# Técnicas para el desarrollo de modelos predictivos de cuantificación a través de espectroscopía NIR

## Diseño de un modelo de cuantificación por calibración multivariada


Dr. Mario Sanhueza Garcias
Centro de Biotecnología
Universidad de Concepción
*msanhuezag@udec.cl*

1

Desarrollo de secuencia para el desarrollo y validación de modelos PLS y PCR predictivos por espectroscopía NIR.

# 1 Organización de los datos

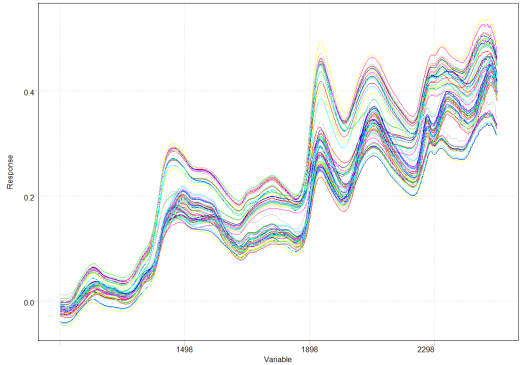
Longitudes/ número de onda

muestras

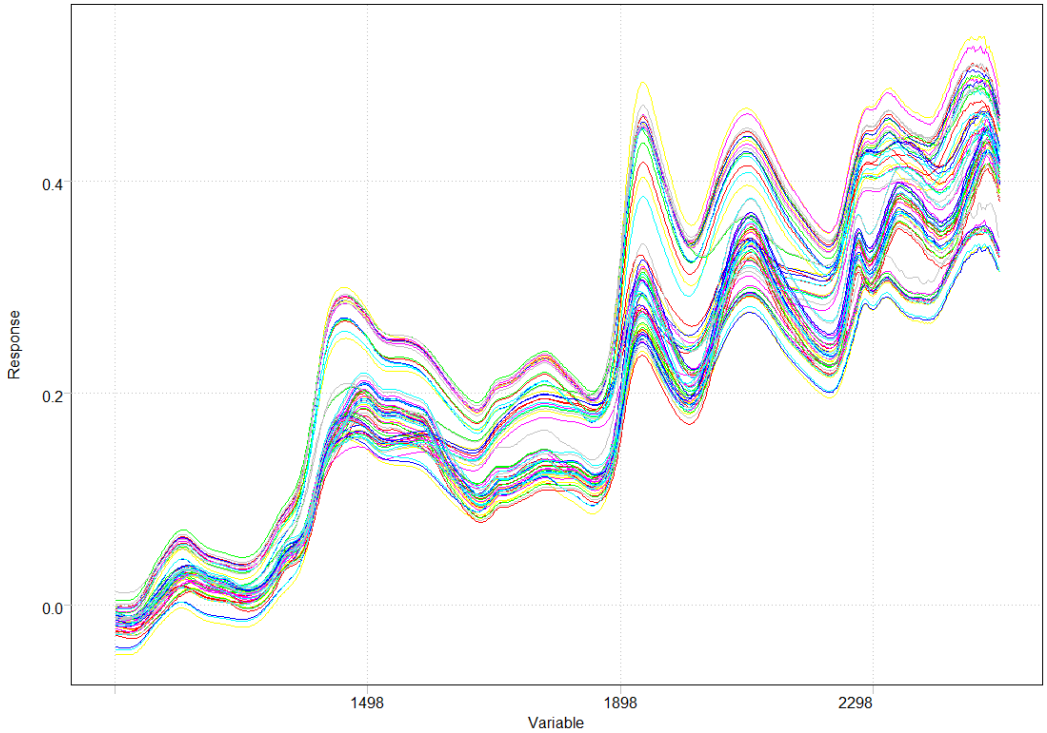
$$X = \begin{bmatrix} x_{1l} & x_{12} & \cdots & \cdots & x_{1j} \\ x_{2l} & x_{22} & \cdots & \cdots & x_{2j} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ x_{I1} & x_{I2} & \cdots & \cdots & x_{IJ} \end{bmatrix}$$

filas: muestras
Columnas: variables (longitudes de onda o números de onda en NIRS)

La data NIR es una data secuencial



# 2 Visualización de los datos

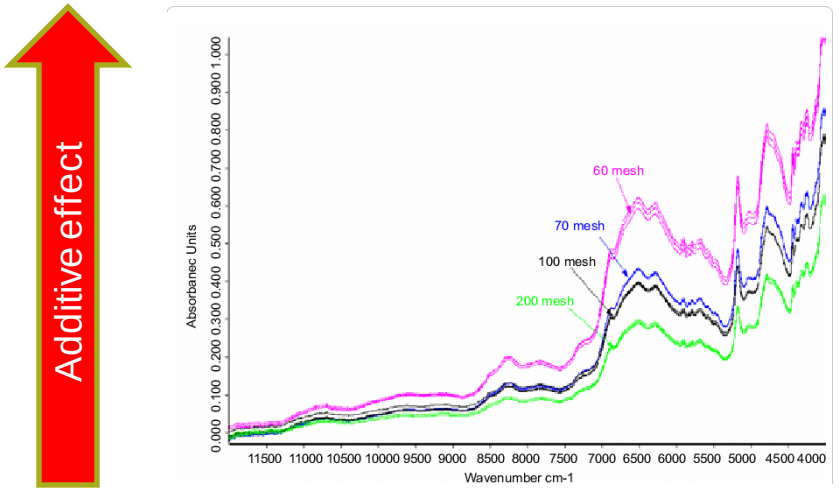



- Colinealidad entre variables
- Dispersión de la luz
- Presencia de espectros anómalos
- Adecuada respuesta (Ley de Lambert Beer)

# Visualización de los datos

La diferencia del tamaño de partícula del material afectará la dispersión y es una muy fuente importante de variación en los espectros NIR

Estos efectos tienen una naturaleza aditiva y multiplicativa y varían de una muestra a otra.



# 3 Técnicas de transformaciones o pre-procesamiento

Señal = Señal verdadera + Ruido azaroso

Información relevante + información irrelevante

Técnicas de corrección de línea base

Técnicas de suavizado

El procesamiento previo se realiza para eliminar los sesgos no químicos de la información espectral y preparar los datos para un análisis posterior

Transformaciones

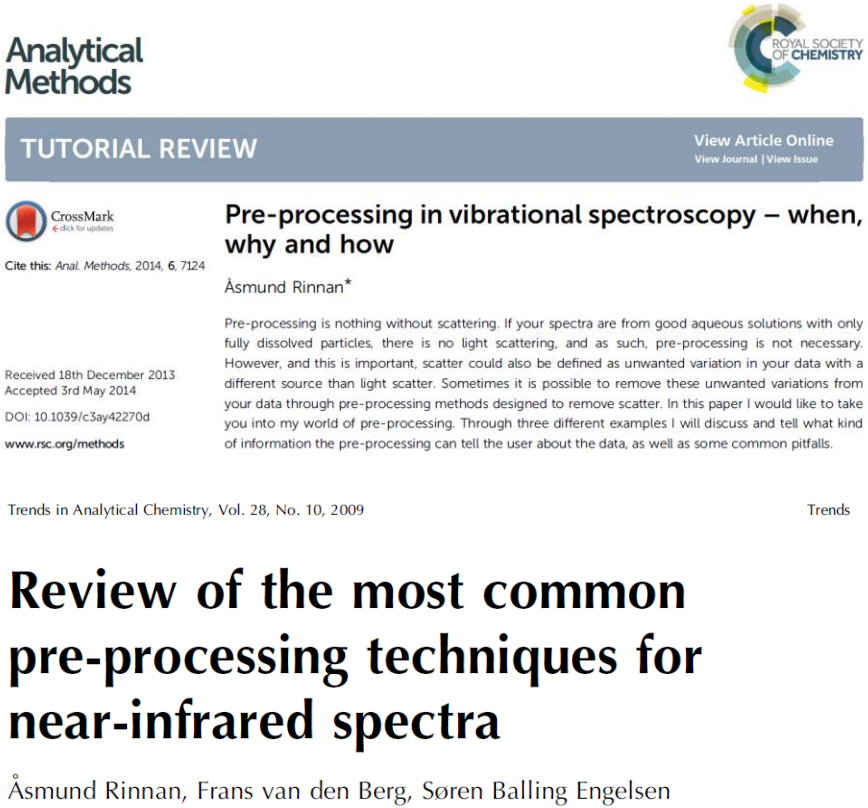
Aplicado en filas

Pre - procesamiento

Aplicado en columnas

# Técnicas de pre-tratamiento de espectros

Suavizados
Derivadas
Corrección de línea base
Corrección multiplicativa de señal (MSC)
Standar normal variate (SNV)
Normalización
Transformaciones kubelka munk
Centrado y autoescalado



Analytical Methods

ROYAL SOCIETY OF CHEMISTRY

TUTORIAL REVIEW

View Article Online
View Journal | View Issue

CrossMark

Cite this: *Anal. Methods*, 2014, **6**, 7124

## Pre-processing in vibrational spectroscopy – when, why and how

Åsmund Rinnan*

Pre-processing is nothing without scattering. If your spectra are from good aqueous solutions with only fully dissolved particles, there is no light scattering, and as such, pre-processing is not necessary. However, and this is important, scatter could also be defined as unwanted variation in your data with a different source than light scatter. Sometimes it is possible to remove these unwanted variations from your data through pre-processing methods designed to remove scatter. In this paper I would like to take you into my world of pre-processing. Through three different examples I will discuss and tell what kind of information the pre-processing can tell the user about the data, as well as some common pitfalls.

Received 18th December 2013
Accepted 3rd May 2014

DOI: 10.1039/c3ay42270d

www.rsc.org/methods

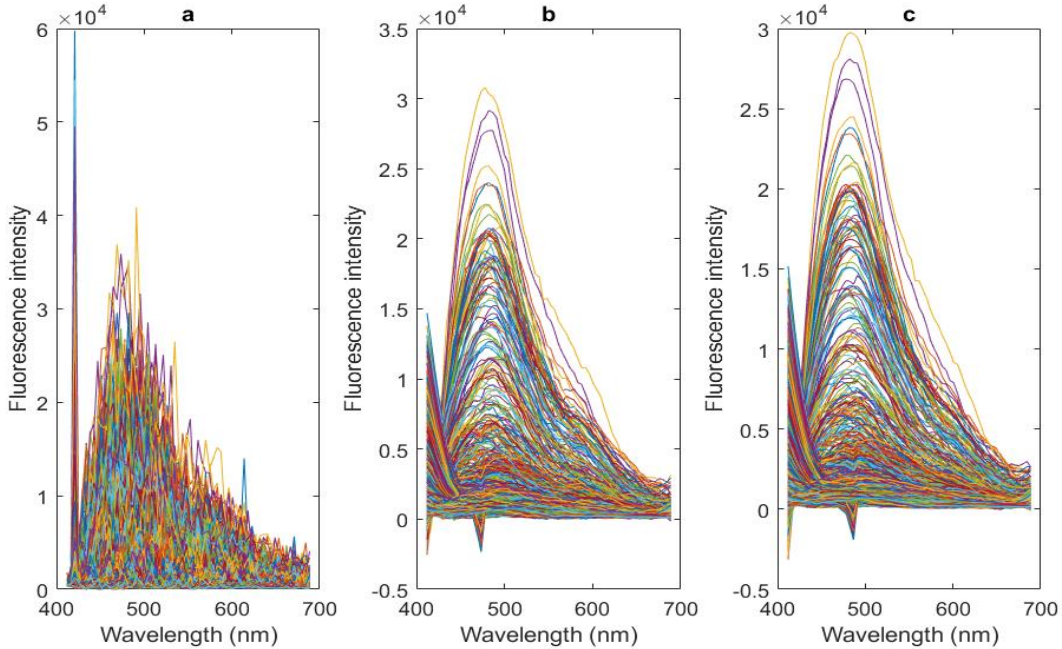
Trends in Analytical Chemistry, Vol. 28, No. 10, 2009 Trends

## Review of the most common pre-processing techniques for near-infrared spectra

Åsmund Rinnan, Frans van den Berg, Søren Balling Engelsen

# Suavizado

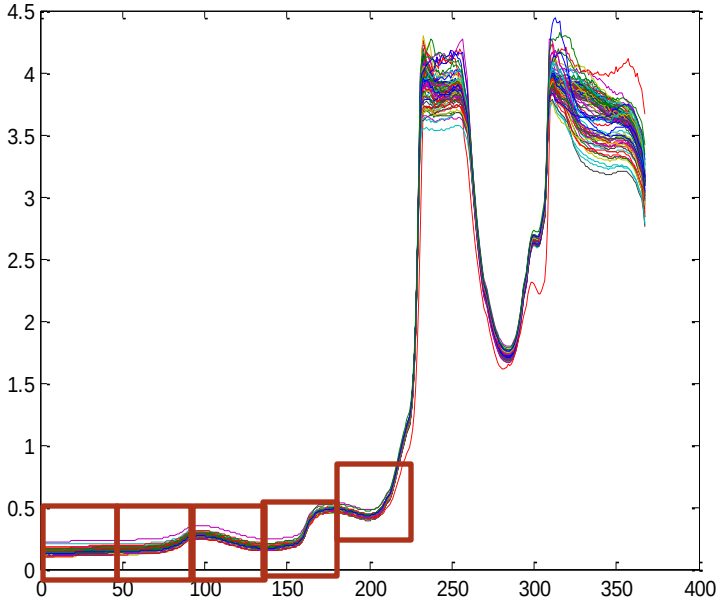
El objetivo es remover el ruido aleatorio



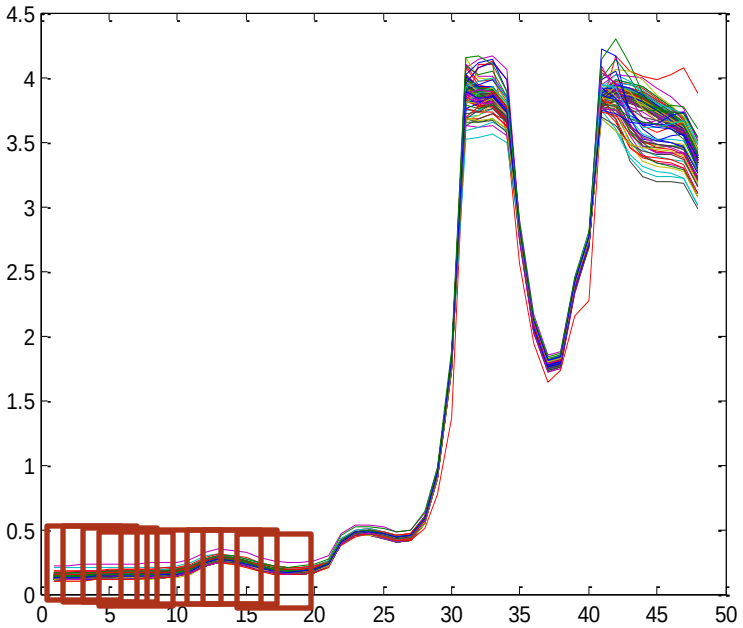

Precaución:

El suavizado excesivo puede reducir la intensidad y la resolución de la señal. Al mismo tiempo, si el suavizado no es suficiente, el ruido permanecerá en los datos

# Métodos de suavizado



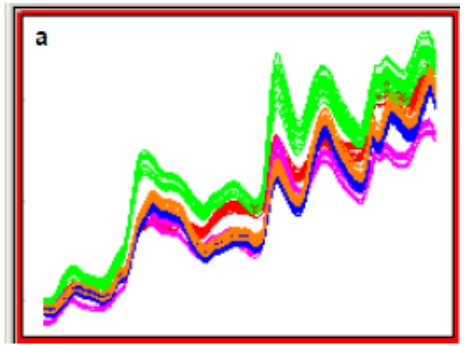

Media fija

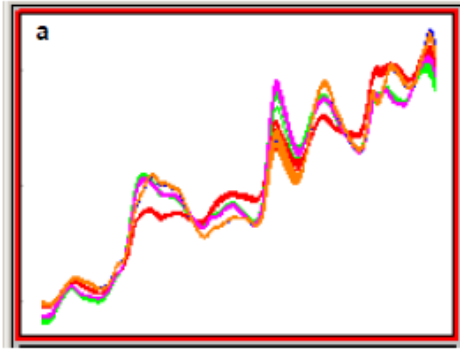



Media móvil

# Corrección multiplicativa de señal

MSC intenta reducir los efectos multiplicativos y aditivos de la línea de base causados por la dispersión en las mediciones NIR

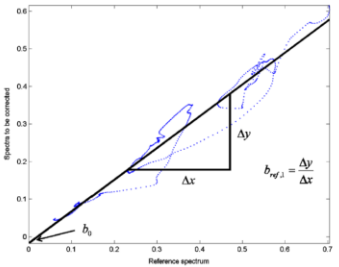



$$x_{org} = b_0 + b_{ref,1}.x_{ref} + \mathrm{e}$$

$$x_{corr} = \frac{x_{org} - b_0}{b_{ref,1}} = x_{ref} + \frac{e}{b_{ref,1}}$$

*X org* = espectros originales
*X corr* =espectros corregidos

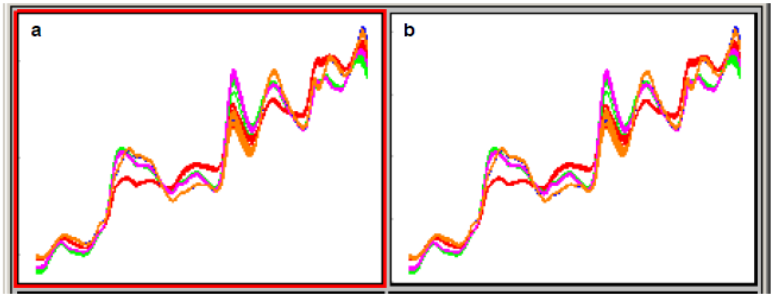


# Standar normal variate (SNV)

SNV se aplica para corregir los efectos de las interferencias multiplicativas de dispersión y tamaño de partícula (al igual que MSC), pero no requiere un espectro de referencia.

Corrección por MSC

Corrección por SNV




$$x_{corr} = \frac{x_{org} - a_0}{a_1}$$

$x_{corr}$ : espectro corregido
$x_{org}$ : espectro original
$a_0$ : promedio de los valores del espectro a ser corregido
$a_1$: desviación estándar de los valores del espectro a ser corregido

# Normalización

Remueve variaciones sistemáticas producidas por la cantidad de muestra producidas por la cantidad de muestra

$\|\mathbf{x}_i\|_\infty = max|x_{ij}|$ Norm sup $l^\infty$

$\|\mathbf{x}_i\|_1 = \sum_{J=1}^{J}|x_{ij}|$ Norm $l_1$

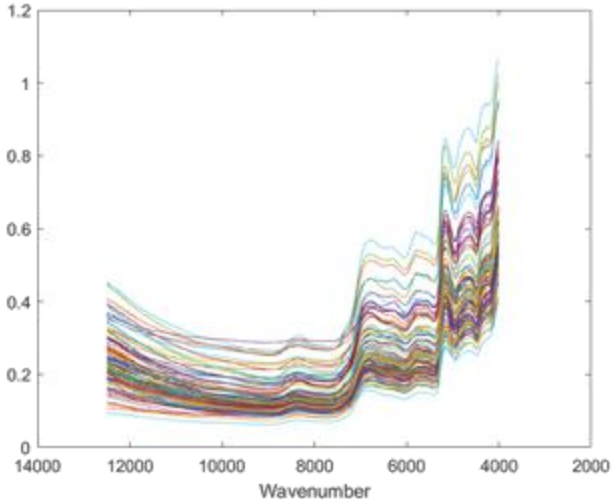
$\|\mathbf{x}_i\|_2 = \sqrt{\sum_{J=1}^{J} x_{ij}^2}$ Norm $l_2$

$$x_{ij(norm)} = \frac{x_{ij}}{\|\mathbf{x}_i\|} \quad j=1,2\ldots J$$

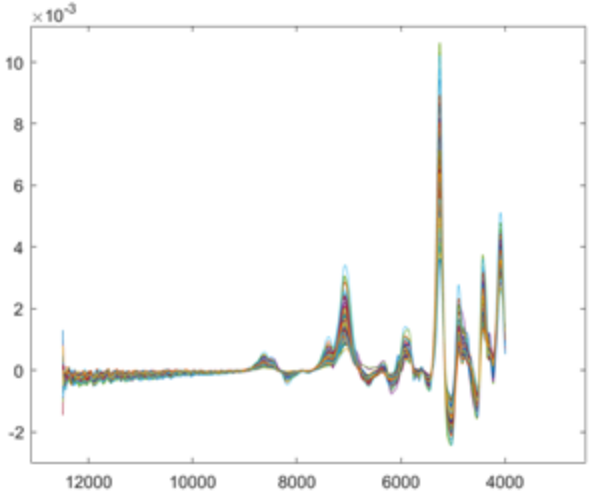
# Derivadas

- Las derivadas se utilizan para reducir los efectos de la línea de base y eliminar las señales de fondo constantes para mejorar la resolución visual.
- Las pequeñas diferencias entre los espectros NIR no son obvias de observar, los picos pueden superponerse.
- Al separar los picos superpuestos, se mejoran las pequeñas diferencias espectrales y se reduce el desplazamiento de la línea de base.
- La primera derivada es un método muy eficaz para eliminar las compensaciones ordinarias lineales (bias).
- La primera derivada es un método muy eficaz para eliminar una línea de base inclinada de un espectro.

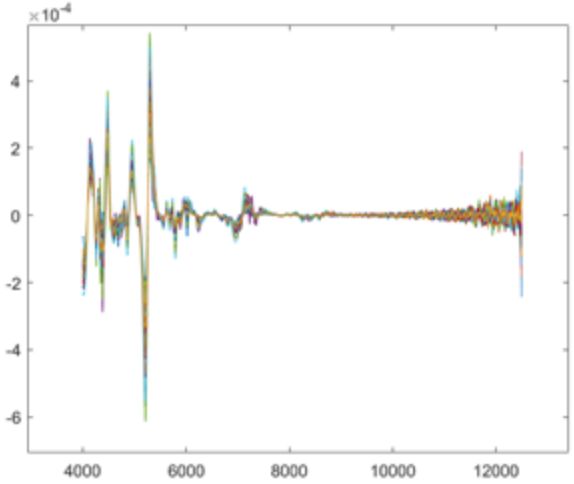
# Derivadas




Raw data




Primera derivada

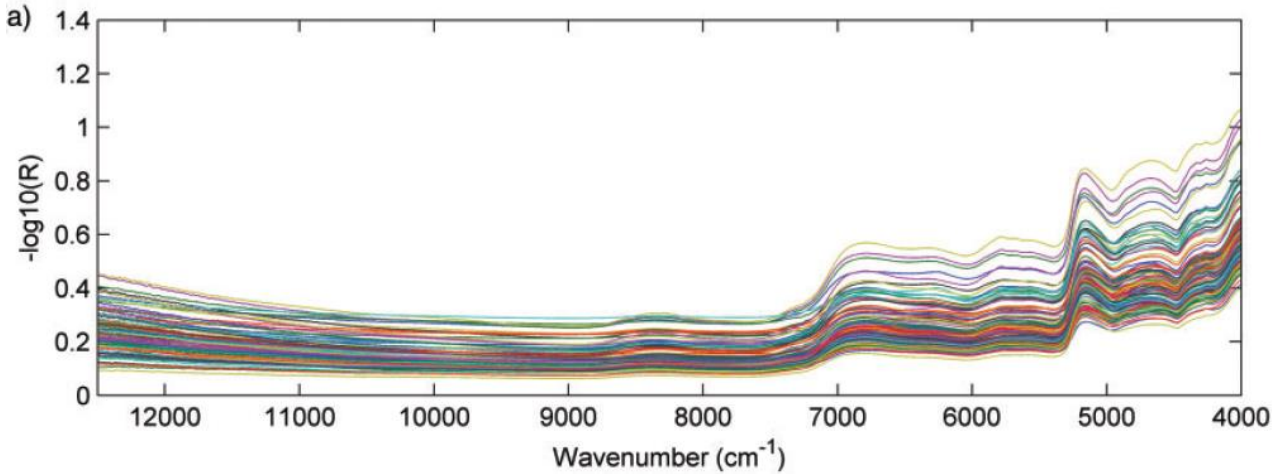



Segunda derivada

# Logaritmos

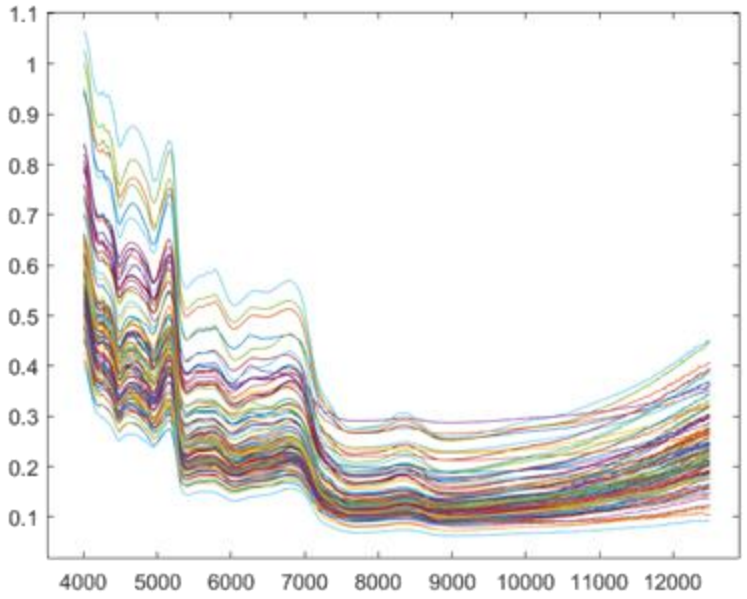
Los datos se linealizan a una escala logarítmica. La transformación de estos datos no altera la interpretación de los resultados..

Ejemplo A=-logT ; pA=-logR

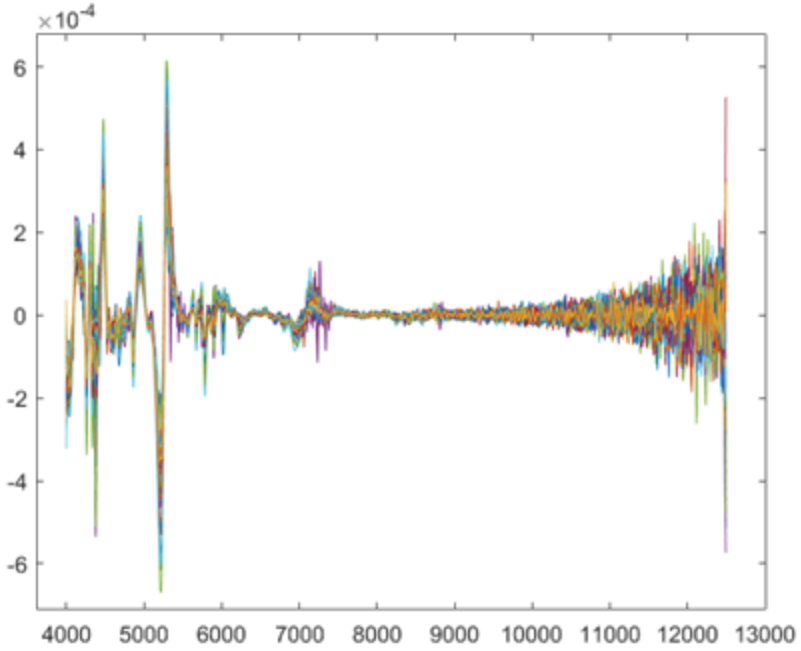



Trabajar con data lineal (respuesta lineal entre la concentración y la señal, para las técnicas como PLS, CLS, PCR)

# Cómo quedará mi data?

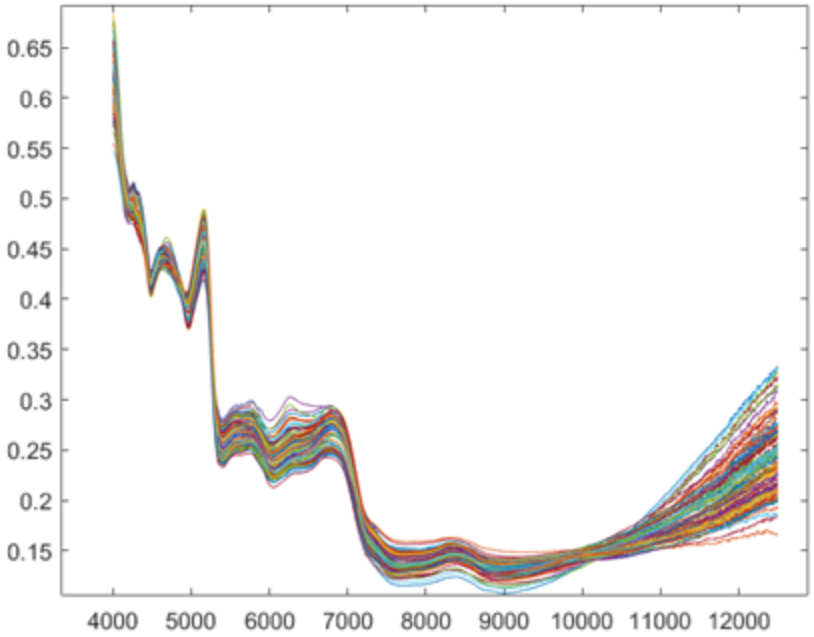



NIR data - 2da derivada(15 points)

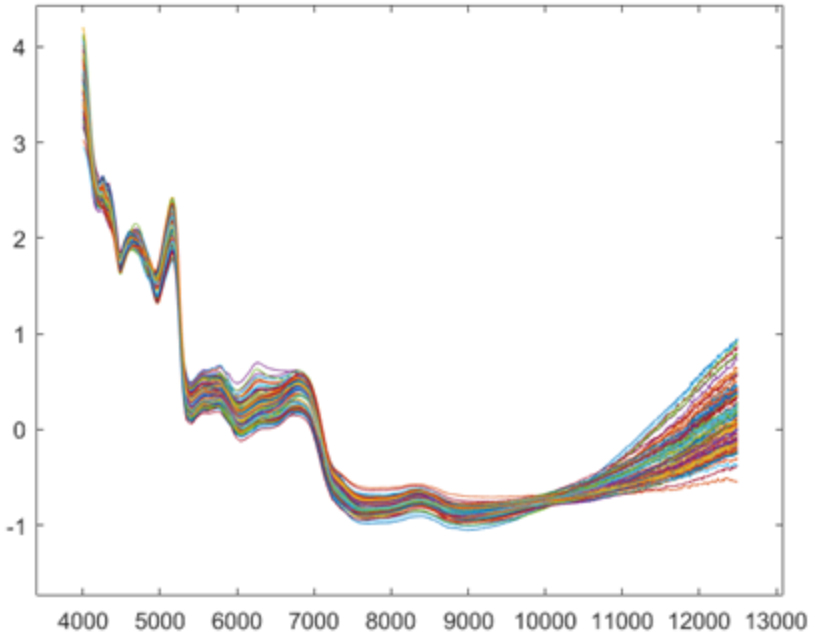


# Cómo quedará mi data?

**MSC (mean)**




**SNV**



# Técnicas de preprocesado:

Mean Center

$$x_{ij(cm)} = x_{ij} - \bar{x}_j$$

Autoscale

$$x_{ij(au)} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Variance scaled

$$x_{ij(sv)} = \frac{x_{ij}}{s_j}$$

**Recomendaciones.**

- Cuando la matriz X tiene variables con unidades diferentes, se recomienda el autoescalado.
- Los métodos de autoescalado son sensibles a la presencia de valores atípicos (especialmente el autoescalado por rango).
- Se recomienda *Centrar en la media* los datos de espectroscopia.

# 4 Análisis Exploratorio (PCA)

**Método de proyección**

Proyecta datos multivias en un espacio dimensional menor

**Método de reducción de variables**

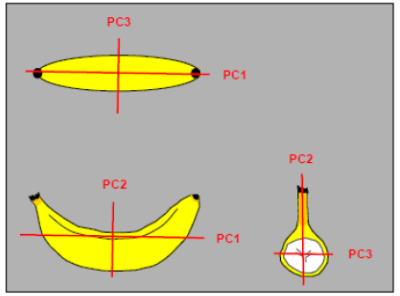
Elimina datos redundantes e innecesarios

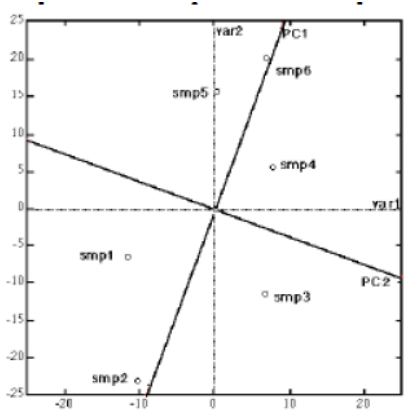
**Detecta datos anómalos**

Muestras atipicas son detectadas

**Establece correlacions**

Permite establecer correlaciones entre variables



# Ejemplos PCA:

Scores (T)

Loadings (L')

12%

66%

33%

8%

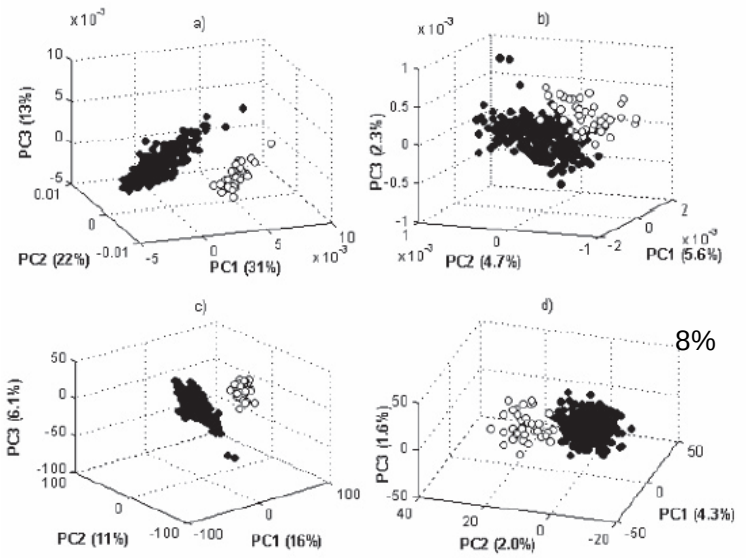
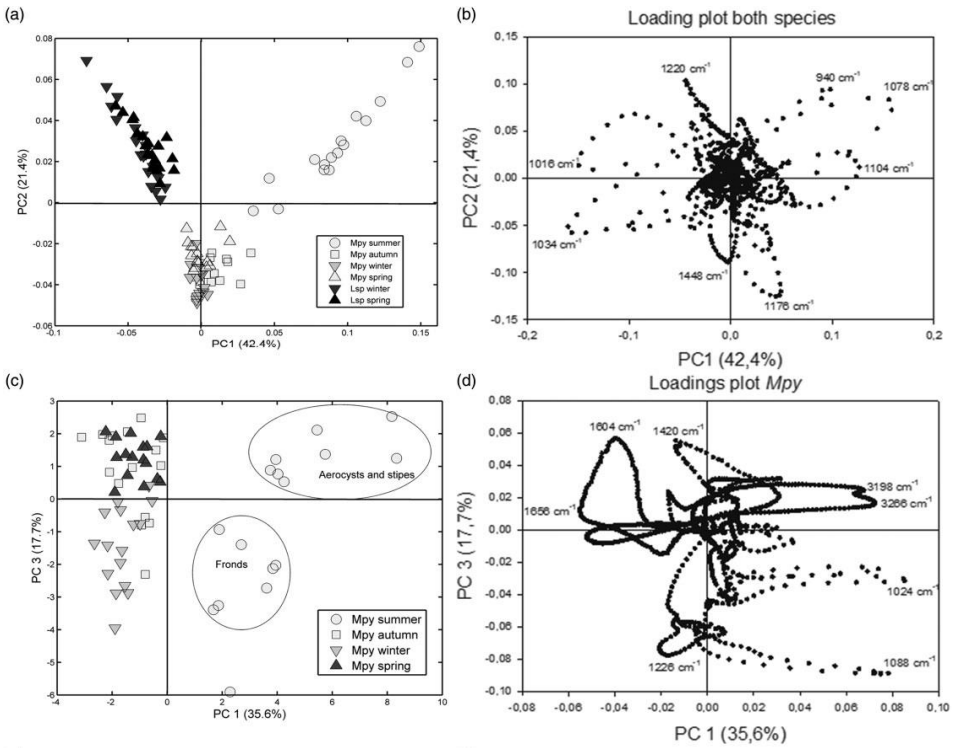



**Fig. 2.** PCA scores plot of different data preprocessing and transformations: a) *MC* and first derivative, b) *MC* and second derivative, c) *AU* and first derivative, d) *AU* and second derivative. Filled circles are *E. globulus* samples and empty circles are *E. nitens* samples.

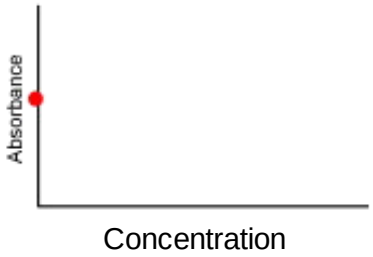
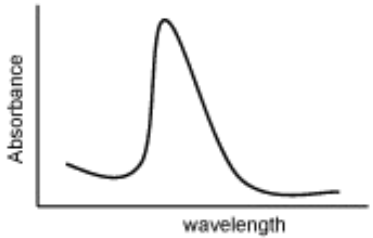
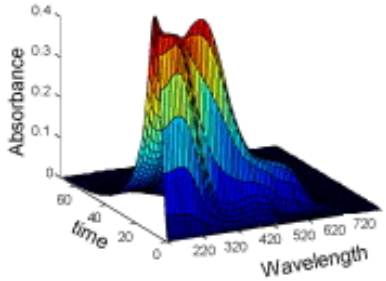
Diferenciación de especies por NIR




Diferenciación de estructuras morfológicas por NIR

# 4 Generar modelo de regresión

# Regresión multivariada vs univariada

| Data type | Scalar | Vector | Matrix |
|---|---|---|---|
| Numerical representation | 0.7 | [ 0.5... 0.8... 1.1... 0.9 ] | $\begin{pmatrix} 0.5... & 0.8... & 1.1... & 0.9 \\ 0.6... & 0.6... & 1.0... & 0.7 \\ \dots & & & \\ 0.8... & 0.4... & 1.2... & 0.6 \end{pmatrix}$ |
| Graphical representation |   |   |   |
| Order | Order O | Order 1 | Order 2 |

# Modelos de regresión

## Objetivo

Modelar una correlación entre datos **medidos** o calculados de forma independiente (X) y alguna **propiedad de la muestra** (Y).

Mediciones

- Cromatografícas
- Espectroscópicas
- Resonancia magnetica nuclear

Property

- Concentración
- Propiedad mecánica
- Actividad biologica

# Pasos para calibraciones multivariadas

## 1. Crear un set de calibración

Construir un modelo a partir de una muestra experimental de la que se conoce el contenido de analitos y las señales multivariantes (espectros, cromatograma, etc).

## 2. Validar el set de datos

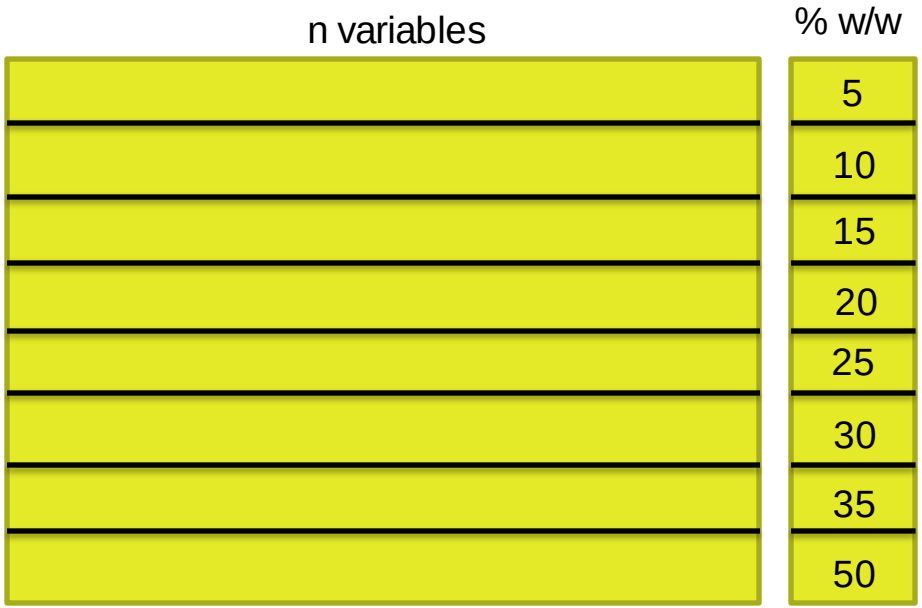
Las muestras se utilizan para medir el poder predictivo del conjunto de calibración actual. Puede realizar validaciones internas o externas.

## 3. Emplear tu set de datos para predecir muestras externas (Set de predicción)

Usar tu modelo para predecir muestras nuevas

# Steps for Calibration

Data Array



# Calibration

Select samples for the “Calibration Set". They should be chosen to be representative of the entire region to be modeled.

Steps:

a) Collect experimental data X
b) Determine experimentally the concentration of the analyte of interest (Y) by some Reference method.
c) Build an appropriate model that correlates X and Y

# Pasos para calibraciones multivariadas

## 1. Crear un set de calibración

Construir un modelo a partir de una muestra experimental de la que se conoce el contenido de analitos y las señales multivariantes (espectros, cromatograma, etc).

## 2. Validar tu set de datos

Las muestras se utilizan para medir el poder predictivo del conjunto de calibración actual. Puede realizar validaciones internas o externas.

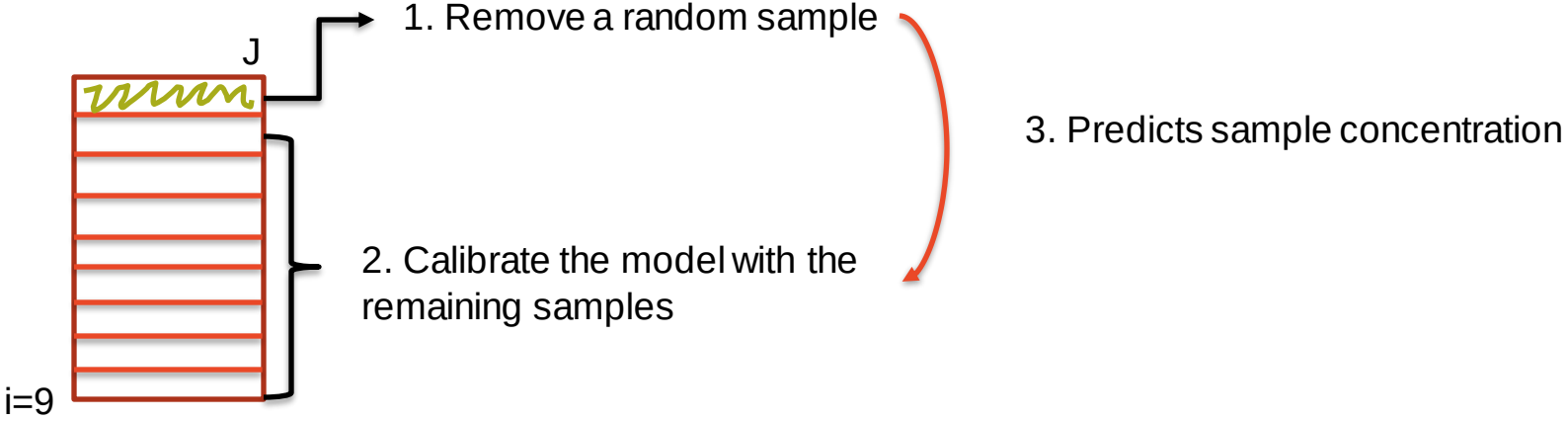
## 3. Emplear tu set de datos para predecir muestras externas (Set de predicción)

Usar tu modelo para predecir muestras nuevas

# Steps for Validation

**M**odeling and **V**alidation

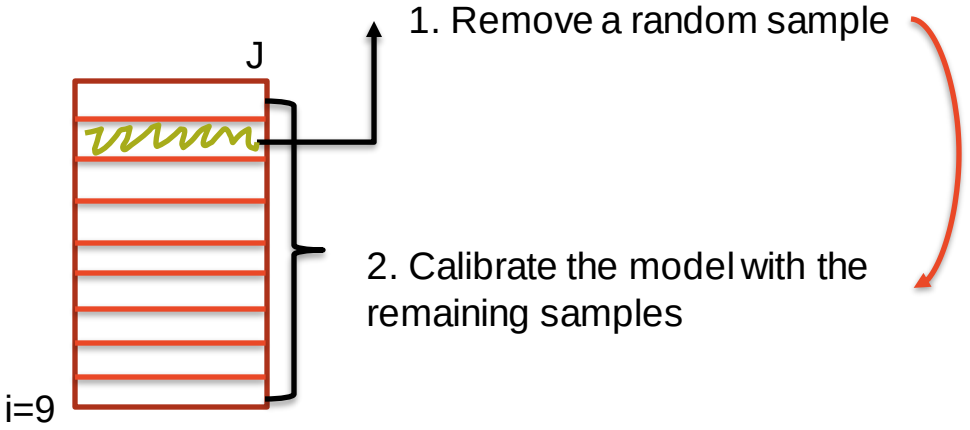
Cross validation : “Leave one out”



# Steps for Validation

Cross validation: Leave one Out

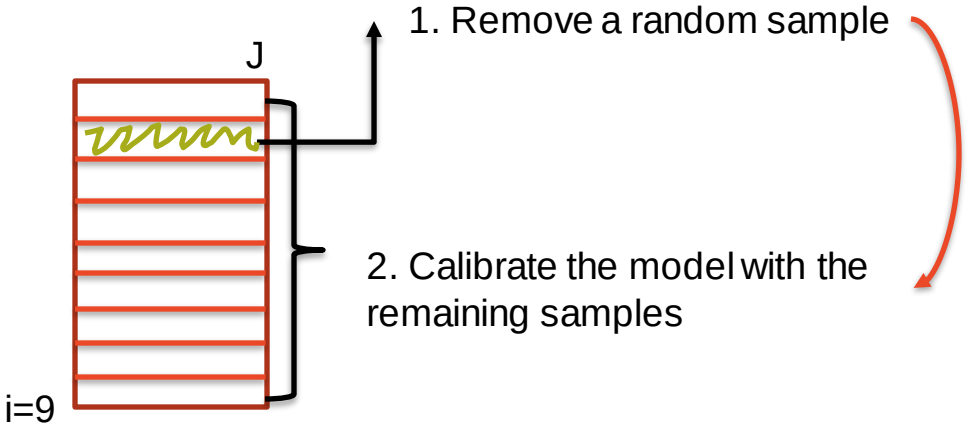
**M**odeling and **V**alidation



# Steps for Validation

Cross validation: Leave one Out

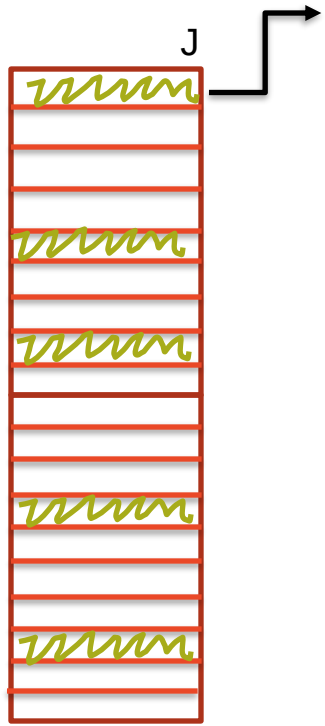
**M**odeling and **V**alidation



# Steps for Validation

External Validation

**M**odeling and **V**alidation




1. Remove the 25% of random simples.

2. The remaining 75% is used to create the calibration set.

# Steps for Validation

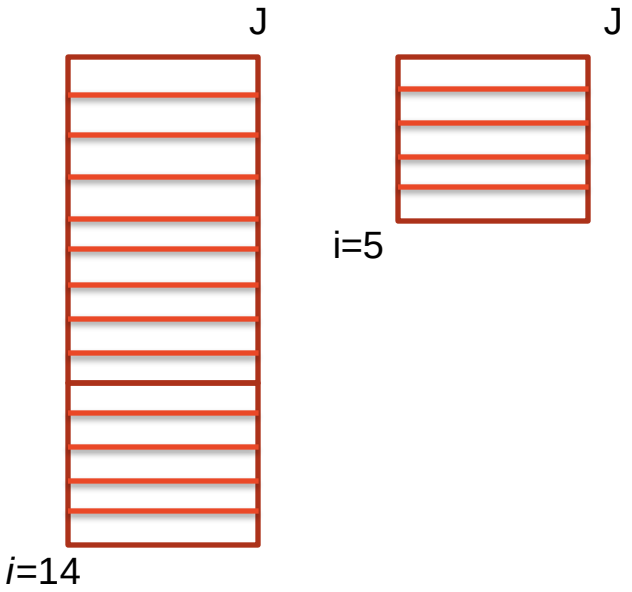
External Validation

**M**odeling and **V**alidation

1. Remove the 25% of random samples.

2. The remaining 75% is used to create the calibration set.

3. The 25% of the samples is used as a Prediction Set.




* Only applicable when the number of samples is high *

You can choose ratios of 75/25 (most recommended) and also 80/20 o 70/30 and 60/40 . It's up to you

# Pasos para calibraciones multivariadas

## 1. Crear un set de calibración

Construir un modelo a partir de una muestra experimental de la que se conoce el contenido de analitos y las señales multivariantes (espectros, cromatograma, etc).

## 2. Validar tu set de datos

Las muestras se utilizan para medir el poder predictivo del conjunto de calibración actual. Puede realizar validaciones internas o externas.

## 3. Emplear tu set de datos para predecir muestras externas (Set de predicción)

Usar tu modelo para predecir muestras nuevas

# Predicción

Una vez construido el modelo, podemos aplicarlo a muestras de concentraciones desconocidas.

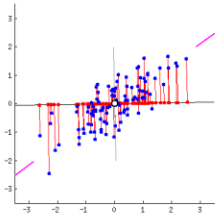
Cada cierto tiempo, hay que repetir el proceso de validación para garantizar que el modelo sigue siendo válido.

# Partial Least Squares (PLS)

PLS es el método multivariado más usado para análisis cuantitativo en química analítica

$$\mathbf{X}_{(nxp)} = \boldsymbol{T}_{(nxd)} . \boldsymbol{P}^{T}_{(dxp)} + \boldsymbol{E}_{(nxp)}$$

$$\mathbf{Y}_{(nxm)} = \mathbf{U}_{(nxd)}\mathbf{Q}^{T}_{(dxm)} + \mathbf{F}_{(nxm)}$$



**X**: medidas espectrales
**Y:** concentraciones
*d* = número de variables latentes

- Es un método de calibración inversa
- Es un métodos de reducción de variables
- Encuentra las variables latentes maximizando su covarianza con la propiedad a medir

## 1º Calibración / Modeling

$$\mathbf{Y} = \boldsymbol{X}_{cal}\mathbf{B} + \mathbf{E}$$

Vector de regresión (**B**)

## 2º Validación

$$\hat{\mathbf{Y}} = \boldsymbol{X}_{val}\mathbf{B} + \mathbf{E}$$

Usa un set de muestras cuya concentración es convalidar el modelo

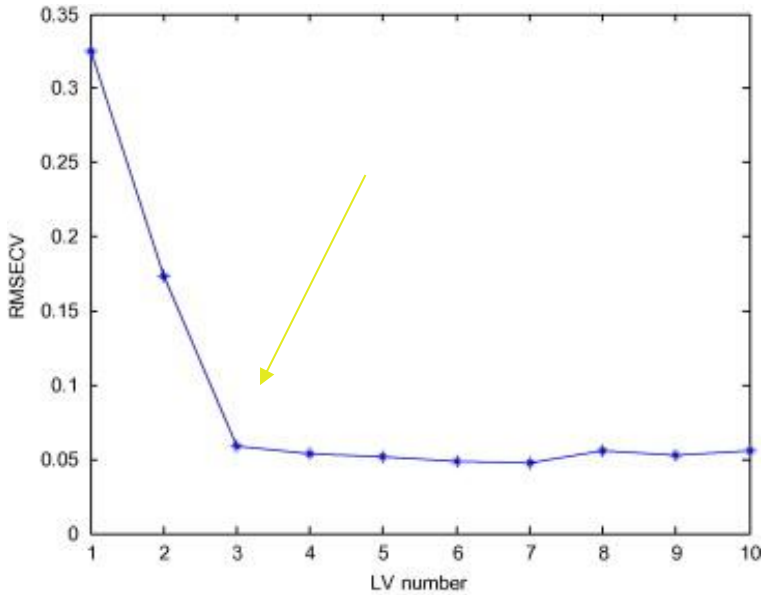
## 3º Predicción

$$\hat{\mathbf{Y}} = \boldsymbol{X}_{new}\mathbf{B} + \mathbf{E}$$

Usa el modelo desarrollado para encontrar la concentración (o propiedad) en una muestra incógnita

Una etapa crucial del desarrollo de un modelo PLS es encontrar el número de variables latentes (factores) a usar

Normalmente se usa el error de validación cruzada




# de Variables latentes

$$\mathbf{Y} = \boldsymbol{X}_{\boldsymbol{cal}}\mathbf{B} + \mathbf{E}$$

Si elegimos 3 variables latentes para neustro modelo

Matemáticamente esto signifca que X tendrá 3 columnas y el vector B tendrá 3 filas, correspondientes a 3 variables latentes

Article

Applied Spectroscopy

Applied Spectroscopy
2017, Vol. 71(10) 2263–2277

DOI: 10.1177/0003702817715654
journals.sagepub.com/home/asp
SAGE

# Chemical Characterization and Determination of the Anti-Oxidant Capacity of Two Brown Algae with Respect to Sampling Season and Morphological Structures Using Infrared Spectroscopy and Multivariate Analyses

Angelo Beratto[1], Cristian Agurto[1], Juanita Freer[2,3], Carlos Peña-Farfal[4], Nicolás Troncoso[1], Andrés Agurto[1], and Rosario del P. Castillo[5]

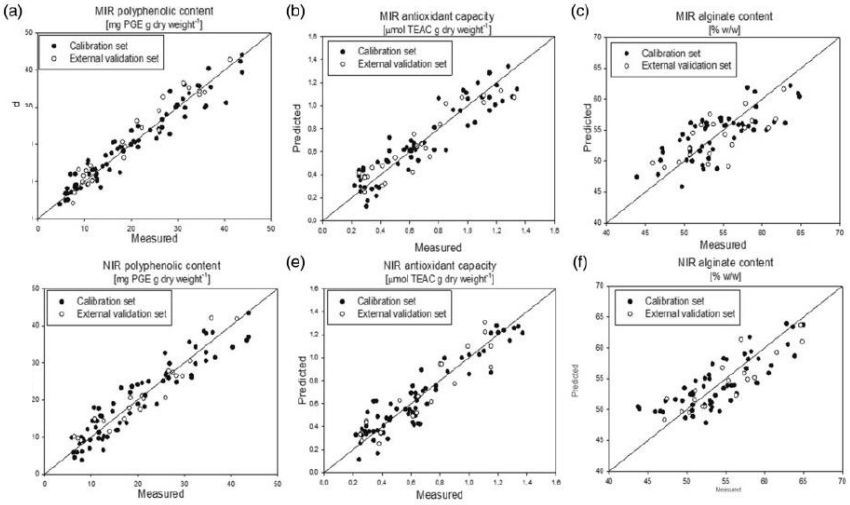
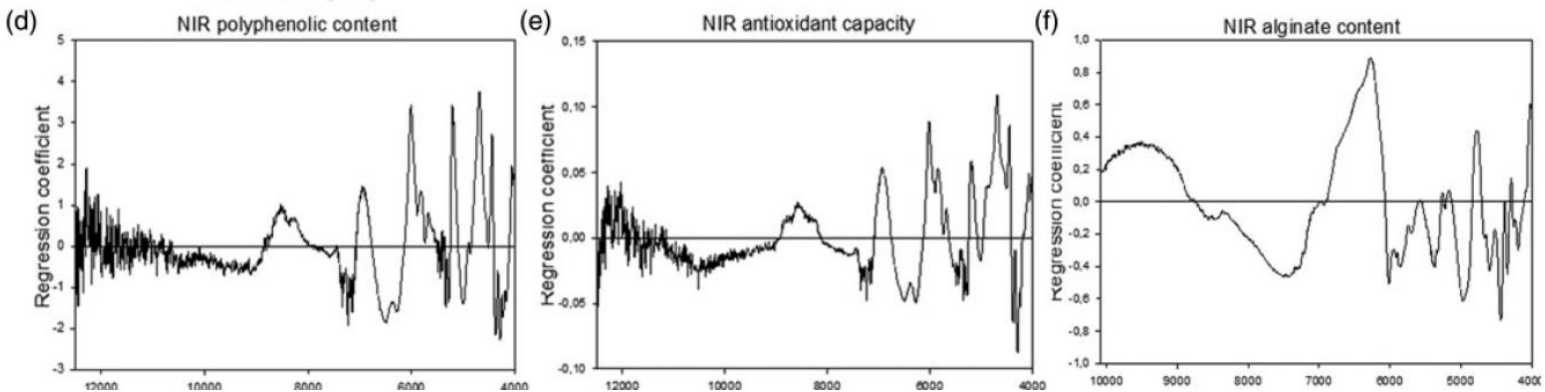


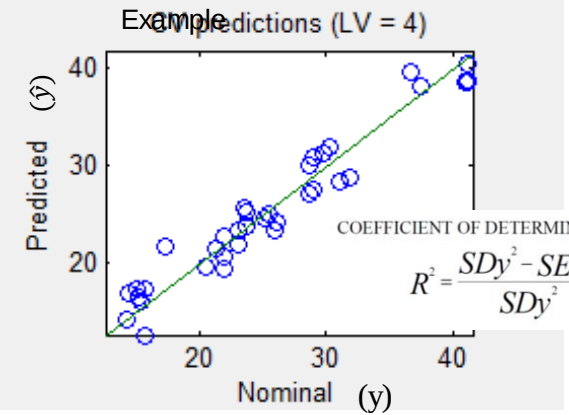

Figure 5. Comparison between the measured values for polyphenolic content, antioxidant capacity, and alginate content using the



# Parámetros de error (cifras de mérito)

## Precisión y exactitud

Example




$$RMSEC = \sqrt{\frac{\sum_{i=1}^{n}(y_{ij} - \hat{y}_{ij})^2}{n-k}}$$

Error de calibración

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(y_{ij} - \hat{y}_{ij})^2}{n}}$$

Error de validación cruzada

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(y_{ij} - \hat{y}_{ij})^2}{n}}$$

Error de validación externa

$n$: número de muestras
$k$: Nro. de LVs

## *Otras cifras de mérito*

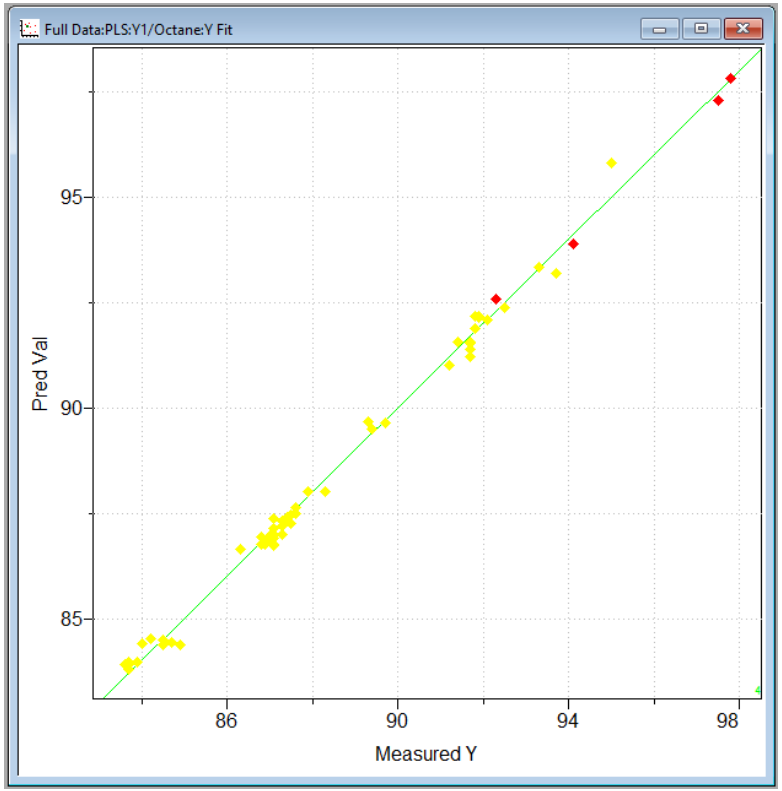
- Sensibilidad $SEN_n = \frac{1}{\|\beta_n\|}$ → Cambio en la señal por cambio en la concentración
- Sensibilidad analítica $\gamma_n = SEN_n/S_y$ → Sirve para comparar diferentes técnicas analíticas
- Límites de detección(LOD) $LOD = 3{,}3S_y/SEN_n$ → Concentración mínima estadísticamente diferente al ruido
- Límites de cuantificación (LOQ) $LOC = 3{,}3S_y/SEN_n$ → Concentración mínima para discernir entre dos concentraciones razonablemente .

# Cifras de mérito

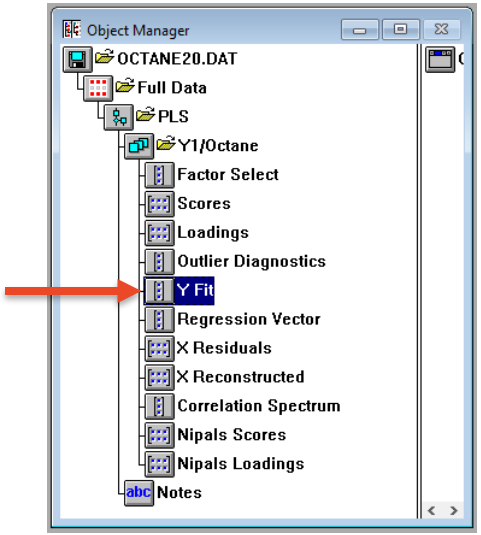
Parámetros básicos requeridos para reportar la bondad de un modelo PLS

| | |
|---|---|
| Número de muestras de calibración | |
| Número de muestras de validación | |
| Rango de concentración | |
| RMSEC | |
| RMSECV | |
| RMSEP | ? |
| $R^2$ Calibración | |
| $R^2$ CV | |
| $R^2$ Predicción | ? |

# 1. Gráfico y fit: valor medido vs valor predicho

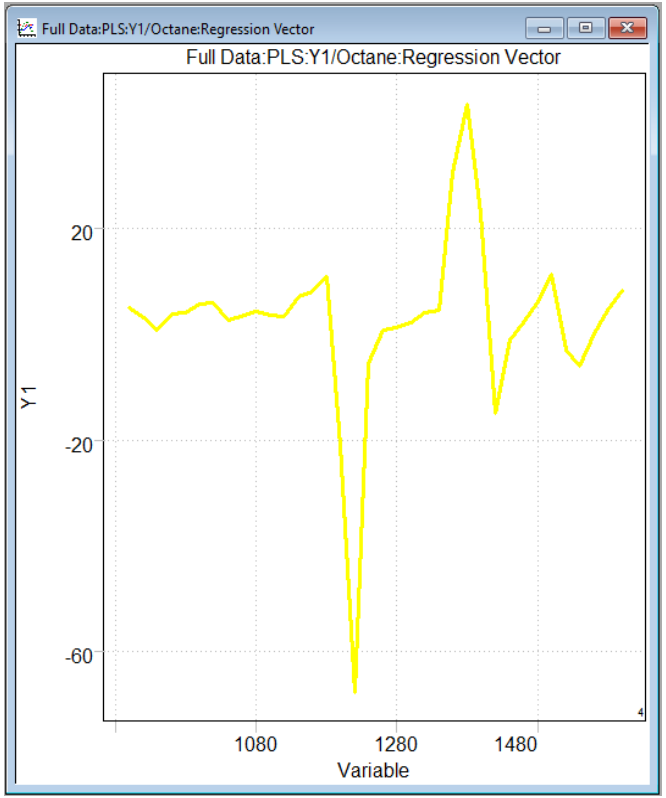



**Y Fit**

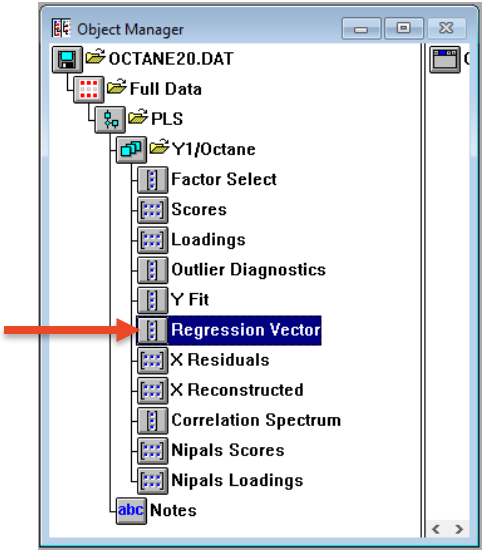



Muestra el ajuste de calibración (y también predicción) realizado por el modelo PLS. Realiza una comparativa entre el valor medido y considerado “verdadero” vs el valor predicho por el modelo para cada muestra.

# 2. Vector de regresión




**Vector de regresión**




Representación de variables importantes que están modelando nuestra calibración .

Nos permite establecer correlaciones en relación a nuestras longitudes de onda

# Pasos para el desarrollo un modelo multivariado

01 Construir un set de datos

- Visualizar la data original
- Escoger un algoritmo
- Establecer un preprocesamiento
- Aplicar transformaciones
- Identificar muestras anómalas

02 Validación del set de datos

- Validación interna
- Validación externa

03 Predecir muestras nuevas

04 Validar !!

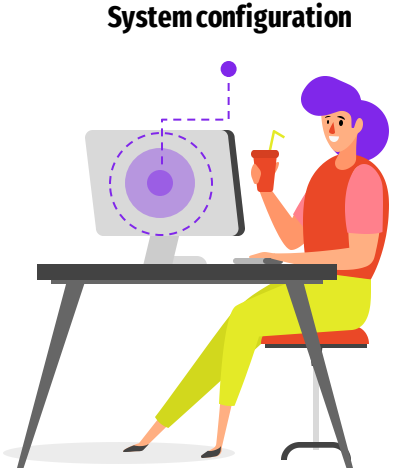
# Modelo PLS

01 **Visualizar la data original**

02 **Escoger un algoritmo**

03 **Establecer un preprocesamiento**

04 **Seleccionar un método de validación interna**




05 **Aplicar transformaciones**
(de ser necesario)

06 **Identificar muestras anómalas**
(remover de ser necesario)

07 **Seleccionar variables latentes**

08 **Observar gráficas: Yfit, Vector de regresión, residuales.**

# Muchas gracias