Revista de la Societat Catalana de Química, núm. **10** (2011), p. 63-70
Filial de l'Institut d'Estudis Catalans
DOI: 10.2436/20.2003.01.30 ISSN: 2013-9853 (ed. electrònica) ISSN: 1576-8961 (ed. impresa)
http://revistes.iec.cat/index.php/RSCQ

Estratègies d'anàlisi de dades en quimiometria: selecció *versus* compressió
Data analysis in chemometrics: selection versus *compression*

Alberto Ferrer
Universitat Politècnica de València. Departament d'Estadística i Investigació Operativa Aplicada i Qualitat

Resum. La quimiometria utilitza eines de mineria de dades per a la modelització empírica de sistemes (bio)químics. El desenvolupament explosiu de les tecnologies de la informació i de les comunicacions ha possibilitat la fabricació d'una gran varietat de sensors que són capaços de recollir grans quantitats de dades i emmagatzemar-les en dispositius informàtics. El repte està a extreure eficientment la informació potencial continguda en aquestes dades, la qual cosa depèn en gran mesura de l'estratègia d'anàlisi usada. Amb tanta quantitat de dades disponible, cal usar algun procediment de reducció del nombre de variables a analitzar. En aquest article es presenten dues estratègies per a aquesta necessària simplificació: compressió *versus* selecció. La gran diferència entre ambdues és que en seleccionar, s'eliminen algunes de les variables mesurades, mentre que en comprimir, no. Si la selecció es realitza al principi de la investigació, es corre el risc d'eliminar variables amb informació útil per resoldre el problema en qüestió. La recomanació és, per tant, comprimir i, si és necessari, seleccionar amb posterioritat. Els beneficis d'aquesta recomanació s'il·lustren amb diversos exemples reals.

Paraules clau: Quimiometria, estructures latents, anàlisi de components principals (PCA), anàlisi discriminant amb mínims quadrats parcials (PLS-DA), compressió, selecció, mineria de dades, model de calibratge, diagnòstic multivariant de processos.

Abstract. Chemometrics uses data mining tools for empirical modeling of biochemical systems. The explosive development of information and communications technology have enabled the manufacture of a wide variety of sensors that are able to register large amounts of data stored on computing devices. The challenge is to efficiently extract the potential information contained in the data, which depends heavily on the strategy of analysis used. With so much data available it is necessary to use a procedure to reduce the number of variables to analyze. In this paper we present two strategies for this necessary simplification: compression* versus *selection. The big difference between them is that with selection some variables are discarded whereas after compression all variables may be recovered. If the selection is made at the beginning of the investigation there is a risk of eliminating variables with useful information to solve the problem at hand. The recommendation is therefore compress and, if it is needed, select. The benefits of this recommendation are illustrated with actual examples.

Keywords: Chemometrics, latent structures, principal component analysis (PCA), partial least squares discriminant analysis (PLS-DA), compression, selection, data mining, soft sensor, multivariate process diagnosis.

Introducció: Enric Casassas i el plaer de les xifres

Abans d'acceptar la invitació a participar en aquest memorial, no sabia res sobre el professor Casassas. No havia tingut l'oportunitat de conèixer-lo en persona ni d'haver llegit algun dels seus treballs, potser perquè la meva aproximació a la quimiometria no és des del món de la química analítica, sinó des de l'enginyeria química. No obstant això, he de reconèixer que, després de llegir algun dels seus treballs i comentaris sobre la seva vida i obra, he descobert una persona inquieta per resoldre nous problemes, oberta a aprendre dels altres, disposada a col·laborar amb d'altres i compromesa amb la formació i la divulgació del seu coneixement, aspectes clau per a un veritable científic.

Enric Casassas es pot considerar un dels principals introductors de la quimiometria a Catalunya. La va aplicar fonamentalment a la química analítica. Va utilitzar tècniques de reconeixement de patrons i d'anàlisi de conglomerats per analitzar mostres d'interès arqueològic i d'altres camps; també va desenvolupar l'aplicació de l'anàlisi factorial i de l'anàlisi de components principals en la interpretació dels resultats experimentals obtinguts en estudis d'equilibris en solució sense la necessitat de postular models químics teòrics previs, els quals sorgeixen de la mateixa anàlisi numèrica.[1]

Correspondència: Alberto Ferrer. Universitat Politècnica de València. Departament d'Estadística i Investigació Operativa Aplicada i Qualitat
Campus de Vera. Camí de Vera, s/n. Edifici 7A. 46022 València
Tel.: +34 963 877 007. Fax: +34 963 877 499
A. e.: *aferrer@eio.upv.es*

En el seu llibre *Del plaer dels sentits al plaer de les xifres*,[2] el professor Casassas ens descobreix la seva passió per la quimiometria. La seva visió d'aquesta disciplina com de «l'art d'extreure informació rellevant des del punt de vista químic de les dades produïdes en els experiments químics quantitatius» és molt semblant a la primera definició de la quimiometria establerta per Svante Wold l'any 1974:[3] «La ciència d'extreure informació de sistemes químics mitjançant models empírics (basats en dades)».

La quimiometria està molt lligada a la incorporació i al desenvolupament de noves tecnologies per al tractament de la informació en àmbits tan rellevants com la química, la bioquímica, la modelització molecular i les tècniques QSAR (*quantitative structure-activity relationships*), la quimioinformàtica, les ciències –*òmiques* (genòmica, proteòmica i metabonòmica), l'estudi global del medi ambient, la modelització de processos o la tecnologia analítica de processos (PAT, *process analytical technology*).

El desenvolupament explosiu de la instrumentació científico-tècnica, de l'automatització i de la informàtica ha permès que en tots aquests àmbits es puguin enregistrar grans quantitats de dades en poc temps, tot generant-se complexes estructures de dades de diversa naturalesa que contenen informació sobre diferents tipus de propietats de les mostres analitzades. Alguns exemples d'aquesta estructura de dades variada són els següents: espectres NIR, UV/visible o de masses, o cromatogrames de mostres de substàncies (relacionats amb propietats químiques); pressions, temperatures i cabals de processos químics (relacionats amb propietats físiques); pH, conductivitats i concentracions d'oxigen en el tractament d'aigües residuals (lligats a paràmetres bioquímics), o bé perfils d'expressió gènica (relacionats amb propietats genètiques de sistemes cel·lulars). Durant l'última dècada, el desenvolupament de la tecnologia dels sensors electrònics (*i-sensing*) ha permès la fabricació de «nassos i llengües electrònics», que proporcionen senyals elèctrics en reaccionar amb els composts de la mostra analitzada. En certs processos, com els de la indústria de l'acer, on no és possible la mesura directa amb sondes ateses les elevades temperatures del procés, s'usa l'«oïda» electrònica, que transforma els senyals acústics en elèctrics, o la termografia, que permet captar variacions de temperatura al forn de fusió. També l'avenç en la tecnologia digital ha permès fer

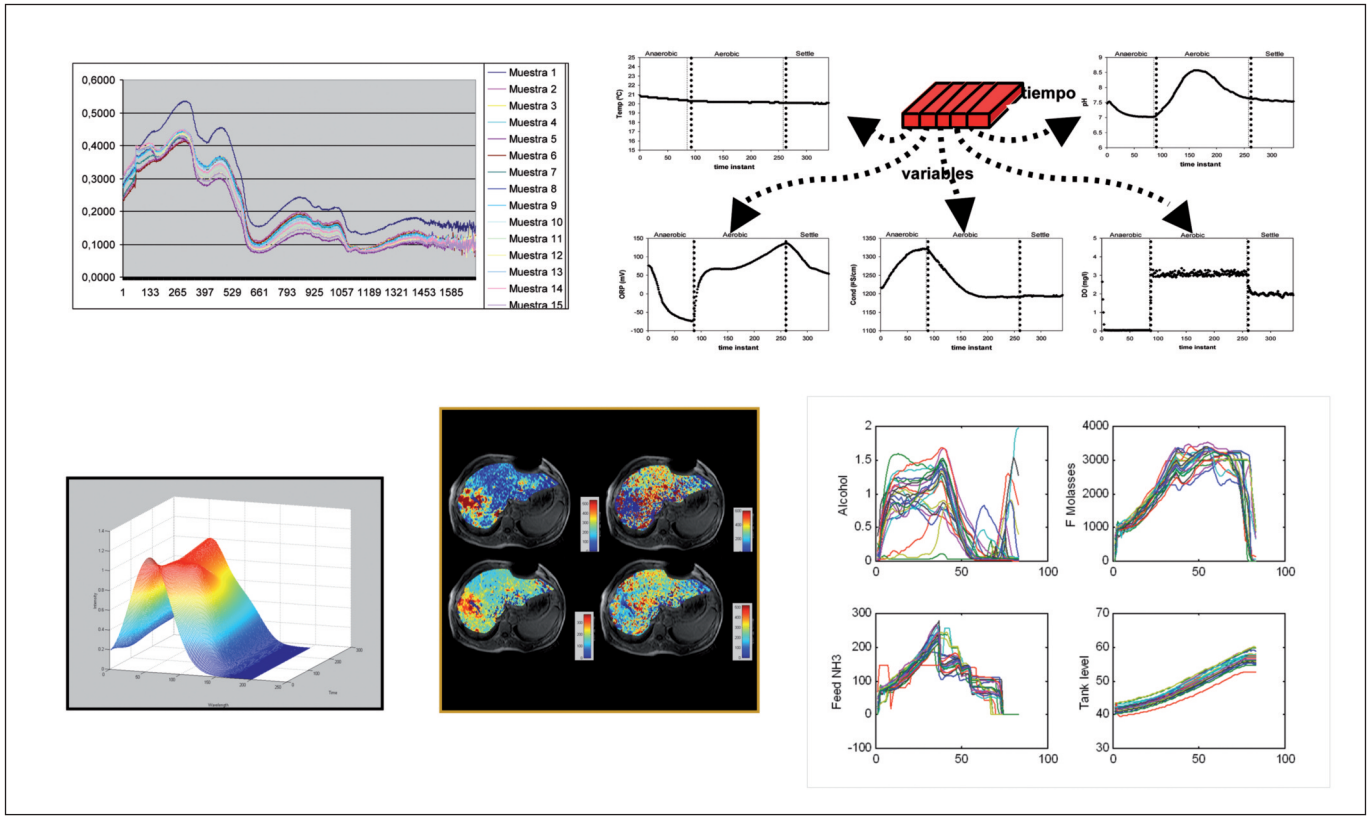

FIGURA 1. Exemples de tipus de dades de diversa naturalesa.

viable a la pràctica l'ús de les càmeres espectrals o hiperespectrals, que proporcionen, per a cada un dels píxels d'una imatge, informació espectral a diferents longituds d'ona. En l'àmbit del radiodiagnòstic mèdic, també són molt usades les imatges de ressonància magnètica nuclear. La figura 1 mostra alguns exemples d'aquestes complexes estructures de dades. La naturalesa multivariant d'aquestes estructures de dades, el seu alt grau de colinealitat, conseqüència de la complexa xarxa de relacions entre variables, i l'existència de valors absents (no enregistrats) en algunes variables constitueixen un verdader repte a l'hora d'extreure eficientment la informació potencial continguda a les esmentades dades. Per tal d'aconseguir-ho, les dades (xifres) no ens han de desconcertar, sinó despertar-nos passió pel tresor (informació) que tanquen. El seu descobriment depèn en gran mesura de l'estratègia d'anàlisi usada, tal com es comenta a l'apartat següent.

Compressió o selecció?

En la tasca d'extreure la informació rellevant, la quimiometria necessita utilitzar mètodes estadístics d'anàlisi de dades. Entre els més coneguts i estudiats, que denominarem *clàssics*, hi ha els mètodes de regressió lineal, l'anàlisi de la variància univariant i multivariant, l'anàlisi de conglomerats o l'anàlisi discriminant de Fisher. El problema d'aquests mètodes clàssics és que van ser desenvolupats en contextos d'escassetat de dades, en els quals s'enregistraven poques variables, poc relacionades, i en els quals el nombre de variables era habitualment molt inferior al d'individus o mostres. De fet, en presència de fortes relacions entre variables o de situacions amb més variables que mostres enregistrades, les tècniques clàssiques són molt ineficients o, fins i tot, inviables.

Tal com ja s'ha comentat, vivim en un nou entorn que ha modificat totalment la naturalesa de les dades disponibles, la qual cosa obliga a un canvi de paradigma: de l'escassetat a la sobreabundància de dades. Davant d'aquesta sobrecàrrega de dades, molts analistes i tècnics se senten desconcertats i, per poder usar les tècniques estadístiques clàssiques, acaben seleccionant *a priori* per a la presa de decisions sols unes quantes de les variables enregistrades atenent criteris subjectius, com ara l'ús de l'experiència prèvia, no comprovada de manera científica. El problema d'aquesta estratègia de selecció és que pot arribar a descartar variables potencialment útils d'una manera fins a un cert punt cega i, per tant, arriscada, la qual cosa pot dificultar enormement la comprensió del problema estudiat. El professor Casassas era molt conscient d'aquest risc i no va dubtar a qüestionar la utilitat de les tècniques estadístiques clàssiques a l'hora d'analitzar aquests nous tipus de dades, pròpies de la quimiometria moderna.[2]

Afortunadament, seleccionar no és l'única estratègia possible i la presència d'estructures complexes de relació entre variables no és realment un problema, sinó la constatació que els processos estan governats per unes quantes estructures no mesurables directament (latents) que afecten les variables que sí que es poden enregistrar (observables). Una cosa semblant li ocorre a una marioneta, en la qual els moviments de les diferents parts del cos (variables observables) no són autònoms, sinó que estan governats a través dels fils o cordes de connexió amb els moviments de les varetes (variables latents), mogudes per les mans del titellaire. La relació entre els moviments de les diferents parts del cos depèn de l'estructura de la marioneta, és a dir, de la connexió de les seves articulacions a les varetes. De la mateixa manera que per poder comprendre el funcionament d'una marioneta cal conèixer-ne l'estructura, per poder comprendre els processos cal descobrir les estructures latents que els governen. Aquest és precisament l'objecte de les tècniques estadístiques multivariants de projecció sobre estructures latents com l'anàlisi de components principals (PCA, *principal component analysis*)[4] i la regressió en mínims quadrats parcials (PLS, *partial least squares*).[5] Aquestes tècniques manegen bé grans matrius de dades mal condicionades (fins i tot en el cas d'existir més variables que individus), són relativament robustes a la presència de dades absents i comprimeixen la informació multidimensional en unes quantes variables latents que expliquen una gran part de la variabilitat de les variables mesurades, així com de les seves relacions, per la qual cosa es poden considerar l'alternativa eficient a l'estratègia de la selecció en els contextos de sobreabundància de dades.

La comprensió dels problemes (i dels processos involucrats) exigeix compressió, no selecció d'informació, almenys *a priori*. En comprimir, l'analista pot analitzar el seu procés en l'espai de les variables latents (de menor dimensió i, en molts casos, ortogonal), la qual cosa el pot ajudar a entendre millor els fenòmens químics, físics, bioquímics, etc., subjacents i fins i tot a usar les eines estadístiques clàssiques. La compressió estableix vincles de relació amb les variables originals, que en qualsevol moment es poden recuperar en l'anàlisi, per la qual cosa les variables originals no es perden, com ocorre amb

aquelles variables descartades en l'estratègia de selecció. Una vegada comprès (comprimit) el procés, en certes aplicacions, com la construcció de models predictius (*soft sensors*) o l'establiment d'esquemes de control estadístic multivariant de processos, pot ser recomanable realitzar una selecció *a posteriori* de les variables originals. Això s'il·lustra a l'apartat següent amb diversos exemples.

Exemple 1. Construcció d'un model de calibratge o predictor (soft sensor)

Es disposa de l'espectre d'absorbància a 1.701 longituds d'ona de quinze mostres d'un aliment (figura 2), així com de valors analítics de dos dels seus paràmetres de qualitat, determinats en laboratori. Es pretén construir un model de calibratge (predictiu) multivariant per poder avaluar indirectament la qualitat de futures mostres de l'aliment a partir del seu espectre d'absorbància, sense la necessitat de recórrer als assaigs analítics de laboratori.

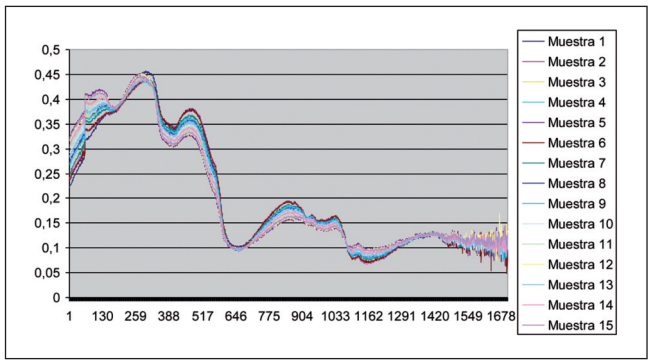

FIGURA 2. Espectres d'absorbància de les quinze mostres analitzades.

La tècnica d'estadística clàssica a usar en aquest cas podria ser la regressió lineal múltiple, però, com que només es disposa de quinze mostres, el model predictiu tan sols podria contenir com a variables predictores, com a màxim, les absorbàncies a 14 longituds d'ona. Això obligaria a seleccionar *a priori* 14 de les 1.701 absorbàncies registrades a l'espectre. Si es fes aquesta selecció de manera automàtica, això donaria lloc a múltiples models. Quin seria el «millor»? El fet que una longitud d'ona no s'elegís com a variable predictora s'hauria d'entendre com que l'absorbància a l'esmentada longitud d'ona no té relació amb les característiques de qualitat? Si s'apliqués el coneixement químic per seleccionar les longituds d'ona que poguessin estar *a priori* més correlacionades amb les característiques de qualitat, com es podria validar científicament l'esmentat coneixement previ? Es podria estar raonablement confiat que no s'haguessin descartat longituds d'ona «informatives»?

Amb l'estratègia de compressió, el problema es resol d'una manera molt simple. Aplicant, en aquest cas, la tècnica PLS, s'obté un model que, amb una única variable latent, té una capacitat predictiva d'aproximadament un 95 %. La figura 3 mostra el grau de relació de les variables predictores X (absorbàncies a 1.701 longituds d'ona) amb les dues variables de qualitat Y. D'aquest gràfic es dedueix que les dues variables de qualitat tenen una forta relació inversa (correlació negativa) i que existeixen zones de l'espectre molt relacionades i d'altres amb molt soroll o poc relacionades amb les característiques de qualitat. Aquesta informació podria ser interessant de comprovar-la amb el coneixement teòric de la relació entre l'absorbància a determinades longituds d'ona i l'estructura química de l'aliment.

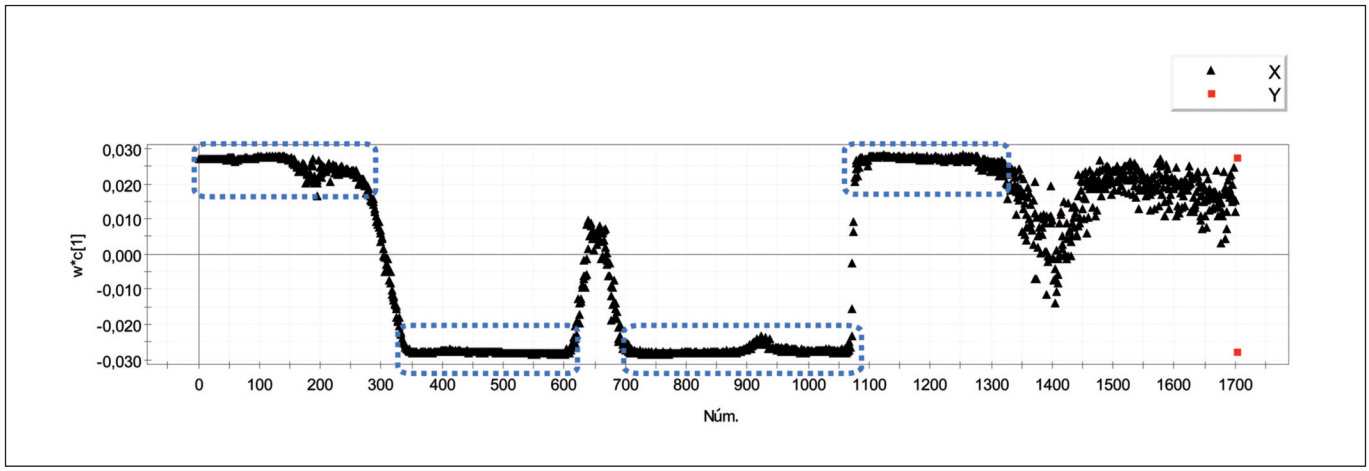

FIGURA 3. Relació de les variables predictores X (absorbàncies) amb les dues variables de qualitat Y en el model PLS original. Es marquen les variables predictores que tenen una relació més forta i consistent amb les dues variables resposta.

En aquest cas, per tal de millorar la qualitat (consistència) de les prediccions, és recomanable seleccionar aquelles zones de l'espectre amb una correlació més forta i consistent (assenyalades a la figura 3) i construir un nou model PLS usant com a predictora aquesta selecció *a posteriori* de variables. La figura 4 mostra les prediccions d'una de les variables de qualitat obtingudes amb aquest nou model usant una part de l'espectre complet original (1.197 de les 1.701 longituds d'ona). La capacitat predictiva continua sent d'aproximadament un 95 %. Altres models més senzills (parsimoniosos) també es podrien derivar a partir dels anteriors.

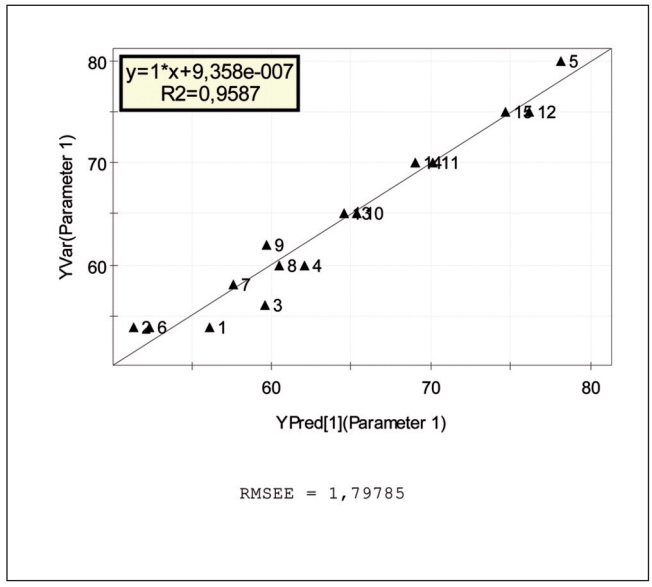

FIGURA 4. Prediccions del model PLS d'una característica de qualitat quan s'utilitza solament una part de l'espectre complet original (1.197 de les 1.701 longituds d'ona).

Exemple 2. Diagnòstic d'un procés químic

El procés objecte d'estudi consisteix en la fabricació d'un alcohol mitjançant la hidrogenació selectiva d'una mescla d'èsters. L'objectiu és estudiar la consistència del procés en diferents campanyes de fabricació. Es disposa de dades de setanta-dues variables de procés mesurades cada hora en diferents unitats del mateix procés, així com de dades d'una variable de rendiment del procés, mesurada cada vuit hores durant dues campanyes de fabricació. La base de dades conté uns cinc mil registres per a les variables de procés i uns cinccents per a la de rendiment durant aproximadament vuit mesos de producció.

La figura 5 mostra l'evolució del rendiment del procés durant les dues campanyes. S'hi observa que, encara que l'engegada d'ambdues campanyes no és similar, una vegada estabilitzat el procés, els valors del rendiment en ambdues segueixen una evolució similar, de manera que no existeixen diferències estadísticament significatives entre els seus valors mitjans (risc de primera espècie = 0,05).

Es pot concloure, per tant, que ambdues campanyes han estat processades en les mateixes condicions i que el producte obtingut és bàsicament el mateix? La resposta és clarament negativa: un mateix rendiment no significa necessàriament un mateix producte ni un mateix procés. Una única característica

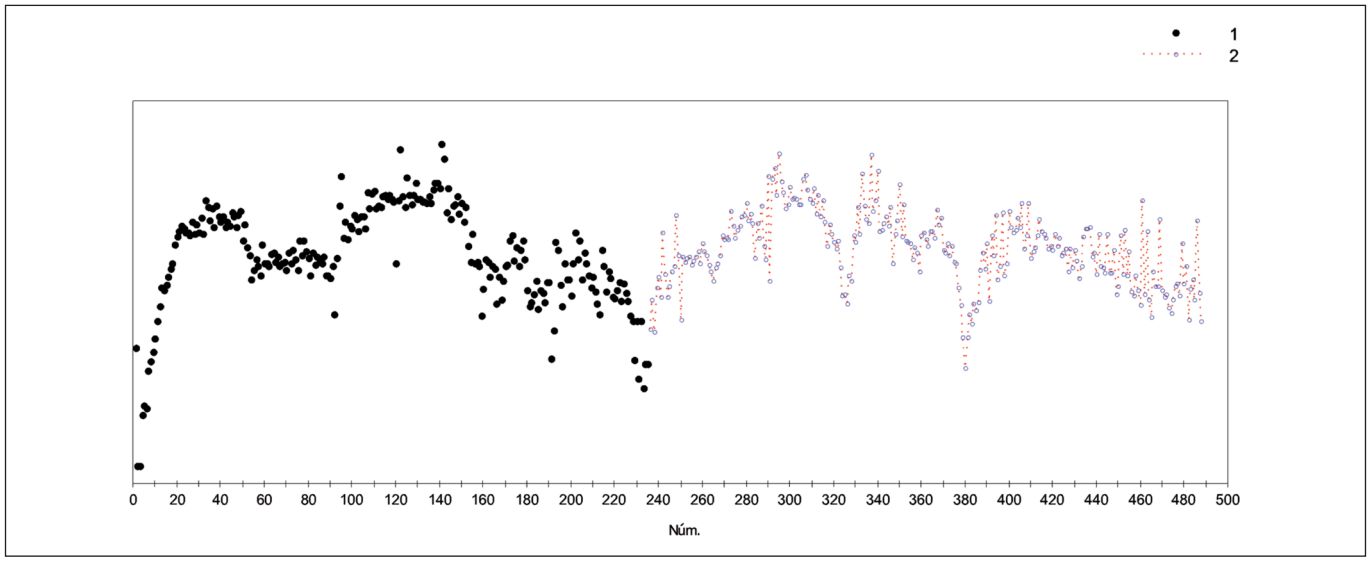

FIGURA 5. Rendiment del procés en les dues campanyes de fabricació.

no defineix completament la «qualitat» del producte acabat. La qualitat és un concepte multidimensional que es defineix en relació amb la manera com el producte satisfà les expectatives del client. Ja que és inviable el fet de provar de caracteritzar completament la qualitat mitjançant múltiples anàlisis del producte acabat, l'alternativa efectiva és provar d'avaluar la consistència en el procés de fabricació. Al cap i a la fi, la forma en la qual el producte ha estat processat constitueix una «empremta dactilar» que es pot utilitzar per garantir la consistència del producte obtingut.

Per respondre les qüestions plantejades, es va procedir a analitzar les dades de les setanta-dues variables del procés durant les dues campanyes de fabricació mitjançant un model PLS discriminant. La figura 6 mostra el diagrama de dispersió dels *scores* (*scores plot*) de les dues campanyes. En aquest cas, dues variables latents són capaces de resumir la informació de les setanta-dues variables del procés amb una bondat de predicció del 94 %. Si l'operativa del procés en les variables analitzades hagués estat la mateixa, els núvols de *scores* d'ambdues campanyes estarien superposats, la qual cosa no ocorre, en aquest cas. La figura 6 indica que ambdues campanyes han estat processades en condicions diferents d'algunes variables del procés i que la campanya 1 és més estable (el núvol de *scores* està més concentrat) que la campanya 2.

Un estudi detallat de la figura 7, en la qual es mostren els coeficients del model PLS que prediu la campanya 1, permet identificar clarament les variables del procés que tenen un comportament diferent en ambdues campanyes i separar-les en dos grups: les de coeficients grans i positius són les que

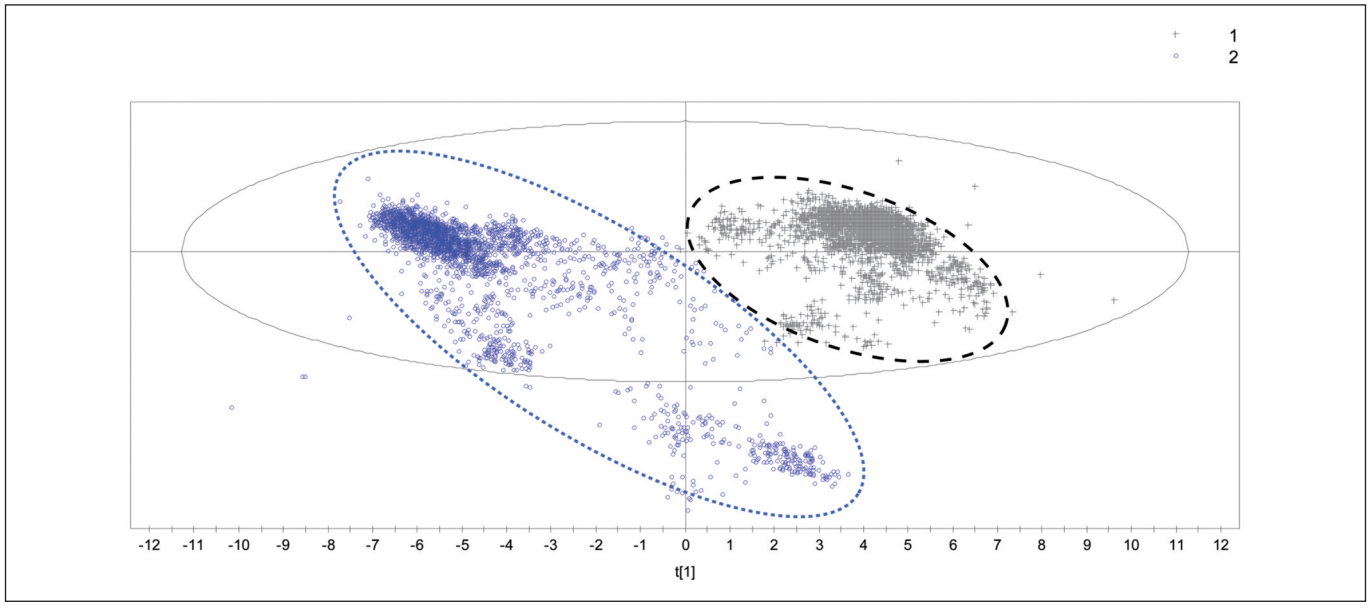

FIGURA 6. *Score plot* de les dues campanyes en el model PLS discriminant.

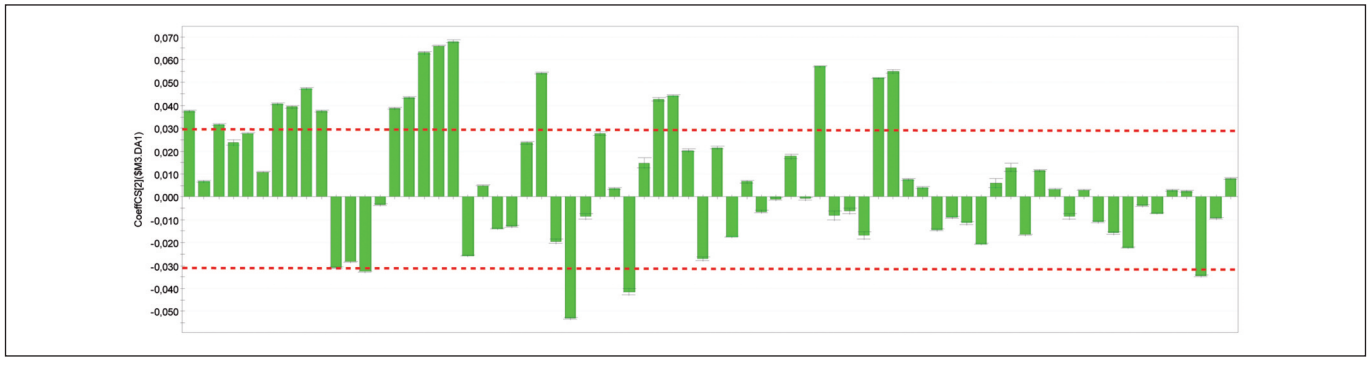

FIGURA 7. Coeficients (*loadings*) de les setanta-dues variables predictores del model PLS de predicció de la campanya 1.

prenen uns valors majors a la campanya 1 respecte de la campanya 2; les de coeficients grans i negatius tenen el comportament contrari, de manera que prenen uns valors majors a la campanya 2 respecte de la campanya 1. Les figures 8 i 9 il·lustren un exemple de cada tipus.

La informació proporcionada per la figura 7 va permetre als tècnics del procés confirmar que la segona campanya, a més de ser més inestable, va tenir un major cost: es va realitzar a major temperatura i amb un major aportament de catalitzador.

La comprensió de les diferències del procés en ambdues campanyes facilitada pel model compressor PLS discriminant ha servit, en aquest cas, per convèncer els tècnics del procés de la necessitat de fer el seguiment no solament del rendiment, sinó també de l'operativa del procés, per garantir un producte rendible de manera consistent. En aquests moments s'està estudiant la possibilitat de desenvolupar un sistema de control estadístic multivariant del procés per provar d'aconseguir operar campanyes futures d'una manera similar a la campanya 1, més estable i de menor cost.

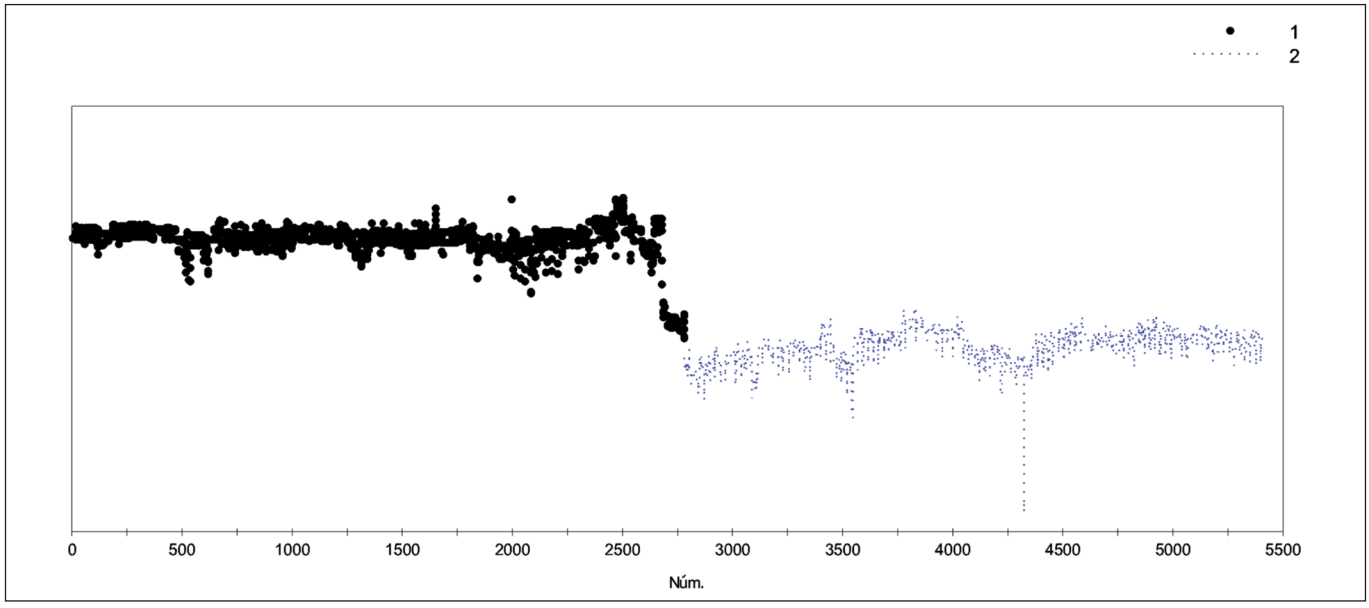

Figura 8. Evolució d'una de les variables del procés que pren uns valors superiors a la campanya 1.

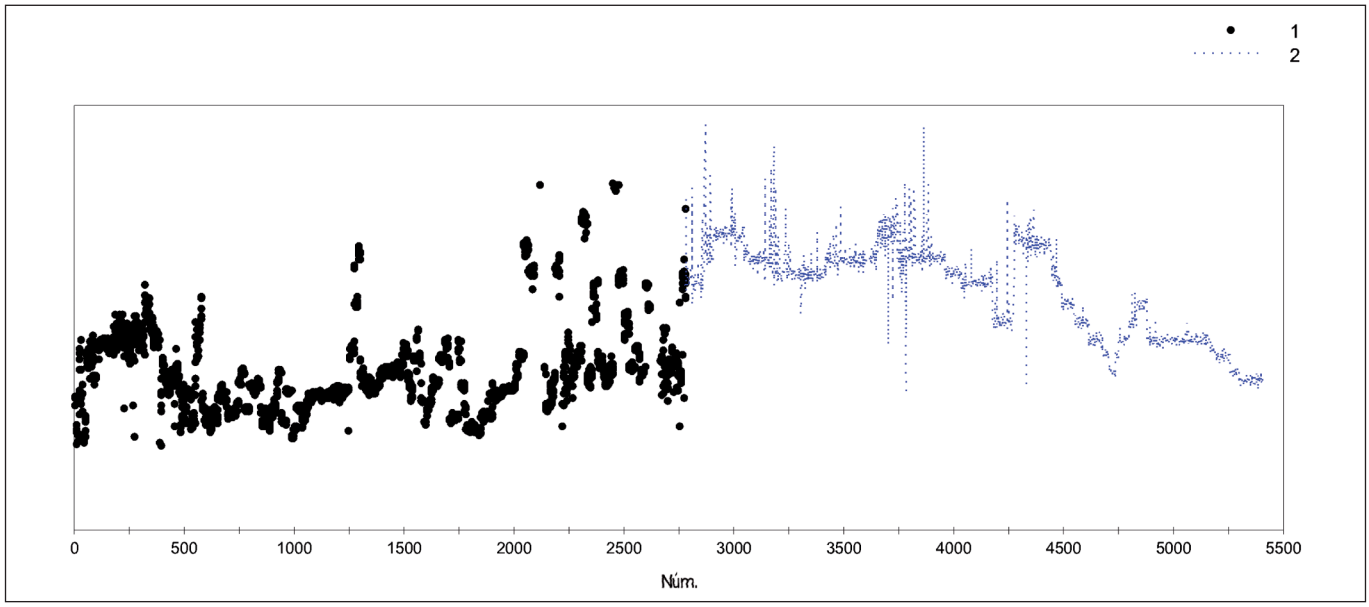

Figura 9. Evolució d'una de les variables del procés que pren uns valors superiors a la campanya 2.

Conclusions

De la mateixa manera que no seleccionem *a priori* les fotos d'un viatge abans d'enviar-les a un amic per no saturar la capacitat màxima del missatge al correu electrònic, sinó que utilitzem programes de compressió que permetin al nostre destinatari, una vegada rebudes totes les fotos comprimides, gaudir del viatge i seleccionar *a posteriori* les que més li agradin, tampoc no és adequat seleccionar *a priori* les variables a utilitzar en les nostres anàlisis de problemes quimiomètrics, malgrat que disposem, en molts casos, de grans bases de dades per analitzar. En seleccionar variables, es corre el risc de perdre informació útil. La comprensió dels processos exigeix no la selecció, sinó la compressió de la informació. Els mètodes estadístics de projecció multivariant, com el PCA (anàlisi de components principals) o el PLS (regressió en mínims quadrats parcials), constitueixen unes excel·lents eines per a aquest propòsit. Una vegada els fenòmens estudiats han estat sotmesos a compressió, tot utilitzant la interacció entre el coneixement tècnic disponible i la informació obtinguda de l'anàlisi multivariant de les dades, si cal, es pot procedir a la selecció *a posteriori* de les variables més rellevants per a l'objecte de l'estudi.

Agraïments

És un plaer per a mi haver pogut participar en el IX Memorial Enric Casassas (2 de desembre de 2009), per la qual cosa vull mostrar el meu agraïment als organitzadors d'aquest esdeveniment, els professors Romà Tauler, Xavier Tomàs i Anna de Juan, per la seva amable invitació.

Referències

[1] Casassas, E. *Sessió en memòria*. Institut d'Estudis Catalans: Barcelona, **2000**.
[2] Casassas, E. «Del plaer dels sentits al plaer de les xifres o de l'alquímia a la quimiometria». A: *Del plaer dels sentits al plaer de les xifres*. Casassas, E.; Simó, E.; Tauler, R. (ed.). Universitat d'Estiu; Institut d'Estudis Catalans: Barcelona, **1997**, p. 7. (Monografies de les Seccions de Ciències; 13).
[3] Wold, S. «Chemometrics: What do we mean with it and what do we want from it». *Chemometrics and Intelligent Laboratory Systems* **1995**, *30*, 109.
[4] Wold, S.; Esbensen, K.; Geladi, P. «Principal component analysis». *Chemometrics and Intelligent Laboratory Systems* **1987**, *2*, 37.
[5] Martens, H.; Naes, T. *Multivariate calibration*. John Wiley & Sons: Nova York, **1992**.

A. Ferrer

Alberto Ferrer és enginyer agrònom i doctor per la Universitat Politècnica de València. Actualment, és catedràtic del Departament d'Estadística i Investigació Operativa Aplicades i Qualitat de la Universitat Politècnica de València, on dirigeix el grup d'investigació en Enginyeria Estadística Multivariant, dedicat al desenvolupament de metodologia estadística per a l'anàlisi, el monitoratge i el diagnòstic de processos complexos. És coordinador del programa de doctorat «Estadística i optimització» de la Universitat Politècnica de València i ha estat editor associat de la revista *Technometrics* (2008-2009). En l'actualitat, és membre del Consell de la International Society for Business and Industrial Statistics (ISBIS), així com de l'European Network for Business and Industrial Statistics (ENBIS) i de la Xarxa Espanyola de Quimiometria.