otra familia de Wavelets, simplemente se hubieran cambiado los valores de ponderación en el cálculo de promedios. No obstante, se deja abierta la posibilidad de adaptar las estrategias para poder utilizar la IWT y los coeficientes de detalle, en cuyo caso el marco ya planteado sí sería adaptable a cambios en las funciones de anásis. Además, debe recordarse que la WT fue utilizada para facilitar los cómputos dadas las limitaciones del instrumental utilizado y la complejidad de los datos analizados, y en este sentido, futuras aplicaciones realizadas con mejores prestaciones de cálculo podrían adaptarse. Ejemplo de esto sería que en lugar de resolver el sistema con las matrices reducidas y en el dominio de los coeficientes de aproximación, se procediera aplicando WT, dejando luego de lado los coeficientes de detalle y aplicando IWT para reconstruir las matrices nuevamente hacia su dominio original, sólo que esta vez el efecto del procedimiento seguiría una lógica de eliminación de ruido previo al análisis, y no de compresión en sí, ya que las matrices serían utilizadas con sus dimensiones originales.

2.6.3 Análisis MCR-ALS de muestras en simultáneo: Generalidades

Antes de comenzar con la discusión específica de los detalles de cada una de las 2 grandes partes en que se divide el trabajo, ya mencionadas y brevemente resumidas en la sección “Datos obtenidos: separación del estudio en partes”, vale la pena discutir asuntos relacionados a algunas

estrategias que han sido generales:

2.6.3.1 Reducción del tamaño

Todas las matrices originales fueron reducidas mediante DWT2 con filtros de Haar en 2 niveles, llegando a tener dimensiones de 127 × 178, y con estas matrices se obtuvieron las resoluciones. La discusión sobre el efecto de la WT ha sido realizada anteriormente.

2.6.3.2 División en regiones

Si bien esto permitió agilizar los cómputos, hay que mencionar potenciales efectos negativos de este proceder. En el fraccionamiento de los tiempos de retención no se tuvo en cuenta si los tiempos elegidos para realizar las divisiones coincidían, al menos para algunas muestras, con picos cromatográficos evidentes. Si esto hubiese ocurrido y suponiendo que existiera sólo un componente dividido (pudieron ser más), dicho componente debería ser resuelto en ambas regiones. Lo anterior implica que:

- Si por alguna razón dicho componente resultara clave en un análisis metabonómico, la concentración resuelta en ambas regiones debería sumarse y esta suma debería ser utilizada en el análisis. Esto no ha sido realizado en este trabajo y si ocurrió la situación descripta, el mismo componente ha sido evaluado como si hubiesen sido 2 por separado.
- Los espectros resueltos para componentes partidos deberían ser iguales para ambas regiones, lo cual es poco probable que haya ocurrido, ya que cada región puede tener particularidades (ruido, cantidad de componentes, entre otros) que afecten el ajuste con MCR-ALS. Tampoco se le dio un tratamiento especial a este asunto.

2.6.3.3 Obtención de matrices aumentadas por apilamiento

La resolución en 3 vías es más efectiva que la resolución de matrices individuales ya que siempre introduce mejoras significativas en la obtención de perfiles de respuestas verdaderos, además del beneficio adicional de proveer capacidades de cuantificación potenciales. Entre las familias de algoritmos dedicados a resoluciones de orden 3 o superior, las del tipo iterativo ponen el foco de atención en la optimización de estimaciones iniciales a través de la utilización de estructuras de datos apropiadas y de restricciones matemáticas y químicas (Smilde y col., 1994). Por ende, realizar análisis simultáneos usando MCR-ALS con datos obtenidos de experimentos

múltiples e independientes puede considerarse una herramienta útil y una estrategia poderosa para las resoluciones. Tal como se ha dicho en la sección de Teoría, la ecuación (6) puede ser extendida para análisis simultáneos de muestras obtenidas con la misma técnica de detección, siendo en nuestro caso señales de MS. En lugar de resolver los datos conformando un arreglo de 3 vías, es posible agrupar las matrices por apilamiento, conformando de esta manera matrices mayores aumentadas por columnas. De esta manera, la resolución de la matriz resultante llevará a que los perfiles espectrales que se obtengan serán comunes a todos los experimentos, más allá de que los perfiles de concentración puedan ser diferentes entre cada muestra.

La resolución completa de matrices aumentadas depende mayoritariamente de 2 características. Una de ellas es la estructura matricial en términos de rango, es decir, si los componentes generadores de varianza son independientes y pueden ser estimados. La estimación de la cantidad de estos componentes fue realizada mediante SVD. La otra característica importante que deben tener las matrices aumentadas es la presencia de variables puras por cada componente modelado, lo cual aumentará la selectividad y mejorará la resolución con MCR-ALS. Las estimaciones iniciales de los espectros, que son el punto de partida en la resolución multivariada, fueron obtenidas con SIMPLISMA.

2.6.3.4 Cálculo del número de componentes mediante SVD

La estrategia utilizada en todas las ejecuciones de SVD impuso que, como máximo, se explicaría un 90% de las variaciones en los datos experimentales (una matriz apilada por región). En este trabajo sólo se puso énfasis en la suma acumulada de valores singulares ordenados de forma decreciente (respecto de la suma total de éstos), es decir, se seleccionaron tantos componentes como los necesarios para obtener una suma acumulada, relativa al total, igual o levemente superior a 0.90. Esto último es cuestionable, porque es probable que se generaran valores singulares muy pequeños, de escasa importancia en relación a la suma de todos los valores singulares, pero que por la decisión de conservar tantos como sea necesario para llegar a la fracción 0.90 hubieran sido conservados. Más aun, es probable también que el último en ser incluido haya tenido como sucesor a otro valor singular inferior pero muy similar, por lo cual decir que uno fue importante y no el otro parece una decisión trivial. Otras veces se impone además que la proporción de un valor singular sobre la suma supere otro umbral, por ejemplo que sea mayor al 1% del total, aunque el establecimiento de una relación entre un componente y la cantidad de varianza mínima que debería representar para ser

considerado como tal, también suele ser trivial. Vale además destacar que la matriz **Σ** contiene a los valores ordenados según su proporción en la generación de la varianza total, pero una vez que los componentes son seleccionados, sus aproximaciones iniciales (obtenidas con SIMPLISMA) pasan a ser utilizadas en condiciones de igualdad dentro de MCR-ALS. Con lo anterior debe entenderse que la información que otorga la SVD acerca de cuál componente podría explicar en mayor o menor grado la varianza no es utilizada, y que de la SVD sólo se obtiene un número concreto de hipotéticos componentes variantes. No obstante a las críticas anteriores, con la SVD y el esquema propuesto sólo se pretendió obtener un indicador común y sistematizable para todas las ejecuciones de MCR-ALS, aunque otras estrategias podrían haberse seguido. Los errores en la estimación de componentes, sean por defecto o por exceso, merecen una mínima reflexión, aun cuando en algunas ocasiones puedan no ser trascendentales. La sobreestimación de componentes forzará el ajuste de forma tal que información que debería ser descartada será modelada y aunque esto puede hacerse adrede con la intención de modelar información no relacionada directamente a los componentes de interés, es probable que en algunas circunstancias se actúe en desmedro del ajuste de estos. Por su parte, la subestimación de componentes conlleva una resolución deficiente, donde algunos perfiles resueltos podrían no ser componentes definidos, sino combinaciones de éstos, o bien donde solamente unos pocos componentes dominantes serán los ajustados.

2.6.3.5 Obtención de estimaciones espectrales iniciales con SIMPLISMA

Debido a la ausencia de bases de datos de las cuales podrían haberse extraído espectros de metabolitos de tomate obtenidos con las mismas técnicas aquí aplicadas, así como también a la imposibilidad de obtener dichos espectros a través de patrones puros dada la alta complejidad de los datos, no se utilizaron espectros puros reales como estimaciones iniciales para MCR-ALS. En su lugar, se utilizaron estimaciones de los espectros de Masa obtenidas con SIMPLISMA a partir del análisis de cada una de las matrices aumentadas que fueran luego resueltas. Dado que el algoritmo lo requiere, se impuso un nivel de ruido en las señales, que en nuestro caso fue de 0,1. Si bien en sistemas sencillos SIMPLISMA puede obtener estimaciones muy similares a lo que luego MCR-ALS resolverá finalmente, los datos aquí analizados no poseen tal característica. Por lo tanto, SIMPLISMA fue utilizado con la única finalidad de obtener una estimación por componente hipotético (según SVD), sin poner énfasis en la calidad de todas estas estimaciones. No obstante, caben algunos comentarios:

- La complejidad de los datos es un factor importante a la hora de evaluar si las estimaciones serán aptas o no. Los datos aquí trabajados son complejos siendo que provienen de un fruto natural y no se podría esperar que todos los espectros estimados con SIMPLISMA tengan las características propias de la espectrometría de Masa. No obstante, se han visto estimaciones iniciales con características adecuadas para la espectrometría en cuestión, es decir, pocas variables con valores significativamente distintos de cero. Aun con eso, la optimización posterior mediante MCR-ALS es lo que termina definiendo los espectros optimizados, pero es lógico pensar que si las estimaciones son buenas, la resolución también podría serlo. Vale decir que uno de los primeros pasos de MCR-ALS es calcular por primera vez a la matriz **C** a partir de las estimaciones iniciales de SIMPLISMA en **S** (obviamente si las estimaciones iniciales están en **C** el primer cálculo será de **S**). Luego de ese paso se continúa con la primera reconstrucción de la matriz apilada, **D'**, y desde ésta y con la matriz apilada original, **D**, se obtienen e informan cifras de mérito sobre la calidad de la reconstrucción antes de la primera iteración de MCR-ALS. Aunque estas cifras no hayan sido transcritas en este texto, se han observado valores de % de varianza explicada no necesariamente bajos e incluso a veces cercanos hasta el 80%, de lo cual se puede suponer que las estimaciones de SIMPLISMA resultaron aceptables.
- El hecho de calcular las estimaciones con matrices aumentadas, así como también el de solicitar el cálculo de varias aproximaciones al mismo tiempo, implica muchos recursos de cómputo y en estos términos el algoritmo ha resultado costoso y lento.

2.6.3.6 Aplicación de restricciones en MCR-ALS

Durante la optimización iterativa se aplicaron básicamente 2 restricciones con el objetivo de obtener resoluciones con significado químico. Así pues, en cada iteración en la que MCR-ALS obtiene nuevas matrices **C** y **S**, el algoritmo incorpora las restricciones impuestas provenientes del conocimiento químico del sistema.

Una de las restricciones en cuestión fue la de no-negatividad, ya que con estos datos no tendría sentido químico obtener perfiles de concentración o espectrales con valores negativos. Como ya se ha explicitado, el umbral de tolerancia para aplicar la restricción fue el provisto por defecto en el algoritmo usado. Este algoritmo utiliza iteraciones internas y evalúa si el criterio es cumplido. En caso de no serlo, realiza más iteraciones, hasta un valor máximo que depende del número de

muestras simultáneas que se están analizando. Vale destacar que en experiencias no reportadas, el algoritmo fue internamente modificado para que el máximo de iteraciones fuera mucho menor a lo que debiera ser por defecto. Al evaluar cifras de mérito para ajustes realizados en igualdad de condiciones y solamente difiriendo en el detalle comentado, se obtuvieron cifras de mérito similares. Por ejemplo, los %LOF EXP observados diferían en el segundo decimal. Esta diferencia tan pequeña no parece merecer la pena, en especial porque con los cambios realizados la velocidad de las iteraciones fue notablemente mayor que lo normal, algo paradójico dado que el algoritmo se denomina "No Negatividad Rápida mediante Mínimos Cuadrados", o FNNLS.

La otra restricción aplicada fue la de unimodalidad en los perfiles de concentración. Esta restricción en los cromatogramas puede ser relacionada con la subestimación de componentes proveniente de la SVD y con la resolución en la captura de las señales, tal y como se esquematiza en la figura 14.

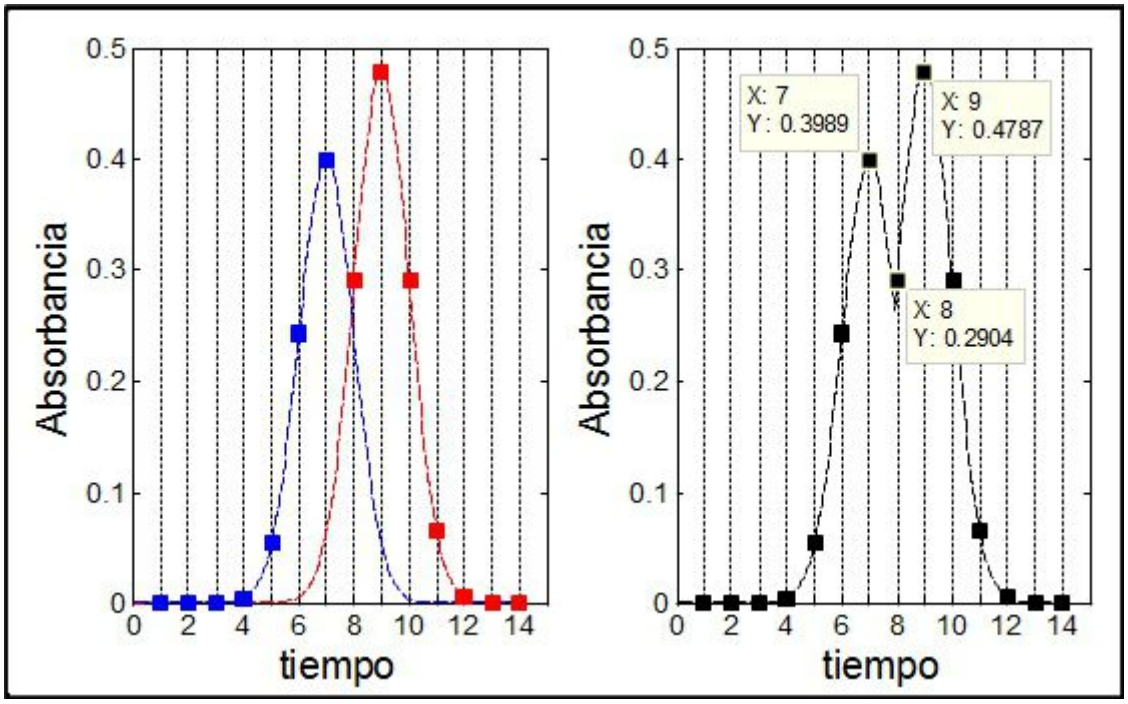


Figura 14: Relación entre restricción de unimodalidad, subestimación de componentes y resolución de captura de las señales

Referencias: Izquierda: Cromatogramas de 2 componentes hipotéticos (rojo y azul), parcialmente solapados. Derecha: Simulación de señal resultante (los puntos señalados con X e Y son de interés, ver el desarrollo del texto a continuación). Los cuadrados y las líneas de puntos verticales indican los tiempos de adquisición de la señal. Las líneas entrecortadas esquematizan las formas que se hubieran obtenido con tiempos de adquisición a mayor resolución.

En la parte izquierda de la figura 14 pueden observarse 2 distribuciones normales simuladas, las cuales representan los cromatogramas de 2 componentes a una determinada longitud de onda. Si no se produce subestimación de componentes, entonces sería posible resolverlos por separado y de forma correcta. La parte derecha muestra la señal capturada desde un equipo, antes de ser procesada. Si bien las líneas entrecortadas indican cómo serían las formas de las campanas si la captura de los datos se hubiese hecho a alta resolución, los puntos realmente adquiridos están representados por los cuadrados sobre cada curva. En este escenario y suponiendo una subestimación de componentes, es probable que lo representado en la gráfica de la derecha sea tomado como un único componente. La restricción de unimodalidad se aplica si en la bajada desde un máximo absoluto se produce un nuevo crecimiento en un punto X de la señal, tal que dicho crecimiento respecto de un punto Y menor y anterior (en dirección desde X hacia el máximo) supera cierto % del valor en Y. En la gráfica el máximo se encuentra en el tiempo 9, el punto Y en el tiempo 8, y el punto X en el tiempo 7. El % de incremento de X sobre Y se aproxima al 37% del valor de Y (100x0,3989/0,2904-100). Por ende, un % de tolerancia para unimodalidad inferior a 37% haría que en este caso la restricción se aplique, por lo cual la forma del pico resuelto no representará ciertamente ni al componente azul ni al rojo. No obstante, debe notarse que con una alta resolución de captura y aunque haya subestimación de componentes, será menos probable que se ejecute la restricción en cuestión (siempre y cuando el % de tolerancia no sea muy bajo) y por tal los perfiles resueltos pueden ser no unimodales. En este trabajo se usó tolerancia cero, por lo cual es posible que se haya perdido información si existió subestimación de componentes. No obstante, analizando el archivo fuente de Matlab (“als.m”) se encontró que si el usuario impone 1 como valor, éste será cambiado a 1.0001 (0.01%) sin ningún tipo de aviso.

Si la restricción de unimodalidad no se aplicara con severidad, cabría la posibilidad de que al finalizar el ajuste con MCR-ALS se pudieran inspeccionar visual o analíticamente los perfiles de concentración resueltos, con el objeto de indicar si alguno de éstos no tiene características de unimodalidad. En dicho caso, cabría un análisis posterior y local en cada región donde existirían estos picos, estimando la cantidad de componentes localmente y con variaciones en el % de tolerancia para unimodalidad, generando así la posibilidad de una mejor resolución.

Existe además otro tipo de restricciones, que no son equivalentes a las recientemente descriptas en el sentido de que no afectan el cálculo de inversas ni modifican valores en los perfiles, pero que limitan la cantidad de iteraciones a utilizar para lograr la convergencia del sistema resuelto, por lo que suelen denominarse condiciones de parada. Estas pueden darse porque se ha alcanzado un

número máximo de iteraciones (impuesto en 300) sin convergencia o porque se ha logrado la convergencia en sí. Como ya se ha hablado, la última vendrá determinada por un parámetro que compara la desviación estándar de los residuos entre iteraciones consecutivas y determina si la diferencia entre ambos casos es superior o inferior al parámetro. En el primer caso habría que realizar una nueva iteración si aún no se hubiera llegado al máximo permitido, mientras que si la diferencia fuera menor al parámetro se habría alcanzado la convergencia. En este trabajo se impuso el valor del mencionado parámetro en 0,1%, aunque este valor podría ser modificado dependiendo de la etapa en que se encuentre la optimización. Todas las ejecuciones de MCR-ALS terminaron antes del máximo de iteraciones.

2.6.4 Análisis MCR-ALS de muestras en simultáneo: Parte 1

Se aplicó MCR-ALS a un grupo de muestras con el objetivo de obtener perfiles de concentración y espectrales de hipotéticos metabolitos endógenos de tomate, para luego poder evaluar cómo fueron evolucionando a través del período de muestreo.

Las matrices utilizadas fueron las del sector A, cultivar Rambo, blancos y tratadas con Carbofurano. En todos los casos se utilizaron las matrices reducidas con WT. Las 16 matrices de dimensiones 127 × 178 fueron divididas en las regiones A, B, C y D, como ya se ha explicado. Por cada región se conformó un apilamiento, posicionando a las muestras tratadas en la parte superior y a los blancos en la inferior, manteniendo el espacio de columnas en común. A su vez, ambas partes contenían a las muestras ordenadas según su día de recolección. De todas formas, este orden establecido no es fundamental a la resolución de las curvas con MCR-ALS, aunque sí resulta funcional a la hora de evaluar las cinéticas de los componentes resueltos.

Como cada región tuvo una cantidad de tiempos definida, los apilamientos resultaron tener dimensiones diferentes en cuanto a filas. A su vez, al aplicar SVD sobre las matrices aumentadas, cada evaluación otorgó un número de componentes a resolver.

Habiendo definido el número de estimaciones iniciales, éstas fueron obtenidas con SIMPLISMA. Posteriormente, se ejecutó MCR-ALS con las restricciones ya discutidas y se obtuvieron los perfiles resueltos. Los espectros estimados en la matriz **S** resultaron comunes (aunque distintos para cada región) a todas las muestras apiladas sin importar el día de recolección ni el carácter de muestra tratada o blanco, mientras que para cada componente en cada muestra en particular se obtuvo un perfil de concentraciones en **C**. Los resultados obtenidos se exponen en la

tabla 4.

Matriz apilada	región A	región B	región C	región D
filas	560	560	480	432
ncomp	109	88	60	55
%LOF EXP	16.480	13.180	18.910	11.840
$\%R^2$	97.280	98.260	96.420	98.600
iter	20	13	5	13

Tabla 4: Detalles y cifras de mérito por región para MCR-ALS (Parte 1)

Referencias: ncomp: número de componentes en MCR-ALS, %LOF EXP: Porcentaje de Falta de Ajuste Experimental, $\%R^2$: Porcentaje de Varianza Explicada, iter: cantidad de iteraciones

En la tabla 4 puede apreciarse que el número de componentes estimados por región siempre ha resultado mayor respecto de la misma estimación cuando ésta fue realizada en el estudio del efecto de la WT. Si se recuerda, dicho análisis se realizó sobre una única muestra y las regiones (en el mismo orden que en la tabla 4) obtuvieron 20, 16, 14 y 11 componentes estimados cuando la matriz analizada había sido reducida con WT, mientras que con la matriz original estos valores fueron 39, 32, 26 y 25. Por lo tanto, puede notarse que el hecho de estimar los componentes con un mayor número de muestras, representando diferentes estadios metabólicos según el día de recolección y además contando con la información contenida en los blancos, dará como resultado un mayor número de componentes que deberían explicar el % de varianza objetivo.

En cuanto a iteraciones, las regiones tuvieron sus particularidades. La región C fue la única que necesitó menos iteraciones respecto de las necesarias durante el estudio del efecto de la WT, las cuales habían sido de 8 para la matriz original y de 6 para la reducida con WT. Es decir, esta región presenta rápida convergencia tanto si se analiza una muestra sola (original o reducida) o 16 en simultáneo, aunque esto no signifique que dicha convergencia será coincidente con los mejores resultados, sino que el criterio de detención de las iteraciones según el cambio en los residuos de iteraciones consecutivas no permite avanzar más allá de unas pocas iteraciones. Para esta región, quizá se lograría un mejor ajuste adaptando el criterio mencionado para que la detención no sea tan temprana y exista una mayor posibilidad de exploración del espacio en estudio. Las otras 3 regiones requirieron un número de iteraciones menor que el necesario para resolver una única matriz original pero mayor al requerido con la matriz reducida. Particularmente, la región B necesitó casi la misma cantidad de iteraciones (13) que con una sola matriz original (14), mientras que las regiones A y D resultaron ser más similares, en estos términos, a las resoluciones con la matriz reducida.

Si bien la cantidad de iteraciones puede relacionarse con el tiempo de cómputo, dicha relación no tiene por qué ser estricta. El tiempo necesario para el pasaje de un paso iterativo actual hacia el siguiente rara vez será constante entre pasos diferentes, más allá de utilizar el mismo sistema de cálculo siempre. En el caso de MCR-ALS, en cada iteración se obtienen perfiles que deberán ser refinados hasta que se cumpla un criterio de convergencia o se de por terminado el proceso. Entre otras cosas, el refinamiento implica evaluar si algunas de todas las restricciones propuestas deben ser impuestas o no a los perfiles. Esto último es una causa de variabilidad del tiempo entre iteraciones. Por ejemplo, puede darse que los perfiles de la iteración N no requieran que se aplique la restricción de no-negatividad, y en cambio ésta puede tornarse necesaria en la iteración N+1, por lo cual el tiempo que demore será mayor. Otro factor de variabilidad en el tiempo es la obtención de inversas matriciales, ya que no todas las inversiones son igualmente costosas en términos de cálculo. A su vez, no existe una relación evidente entre cantidad de iteraciones y calidad de resultados. Por ejemplo, es más probable que se obtengan mejores resultados con buenas estimaciones iniciales y pocas iteraciones para su refinamiento a que sean obtenidos partiendo de malas estimaciones y con la posibilidad de iterar tantas veces como se desee.

Los resultados de %LOF EXP y $\%R^2$ serán analizados conjuntamente. Antes, vale destacar que todas las regiones obtuvieron $\%R^2$ superiores al 96,4%, lo cual resulta aceptable dado que el objetivo consistía en explicar al menos un 90% de la variabilidad. Para la región A, ambas cifras de mérito resultaron levemente peores que las obtenidas en las experiencias con una sola muestra (reducida y original). Contrariamente, la región B mejoró notablemente si se realiza la misma comparación. Por su parte, las regiones C y D obtuvieron resultados mejores si se los compara con los obtenidos con la matriz reducida, pero levemente peores que aquellos obtenidos con la matriz original, aunque más cercanos a los últimos que a los primeros. Sin dudas, la región D obtuvo grandes mejorías con la resolución conjunta de muestras, ya que no sólo obtuvo las mejores cifras de mérito en cuestión respecto de las otras regiones, sino que además en las experiencias con una sola muestra y con la matriz reducida con WT esta misma región presentó las cifras de mérito de menor calidad entre todas las analizadas. En resumen, viendo que estas cifras de mérito para todas las regiones se mantuvieron similares o en general mejoraron respecto de los obtenidos con una sola matriz reducida, sumado al hecho de que se modelaron más componentes en varias muestras en simultáneo, el apilamiento resultó benéfico para los ajustes.

2.6.4.1 Comparación de perfiles de evolución durante los días de muestreo en Muestras Tratadas y Blancos

Luego de la resolución con MCR-ALS, además de los perfiles en **C** y **S**, se obtuvo una matriz **A** conteniendo las áreas bajo los perfiles de concentración para cada componente manteniendo el orden según el día de muestreo y separando Muestras Tratadas (MT) de Muestras Blanco (MB). De esta forma, la extracción de una fila de **A** resulta en los perfiles evolutivos de cierto componente para ambos tipos de muestra. La figura 15 esquematiza lo dicho.

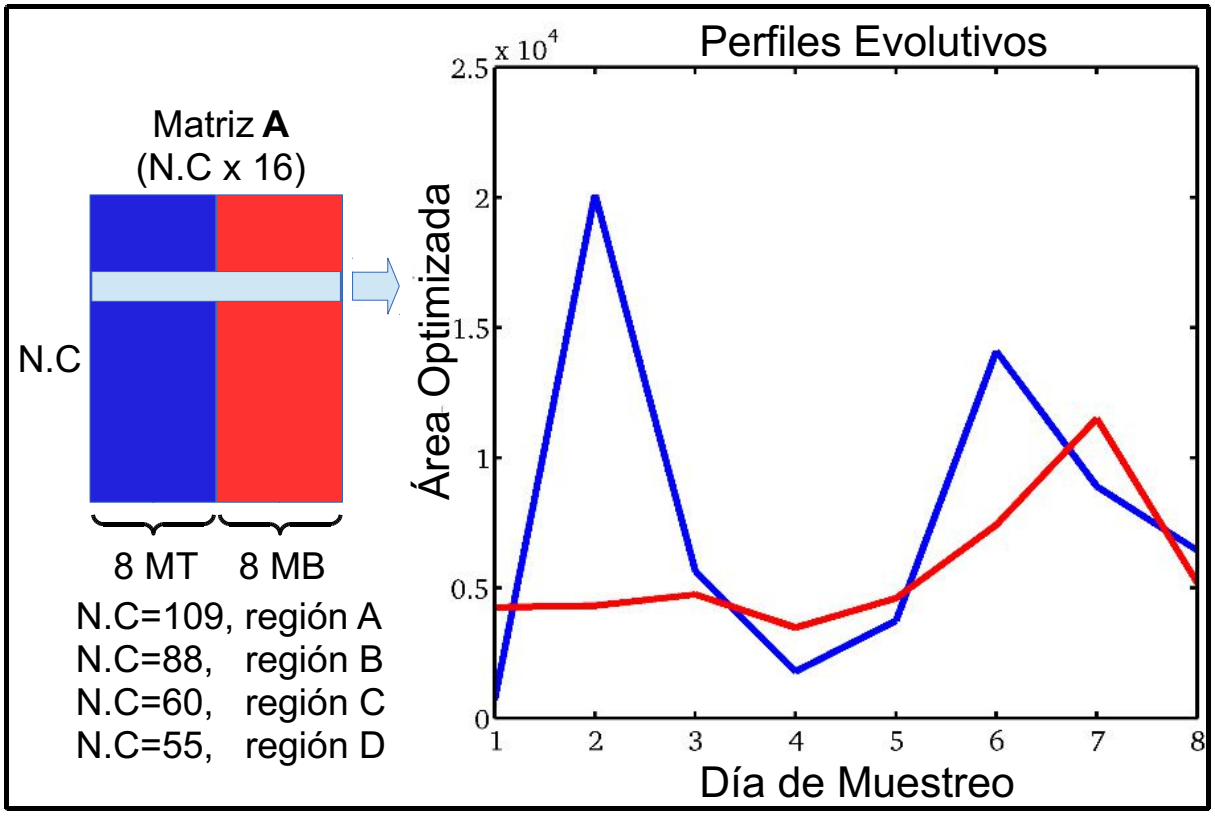


*Figura 15: Conformación de la matriz **A** de áreas bajo perfiles de concentración para una de las cuatro regiones (Izquierda) y gráfica respectiva a los 2 perfiles evolutivos seleccionados (Derecha)*

Referencias: N.C: Número de Componentes propios de cada región, MT (Azul): Muestras Tratadas, MB (Rojo): Muestras Blanco.

Específicamente, en cada submatriz de **A** (primeras 8 columnas para MT o últimas 8 columnas para MB) cada elemento a_{ij} se corresponde con el área del i-ésimo componente en el j-ésimo día de muestreo. Por ende, si se extrae una fila de **A** y se divide en 2 partes iguales, se tendrá una representación del perfil de evolución de un determinado componente en el tiempo, lo cual dejará

ver cómo varió la concentración de dicho componente a través de los 8 días de muestreo tanto en MT como en MB.

En base a lo anterior, se realizó una comparación exhaustiva de los perfiles evolutivos de cada componente en las MT y en las MB. Dicha comparación se llevó a cabo mediante el cálculo de un coeficiente de correlación de Pearson para cada par de perfiles evolutivos. Vale destacar que de esta manera se intentaron encontrar relaciones del tipo lineal, aunque bien podrían haberse buscado de otros tipos. Como cada coeficiente provee la correlación entre ambos tipos de perfiles, esto permite realizar una clasificación cualitativa de diferentes comportamientos observados. Ya que las MT y las MB fueron recolectadas bajo el mismo procedimiento y en un estado similar respecto de la maduración de los frutos, algunos de los comportamientos observados podrían postularse como debidos a la presencia/ausencia del pesticida.

En la figura 16 pueden observarse algunos ejemplos de perfiles evolutivos que resultaron ser similares tanto en términos cinéticos como en términos de los valores de las áreas resueltas. Los coeficientes de correlación cercanos a 1 indican un grado alto de similitud en la cinética evolutiva a través de los días de muestreo, mientras que la similitud entre valores resueltos sugiere que los niveles de concentración no se vieron severamente afectados. En este sentido, puede decirse que estos componentes (y otros no mostrados) no sufrieron cambios relevantes en su metabolismo natural, observado a través de los blancos, que pudieran relacionarse a un efecto de *stress* fisiológico derivado de la aplicación de Carbofurano. Cabe destacar que las similitudes cinéticas pueden ser mejor observadas en los gráficos insertos con escalado por máximos, pero esta información sería insuficiente para caracterizar los niveles de concentración en sí.

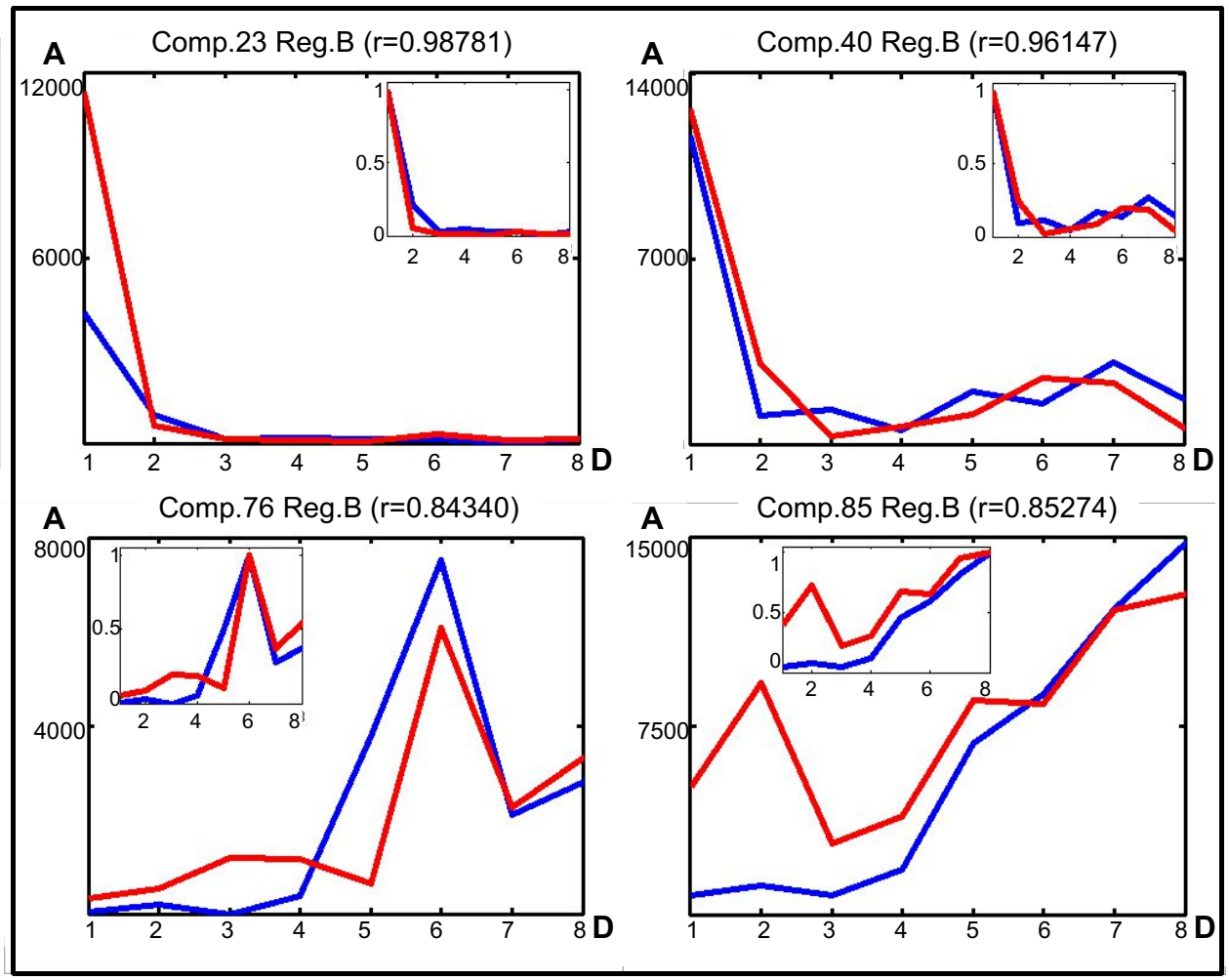


Figura 16: Perfiles evolutivos similares a través de los días de muestreo para algunos componentes resueltos en MT y MB, cultivar Rambo

Referencias: MT (Azul): Muestras Tratadas, MB (Rojo): Muestra Blanco, A: Área resuelta, D: día de muestreo, Comp.N: Nro de componente en su Región, Reg.X: Región X (A-D), r: coeficiente de correlación de Pearson entre ambos perfiles. Los rectángulos insertos en cada figura corresponden a cada perfil escalado por su máximo.

Con lo visto en el análisis anterior, es esperable que existan casos en los cuales los perfiles estén compuestos de valores en niveles de concentraciones apreciablemente diferentes, más allá de obtener coeficientes de correlación altos y por ende cinéticas similares. En estos casos, el efecto de la aplicación de Carbofurano estaría en una modificación de los niveles metabólicos normales, por exceso o por defecto respecto de los Blancos. La figura 17 ejemplifica algunos casos observados.

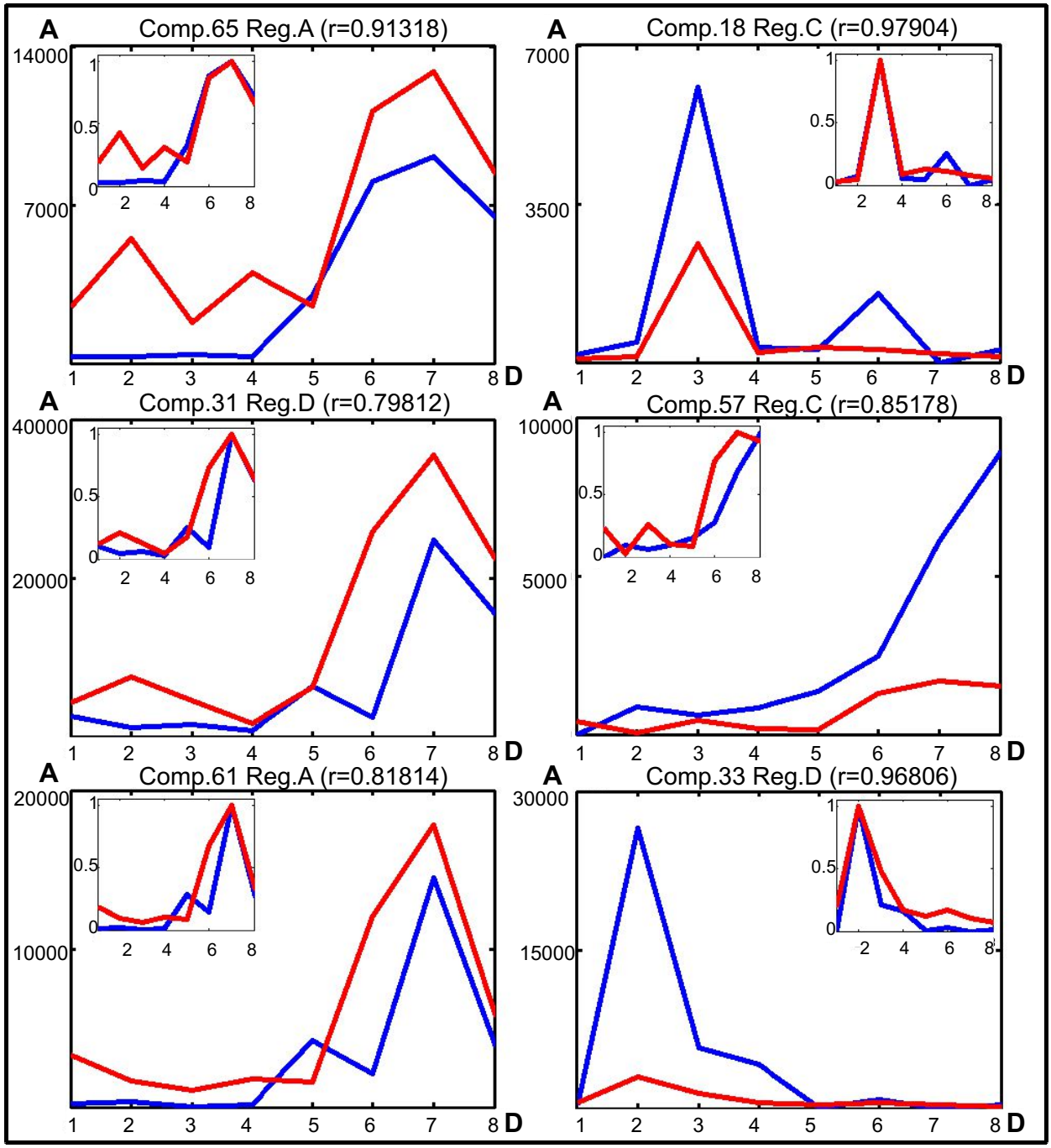


Figura 17: Perfiles evolutivos de algunos componentes con cinéticas similares pero con niveles de concentración diferentes, generalmente superiores en MB (Izquierda) o en MT (Derecha), cultivar Rambo

Referencias: MT (Azul): Muestras Tratadas, MB (Rojo): Muestra Blanco, A: Área resuelta, D: día de muestreo, Comp.N: Nro de componente en su Región, Reg.X: Región X (A-D), r: coeficiente de correlación de Pearson entre ambos perfiles. Los rectángulos insertos en cada figura corresponden a cada perfil escalado por su máximo.

En primer lugar, cabe aclarar que en la figura 17 no en todos los días de muestreo un componente resultó en niveles mayores en MB o MT exclusivamente. Sin embargo, el componente supuestamente mayoritario lo fue en al menos 5 de los 8 días de muestreo.

Las gráficas de la izquierda muestran una clara superioridad de los niveles en MB, más allá de algunos pocos puntos de muestreo. Si bien las cifras de correlaciones no resultaron tan altas, en los gráficos adjuntos se puede observar cierto nivel de similitud entre las cinéticas. Estas gráficas pueden dar a pensar que existieron vías metabólicas cuyas velocidades fueron modificadas, de forma tal que en las MT haya aumentado la velocidad en procesos catabólicos y/o disminuido en procesos anabólicos relacionados a la eliminación y síntesis, respectivamente, de cada componente en cuestión. Esta modificación de la velocidad no debe confundirse con lo que sería un metabolismo acelerado o retrasado en términos de días, lo cual será analizado posteriormente.

De forma inversa, los gráficos de la derecha muestran un comportamiento similar, solo que las MT presentan niveles mayores en general. En estos casos los coeficientes de correlación fueron generalmente superiores a los anteriores, lo cual queda también plasmado en las gráficas adjuntas, en las cuales puede observarse un mayor solapamiento de los perfiles escalados. Para este tipo de metabolitos, el tratamiento con pesticida podría haber causado aumentos de las velocidades de síntesis y/o bajadas en las de eliminación.

Por otro lado, muchos perfiles de evolución mostraron bajos coeficientes de correlación, lo cual usualmente indica comportamientos diferentes e independientes del tratamiento con pesticida. Sin embargo, cuando estos mismos perfiles fueron observados, pudo notarse que para algunos pares comparados se podría obtener una buena correlación si uno de los perfiles del par se trasladara respecto del otro, vistos ambos como 2 vectores horizontales de 8 elementos cada uno. Este tipo de movimientos relativos implica que de los 8 datos que se tienen para cada tipo de perfil, sólo N de ellos vaya a ser utilizado en la comparación. El valor de N depende de cuántos días de muestreo se traslada un perfil sobre el otro. Por ejemplo, si se traslada el perfil de MT 2 días respecto del de MB, las áreas de los días de muestreo 3 a 8 para MB y las áreas de los días de muestreo 1 a 6 para MT serán las comparadas, obviando al resto. Si el movimiento relativo es muy grande se corre el riesgo de realizar el cálculo con muy pocos datos participando y por lo tanto los coeficientes de correlación tendrían menor sentido empírico, por lo cual se realizaron comparaciones con movimientos de hasta 3 días. Ya que los Blancos representan lo que deberían ser los perfiles normales, se decidió conservarlos como parámetros fijos en los movimientos relativos, por lo cual cuando se indique un movimiento, se hará referencia a los perfiles de MT (si los valores son

negativos, el movimiento del perfil de MT debe ser concebido hacia la izquierda).

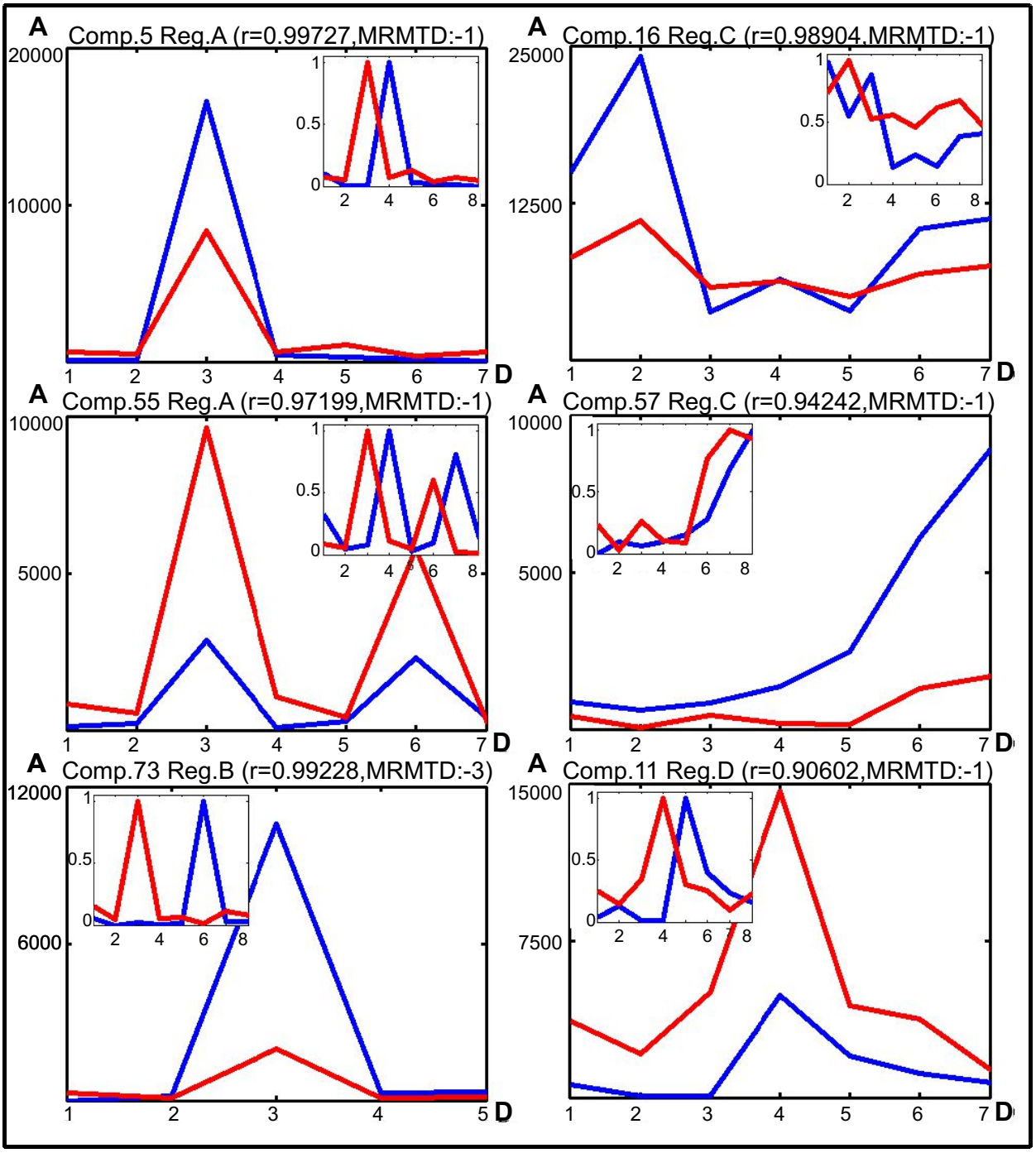


Figura 18: Perfiles evolutivos de componentes con metabolismos retrasados en MT respecto de MB, cultivar Rambo

Referencias: MT (Azul): Muestras Tratadas, MB (Rojo): Muestra Blanco, A: Área resuelta, D: día de muestreo, Comp.N: Nro de componente en su Región, Reg.X: Región X (A-D), r: coeficiente de correlación de Pearson entre ambos perfiles. MRMTD: Movimiento Relativo de Muestras Tratadas en Días. Los rectángulos insertos en cada figura corresponden a cada perfil escalado por su máximo y sin traslados.

La figura 18 muestra algunos ejemplos de traslados hacia la izquierda. Encontrar altas correlaciones significaría que las vías metabólicas afectadas podrían haber sido retrasadas. Como en figuras anteriores, los gráficos insertos en cada gráfica muestran a los perfiles originales escalados por sus máximos y sin movimientos relativos, con el objeto de mostrar cómo a simple vista son visibles los corrimientos. Los valores negativos de MRMTD indican cuántos días de muestreo han tenido que ser trasladados los perfiles de las MT para encontrar correlaciones aceptables con los de las MB. Por cuestiones de espacio no se muestran más ejemplos, pero los exhibidos son sólo unos pocos de los encontrados, los cuales también incluyen traslados de 2 días que no han sido ejemplificados. En general, los valores de correlación encontrados resultaron altos. Lo que pudo observarse con estos traslados es que en las MT suceden eventos que en las MB suceden antes, de lo cual se puede inferir que es posible que las vías metabólicas de los componentes seleccionados hayan sufrido un retraso luego de la aplicación del pesticida, pero sin modificar severamente la forma del perfil cinético. A su vez, los efectos de traslado no anulan la presencia de efectos vistos en gráficas anteriores. Por ejemplo, en las gráficas izquierda-media y derecha-inferior puede observarse que los niveles en MB son superiores en general, mientras que en la gráfica derecha-media sucede lo opuesto.

De manera similar pero opuesta a la anterior, en la figura 19 se presentan algunos casos observados con traslados de los perfiles evolutivos de MT hacia la derecha, por lo cual las cifras de MRMTD son positivas. En general, los valores de correlación encontrados resultaron aceptables, aunque en promedio levemente inferiores que los analizados previamente. Contrariamente a los últimos, con los movimientos relativos (y con otros no mostrados) pudo notarse que las cinéticas de los componentes en las MT pudieron haber sido adelantadas en el tiempo respecto de las MB luego de la aplicación de Carbofurano, sin que por esto cambie mucho la forma del perfil cinético. También como antes, los efectos de traslado no se presentan sólos, sino que efectos vistos en gráficas anteriores también son visibles en estos casos, como ser disparidades muy en favor de uno de los 2 perfiles comparados, o niveles de concentración muy similares, lo cual es notable en el ejemplo inferior de la izquierda. Como se ha visto, existen algunos comportamientos metabólicos que bien podrían provenir de la aplicación de un pesticida sobre frutos naturales. Las tendencias analizadas permiten pensar que efectivamente existen diferencias en la resolución de MT y de MB, por lo cual se desprende que a través del análisis de los perfiles evolutivos resueltos sería posible detectar si existió o no *stress* fisiológico debido al tratamiento. Esto podría realizarse con algoritmos de clasificación y de éstos podría obtenerse información relevante para la búsqueda de posibles

biomarcadores relacionados, lo cual será informado y discutido en la siguiente parte de este trabajo.

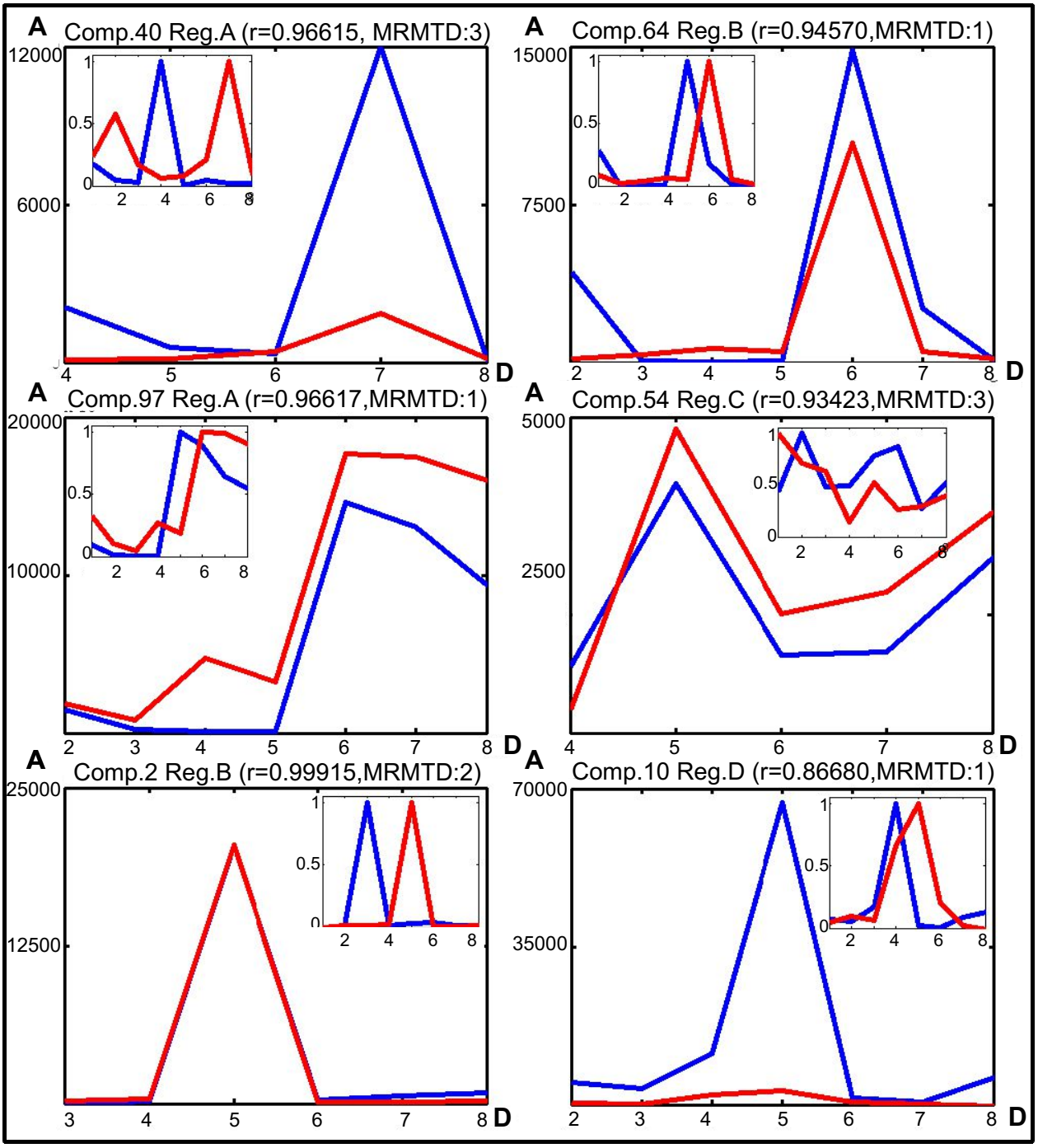


Figura 19: Perfiles evolutivos de componentes con metabolismos adelantados en el tiempo en MT respecto de MB, cultivar Rambo

Referencias: A: Área resuelta, D: día de muestreo, Comp.N: Nro de componente, Reg.X: Región X, r: coeficiente de correlación de Pearson entre ambos perfiles, MRMTD: Movimiento Relativo de Muestras Tratadas en Días. Los rectángulos insertos en cada figura corresponden a cada perfil original sin movimientos relativos, escalados por su máximo.

2.6.5 Análisis MCR-ALS de muestras en simultáneo: Parte 2

La descripción de lo hecho inicialmente es similar a la de la parte anterior, sólo que en este caso se utilizaron 2 cultivares distintos y frutos de todos los sectores. Para diferenciar a los 4 tipos de tomates, la nomenclatura fue la siguiente: R para Rambo Tratado, Rb para Rambo Blanco, F para RAF Tratado y Fb para RAF Blanco. Como cada sector de la zona de cultivo aportó 32 muestras, a razón de 1 de cada clase por día de muestreo durante 8 días, el total de muestras ascendió a 96. En todos los casos se utilizaron las matrices reducidas con WT. Las 96 matrices de dimensiones 127 × 178 fueron divididas en las regiones A, B, C y D, como ya se ha explicado. Por cada región se conformó un apilamiento manteniendo el espacio de columnas en común, con el siguiente orden desde arriba hacia abajo: 32 regiones del Sector A (8R, 8F, 8Rb, 8 Fb), 32 regiones del Sector B (8R, 8F, 8Rb, 8 Fb) y 32 regiones del Sector C (8R, 8F, 8Rb, 8 Fb). Lo anterior puede observarse en la figura 11 sobre la leyenda "Matriz Aumentada (una por región)". A su vez, cada grupo de 8 regiones (por ejemplo 8R u 8Fb) se mantuvo en el mismo orden que su día de recolección. Cada apilamiento resultó con su propia cantidad de filas, de cada uno se estimó la cantidad de componentes mediante SVD y con cada uno se aplicó SIMPLISMA para obtener estimaciones iniciales. Posteriormente, se ejecutó MCR-ALS con las restricciones ya discutidas y se obtuvieron los perfiles resueltos. Los resultados obtenidos se exponen en la tabla 5.

Matriz apilada	región A	región B	región C	región D
filas	3360	3360	2880	2592
ncomp	139	127	95	80
%LOF EXP	22.070	19.770	19.460	13.840
$\%R^2$	95.120	96.090	96.210	98.080
iter	13	12	32	15

Tabla 5: Detalles y cifras de mérito por región para MCR-ALS (Parte 2)

Referencias: ncomp: número de componentes en MCR-ALS, %LOF EXP: Porcentaje de Falta de Ajuste Experimental, $\%R^2$: Porcentaje de Varianza Explicada, iter: cantidad de iteraciones

Al igual que cuando se compararon los resultados de la Parte 1 con los obtenidos durante el estudio del efecto de la WT, en la tabla 5 puede apreciarse que el número de componentes estimados para explicar un 90% de la varianza por región siempre ha resultado mayor que el obtenido en la Parte 1. Esto pudo deberse fundamentalmente a dos cosas. La primera de ellas es que el número de muestras en estudio se incrementó 6 veces, por lo cual es probable que haya existido

nueva información sobre componentes variantes que antes no fue presentada. La segunda es que además se incluyó a la clase RAF, que seguramente debería tener metabolitos similares a los Rambo por ser ambos tomates, pero que a su vez debería tener los suyos en particular. La región que presentó el mayor aumento absoluto fue la B, con 39 componentes más que antes, mientras que la región C fue la que tuvo el mayor incremento relativo a lo que tenía anteriormente (aproximadamente 58,3%). Observando las iteraciones, las regiones B y D obtuvieron resultados similares a los de la primera parte. La región A se detuvo 7 iteraciones antes, y la C lo hizo 27 después, siendo que en los análisis anteriores era la región que normalmente convergía con rapidez.

Para todas las regiones se obtuvieron desmejoras en %LOF EXP y $\%R^2$, aunque estos últimos fueron todos mayores que 95%, superando el 90% establecido originalmente. Las regiones más afectadas en el aumento de %LOF fueron la A (aproximadamente 34%) y la B (50%). La región C fue la que sufrió de menor forma estos aumentos tanto en términos absolutos (0,55) como en términos relativos (aproximadamente 2,9%). Si se observan los $\%R^2$, los detrimentos parecen menos graves. Las regiones C y D apenas cayeron aproximadamente 0,2% y 0,5%, respectivamente. En resumen, en términos de ajuste la calidad fue inferior a la anterior en general, pero a su vez el sistema tiene más componentes diferentes, algunos de los cuales sólo deberían estar en uno de los tipos de cultivar y no en el otro, lo cual resulta en una situación más compleja y desafiante para resolver. No obstante, una comparación directa en ese sentido no parece estrictamente determinante, y lo más importante es la información que de esta nueva resolución pueda obtenerse.

2.6.5.1 Modelos de clasificación con PLS-DA: Generalidades

Luego de la resolución con MCR-ALS, las áreas bajo los perfiles de concentración de todos los componentes resueltos en las 4 regiones fueron compiladas en una matriz **A** de tamaño 441 × 96, es decir, “cantidad de componentes resueltos” × “cantidad de muestras disponibles”. El orden de las filas se correspondió con el de las regiones: componentes de la región A en la parte superior, debajo los de B, luego los de C y finalmente los de D. Por su parte, el orden en columnas situó a las 32 muestras del Sector A, luego a las 32 del Sector B y finalmente a las 32 del Sector C, teniendo en cuenta que cada Sector posee 4 tipos de muestras ordenadas: 8R, 8F, 8Rb y 8Fb . Esta matriz fue utilizada para obtener distintos modelos de clasificación.

La posibilidad de obtener modelos PLS-DA que pudieran diferenciar a las distintas clases de las muestras en juego resultó de interés por las siguiente razones:

- Un modelo para Blancos y Muestras Tratadas permitiría establecer una forma de evaluar si existieron efectos de *stress* debidos al tratamiento con pesticida. Lo anterior está directamente relacionado con los objetivos de este estudio .
- Aunque sin tener el mismo nivel de prioridad que el punto anterior, un modelo capaz de diferenciar tomates Rambo de RAF sería útil para verificar la autenticidad de estos cultivares, siendo la de RAF de mayor interés debido a que tienen un mayor valor y es posible adulterar productos derivados de RAF a través de su mezcla con derivados de otros cultivares de menor valor. Dicho modelo no se plantea para diferenciar frutos sin procesar, ya que sus formas son distintas y eso es suficiente para distinguirlos.

Para elaborar los modelos PLS-DA, un grupo de muestras fue seleccionado para ser parte del conjunto de calibración (CAL). Para esto, en el orden en que se encontraban tabuladas las 96 muestras, se tomaron 1 de cada 3. Como resultado, el conjunto de calibración resultó con 32 muestras y las restantes 64 formaron parte del conjunto de validación (VAL) con el que se pretendió evaluar las capacidades predictivas de los modelos. Vale destacar algunas cosas al respecto:

- Si bien el número de calibradores no es pequeño, el número de validadores lo duplica. Esto no suele ser muy común, ya que normalmente se suelen asignar más muestras a CAL que a VAL. En este sentido, la partición de los datos disponibles representa cierto desafío.
- Al tomar 1 de cada 3 muestras disponibles para CAL, este conjunto y VAL contienen muestras de todos las clases y Sectores de forma balanceada.
- Los mismos conjuntos de CAL y VAL fueron utilizados en todos los modelos, sin importar si éstos eran binarios (MB/MT o R/F) o cuaternarios (R/F/Rb/Fb).
- Según cada modelo, valores codificantes de clase fueron asignados a cada muestra en CAL:
  - En los modelos MB/MT y R/F, las clases fueron representadas en vectores **y** con ceros (MB y F) o con unos (MT o R), y en cada caso se aplicó el algoritmo de PLS1 para obtener modelos binarios de decisión.
  - En el caso del modelo de 4 clases (R/F/Rb/Fb) la codificación fue realizada utilizando una matriz **Y** compuesta de unos y ceros, con tantas filas como muestras en CAL y tantas columnas como clases disponibles. En este sentido, cada fila de **Y** contendría solamente un valor igual a 1 en la columna que indicara la clase de la muestra en cuestión. Para este modelo, el algoritmo utilizado en la etapa de calibración fue PLS2.

- Cuando fue necesario, las clases R, F, Rb y Fb fueron codificadas con los enteros 1, 2, 3 y 4, respectivamente.

A su vez, se recuerdan cuestiones comunes:

- Todos los modelos fueron obtenidos con y sin selección de variables (componentes), según el procedimiento ya explicado
- Sólo se aplicó centrado como preprocesamiento a las matrices **X** de CAL y a sus respectivos vectores/matrices **y**/**Y**. Cuando se realizaron predicciones sobre los datos en la matriz VAL pertinente, éstos datos fueron previamente centrados con la media de CAL respectiva.
- La selección del número de Variables Latentes (LV) para cada modelo se realizó evaluando la curva RMSECV en función de LV, seleccionando como número de LV óptimo a aquel que presentara el primer mínimo, como ya se ha explicado. Debe recordarse que en los modelos cuaternarios dicho valor correspondió al primer mínimo de la clase peor predicha.

La tabla 6 expone un resumen de lo obtenido con los distintos modelos PLS-DA.

	Sin selección de componentes			Con selección de componentes		
	R/F	MB/MT	4 clases	R/F	MB/MT	4 clases
Clases	(R+Rb) (F+Fb)	(Rb+Fb) (R+F)	(R)(F) (Rb)(Fb)	(R+Rb) (F+Fb)	(Rb+Fb) (R+F)	(R)(F) (Rb)(Fb)
LV	3	2	6	3	2	5
NC	441	441	441	41	63	127
CCC	32/32 (100%)	32/32 (100%)	32/32 (100%)	31/32 (96.88%)	32/32 (100%)	31/32 (96.88%)
CCV	60/64 (93.75%)	63/64 (98.44%)	52/64 (81.25%)	58/64 (90.63%)	63/64 (98.44%)	57/64 (89.06%)

Tabla 6: Resultados obtenidos de diferentes modelos PLS-DA, con y sin selección de componentes, a partir de las matrices de Áreas resueltas

Referencias: LV: Variables Latentes en PLS-DA, NC: Número de Componentes en el modelo, CCC: Clasificaciones Correctas en Calibración, CCV: Clasificaciones Correctas en Validación . Las letras entre paréntesis indican las clases que son utilizadas como una única clase en cada modelo

En la tabla 6 puede apreciarse que el proceso de selección de componentes redujo significativamente la cantidad de éstos activos en los modelos, dejando cerca de 10%, 15% y 30% de los originalmente disponibles para los modelos “R/F”, “MB/MT” y “4 clases”, respectivamente.

A la vista de los resultados de CCC y CCV, el procedimiento pareció influir de manera más relevante en los modelos de 4 clases, provocando al mismo tiempo una disminución en la cantidad de LV necesarias. Los modelos con menos variables, originales y/o latentes, son considerados más parsimoniosos y se supone que por depender de menos componentes variantes también suelen ser más robustos.

También puede observarse en la tabla 6 que todos los modelos presentan resultados muy buenos y similares para el conjunto de calibración respectivo, lo cual queda plasmado en las altas cifras de CCC. Esto debe interpretarse con cautela, ya que dichos resultados podrían provenir de situaciones de sobreajuste a los calibradores, una situación muy común cuando el modelado se realiza sin demasiados cuidados.

Si se presta atención a los modelos binarios, y en especial a los de MB/MT, se puede apreciar que con o sin selección de componentes se han requerido solamente 2 LV para contemplar las diferencias entre clases. Si a su vez se observan las cifras de CCV, las cuales han resultado ser las más altas logradas, y si se contempla que una de las causas normales de sobreajuste a los calibradores se da cuando se opta por un número de LV alto, donde las LV de mayor orden son dedicadas a modelar detalles de los calibradores poco relevantes para representar tendencias generales, se puede inferir que en los modelos MB/MT es poco probable que haya existido sobreajuste.

Mayores detalles de cada uno de los modelos se exponen a continuación.

2.6.5.2 Modelos de clasificación con PLS-DA: 4 clases

La hipótesis de sobreajuste inexistente no sería válida para otros modelos, en especial para el de 4 clases sin selección. Éste obtuvo 100% en CCC, utilizó el mayor número de LV y predijo mal casi el 20% de VAL, con lo cual es más probable que haya existido sobreajuste. Si este modelo es comparado con su par con selección de componentes, puede notarse que se utilizó una LV menos y que se cometió un error en la predicción de CAL, pero a costa de esto aumentó la cifra de CCV, con lo cual el modelo parece ser más generalista. En promedio, los modelos cuaternarios obtuvieron resultados peores los otros modelos. Una razón por la cual el desempeño de los modelos de 4 clases fue menor podría ser que el número de muestras representando a cada clase distinta en CAL no hubiese sido suficiente para exponer las diferencias entre clases, pero si este hubiese sido el caso, tampoco se hubiesen obtenido buenos resultados en CAL. Más aun, aunque no se muestra el detalle, las muestras mal predichas en VAL fueron todas MT, por lo que puede suponerse que las clases Rb

y Fb estaban lo suficientemente representadas en CAL con sólo 8 muestras por clase. Probablemente el menor desempeño de los modelos cuaternarios haya provenido del hecho de que el modelado con PLS2 tiene la dificultad extra de que debe seleccionarse un único número de LV como óptimo para todas las clases. Para estos modelos, se optó por predecir de la mejor manera posible a la peor clase predicha según RMSECV, aunque esto llevó a un aumento del número de LV sin mejorar con esto los resultados en VAL. No obstante a las críticas realizadas a PLS2, se destaca que el algoritmo es capaz de obtener ventajas de posibles correlaciones entre variables en **Y** (Galtier y col., 2011). Lo anterior suele no ser tan llamativo cuando las clases son totalmente independientes, pero a diferencia de esto, en nuestro caso las clases efectivamente mantienen ciertas relaciones. Por ejemplo las muestras R y Rb, tomadas como clases distintas, deberían tener propiedades en común derivadas de su naturaleza Rambo. Por ende, más allá de obtener estos resultados con modelos de 4 clases, no resulta tan inapropiada la aplicación de PLS2.

En la figura 20 pueden compararse ambos modelos cuaternarios y pueden observarse los valores predichos para cada muestra en juego antes de su clasificación definitiva. Los círculos que representan valores predichos para los calibradores son los que dan origen a los intervalos de confianza (en líneas angostas) alrededor de la media (en líneas gruesas) predicha en CAL para cada clase. Los intervalos fueron obtenidos sumando y restando a la media el valor $ts/n^{1/2}$, siendo n el número de muestras que dieron origen a cada media (8), t el coeficiente de Student para n-1 grados de libertad (95% de confianza) y s la desviación estándar respectiva. Estos intervalos son solamente indicativos de la media y dispersión de los calibradores, pero no son determinantes a la hora de clasificar una muestra. Es decir, aun cuando una muestra esté muy por fuera del intervalo de su clase, esto no significa que la muestra será mal clasificada, ya que la clasificación se lleva a cabo por otro mecanismo que implica estadística Bayesiana y la superación de umbrales de clase. Más aun, el que una muestra se encuentre dentro del intervalo de otra clase tampoco significa que esa será su predicción, pues también es posible que el modelo dictamine que algunas muestras son inclasificables. Aun con lo anterior, puede apreciarse que la extensión de los intervalos y su posición no han sido exactamente las mismas si se observa una misma clase en ambos casos. La clase Fb (4) posee resultados similares, aunque viendo las proyecciones puede observarse que el proceso de selección de componentes produjo una separación de las muestras en CAL, lo cual repercutió luego en las predicciones de las muestras en VAL de igual forma. Observando la clase Rb (3), el modelo sin selección agrupó a las muestras en CAL en el menor de todos los intervalos, mientras que el otro modelo las mantuvo más dispersas. Esto da a entender que la utilización de las

áreas de todos los componentes para predecir las clases repercutió en una homogeneización de las muestras que las representaban, mientras que la selección de componentes impuso mayores diferencias entre éstas muestras.

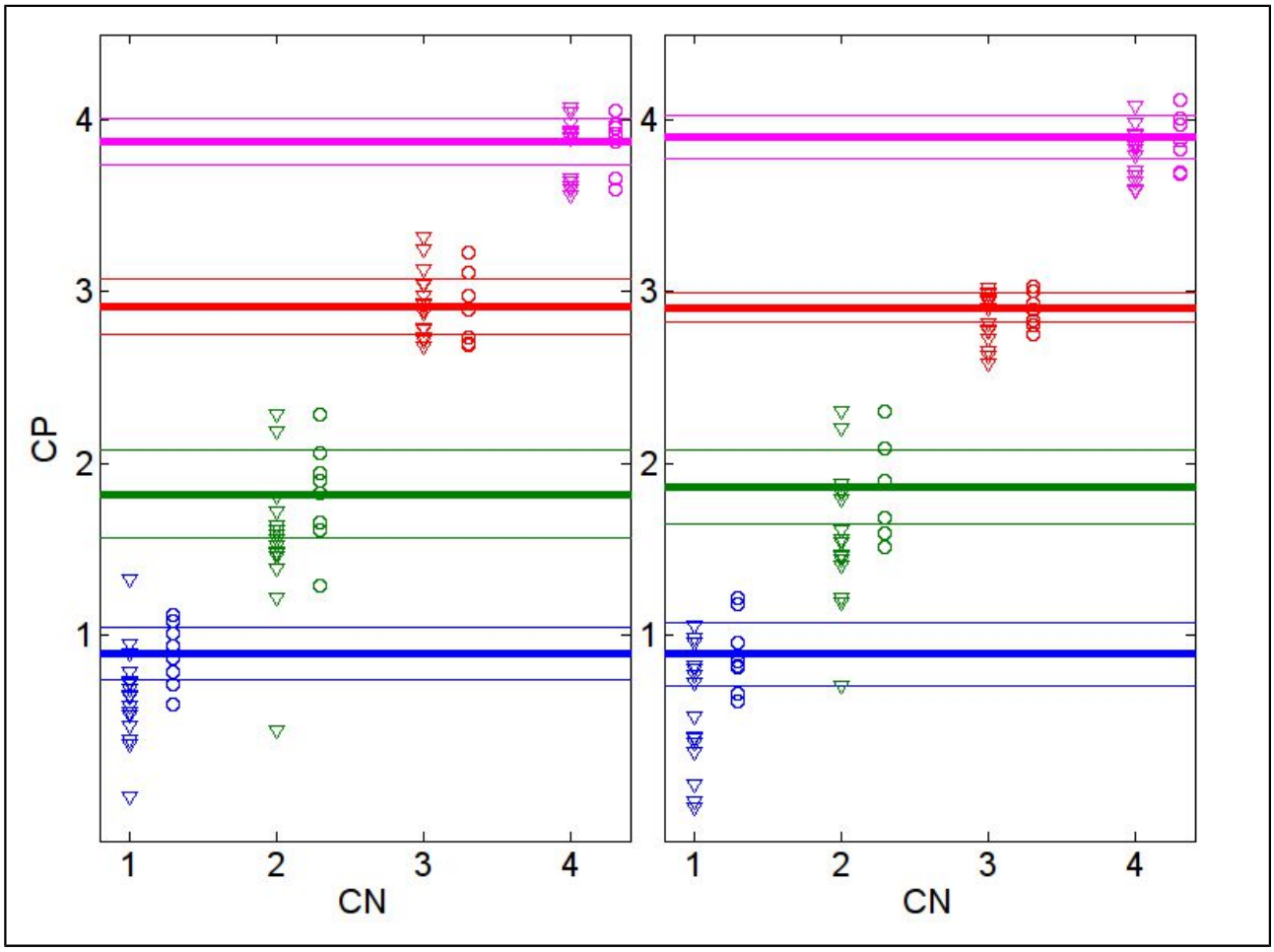


Figura 20: Representación de las predicciones de clase para los modelos PLS-DA cuaternarios con (izquierda) y sin (derecha) selección de componentes

Referencias: Azul: R(1), Verde: F(2), Rojo: Rb(3), Magenta: Fb(4), Triángulos: muestras en VAL, Círculos: muestras en CAL, CN: Clase Nominal, CP: Clase Predicha. Los círculos de CAL fueron trasladados hacia la derecha de sus respectivos triángulos de VAL para mayor claridad. Las líneas de color gruesas señalan datos promedio y las finas delimitan un intervalo de confianza (ver texto)

Las 2 clases de MB descriptas no presentaron ninguna clasificación incorrecta, por lo que en este sentido, el proceso de selección no ha sido demasiado relevante. Por otro lado, dentro de las MT se encontraron todos los errores de predicción. La clase F (2) deja ver a ambos tipos de calibradores muy dispersos y lo mismo se observa en VAL. Sin embargo vale destacar que en la

gráfica con selección, la muestra en CAL que se presenta más abajo que el resto es el único calibrador mal predicho, lo cual representó finalmente una ventaja por cuanto el modelo tuvo mejores aptitudes de generalización, ya que tuvo sólo 3 errores en VAL, situándose 2 de ellos por debajo de la muestra de CAL señalada. En cambio en el modelo sin selección existieron 5 predicciones incorrectas en VAL, todas ellas ubicadas por debajo de la muestra de CAL más baja de ese conjunto. Finalmente, en la clase R (1) los intervalos fueron similares, pero hubo menos errores en VAL con selección (4) que sin ésta (7). En ambos casos, todos los errores se encontraron en muestras que se proyectan por debajo de la muestra de CAL inferior respectiva, aunque lo contrario no sea cierto, es decir, no todas las muestras de VAL observadas por debajo de su CAL inferior respectiva han sido mal predichas.

Habiendo descripto algunos aspectos de los modelos cuaternarios, cabría la posibilidad de analizar cómo fueron proyectadas todas las muestras en los planos conformados por las distintas LV, pero siendo que éstas son 5 y 6, los modelos son complejos y no resulta de utilidad ver las mencionadas proyecciones. Se concluye que estos modelos resultaron menos eficaces y a su vez que requieren más cuidado en su análisis y elaboración, fundamentalmente en el acuerdo del número de LV utilizadas para modelar a todas las clases en simultáneo.

2.6.5.3 Modelos de clasificación con PLS-DA: Muestras Blanco/Muestras Tratadas

Más allá del hecho de que la selección de componentes no afectó de manera radical el desempeño de ninguno de los modelos realizados, no son sólo esas cifras las que pueden brindar información y algunos otros aspectos deberían ser tenidos en cuenta.

La figura 21 muestra la distribución de *scores* de todas las muestras puestas en juego, para los modelos MB/MT, en los planos compuestos por sus respectivas LV. Ante todo, vale considerar que la varianza capturada en **Y**, que no es otra cosa que la información de clases a partir de los calibradores, alcanzó valores altos y similares para los dos modelos, siendo levemente superior para el que utilizó selección de componentes. Este modelo, a su vez, capturó 57.25% de la varianza en **X**, mientras que su homólogo sin selección sólo alcanzó el valor de 20.16%. No obstante, aunque las vcX se mostraron bajas, los resultados de las clasificaciones fueron muy buenos.

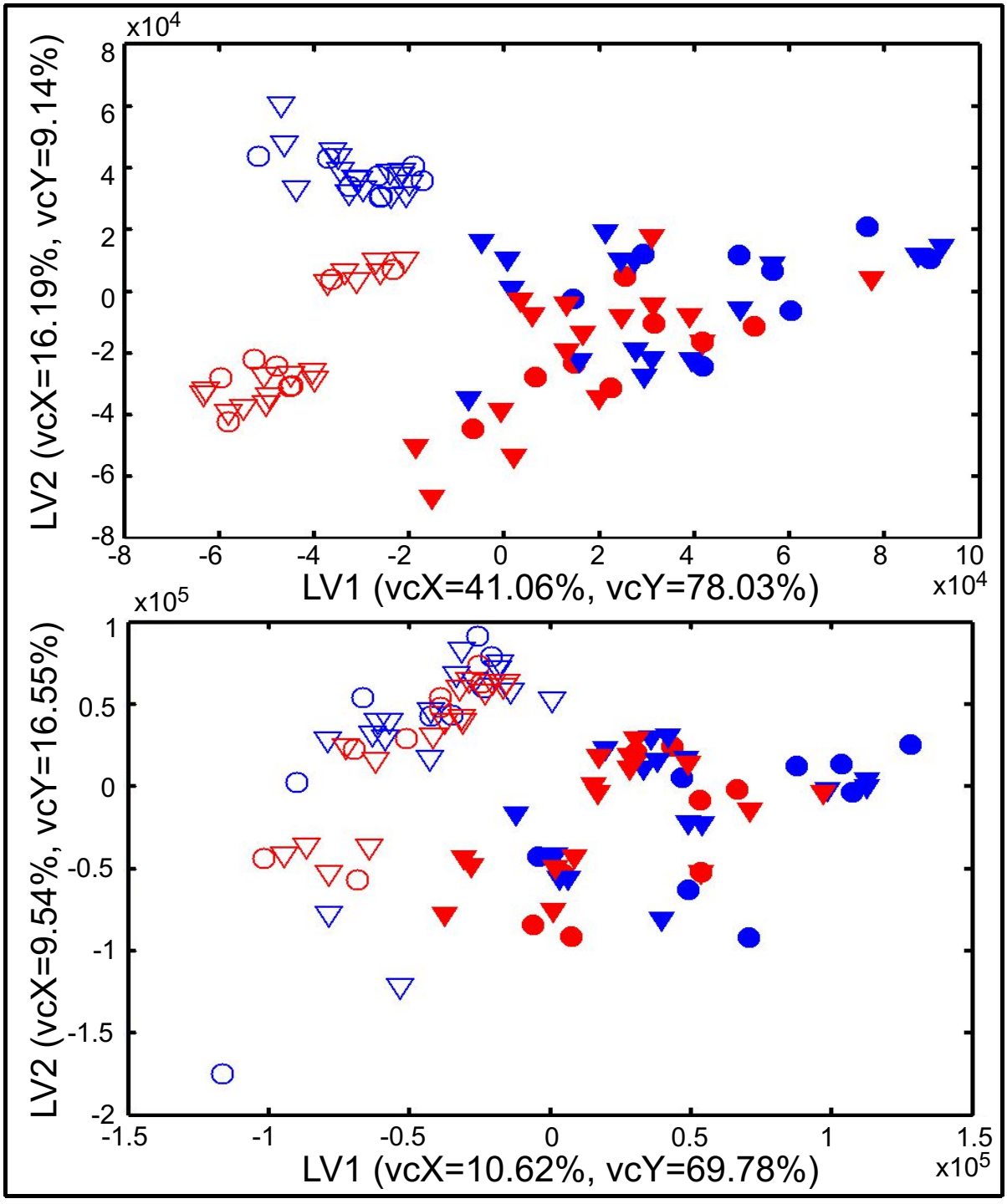


Figura 21: Distribución de scores para muestras de Calibración (círculos) y Validación (triángulos) en modelos PLS-DA para MB/MT, con (arriba) y sin (abajo) Selección de Componentes

Referencias: LVn: Variable Latente n, vcX: Varianza Capturada en **X**, vcY: Varianza Capturada en **Y**, Símbolos rellenos: MT, Símbolos vacíos: MB, Azul: Rambo, Rojo: RAF

En la figura 21 se aprecia que para ambos modelos las MB se sitúan en la parte izquierda de los planos, con valores de LV1 fundamentalmente negativos, mientras que las MT lo hacen a la derecha. Estas separaciones son coherentes con el hecho de que en la LV1 se modeló gran parte de la varianza correlacionada a la información de clase. Si se analiza la LV2, se puede observar que la selección de componentes produjo una separación entre muestras Blanco de ambos cultivares, aun cuando no se proveyó este tipo de información durante la construcción del modelo, ya que en esta etapa las MB representaron a un único grupo independientemente de su naturaleza R o F. A su vez, el proceso de selección en sí tampoco contó con datos de este tipo. Entonces, bajo este punto de vista, la selección de componentes parece haber preservado características naturales y no modeladas de las MB. Conclusiones similares fueron obtenidas en (Westerhuis y col., 2008), donde además se dijo que aunque las gráficas de este tipo (planos de *scores* en las LV) no deberían ser utilizadas para inferir separación entre clases porque podrían existir problemas de sobreajuste, efectivamente podrían revelar estructuras (subgrupos) dentro de las clases, ya que si el modelo no es forzado a contemplar estas diferencias, como fue en nuestro caso, estos resultados no podrían provenir de sobreajustes. La misma situación no se aprecia en la gráfica sin selección de componentes, y aunque los desempeños fueron similares, los niveles de interpretación no son los mismos, lo cual no es un dato menor. También debe notarse que en ambos modelos no se produjo una separación clara de las MT en términos de naturaleza R o F. Esto podría deberse a que el tratamiento con pesticida podría provocar una homogeneización parcial, en el sentido de metabolitos presentes o mayoritarios, en las MT. Como resultado, las MT parecen ser no diferenciables entre sí, al menos a nivel visual. La hipótesis de homogeneización parcial debida al tratamiento se muestra de acuerdo con lo visto en el análisis de los modelos de 4 clases, donde se observó que todas las muestras en VAL mal predichas fueron siempre MT y que no existieron confusiones para los blancos de ambos cultivares. A su vez, cuando se realizó selección de componentes para los modelos R/F, el número de variables conservadas fue menor que en las restantes selecciones (63 para MB/MT, 121 para 4 clases), aun cuando se utilizaron los mismos parámetros o restricciones. Lo último sugiere que en los modelos R/F menos componentes podrían proveer perfiles de concentración (a través de las áreas bajo estos) que pudieran ser considerados estadísticamente diferentes o, dicho en otras palabras, que el tratamiento con Carbofurano llevaría a una situación en la cual muchos de los metabolitos se comportan similarmente entre ambos cultivares.

La figura 22 expone información sobre los valores de *loadings* en cada LV, así como también el vector de regresión para el modelo MB/MT con selección de componentes.

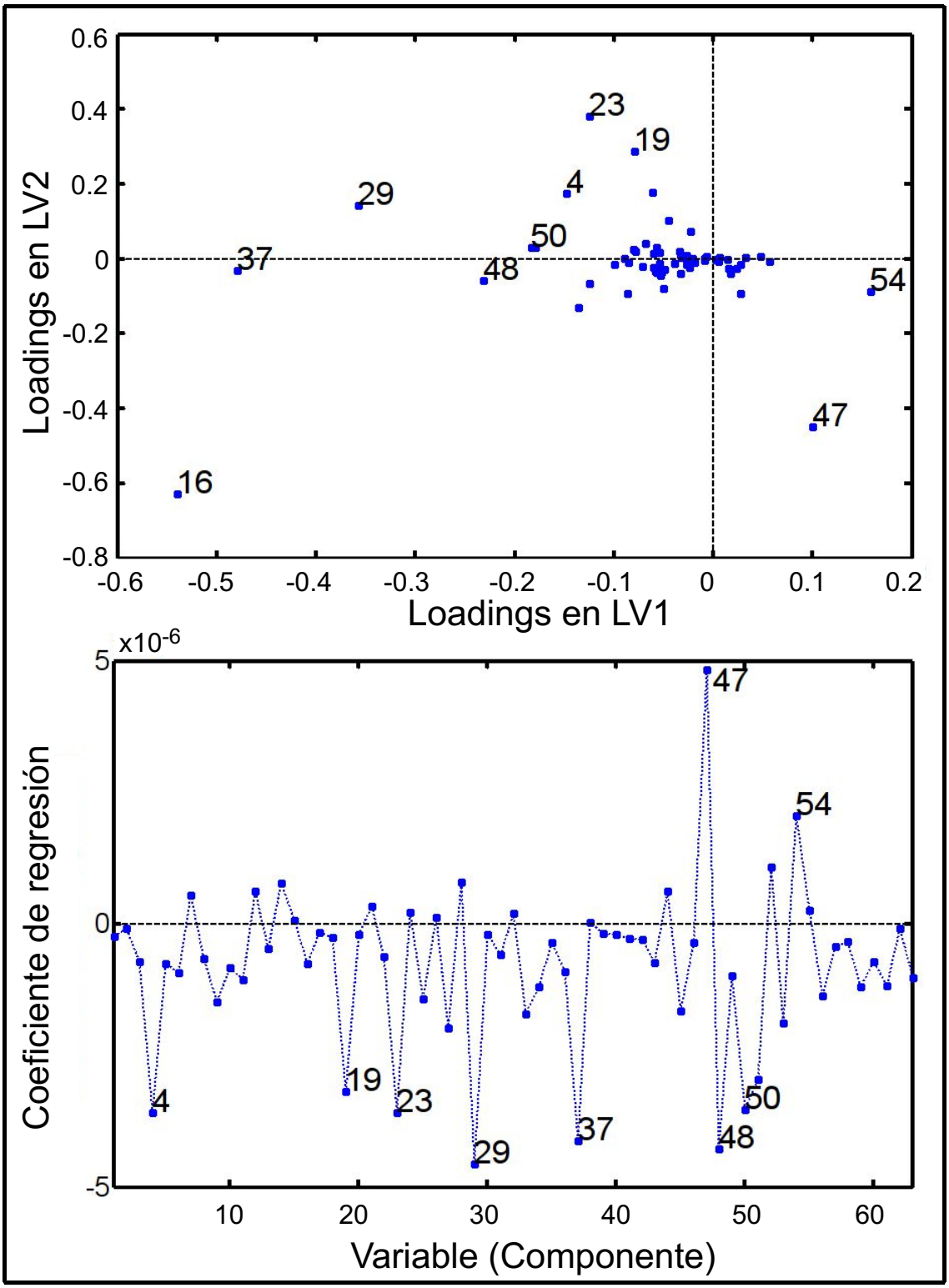


Figura 22: Gráfica de loadings (arriba) y vector de regresión (abajo) para el modelo MB/MT con selección de componentes

Referencias: Los números dentro de cada gráfica señalan componentes potencialmente relevantes en el modelado (ver texto)

La figura 22 expone la importancia relativa de cada componente en el modelo MB/MT con selección. En la parte superior se destacaron *loadings* de componentes que, por no estar cercanos al origen de coordenadas del mencionado gráfico, se supone que han sido relevantes durante la etapa de modelado. En la parte inferior pueden verse básicamente los mismos componentes destacados en el vector de regresión del modelo. En ambos gráficos, los puntos numerados podrían ser considerados como potenciales biomarcadores responsables de la diferenciación de clases. Debe tenerse en cuenta que estos gráficos provienen de información solamente relacionada al conjunto CAL, por lo cual no todos estos componentes necesariamente presentarán importancia al analizar las muestras en VAL. Para corroborar lo anterior, la figura 23 expone las áreas resueltas con MCR-ALS en todas las muestras para algunos de estos componentes numerados. En dicha figura debería notarse que las muestras 1-32, 33-64 y 65-96 corresponden a diferentes sectores (A, B y C, respectivamente). Como se puede observar, los resultados suelen ser diferentes entre sectores, por lo que es cuestionable que éstos sean interpretados como replicados entre sí. A primera vista, los sectores A y B muestran resultados más similares entre sí que con respecto a los del sector C. A su vez, las diferencias vistas son de distintos tipos. Por ejemplo, la gráfica del componente 4/83 deja ver que en los sectores A y B las MT tienen generalmente menores áreas resueltas que sus respectivas MB, mientras que en el sector C los blancos RAF se encuentran en niveles similares a ambos tipos de MT. Otro ejemplo de diferencia de niveles puede ser visto en la gráfica del componente 19/190, donde todos los sectores muestran a las muestras de la clase Rb separadas del resto, pero en el sector C la diferencia no es tan notoria como en los sectores A y B. Por otro lado, en el sector C se observaron comportamientos invertidos. Ejemplo de lo anterior son las gráficas inferiores, donde en ambos componentes puede verse que en dicho sector las muestras Rb poseen áreas mayores que las Fb, mientras que lo opuesto sucede en los otros sectores. No obstante a este evidente problema de diseño, el modelo ha presentado robustez en las clasificaciones. Esto último probablemente sea debido a que el conjunto de calibración fue obtenido a partir del agrupamiento de todas las muestras en juego, es decir, incluyendo información de los 3 sectores. La misma robustez no podría esperarse si, por ejemplo, el modelo hubiese sido obtenido a partir de las muestras del sector A, actuando las de los sectores B y C como validadores. Finalmente, aunque es evidente que cada sector tuvo sus particularidades, puede notarse en todos que, habiendo seleccionado un determinado componente, las muestras Rb están generalmente separadas de las Fb. Esto rememora lo visto en las gráficas de *scores*, donde en el modelo para MB/MT con selección de componentes se observaba la diferencia de blancos aunque no hubiese sido modelada.

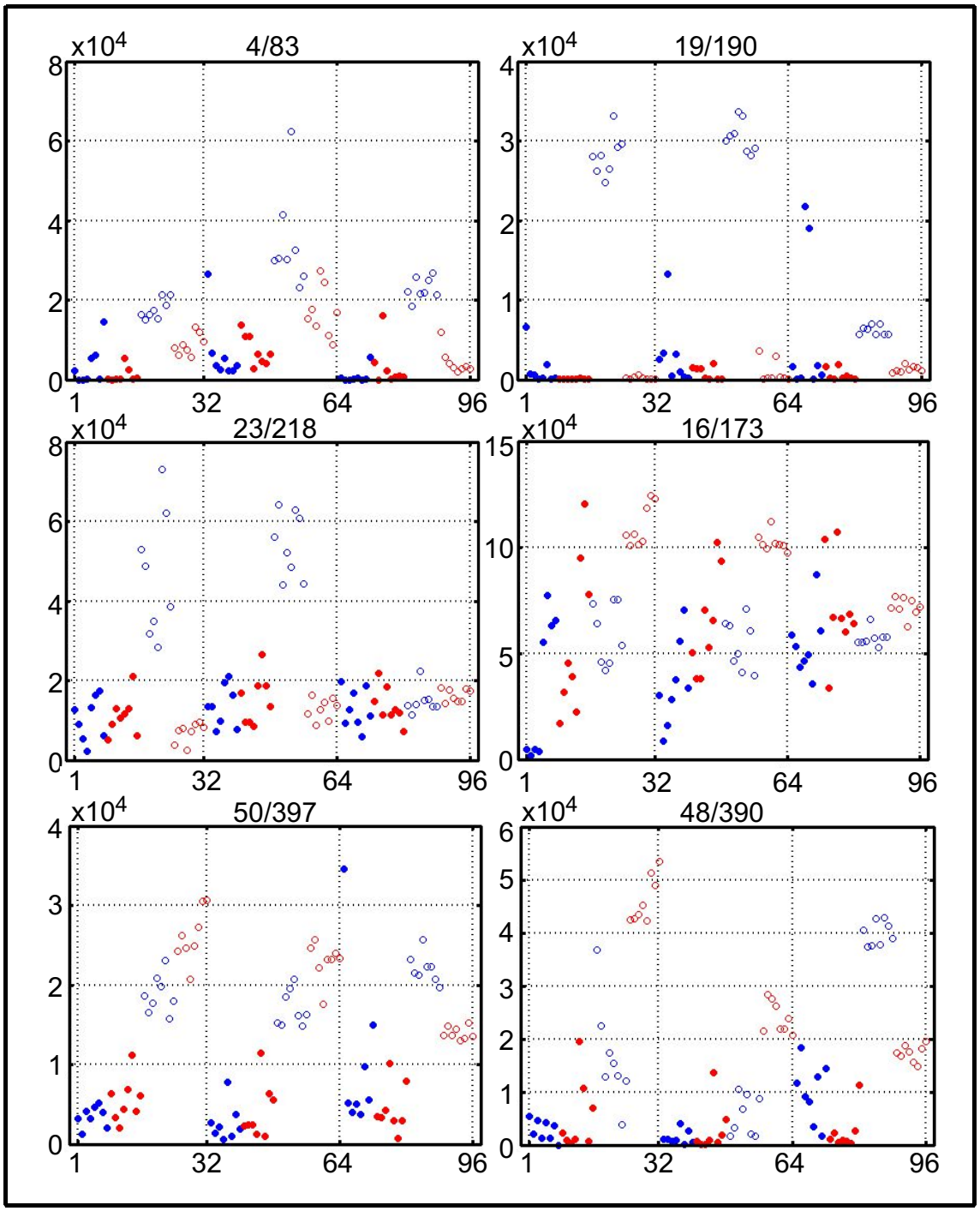


Figura 23: Áreas resueltas en las 96 muestras disponibles para algunos componentes seleccionados en base a su aporte al vector de regresión del modelo MB/MT con Selección de Componentes

Referencias: Símbolos rellenos: MT, Símbolos vacíos: MB, Azul: muestras Rambo, Rojo: muestras RAF. Los valores observados en el eje de las abscisas delimitan a las muestras de cada sector y en cada uno las muestras de cada clase se encuentran graficadas según su día de recolección. Los números del tipo A/B en la parte superior de cada gráfico indican el orden de componente en la selección (A) y el orden del mismo componente entre los 441 originales (B).

Ya que la selección de componentes impuso que la media de MB en CAL sea distinta a la de las MT en CAL, es lógico esperar que esto pueda ser intuido desde las gráficas expuestas en la figura 23. A su vez, todos los ejemplos mostrados corresponden a componentes que en el gráfico de *loadings* se situaban en la parte izquierda, y debajo en el gráfico del vector de regresión. Si se recuerda cómo eran proyectadas estas muestras en la gráfica de *scores* de la figura 21, debe entenderse que en los ejemplos expuestos las áreas de las MB son mayores que las de MT a nivel promedio (de los 63 componentes selectos disponibles, 49 de ellos presentaron esta característica), aunque esta diferencia debió haber sido clara en CAL pero no necesariamente en VAL. Es decir, estos metabolitos estarían siendo suprimidos con el tratamiento con pesticida, sea por una menor síntesis o por un aumento de lisis. Y aunque se desconoce la cinética del Carbofurano durante los días de muestreo (se hablará de esto más adelante), puede deducirse que los ejemplos mostrados no podrían corresponder a este compuesto, pues no sería posible que las MB tuvieran niveles mayores que las MT.

Otro detalle a tener en cuenta es la fuente de datos que ha originado la diferencia de medias entre MB y MT. En las gráficas de los componentes 50/397 y 48/390 fundamentalmente se puede observar que ambos tipos de MB (Rb y Fb) se encuentran en niveles superiores a las MT, con excepción del sector B para el componente 48/390 y de algunos puntos aislados. A nivel general dentro de lo expuesto en la gráfica, estos 2 componentes serían los de mayor potencial como marcadores de *stress* debido al tratamiento. En cambio, para los componentes 19/190 y 23/218 (exceptuando el sector C) los blancos Rambo fueron los que presentaron niveles notablemente mayores, mientras que para el componente 16/173 los blancos RAF se diferenciaron del resto. Es decir, en algunos casos sólo un miembro del par que conformó la clase MB presentó mayores diferencias respecto de las MT. Esto último, visto fuera de contexto, daría a entender que esa información no sería suficiente para diferenciar MB de MT, o que cierto tipo de MB podrían ser confundidas con MT. Sin embargo, cualquier modelo PLS-DA surgirá de combinar los aportes de diferentes variables (componentes) con el objeto de maximizar la covarianza con la información de clases. Obviamente, algunas de estas variables realizarán aportes más “puros”, mientras que otras lo harán de manera parcializada pero con sinergia potencial una vez que todas sean combinadas en un modelo. En los casos en los que sólo una clase de blanco haya mostrado diferencias significativas en un determinado componente, se podría pensar también que el metabolismo de dicho componente sólo en tomates de ese mismo cultivar fue afectado con el tratamiento, ya que la otra clase de blanco estaría en un nivel similar al de su respectiva muestra tratada, de lo que podría deducirse que el

pesticida afecta de manera diferente vías metabólicas de un mismo componente (espectros resueltos en común), dependiendo del tipo de cultivar.

Se hace hincapié en el hecho de que los componentes con números resaltados en el vector de regresión han resultado relevantes a partir de la información de CAL solamente. Si se observan uno a uno todos los componentes dentro de los 63 seleccionados, se pueden observar separaciones entre MB/MT tanto en CAL como en VAL más claras que las expuestas. Es decir, existen componentes mejores para la diferenciación, pero estos componentes en las muestras de CAL no han resultado ser precisamente los más significativos según los criterios de ajuste de PLS-DA. No obstante, este riesgo se corre siempre que uno elije a un conjunto de muestras para representar a todo un universo de estudio y a ciertos pretratamientos y algoritmos para obtener conclusiones de dicho universo a partir de los datos conocidos.

También se remarca que PLS-DA pone énfasis en la correlación con la información de clases, aún a costa de explicar menos varianza a nivel de señales (áreas en este caso). Si se comparan los componentes expuestos en las gráficas en cuanto a sus niveles de área, puede notarse que no todos alcanzan los mismos valores. No obstante, lo último se relaciona a la varianza a nivel de señales, no de clases, y por consiguiente para PLS-DA esos valores en sí pueden resultar no tan relevantes, ya que un buen marcador para este algoritmo sería aquel cuyos niveles de áreas pueden ser altos o bajos, pero siempre correlacionados con los valores de clase. Al respecto, estos últimos han sido 1 para MT y 0 para MB, algo ciertamente muy estricto, ya que áreas quizá muy distintas en MB han sido correlacionadas al mismo tiempo con el valor 0, y de la misma forma áreas dispares en MT han sido forzadas a tener correlaciones con el valor 1. Esto no es evitable (aunque se elijan otros valores codificantes el problema persistiría) y proviene del hecho de adaptar un algoritmo de uso cuantitativo como PLS a operaciones cualitativas como estas clasificaciones.

Prosiguiendo con el análisis de componentes, corresponde ahora evaluar a aquellos que tienen medias de área mayores en MT que en MB, los cuales han sido 14 en total (de los 63 posibles). Dos de estos han sido destacados en el vector de regresión con valores positivos y en el gráfico de *loadings* fueron observados a la derecha. La figura 24 expone los resultados para estos componentes.

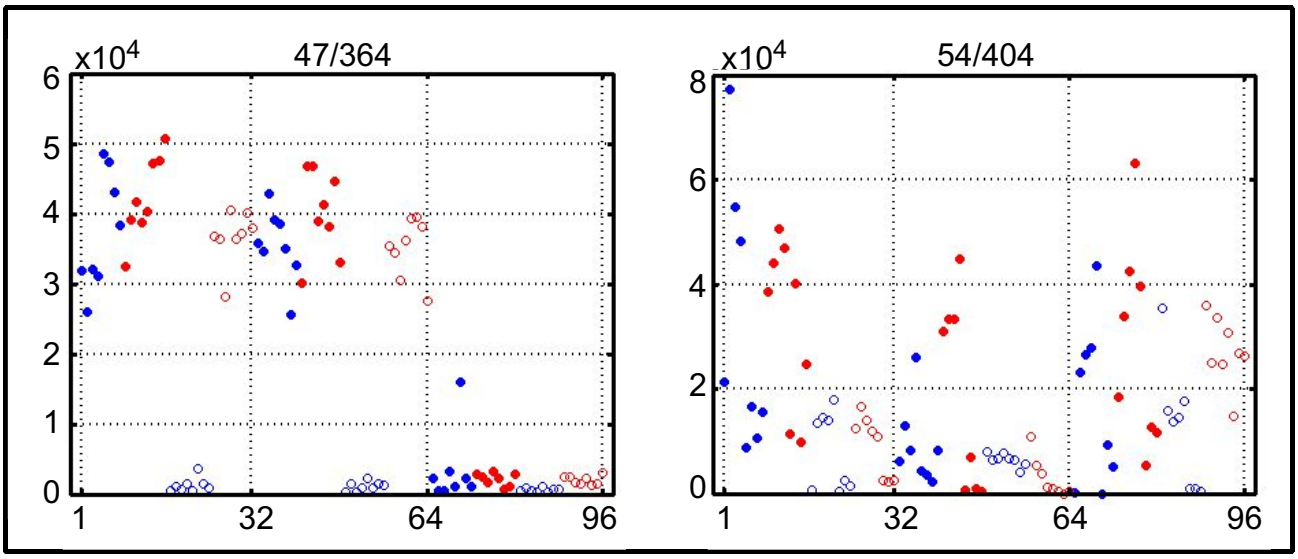


Figura 24: Áreas resueltas en las 96 muestras disponibles para algunos componentes seleccionados en base a su aporte al vector de regresión del modelo MB/MT con Selección de Componentes

Referencias: Símbolos rellenos: MT, Símbolos vacíos: MB, Azul: muestras Rambo, Rojo: muestras RAF. Los valores observados en el eje de las abscisas delimitan a las muestras de cada sector y en cada uno las muestras de cada clase se encuentran graficadas según su día de recolección. Los números del tipo A/B en la parte superior de cada gráfico indican el orden de componente en la selección (A) y el orden del mismo componente entre los 441 originales (B).

En la figura 24 se aprecia que ninguno de los componentes logra una separación clara y total de MB y MT. Viéndolos en particular, el componente 47/364 establece con nitidez una diferencia entre Rb y el resto de las clases en los sectores A y B. A su vez es el que tiene el mayor valor absoluto de todos los coeficientes de regresión, por lo que es evidente que este componente juega un rol clave en el modelo de clasificación. Por su parte el componente 54/404 no aporta información a primera vista, al mismo tiempo que su coeficiente de regresión no es particularmente importante. Podría haberse esperado más de este último, ya que su proyección en el plano de *loadings* deja ver que tiene el valor más positivo en la LV1, la cual capturó la mayor parte de la varianza de clases.

El hecho de que existan menos componentes donde las áreas, representando concentraciones metabólicas, sean mayores estadísticamente en MT que en MB, sumado al hecho de que las MB representan a la situación metabólica normal, sugieren que el tratamiento con Carbofurano produce mayoritariamente depresión de las vías de síntesis afectadas y minoritariamente aumento de éstas, o a la inversa en el caso de las vías de eliminación de componentes.

Vale destacar otro aspecto de los componentes donde las áreas son predominantemente mayores

en MT que en MB. Sólo este tipo de componentes podría contener la resolución de residuos de Carbofurano, ya que no puede esperarse que éste estuviera presente en mayor cantidad en las MB que en las MT. Si bien por cuestiones relativas al ajuste múltiple de curvas con cierto grado de ruido sería posible obtener en las MB áreas mayores a 0, estrictamente hablando los blancos deberían haberse resuelto con valores de área iguales a 0 en todos los casos. Por esta razón ninguno de los componentes expuestos podría ser Carbofurano, dado que en las MB existen niveles apreciables de áreas resueltas. Lo mismo se dedujo de la inspección de todos los componentes que no han sido mostrados. No obstante, los componentes graficados sirven para ejemplificar 2 grandes grupos de cinéticas que podrían ser posibles para el Carbofurano:

- El componente 47/364 en los sectores A y B podría asociarse a cinéticas de niveles altos, sin decaimiento en general. Obviamente los puntos son muy variables y no podría pensarse, más allá de posibles errores de ajuste, que representaran un valor constante en los días de muestreo. Más aun, algunas MT presentan sus valores más altos en sus últimos días de muestreo respectivos, sumado al hecho de que las muestras Fb tienen niveles altos. Pero a nivel general, se podría postular que con este modo de muestreo, no se vería un decaimiento del pesticida.
- El componente 54/404 deja ver que en las MT en general los valores de área para los primeros días de muestreo son mayores que para los últimos. Esto claramente no es estricto, sumado al hecho de que también se da en las MB, pero representaría un decaimiento posible de ser apreciado durante el muestreo.

Se proseguirá con la discusión sobre el Carbofurano posteriormente. Ahora es conveniente situar el foco en el modelo MB/MT sin selección de componentes. La imposición en la selección respecto de la diferencia de medias podría haber eliminado componentes con cinéticas que no cumplieran el criterio pero que sí podrían haber tenido áreas apropiadas para modelar una hipotética cinética de decaimiento del Carbofurano. Por ejemplo, si el compuesto se hubiese mantenido en un nivel alto de concentración tan solo unos pocos días, y luego hubiese decaído completamente, entonces la media de las áreas en MT podría dar baja y aproximarse significativamente a la media de las áreas del mismo componente pero en MB, con lo cual este componente hubiese sido descartado. Vale destacar que la selección de componentes no hubiera eliminado a los componentes que tuvieran cinéticas de nivel alto constante o con leve bajada, pero ya se dijo que entre los selectos no se vieron perfiles apropiados para modelar al pesticida. Por lo tanto, se realizó un

análisis de los componentes del vector de regresión sin selecciones previas. Aquí se observaron 2 cosas:

- En primer lugar, que los componentes que habían sido significativos en el vector con selección también lo eran sin selección. Se observó que no eran los de máxima importancia en términos de valor absoluto en el vector de regresión, por lo cual podrían existir otros componentes más correlacionados a la diferenciación de clases en cuestión. Estos otros componentes fueron analizados, obteniendo resultados similares que con la selección de componentes. Es decir, se vieron algunos en los cuales las áreas de al menos uno de los subcomponentes de cada clase se proyectaba alejado del resto. A su vez, las gráficas de estos componentes resultaron más confusas, lo cual está de acuerdo con el hecho de que sin restricción en la diferencia de medias es posible encontrar todo tipo de relaciones entre las áreas. Y aunque los resultados en términos de CCC y CCV fueron iguales a los obtenidos con selección de componentes, estas confusiones, sumadas a las ya vistas en el gráfico de *scores*, dan a entender que el modelo con selección de componentes es más fácilmente interpretable.
- Se observó también que ninguna de las cinéticas vistas a partir de las áreas de estos componentes podrían relacionarse con el decaimiento propuesto para el Carbofurano.

También vale destacar que el modelo sin selección de componentes podría haber tenido en cuenta a un componente con este tipo de cinética sin otorgarle demasiada relevancia, por lo cual no hubiese quedado indicado en las gráficas que se supone muestran a los metabolitos más importantes para la diferenciación. Es decir, los componentes que se consideraron relevantes fueron aquellos que en el gráfico de su vector de regresión tenían valores absolutos altos para sus coeficientes, o bien valores alejados del origen de coordenadas en el gráfico de *loadings*. Estos componentes fueron los que deberían haber tenido la mayor correlación con la información de clase y por eso habrían resultado decisivos en la clasificación, pero el detalle de cada evolución cinética será propio de cada componentes, incluso para los no selectos, y no puede deducirse directamente de las 2 gráficas en cuestión. Si el Carbofurano hubiese tenido una cinética en decaimiento como la propuesta en el ejemplo, entonces varias de las MT tendrían sus valores de área en niveles cercanos a los de las MB (que deberían ser cero). Por lo tanto, este componente podría resultar no significativo en el modelado, ya que la correlación entre “nivel de área” y “clase” no debería ser alta, sino más bien baja y proveniente de asociar un mismo nivel de áreas (en blancos o en MT ya

decaídas) a dos valores de clase diferentes. Por lo tanto, los perfiles de todos los componentes fueron inspeccionados en búsqueda de una cinética de decaimiento apta, sin encontrar resultados convincentes.

Para concluir, más allá de los detalles de cada uno, los modelos PLS-DA del tipo MB/MT derivados de la resolución con MCR-ALS resultaron aptos para determinar si existió *stress* fisiológico en ambos tipos de cultivares.

2.6.5.4 Modelos de clasificación con PLS-DA: Rambo/RAF

Los últimos modelos que resta analizar son los destinados a diferenciar tomates Rambo de RAF. Estos modelos resultaron ser levemente más complejos que los de MB/MT, ya que necesitaron una LV más. Según lo visto hasta ahora, esto podría haberse debido a que en la etapa de calibración los cultivares no fueron simplemente representados por blancos, los cuales representan la verdadera naturaleza de los frutos, sino también por MT. Como ya se ha dicho, algunos resultados sugirieron que el tratamiento con pesticida podría haber producido una homogeneización de las MT en términos de variados metabolismos. Por lo tanto, es posible que esas MT no hayan sido muy diferentes a nivel de áreas en muchos componentes, como aparentemente sí lo fueron las MB de ambos cultivares. En otras palabras, algunas MT de ambas clases estarían actuando como calibradores en grupos diferentes o representando naturalezas distintas, pero lo harían aportando información similar. Así pues, esto podría resultar en un problema en la etapa de calibración, probablemente solucionado con el agregado de una LV más al modelo. En la figura 25, el gráfico de *scores* de estos modelos deja ver que las clases en cuestión se encuentran relativamente separadas. En principio, se hace notar que los % de varianza explicados en **X** e **Y** se obviaron de las gráficas para no sobrecargarlas. Para el modelo con selección de componentes, los valores en **X** según el orden de LV fueron 40.91%, 15.83% y 12.66%, dando un total de 69.40%, mientras que para **Y** fueron 68.53%, 5.66% y 2.97%, con un total de 77.16%. Similarmente, para el modelo sin selección de componentes los valores fueron de 12.88%, 20.49% y 5.99% para un total de 39.36% en **X**, mientras que para **Y** sumaron 86.61% con aportes de 59.74%, 13.84% y 13.03%. De esto se deduce que la selección modeló más varianza (casi 30%) a nivel de áreas y concentró como en ninguna otra LV la varianza en **Y** (en la LV1 obtuvo el máximo de 68.53%), modelando poca información en las restantes LV. El modelo sin selección obtuvo mayor correlación con la información de clases (86.61%) y en sus últimas 2 LV presentó al menos cifras superiores al 13%. Esta mayor correlación puede que explique la leve diferencia si se comparan las cifras de CCC y CCV pertinentes.

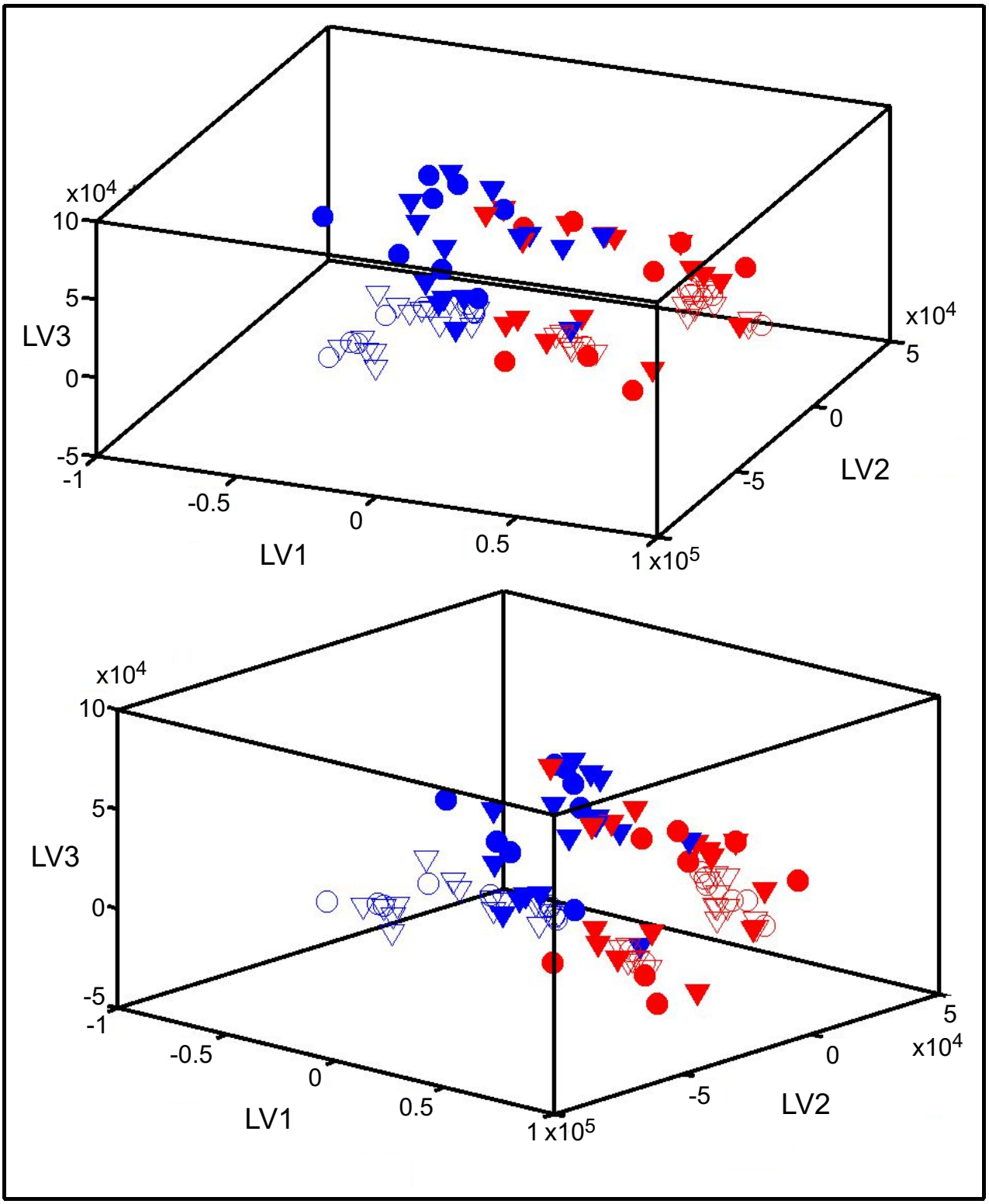


Figura 25: Distribución de scores de muestras de Calibración (círculos) y Validación (triángulos) en modelos PLS-DA para R/F, con (arriba) y sin (abajo) Selección de Componentes

Referencias: LVn: Variable Latente n, Símbolos rellenos: MT, Símbolos vacíos: MB, Azul: muestras Rambo, Rojo: muestras RAF

También debe notarse que para el modelo sin selección la varianza a modelar es mayor por cuanto existe información de todos los componentes y, no obstante, sólo modelando 39.36% de la varianza en **X** el modelo alcanzó resultados aceptables.

Respecto de lo expuesto en las gráficas de la figura 25, a nivel visual las clases R y F se separan de forma aceptable, pero no se observan separaciones extra de forma tan clara como en el caso de los modelos MB/MT con selección de componentes. Sí vale destacar que se observan agrupamientos de muestras similares y que desde las perspectivas mostradas los grupos de blancos parecen situarse en extremos opuestos de los cubos, representando lo que verdaderamente debería ser un tomate R o F, pero los límites no son tan definidos.

Con lo visto en los modelos MB/MT puede entenderse que evidentemente existen componentes diferentes entre blancos de ambos cultivares. De hecho, estos componentes ni siquiera habían sido modelados e igualmente fueron notorios. Por consiguiente, vale destacar algunas cosas:

- A sabiendas de que estos componentes existían, el uso de estrategias similares a las utilizadas en el análisis de los modelos MB/MT derivó en resultados con interpretaciones parecidas. En relación a los vectores de regresión, se encontraron componentes significativos para las diferenciaciones con y sin selección, en algunos casos coincidentes, como en los casos previamente detallados. También se observó que algunos componentes relevantes según sus coeficientes de regresión además presentaban relevancia en los modelos MB/MT. Evaluando las áreas de los componentes por separado, pudo notarse que aquellos que únicamente eran significativos en los vectores R/F y no en los MB/MT producían las mejores separaciones entre muestras de ambos cultivares. Estos últimos componentes serían los más apropiados para actuar como marcadores de clase R/F, ya que estarían intrínsecamente relacionados a la naturaleza de cada cultivar.
- Un mejor modelo no sólo en el sentido de % de varianzas explicadas sino también en su potencial para ser interpretado podría obtenerse directamente con las MB, sin participación de las MT en la calibración, porque las últimas resultan ser menos distinguibles entre sí. Aunque seguramente existieron vías metabólicas inalteradas o poco alteradas con el tratamiento que, por lo tanto, podrían representar a las vías naturales tal y como lo hacen las MB, en general no muestran las diferencias entre cultivares tanto como evidentemente lo hacen las MB. A su vez, se dijo que la intención de este tipo de modelo sería detectar adulteraciones en productos procesados derivados fundamentalmente de los tomates RAF, ya que en términos de forma de los frutos sería suficiente verlos para diferenciarlos de otros.

Como se supone que los productos elaborados tienen un tiempo extenso de permanencia desde que pudieron ser tratados con Carbofurano hasta antes de ser consumidos, se supone que el Carbofurano no podría ser lo suficientemente estable como para permanecer en estos productos. Por ende, buscarlo sería en vano, e incluir muestras tratadas en el diseño no parece ser lo óptimo en estas circunstancias.

2.6.5.5 Acerca del componente Carbofurano

Para poder determinar con certeza si hubo o no tratamiento con Carbofurano, debería poder probarse su existencia a través de una cuantificación apropiada a tal fin. A lo largo de esta discusión se ha hecho mención a la búsqueda de cinéticas en el período de muestreo para componentes que pudieran representar al Carbofurano. Dado que el objetivo principal de estas clasificaciones fue determinar el comportamiento metabólico de los cultivares con y sin tratamiento con el pesticida, es válido destacar algunos aspectos relacionados:

- Los datos analizados indican que en ninguno de los modelos para MB/MT se obtuvo una cinética convincente, similar entre sectores y apropiada para el pesticida. No obstante, se han podido diferenciar las MB de las MT con excelentes resultados. Esto da a entender que aunque no se haya podido hallar el componente que definiría sin dudas si hubo tratamiento o no, los efectos metabólicos que este habría causado son tanto o más importantes, en términos de varianza generada, que su sola presencia o ausencia. Esto es lógico, pues la varianza que podría estar relacionada a un único componente en relación a un gran conjunto de metabolitos naturales debería ser casi despreciable. Si la cinética fuera de nivel casi constante (o sin mucho decaimiento) durante el muestreo y para el componente en las MB se hubiera obtenido un nivel promedio también casi constante pero indefectiblemente cercano a cero, entonces seguramente el componente hubiese sido destacado en los modelos porque habría tenido una alta correlación con la información de clase, más allá de que su nivel pseudo constante en MT podría ser bajo y por lo tanto su aporte a la varianza en **X** sería escaso. El hecho de no poder identificar al Carbofurano impide decir con certeza que una muestra ha sido tratada específicamente con este pesticida, más allá de que los efectos de *stress* sí puedan ser notados.
- Los esfuerzos dedicados a intentar encontrar cinéticas apropiadas para el pesticida también fueron realizados en base a los resultados de la Parte 1, aunque no fueron documentados. En efecto, en esas experiencias, así como en otras donde no se mezclaron muestras entre sectores o

cultivares durante la resolución con MCR-ALS, tampoco se encontró la cinética buscada.

- Tener disponibilidad de un patrón de Carbofurano obtenido en las mismas condiciones cromatográficas y de detección que las muestras no sólo permitiría realizar identificación y cuantificación, sino que además se estaría contando con ambos perfiles (sus vectores en **C** y **S**) para el componente puro, pero a su vez afectado por características propias de los experimentos realizados, como ser defectos del HPLC o del detector, entre otros. Este tipo de características pueden parecer problemáticas a primera vista, pero como todas las muestras sufrirían hipotéticamente situaciones similares y dentro de éstas podría estar el analito, la presencia de las mencionadas características puede resultar crucial para obtener buenos resultados.
- Si durante las experiencias no se obtiene información desde patrones, la cuantificación resulta imposibilitada, pero existen escenarios alternativos que pueden mejorar otros aspectos. Por ejemplo, se podría contar con el espectro del componente a partir de su análisis en un instrumento similar, aunque no igual. O podría obtenerse desde bases de datos, aunque en este caso existe la posibilidad de que no se informe todo el espectro de MS, sino sólo cuáles variables de m/z fueron significativas y qué intensidades relativas obtuvieron unas respecto de otras. En los 2 ejemplos mencionados, la información espectral podría representar ventajas a la hora del modelado y resolución. No obstante, deberían realizarse adaptaciones para que esa información fuera apta para ser incluida en un análisis simultáneo con muestras como las aquí obtenidas. En este contexto, la WT (no necesariamente con filtros de Haar) sería nuevamente útil, especialmente por su capacidad de extraer información en sus coeficientes de detalle, ya que éstos pueden ser usados para modelar las respectivas estructuras de ruido en los patrones y en las muestras, de forma que los primeros se adapten a lo impuesto por las segundas, similarmente a como ocurre en el procedimiento de estandarización de señales denominado WHDS (del inglés *Wavelet Hybrid Direct Standardization*) (Tan y Brown, 2001). Lo importante aquí es resaltar que este tipo de información espectral podría ser incluida explícitamente en el modelo, bien a través de su sólo uso como estimación inicial de uno de los componentes que deberían tenerse en cuenta, o bien a través del uso en simultáneo de una restricción asociada a la denominada Matriz de Selectividad durante el ajuste experimental realizado por MCR-ALS. Ésta última no ha sido utilizada en este trabajo, pero vale la pena mencionarla como una opción para mejorar los resultados. Se trata de construir una matriz para informar que para determinados componentes de los que han de resolverse, existe información espectral que debe ser tenida en cuenta al obtener la matriz **S** producida en cada iteración. Esta matriz tendrá el

mismo tamaño que **S** (“componentes a resolver” × “variables espectrales”) y presenta valores codificados que pueden indicar que en cierta variable y componente no deben existir valores significativos (0), o que pueden existir y se desconocen por lo que pueden ser optimizados con libertad (se informan con Inf o NaN en Matlab), o bien que el valor debe ser el informado en la Matriz de Selectividad (cualquier valor distinto de 0). En el último caso, también cabe la posibilidad de explicitar si ese valor debe respetarse estrictamente (se dice “conocido”), o si sólo es un máximo posible que puede resultar menor luego de la optimización (se dice “selectivo”). Por lo tanto, incluir este tipo de información espectral y aplicar la restricción en cuestión puede mejorar la resolución. A su vez, lo explicado para la Matriz de Selectividad de espectros también puede aplicarse con la de perfiles de concentración. Para ambos casos, la aplicación de estas restricciones a través de información incorrecta forzará notablemente al algoritmo y hará que se requieran recursos de cómputo pura y exclusivamente para ajustar los datos experimentales a referencias que no deberían ser tenidas en cuenta, a la vez que las resoluciones definitivamente resultarían erróneas.

- Al respecto de la cinética buscada y no encontrada, debe tenerse en cuenta que el único dato certero es que en las MB no debería existir dicho componente (la certeza de su presencia en MT será discutida posteriormente), pero esta condición tampoco brinda suficiente información para resolver la cuestión. De hecho, se realizaron experiencias en las cuales las áreas de todos los componentes fueron ordenadas según la media de áreas para MB, siendo de interés aquellas lo más cercanas posibles a cero durante todos los días de muestreo. Los componentes fueron visualmente inspeccionados en ese orden, pero no se encontraron cinéticas en las MT que podrían corresponderse con las postuladas para el Carbofurano. Vale destacar que estrictamente hablando las áreas en las cinéticas para MB no deberían ir cerca del nivel cero, sino en cero en sí. No obstante, por errores en el ajuste sí se podrían esperar valores bajos. Aun habiendo realizado el análisis anterior con cada sector por separado para evitar un incorrecto orden de medias producto de la no repetibilidad entre sectores, los resultados tampoco mejoraron.
- En relación a la certeza de ausencia en MB y a la posibilidad de presencia en MT, otra restricción que podría utilizarse en MCR-ALS es la asociada a la denominada Matriz de Correspondencia entre especies modeladas y muestras experimentales. Esta restricción se aplica a través de la construcción de una matriz de datos binarios, con dimensiones coincidentes con el número de muestras independientes en las filas y con el número de componentes modelados en

las columnas. El valor 0 en ciertas filas y columnas de la matriz indica que en las muestras de dichas filas no pueden existir los componentes de dichas columnas. Similarmente, el valor 1 indica posibilidad de existencia de un componente, aunque no certeza. Esto significa que durante la optimización el perfil de concentración del componente no debe ser restringido a tener un valor nulo y constante como cuando se garantizan ausencias de componentes, más allá de que en efecto el componente pueda no haber estado y que por ende su perfil de concentración haya sido resuelto como nulo y constante. Al igual que con la Matriz de Selectividad, debe existir mucha seguridad sobre estas imposiciones, pues de no ser reales, interferirán en detrimento de mejores resultados. Imponer incorrectamente certeza sobre la inexistencia de un componente hará que en el ajuste éste no sea directamente tenido en cuenta, mientras que posibilitar la existencia de componentes que podrían no existir no influirá tan negativamente, ya que en el mejor de los casos su presencia podrá ser aceptada por el modelo y éste determinará que el nivel de existencia, más allá de lo paradójico, pueda ser nulo. Cuando no se especifica esta restricción, como en nuestro caso, la matriz se conforma totalmente de unos, es decir, no se elimina la posibilidad de encontrar a cualquiera de los componentes en cualquiera de las muestras. Bajo este punto de vista, haber obviado usar la restricción parece ser incorrecto. Sin embargo, esta restricción no se sustenta siempre sobre sí misma y deben analizarse otros aspectos. En el caso de nuestra experiencia con Carbofurano, ya que absolutamente todas las estimaciones iniciales fueron obtenidas matemáticamente (con SIMPLISMA) y no se contó con información espectral experimental, no es posible identificar cuál de estas estimaciones sería la del Carbofurano y por ende no se puede especificar, a excepción de una elección trivial sobre cualquiera de los disponibles, que ese componente no debería estar presente en las MB. Por lo tanto, este problema repercute en el armado de la Matriz de Correspondencias, ya que no se puede determinar para cuál de todos los componentes las columnas deben tener el valor 0, aun cuando sí se puedan identificar las MB y las MT. Esa fue la razón de no aplicar esta restricción. Si lo hubiésemos hecho, simplemente hubiéramos logrado que un componente se resolviera en niveles nulos para las MB, lo cual en sí no hubiese aportado información relacionada al Carbofurano.

- Aun suponiendo que se tuviera un espectro apropiado de Carbofurano y entonces se pudiera construir la Matriz de Correspondencias (además de la de Selectividad), especificar que en las MT efectivamente debería existir el componente en cuestión, más allá de que intuitivamente parecería algo lógico, merece una reflexión adicional. Aunque uno garantice la posibilidad de

presencia basándose en la seguridad de que durante los experimentos sólo las MT fueron efectivamente rociadas con el pesticida, existe otro factor relevante que determinará si a nivel de ajuste el componente resultará con un perfil de concentración significativamente distinto de cero o no. Este factor tiene relación con el hecho de si con el sistema de detección utilizado sería posible percibir al pesticida en los niveles de concentración en que este se encontraba realmente en las muestras tratadas. Es decir, si los niveles de señal provocados por el componente no hubiesen superado ciertos límites mínimos, con seguridad no podría haber sido detectado. Entonces, si se va a especificar la presencia de un componente selectivamente en ciertas muestras y se pretende sacar verdadera ventaja de esto, el componente deberá estar en niveles apropiados para confirmar su presencia. Caso contrario, la especificación no tendrá demasiada utilidad.

En este trabajo y por cuestiones ya descriptas, se intentó encontrar al Carbofurano a través del análisis de múltiples cinéticas evolutivas que podrían representarlo. No obstante, ninguna de las reportadas (ni muchas otras que fueron observadas) resultó apropiada por razones diversas, como ser la presencia de áreas significativas (no asociables a errores de ajuste) en al menos un día de recolección en MB, o la ausencia de repetibilidad entre las cinéticas observadas en distintos sectores, entre otras. No haber encontrado una cinética apta da a entender 2 cosas:

- Las resoluciones no fueron satisfactorias para encontrar al Carbofurano (sí para detectar efectos de *stress*), porque al menos un componente con las características cinéticas buscadas debería haberse hallado en MT y sobre todo no encontrado en MB. De todos los componentes modelados, se presupone que es el único con presencias o ausencias selectivas.
- Dadas las condiciones experimentales llevadas a cabo, el Carbofurano no sería detectable, aunque sí efectos de su aplicación. Esto podría deberse a una gran inestabilidad del compuesto, pero si éste fuera el caso, entonces su función de proteger a los tomates de sus plagas no tendría sentido. Por lo tanto, deben quedar al menos trazas del pesticida, y esto está de acuerdo con la posibilidad de que el nivel de concentración en los frutos hubiese sido muy bajo, suficiente para protegerlos, pero no para ser detectado.

El último punto expuesto motivó una revisión bibliográfica adicional. No se tienen datos para corroborar si los niveles de concentración final en los frutos hubiesen sido detectables durante todos los días del muestreo realizado. Varios textos revisados (Abad y col., 1997; Pacioni y Veglia, 2003;

Abad y col., 1999; Gui y col., 2009) mostraron estudios realizados en matrices de origen vegetal, como ser frutos o jugos de manzana, uva, ananá, banana, papa, frutilla, coco, lechuga y zanahoria. Con técnicas diferentes (HPLC, Espectroscopías de fluorescencia y UV-VIS, Anticuerpos monoclonales/ELISA, entre otras) cada uno obtuvo determinados límites de cuantificación y de detección para el Carbofurano. Debe destacarse que en general las muestras eran fortificadas con determinadas concentraciones del pesticida y que estas fortificaciones podrían no representar una aplicación normal del pesticida sobre frutos como la realizada en el presente trabajo. Ejemplo de esto fueron los estudios sobre jugos comprados en supermercados y fortificados en laboratorio, algo nada representativo de nuestra realidad. De los resultados en estos textos se desprendió sobre todo que los métodos propuestos flanqueaban los potenciales problemas que este tipo de matriz podría acarrear, pero se dificultó realizar una extrapolación a nuestro caso. Otro texto (Ling y col., 1993) fue más acorde a nuestra situación por cuanto los frutos analizados también fueron de tomates. En este trabajo, además de proponer una cinética de decaimiento de primer orden para los residuos del pesticida, los frutos fueron obtenidos con un sistema hidropónico, la detección se realizó con espectroscopía UV y el Carbofurano no fue rociado, sino que fue adicionado en la solución hidropónica. Los autores detectaron residuos del pesticida pero nunca en valores superiores a su límite permitido (MLR), por lo cual culminaron concluyendo que el período de tiempo de seguridad impuesto entre la aplicación del pesticida y el consumo de los frutos podría ser menor al determinado por la ley, que en ese entonces era de 60 días y no de 45. No obstante, además de la diferencia con nuestro método de detección y de que no existió aplicación por rociado, ya que el cultivo fue hidropónico no se supone que para un cultivo en tierra como el que dio origen a nuestros datos se obtendrían resultados similares, debido a que la absorción de componentes por parte de los vegetales no se produciría de la misma forma.

Finalmente, uno de los textos más sugerentes en este contexto fue (OMS/FAO, 2009), donde se reportan conclusiones acerca del uso de Carbosulfán y de Carbofurano en productos alimenticios. El Carbosulfán, más allá de poder ser utilizado directamente, es un precursor que puede dar origen a Carbofurano. Antes de 1997 se habían realizado estudios del primero en cítricos como naranjas y mandarinas. En 1997 se realizaron otros estudios en Brasil, México y España, incluyendo otros cítricos. En una nueva evaluación de los datos de 1997 realizada en 2004 sobre los residuos de Carbosulfán en pulpas de naranja se determinó que era poco probable que residuos de carbamatos como Carbosulfán se encontraran en pulpas de naranja en niveles mayores al LOQ (Límite de Cuantificación) de 0,05 mg/Kg. El mismo valor se estimó para el Carbofurano en naranjas. Esta

estimación se basó en un estudio metabólico de los realizados en 1997, en el cual la pulpa de las naranjas tratadas con Carbofurano marcado con ^{14}C contenía no más de 0.3% de los residuos radioactivos 30 días después del tratamiento. En 2009, una nueva discusión determinó que menos del 0.3% de los residuos radioactivos pueden ser encontrados en la pulpa comestible de la fruta 0, 7, 15 y 30 días después del tratamiento. Más allá de que nuestro trabajo no se basó en naranjas y no utilizó marcadores radioactivos, el texto en cuestión sugiere que el Carbofurano no parece ser un analito estable, de sencilla detección y/o cuantificación, incluso el mismo día de su aplicación, si es que fue utilizado en concentraciones normales para prácticas agrícolas. A su vez, es posible que sí pueda traspasar las capas externas de ciertos frutos y alojarse en el interior de éstos, en concentraciones muy bajas, pero quizá suficientes para producir cambios metabólicos variados y posteriormente perceptibles. Por todo lo anterior, aunque pudieron haberse cometido múltiples errores experimentales y/o estratégicos que repercutieran en la no resolución del componente Carbofurano, es posible que no haya sido encontrado simplemente porque no se encontraba en un nivel de concentración apropiado para su detección.

2.7 Conclusiones

– La estrategia quimiométrica basada fundamentalmente en WT y MCR-ALS para la resolución e interpretación posterior de arreglos de datos de LC-MS se mostró adecuada para realizar un estudio metabonómico en frutos de tomates tras su tratamiento con Carbofurano.

– Los perfiles evolutivos de supuestos compuestos endógenos a través de los días de muestreo pudieron ser observados y comparados, gráfica y analíticamente, en Muestras Tratadas y en Blancos, ya que la metodología se mostró apropiada para capturar las trayectorias individuales de estos metabolitos en el tiempo. De esto se desprendió que las cinéticas de varios componentes difirieron a causa del *stress* fisiológico que representó el tratamiento. Las diferencias observadas pudieron relacionarse a niveles de concentraciones en los metabolitos, así como a retrasos/adelantos de las vías metabólicas durante el muestreo.

– Los resultados obtenidos con las matrices reducidas mediante WT y filtros de Haar mostraron coherencia con lo que podría esperarse en un estudio de muestras complejas con metabolismos variados. Al mismo tiempo, la compresión conservó la suficiente cantidad de información como para poder distinguir comportamientos diversos que fueron observados y que

resultaron útiles para realizar discernimientos. Si bien es probable que la resolución de las señales en sus dominios originales hubiese producido resultados mejores dado que no se pierde nada de la información obtenida, la utilización de WT se destacó por una disminución notable en los recursos de cómputo. Esto no es un detalle menor en los tiempos actuales, donde el uso de sistemas embebidos permite realizar instrumental portátil y mediciones por procesamiento inmediato de las muestras en campo. Los sistemas embebidos suelen no contar con los recursos de cómputo de una computadora común. Por lo tanto, la WT se muestra apta para este tipo de situaciones, en las cuales la reducción de la cantidad de información antes de su procesamiento es imperiosa..

- La utilización de matrices apiladas en MCR-ALS permitió la resolución simultánea de varias muestras, permitiendo a su vez la obtención de conjuntos comunes de especies resueltas a nivel espectral, con lo cual fue posible comparar cinéticas. La calidad de los ajustes, aunque mejorable, en términos de varianza explicada siempre fue superior a la mínima impuesta de 90%.
- La SVD como método de estimación del número de especies variantes relevantes, si bien no garantiza la obtención del número exacto y necesario desde un punto de vista metabonómico, en el esquema propuesto permitió mantener consistencia entre los distintos análisis realizados.
- Ya que no se contó con espectros puros para los metabolitos modelados, SIMPLISMA resultó de gran utilidad para obtener las estimaciones iniciales. Se observó que en algunos casos dichas estimaciones resultaron ser similares a lo obtenido tras su optimización en MCR-ALS, por lo cual la extracción de las variables más puras fue viable a través del método. A su vez en ocasiones se observó que antes de la primera iteración de MCR-ALS, los porcentajes de varianza explicada para los datos modelados solamente a través de las aproximaciones de SIMPLISMA fueron altos, lo que significó buenos puntos de partida para los procesos de refinamiento en MCR-ALS.
- Algunas restricciones fueron utilizadas en los procedimientos de ajuste y otras sólo discutidas o propuestas, pero todas pueden mejorar la calidad de los resultados y la forma en que éstos pueden ser interpretados desde un punto de vista experimental, siempre que se utilicen en el marco apropiado.
- Los modelos de clasificación PLS-DA fueron útiles para extraer información desde las áreas resueltas en MCR-ALS. La selección de componentes no afectó de manera determinante el desempeño de las clasificaciones, pero facilitó su interpretación.
- Los modelos para Rambo/RAF y “4 clases”, si bien no eran prioritarios, otorgaron cifras de mérito aceptables para las predicciones.

– Los modelos PLS-DA del tipo MB/MT resultaron aptos para determinar si existió stress fisiológico en ambos tipos de cultivar.

– Algunos de los efectos de *stress* se mostraron más visibles en un tipo de cultivar que en el otro, aunque a nivel de los modelos de clasificación pudo observarse una homogeneización parcial de ambos tipos de MT.

– La detección de MT pudo ser realizada probablemente porque las diferencias mayoritarias radican en los efectos del tratamiento sobre los metabolitos naturales y no sobre la detección específica de residuos de Carbofurano. Algunos análisis sugirieron que el tratamiento produce mayoritariamente depresión de las vías de síntesis afectadas y minoritariamente aumento de éstas, o a la inversa en el caso de las vías de eliminación de componentes.

– El hecho de no poder identificar al Carbofurano impide decir con certeza que una muestra ha sido tratada específicamente con este pesticida, más allá de que los efectos de stress sí puedan ser notados.

– El pesticida no pudo ser hallado a través del análisis de cinéticas evolutivas resueltas. Más allá de potenciales errores de ajuste o experimentales, y de que la concentración en las MT pudo no ser suficiente para el sistema de detección utilizado, también es posible que la reducción con WT haya eliminado la posibilidad de detectarlo, aunque quizá sí sería posible en el dominio original de los datos. Se vio en algunas experiencias que al comparar los componentes estimados con SVD en matrices originales y reducidas, las últimas mostraban menor cantidad siempre. El insecticida pudo haber sido un componente minoritario, con escaso aporte individual a la varianza general, descontando la varianza que de su aplicación puede derivarse en otros metabolitos modelados. Así pues, es posible que la WT lo haya eliminado y éste no haya sido modelado. Otra opción pudo ser que el mismo componente (sin la necesidad de que haya sido minoritario) y otros pudieron haber tenido espectros similares y/o estar co-eluyendo, por lo cual el proceso de promediarlos a través de su representación como coeficientes de aproximación pudo haber acrecentado la posibilidad de que a nivel de ajuste, se comportaran como un único componente. De esta última opción surge que de haberse resuelto el Carbofurano, pudo haberlo hecho en conjunción con otros componentes, los cuales pudieron tener cinéticas de evolución ciertamente diferentes y en dicho caso la cinética promedio pudo no haber sido la apropiada para detectar al insecticida. No obstante a las críticas señaladas, la WT se mostró apropiada para conservar una cantidad de información suficiente para evaluar la situación de stress fisiológico en cuestión, objetivo del trabajo.

– Un estudio como el realizado podría ser notablemente mejorado si se cuenta con los espectros de los metabolitos modelados. Más allá de las ventajas del uso de estos espectros en combinación con las restricciones ya discutidas, el salto de calidad se daría en que al poder identificar a los componentes sería posible relacionarlos en vías metabólicas documentadas y entonces así se podría evaluar específicamente cuáles vías experimentaron cambios, cuales surgieron o cuales fueron eliminadas, todo en base al tratamiento con Carbofurano. También podrían obtenerse mejoras a través de la inclusión de técnicas de detección más sofisticadas que las aquí utilizadas, como ser Tiempo de Vuelo (TOF), Quadrupolo-TOF o espectroscopía de MS en tandem (MS/MS), entre otras.

CAPÍTULO 3: Obtención automatizada de muestras y lecturas fluorimétricas mediante hardware y software de código abierto. Aplicación en el laboratorio quimiométrico.

3.1 Resumen

Varias muestras conteniendo las fluoroquinolonas Ofloxacina, Ciprofloxacina y Danofloxacina fueron resueltas mediante HPLC y registradas a través de espectroscopia UV, otorgando sendas matrices para su resolución numérica posterior.

El equivalente al descarte de estas cromatografías fue fraccionado en pocillos de placas de ELISA. Para realizar esta tarea, fue necesario diseñar, construir y programar un recolector automatizable, lo cual se hizo posible a través de la implementación de hardware de código abierto y del reciclado de varios componentes de tecnologías en desuso.

Desde cada pocillo con muestras recolectadas en cada placa de ELISA se obtuvieron matrices de Excitación-Emisión mediante la utilización de un fluorímetro conectado y controlando a un lector de placas automatizado. No obstante, por limitaciones en el diseño del software de operación del instrumento, fue necesario elaborar una interfaz gráfica de comunicación con éste, lo cual dio lugar a la obtención de las matrices señaladas. De forma similar, también se generó una interfaz para controlar al recolector de muestras. Ambas interfaces fueron diseñadas y desarrolladas a partir de software de código abierto.

La resolución de las señales obtenidas mediante MCR-ALS condujo a diversos análisis, algunos relacionados a la evaluación del dispositivo recolector, y otros a la calidad de los ajustes y de cuantificaciones derivadas desde los últimos.

Dentro de un marco de potenciales mejoras, los resultados alcanzados pueden considerarse aceptables para la obtención de datos de orden superior.

3.2 Introducción

El presente trabajo trata sobre un problema que se presentó en el laboratorio de Desarrollo Analítico y Quimiometría (LADAQ) durante una investigación y está relacionado a la imposibilidad de realizar experimentos aún cuando el instrumental necesario esté realmente disponible pero, ante limitaciones de software y/o hardware, los experimentos se tornan en sí inaccesibles. Esto suele ser común en laboratorios de Investigación y Desarrollo fundamentalmente de países con escaso nivel de desarrollo tecnológico, y el previsible resultado termina siendo el abandono de las aspiraciones

científicas. También trata sobre la elaboración de instrumental reciclando partes de tecnología en desuso y aplicando tecnologías filosóficamente diferentes a las convencionales.

Gran parte del conocimiento necesario para desarrollar las tareas reportadas no proviene de fuentes académicas clásicas, oficiales y donde los trabajos son revisados por pares designados por un editor, sino de proyectos y/o comunidades de software y hardware libre, donde el avance colectivo del conocimiento se concibe de otra forma. Esto último no elude el proceso de revisión, de hecho en algunos foros virtuales existe amplia participación de académicos, sólo que no están organizados de la misma manera. A su vez, el espíritu de compartir el conocimiento aboliendo en lo posible toda barrera de acceso a éste, es motivo común de ayuda rápida y precisa ante las solicitudes de los miembros de estas comunidades. Muy frecuentemente, lo que uno consulta de estos foros ya fue previamente preguntado, y las respuestas suelen estar amablemente adaptadas más al nivel del entendimiento que haya demostrado quien realizó la consulta que a formalismos sobre cómo debe ser expuesta y fundamentada la información. Esto último por un lado resulta inapropiado para el contexto académico de este escrito, pero por otro representa una gran ventaja cuando uno se adentra en prácticas que le son poco frecuentes, máxime cuando parece innegable notar que muchas veces son los formatos y las reglas estrictas del academicismo los que impiden en sí el acceso a una mínima compresión para lograr ciertos objetivos prácticos, más allá de un entendimiento total de las cosas (ilusamente pretendido). Otra característica de estos entornos es la ausencia casi total de material impreso, debida fundamentalmente a que no tiene mucho sentido imprimir conocimiento que se actualiza frecuentemente y a medida de los miembros, además de que todo se aloja en Internet y este libre acceso de alguna manera también desalienta emprendimientos clásicos con fines de lucro, como suelen ser muchos textos impresos.

En relación al tema reportado, en el laboratorio existía la necesidad de generar matrices de Excitación-Emisión a partir de un fluorímetro, con el objeto de poner a prueba algoritmos para datos de tercer orden (este objetivo no es parte del presente escrito, donde los datos serán resueltos por MCR-ALS). La obtención de las mencionadas matrices debía realizarse recolectando muestras ordenadamente a la salida de un HPLC, por lo cual se hizo necesaria la automatización de esta tarea.

El fluorímetro en cuestión posee un accesorio para obtener señales de fluorescencia de forma automática en distintos pocillos de placas de ELISA de 96 unidades, lo cual determinó que éste fuera el recipiente de recolección de muestras cromatografiadas. No obstante, el software provisto por el fabricante, desde el cual se opera el fluorímetro, sólo permite seleccionar hasta 20 pares de

longitudes de onda de Excitación-Emisión, y no permite automatizar la tarea de cambiar estos pares, por lo cual no era posible obtener las matrices deseadas, aún cuando a nivel instrumental se podía pensar que efectivamente lo necesario se encontraba presente. Por lo último, se consultó el manual del fluorímetro y se verificó que así como el software provisto se comunica con el instrumento, es posible elaborar una interfaz propia de comunicación, con lo cual se hace posible controlar más parámetros que los permitidos y de esto se deriva la posibilidad de obtener a las matrices buscadas. A su vez, cualquier mejora que pudiera obtenerse para el fluorímetro podría ser aplicada en otro laboratorio de uso común (de nuestra Facultad) donde existe el mismo instrumento, lo cual también resultaba motivador.

Tanto para la elaboración de interfaces de comunicación, como para la automatización de la recolección de muestras en placas de ELISA, y para el modelado y obtención de circuitos electrónicos necesarios, las herramientas que fueron utilizadas poseen una característica filosófica común, la de ser de código abierto y, en el caso de lo referido solamente a software, de descarga gratuita (tal y como Linux, el más popular de este tipo de proyectos de código abierto). Estas herramientas fueron Processing [http://processing.org/], Arduino [http://arduino.cc/] y Fritzing [http://fritzing.org/].

Processing es un lenguaje de programación, un entorno de desarrollo y una comunidad online. Fue creado en 2001 por Ben Fry y Casey Reas en el MIT (*Massachusetts Institute of Technology*). Aunque originalmente fue destinado a facilitar la enseñanza de artes visuales y sus aplicaciones, como conjunto evolucionó hasta convertirse en una herramienta también apta para profesionales. Actualmente existe una gran comunidad de estudiantes, artistas, diseñadores, investigadores y aficionados usando Processing para aprender, prototipar y producir. Muchas de estas personas son las que producen código en forma de bibliotecas, lo cual aumenta las capacidades provistas por los responsables del proyecto Processing. Según la página web oficial, sus aplicaciones más habituales son el prototipado de software (por empresas como Google, Intel, Nokia, entre otras) y la visualización de datos, como fue el caso en el cual la NSF y la NOAA (*National Science Foundation* y *National Oceanic and Atmospheric Administration*, Estados Unidos) financiaron investigaciones sobre diversidad de fito y zooplancton en la Universidad de Washington, donde Processing fue utilizado para simulaciones de ecología dinámica. En el presente trabajo, Processing fue utilizado para crear interfaces gráficas que permitiesen comunicación tanto con el fluorímetro como con una placa Arduino.

El proyecto Arduino nació en 2005 para estudiantes del Instituto de Diseño Interactivo de Ivrea, Italia. Si bien desde hace un tiempo los responsables del proyecto (implementación, documentación, etc.) son David Cuartielles, Gianluca Martino, Tom Igoe, David Mellis y Massimo Banzi, Arduino es en sí una gran comunidad de usuarios de todo el mundo. El término también se utiliza para denominar al Entorno Integrado de Desarrollo (IDE) donde se realiza la escritura de código, y para nombrar genéricamente a varios modelos de plataformas para prototipado electrónico (placas) basadas en hardware de código abierto (o directamente hardware abierto). Los proyectos prototipados pueden funcionar de forma autónoma siempre que se los provea de energía y/o conectados a una PC u otros dispositivos, incluyendo a otras placas Arduino. En cuanto a la filosofía de hardware abierto, desde la sección de preguntas comunes (FAQ) del sitio oficial de Arduino se puede leer lo siguiente: "El hardware de código abierto comparte muchos de los principios y modos del software libre y de código abierto. En particular, creemos que la gente debe ser capaz de estudiar nuestro hardware para entender cómo funciona, hacerle cambios, y compartirlos. Para facilitar esto, liberamos todos los archivos de diseño originales (Eagle CAD) para un hardware Arduino. Estos archivos existen bajo una licencia *Creative Commons Attribution Share-Alike*, la cual permite la realización de trabajos derivados tanto personales como comerciales, siempre que se reconozca a Arduino y se liberen los nuevos diseños bajo la misma licencia".

Como piezas de hardware, las placas fueron concebidas para permitir el acercamiento de personas a la electrónica sin conocimientos específicos sobre ésta, más allá de que los diseñadores desde luego estén muy capacitados, y manteniendo bajos los costos de fabricación aún a expensas de disminuir algunas capacidades (que pueden ser aumentadas mediante estrategias que requieren inversiones tampoco muy grandes). El proyecto está dirigido para que tanto las placas Arduino como los componentes electrónicos sean utilizados como piezas con funciones determinadas. Por limitaciones propias de quien escribe, sumado al hecho de que esta tesis doctoral no representa el contexto apropiado para profundizar en temas de electrónica, este enfoque funcional será el utilizado en las descripciones posteriores, sin ahondar demasiado en los principios de funcionamiento de los componentes. Las placas Arduino pueden ser conectadas a sensores, procesar señales, emitir salidas eléctricas y comunicarse con otros dispositivos, entre otras características útiles para el presente trabajo. Poseen un microcontrolador que puede ser programado utilizando un lenguaje (derivado de Wiring) creado para Arduino y un IDE basado en Processing. Las placas pueden ser fabricadas y/o comercializadas por cualquier persona, y el software necesario es gratuito. Existen varios modelos de placas (Mega, Nano, Mini, Leonardo, entre otras) que se

diferencian en mayores o menores capacidades y tamaños, siendo el estándar actual el modelo denominado UNO, utilizado en el presente trabajo. Todos estos modelos están basados en microcontroladores de la familia Atmel AVR de 8 bits, mientras que el más reciente modelo DUE se basa en Atmel ARM de 32 bits, con muchas más prestaciones de cálculo. Las placas suelen ser de muy bajo costo y normalmente no están sujetas a restricciones de importación. Si bien el grupo original produce placas en Italia, éstas son también producidas alrededor de todo el mundo por otros. En la propia experiencia de quien escribe, habiendo probado varios modelos y más de una decena de placas para diversos prototipos, el origen de fabricación no parece impactar en las funcionalidades.

Entre algunos ejemplos de instrumental, el proyecto Spectruino [http://myspectral.com/] consiste de un espectrómetro de código abierto basado en Arduino, el cual recientemente ha sido utilizado por la NASA en el espacio. En la referencia (D'Ausilio, 2012) se realizaron experiencias psico y neurológicas utilizando Arduino, mientras que (Pearce, 2012) realizó una alusión a Arduino a nivel académico, en relación a la generación de instrumental de investigación con hardware de código abierto.

A su vez, dentro de la filosofía de hardware abierto y en relación a la elaboración de instrumental, vale destacar a la impresora RepRap [http://reprap.org], electrónicamente controlada por placas Arduino. Este tipo de impresoras (existen variantes) tiene entre sus objetivos aumentar al máximo la capacidad de autoreplicación (impresión de las piezas necesarias para construir una nueva impresora, excluyendo componentes metálicos y electrónicos, entre otros) y permiten la impresión tridimensional de objetos con formas muy específicas, los cuales entre otras cosas son utilizados para elaborar mecanismos de automatización de tareas. Muchos diseños (*things*) puestos a disposición por usuarios de estas impresoras pueden encontrarse y descargarse gratuitamente en [http://www.thingiverse.com/]. Entre éstos existen elementos simples de laboratorio como por ejemplo placas de ELISA (*thing*:53671) y gradillas para tubos Eppendorf (*thing*:49814), así como algunos otros más complejos como los necesarios para convertir un típico taladro Dremel en una centrífuga (Dremelfuge) para seis de los tubos mencionados (*thing* 1483), o una tobera especial que, adaptada sobre la cámara de ciertos teléfonos con sistema operativo Android, puede convertirlo en un espectrómetro (*thing*:49934). El último fue desarrollado por otra comunidad de corte filosófico similar, denominada Public Lab [http://publiclab.org/].

Fritzing es un proyecto pensado para dar soporte sobre hardware de código abierto a diseñadores, artistas, investigadores y aficionados que quieren trabajar con interacciones electrónicas. Está conformado en una comunidad virtual que da alojamiento para que los usuarios documenten sus prototipos, los compartan con otros, e incluso lleguen a etapas de fabricación profesional. Como herramienta de software, Fritzing facilita el modelado y la creación de circuitos que posteriormente pueden imprimirse en placas.

En cuanto a los fármacos utilizados como analitos en este trabajo, fueron tres fluoroquinolonas denominadas Ofloxacina, Ciprofloxacina y Danofloxacina. Los miembros de esta familia son agentes antibacterianos muy útiles, los cuales se administran en grandes cantidades a humanos y animales, finalizando como desechos fundamentalmente en aguas residuales de hospitales y municipios, y regresando a la tierra cuando los animales excretan (Jjemba, 2006). Residuos de estos antibióticos fueron reportados en ambientes naturales de muchos países (Tamtam y col., 2009). Desde luego, el monitoreo de pequeñas cantidades de estos compuestos en diversas matrices ambientales se torna necesario para la protección de la salud humana y el control del ambiente. La literatura revela una gran cantidad de métodos reportados para la determinación de fluoroquinolonas en aguas, especialmente incluyendo HPLC con detección de fluorescencia o de masa (Speltini y col., 2010). Recientemente se reportó la cuantificación simultánea de ocho quinolonas en aguas, analizadas mediante HPLC-fluorescencia (Vázquez y col., 2012).

Finalmente, vale destacar que en un trabajo reciente (Lozano y col., 2013) se realizó una experiencia similar a la que será descripta, en la cual se obtenían matrices de Excitación-Emisión en función del tiempo de elución cromatográfico. En dicho trabajo, en lugar de recoger la misma muestra cromatografiada en fracciones separadas y luego obtener matrices de fluorescencia como en el presente, cada muestra fue inyectada ocho veces en el sistema cromatográfico, registrando en cada caso espectros de Emisión con longitudes de onda de Excitación variables. Este es solo un ejemplo de una tendencia en el mundo quimiométrico, referida a combinar señales desde diferentes fuentes con el objeto de realizar procesamientos algorítmicos para datos de orden cada vez mayor. Vale mencionar que en el citado trabajo la resolución se realizó mediante una Calibración de cuarto orden, y que aunque los datos registrados en el presente trabajo (tiempo de elución, UV, Emisión-Excitación) permitirían evaluar diferentes combinaciones para calibrar, no es objetivo del presente ahondar en esas cuestiones, por lo que los datos serán tratados simplemente con MCR-ALS.

3.3 Objetivos

- A través del uso de hardware de código abierto y de partes fundamentalmente recicladas desde tecnología en desuso, elaborar un recolector de muestras automático y programable para placas de ELISA y aplicarlo en la salida de un HPLC.
- A través del uso de software de código abierto, diseñar los circuitos electrónicos necesarios para el recolector, y elaborar interfaces gráficas de comunicación para controlar al recolector y para obtener matrices de Emisión-Excitación desde un fluorímetro con lector de placas de ELISA.
- Evaluar los resultados de las tecnologías de código abierto a través de la resolución de señales mediante MCR-ALS aplicado en muestras conteniendo tres fármacos.

3.4 Teoría

La teoría sobre MCR-ALS fue expuesta en el capítulo 2. Además de las cifras de mérito ya definidas en los capítulos previos, se agregaron algunas adicionales.

3.4.1 Cifras de mérito

- Error Relativo de las Predicciones % (REP%, del inglés *Relative Error of Prediction*)

$$\mathrm{REP}\,\% = \frac{\mathrm{RMSE}}{\mathrm{mConcCal}} \times 100 \quad (1)$$

donde RMSE corresponde a un grupo de predicciones y mConcCal es la media de las concentraciones de la Calibración utilizada para obtener dichas predicciones.

Para las calibraciones pseudo-univariadas basadas en áreas resueltas con MCR-ALS, teniendo en cuenta a su vez el grado de solapamiento de los perfiles espectrales resueltos y la cantidad de variables en el orden de apilamiento (tiempos o pocillos en este trabajo) por matriz individual se utilizó la siguiente expresión de Sensibilidad, obtenida desde la referencia (Bauza y col., 2012):

- Sensibilidad (SenMCR):

$$\mathrm{SenMCR} = m_n \left[J \left(\mathbf{S}^{\mathbf{T}} \mathbf{S} \right)^{-1}_{nn} \right]^{-1/2} \quad (2)$$

donde n es el índice del analito de interés en una mezcla de múltiples componentes, m_n es la pendiente de una recta de calibración pseudo-univariada para el mismo analito según áreas resueltas con MCR-ALS, J representa la cantidad de canales por matriz en el sentido de apilamiento (tiempos

o pocillos) y $\mathbf{S^T}$ es la matriz de perfiles espectrales resueltos para todos los componentes de la mezcla. El subíndice *nn* indica que de la matriz $(\mathbf{S}^T\mathbf{S})^{-1}$, debe tomarse el elemento (*n*,*n*).

Con esta cifra definida, se derivan las siguientes:

- Inversa de Sensibilidad Analítica (InvSenAn):

$$\text{InvSenAn} = \frac{\text{desvSt}}{\text{SenMCR}} \quad (3)$$

donde desvSt es la desviación estándar de las predicciones para muestras replicadas.

- Límite de Detección (LOD, del inglés *Limit Of Detection*):

$$\text{LOD} = 3.3\,\text{InvSenAn} \quad (4)$$

- Límite de Cuantificación (LOQ, del inglés *Limit Of Quantitation*):

$$\text{LOQ} = 10\,\text{InvSenAn} \quad (5)$$

3.5 Materiales y Métodos

La determinación de las condiciones experimentales y sus parámetros, en cuanto a la separación cromatográfica de las muestras mediante HPLC (condiciones de corrida, longitudes de onda UV), sus concentraciones y su obtención (muestras de Calibración y Validación), la elección de los analitos, las condiciones en las lecturas de fluorescencia (velocidad de adquisición, longitudes de onda de Excitación-Emisión, *slits*, entre otros) y la decisión de interpolar datos de fluorescencia mediante *spline* y suavizarlos con polinomios de Savisky-Golay, fueron producto del trabajo de investigación desarrollado por Mirta Alcaraz.

La construcción del esqueleto mecánico del recolector de muestras, la selección de partes a reciclar y la disposición de éstas (y otras compradas) en el esqueleto, el diseño y construcción de un puente móvil de Aluminio, de una pieza de acrílico necesaria para enhebrar el capilar de recolección y de una base metálica para actuar como soporte del dispositivo, fueron producto del trabajo de Gabriel Gómez.

3.5.1 Reactivos y solventes

Todos los estándares fueron de grado analítico. El fármaco Ofloxacina (OFL) fue provisto por Sigma (Alemania) y los fármacos Ciprofloxacina (CPF) y Danofloxacina (DNF) fueron provistos

por Fluka (Suiza).

Metanol (MeOH) y Acetonitrilo (ACN), ambos de grado LC, fueron obtenidos desde J.T. Baker (Holanda). Se obtuvo agua ultrapura con un sistema de purificación de agua Mili-Q Millipore (Estados Unidos). Ácido Acético (HAc) fue provisto por Cicarelli (Argentina) y Acetato de Sodio tri-hidratado (NaAc) fue provisto por Anedra (Argentina).

3.5.2 Soluciones y muestras

A partir de los reactivos sólidos, se prepararon soluciones stock de cada fármaco en MeOH con concentraciones de 200.00 ppm, las cuales se mantuvieron refrigeradas a 4ºC en oscuridad. Las soluciones estándar de trabajo se obtuvieron cada día en el cual se realizaron experiencias, en base a diluciones apropiadas con agua y volúmenes finales de 2.00 mL. De esta forma se obtuvieron las muestras de Calibración y Validación expuestas en la tabla 1.

Calibración				Validación			
Muestra	OFL (ppm)	CPF (ppm)	DNF (ppm)	Muestra	OFL (ppm)	CPF (ppm)	DNF (ppm)
OFL2	2.00	0.00	0.00	Val1	8.83	13.24	0.79
OFL4	4.00	0.00	0.00	Val2	6.00	9.00	1.50
OFL6	6.00	0.00	0.00	Val3	2.00	9.00	1.50
OFL8	8.00	0.00	0.00	Val4	3.17	9.00	1.04
OFL10	10.00	0.00	0.00	Val5	10.00	9.00	1.50
CPF3	0.00	3.00	0.00	Val6	3.17	13.24	2.21
CPF6	0.00	6.00	0.00	Val7	6.00	13.24	0.50
CPF9	0.00	9.00	0.00	Val8	6.00	15.00	2.50
CPF12	0.00	12.00	0.00	Val9	6.00	9.00	1.50
CPF15	0.00	15.00	0.00	Val10	8.83	4.76	2.21
DNF05	0.00	0.00	0.50	Val11	6.00	15.00	1.50
DNF10	0.00	0.00	1.00	Val12	6.00	9.00	1.50
DNF15	0.00	0.00	1.50	-	-	-	-
DNF20	0.00	0.00	2.00	-	-	-	-
DNF25	0.00	0.00	2.50	-	-	-	-

Tabla 1: Composición de muestras de Calibración y Validación

Referencias: Los valores subrayados representan a las muestras utilizadas como replicados para el cálculo de cifras de mérito según calibraciones pseudo-univariadas

Las concentraciones para las muestras de Calibración se obtuvieron empíricamente de forma tal de poder obtener señales apropiadas de fluorescencia tras una separación en HPLC, teniendo en cuenta que la cromatografía diluye también a las muestras. Definidas las concentraciones mínimas de cada analito, se obtuvieron cuatro niveles de concentración extra para cada uno y luego la composición de las muestras de Validación se derivó a través de un Diseño Central Compuesto. A su vez, en éste último conjunto existen muestras con concentraciones repetidas para cada fármaco, y dichas muestras fueron utilizadas como replicados para obtener desde sus predicciones (en calibraciones pseudo-univariadas) información de variabilidad utilizada para obtener cifras de mérito.

3.5.3 Programas

Para la operación del HPLC se utilizó ChemStation, programa provisto por el fabricante. El tratamiento de datos y los cálculos en general fueron realizados en Matlab (MATLAB *7.6.0*, 2008), donde además se aplicó MCR-ALS a través de una interfaz gráfica (Jaumot y col., 2005) o bien desde línea de comandos. También se utilizaron tres programas de descarga gratuita y código abierto: Arduino IDE 1.0.1 para programar a la placa Arduino, Processing IDE 2.0b6 para la elaboración de interfaces gráficas y comunicación en serie con la placa Arduino y con el fluorímetro, y Fritzing 0.7.12b para el modelado de circuitos y la obtención de impresiones para transferir circuitos a placa. Los tres programas pueden ser descargados desde sus respectivas comunidades virtuales, donde además son permanentemente actualizados. Todos los enlaces a Internet mencionados en el presente texto se encuentran funcionales a la fecha (Septiembre, 2013).

3.5.4 HPLC-UV y recolección de fracciones en placas de ELISA

Ante todo, la velocidad de flujo fue determinada no sólo para obtener una separación cromatográfica aceptable, sino también teniendo en cuenta que en el capilar de descarte del HPLC debían obtenerse gotas separadas, pues de obtenerse un flujo continuo se perdería muestra al cambiar los pocillos de recolección en la placa de ELISA. Este valor fue fijado en 1.8 mL/min (30 μL/s).

Los estudios cromatográficos se realizaron en un instrumento Agilent modelo 1100 (Alemania), con detector de arreglo de diodos UV-visible, y controlado con el programa ChemStation, también

utilizado para adquirir matrices con espectros UV entre 200 nm y 400 nm, cada 1 nm.

La columna analítica fue una Zorbax Eclipse XDB-C18, de 75 mm × 4.6 mm, con un tamaño de partícula de 3.5 μm. La temperatura de la columna fue controlada en 35ºC. La fase móvil consistió de 10 mmol/L de buffer de HAc (pH=4.0) – MeOH – ACN (71:9:20, v/v). Todas las muestras fueron filtradas a través de un filtro de membrana de nylon de 0.22 μm, se inyectaron 100 μL y la elución se realizó en modo isocrático. El análisis completo requirió 2 minutos.

La recolección automatizada de muestras a partir del descarte del HPLC se realizó adjuntando un capilar de aproximadamente 20 cm a la salida del instrumento. La longitud del capilar de recolección fue la mínima necesaria para poder depositar líquido en el más lejano de los pocillos de una placa de ELISA según el diseño del recolector. No se optó por un capilar de mayor longitud para evitar o minimizar posibles efectos de mezcla/dispersión que podrían producirse en las eluciones, ya que el capilar no posee el mismo relleno que la columna analítica, sino que se utilizó vacío. Las recolecciones definitivas fueron realizadas cada 2 segundos, con 17 pocillos por muestra, y para cada muestra se utilizó una placa de ELISA diferente. Todas las recolecciones comenzaron 47 segundos después del momento de la inyección en el HPLC. Esto se determinó empíricamente y con el objetivo de que existieran concentraciones apreciables del primer analito en eluir (OFL) en el pocillo 2 de cada placa (es decir, el pocillo 1 se incluyó por precaución, pero era sabido que si no existían corrimientos severos en la cromatografía, en el pocillo 1 no deberían encontrarse los analitos, sino recién en el 2).

Cada vez que una recolección finalizaba se procedía a obtener datos en el fluorímetro. En el caso en que éste estuviera aún ocupado con una placa previa, las placas en espera eran recubiertas con un film para evitar evaporaciones.

Vale aclarar que el HPLC en cuestión también puede obtener datos de Emisión-Excitación a través del uso de un detector de fluoresencia incorporado cuya sensibilidad respecto de los analitos en estudio fue superior a la del fluorímetro utilizado (datos no mostrados) debido fundamentalmente a que es más moderno y su fotomultiplicador es mejor. Sin embargo, esto sólo es útil para obtener espectros de Excitación o de Emisión, pero no matrices.

3.5.5 Lectura de fluorescencia

Todas las medidas de fluorescencia se realizaron en un Espectrómetro de Luminiscencia Perkin-Elmer LS55 (Reino Unido), equipado con un accesorio para leer placas de ELISA acoplado a una fibra óptica y a un fotomultiplicador, y operado desde una PC a través de una conexión en serie RS232C. El lector de placas consta de una base para insertar una placa y posee un sistema electromecánico (operable a través de la PC de control) de posicionamiento de la placa bajo la fibra óptica, lo cual permite leer los pocillos por separado. Para cada uno de los 17 pocillos recolectados en cada placa de ELISA se obtuvieron matrices de Excitación-Emisión mediante la interfaz gráfica creada en este trabajo. Las señales se obtuvieron entre 260 nm y 340 nm (cada 5 nm) para Excitación y entre 380 nm y 500 nm (cada 5 nm) para Emisión. Los *slits* de ambos monocromadores fueron configurados en 10 nm, el voltaje del detector fue de 600V y la velocidad de adquisición fue de 900 nm/min.

Las placas de ELISA utilizadas fueron blancas y aptas para estudios de fluorescencia. Su tamaño es el mismo que el del estándar clásico de 96 pocillos.

3.5.6 Componentes electrónicos y electromecánicos

De una impresora Canon BJC 1000 se obtuvieron algunos componentes para el recolector de muestras, como ser la fuente incorporada que transforma 220V de corriente alterna en 24V de corriente continua, y sus dos motores paso a paso (pap) de 24V. El pap responsable del movimiento a través de las columnas de una placa de ELISA puede dar 96 pasos/vuelta y es del tipo bipolar, mientras que el encargado de las filas es unipolar con 48 pasos/vuelta. A través de bibliotecas de programación se optó por realizar pasos medios, por lo cual ambos pap duplican la cantidad de pasos/vuelta (esto será descripto oportunamente). También se recicló la estructura metálica que permite el movimiento del carro de la impresora, lo cual incluyó a su sistema de correas y engranajes, y a un eje auxiliar. El carro fue modificado para mover un puente móvil de Aluminio en lugar de cartuchos de tinta.

Componentes varios como LEDs, cables, mangueras para datos y fichas hembra para conectores en placa fueron reciclados fundamentalmente de impresoras y PC en desuso. Vale destacar que no quedó otra opción que extraer y reciclar las fichas hembra pues la disponibilidad para su compra en la ciudad (Santa Fe) fue nula. También desde impresoras fueron rescatados sensores de presencia de

hoja, los cuales fueron puestos a prueba para contar gotas a la salida del capilar de recolección, aunque finalmente se decidió excluirlos (ver texto). Otros componentes fueron obtenidos en casas de electrónica de la ciudad, como resistencias de carbón, diodos de bajo voltaje, hileras de pines, botones tipo *switch,* el integrado ULN2803A y la placa virgen de PCB-Cobre (PCB proviene del inglés *Printed Circuit Board*, Ciruito Impreso en Placa).

La placa Arduino UNO (revisión 3) y el controlador de motores (*driver*, con integrado L293D) fueron adquiridos en Mercadolibre [www.mercadolibre.com.ar] desde Buenos Aires a través de un vendedor particular.

3.5.6.1 Metodología para obtención de placa tipo *shield*

En este contexto, un *shield* es un tipo de circuito impreso en placa cuyo diseño permite un encastre adecuado a la forma y pines de una placa Arduino. A través de Fritzing y teniendo en cuenta la distancia entre contactos de componentes que fueron reciclados fundamentalmente desde impresoras, se modeló el circuito y se obtuvo una impresión en papel fotográfico con impresora láser a tóner. Dicha impresión debe ser transferida a una placa virgen de PCB-Cobre previamente lijada, limpiada con alcohol y posteriormente secada. La transferencia se realiza a través del uso de una plancha común en el máximo de temperatura posible. Luego de esperar unos 15 minutos, se pone un objeto pesado y plano sobre el conjunto placa-impresión, y se espera hasta que enfríe. Se sumerge en agua y se extrae el papel con cuidado de no levantar las pistas transferidas. El secado con aire caliente evidenciará si quedaron restos de papel que deben ser removidos, ya que de quedar intactos entre dos pistas éstas harán contacto. En caso de que una parte necesaria se haya desprendido, es posible completar manualmente el circuito con un fibrón indeleble. El conjunto debe ser sumergido en solución para quemado de placas ($FeCl_3$, “ácido para placas” en las ferreterías). El recipiente contenedor debe ser plástico y de ser posible se puede agregar un baño María con agua caliente para acelerar el proceso. Todo es removido sutilmente cada tres minutos aproximadamente. Luego de unos doce minutos se extrae y lava con agua. Se verifica que el Cobre haya sido correctamente atacado, y de no darse esto, se vuelve al ataque ácido. Logrado lo anterior, se lija con lija al agua, se realizan perforaciones con taladro (se utilizaron mechas de 0.75 mm y 1 mm, según los contactos de los componentes reciclados), se controla con multímetro (*tester*) que no existan contactos indebidos y finalmente se aplica decapante en aerosol, el cual protegerá al Cobre de la oxidación y favorecerá el último paso necesario de soldadura de componentes con puntos de

Estaño.

Una vez que todos los componentes fueron soldados, se corroboró el correcto encastre del *shield* en la placa Arduino UNO y finalmente se conectaron con cables los contactos en el *shield* con los componentes apropiados (motores, botones, fuente de 24V, entre otros).

3.5.7 Arduino: IDE, bibliotecas y modelo UNO

Todos los modelos de placas (UNO, Mega, Nano, DUE, entre otras variantes) utilizan una conexión USB para comunicarse con una PC y a través de ésta pueden ser programadas. Es necesario obtener el Entorno Integrado de Desarrollo (IDE), el cual puede ser libremente descargado desde [http://arduino.cc/en/Main/Software] y luego descomprimir los archivos en una carpeta a elegir (es decir, el IDE no se instala, sino que se copia y ejecuta). Su funcionamiento es posible en Linux, Windows y Mac OS, y en caso de que se requieran controladores estos son provistos. La versión del IDE utilizado fue la 1.0.1. Una vez que el IDE está listo y la placa está conectada a la PC, es necesario seleccionar el puerto de conexión y el modelo de placa conectada. Hecho lo anterior, ya será posible programar en el IDE a través de la escritura en archivos denominados *sketchs* (de extensión PDE y antiguamente INO), cuyo contenido podrá ser grabado en la placa Arduino en uso previa compilación del código a lenguaje binario (se realiza simplemente pulsando un botón, tanto la compilación como el grabado del código binario). Como cualquier otro entorno de programación (por ejemplo Matlab) está dividido en regiones (menúes, consola, región de accesos directos a operaciones comunes como verificar código o grabarlo en la placa, entre otras), dentro de las que se destaca el editor de textos donde se escriben las sentencias de programación y donde ciertas palabras clave propias del lenguaje son resaltadas en colores o tonos diferentes.

A su vez, a través del IDE es posible acceder a un gran conjunto de ejemplos para aprender a utilizar distintas capacidades de las placas y las sentencias apropiadas. Además de las funciones digitalRead y digitalWrite que serán posteriormente mencionadas, vale destacar que se utilizó la función delay(t), con t siendo un entero representando milisegundos. Esta función es útil para realizar esperas luego de realizar lecturas digitales (50 milisegundos suelen ser suficientes), para hacer que el capilar de recolección repose sobre un pocillo determinado de la placa de ELISA, y en general para detener la ejecución de código por un determinado tiempo. En relación a estos tiempos

de espera, es necesario aquí entender una diferencia entre la forma en que se programa en Arduino respecto de formas clásicas de programación como en Matlab. En el último la programación de una tarea (usualmente un cálculo) implica una sucesión de líneas de programación ordenadas. La culminación de una línea implica el inmediato comienzo de la siguiente, y no hay razón para pensar que se requiera un tiempo de espera entre un comando y el siguiente. Cuando se terminan las líneas, culmina el proceso. En cambio en Arduino existe una división obligatoria en cada *sketch* que contiene a las sentencias, ya que cada uno debe tener una función denominada *setup* y otra denominada *loop*. Antes de *setup* las variables que van a utilizarse suelen declararse (reservar nombres) o inicializarse (reservar nombres y asignar un primer valor a cada variable). En *setup* se establecen los modos de salida/entrada de los pines, el inicio de comunicaciones en serie con una PC y cualquier otra cosa que requiera ser configurada antes de ser usada. Todo el código que se encuentre dentro de la función *setup* se ejecutará una única vez cada vez que la placa sea encendida. En cambio, las líneas programadas dentro de *loop* se ejecutarán indefinidamente hasta que se corte el suministro eléctrico. Una vez que se culmine la ejecución de la última, se comenzará nuevamente por la primera y el ciclo volverá a repetirse (de allí *loop*). A su vez, entre línea y línea puede ser necesario un tiempo apropiado para esperar respuestas de sensores que aunque a veces lo parezcan, no son inmediatas. En algunos casos existen funciones denominadas bloqueantes, es decir que bloquean la ejecución hasta que la función culmine su proceso.

Una biblioteca (*library*) es un conjunto de instrucciones relacionadas a una determinada tarea, de forma similar a los *toolboxes* de Matlab. La bibliotecas permiten reutilizar código de forma prolija, evitando errores, y permiten la aplicación de funciones de forma muy sencilla, lo cual no sería así si uno mismo cada vez tuviera que programar todo desde cero. Su inclusión dentro de un *sketch* es muy sencilla, ya que sólo bastará escribir al comienzo la sentencia #include <nombre.h>, donde nombre representa el nombre de la biblioteca. Existen muchas bibliotecas estándar incluidas en el IDE, y también es posible agregar bibliotecas programadas por usuarios de la comunidad internacional.

Para entender el concepto, vale tomar como ejemplo una de las muchas bibliotecas estándar que incluye Arduino y que fue utilizada en este trabajo, la biblioteca *Stepper* [http://arduino.cc/en/Reference/Stepper?from=Tutorial.Stepper], apta para controlar motores paso a paso (denominados *steppers*). Para realizar movimiento (pasos) con estos motores y sin usar

bibliotecas, es necesario que determinadas señales digitales permitan que las bobinas del motor se energicen siguiendo cierta secuencia para un determinado sentido de giro, mientras que otra secuencia será necesaria solamente para invertir el giro. A su vez, la velocidad de giro estará determinada por el tiempo entre paso y paso, por lo cual son necesarias funciones de control de tiempo (*timing*). Además debería escribirse código para asociar cada pin de Arduino con una bobina respectiva. Como puede intuirse, programar desde cero todo esto sería una tarea tediosa y muy sujeta a errores durante la escritura de código. En cambio, una vez que en un *sketch* se ha declarado el uso de la biblioteca *Stepper*, sólo serán necesarias las siguientes sentencias:

- Antes de *setup* (declaración/inicialización de variables): Stepper mimotor(pasosXvuelta, 8,9,10,11). La sentencia anterior creará una variable llamada "mimotor" (puede ser otro nombre) que estará asociada a un motor paso a paso con "pasosXvuelta" pasos/vuelta y cuyas bobinas se encontrarán conectadas a nivel lógico a los pines 8, 9, 10 y 11 de Arduino.

- En *setup* (configuración): mimotor.setSpeed(60). Esto hará que el motor controlado se mueva a 60 revoluciones/minuto. Llamados posteriores a mimotor.setSpeed(velocidad) permitirán el cambio de velocidad. Estos cambios pueden realizarse en *loop*, pero al menos debe existir una primera velocidad configurada en *setup*.

- En *loop* (repeticiones): mimotor.step(N). Esto hará que el motor realice N pasos en sentido horario (-N para movimientos antihorarios). La función step es un ejemplo de función bloqueante, por lo que la siguiente sentencia luego de step no será ejecutada hasta que los N pasos se hayan realizado.

Como puede apreciarse, con unas pocas sentencias será posible realizar todo tipo de movimientos a cualquier velocidad físicamente permitida. De aquí la gran utilidad de las bibliotecas.

En cuanto a la placa Arduino en sí, la figura 1 obtenida mediante Fritzing esquematiza al modelo UNO utilizado. Un detalle minucioso de todos los componentes y funciones puede ser obtenido desde [http://arduino.cc/en/Main/arduinoBoardUno]. En este contexto se pretende destacar a los siguientes:

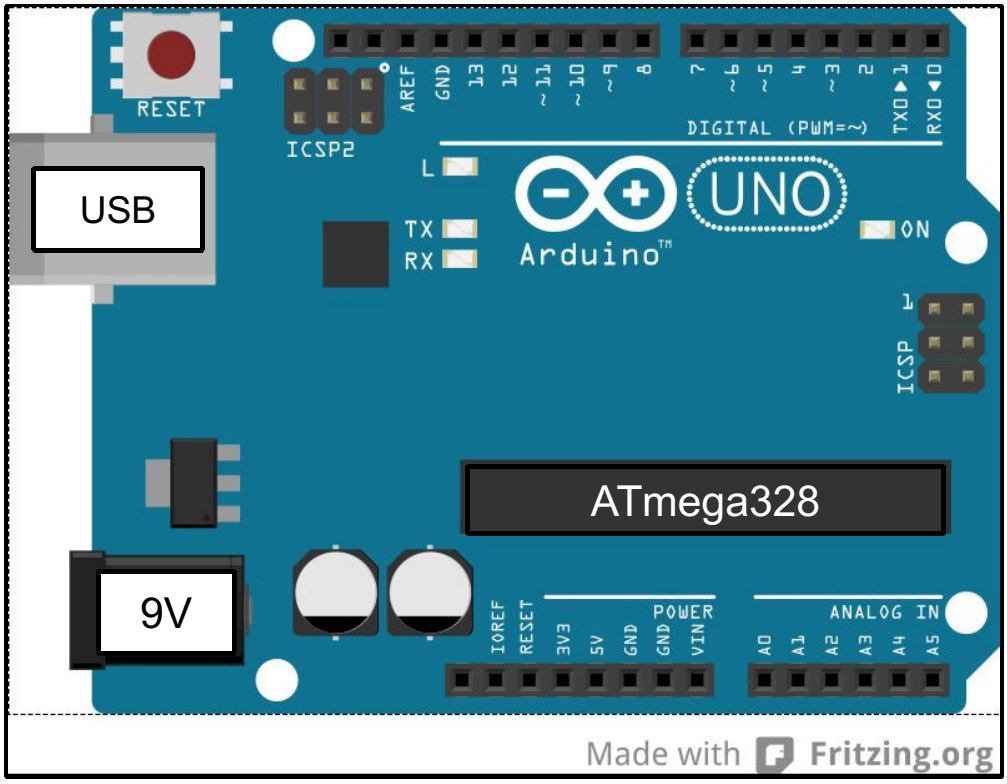


Figura 1: Esquema de Arduino UNO revisión 3

- Conexión USB: Posee un integrado Atmega16U2 que permite realizar la conversión entre USB y serie (RS-232). Con un cable común USB es posible conectar la placa a cualquier computadora. Por esta vía los *sketchs* serán grabados en la memoria de la placa para su posterior ejecución. También por este medio es posible realizar una comunicación en serie entre la placa y una PC, tal y como con el fluorímetro, y utilizarla desde una interfaz gráfica para enviar instrucciones a la placa. La conexión USB a su vez permite alimentar de energía a los circuitos básicos.

- Entrada de alimentación (9V en el esquema): Una vez que se ha grabado un programa, la conexión USB es prescindible y la alimentación puede realizarse directamente con un transformador. Según las especificaciones, se recomiendan voltajes de entrada de entre 7V y 12V, aunque se admiten entre 6V y 20V, más allá de que el voltaje interno de operación es de 5V (por ese motivo la conexión USB con 5V es suficiente).

- Microcontrolador ATmega328: Todas las funciones principales de la placa estarán comandadas por este microcontrolador, el cual será el encargado de ejecutar el código programado en su

memoria (re-escribible). Vale destacar que las placas como Arduino UNO están pensadas para elaboración de prototipos, pero una vez que se ha definido la configuración final de los componentes envueltos en un proyecto, es probable que no vayan a ser utilizadas muchas de las capacidades de Arduino (por citar un ejemplo, si el proyecto sólo consta de prender y apagar un LED, entonces no sería necesaria una conexión USB). En estos casos, existe la opción de grabar las instrucciones en múltiples microcontroladores ATmega328 y luego éstos son utilizados como cualquier otro componente en placas elaboradas con los requisitos mínimos para la ejecución de las instrucciones programadas, lo cual reducirá los costos significativamente si el proyecto debe producirse de forma masiva. Dadas las características filosóficas del proyecto Arduino, las placas ensambladas suelen tener muy bajo costo y entonces se las suele utilizar directamente para realizar dispositivos terminados cuando éstos son de fabricación única (como en este trabajo).

- Pines Digitales de Entrada y Salida (I/O, *Digital Input/output pins*): En el esquema, éstos se encuentran en la parte superior, numerados del 0 al 13. Para configurar a los pines como salidas se usa la función pinMode(PIN, 'OUTPUT'), y para hacerlo como entradas su equivalente pinMode(PIN,'INPUT'), siendo PIN el número que lo representa. La escritura y la lectura de un pin digital se realiza con las funciones digitalWrite(PIN, VALOR) y VALOR=digitalRead(PIN), respectivamente. En ambos casos VALOR será HIGH (5V) o LOW (0V). En el caso de las salidas, éstas podrán proveer hasta 40mA de corriente continua (apto para alimentar componentes de bajo consumo como LEDs). En la programación del recolector de muestras todos los pines fueron utilizados como digitales. Vale destacar que los pines marcados con " ~ " (pines 3, 5, 6, 9, 10 y 11) son capaces de emitir salidas de voltajes intermedios entre 0V y 5V, pudiendo dividir ese intervalo en 2^8 partes (8 bits de resolución), a través de Modulación por Ancho de Pulsos (PWM). Esta técnica consiste en alternar el estado de un pin entre dos posibles, 0V y 5V, con diferentes tiempos de permanencia entre uno y otro. Así por ejemplo si el tiempo es el mismo para ambos, el voltaje obtenido será de 2.5V. Otros pines digitales merecen ser comentados. Se recomienda tratar de evitar el uso de los pines 0 (R, receptor) y 1 (T, transmisor), ya que éstos son necesarios para establecer la comunicación en serie con una PC, aunque serán posibles de ser utilizados si durante la ejecución de un programa no se requiere tal comunicación. Finalmente, vale destacar al pin 13, el cual ya tiene incorporado un LED (L en la placa del esquema), por lo que éste pin suele utilizarse para verificar lógicas programadas, por ejemplo que si se supera un valor en determinado sensor entonces el LED sea encendido como señal.

- Pines Analógicos de Entrada (*Analog Input pins*): En el esquema son vistos en la parte inferior derecha y son denominados A0-A5. Permiten leer voltajes entre 0V y 5V con una resolución de 10 bits y suelen utilizarse para medir señales no digitales (por ejemplo la intensidad de luz en un ambiente). Estos pines también pueden ser utilizados para lectura/escritura digital y son numerados desde el 14 (A0) hasta el 19 (A5).

- Pines "Power": En el esquema se observan en la parte inferior, a la izquierda de los pines de Entrada Analógica. GND es la conexión de masa, 5V es un contacto por el que constantemente se emiten 5V y 3v3 lo hace con 3.3V (así como existe lógica de programación entre 0V y 5V, existen dispositivos cuya lógica equivalente es entre 0V y 3.3V, y de hecho éste es el caso de la placa Arduino DUE, posterior al modelo UNO). Si la alimentación de la placa no se realiza por USB sino por la entrada marcada con 9V en el esquema, entonces el pin VIN puede emitir dicho voltaje de alimentación constantemente.

3.5.8 Processing: IDE y bibliotecas

El IDE utilizado corresponde a la versión 2.0b6 y se obtuvo de forma gratuita a través de [https://processing.org/download/?processing]. Existen versiones para Linux, Windows y Mac OS, para arquitecturas de 32 y 64 bits. El IDE puede ser utilizado para ejecutar las interfaces diseñadas, o una vez finalizado el diseño es posible elegir una arquitectura y un sistema operativo objetivo, lo cual seguido de su compilación resulta en un programa ejecutable, independiente de Processing.

Desde este IDE también se puede acceder a un gran conjunto de ejemplos para ejecutar distintas funciones. Vale destacar que así como en los *sketchs* de Arduino existe la función *loop*, en Processing debe existir obligatoriamente la función *draw* que también se repetirá indefinidamente. En nuestro caso, como Processing fue utilizado para elaborar interfaces gráficas (una para comunicación en serie con una placa Arduino y otra para el fluorímetro), lo que se repite constantemente en la función *draw* está relacionado con cambios en los controles gráficos (por ejemplo que el usuario haya insertado un nuevo valor para alguno de los parámetros, que haya dado la orden de iniciar una recolección o la obtención de una matriz de fluorescencia) y para evaluar permanentemente si existe algún mensaje que deba ser enviado o recibido en las comunicaciones en serie.

En cuanto a bibliotecas, se utilizaron dos. Una de ellas fue la denominada Serial, estándar en

Processing, necesaria para comunicar a las interfaces gráficas tanto con Arduino como con el fluorímetro. En ambos casos se requiere saber en qué puerto de la PC se encuentran conectados los dispositivos, y a su vez es necesario establecer la velocidad de transferencia, la cual fue para ambos de 9600 bps (bits/segundo). La otra biblioteca utilizada fue escrita por Andreas Schlegel y se denomina controlP5 [http://www.sojamo.de/libraries/controlP5/download/controlP5-2.0.4.zip]. Esta biblioteca puede adjuntarse a las estándar de Processing y provee componentes para realizar interfaces de usuario, como ser barras deslizantes, listas desplegables, botones, entre otros.

3.6 Resultados y Discusión

3.6.1 Obtención de una interfaz gráfica para operar el fluorímetro

Lo primero en ser realizado fue la lectura del manual del instrumento, específicamente un capítulo dedicado a la comunicación con éste. De esto se obtuvo la siguiente información relevante:

- Comunicación: normas RS-232C a 9600 bps (bits por segundo)
- Longitudes de onda válidas: Excitación entre 200 y 800 nm, Emisión entre 200 y 900 nm.
  - Según el manual, deben ser múltiplos de 0.1 nm. No obstante, al configurar valores y leer posteriormente la real posición de los monocromadores, se verificó que éstos no pueden ser dirigidos a cualquier posición, ya que la posición final de ambos será múltiplo de 0.2 nm, no de 0.1 nm. En el caso de Excitación, los valores siempre resultan disminuidos (por ejemplo, si se decide enviarlo a 300.1nm o a 300.9 nm, será dirigido a 300.0 nm y a 300.8 nm). En el caso de Emisión no se encontró una lógica de redondeo (a veces hay disminución, otras aumento) pero siempre se dirige al monocromador a un múltiplo de 0.2 nm. Esto no significa que los monocromadores no estén posicionados realmente donde se los envía, pero al menos significa que el comando utilizado para consultar la posición no se condice con lo dicho en el manual. Por esta razón al recolectar datos y realizar un reporte automático de la recolección se explicitan los valores donde estuvieron los monocromadores según el comando de consulta, y no donde supuestamente quizá pretendía el usuario. Para evitar inconvenientes, lo mejor es configurar los barridos con intervalos múltiplos de 0.2 nm.
- Velocidad de barrido: entre 10 y 1500 nm/min
- Los datos de salida tienen corregidos los valores de Excitación (a través de una tabla de

factores de corrección alojada en la memoria no volátil del instrumento), son promediados en bloque, y pueden ser pre-procesados con otros métodos que no fueron utilizados.
  - Entre éstos, es posible aplicar filtros de Savitzky-Golay para suavizado, pero esto fue realizado con los datos ya colectados.
- Los *slits* pueden configurarse independientemente. Para el caso de Excitación, puede tomar valores de 0 o entre 2.5 y 15 nm, mientras que para Emisión estos valores pueden ser 0 o entre 2.5 y 20 nm, siempre en múltiplos de 0.1 nm.
- Existen comandos para realizar calibraciones de partes del instrumento de forma automática
  - Es válido mencionar que estos comandos no se encuentran accesibles para el usuario si se utiliza la interfaz gráfica provista por el fabricante. Cuando se realizaron pruebas con estos comandos, ni el slit ni el monocromador de Emisión pudieron se calibrados porque el instrumento requería una lámpara de mercurio para tal fin. Sin embargo, para calibrar el slit y el monocromador de Excitación solamente fue necesario insertar agua destilada en una cubeta en el momento requerido por el instrumento y esperar aproximadamente 10 minutos hasta recibir en pantalla un mensaje de finalización. Tras este procedimiento las lecturas comenzaron a ser un poco mejores que antes de éste, por lo cual se diseñó una pequeña interfaz gráfica (no se discutirá) para aplicar las calibraciones en el futuro.
- El voltaje del fotomultiplicador se configura automáticamente en función de otro voltaje dependiente de la posición del slit de Excitación y de la posición del monocromador de Emisión.
- El lector de placas ELISA puede ser dirigido por pocillo (1-96) o por posición (valores X e Y en mm)

La comunicación con el fluorímetro se realizó a través de una conexión según normas del estándar RS-232C (Electronic Industries Association, 1969), antiguamente utilizado en dispositivos de computadoras (*mouse*, MoDem, entre otros) y actualmente desplazado por el estándar USB, aunque en equipos de laboratorio siga utilizándose ampliamente. De forma muy resumida se puede decir que éste estándar establece características eléctricas que deben cumplir las señales a ser recibidas y transmitidas, como ser sus valores binarios de voltaje y la velocidad del traspaso entre dispositivos comunicados; características físicas para la construcción de interfaces, como la disposición de conectores ("pines") machos y hembras; y características funcionales para cada pin.

De los últimos, interesan especialmente dos de ellos, el de Transmisión (T) y el de Recepción (R). Para que dos dispositivos se conecten en serie, el T de uno debe conectarse al R del otro y viceversa si se pretende que ambos dispositivos puedan tanto transmitir como recibir. En nuestro caso, el fluorímetro está conectado a una PC para su control, por lo cual la interfaz gráfica diseñada y ejecutada en la misma PC no hace más que acceder a los pines T y R del respectivo puerto serie de la PC, con lo cual puede comunicarse con el fluorímetro. A través de éstas conexiones se enviarán valores altos y bajos de voltaje en ambos sentidos. Si ambos dispositivos cumplen el estándar, podrán decodificar estos valores binarios (bits) y será posible el intercambio de mensajes. No obstante, será necesario establecer la velocidad de transferencia en bits por segundo (bps o b/s), la cual tomó el valor de 9600 bps dado que así lo indicaba el manual del fluorímetro. Ya que es posible que una PC esté conectada por más de un puerto a distintos dispositivos, será necesario saber qué puerto de la PC se encuentra conectado al fluorímetro (en PC con Windows, los puertos serie son denominados “COM”) y pasarlo como parámetro a la interfaz gráfica.

Desde luego, aunque los dispositivos se encuentren conectados no podrán comunicarse si desconocen el formato de los mensajes. En nuestro caso, esto lo impuso el fabricante del fluorímetro, por lo cual fue necesario realizar ensayos preliminares para verificar el formato en que debían ser enviados/recibidos los mensajes y las sentencias específicas necesarias para obtener matrices de Emisión-Excitación de cada pocillo de una placa de ELISA (el instrumento tiene muchas otras capacidades con sentencias respectivas, las cuales no serán reportadas). Según el manual, se obtuvo la siguiente información relevante:

- El instrumento siempre está listo para recibir datos y todas las respuestas obtenidas desde éste son producto de comandos previamente enviados. El instrumento arroja códigos como respuesta y las consultas deben realizarse a través de la sintaxis “$COMANDO PARÁMETROS TERMINADOR”. COMANDO identifica una operación a través de 2 letras (por ejemplo GM para posicionar a los monocromadores), parámetros es una lista de valores separados por coma (propios de cada comando, como ser las longitudes de onda objetivo para los monocromadores) y terminador es el código que representa una nueva línea (\n)
- La salida de datos desde el instrumento podrá estar compuesta de hasta 126 caracteres ASCII, terminados en “\n”. La primera y la última respuesta a cualquier comando es un código de error de 4 dígitos, y entre ambas pueden recibirse datos. Cualquier salida que no sólo conste de un código de error comienza con el carácter DEL (suprimir).

- A excepción de algunos comandos indicados, cuando un comando que requiera parámetros sea enviado sin éstos, la respuesta del instrumento será el valor actual de dichos parámetros, separados por coma de ser necesario, comenzando y terminando la cadena de caracteres con 0000. Esto es útil para realizad consultas al instrumento.

Esta información debió ser tenida en cuenta para la elaboración de cadenas de caracteres durante el envío de consultas, así como para interpretar y procesar las respuestas recibidas. A su vez, el manual dispone de una lista de códigos de error, algunos de los cuales fue necesario conocer para elaborar la lógica de comunicación:

- 0000: Sin error
- 0110: Instrumento ocupado (busy)
- Muchos otros específicos

Habiendo establecido adecuadamente el puerto y la velocidad (9600 bps), se utilizó un archivo de pruebas en el cual se realizó la comunicación en serie entre Processing y el fluorímetro. Processing posee una zona gráfica denominada Consola, en la cual se pueden imprimir en pantalla los comandos enviados y las respuestas recibidas. A través de la Consola, se pusieron a prueba comandos potencialmente útiles para lograr los objetivos, algunos de los cuales son destacados en la tabla 2. En esta tabla se puede apreciar a los comandos agrupados según sus funciones generales. Los primeros son simples configuraciones que permitirán la adquisición de espectros con la orden SC (ver aparte). En base a lo observado sobre la posición deseada y la real de los monocromadores, es conveniente que los valores para AH, AL y SI sean múltiplos de 0.2 nm. Los comandos misceláneos fueron muy útiles. ES permite usar eficientemente a la fuente, de forma tal de mantenerla encendida lo mínimo indispensable. AB sirve para cancelar las operaciones si se detectan errores o si el usuario lo determina. ST retorna una cadena de caracteres ordenados según el fabricante (que denomina a esta cadena *status record*), desde la cual puede obtenerse información actualizada sobre el instrumento, como se cuál es el monocromador actualmente seleccionado, cuáles son las posiciones de los monocromadores, sus *slits*, etc. De gran utilidad fue el comando BE 6161, el cual retorna los últimos 100 comandos recibidos por el instrumento, sin importar si éstos fueron o no enviados desde la interfaz gráfica provista por el fabricante. Esto fue muy relevante cuando se quisieron poner a prueba los comandos para movimiento automatizado de las placas.

Comando	Función
Configuración para adquirir espectros	
$MX 0	Seleccionar un monocromador (0=Ex, 1=Em, 2=Ambos)
$AH 800.0	Establecer longitud de onda superior para un monocromador
$AL 200.0	Establecer longitud de onda inferior para un monocromador
$SI 3.0	Establecer incremento para barridos (entre 0.1 y 5 nm)
$SS 400	Establecer velocidad de barrido (entre 10 y 1500 nm/min)
$SX 10	Establecer *slit* de Excitación (0, o entre 2.5 y 15 nm, cada 0.1 nm)
$SM 20	Establecer *slit* de Emisión (0, o entre 2.5 y 20 nm, cada 0.1 nm)
Misceláneos	
$ES 1	Prender (0) o apagar (1) la fuente
$ST	Informar status
$AB	Abortar última operación
$BE 6161	Informa últimos 100 comandos recibidos. $BE sale de este modo
Movimiento automatizado de la placa de ELISA	
$WP 96	Posicionar en el pocillo 1-96
$PP X,Y	Posicionar en X, Y en mm. Envía a posición Park sin argumentos
Posicionamiento de monocromadores y solicitud de espectros	
$GM Ex,Em	Enviar el monocromador de Excitación a Ex (nm) y el de Emisión a Em (nm). $GM devuelve los valores actuales de los monocromadores
$SC 1	Obtener espectro (1=inmediatamente, 2=inicio remoto)

Tabla 2: Comandos destacados del fluorímetro para comunicación RS-232C

En efecto, si bien el manual indicaba que WP era el comando necesario para dirigirse a un pocillo determinado, la orden no se ejecutaba, y a cambio se obtenía un error indicando que el accesorio de lectura de placas no estaba conectado, cuando realmente lo estaba. Ante esto, se decidió realizar movimientos desde la interfaz del fabricante (con simples clic sobre un modelo virtual de placa) y posteriormente se consultó cuáles comandos habían sido recibidos para efectuar los movimientos. Con lo anterior se corroboró que ni siquiera la interfaz del fabricante utiliza en realidad el comando WP, sino que lo hace con PP, es decir, introduciendo valores de X e Y en mm, según cada pocillo. Con esto fue posible controlar sin inconvenientes la posición de la placa, y a su vez se pudo realizar un mapeo de la posición X e Y de cada pocillo, para luego utilizar este mapeo en la elaboración de la interfaz gráfica propia. Vale destacar que el comando PP existe porque el lector, además de poder leer placas de ELISA, puede realizar barridos sobre TLC (Cromatografía de

Capa Delgada), en cuyo caso es necesario enviar al instrumento entre cuáles valores (en mm) de X e Y deben realizarse dichos barridos. A su vez, el comando PP sin más argumentos da la orden de enviar la sonda hacia la posición “Park”, equivalente a la esquina superior izquierda de la placa, pero por fuera de ésta.

El comando GM permite enviar a los monocromadores a determinadas posiciones y con este se establecen las posiciones iniciales de ambos antes de la solicitud de un espectro. Cuando los monocromadores alcanzan sus destinos y habiendo enviado todas las configuraciones previamente, es posible solicitar un espectro inmediatamente con el comando SC 1 (SC 2 sirve para que el espectro sea adquirido cuando llegue una señal eléctrica remota al instrumento, lo cual no fue aplicado en este trabajo). El fluorímetro primero responderá con un código de error, que si es 0000, indicará que a continuación se enviarán más datos. Los primeros en llegar serán los denominados *status records* (que se pueden obtener con ST), los cuales indicarán en qué condiciones se darán las adquisiciones. Luego seguirán los espectros propiamente dichos, a través de valores separados por coma (un valor por longitud de onda en el espectro) y culminados en “-9999” según el fabricante.

Vale destacar que cuando un comando es enviado hacia el fluorímetro, su respuesta puede ser casi inmediata, o bien tardía debido a tareas aún pendientes (por ejemplo, que aún se esté moviendo un monocromador). Por lo tanto, esto se contempla a través del envío repetitivo de consultas cada cierto tiempo. Esta demora (*delay*) fue establecido empíricamente en 250 milisegundos. Si la respuesta obtenida era 0110 entonces se debía esperar y consultar nuevamente. Si la respuesta era 0000 significaba el fin de la espera y la continuación con las siguiente líneas de comandos. Si el código de error no era ninguno de los anteriores, entonces todo el proceso era abortado.

En cuanto al envío y recepción de comandos y respuestas a través de cadenas de caracteres codificantes, su implementación en Processing es muy sensilla una vez creada la comunicación en serie, que a modo de ejemplo estará asociada a una variable llamada Puerto, creada a partir de la biblioteca Serial incluida. Para enviar una cadena como “GM 300,400” basta con escribir Puerto.write(“$GM 300,400\n”). Para recibir una respuesta, dado que el instrumento siempre finaliza con '\n' (nueva línea), entonces se utiliza la función readStringUntil (“leer cadena hasta”) de la siguiente forma: respuesta= Puerto.readStringUntil('\n'). Luego “respuesta” deberá ser procesada para extraer datos.

Teniendo en cuenta todo lo anterior y aplicando conceptos clásicos de programación (condicionales, ciclos iterativos, concatenación de cadenas de caracteres para formar cadenas

mayores y con esto comandos, asociación entre elementos gráficos y variables de programación, entre otros y según las necesidades) se elaboró una interfaz gráfica como la expuesta en la figura 2.

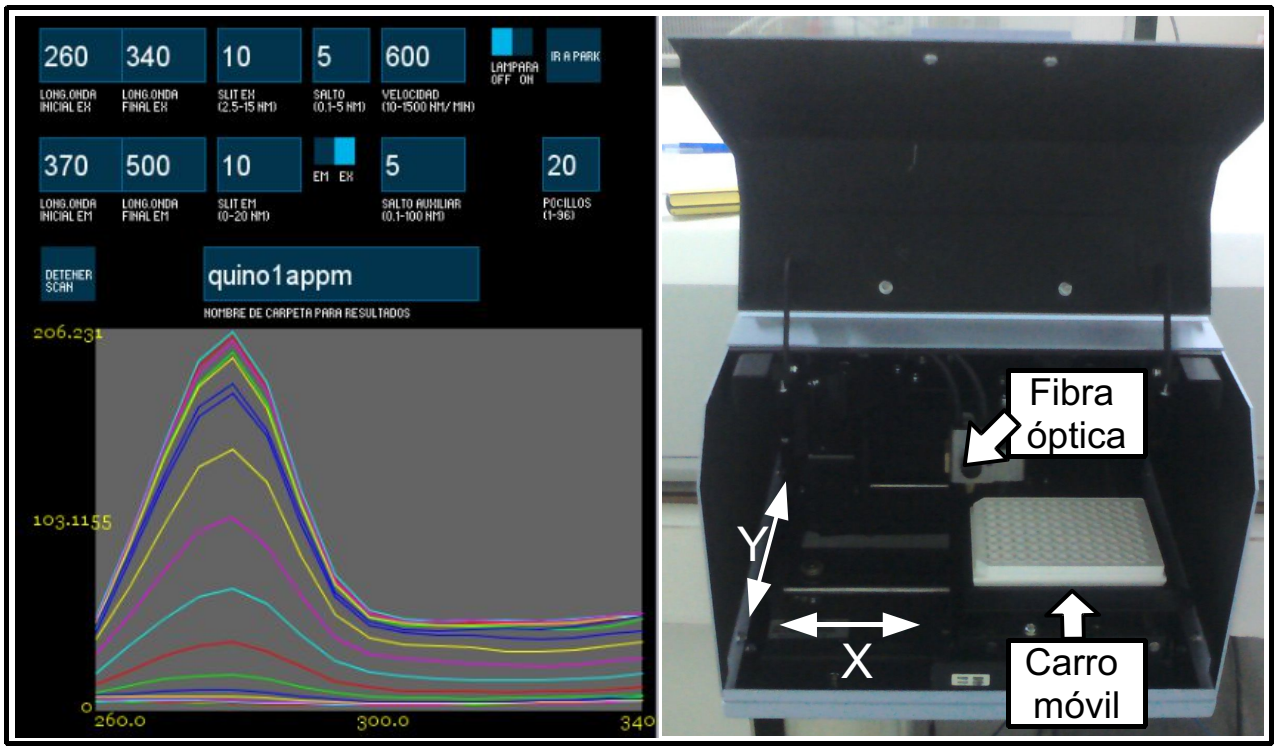


Figura 2: Interfaz gráfica para obtención de matrices de Excitación-Emisión durante una adquisición de datos (Izquierda) y Accesorio lector de placas de ELISA (Derecha)

La figura 2 muestra al accesorio controlado (se utiliza con la tapa superior cerrada) y a la interfaz durante una adquisición real, por lo que pueden observarse los espectros mientras van siendo obtenidos. En la interfaz gráfica deben introducirse valores para todos los parámetros necesarios (longitudes de onda, velocidad, *slits*). Cada valor introducido por el usuario es automáticamente analizado para verificar su aptitud, por ejemplo si una longitud de onda no excede a los máximos permitidos, y en caso contrario no se acepta el valor. También deberá introducirse un valor N válido (entre 1 y 96) que indique cuántos pocillos de la placa deben ser leídos (en el mismo orden en que fueron recolectados). Además deberá introducirse un nombre de carpeta, dentro de la cual serán guardadas las matrices (vectorizadas en realidad) obtenidas en archivos del tipo “n.txt”, con n desde 1 hasta N. En esta misma carpeta se creará automáticamente un archivo denominado “reporte.txt”, en el cual habrá información general sobre la adquisición de datos: intervalos de Excitación y Emisión junto a sus respectivos *slits*, velocidad de barrido, cantidad de pocillos y un

detalle de las longitudes de onda en las cuales hayan sido posicionados realmente los monocromadores.

Cuando todos los parámetros sean aptos y se dé la instrucción de comenzar con las lecturas, los parámetros serán enviados al fluorímetro con sus comandos respectivos, y en cada caso se esperará la respuesta 0000 tras el envío. Tomando como ejemplo la captura de espectros de Excitación con longitudes de onda de Emisión variables, la secuencia de comandos sería la siguiente (el signo + simboliza concatenación de caracteres):

- Seleccionar monocromador de Excitación: "$MX 0\n"
- Establecer longitud de onda mínima, MiniEx: "$AL "+MiniEx+'\n'
- Establecer longitud de onda máxima, MaxiEx: "$AH "+MaxiEx+'\n'
- Establecer slit de Excitación, slitEx: "$SX "+slitEx+'\n'
- Establecer slit de Emisión, slitEm: "$SM "+slitEm+'\n'
- Establecer incremento de Excitación, saltoEx: "$SI "+saltoEx+'\n'
- Establecer velocidad de barrido, velScan: "$SS "+velScan+'\n'
- Prender fuente: "$ES 0\n"
- Posicionar en posición Park: "$PP\n"

Hasta aquí se habrá obtenido la configuración mínima necesaria para solicitar espectros. La siguiente secuencia de comandos será repetida para cada pocillo:

1- Posicionar la sonda del fluorímetro.en X e Y mm, según el pocillo en cuestión

- "$PP "+X+','+Y+'\n'

- Previamente se obtuvieron éstos valores para cada uno de los 96 pocillos y se los puede consultar en cualquier momento durante una ejecución

2- Llevar a los monocromadores a sus posiciones iniciales correspondientes (longitudes de onda, lamEx y lamEm)

- "$GM "+lamEx+','+lamEm+'\n';

- Una vez realizados los movimientos, consultar posiciones reales (en nm) y adjuntarlas a registro.txt

3- Solicitar espectro (*start scan*)

- "$SC 1\n"

- A continuación el fluorímetro comenzará a enviar cadenas de caracteres ASCII que culminarán en '\n', por lo cual en Processing se da la orden de leer el puerto en serie y acumular caracteres hasta

que se encuentre '\n'. Algunas de las cadenas serán espectros (y son llamadas *data records*), pero no todas. Por lo tanto, éstas deben ser analizadas para tomar decisiones en base a su contenido. Sólo si la cadena contiene el valor -9999 (impuesto por el fabricante) es señal de que los caracteres anteriores a ese valor corresponden a espectros. Por ende, deben ser acumulados en variables para su posterior registro en un archivo. Los *data records* consisten de hasta 10 puntos sucesivos, separados por coma, de un espectro, y pueden transmitirse en ASCII, como en el presente trabajo, o en código binario.

4- Actualizar gráficos en pantalla a media que se obtienen los espectros.

5- Si aún existen espectros para completar la matriz del pocillo en cuestión, volver al paso 2, establecer una nueva posición inicial para los monocromadores (para el ejemplo, el de Excitación volverá siempre a su menor longitud de onda, mientras que el de Emisión tendrá que ir variando según el incremento impuesto) y repetir los pasos 3-5.

Una vez que se completa la matriz de Excitación-Emisión para el pocillo n, se guarda vectorizada en un archivo “n.txt”. La interfaz gráfica es reiniciada a la espera de nuevos espectros. Los pasos anteriores se repetirán hasta culminar con el último pocillo. Finalmente se guarda información general en “reporte.txt”.

Vale destacar que la comunicación en serie, el intercambio de información y la elaboración de una interfaz gráfica podría hacerse también con Matlab, pero se decidió utilizar Processing por ser libre, y para poner esa característica a prueba a nivel de las necesidades de un laboratorio de investigación quimiométrica. Una vez superadas las dificultades relacionadas a aprender la sintaxis del lenguaje (que tiene muchas similitudes con C++ o Matlab), Processing resulta sencillo y amigable. A su vez, dadas sus bases filosóficas, es realmente importante la comunidad en Internet y sus contribuciones, como ser la ayuda desinteresada ante consultas en foros, o el aporte gratuito de herramientas pre-diseñadas de uso sencillo y buena calidad como en este trabajo fue la biblioteca para elementos de interfaz gráfica controlP5.

3.6.2 Ensamble de hardware, programación de Arduino y elaboración de una interfaz gráfica para recolección de muestras

En base a restos fundamentalmente de impresoras como motores paso a paso, rieles, fuentes de

corriente, correas para convertir movimiento circular en lineal, LEDs, sensores de presencia de hoja (para detección de gotas, ver aparte), sumando algunos componentes comprados como botones tipo *switch* (de 2 posiciones) y fabricando piezas especiales como un puente móvil de aluminio, una base metálica para el dispositivo y un sostén de acrílico con un ojal para sujetar el capilar de recolección a la salida del HPLC, se diseñó y construyó el esqueleto electromecánico del recolector de muestras, cuyo esquema se expone en la figura 3.

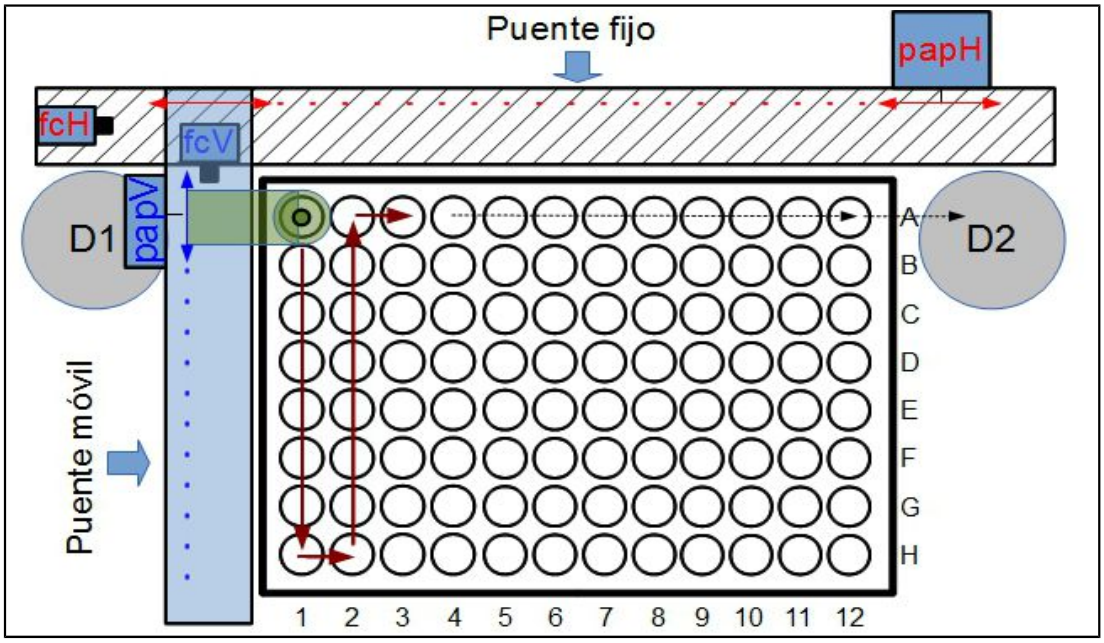


Figura 3: Esquema del recolector de muestras para una placa de ELISA

Referencias: H: Horizontal, V: Vertical, pap: Motor paso a paso, fc: *switch* de final de carrera, D1: depósito original de descarte, D2: depósito final de descarte. En rojo y azul lo relativo a los movimientos horizontales y verticales, respectivamente. La pieza verde posee un ojal para enhebrar el capilar de recolección.

En la figura 3 puede apreciarse un esquema del recolector de muestras visto desde arriba. La pieza graficada en verde posee un ojal en el cual el capilar de recolección es enhebrado. El movimiento de ésta pieza en ambos ejes deposita las gotas colectadas en los distintos pocillos. En el esquema también se señaló el recorrido de recolección que fue utilizado en experimentos descriptos posteriormente. Este recorrido, en flechas marrones, implica recolectar muestras en 17 pocillos (desde A1 hasta A3). También se señaló el recorrido final una vez que el pocillo en A3 se encuentra listo. Esta salida se programó de forma tal que al terminar con los pocillos solicitados, el capilar de

recolección será movido automáticamente hasta encontrar la columna 12, luego será movido hasta encontrar la fila A, y finalmente irá al depósito final de descarte 2 (D2), donde permanecerá inmóvil. De esta forma, las gotas que caigan al ir hacia D2 no contaminarán pocillos previamente colectados. De manera similar, el depósito original de descarte (D1) corresponde a la posición en la cual permanecerá el capilar antes de recibir la orden de comenzar con las recolecciones (es decir, ir al pocillo A1). Tanto D1 como D2 son simples vasos de precipitado o cualquier recipiente cuya forma permita residir en esas posiciones.

Se determinó que el movimiento del capilar fuera tal que en cada placa se fueran completando columnas ordenadamente y no filas. Esto fue así debido a características constructivas, ya que de recolectar filas el puente móvil vibraba mucho más de lo necesario y esto causaba desprendimientos de gotas y caída fuera de lugar.

En el denominado “Puente fijo” se encuentra anclado el papH. La rotación del último, correa de fuerza y eje guía de por medio, se traducirá en movimientos lineales (esquematizados con línea de puntos rojos) y esto hará que todo el “Puente móvil” se mueva, determinando la columna de la placa. De forma similar, en el “Puente móvil” se encuentra anclado el papV, el cual al rotar y con otra correa y otro eje guía de por medio permitirá mover un carro móvil dentro del puente móvil. Ya que el carro móvil es el que porta la pieza donde se enhebra el capilar de recolección, controlando sus movimientos se obtendrá posicionamiento lineal vertical (en línea de puntos azules), seleccionando de esta forma la fila de la placa.

Los finales de carrera (fcH y fcV) son botones tipo *switch* con dos contactos metálicos paralelos, los cuales quedan unidos cuando se presiona el botón, ya que un contacto metálico perpendicular los une. En uno de los contactos paralelos deberá haber una unión (un cable) hacia 0V (masa, tierra, *ground* o GND en la placa Arduino), usualmente en serie con una resistencia pequeña cuya función será estabilizar las lecturas posteriores de voltaje (existen corrientes espurias que interfieren con las lecturas digitales), y que por su posición (cerca de GND) y efecto suele denominarse *pull down*. El otro contacto paralelo deberá estar conectado a 5V. Tanto para fcH como para fcV existirá un pin digital de lectura reservado en Arduino. Cada pin deberá estar cableado hasta llegar al contacto de GND en el botón. Cuando el botón no esté presionado, la lectura de voltaje en el pin será de 0V (0 o valor lógico Falso), y cuando lo esté será de 5V (1 o valor lógico Verdadero). A través del uso de condicionales de programación, es posible evaluar cuál de las dos posibilidades ocurre en un momento dado y de esto surge la función de los finales de carrera, es decir, representar una posición específica a través del cambio de voltaje. Por un lado, esto es útil

cuando se realizan movimientos, ya que antes de cada movimiento puede verificarse si el final de carrera se encuentra presionado, en cuyo caso un movimiento hacia el botón no debe ser permitido para no dañar a los motores. En este dispositivo la función fundamental es similar a la que ejecutan algunas impresoras con motores pap al ser encendidas. Los pap no suelen poseer un mecanismo por el cual uno pueda consultar la posición absoluta. Como no es posible determinar en qué posición se encuentran los pap cuando se inicia el suministro eléctrico (por ejemplo pudo haberse cortado la luz durante una ejecución de movimiento), es necesario encontrar un punto de referencia, usualmente denominado paso cero. Por esta razón, tanto en algunas impresoras con el carro que porta los cartuchos como en este dispositivo respecto del ojal para el capilar, al inicio se ejecuta una secuencia que permite encontrar a los finales de carrera y de esta manera se obtiene la posición a través de la cual serán relativizadas todas las restantes posteriores. Estos movimientos se realizan a baja velocidad y luego de cada paso se verifica reiteradamente si los finales de carrera cambian de voltaje, en cuyo caso se detiene el proceso asumiendo que se ha llegado a la posición deseada y que por ese motivo el final de carrera ha sido presionado. En el caso del esquema, esto equivaldría a mover lentamente (de a un paso en papH) el puente móvil hacia la izquierda hasta que éste presione a fcH y se lean 5V, y luego a mover de la misma forma el ojal hasta que en fcV se lean 5V. Habiendo determinado los respectivos “paso 0”, a partir de aquí cualquier orden de movimiento actualizará la cantidad de pasos en una variable correspondiente, y de esta manera se sabrá cuántos pasos serán necesarios para realizar cualquier movimiento deseado. Vale destacar que una consulta a este tipo de variables entregará un valor de pasos actuales, pero no necesariamente la posición real de los objetos móviles. Cuando se automatizan tareas de éste tipo, es probable que en la práctica alguna vez se produzcan inconvenientes, por ejemplo que un objeto trabe el movimiento de los motores y los detenga mecánicamente, por lo cual debería detenerse la tracción. En estos casos, a nivel de variables de programación se podrá deducir falsamente la posición real, porque no existe una retroalimentación que indique si los movimientos llegaron realmente a su destino. En este trabajo no se utilizaron sistemas para corroborar las posiciones reales, ya que a la velocidad en que se desarrollaban los movimientos y con el debido cuidado experimental las recolecciones automáticas resultaban apropiadas. Sin embargo, vale destacar el uso de *encoders*, normalmente utilizados en impresoras y en cualquier *mouse*, y por lo tanto reciclables. Los *encoders* usualmente tienen principios ópticos de funcionamiento y suelen estar construidos de forma tal de intercalar ranuras en un material opaco, o bien barras opacas en un material traslúcido. Sin importar si el *encoder* es lineal (puede verse como una tira traslúcida con barras negras en las impresoras) o

circular (usualmente una rueda ranurada para el *mouse*), de un lado de éste existe una fuente de luz, y del otro lado hay un componente sensible a la luz y capaz de emitir dos valores diferentes de voltaje, según tenga acceso o no a la luz. Los movimientos que quieren controlarse producirán el movimiento de partes de los *encoders*, lo cual generará intermitencia en los voltajes. Habiendo calibrado la cantidad de pasos necesarios para generar intermitencia, si éstos valores son leídos con frecuencia luego de ordenar un movimiento, no sólo será posible evaluar realmente el avance de los pap sino también sus posiciones definitivas y, en caso de emergencia, se podrá detener la tracción de los pap a tiempo antes de producir daños.

Mientras que para los finales de carrera sólo son necesarios tres cables para cada uno (GND, 5V y uno para leer la señal digital conectado a un pin de Arduino), la conexión con los pap es más compleja y requiere explicaciones previas. Aunque a nivel constructivo realmente no sea así, los pap pueden imaginarse construidos con un eje rotativo al cual existen adheridos imanes permanentes, rodeado de 4 bobinas. Cuando por una (o 2 en algunos casos) se hace circular una corriente eléctrica apropiada en intensidad y en polaridad, se producirá un campo magnético en la bobina con un determinado sentido, lo cual perturbará la posición del eje para que los campos magnéticos sean alineados y de esta forma se constituirá lo que se denomina un paso. De aquí debe entenderse que si la corriente permanece activa en la misma bobina, no se producirá ningún paso más, y sólo será posible un nuevo movimiento una vez que otra bobina sea polarizada. Por esta característica los pap son considerados precisos y ésta precisión viene determinada por el ángulo mínimo que determinará un paso (depende del fabricante). Desde luego, el movimiento en cualquier sentido será determinado por la ordenada polarización de las bobinas, y la velocidad dependerá de la frecuencia de llegada de éstos impulsos eléctricos.

La figura 4 esquematiza modos de realizar/interpretar movimientos con los pap y a partir del esquema es posible concebir dos tipos de movimientos. En el primero de ellos la activación de bobinas se da de a una, mientras que en el segundo lo hacen dos en simultáneo. El resultado de ambos será movimiento basado en cuatro pasos, sólo que en el primer caso el movimiento consumirá menos energía eléctrica pero ejercerá menor torque. La elección entre un modo y el otro dependerá de las necesidades y límites prácticos. Es sencillo deducir que existe un tercer modo para realizar pasos, el cual ha sido utilizado en la construcción del recolector de muestras. Este modo se basa en la combinación de los anteriores, por lo cual existirán ocho pasos distintos (A, A-B, B, B-C, C, C-D, D y D-A) y con esto aumentará la resolución del posicionamiento, es decir, los ángulos

entre paso y paso valdrán la mitad que en los modos anteriores y por ende los movimientos lineales asociados serán menores y más precisos.

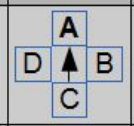
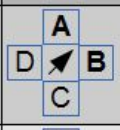
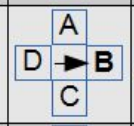
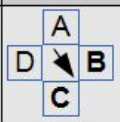
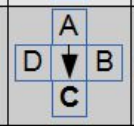
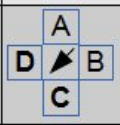
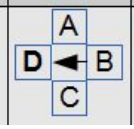
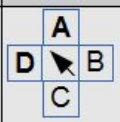
Paso	A	B	C	D		A	B	C	D	
1	1	0	0	0		1	1	0	0	
2	0	1	0	0		0	1	1	0	
3	0	0	1	0		0	0	1	1	
4	0	0	0	1		1	0	0	1	

Figura 4: Esquema de movimiento en motores paso a paso

Referencias: Las bobinas (A-D) energizadas en cada paso (1-4) son las que se encuentran en negrita o con un valor de 1 en su respectiva columna. Las flechas señalan la alineación de campo magnético que moverá al eje.

Sin ahondar demasiado en detalles, existen dos tipos de motores pap, unipolares (como papV) y bipolares (como papH), ambos sumamente utilizados en la construcción de impresoras y *scanners* (actualmente son reemplazados por motores de corriente continua, más económicos, pero con los cuales no se podía obtener la precisión necesaria, hasta el actual avance de la electrónica), y también en instrumental de laboratorio como bombas de jeringa. La intensidad de corriente que puede obtenerse a partir de un pin digital de Arduino es de tan solo 40 mA, insuficiente para energizar a las bobinas de los pap y generar movimiento. Por lo tanto, se requiere de otra fuente más potente de energía eléctrica, que en este caso fue la fuente de la misma impresora desde la que se obtuvieron los pap. No obstante, esta fuente no puede tomar contacto directo con los pines de Arduino ya que éstos toleran voltajes de hasta 5V (contra 24V de la fuente), por lo cual las conexiones se tornan más complejas. La figura 5, obtenida con Fritzing, ejemplifica la conexión a un pap unipolar.

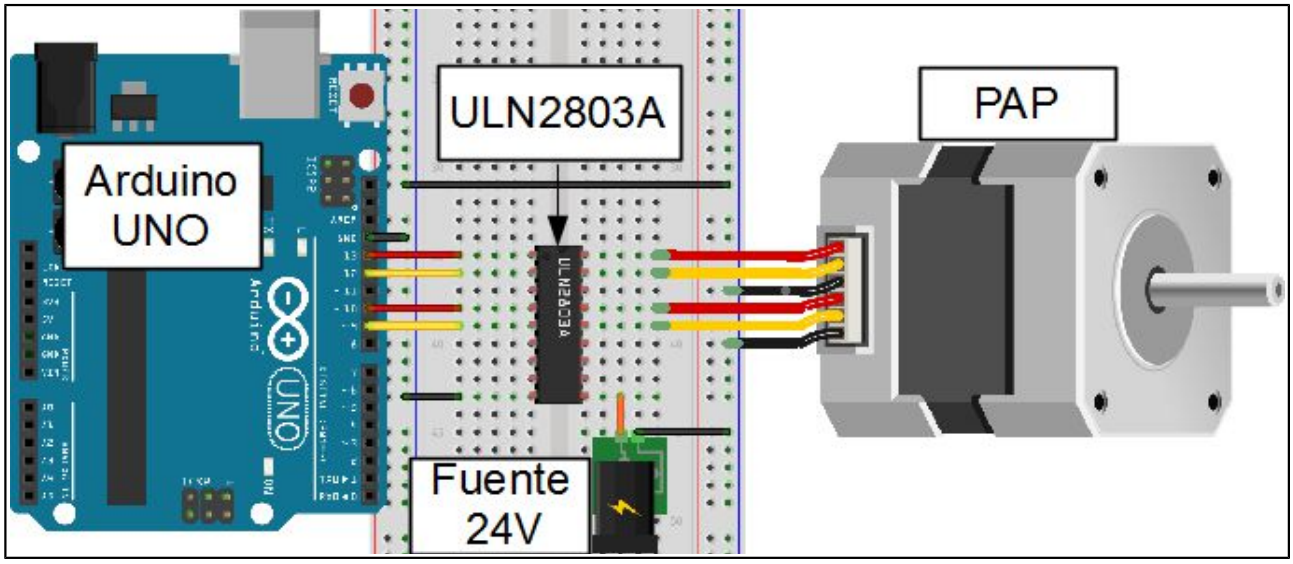


Figura 5: Esquema de conexiones entre una placa Arduino y un pap unipolar a través del arreglo de transistores Darlington ULN2803A

En la figura 5 las masas de Arduino, de la fuente de 24V, del pap y del integrado ULN2803A se encuentran representadas por cables negros y a su vez unidas entre todas, independientemente de si sus respectivos positivos eran de voltajes diferentes. Esto se realizó con todas las masas del recolector de muestras. En relación, en el esquema el pap tiene dos cables de masa, contabilizando un total de seis cables, pero suelen encontrarse de cinco, es decir, las dos masas se encuentran unidas en un único contacto. También en el esquema puede apreciarse la función del integrado ULN2803A, cuya hoja de datos con todo tipo de detalles puede obtenerse desde Internet a través de cualquiera de sus fabricantes (no existe uno en exclusivo), por ejemplo en [http://www.ti.com/lit/ds/symlink/uln2803a.pdf]. El integrado es un arreglo de 8 transistores Darlington, cada uno de los cuales puede transmitir hasta 500 mA desde un voltaje no superior a 30V, lo cual resultó suficiente para los pap en cuestión. Cada uno de estos 8 elementos funcionará como compuerta, estando compuestos por un pin de entrada para 5V (son los situados a la izquierda, conectados directamente a pines digitales de salida en la placa Arduino) y por un respectivo pin de salida (a la derecha, conectados al pap), que transmitirá la corriente provista por la fuente. La conexión entre la última y el integrado se da en los pines inferiores del esquema, siendo el de la derecha para 24V (cable naranja) y el de la izquierda para masa (0V).

Cada vez que a una entrada de 5V llegan 5V (escritura digital ordenada con Arduino) estará permitido en la salida respectiva el pasaje de hasta 500 mA a 24V provenientes de la fuente.

Mientras esto se prolongue, la bobina conectada a la salida permanecerá energizada y su campo magnético estará activo. Controlando la frecuencia y el orden con los cuales las bobinas son energizadas/desenergizadas se hace posible realizar movimientos.

Ya que el movimiento de los pap se da cuando este recibe energía, si por alguna razón el pap es sometido a movimiento entonces generará energía. Esta es otra de las funciones del integrado, es decir, aislar la conexión entre lógica de 5V y corriente de 24V. Estos integrados poseen protecciones internas (resistencias de 2.7 kΩ para cada par del arreglo) de forma tal que si se da un sobrevoltaje y en efecto la energía intenta llegar hasta la placa Arduino, antes se produzca la rotura de la protección. No obstante, para una mayor protección, existen los denominados controladores de motores, o *drivers*. En estos cada pin del integrado es protegido con resistencias y diodos que impiden el pasaje de corrientes en sentido inverso. A su vez, ya que puede existir sobrecalentamiento durante las operaciones, los *drivers* suelen proteger al integrado con un disipador metálico de calor. Para operar al papH bipolar, el integrado necesario L293D [http://www.ti.com/lit/ds/symlink/l293d.pdf] fue utilizado como parte de un *driver* de motores (el ULN2803A no cumple con los requisitos). Los motores bipolares, los cuales tienen normalmente sólo cuatro cables ya que no se les suministra contacto directo con masa, requieren no sólo una determinada intensidad de corriente para sus bobinas, sino también una polaridad apropiada para cada paso, y es el integrado L293D el que cumple esta función a través de circuitos denominados "puentes H". El integrado es capaz de proveer hasta 600 mA a voltajes entre 4.5V y 36V en cada una de sus cuatro salidas, cada una de las cuales estará cableada con una bobina del pap. Por el lado de las entradas, éstas se encuentran protegidas y por lo tanto la conexión entre los respectivos cuatro pines digitales de salida en Arduino se realiza de forma directa a las cuatro entradas del *driver*. A su vez el último posee una entrada para la fuente de 24V, una para masa, y otra para 5V. Finalmente, existe en el *driver* la posibilidad de cortar/permitir el suministro eléctrico a las bobinas a través de un pin digital (se conoce como pin *enable*). Esto fue utilizado en el recolector de muestras para evitar sobrecalentamientos, por lo cual cuando papH no debía realizar movimiento el suministro era cortado enviando 0V al mencionado pin. Vale aquí destacar que el ULN2803A no posee esta característica, por lo que para cortar el suministro a papV se escribió una función en Arduino que indicaba el envío simultáneo de 0V a todos los pines de entrada en el integrado.

En el recolector también se utilizaron LEDs reciclados, uno verde y otro rojo, para indicar estados de reposo, espera para comenzar una recolección, entre otros. La conexión de LEDs es sumamente sencilla, ya que uno de sus contactos (por convención es el más corto) debe ser

conectado a masa y el otro a un pin digital de salida en Arduino, el cual al proveerlo de 5V lo encenderá. En cualquiera de las dos conexiones es recomendable poner en serie una resistencia pequeña (por ejemplo de 220 Ω) para proteger al LED.

A nivel de descripción de componentes, finalmente vale comentar que se intentó que el sistema pudiera contar las gotas a medida que iban siendo desprendidas del capilar de recolección. Para esto se utilizaron sensores de presencia de hoja rescatados de impresoras, cuyos principios de funcionamiento son similares a los expuestos al hablar de *encoders*. Estos sensores tienen forma de "U", con una fuente de luz en uno de los brazos (suelen ser UV) y con un detector de luz en el otro. Cuando entre ambos se interpone una hoja (una gota en nuestro caso) se interrumpe el haz de luz y esto da un determinado voltaje de salida, distinto al que puede ser leído cuando no existe interrupción. Por lo tanto, controlando la variación de esta salida es posible contabilizar las interrupciones, y con esto la cantidad de gotas. En pruebas preliminares donde los sensores fueron probados con agua inyectada manualmente con jeringas, el conteo fue correcto. Sin embargo, una vez que todo el recolector fue montado, se comprobó que las gotas que salían del HPLC no eran apropiadas para pasar por el detector, sumado al hecho de que al existir movimiento y estar cercanos el final del capilar con los pocillos de la placa se producían problemas de tensión superficial que terminaban estancando gotas en el detector. Esta fue la razón por la cual se decidió extraer este sensor, más allá de que el circuito eléctrico para leerlo sí fue realizado y en la lógica de programación haya podido ser incluido. Es posible cuestionar el uso de un contador de gotas dado que requiere una interacción con la muestra y esto podría interferir en la posterior lectura de la muestra en el fluorímetro. No obstante, en este trabajo las muestras eran irradiadas con luz UV dentro del HPLC, por lo que el efecto de un contador con luz UV no debería ser perjudicial.

Desde luego que antes de realizar un montaje en posiciones definitivas, los componentes son puestos a prueba de forma separada. Para evitar tener que realizar soldaduras en estas etapas, se utiliza una placa de pruebas o *protoboard*, presente en el esquema de la figura 5, donde los cables, integrados y otros componentes se insertan directamente y las uniones entre éstos pueden realizarse de forma ordenada.

Una vez que la etapa de pruebas culminó en el *protoboard*, hay que decidir si se hará una plaqueta para un circuito que incluya a un ATmega328 programado, o bien si se reutilizará la misma placa Arduino con la que se realizaron las pruebas. En nuestro caso, se decidió hacer lo último, fundamentalmente porque la placa ya posee un sistema para ser reconocida como dispositivo USB

en cualquier PC, con lo cual la operación del recolector puede realizarse sin depender de una PC específica. Esto último es algo que suele suceder en los laboratorios y en ocasiones se torna difícil o hasta imposible operar instrumental sólo por la obsolescencia de las computadoras que los controlan.

La estrategia utilizada fue similar a lo que en la jerga de Arduino se conoce como *shield*. Las placas Arduino tienen también como objetivo resultar accesibles desde lo económico, razón por la cual en su construcción existen limitaciones que, de superarse, la harían más útil, pero que no se llevan a cabo para mantener bajos los costos. Para aquellos usuarios que deseen aumentar las capacidades de su placa invirtiendo dinero, existen placas con funciones relacionadas a ciertas tareas, fabricadas de forma tal que se ajustan adecuadamente al diseño de pines de una placa Arduino. Estos son los *shields* y se puede realizar una analogía entre Arduino-*shield* y Placa Madre de una PC - Cualquiera de las tarjetas insertas. Los *shields* suelen ser provistos de una biblioteca para su funcionamiento. Ejemplos de éstos son los *shields* que permiten conexión de Arduino a internet (Ethernet y Wifi), que dan posicionamiento (GPS), que permiten manejar varios y diversos motores al mismo tiempo (similarmente al *driver* de motores utilizado), entre muchísimos otros. Una extensa lista puede obtenerse desde [http://shieldlist.org/]. En nuestro caso, el *shield* fabricado se ajusta a una placa UNO y contiene elementos de protección (diodos y resistencias), al integrado ULN2803A y a todos los componentes necesarios para transmitir corrientes y señales desde y hacia finales de carrera, motores pap y LEDs (también para el cuentagotas que finalmente fue descartado). Para las transmisiones entre el *shield* y los componentes se reciclaron cables simples (obtenidos desde PC, impresoras, etc.), o bien un cable plano de múltiples canales tipo manguera (se rescató de una impresora, así como sus conectores hembra, ambos reciclados en el recolector). Vale destacar que el hecho de adoptar la modalidad de *shield* tiene otra ventaja, y es que la placa UNO puede ser utilizada con otros propósitos simplemente extrayendo el *shield* y reprogramándola. Debe entenderse que la placa en sí constituye una herramienta de laboratorio para otros prototipos.

Para modelar el circuito impreso en Cobre al cual se sueldan los conectores del *shield* se utilizó Fritzing, el cual dispone de modelos tamaño real de placas Arduino, por lo cual se simplifica notablemente la tarea de escalar el diseño a un tamaño apropiado para su encastre. En Fritzing uno realiza las conexiones entre componentes con cables y un *protoboard* virtual (como se ha visto esquematizado). Una vez que todo está listo, Fritzing realiza automáticamente las rutas de Cobre (*rounting*) para suplantar a los cables. De este diseño se obtiene una impresión, la cual se transfiere a una placa virgen que luego será procesada para obtener el circuito finalizado. La figura 6 expone

fotos de varios elementos nombrados hasta aquí.

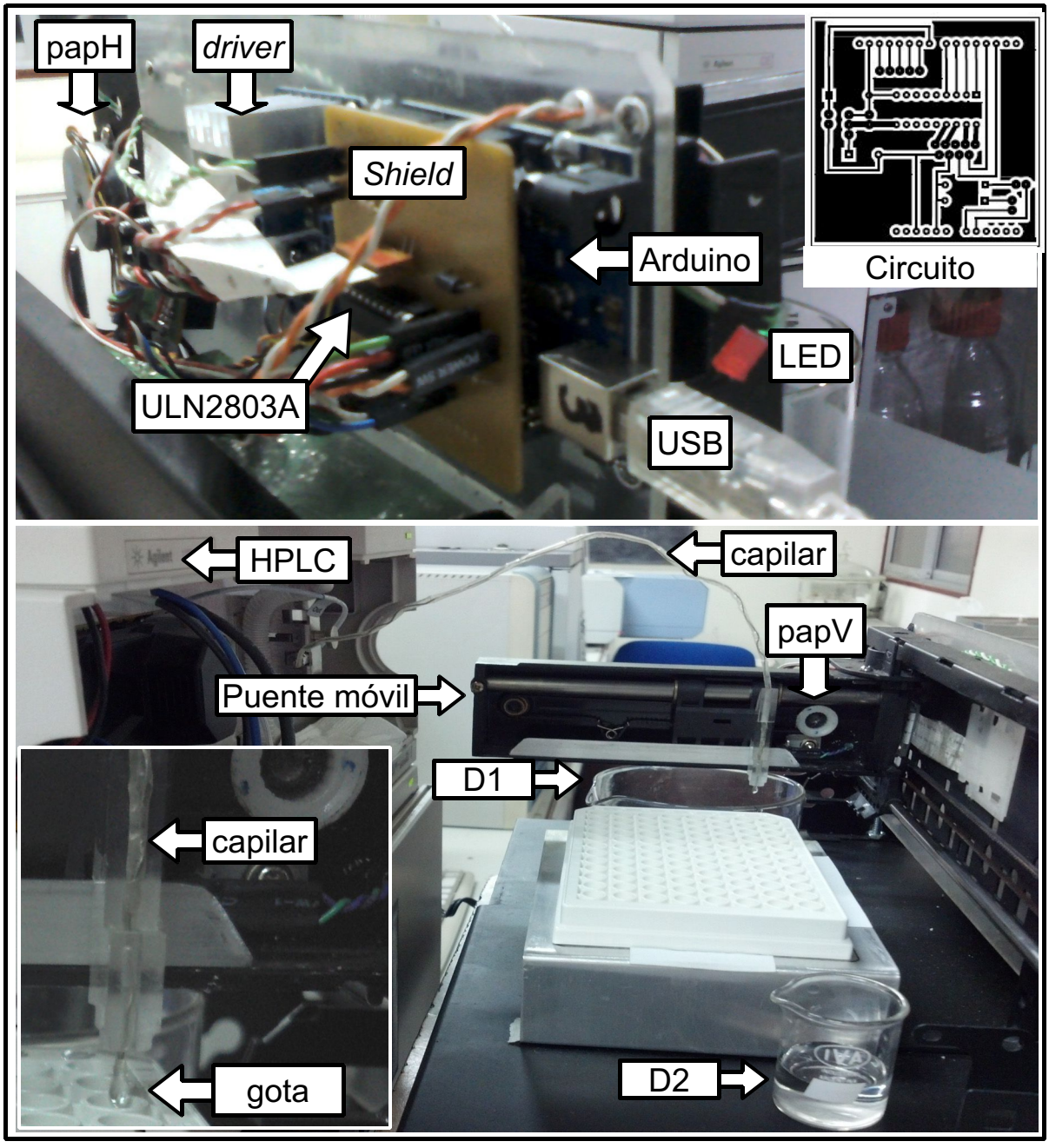


Figura 6: Recolector de muestras: Componentes y circuito

Referencias: D1 y D2: Depósitos Inicial y Final, respectivamente.

De la figura 6, vale decir que del *driver* sólo se puede observar el disipador de calor (el resto se encuentra debajo) y que en el zoom inferior izquierdo no se aprecian mayores detalles, pero aunque no se señaló, debajo de la palabra “capilar” también está fcV.

En cuanto a la elaboración del *shield*, una vez impreso el esquema en papel fotográfico, se elaboró la placa como se describió en Materiales y Métodos, y se soldaron los componentes con puntos de Estaño.

Una vez que todo fue finalmente ensamblado, se procedió a programar a la placa UNO y a elaborar una interfaz gráfica en Processing para comunicarse con ésta, con el objeto de poder controlar el movimiento del capilar de recolección.

En cuanto a la interfaz, su elaboración fue similar a la del fluorímetro a nivel de componentes de interfaz gráfica, tal como puede observarse en la figura 7.

Figura 7: Interfaz gráfica para controlar el recolector

En la primera fila de componentes debe insertarse el pocillo inicial (1), el final (17) y el tiempo en segundos (2) en el que el capilar de recolección deberá ser mantenido sobre un pocillo. Si se desea dar inicio a la recolección en ese momento, se debe pulsar el botón INICIAR. Si se introduce un tiempo en segundos (47) como *delay* de inicio, entonces al pulsar al último un LED comenzará a titilar una vez por segundo hasta que finalmente comenzará la recolección (el capilar pasará desde el depósito de descarte inicial hacia el pocillo indicado como inicial). En este trabajo se utilizaron 47 segundos y se pulsaba Iniciar en el momento en que el HPLC realizaba la inyección. Dado el flujo de trabajo, con ese tiempo era suficiente para obtener al primer analito eluido (OFL) en el segundo pocillo. El botón STOP de la fila superior da la orden de cancelar una recolección en curso, haciendo que inmediatamente el capilar sea dirigido hacia el depósito final de descarte (misma posición que alcanzará automáticamente si la recolección no es detenida y logra completarse). Esta primera fila de componentes está dedicada pura y exclusivamente a una recolección sobre placas de ELISA de forma automática. Sin embargo, el recolector puede recorrer cualquier punto que no

supere la superficie de una de estas placas, por lo cual debajo pueden utilizarse otro tipo de contenedores en lugar de pocillos (tubos Eppendorf, de ensayo, etc.) y/o recorridos distintos. Para todo esto están los controles de las filas siguientes.

La segunda fila de controles comienza con un selector a través del cual es posible pasar desde el modo automático de recolección hacia un control manual. El control a la derecha permite introducir un número entero de pasos (1) y una velocidad en rpm (revoluciones por minuto, 60). De esta forma, con las flechas del teclado de la PC en la que se ejecuta la interfaz es posible dirigir el movimiento y cada pulsación realizará la cantidad configurada de pasos a la velocidad establecida (se controla previamente que la cantidad de pasos actuales y la propuesta no resulten en posiciones prohibidas). El botón denominado ORIGEN dirige al capilar desde donde se encuentre hacia el depósito inicial.

El modo manual es útil para realizar pruebas de respuesta y reacción del dispositivo, como por ejemplo a qué velocidad máxima no se obtienen caídas de gotas fuera de lugar, pero también para obtener posiciones específicas informadas en pasos. Cada vez que se realice un movimiento manualmente, los dos componentes más a la izquierda de la tercera fila indicarán la cantidad de pasos en el eje horizontal y en el vertical. Tomando nota de estas coordenadas es posible realizar un recorrido propio, siempre que se respete el área realmente cubierta por el recolector (tanto la interfaz como el código de Arduino contienen la información de la cantidad real de pasos máximos permitidos desde fcH y fcV, determinados empíricamente en 850 y 166, respectivamente). Cualquier recorrido es revisado en todas sus coordenadas y no se acepta en caso de encontrar al menos una no válida, lo cual es informado al usuario. Para que sea válido, el archivo debe contener líneas de pasos en X e Y, separados por coma. Se asumirá que las últimas dos coordenadas corresponden a la posición de descarte final donde deberá reposar al culminar. Si se provee a la interfaz de un archivo txt de estas características, puede utilizarse el botón RECORRIDOTXT para desplegar una ventana de exploración y seleccionarlo. Esto ejecutará los controles necesarios y alertará lo que sea requerido. Si las coordenadas son aptas, al pulsar INICIAR serán recorridas esperando el *delay* (47s) y el tiempo entre pocillos (2s).

La parte inferior de la figura 7 deja ver una zona de mensajes. Cuando exista algún error, como una coordenada incorrecta, será informado en esa zona.

En cuanto al *sketch* en la placa UNO, inicialmente se reservan nombres de variables y se las inicializa según corresponda a las conexiones reales. Sin entrar en grandes detalles, se definen

relaciones entre pines de la placa UNO y componentes de salida como los pap (cada uno asociado a 4 pines, 1 por bobina), encendido/apagado del *driver y* LEDs, y de entrada como fcH y fcV (originalmente también para el pin del cuentagotas). Cada entrada debe también tener una variable asociada, donde se irá actualizando el valor leído en cada una. También se establecen valores constantes, como pasos máximos permitidos, pasos/vuelta para los pap, pasos necesarios en cada pap para ir desde un pocillo hacia uno vecino, entre otros.

En la función *setup* se establecerá qué pines son de salida y de entrada, y se configurará el inicio de la comunicación en serie, se ejecutará una rutina de comprobación de la comunicación (*handshacking*) y se establecerán velocidades iniciales para los pap. Luego se utilizará una función que enviará al recolector hacia su punto (0,0) a baja velocidad, a través del cambio de señal en fcH y en fcV, como ya se explicó.

A partir de aquí comienza la función *loop*. Se realiza una consulta permanente en el puerto en serie. Si no se recibe nada, no se hace nada. Se recibirá cuando la interfaz envíe datos y en dicho caso los caracteres enviados como ASCII serán condicionalmente evaluados.

La comunicación en serie entre la interfaz gráfica y la placa UNO envuelve un intercambio de mensajes entre ambas, de forma similar a lo explicado con el fluorímetro, pero con la diferencia de que al haber construido el recolector, no existe un código para los mensajes al cual haya que adaptarse, sino que éste fue creado a conveniencia de las operaciones necesarias. El código utilizado se basó en el intercambio de una cadena ASCII elaborada por la interfaz según cada operación.

Se toma el primer carácter de la cadena, el cual indica el modo. El valor 1 indica manual, 2 automático, 3 cambio de velocidad y 4 ir hacia (0,0) lentamente.

Si el modo es manual, se lee el carácter 2 para deducir qué pap debe moverse y el carácter 3 para obtener la cantidad de pasos a dar. Con esto es suficiente para dar los pasos, por lo cual éstos son ejecutados, la cuenta de pasos es correspondientemente actualizada (en memoria de Arduino y en la interfaz) y las bobinas son finalmente apagadas.

Si el modo es automático, se leen los caracteres 2, 3 y 4 para obtener la cantidad total de pocillos (o la cantidad de posiciones para detenerse si el recorrido provino de un txt), el tiempo entre pocillos, y el *delay* de inicio (cero en su defecto). Aquí se evalúa si la cadena culmina en '\n' o si prosigue. Si culmina, es la indicación de que debe comenzar la recolección luego del debido *delay* (en dicho caso, el LED verde titila una vez por segundo, hasta quedar finalmente encendido cuando termina la espera), lo cual será realizado con las esperas programadas para cada pocillo y culminando en el depósito final de descarte. Si la cadena de caracteres no culmina, significa que a

continuación serán enviadas las coordenadas. Éstas son leídas e interpretadas (es algo básicamente instantáneo) y luego comienza el *delay* y/o la recolección.

Si el modo es cambio de velocidad, el carácter 2 portará su nuevo valor. La misma velocidad será aplicada a ambos pap.

3.6.3 Determinación de tiempo de recolección por pocillo y de *delay* inicial

Para determinar cuánto tiempo debía residir el capilar de recolección sobre un pocillo de ELISA, así como también para establecer el número óptimo de pocillos a recolectar (lo cual, para evitar pérdida de información, incluyó pocillos antes y después de los tiempos analizados), se realizaron experiencias en las cuales el cromatograma (a 280 nm) de una mezcla de los tres fármacos fue utilizado para estimar el tiempo de llegada a las placas de ELISA, el cual fue establecido en 47 segundos después del momento de la inyección en el HPLC. En estas experiencias también se determinaron las condiciones óptimas para las corridas cromatográficas (fase móvil, velocidad de flujo, etc.) y las concentraciones mínimas de cada fármaco aptas para obtener una señal aceptable posteriormente en el fluorímetro. Vale aclarar que aunque la velocidad de flujo sea constante y el tiempo en cada pocillo definido, el producto de ambos no necesariamente daría el volumen depositado, puesto que es posible que algunas gotas en movimiento caigan fuera del pocillo donde técnicamente deberían hacerlo. Esto último, que puede resultar cuestionable, será evaluado posteriormente y se verá qué tan determinantes pueden ser estas caídas.

Desde luego que la determinación del tiempo de residencia en cada pocillo tuvo como objetivo recolectar tantos pocillos como fuera posible, para preservar los beneficios de la cromatografía y para mantener a los analitos separados. El aumento en el número de pocillos puede lograrse disminuyendo el tiempo de residencia en cada uno, pero a una velocidad de flujo constante, existe un límite mínimo de tiempo. A su vez, tiempos menores representan (no exactamente) volúmenes menores y algunos no son aptos para obtener luego lecturas con la fibra óptica del fluorímetro dada su sensibilidad. Por todo lo anterior, el tiempo entre pocillos fue establecido en 2 segundos (aproximadamente 60 μL/pocillo), para un total de 17 pocillos, recolectados con un *delay* de 47 segundos desde la inyección. La figura 8 expone todos los espectros de Excitación obtenidos para la muestra VAL2 en cada pocillo, donde resulta perceptible la separación de los analitos.

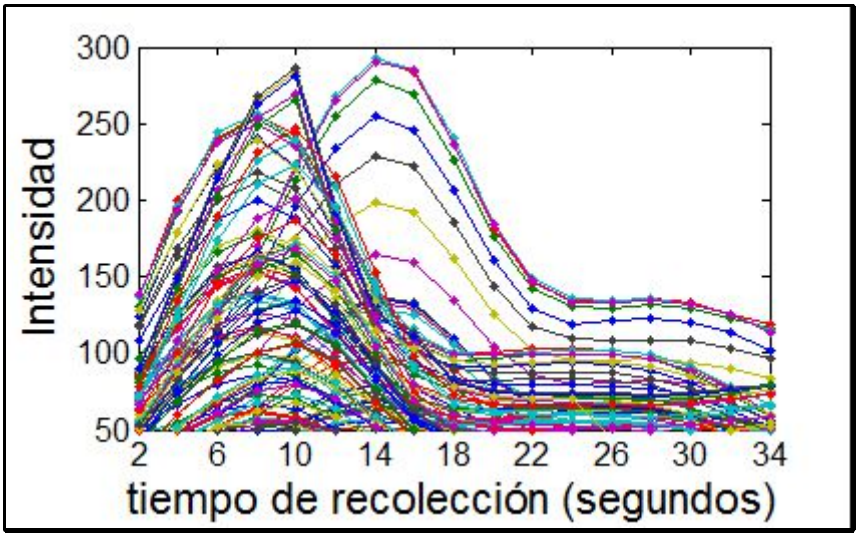


Figura 8: Espectros de Excitación para la muestra VAL2 en recolecciones cada 2 segundos

3.6.4 Obtención de datos para cuantificaciones

Las 15 muestras de Calibración y las 12 de Validación fueron cromatografiadas en el HPLC, de lo cual se obtuvieron matrices de espectros UV. Cada muestra fue recolectada en 17 pocillos de ELISA, tal como fue explicado.

Las matrices UV originales contenían 293 tiempos diferentes. De cada una se tomó como tiempo cero al equivalente a 47 segundos (*delay* de recolección) y se conservaron luego 64 tiempos totales, suficientes para abarcar los perfiles cromatográficos completos de los tres analitos. Luego las 27 matrices fueron apiladas.

Cada recolección originó dos tipos de datos de fluorescencia. Un conjunto corresponde a los datos crudos obtenidos, denominado Fcr (Fluorescencia crudo), mientras que de cada matriz Fcr se obtuvo otra aplicando *spline* (interpolación cúbica por partes) y suavizado con polinomios de Savistky-Golay (Savitzky y Golay, 1964), dando origen a Fss (Fluoresencia *spline/*Savitsky-Golay).

Cada matriz Fcr tuvo dimensiones de 17×25 (Excitación-Emisión) por pocillo. Cada una fue vectorizada concatenando espectros de Excitación, obteniendo vectores de 425 elementos, los cuales corresponden tanto a un pocillo como a un tiempo de recolección. Por lo tanto, los vectores de los 17 pocillos de cada muestra fueron dispuestos para conformar 27 matrices de 17×425, las cuales fueron apiladas para su posterior resolución mediante MCR-ALS.

Para obtener las matrices Fss, cada matriz Fcr de 17×25 se procesó con *spline* y la interpolación

en ambas dimensiones resultó en matrices de 33×49, un punto intermedio por cada punto original. Luego se aplicaron suavizados mediante polinomios de Savistky-Golay de grado tres en ambos órdenes. Nuevamente, cada matriz fue vectorizada concatenando espectros de Excitación, y estos vectores de 1617 elementos fueron apilados por placa, obteniendo matrices con dimensiones de 17×1617. Finalmente se apilaron estas 27 matrices para aplicar MCR-ALS. La figura 9 esquematiza los procedimientos realizados.

Vale mencionar que al analizar visualmente los datos de fluorescencia obtenidos, se percibió que algunas muestras de Calibración, específicamente dos de OFL (8 y 10 ppm) y una DNF (2.5 ppm), poseían espectros de menor intensidad a la que deberían tener en relación a otras muestras de Calibración para los mismos analitos. Esto pudo ser debido a que cuando las muestras eran procesadas en HPLC demoraban menos que cuando eran procesadas en el fluorímetro, ya que la adquisición de datos en el último es mucho más lenta. La razón de esta diferencia es que el HPLC cuenta con un Detector de Arreglo de Diodos (DAD) que adquiere los espectros en todas las longitudes de onda (201) casi instantáneamente, mientras que en el fluorímetro los espectros de Excitación se obtienen uno a uno, variando entre cada uno la longitud de onda de Emisión, todo lo cual requiere tiempo para posicionar a los monocromadores. Como consecuencia de esta menor velocidad de procesamiento en el fluorímetro, las placas de Elisa que eran colectadas a la salida del HPLC requerían un tiempo de espera antes de poder ser leídas y esto produjo acumulaciones de placas. Si bien éstas eran inmediatamente recubiertas con un *film* para evitar evaporaciones o contacto con la atmósfera, es posible que hayan existido fenómenos de cinética desconocida que podrían haber causado la disminución en las señales. No obstante, ya que los cambios apreciados eran de intensidad pero no de “forma espectral”, se decidió incluir a las muestras de calibración conflictivas en los apilamientos y tenerlas en cuenta para los ajustes con MCR-ALS. En una etapa posterior de Calibración pseudo-univariada las áreas resueltas correspondientes a las muestras en cuestión no fueron tenidas en cuenta. Para datos UV no existieron exclusiones de ningún tipo.

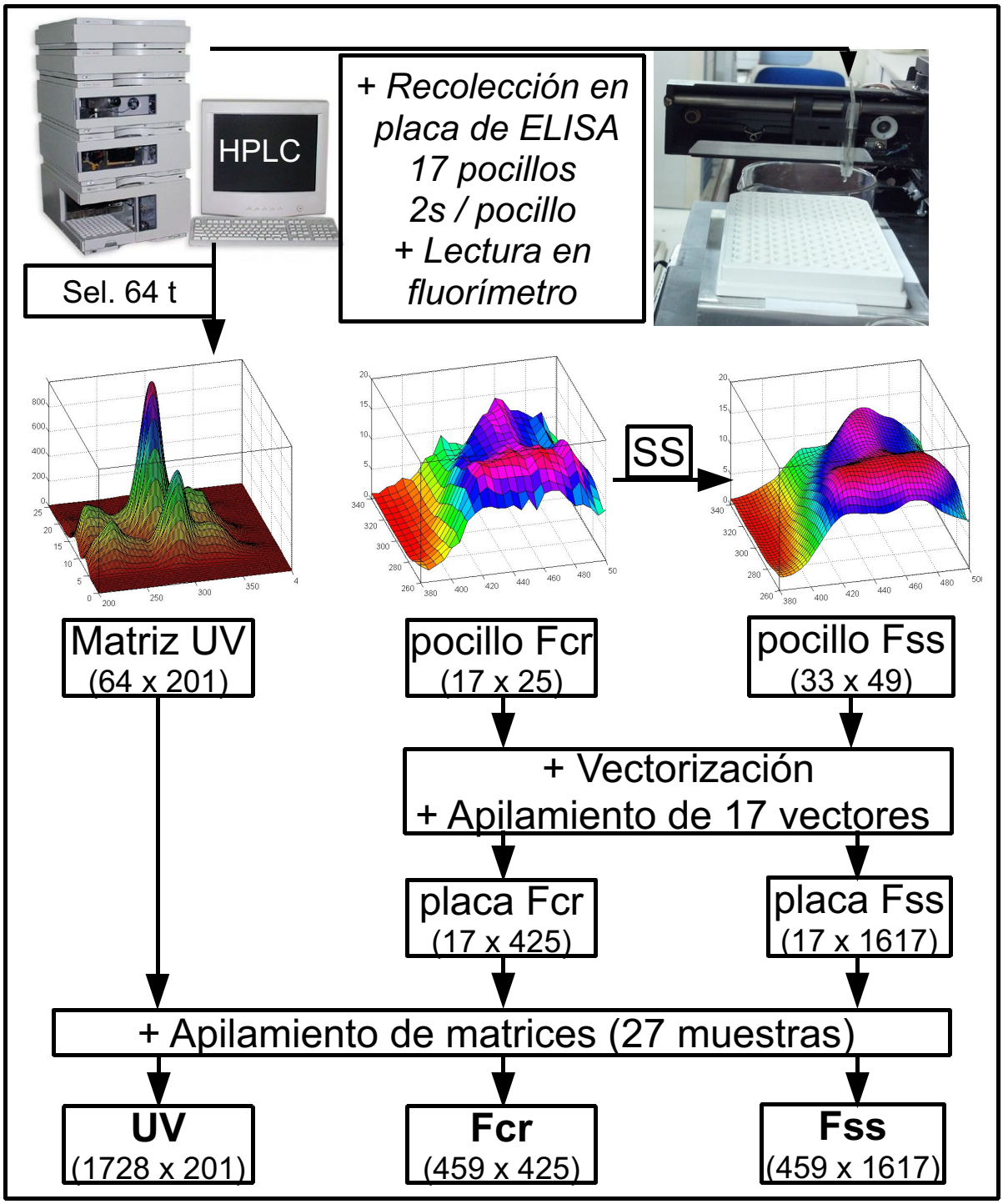


Figura 9: Esquema de la metodología realizada para obtener datos a procesar con MCR-ALS

Referencias: Sel. 64 t: Selección de 64 tiempos, SS: Spline y suavizado de Savitsky-Golay

3.6.5 Análisis MCR-ALS de muestras en simultáneo

Una vez que se obtuvieron las matrices apiladas UV, Fcr y Fss, se procedió a determinar mediante SVD la cantidad de componentes recomendados para explicar un 90% de la varianza presente, los cuales resultaron ser 2, 13 y 5, respectivamente. En primer lugar conviene destacar que ninguno coincide con 3, que es la cantidad real de analitos incluidos en las mezclas. En el caso de UV, la estimación menor a 3 sugiere la existencia de solapamientos reales entre analitos. El gran exceso de componentes para Fcr puede ser debido a que al no existir procesamientos, los datos crudos presentan variaciones abruptas en algunas regiones (falta de suavidad) y como resultado estas fuentes de varianza no son asociadas entre sí y se aproximan como fuentes independientes. Finalmente, las estimación para Fss parece más cercana a la realidad, ya que el exceso de componentes respecto de 3 es escaso y podría representar fenómenos presentes pero no debidos a los analitos.

En base a los resultados anteriores y a sabiendas de la cantidad de componentes reales en las mezclas, se decidió que UV fuera resuelto con 3 componentes. Por el lado de los datos de fluorescencia, ambos fueron resueltos con 4 y 5 componentes, verificando que con 4 se obtuvieron mejores resultados y estableciendo de esta manera la cantidad definitiva que dio origen a los resultados que serán reportados y analizados. Vale desde ya destacar que el componente extra en Fcr y Fss se interpretó como un efecto de línea de base no presente en los datos UV.

Para las aproximaciones iniciales de MCR-ALS, no se utilizó SIMPLISMA, sino que se sacó provecho de la presencia de muestras con un sólo componente, de las cuales pueden obtenerse buenas estimaciones de los espectros puros de los analitos. Las muestras elegidas para obtener estas aproximaciones fueron las respectivas a la concentración media disponible para cada analito, es decir, OFL6, CPF 9 y DNF15. En el caso de UV se tomaron los espectros en los tiempos donde se presentaron los máximos de absorbancia. Para los datos de fluorescencia, el equivalente a un determinado tiempo es un pocillo, y las matrices de Excitación/Emisión de cada pocillo se resuelven vectorizadas, de forma tal que cada vector es en realidad la concatenación de varios espectros de Excitación (uno por cada longitud de onda de Emisión). Por lo tanto, para OFL6, CPF9 y DNF15 se analizaron los 17 vectores pertinentes y se obtuvo la aproximación del espectro puro (concatenación de espectros de Excitación) con aquel vector que resultara poseer el valor máximo. A su vez, fue necesaria una estimación para el cuarto componente extra, la cual se obtuvo a partir del pocillo 17 de OFL6. Claramente a nivel visual los pocillos 17 básicamente ya no contenían

señales relacionadas a los analitos, por lo cual las señales apreciadas podrían atribuirse a líneas de base y de allí la decisión de usar esta información para modelar al cuarto componente. A su vez, la elección de OFL6 y no de otra muestra para proveer la información del pocillo 17 fue realizada teniendo en cuenta que en la etapa de calibración pseudo-univariada (luego de la resolución con MCR-ALS) no serían utilizadas 2 de las 5 muestras de Calibración para OFL, y en un intento de contrarrestar este efecto negativo, aportar la aproximación inicial podría resultar positivo.

En cuanto al uso de restricciones en MCR-ALS, se utilizaron las siguientes:

- No-negatividad en espectros y concentraciones de todos los componentes
- Unimodalidad para perfiles de concentración de todos los componentes
- Matriz de Correspondencia entre muestras experimentales y especies modeladas: esta matriz se construyó con una cantidad de filas igual a la cantidad de matrices apiladas (una por mezcla analizada) y con tantas columnas como componentes modelados. La presencia de 0 para la muestra "i" y para el componente "j" asegura la ausencia del último en la muestra. Por lo tanto, originalmente se obtuvo una matriz de "unos" (posibilidad de presencia) y luego se intercambiaron "ceros" para las muestras de calibración que sólo contenían un componente. En el caso de los datos de fluorescencia, el componente extra modelado (línea de base) se postuló como potencialmente presente en todas las muestras, es decir, con valor de 1 en todos los valores de su respectiva columna.
- Normalización de los espectros puros
- Criterio de convergencia del 0,1% en la diferencia de la desviación estándar de los residuos entre iteraciones sucesivas
- Cantidad de iteraciones en 50 permitidas como máximo.

3.6.5.1 Cifras de mérito de los ajustes

Las cifras de mérito de los ajustes para las distintas resoluciones de las matrices apiladas y otros detalles de las últimas se exponen en la tabla 3. Como puede apreciarse en dicha tabla, todas las resoluciones alcanzaron altos % de Varianza Explicada. Sin embargo, la calidad del ajuste con los datos UV fue superior, lo cual queda especialmente plasmado comparando los %LOF EXP. En este sentido, un factor que podría ser importante es la calidad de las señales originales, siendo las de UV matrices con superficies suaves, no siéndolo las de Fcr, y siéndolo parcialmente las de Fss. De

hecho, comparando a Fcr con Fss, puede intuirse que las mejorías en las cifras de mérito provienen del acondicionamiento de las señales espectrales que solamente posee Fss, ya que ambas poseen el mismo nivel de resolución en el sentido de las concentraciones (17 pocillos por muestra). En relación a lo último, otro factor que también debe considerarse influyente es que en UV la separación de los analitos mediante HPLC se realizó en condiciones controladas por el equipo (aun así se apreciaron diferentes tiempos de retención resueltos por MCR-ALS entre distintas muestras para los mismos analitos, por lo que el control no fue absoluto) y la resolución de captura espectral fue mayor. En cambio, en Fcr (y por consiguiente en Fss) la acumulación de gotas en un mismo pocillo implica mezcla, disminución de la resolución cromatográfica, más allá de posibles caídas de gotas "fuera de lugar" y de efectos de dispersión u otros que pudieron haberse dado desde la salida del HPLC, por el capilar de recolección, hasta cada pocillo en las placas de ELISA.

Matriz apilada	UV	Fcr	Fss
filas	1728	459	459
columnas	201	425	1617
ncomp	3	4	4
%LOF EXP	1.717	15.197	12.900
$\%R^2$	99.971	97.691	98.336
iter	18	3	3

Tabla 3: Detalles y cifras de mérito por matriz apilada para MCR-ALS

Referencias: ncomp: número de componentes en MCR-ALS, %LOF EXP: Porcentaje de Falta de Ajuste Experimental, $\%R^2$: Porcentaje de Varianza Explicada, iter: cantidad de iteraciones

Vale destacar que se realizaron algunas variantes de cálculo no reportadas para Fcr y Fss, como ser resolver con 5 componentes o utilizando otras aproximaciones para el cuarto componente, pero no se obtuvieron mejorías significativas. También vale mencionar que las resoluciones para Fcr y Fss que fueron expuestas no alcanzaron el criterio de convergencia por defecto (cambio en la desviación estándar de los residuos mayor a 0.1% entre iteraciones consecutivas) y que los cómputos fueron detenidos automáticamente al no encontrarse mejorías durante 20 iteraciones consecutivas posteriores a la número 3, informada como óptima.

3.6.5.2 Perfiles resueltos

Usualmente, la evaluación de los perfiles espectrales resueltos suele darse en comparación con las estimaciones que se posean para los espectros puros. En estas experiencias, los últimos a su vez sirvieron como aproximaciones iniciales de cada analito para la resolución con MCR-ALS, por lo que tras las optimizaciones ambos resultaron muy similares en todos los casos. Más aún, si se calcula el coeficiente de correlación entre cada espectro puro y su respectivo perfil resuelto, el menor de ellos resulta de por sí bastante elevado (0.9997, 0.9851 y 0.9871 para UV, Fcr y Fss, respectivamente). Por consiguiente y para no recargar más aun a las gráficas, en la figura 10 sólo se exponen los perfiles espectrales resueltos y detalles derivados de éstos. Como ya se comentó, dado que los coeficientes de correlación respecto de espectros puros resultaron en altos valores, la forma de los espectros resueltos es aceptable y no merece mayores comentarios. Sí vale la pena analizar las gráficas de Fcr y Fss. En principio es conveniente destacar la similitud de las formas resueltas. Dado que en Fss se han agregado puntos entre los de Fcr, las curvas resultan más suaves y densas, pero no más informativas a primera vista. En cuanto a la diferencia de valores se aprecia que el máximo de Fcr es básicamente el doble que el de Fss, lo cual proviene de que los espectros expuestos son la salida directa de MCR-ALS, donde se los ha normalizado, aun cuando el número de variables no coincida. Si se aprecia cualquier perfil resuelto (en línea de puntos) se observan formas repetidas, que son los espectros de Excitación concatenados, cada uno pertinente a una longitud de onda de Emisión. Si se toma el máximo, puede deducirse qué longitud de onda de Excitación lo presentó. Realizando esto individualmente para cada componente, se obtuvo un valor para cada uno y los valores de intensidad de fluorescencia normalizada de cada espectro de Excitación en el mencionado valor fueron unidos (líneas continuas) conformando estimaciones de los espectros de Emisión de cada analito. Con la misma lógica se procedió a obtener estimaciones de los espectros de Excitación en la longitud de onda donde fue máxima la Emisión, lo cual queda plasmado en los recuadros insertos, donde en el eje horizontal a su vez se indicaron los valores de longitud de onda reales en nm. Finalmente, sobre el cuarto componente (c4) se aprecia que su aporte no resulta constante con el cambio de variables.

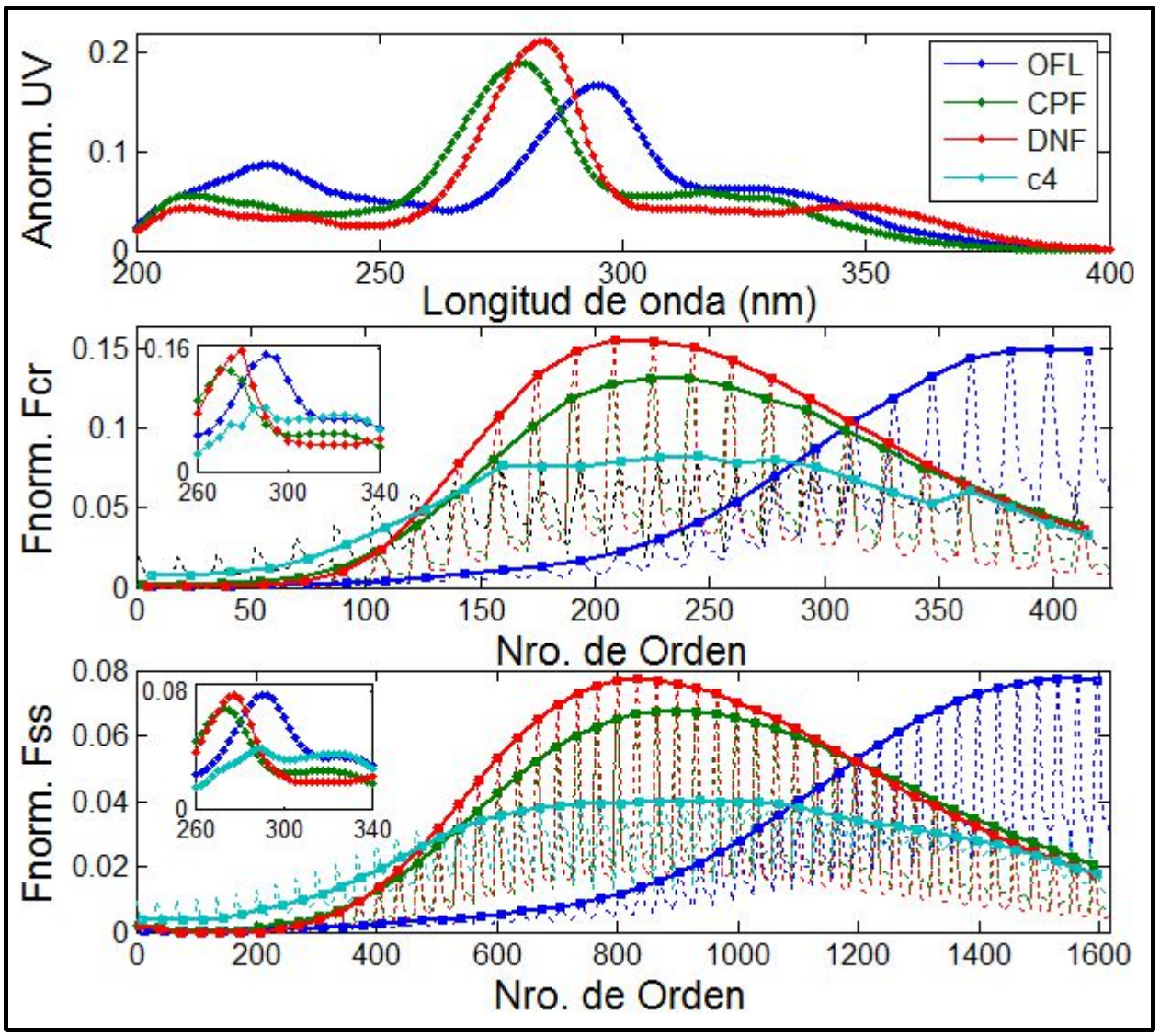


Figura 10: Perfiles espectrales resueltos con MCR-ALS para datos apilados UV, Fcr y Fss

Referencias: Anorm: Absorbancia normalizada, Fnorm: intensidad de Fluorescencia normalizada, c4: componente 4 (sólo fluorescencia). Para Fcr y Fss, las líneas de trazo discontinuo representan a los perfiles espectrales resueltos, las de trazo continuo a espectros de Emisión y las presentes en el recuadro inserto a espectros de Excitación (ver texto).

Respecto de los perfiles de concentración resueltos, la figura 11 expone resultados para las muestras originalmente concebidas como de Calibración.

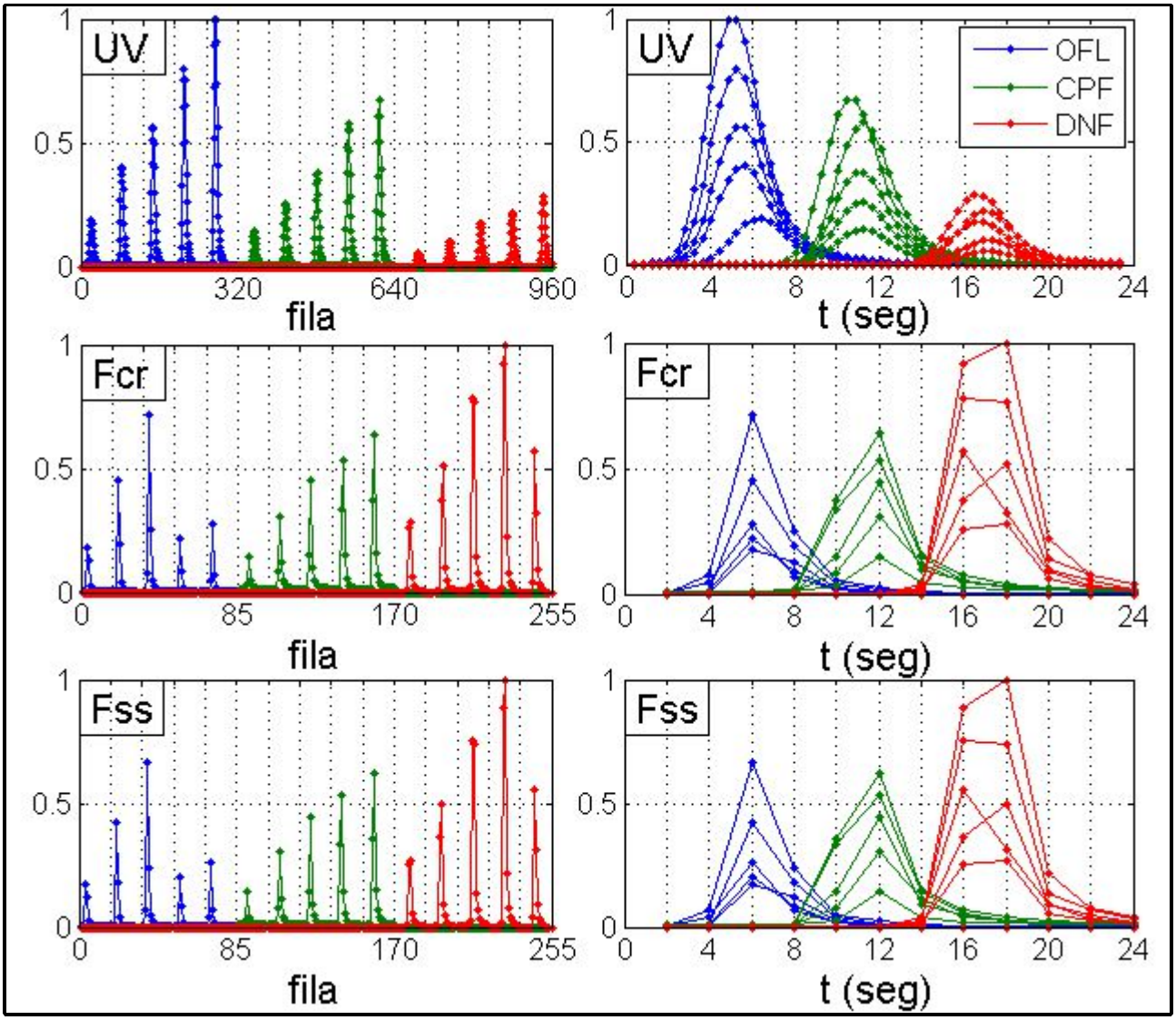


Figura 11: Perfiles de concentración resueltos con MCR-ALS para datos apilados UV, Fcr y Fss, en las muestras originales de Calibración

Referencias: fila: número de fila en cada matriz apilada, t (seg): tiempo en segundos. Los valores en el eje vertical de cada gráfica corresponden a los perfiles resueltos escalados por el máximo de cada gráfica. En las gráficas de la izquierda, las muestras de cada analito se encuentran ordenadas según concentración creciente. Se obvió el cuarto componente de fluorescencia.

Las gráficas de la izquierda en la figura 11 permiten apreciar claramente cómo con UV las áreas resueltas para las 15 muestras de Calibración resultaron mantener una relación lineal aparentemente apropiada entre la concentración de cada analito y el área resuelta bajos los perfiles (por la escala utilizada, la linealidad se aprecia más evaluando los picos máximos). En cambio, para Fcr y Fss las dos muestras más concentradas de OFL y la más concentrada de DNF obtuvieron áreas menores a

lo que se esperaría. Esto está de acuerdo con la observación sobre las señales originales de esas matrices, por lo cual se decide excluirlas de las posteriores calibraciones pseudo-univariadas basadas justamente en las áreas bajos las respectivas curvas. Como ya se ha dicho, una razón para explicar estas disminuciones podría radicar en el tiempo de espera entre colecciones y lecturas en fluorímetro, que no fue estrictamente el mismo para todas las muestras analizadas, lo cual ciertamente debería estandarizarse para experiencias futuras similares. En las gráficas de la izquierda también se dejaron señalados los límites de recolección entre muestra y muestra (líneas de puntos verticales). La comparación entre cualquier muestra en UV respecto de una inferior en Fcr y/o Fss deja ver la diferencia de suavidad debida a la resolución de cada curva (64 tiempos en UV, 17 pocillos/2 segundos en Fcr y Fss). También es posible apreciar que los picos en Fcr/Fss parecen estar antes en el tiempo en relación a los de UV, lo cual no podría ser posible ya que el detector UV se encuentra antes en el camino del flujo. En efecto, esto sólo es producto de que los datos UV fueron recortados y limitados en el tiempo, mientras que para los de fluorescencia los últimos pocillos básicamente no contenían señales de los analitos pero aún así fueron resueltos y por ende graficados. Dicho de otra manera, si bien los datos UV y los de fluorescencia coinciden en el tiempo cero de cada perfil, los datos en UV terminan aproximadamente a los 25 segundos, mientras que los restantes lo hacen en 34 segundos (17 pocillos, 2 segundos en cada uno).

Las gráficas a la derecha de la figura 11 dejan ver otro tipo de detalles. Aquí cada perfil puede apreciarse en función del tiempo en segundos, y las líneas discontinuas verticales coinciden con un cambio de pocillo (2 segundos). A su vez, el límite superior de tiempo coincide con el real de UV y aún así los perfiles en Fcr/Fss bajo cuyas curvas existe área integrable (con utilidad real para las cuantificaciones) se aprecian básicamente enteros, lo cual evidencia que los últimos 5 pocillos podrían haber sido descartados. De lo anterior podría surgir la opción de eliminar de cada muestra los últimos pocillos, con lo cual podrían obtenerse mejoras en los ajustes, pero más importante aún, si las muestras volvieran a ser recolectadas, se obtendría reducción del tiempo necesario para procesar cada muestra y del tiempo de uso de la lámpara.

Se observa que los perfiles de UV en mayor resolución son claramente más suaves que los de fluorescencia. Si se comparan los últimos entre sí, a simple vista resultan muy similares. En cuanto a las formas, sólo vale destacar que los perfiles de OFL y CPF presentan buen alineamiento en sus picos, mientras que para DNF el cambio de pocillo entre los tiempos de 16 y 18 segundos coincidió con el momento en el cual se recolectaban las concentraciones máximas de DNF, por lo cual existen

máximos en ambos tiempos y esto da la impresión de un ensanchamiento de pico.

Un detalle interesante para analizar es la correlación de tiempos entre datos en UV y fluorescencia. En principio vale resaltar que ya en el HPLC los perfiles resueltos muestran disparidades respecto de los tiempos donde deberían presentarse los máximos de cada analito. Por un lado, esto habla de defectos de repetibilidad en el instrumental, aunque tolerables. Por otro lado, podría asumirse que estos corrimientos deberían verse reflejados en los perfiles resueltos en fluorescencia. Esto no se aprecia, lo cual es producto de la disminución en la resolución de tiempos durante las colecciones. No obstante, esta necesidad de alineamiento no es trascendental, ya que las cuantificaciones se realizarán en base a áreas resueltas y no a tiempos específicos. También vale comentar que el hecho de que las resoluciones numéricas hayan sido posibles a pesar de los corrimientos de algunos picos es mérito de MCR-ALS. Otros algoritmos más dependientes de alineamiento pueden ser menos tolerantes a este tipo de rupturas de la trilinearidad. Igualmente, aunque no es necesario que la relación de tiempos sea estricta, vale mencionar que a groso modo los picos en fluorescencia se aprecian aproximadamente un segundo más tarde que en UV, tiempo que podría pensarse como el necesario para atravesar el capilar de colección. Al respecto del último, es posible que al no existir más el relleno de la columna cromatográfica los componentes cromatografiados incurran en difusiones o fenómenos similares, aunque esto no se verifica al comparar datos UV y de fluorescencia, lo cual puede ser debido a que la longitud del capilar de colección fue la mínima necesaria para conectar la salida del HPLC con el pocillo más lejano de cada placa.

Por razones ya expuestas, debe entenderse que el objetivo de estas experiencias no radica en una coordinación perfecta, sino en una lo suficientemente aceptable para aprovechar los beneficios de la separación cromatográfica previa. De hecho, existen varias otras razones contrarias a una coordinación real:

- Coordinación entre inyección y recolección: el inicio de una recolección (en realidad de los 47 segundos de *delay*) se realizó manualmente una vez que el HPLC daba la señal de haber inyectado cada muestra. Si bien se puso mucha atención en este paso, es una fuente posible de variabilidad. Una forma de evitar esto consistiría en realizar una intervención tal que la señal de inyección sea automáticamente interceptada y utilizada para comenzar la espera antes de la primera recolección. A nivel de software y hardware, esto podría realizarse mediante análisis de imágenes (el software Chemstation que opera al HPLC cambia de color un ícono al realizar la inyección, lo cual es

sencillo de detectar) o bien escribiendo rutinas de programación en el lenguaje propio del HPLC (esto ha sido realizado para trabajos no reportados en la presente tesis) que generen una señal eléctrica, en los puertos traseros del HPLC mismo o en algún puerto de la computadora a cargo, detectable y procesable por Arduino. Los instrumentos analíticos suelen tener conectores cuyos voltajes, leídos según instrucciones del fabricante, pueden ser decodificados para proporcionar información sobre el status de distintas operaciones, como bien podría ser "comienzo de inyección" en el HPLC. A modo de ejemplo y aunque no fue puesto en práctica, en el caso del fluorímetro utilizado existe un conector que otorga un voltaje cuando el instrumento se encuentra adquiriendo espectros y otro cuando está libre para recibir nuevas peticiones. Este tipo de "mensajes" permiten el desarrollo de automatizaciones.

- Curvatura del capilar de recolección: si bien la longitud del recorrido es siempre la misma, al cambiar de pocillo cambia la curvatura del capilar (el extremo a la salida del HPLC permanece siempre fijo y el otro es móvil), lo cual puede tener influencia

- Tiempo entre movimientos: aquí existen varios casos. El primero de ellos es el movimiento inicial desde el Depósito 1 hacia el pocillo 1 y éste tiempo debería adicionarse al supuesto necesario para atravesar el capilar de colección. En segundo lugar existe el movimiento genérico necesario para ir desde el pocillo N al N+1. En tercer lugar existen casos particulares en los que al cambiar de pocillo, cambia la columna de la placa, con lo cual el movimiento debe ser realizado por todo el puente al mismo tiempo y por razones constructivas (para evitar temblores y el desprendimiento de gotas) estos movimientos se realizaron a velocidades menores que las implicadas en los movimientos dentro de una misma columna de cada placa. Esto último puede tener relación con los máximos de DNF en los tiempos de 16 y 18 segundos. Precisamente, al concluir los 16 segundos se habrán colectado los 8 pocillos de la primera columna de la placa y por lo tanto se produce el cambio hacia la segunda. Este movimiento es más lento, por lo cual es más probable que algunas gotas caigan en el camino entre pocillos. Si se observan los perfiles UV, los máximos aparecen entre 16 y 17 segundos, por lo que con la demora que implica el traspaso por el capilar, deberían llegar a la placa a los 17 y 18 segundos o poco más. Por ende, es más probable que los máximos en fluorescencia deberían presentarse a los 18 segundos y de no ser así, podría asumirse pérdida de gotas concentradas. Entre los perfiles de DNF, se aprecia uno que claramente tiene su máximo en 16, no en 18. Esa muestra es a su vez la más concentrada que será descartada para las calibraciones pseudo-univariadas. Como su área debería ser la mayor de todas, es aún más probable que se haya

perdido analito en esa recolección. Un razonamiento similar podría explicar la exclusión de las dos muestras de OFL que presentaron señales y áreas menores a lo esperado, es decir, bien pudieron haberse perdido gotas durante la recolección de estas muestras, aunque para el caso de OFL igualmente todos los máximos resultaron alineados a los 6 segundos.

Todos estos movimientos descriptos se simplifican como instantáneos pero en realidad no lo son, ya que requieren tiempos que incluso pueden variar levemente. De hecho, para evitar que los motores eleven demasiado su temperatura, se decidió que al llegar a cada pocillo, las bobinas fueran desenergizadas mediante corte del suministro eléctrico. Al retomar los movimientos, existe por tal un tiempo necesario para re-energizar las bobinas apropiadas y un tiempo mínimo para vencer a la inercia, todo lo cual puede afectar la coordinación si no se tiene repetibilidad. También en relación a los motores existen las velocidades asociadas a cada uno, las cuales fueron escogidas con el objetivo de evitar otra de las posibles fuentes de falta de coordinación, que es la caída de gotas fuera de tiempo (antes o después de donde deberían) y lugar (dentro o fuera del pocillo donde deberían caer), tal y como se discutió recientemente para la muestra DNF25. Experimentalmente se comprobó que si las velocidades eran altas, al llegar a un pocillo de destino el frenado era muy brusco y esto provocaba el balanceo de gotas y su caída fuera de lugar. De forma similar, cuando las velocidades eran bajas las gotas caían entre pocillos al pasar desde uno a otro, por lo cual se perdía parte de las muestras. Así pues, se determinaron empíricamente velocidades pseudo-óptimas para cada motor, siendo de 100 rpm para el movimiento vertical y de 30 rpm para el horizontal (a mayores velocidades mayores temblores del puente entero), con ambos valores constantes durante todas las experiencias.

El conteo de gotas no sólo podría mejorar la coordinación sino también haría que las recolecciones no fueran por tiempo sino por cantidad de gotas. Con esto también existiría la posibilidad de minimizar la probabilidad de perder gotas durante un movimiento, de certificar si aún así hubo gotas perdidas y en tal caso entre qué pocillos se habrían perdido. Con lo último podría evitarse la exclusión de calibradores, ya que al ser éstos conocidos y al tener datos de los pocillos vecinos, sería posible realizar una reconstrucción matemática (estandarización espectral) de cómo debería haber sido la señal si no se hubieran perdido gotas. Como se ha visto, el problema de la caída de gotas tiene distintas aristas. Por un lado, las gotas que caigan en pocillos considerados incorrectos (antes o después de tiempo) afectarán la correlación de tiempos entre UV y

fluorescencia, pero como las áreas son las que serán utilizadas para las cuantificaciones, no habrá mayores inconvenientes con MCR-ALS. En cambio, las gotas que caigan fuera de cualquier pocillo implicarán pérdida real de analito y la potencial exclusión de la muestra en cuestión, a la vez que visualmente podrá parecer que los máximos difieren si es que las gotas perdidas correspondían al tiempo en que deberían haberse hallado los máximos.

Incluso si se lograran gotas de menor tamaño y cada una pudiese ser recolectada en un único pocillo, se supone que se mantendría en mayor medida la resolución cromatográfica. No obstante, esto no siempre sería aprovechable, ya que existen otros límites. Para el caso de estas experiencias, la sonda de fluorescencia utilizada en el lector de placas de ELISA tiene un tamaño definido por el fabricante, apto para leer pocillos de tamaño estándar (96 por placa, existen de mayor número y menor tamaño) y esto implica la recolección de un volumen mínimo, mayor a una gota, para obtener señales aceptables. A su vez, el conteo de gotas requiere alguna intervención de la muestra con una señal (óptica, eléctrica, etc.), lo cual podría ser cuestionado para analitos potencialmente reactivos con dicha señal.

3.6.5.3 Calibraciones pseudo-univariadas

Previa exclusión de los perfiles de las muestras OFL8, OFL10 y DNF25, los valores de área de cada analito en las muestras de Calibración fueron utilizados para construir modelos lineales pseudo-univariados en relación a las concentraciones nominales de cada analito en las muestras, de los cuales se derivaron las respectivas pendientes y ordenadas al origen. Estos valores fueron utilizados para interpolar las áreas resueltas con cada muestra de validación y de esta forma se obtuvieron las predicciones de cada analito. Vale destacar que aunque las muestras más concentradas de 2 de los 3 analitos fueron excluidas, no se excluyeron del conjunto de Validación aquellas que por diseño hayan resultado tener concentraciones mayores y por ende fuera del intervalo calibrado para cada analito. Los resultados para las predicciones de Validación, de calidad muy diferente según cada analito, se exponen en la tabla 4. Si se analizan las predicciones individualmente para cada muestra (no mostrado) se observa que existen algunas de gran calidad y otras con errores muy grandes. En el caso de OFL las cifras resultaron aceptables, con valores de RMSEV similares para los tres tipos de datos procesados. La diferencia más notable en REP% en parte se da porque con UV no hubo exclusiones de muestras, por lo que la media de concentraciones (denominador del REP%) con un valor de 6.00 ppm resulta mayor que en el caso

de los datos de fluorescencia, donde las mayores concentraciones de OFL fueron excluidas y por ende su concentración media resultó en 4.00 ppm. En cuanto a las recuperaciones medias porcentuales, resultaron muy parecidas.

En el caso de CPF, teniendo en cuenta el intervalo calibrado entre 3.00 ppm y 15.00 ppm, los valores de RMSEV resultan inadmisibles. Esto ya se manifiesta con UV, por lo que no se pueden esperar mejorías con fluorescencia, lo cual invalida que sus cifras de mérito sean mejores.

Finalmente, con DNF se aprecian valores de RMSEV bajos nuevamente, aunque no en relación al intervalo calibrado, tal como indican los REP% (también debe tenerse en cuenta un menor denominador para los REP% de fluorescencia). Los resultados en general poseen una aceptabilidad intermedia entre la de OFL y la de CPF.

Matriz	UV	Fcr	Fss
OFL			
RMSEV	0.52	0.75	0.76
REP%	8.62	18.73	18.91
mRec	110.02	108.41	108.90
CPF			
RMSEV	9.18	4.34	3.85
REP%	101.99	48.19	42.78
mRec	157.01	114.42	115.21
DNF			
RMSEV	0.55	0.38	0.36
REP%	36.98	30.20	28.76
mRec	123.97	128.09	126.13

Tabla 4: Resultados analíticos para predicciones de Validación según calibraciones pseudo-univariadas con áreas UV, Fcr y Fss provenientes de MCR-ALS

Referencias: mRec%. media de Recuperaciones porcentuales

Un factor de variabilidad extra reside en que las muestras de Calibración y de Validación se procesaron en días distintos, pero si esto fuera muy relevante, OFL debería haber dado peores resultados. También, tras discusiones posteriores a las recolecciones y tras el análisis de los resultados, surgió la hipótesis de que pudieron cometerse errores durante la preparación de soluciones de CPF y DNF, específicamente relacionados a fallas en la pipetas capilares utilizadas.

Vale notar que en general los resultados de Fcr no difieren demasiado de los de Fss, por lo que el pre-tratamiento con *spline* y suavizado de Savitsky-Golay podría obviarse desde este punto de vista.

Finalmente, la tabla 5 expone cifras de mérito en relación a las calibraciones pseudo-univariadas.

Matriz	UV	Fcr	Fss
OFL			
SenMCR	265.40	74.32	143.56
InvSenAn	8.25E-4	6.89E-3	3.61E-3
LOD	2.72E-3	2.28E-2	1.19E-2
LOQ	8.25E-3	6.89E-2	3.61E-2
CPF			
SenMCR	61.15	9.24	16.88
InvSenAn	5.98E-2	3.84E-1	1.88E-1
LOD	1.97E-1	1.27E+0	6.21E-1
LOQ	5.98E-1	3.84E+0	1.88E+0
DNF			
SenMCR	155.80	136.31	253.66
InvSenAn	4.68E-3	1.32E-3	7.42E-4
LOD	1.54E-2	4.36E-3	2.45E-3
LOQ	4.68E-2	1.32E-2	7.42E-3

Tabla 5: Cifras de mérito para calibraciones pseudo-univariadas con áreas UV, Fcr y Fss provenientes de MCR-ALS

Referencias: SenMCR: Sensibilidad (Unidades de Señal · L/mg) , InvSenAn: Inversa de Sensibilidad Analítica (mg/L), LOD: Límite de Detección (mg/L), LOQ: Límite de Cuantificación (mg/L), E+n=x10^n

A diferencia de lo dicho recientemente, Fss obtuvo valores de sensibilidad aproximadamente dos veces superiores a los respectivos de Fcr, mientras que en el resto de las cifras de mérito (InvSenAn, LOD y LOQ) fueron aproximadamente la mitad. La relación entre cifras de mérito es lógica, ya que sensibilidad es el denominador común para el cálculo de las restantes cifras. Por lo tanto, el tratamiento que da origen a Fss debe interpretarse solamente como factor de aumento en la sensibilidad. Si se observa la ecuación (2) en la que se define a esta cifra de mérito, dado que tanto para Fcr como para Fss el valor *J* (cantidad de pocillos en este caso), un aumento de sensibilidad

con Fss podría ser debido a un aumento en la pendiente de la recta pseudo-univariada, o bien a una disminución del factor relativo al grado de solapamiento entre perfiles. Evaluando a éstos últimos, los valores en Fcr y Fss son, respectivamente, 2.06 y 2.09 para OFL, 26.60 y 32.53 para CPF, 20.17 y 24.60 para DNF. Como puede apreciarse, en el caso de Fss los valores siempre resultaron mayores, por lo que si el grado de solapamiento fue incluso mayor, la diferencia entre las sensibilidades se debió fundamentalmente a pendientes mucho mayores en Fss que en Fcr.

Si se comparan los resultados entre UV y Fss como mejor alternativa de fluorescencia, UV obtuvo mejores valores de sensibilidad (y por ende del resto de la cifras de mérito, comparativamente hablando) para OFL y CPF, y lo inverso ocurrió con DNF. En este último caso vale destacar que se obtuvo una de las mayores de todas las sensibilidades y que incluso Fcr obtuvo un resultado similar a UV, aún cuando DNF fue el analito presente en menores concentraciones que el resto. En todos los casos debe recordarse que *J* fue 64 para UV y sólo 17 para fluorescencia. Por ende, puede pensarse que aunque los datos de fluorescencia poseían una menor resolución en términos cromatográficos, ésta fue suplida por una buena (y hasta mejor) resolución espectral. A su vez, si se aprecia en la figura 11 que las concentraciones de los tres analitos son básicamente nulas en los pocillos 1 y 13 al 17, se podría pensar que una resolución sin la información de estos pocillos arrojaría perfiles similares a los ya obtenidos (con un grado de solapamiento similar), por lo que las áreas y por ende las pendientes básicamente no cambiarían, pero sí disminuiría *J*, con lo cual aumentaría la sensibilidad (desde luego sólo en términos matemáticos). Si estos 6 pocillos no fueran tenidos en cuenta, *J* valdría 13 y las sensibilidades de fluorescencia serían aproximadamente 14.35% mayores que las tabuladas. Con este valor, por ejemplo, la sensibilidad en Fcr también superaría levemente a UV para DNF.

3.7 Conclusiones

– Se pudo construir y programar un dispositivo para recolectar muestras automáticamente en placas de ELISA. Se concluye que el reciclaje de componentes y la aplicación de hardware de código abierto resultó apropiado para lograr estos objetivos.

– Se elaboraron interfaces gráficas para comunicación entre una PC con el recolector o con el fluorímetro. Esto permitió obtener matrices de Excitación-Emisión que luego fueron procesadas. La aplicación de software de código abierto proveyó buenas herramientas tanto para el desarrollo de

interfaces gráficas como para el de circuitos impresos.

– La resolución mediante MCR-ALS tanto con datos UV como de fluorescencia permitió obtener información de diferentes aspectos del procedimiento completo, desde la evaluación de ajustes y cuantificaciones mediante sus cifras de mérito hasta detalles relativos a la coordinación mecánica. De alguna manera lo último resulta en una aplicación novedosa de MCR-ALS, es decir, la aplicación del algoritmo para evaluar el funcionamiento de un dispositivo.

– Se vio que existen varias razones para que la coordinación entre la separación cromatográfica y la recolección en pocillos no sea ideal (movimientos, velocidades, fenómenos dentro del capilar de recolección, entre otros), pero que esto no resulta en impedimentos para MCR-ALS, con lo cual los perfiles resueltos permiten realizar cuantificaciones aunque puedan encontrarse desalineados en el tiempo. No obstante, algunos de estos factores de variabilidad podrían ser minimizados y esto posibilitaría la aplicación de algoritmos más sensibles a la falta de coordinación.

– Si las cuantificaciones se realizan en base a las áreas bajo los perfiles de concentración resueltos, entonces sólo la caída de gotas fuera de todo pocillo sería un factor determinante para la exclusión de predicciones. De aquí se deriva la utilidad de un sistema de conteo de gotas con el cual coordinar flujo a recolectar y movimientos entre pocillos.

– Las herramientas de código abierto representan mayores potencialidades para el laboratorio de investigación, brindando soporte a través de comunidades de usuarios predispuestos a compartir el conocimiento, y permitiendo mayores niveles de concreción experimental. Es digna de destacar la utilidad de placas del tipo Arduino, capaces de controlar muchos otros sensores y actuadores que los aquí utilizados, con lo cual la fabricación de instrumental más avanzado que el reportado se torna viable.

– Los desafíos propuestos fueron superados de manera aceptable, aun cuando gran parte del conocimiento necesario no estaba originalmente al alcance de alguien formado fundamentalmente en cálculos quimiométricos. Quien escribe comenzó estas tareas sabiendo programar en algunos lenguajes, pero no en ninguno de los utilizados (Arduino y Processing), y básicamente sin conocimientos de electrónica. Luego de un breve tiempo de familiarización con los entornos, comandos y componentes, no sólo que se logró el cumplimiento de objetivos, sino que a su vez las barreras experimentales, una vez superadas, dejan de ser percibidas como algo que debería ser resuelto sólo por terceros calificados.

Conclusión general del trabajo de tesis

Más allá de las conclusiones ya expuestas en los respectivos capítulos del texto, se puede decir que la aplicación de herramientas quimiométricas pre-existentes, así como su desarrollo, permitieron resolver muestras de origen biológico y/o químico desde un punto de vista químico-analítico. Estas herramientas pudieron ser aplicadas a datos de distinto orden obtenidos de diversas fuentes e instrumentales. La concentración de analitos y/o la cuantificación/calificación de otras propiedades pudieron ser determinadas y analizadas en distintos contextos. Oportunamente se extrajeron múltiples conclusiones relativas a la determinación de Etanol, de contenido proteico en muestras de maíz, de presencia/ausencia del pesticida Carbofurano en muestras derivadas de frutos de tomate, y de concentración de fluoroquinolonas en mezclas. Lo analizado reviste importancia desde puntos de vista de producción industrial y agraria, así como también a nivel de contaminación ambiental con pesticidas o antibióticos. Finalmente, vale destacar el uso de tecnologías de código abierto para potenciar las capacidades del laboratorio analítico.

Bibliografía

Abad, A.; Moreno, M.J. y Montoya, A. (1997) *A monoclonal immunoassay for carbofuran and its application to the analysis of fruit juices*. Anal. Chim. Acta 347, 103–110.

Abad, A.; Moreno, M.J.; Pelegrí, R.; Martínez, M.I.; Sáez, A.; Gamón, M. y Montoya, A. (1999) *Determination of carbaryl, carbofuran and methiocarb in cucumbers and strawberries by monoclonal enzyme immunoassays and high-performance liquid chromatography with fluorescence detection: An analytical comparison*. J. Chromatogr. A 833, 3–12.

Anastassiades, M.; Lehotay, S.J.; Stajnbaher, D. y Schenck, F.J. (2003) *Fast and easy multiresidue method employing acetonitrile extraction/partitioning and "dispersive solid-phase extraction" for the determination of pesticide residues in produce*. J. AOAC Int. 86, 412–431.

Anderson, C.E. y Kalivas, J.H. (1999) *Fundamentals of Calibration Transfer Through Procrustes Analysis*. Appl. Spectrosc. 53, 1268–1276.

Anderssen, E.; Dyrstad, K.; Westad, F. y Martens, H. (2006) *Reducing over-optimism in variable selection by cross-model validation*. Chemom. Intell. Lab. Syst. 84, 69–74.

Andries, E.; Hagstrom, T.; Atlas, S.R. y William, C. (2007) *Regularization strategies for hyperplane classifiers: application to cancer classification with gene expression data*. J. Bioinform. Comput. Biol. 5, 79–104.

Arancibia, J.A.; Boschetti, C.E.; Olivieri, A.C. y Escandar, G.M. (2008) *Screening of Oil Samples on the Basis of Excitation−Emission Room-Temperature Phosphorescence Data and Multiway Chemometric Techniques. Introducing the Second-Order Advantage in a Classification Study*. Anal. Chem. 80, 2789–2798.

Aster, R.C.; Borchers, B. y Thurber, C.H. (2005) *Parameter Estimation and Inverse Problems*. Elsevier, Amsterdam, Países Bajos.

Barker, M. y Rayens, W. (2003) *Partial least squares for discrimination*. J. Chemom. 17, 166–173.

Bauza, M.C.; Ibañez, G.A.; Tauler, R. y Olivieri, A.C. (2012) *Sensitivity Equation for Quantitative Analysis with Multivariate Curve Resolution-Alternating Least-Squares: Theoretical and Experimental Approach*. Anal. Chem. 84, 8697–8706.

Bechtel, K.L. (1997) *Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares*. J. Chemom. 11, 181–238.

Booksh, K.S. y Kowalski, B.R. (1994) *Theory of Analytical Chemistry*. Anal. Chem. 66, 782A–791A.

Bos, M. y Vrielink, J.A.M. (1994) *The wavelet transform for preprocessing IR spectra in the identification of mono- and di-substituted benzenes*. Chemom. Intell. Lab. Syst. 23, 115–122.

Bro, R. y De Jong, S. (1997) *A fast non-negativity-constrained least squares algorithm*. J. Chemom. 11, 393–401.

Bro, R. y Sidiropoulos, N.D. (1998) *Least squares algorithms under unimodality and non-negativity constraints*. J. Chemom. 12, 223–247.

Capron, X.; Walczak, B.; de Noord, O.E. y Massart, D.L. (2005) *Selection and weighting of samples in multivariate regression model updating*. Chemom. Intell. Lab. Syst. 76, 205–214.

Censor, Y. (1977) *Pareto Optimality in Multiobjective Problems*. Appl Math Optim. 4, 41–59.

Chau, F.-T.; Liang, Y.-Z.; Gao, J. y Shao, X.-G. (2004) *Chemometrics: From Basics to Wavelet Transform*. John Wiley & Sons, Estados Unidos.

Chen, G. y Harrington, P.B. (2003) *SIMPLISMA applied to two-dimensional wavelet compressed ion mobility spectrometry data*. Anal. Chim. Acta 484, 75–91.

Claerbout, J.F. y Muir, F. (1973) *Robust modeling with erratic data*. Geophysics 38, 826–844.

Cocchi, M; Durante, C.; Foca, G.; Manzini, D.; Marchetti, A. y Ulrici, A. (2004) *Application of a wavelet-based algorithm on HS-SPME/GC signals for the classification of balsamic vinegars*. Chemom. Intell. Lab. Syst. 71, 129–140.

Cogdill, R.P.; Anderson, C.A. y Drennen, J.K. (2005) *Process Analytical Technology Case Study, Part III: Calibration Monitoring and Transfer*. AAPS PharmSciTech 6, 284–297.

Collantes, E.R.; Duta, R.; Welsh, W.J.; Zielinski, W.L. y Brower, J. (1997) *Preprocessing of HPLC Trace Impurity Patterns by Wavelet Packets for Pharmaceutical Fingerprinting Using Artificial Neural Networks*. Anal. Chem. 69, 1392–1397.

Cuesta Sánchez; F., Khots, M.S. y Massart, D.L. (1994) *Algorithms for the assessment of peak purity in liquid chromatography with photodiode-array detection. Part II*. Anal. Chim. Acta 290, 249–258.

D'Ausilio, A. (2012) *Arduino: a low-cost multipurpose lab equipment*. Behav. Res. Methods 44, 305–313.

Da Cunha, N.O. y Polak, E. (1967) *Constrained Minimization Under Vector-valued Criteria in Finite Dimensional Spaces*. J. Math. Anal. Appl. 19, 103–124.

Daubechies, I. (1992). Ten Lectures on Wavelets. SIAM.

Dax, A. (1992) *On Regularized Least Norm Problems*. SIAM J. Optim 2, 602–618.

De Jong, S. (1993) *SIMPLS: An alternative approach to partial least squares regression*. Chemom. Intell. Lab. Syst. 18, 251–263.

De Juan, A.; Maeder, M.; Martínez, M. y Tauler, R. (2000) *Combining hard- and soft-modelling to solve kinetic problems*. Chemom. Intell. Lab. Syst. 54, 123–141.

De Juan, A. y Tauler, R. (2003) *Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution*. Anal. Chim. Acta 500, 195–210.

De Juan, A.; Van der Heyden, Y.; Tauler, R. y Massart, D. (1997) *Assessment of new constraints applied to the alternating least squares method*. Anal. Chim. Acta 346, 307–318.

De Noord, O.E. (1994) *Multivariate calibration standardization*. Chemom Intell Lab Syst 25, 85–97.

Den, W. y Malinowski, E.R. (1993) *Investigation of copper(II)-Ethylenediaminetetraacetate complexation by window factor analysis of ultraviolet spectra*. J. Chemom. 7, 89–98.

DiFoggio, R. (2005) *Desensitizing models using covariance matrix transforms or counter-balanced distortions*. J. Chemom. 19, 203–215.

DiFoggio, R. (2007) *Influencing models to improve their predictions of standard samples*. J. Chemom. 21, 208–214.

Du, Y.P., Kasemsumran; S., Maruo, K.; Nakagawa, T. y Ozaki, Y. (2005) *Improvement of the Partial Least Squares Model Performance for Oral Glucose Intake Experiments by Inside Mean Centering and Inside Multiplicative Signal Correction*. Anal. Sci. 21, 979–984.

Ducruix, C.; Vailhen, D.; Werner, E.; Fievet, J.B.; Bourguignon, J.; Tabet, J.-C.; Ezan, E. y Junot, C. (2008) *Metabolomic investigation of the response of the model plant Arabidopsis thaliana to cadmium exposure: Evaluation of data pretreatment methods for further statistical analyses*. Chemom. Intell. Lab. Syst. 91, 67–77.

Dunn, W.B. y Ellis, D.I. (2005) *Metabolomics: Current analytical platforms and methodologies*. TrAC Trends Anal. Chem. 24, 285–294.

Eilers, P.H.C. (2003) *A Perfect Smoother.* Anal. Chem. 75, 3631–3636.

Eilers, P.H.C. y Marx, B.D. (2003) *Multivariate calibration with temperature interaction using two-dimensional penalized signal regression*. Chemom. Intell. Lab. Syst. 66, 159–174.

Faber, K. y Kowalski, B.R. (1996) *Prediction error in least squares regression: Further critique on the deviation used in The Unscrambler.* Chemom. Intell. Lab. Syst. 34, 283–292.

Faber, N.M.; Song, X.H. y Hopke, P.K. (2003) *Sample-specific standard error of prediction for partial least squares regression*. TrAC Trends Anal. Chem. 22, 330–334.

Fearn, T. (2001) *Standardisation and calibration transfer for near infrared instruments: a review.* J. Infrared Spectrosc. 9, 229–244.

Fernández Pierna, J.A.; Jin, L.; Wahl, F.; Faber, N.M. y Massart, D.L. (2003) *Estimation of partial least squares regression prediction uncertainty when the reference values carry a sizeable measurement error.* Chemom. Intell. Lab. Syst. 65, 281–291.

Feudale, R.N.; Woody, N.A.; Myles, A.J.; Tan, H.W.; Brown, S.D. y Ferré, J. (2002) *Transfer of Multivariate Calibration Models: A Review. Chemom*. Intell. Lab. Syst. 64, 181–192.

Forrester, J.B. y Kalivas, J.H. (2004) *Ridge regression optimization using a harmonious approach*. J. Chemom. 18, 372–384.

Galtier, O.; Abbas, O.; Le Dréau, Y.; Rebufa, C.; Kister, J.; Artaud, J. y Dupuy, N. (2011) *Comparison of PLS1-DA, PLS2-DA and SIMCA for classification by origin of crude petroleum oils by MIR and virgin olive oils by NIR for different spectral regions*. Vib. Spectrosc. 55, 132–140.

Garrido, M.; Lázaro, I.; Larrechi, M.S. y Rius, F.X. (2004) *Multivariate resolution of rank-deficient near-infrared spectroscopy data from the reaction of curing epoxy resins using the rank augmentation strategy and multivariate curve resolution alternating least squares approach*. Anal. Chim. Acta 515, 65–73.

Geladi, P. y Kowalski, B.R. (1986) *Partial least-squares regression: a tutorial*. Anal. Chim. Acta 185, 1–17.

Gemperline, P.J. y Cash, E. (2003) *Advantages of Soft versus Hard Constraints in Self-Modeling Curve Resolution Problems. Alternating Least Squares with Penalty Functions*. Anal. Chem. 75, 4236–4243.

Gong, F.; Liang, Y.-Z.; Cui, H.; Chau, F.-T. y Chan, B.-T. (2001a) *Determination of volatile components in peptic powder by gas chromatography–mass spectrometry and chemometric resolution*. J. Chromatogr. A 909, 237–247.

Gong, F.; Liang, Y.-Z.; Xu, Q.-S. y Chau, F.-T. (2001b) *Gas chromatography–mass spectrometry and chemometric resolution applied to the determination of essential oils in Cortex Cinnamomi*. J. Chromatogr. A 905, 193–205.

Grande, B.V. y Manne, R. (2000) *Use of convexity for finding pure variables in two-way data from mixtures*. Chemom. Intell. Lab. Syst. 50, 19–33.

Gui, W.; Jin, M.; Sun, L.; Guo, Y. y Zhu, G. (2009) *Residues determination of carbofuran in vegetables based on sensitive time-resolved fluorescence immunoassay*. Food Agric. Immunol. 20, 49–56.

Haaland, D.M. (2000) *Synthetic Multivariate Models to Accommodate Unmodeled Interfering Spectral Components during Quantitative Spectral Analyses*. Appl. Spectrosc. 54, 246–254.

Haaland, D.M. y Thomas, E.V. (1988) *Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information*. Anal. Chem. 60, 1193–1202.

Haar, A. (1910) *Zur Theorie der orthogonalen Funktionensysteme*. Math. Ann. 69, 331–371.

Halket, J.M.; Przyborowska, A.; Stein, S.E.; Mallard, W.G.; Down, S. y Chalmers, R.A. (1999) *Deconvolution gas chromatography/mass spectrometry of urinary organic acids – potential for pattern recognition and automated identification of metabolic disorders*. Rapid Commun. Mass Spectrom. 13, 279–284.

Hansen, P.C. (1998) *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. Philadelphia, Estados Unidos

Hastie, T.J.; Tibshirani, R.J. y Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining*, Inference, and Prediction. New York, Estados Unidos

Hoerl, A.E. y Kennard, R.W. (1970) *Ridge Regression: Biased Estimation for Nonorthogonal Problems*. Technometrics 12, 55–67.

Jalali-Heravi, M.; Zekavat, B. y Sereshti, H. (2006) *Characterization of essential oil components of Iranian geranium oil using gas chromatography–mass spectrometry combined with chemometric resolution techniques*. J. Chromatogr. A 1114, 154–163.

Jalali-Heravi, M.; Zekavat, B. y Sereshti, H. (2007) *Use of gas chromatography–mass spectrometry combined with resolution methods to characterize the essential oil components of Iranian cumin and caraway*. J. Chromatogr. A 1143, 215–226.

Jaumot, J.; Gargallo, R.; de Juan, A. y Tauler, R. (2005) *A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB*. Chemom. Intell. Lab. Syst. 76, 101–110.

Jjemba, P.K. (2006) *Excretion and ecotoxicity of pharmaceutical and personal care products in the environment*. Ecotoxicol. Environ. Saf. 63, 113–130.

Jonsson, P.; Bruce, S.J.; Moritz, T.; Trygg, J.; Sjöström, M.; Plumb, R.; Granger, J.; Maibaum, E.; Nicholson, J.K.; Holmes, E. y Antti, H. (2005) *Extraction, interpretation and validation of information for comparing samples in metabolic LC/MS data sets*. Analyst 130, 701–707.

Jonsson, P.; Gullberg, J.; Nordström, A.; Kusano, M.; Kowalczyk, M.; Sjöström, M. y Moritz, T. (2004) *A Strategy for Identifying Differences in Large Series of Metabolomic Samples Analyzed by GC/MS*. Anal. Chem. 76, 1738–1745.

Jonsson, P.; Johansson, E.S.; Wuolikainen, A.; Lindberg, J.; Schuppe-Koistinen, I.; Kusano, M.; Sjöström, M.; Trygg, J.; Moritz, T. y Antti, H. (2006) *Predictive Metabolite Profiling Applying Hierarchical Multivariate Curve Resolution to GC−MS DataA Potential Tool for Multi-parametric Diagnosis*. J. Proteome Res. 5, 1407–1414.

Kalivas, J.H. (2001) *Basis sets for multivariate regression*. Anal. Chim. Acta 428, 31–40.

Kalivas, J.H. (2004) *Pareto calibration with built-in wavelength selection*. Anal. Chim. Acta 505, 9–14.

Kalivas, J.H. (2008) *Learning from Procrustes analysis to improve multivariate calibration*. J. Chemom. 22, 227–234.

Kalivas, J.H. y Green, R.L. (2001) *Pareto Optimal Multivariate Calibration for Spectroscopic Data*. Appl. Spectrosc. 55, 1645–1652.

Kalivas, J.H. y Kowalski, B.R. (1982) *Compensation for Drift and Interferences in Multicomponent Analysis*. Anal. Chem. 54, 560–565.

Kalivas, J.H. y Lang, P.M. (1994) *Mathematical Analysis of Spectral Orthogonality.* New York, Estados Unidos

Kennard, R.W. y Stone, L.A. (1969) *Computer Aided Design of Experiments*. Technometrics 11, 137–148.

Kramer, K.E. y Small, G.W. (2007) *Blank Augmentation Protocol for Improving the Robustness of Multivariate Calibrations*. Appl. Spectrosc. 61, 497–506.

Kvalheim, O.M., Liang, Y.Z., 1992. Heuristic evolving latent projections: resolving two-way multicomponent data. 1. Selectivity, latent-projective graph, datascope, local rank, and unique resolution. Anal. Chem. 64, 936–946.

Lawson, C.L. (1995). *Solving Least Square Problems*. Philadelphia, Estados Unidos.

Lawton, W.H. y Sylvestre, E.A. (1971) *Self Modeling Curve Resolution*. Technometrics 13, 617–633.

Leger, M.N. y Wentzell, P.D. (2002) *Dynamic Monte Carlo self-modeling curve resolution method for multicomponent mixtures*. Chemom. Intell. Lab. Syst. 62, 171–188.

Lehotay, S.J.; Mastovská, K. y Lightfield, A.R. (2005) *Use of buffering and other means to improve results of problematic pesticides in a fast and easy method for residue analysis of fruits and vegetables*. J. AOAC Int. 88, 615–629.

Lenz, E.M. y Wilson, I.D. (2007) *Analytical strategies in metabonomics*. J. Proteome Res. 6, 443–458.

Liang, Y. y Kvalheim, O.M. (1994) *Diagnosis and resolution of multiwavelength chromatograms by rank map, orthogonal projections and sequential rank analysis*. Anal. Chim. Acta 292, 5–15.

Liang, Y.Z.; Kvalheim, O.M.; Keller, H.R.; Massart, D.L.; Kiechle, P. y Erni, F. (1992) *Heuristic evolving latent projections: resolving two-way multicomponent data. 2. Detection and resolution of minor constituents*. Anal. Chem. 64, 946–953.

Lilliefors, H.W. (1967) *On the Kolmogorov-Smirnov test for normality with mean and variance unknown*. J. Am. Stat. Assoc. 62, 399–402.

Ling, C.F.; Melian, G.P.; Jiminez-Conde, F. y Revilla, E. (1993) *High-performance liquid chromatographic analysis of carbofuran residues in tomatoes grown in hydroponics*. J. Chromatogr. A 643, 351–355.

Lorenz, E.N. (1956) *Empirical orthogonal functions and statistical weather prediction (No. 1)*, Statistical Forecasting Project. M.I.T., Cambridge, Massachusetts.

Lozano, V.A.; Muñoz de la Peña, A.; Durán-Merás, I.; Espinosa Mansilla, A. y Escandar, G.M. (2013) *Four-way multivariate calibration using ultra-fast high-performance liquid chromatography with fluorescence excitation–emission detection. Application to the direct analysis of chlorophylls a and b and pheophytins a and b in olive oils*. Chemom. Intell. Lab. Syst. 125, 121–131.

Maeder, M. (1987) *Evolving factor analysis for the resolution of overlapping chromatographic peaks*. Anal. Chem. 59, 527–530.

Maeder, M. y Zilian, A. (1988) *Evolving factor analysis, a new multivariate technique in chromatography*. Chemom. Intell. Lab. Syst. 3, 205–213.

Malinowski, E.R. (1982). *Obtaining the key set of typical vectors by factor analysis and subsequent isolation of component spectra*. Anal. Chim. Acta 134, 129–137.

Malinowski, E.R. (1992) *Window factor analysis: Theoretical derivation and application to flow injection analysis data*. J. Chemom. 6, 29–40.

Malinowski, E.R. (2002) *Factor Analysis in Chemistry*, 3rd ed. Wiley-Blackwell, New York.

Mallat, S.G. (1989) *A theory for multiresolution signal decomposition: the wavelet representation*. IEEE Trans. Pattern Anal. Mach. Intell. 11, 674–693.

Manne, R.; Shen, H. y Liang, Y. (1999) *Subwindow factor analysis*. Chemom. Intell. Lab. Syst. 45, 171–176.

Maruo, K.; Oota, T.; Tsurugi, M.; Nakagawa, T.; Arimoto, H.; Hayakawa, M.; Tamura, M.; Ozaki, Y. y Yamada, Y. (2006) *Noninvasive Near-Infrared Blood Glucose Monitoring Using a Calibration Model Built by a Numerical Simulation Method: Trial Application to Patients in an Intensive Care Unit*. Appl. Spectrosc. 60, 1423–1431.

Marx, B.D. y Eilers, P.H.C. (2002) *Multivariate calibration stability: a comparison of methods*. J. Chemom. 16, 129–140.

MATLAB 7.6.0 r2008a (2008) The MathWorks Inc., Natick, Massachusetts.

Mevik, B.H.; Segtnan, V.H. y Næs, T. (2004) *Ensemble methods and partial least squares regression*. J. Chemom. 18, 498–507.

Moco, S.; Vervoort, J.; Moco, S.; Bino, R.J.; De Vos, R.C.H. y Bino, R. (2007) *Metabolomics technologies and metabolite identification*. TrAC Trends Anal. Chem. 26, 855–866.

Muñoz, G. y de Juan, A. (2007) *pH- and time-dependent hemoglobin transitions: A case study for process modelling*. Anal. Chim. Acta 595, 198–208.

Næs, T.; Isaksson, T.; Fern, T. y Davies, T. (2002) *A User Friendly Guide to Multivariate Calibration and Classification*, NIR Publications. Chichester.

Navea, S.; de Juan, A. y Tauler, R. (2001) *Three-way data analysis applied to multispectroscopic monitoring of protein folding*. Anal. Chim. Acta 446, 185–195.

Ni, W.; Brown, S.D. y Man, R. (2009) *Wavelet Orthogonal Signal Correction-Based Discriminant Analysis*. Anal. Chem. 81, 8962–8967.

OMS/FAO (2009) *Pesticide residues in food 2009 - Report of the Joint Meeting of the FAO Panel of Experts on Pesticide Residues in Food and the Environment and the WHO Core Assessment Group on Pesticide Residues*. Geneva, Suiza.

Pacioni, N.L. y Veglia, A.V. (2003) *Determination of carbaryl and carbofuran in fruits and tap water by β-cyclodextrin enhanced fluorimetric method*. Anal. Chim. Acta 488, 193–202.

Páramo, C. (2011) *Singular attractors in speech control,* J. Id. Du.. 69, 1944-2013.

Pearce, J.M. (2012). *Building Research Equipment with Free, Open-Source Hardware*. Science 337, 1303–1304.

Peré-Trepat, E.; Lacorte, S. y Tauler, R. (2007) *Alternative calibration approaches for LC–MS quantitative determination of coeluted compounds in complex environmental mixtures using multivariate curve resolution*. Anal. Chim. Acta 595, 228–237.

Peré-Trepat, E.; Petrovic, M.; Barceló, D. y Tauler, R. (2004) *Application of chemometric methods to the investigation of main microcontaminant sources of endocrine disruptors in coastal and harbour waters and sediments*. Anal. Bioanal. Chem. 378, 642–654.

Peré-Trepat, E. y Tauler, R. (2006) *Analysis of environmental samples by application of multivariate curve resolution on fused high-performance liquid chromatography–diode array detection mass spectrometry data*. J. Chromatogr. A 1131, 85–96.

Perrin, C.; Walczak, B. y Massart, D.L. (2001) *The Use of Wavelets for Signal Denoising in Capillary Electrophoresis*. Anal. Chem. 73, 4903–4917.

Plumb, R.S.; Stumpf, C.L.; Gorenstein, M.V.; Castro-Perez, J.M.; Dear, G.J.; Anthony, M.; Sweatman, B.C.; Connor, S.C. y Haselden, J.N. (2002) *Metabonomics: the use of electrospray mass spectrometry coupled to reversed-phase liquid chromatography shows potential for the screening of rat urine in drug development*. Rapid Commun. Mass Spectrom. RCM 16, 1991–1996.

Riley, M.R.; Arnold, M.A. y Murhammer, D.W. (1998) *Matrix-Enhanced Calibration Procedure for Multivariate Calibration Models with Near-Infrared Spectra*. Appl. Spectrosc. 52, 1339–1347.

Ruckebusch, C.; De Juan, A.; Duponchel, L. y Huvenne, J.P. (2006) *Matrix augmentation for breaking rank-deficiency: A case study.* Chemom. Intell. Lab. Syst. 80, 209–214.

Sáiz-Abajo, M.J.; Mevik, B.H.; Segtnan, V.H. y Næs, T. (2005) *Ensemble methods and data augmentation by noise addition applied to the analysis of spectroscopic data*. Anal. Chim. Acta 533, 147–159.

Sánchez, F.C.; Toft, J.; van den Bogaert, B. y Massart, D.L. (1996) *Orthogonal Projection Approach Applied to Peak Purity Assessment*. Anal. Chem. 68, 79–85.

Savitzky, A. y Golay, M.J.E. (1964) *Smoothing and Differentiation of Data by Simplified Least Squares Procedures*. Anal. Chem. 36, 1627–1639.

Schauer, N. y Fernie, A.R. (2006) *Plant metabolomics: towards biological function and mechanism*. Trends Plant Sci. 11, 508–516.

Shao, X.; Cai, W.; Sun, P.; Zhang, M.; Zhao, G. (1997) *Quantitative Determination of the Components in Overlapping Chromatographic Peaks Using Wavelet Transform*. Anal. Chem. 69, 1722–1725.

Shen, H.; Grung, B.; Kvalheim, O.M. y Eide, I. (2001) *Automated curve resolution applied to data from multi-detection instruments*. Anal. Chim. Acta 446, 311–326.

Shen, H.; Manne, R.; Xu, Q.; Chen, D. y Liang, Y. (1999) *Local resolution of hyphenated chromatographic data*. Chemom. Intell. Lab. Syst. 45, 323–328.

Shih, W.; Bechtel, K.L. y Feld, M.S. (2007) *Constrained Regularization: Hybrid Method for Multivariate Calibration*. Anal. Chem. 79, 234–239.

Smilde, A.K.; Tauler, R.; Henshaw, J.M.; Burgess, L.W. y Kowalski, B.R. (1994) *Multicomponent Determination of Chlorinated Hydrocarbons Using a Reaction-Based Chemical Sensor. 3. Medium-Rank Second-Order Calibration with Restricted Tucker Models*. Anal. Chem. 66, 3345–3351.

Speltini, A.; Sturini, M.; Maraschi, F. y Profumo, A. (2010) *Fluoroquinolone antibiotics in environmental waters: Sample preparation and determination*. J. Sep. Sci. 33, 1115–1131.

Stork, C.L. y Kowalski, B.R. (1999) *Weighting schemes for updating regression models—a theoretical approach*. Chemom. Intell. Lab. Syst. 48, 151–166.

Stout, F.; Kalivas, J.H. y Héberger, K. (2007) *Wavelength Selection for Multivariate Calibration Using Tikhonov Regularization*. Appl. Spectrosc. 61, 85–95.

Stout, F. y Kalivas, J.H. (2006) *Tikhonov regularization in standardized and general form for multivariate calibration with application towards removing unwanted spectral artifacts*. J. Chemom. 20, 22–33.

Sulub, Y. y Small, G.W. (2007) *Spectral Simulation Methodology for Calibration Transfer of Near-Infrared Spectra*. Appl. Spectrosc. 61, 406–413.

Swieranga, H.; Haanstra, W.G.; de Weijer, A.P. y Buydens, L. (1998) *Comparison of Two Different Approaches toward Model Transferability in NIR Spectroscopy.* Appl. Spectrosc. 52, 7–16.

Tamtam, F.; Mercier, F.; Eurin, J.; Chevreuil, M. y Le Bot, B. (2009) *Ultra performance liquid chromatography tandem mass spectrometry performance evaluation for analysis of antibiotics in natural waters*. Anal. Bioanal. Chem. 393, 1709–1718.

Tan, H.W. y Brown, S.D. (2001) *Wavelet hybrid direct standardization of near-infrared multivariate calibrations*. J. Chemom. 15, 647–663.

Tauler, R.; Izquierdo-Ridorsa, A. y Casassas, E. (1993a) *Simultaneous analysis of several spectroscopic titrations with self-modelling curve resolution*. Chemom. Intell. Lab. Syst. 18, 293–300.

Tauler, R.; Kowalski, B. y Fleming, S. (1993b) *Multivariate curve resolution applied to spectral data from multiple runs of an industrial process*. Anal. Chem. 65, 2040–2047.

Tauler, R.; Marqués, I. y Casassas, E. (1998) *Multivariate curve resolution applied to three-way trilinear data: Study of a spectrofluorimetric acid–base titration of salicylic acid at three excitation wavelengths*. J. Chemom. 12, 55–75.

Tibshirani, R. (1996) *Regression Shrinkage and Selection via the Lasso*. J. R. Stat. Soc. Ser. B Methodol. 58, 267–288.

Tikhonov, A.N. (1943) *On the stability of inverse problems*. Dokl Akad Nauk SSSR 39, 195–198.

Tikhonov, A.N. (1963) *Solution of incorrectly formulated problems and the regularization method*. Sov. Math Dokl 4, 1035–1038.

Trygg, J. y Wold, S. (1998) *PLS regression on wavelet compressed NIR spectra*. Chemom. Intell. Lab. Syst. 42, 209–220.

Vademécum de Variedades Hortícolas. Portagranos 2005-2006 (2005) Escobar Impresiones, Almería, España.

Vandeginste, B.; Essers, R.; Bosman, T.; Reijnen, J. y Kateman, G. (1985) *Three-component curve resolution in liquid chromatography with multiwavelength diode array detection*. Anal. Chem. 57, 971–985.

Vázquez, M.M.P.; Vázquez, P.P.; Galera, M.M. y García, M.D.G (2012) *Determination of eight fluoroquinolones in groundwater samples with ultrasound-assisted ionic liquid dispersive liquid–liquid microextraction prior to high-performance liquid chromatography and fluorescence detection*. Anal. Chim. Acta 748, 20–27.

Viant, M.R. (2008) *Recent developments in environmental metabolomics*. Mol. Biosyst. 4, 980–986.

Vogt, F.; Rebstock, K. y Tacke, M. (2000) *Correction of background drifts in optical spectra by means of "pseudo principal components"*. Chemom. Intell. Lab. Syst. 50, 175–178.

Vogt, F.; Steiner, H.; Booksh, K. y Mizaikof, B. (2004) *Chemometric Correction of Drift Effects in Optical Spectra*. Appl. Spectrosc. 58, 683–692.

Walczak, B. (2000), *Wavelets in Chemistry*, Volume 22, 1st ed. Elsevier Science.

Walczak, B. y Massart, D.L. (1997a) *Wavelet packet transform applied to a set of signals: A new approach to the best-basis selection*. Chemom. Intell. Lab. Syst. 38, 39–50.

Walczak, B. y Massart, D.L. (1997b) *Wavelets — something for analytical chemistry?* TrAC Trends Anal. Chem. 16, 451–463.

Wang, Y.; Veltkamp, D.J. y Kowalski, B.R. (1991) *Multivariate instrument standardization*. Anal. Chem. 63, 2750–2756.

Wang, Z.; Dean, T. y Kowalski, B. (1995) *Additive Background Correction in Multivariate Instrument Standardization*. Anal. Chem. 67, 2379–2385.

Wehlburg, C.M.; Haaland, D.M. y Melgaard, D.K. (2002a). *New Hybrid Algorithm for Transferring Multivariate Quantitative Calibrations of Intra-vendor Near-Infrared Spectrometers*. Appl. Spectrosc. 56, 877–886.

Wehlburg, C.M.; Haaland, D.M.; Melgaard, D.K. y Martin, L.E. (2002b) *New Hybrid Algorithm for Maintaining Multivariate Quantitative Calibrations of a Near-Infrared Spectrometer.* Appl. Spectrosc. 56, 605–614.

Westerhaus, M.O. (1991) *Improving Repeatability of NIR Calibrations Across Instruments*, Proceedings of the Third International Near Infrared Spectroscopy Conference. Gembloux, Bélgica.

Westerhuis, J.A.; Hoefsloot, H.C.J.; Smit, S.; Vis, D.J.; Smilde, A.K.; van Velzen, E.J.J.; van Duijnhoven, J.P.M. y van Dorsten, F.A. (2008) *Assessment of PLSDA cross validation*. Metabolomics 4, 81–89.

Windig, W. y Guilment, J. (1991) *Interactive self-modeling mixture analysis*. Anal. Chem. 63, 1425–1432.

Wise, B.M.; Gallagher, N.B.; Bro, R.; Shaver, J.M.; Windig, W. y Koch, R.S. (2005) *PLS Toolbox 3.52 for use with MATLAB*.

Wold, H. (1966) *Estimation of Principal Components and Related Models by Iterative Least squares, in: Multivariate Analysis*. Academic Press, New York, pp. 391–420.

Wold, S.; Ruhe, A.; Wold, H. y Dunn, W.J. (1984) *The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses*. SIAM J. Sci. Stat. Comput. 5, 735–743.

Wu, W. (1998) *ChemoAC Toolbox version 2.0*, FABI. Bruselas, Bélgica.

Wülfert, F.; Kok, W.T.; de Noord, O.E. y Smilde, A.K. (2000a) *Linear techniques to correct for temperature-induced spectral variation in multivariate calibration*. Chemom. Intell. Lab. Syst. 51, 189–200.

Wülfert, F.; Kok, W.T.; de Noord, O.E. y Smilde, A.K. (2000b) *Correction of temperature-induced spectral variation by continuous piecewise direct standardization*. Anal. Chem. 72, 1639–1644.

Wulfert, F.; Kok, W.T. y Smilde, A.K. (1998) *Influence of Temperature on Vibrational Spectra and Consequences for the Predictive Ability of Multivariate Models*. Anal. Chem. 70, 1761–1767.

Xu, C.-J.; Jiang, J.-H. y Liang, Y.-Z. (1999) *Evolving window orthogonal projections method for two-way data resolution*. Analyst 124, 1471–1476.

Zachariassen, C.B.; Larsen, J.; van den Berg, F.; Bro, R.; de Juan, A. y Tauler, R. (2006) *Comparison of PARAFAC2 and MCR-ALS for resolution of an analytical liquid dilution system*. Chemom. Intell. Lab. Syst. 83, 13–25.

Zhu, M. (2008) *Kernels and Ensembles*. Am. Stat. 62, 97–109.