
**UNIVERSIDAD DEL AZUAY**

**DEPARTAMENTO DE POSGRADOS**
**MAESTRIA EN MATEMÁTICAS APLICADAS**

**TEMA:**
**"TRATAMIENTO MATEMÁTICO DE SEÑALES ESPECTROFOTOMÉTRICAS INFRARROJAS PARA LA CUANTIFICACIÓN DE AZÚCARES"**

**TRABAJO DE GRADUACIÓN PREVIO A LA OBTENCIÓN DEL TITULO DE MAGISTER EN MATEMÁTICAS APLICADAS**

**AUTOR:**
**ING. BOLÍVAR ANDRÉS PÉREZ GONZÁLEZ**

**DIRECTOR:**
**PROF. ROBERTO TODESCHINI**

**CUENCA-ECUADOR**

**2017**

*Bolívar Andrés Pérez González*

*Trabajo de Graduación*

*Prof. Roberto Todeschini*

*Febrero 2017*

# *"TRATAMIENTO MATEMÁTICO DE SEÑALES ESPECTROFOTOMÉTRICAS INFRARROJAS PARA LA CUANTIFICACIÓN DE AZÚCARES"*

# DEDICATORIA

El presente trabajo está dedicado a la memoria de mi madre Bertha González, quien siempre supo estar presente en mi desarrollo personal. Este es un logro que ella lo hubiese querido ver y participar. Amor siempre.

# AGRADECIMIENTOS

Al Milano Chemometrics and Research Group de la Universidad Bicocca Milan-Italia, en especial al Profesor Roberto Todeschini, por haberme permitido realizar el presente trabajo bajo su dirección.

# RESUMEN

En el presente trabajo, se estudiaron diferentes métodos matemáticos que permitan cuantificar la cantidad de Glucosa, Fructosa, Sacarosa y Lactosa, presente en muestras acuosas, a partir de espectros infrarrojos (FTIR). De todos los modelos calculados, la lactosa no presentó un modelo efectivo para predecir su concentración. En los otros casos el mejor pretratamiento utilizado, fue la primera derivada, sin importar el filtro que se utilice, la selección de variables óptimas fue el Algoritmo Genético utilizando el método Mínimos cuadrados parciales (PLS) y los mejores modelos de regresión fueron PLS para la fructosa y PCR para la Glucosa y Sacarosa.

# ABSTRACT

This work dealt with the study of different mathematical methods to quantify the amount of glucose, fructose, sucrose and lactose present in aqueous samples, from infrared spectrum (FTIR). Of all the calculated models, lactose did not present an effective model so as to predict its concentration. In the other cases, the best pre-treatment used was the first derivative, regardless the filter. The selection of optimal variables was the Genetic Algorithm using the Partial Least Squares method (PLS). The best regression models were PLS for fructose and CRP for glucose and sucrose.

UNIVERSIDAD DEL AZUAY
Dpto. Idiomas

Translated by,
Lic. Lourdes Crespo

# ÍNDICE

## ÍNDICE DE TABLAS



## ÍNDICE DE FIGURAS

## 1. INTRODUCCIÓN

La cuantificación de azúcares de un alimento, en general, tiene varias limitaciones tanto operacionales como de costo. En la química tradicional el método más conocido para determinar azúcares reductores, es el método de Fehling (Fehling), que se limita únicamente a la cuantificación de azúcares reductores. En la actualidad, y con el desarrollo de la cromatografía líquida de alta resolución (HPLC), la cuantificación de azúcares por este método constituye el ensayo estándar reconocido por la Association Of Analytical Communities (AOAC Internacional). Sin embargo, la limitación en este método es el costo que se genera, debido al costo del equipo como tal sumado a los costos adicionales de laboratorio (Thermo Scientific).

Debido a esto los investigadores alrededor del mundo han buscado otras formas de analizar azúcares de forma rápida, económica y exacta, siendo una alternativa a este problema la Espectroscopía Infrarroja, con la cual se puede medir la absorción inducida por los movimientos vibracionales producidos por grupos químicos específicos en la región del espectro infrarrojo. La región del espectro infrarrojo se encuentra entre los 12500-100 $cm^{-1}$, la cual se divide en 3 regiones denominadas NIR (Near-Infrared), MIR (Mid-Infrared) y FIR (Far-Infrared) (Dufour). La región del NIR que se encuentra entre los 12500 y 4000 $cm^{-1}$ se caracteriza por ser la primera región en la cual se puede observar bandas de absorción correspondientes a la vibración molecular. La región del MIR (4000 y 400$cm^{-1}$) es la región más importarte del espectro vibracional de las moléculas, dado que aquí se puede obtener toda la información relacionada con las moléculas orgánicas. La región del FIR comprendida entre 400 y 100 $cm^{-1}$ entrega información conformacional de las estructuras (Dufour). Debido a que el estudio sobre los espectros infrarrojos, inicialmente, no permitan interpretaciones muy exactas, se desarrollaron espectrofotómetros infrarrojos que utiliza el interferómetro de Michelson. Michelson descubrió que al aplicar las transformadas de Fourier al espectro infrarrojo obtenido utilizando un interferómetro las señales tienen mejor resolución y la interpretación de las mismas es más sencilla. Desde este acontecimiento, en el desarrollo de la espectroscopía infrarroja, se conoce a este sistema como espectrometría FTIR (Fourier Transform Infrared) (Subramanian y Rodriguez-Saona). Con base en este concepto se ha identificado, la región del espectro infrarrojo, en la cual se pueden medir los hidratos de carbono. Esta región está comprendida entre los 1500 y 800 $cm^{-1}$ (Wang , Kliks y Jun). En esta región los diferentes hidratos de carbono tienen diferentes bandas de absorción, siendo un problema generalizar modelos en el estudio de alimentos que presenten combinaciones de hidratos de carbono. Debido a que los espectrofotómetros, registran bandas de absorción cada 2 $cm^{-1}$, la cantidad de puntos del espectro es grande. Teniendo en cuenta que cada uno de los puntos del espectro representa una variable, la cantidad de variables presentes es elevada.

Para alcanzar el objetivo de obtener un modelo matemático que permita cuantificar la presencia de azúcares a partir de espectros infrarrojos FTIR, se ha considerado conveniente

utilizar 4 tipos de tratamientos matemáticos, que son: el pre-tratamiento de datos, la selección de variables, el desarrollo de modelos multivariante de regresión y la validación de los modelos. Teniendo en cuenta que cada tipo de tratamiento posee diferentes métodos, se pretende generar diferentes combinaciones de estos métodos para encontrar, el mejor modelo aplicable a espectros infrarrojos de mezclas de azúcares en solución acuosa.

El desarrollo de modelos para análisis de azúcares en muestras líquidas utilizando la espectroscopía infrarroja FTIR, tiene gran difusión, sin embargo, los modelos desarrollados no siempre cumplen con los parámetros de calidad del modelo. Los métodos a los cuales se hace referencia son el pre-tratamiento de datos, desarrollo del mejor modelo de regresión que se adapte a los datos obtenidos, y su posterior validación. Por esta razón, los modelos publicados son, en algunos casos, ambiguos y no reproducibles en otras condiciones de trabajo. (Sorol, Arancibia y Bortolato), (Wang, Kliks y Soojin), (Gabriel, Prestes y Pinheiro)

El objetivo del presente trabajo es desarrollar un modelo predictivo de regresión que permita cuantificar azúcares, utilizando la información de la espectroscopía infrarroja (FTIR) con técnicas de regresión multivariable

Para lograr este objetivo, es necesario conseguir cumplir con otros objetivos que son:

- Evaluar los métodos de pre-tratamiento de datos conocidos a espectros infrarrojos de azúcares y seleccionar el que mejor se adapte a los mismos.
- Desarrollar un método de regresión multivariable basado en selección de variables
- Validar el modelo de regresión

# 2. CAPITULO 1: MATERIALES Y MÉTODOS

En el capítulo 1, se describirán los tratamientos matemáticos que se utilizarán para el desarrollo del trabajo, la forma en la que se generará la base de datos y finalmente se indicará el software que se utilizará.

## 2.1. TRATAMIENTOS MATEMÁTICOS

Se refiere a los algoritmos y funciones matemáticas que se utilizarán, sobre la base de datos, para el desarrollo del presente trabajo.

### 2.1.1. PRETRATAMIENTOS DE DATOS

Los espectros que se obtienen en FTIR son únicos para cada medición realizada, por lo cual es necesario realizar el pre-tratamiento de los espectros antes de intentar desarrollar cualquier tipo de tratamiento matemático para el desarrollo del modelo. La utilización del pre-tratamiento de datos, debe ser hecha de tal forma que se pueda realizar una interpretación adecuada de la información original de los espectros obtenidos de las mediciones de muestras o patrones (Fearn, The effect of spectral pre-treatments on interpretation).De todas

estas opciones de pre-tratamientos no se puede asegurar que una sea mejor que otra, por lo que es necesario probar cual es la que mejor se adapta a los datos obtenidos, y de esta forma desarrollar cualquier tipo de cálculo (Fearn, The effect of spectral pre-treatments on interpretation).

Los pre-tratamientos de datos utilizados, se realizaron en el software OMNIC, (Thermo Electron Corporation), el cual está desarrollado para el tratamiento de espectros. Siendo los pre-tratamientos generales utilizados, la corrección de la línea de base, la suavización del espectro y la normalización de la escala.

#### 2.1.1.1. CORRECCIÓN DE LA LINEA DE BASE

La función Automatic Baseline Correct, del software OMNIC, (Thermo Electron Corporation), consiste en realizar primero un ajuste cuadrático del espectro *Y(x)*, utilizando mínimos cuadrados lineales

$$Y\left(x\right) \sim Y'\left(x\right) = ax^2 + bx + c \quad (1)$$

A continuación se encuentra la diferencia máxima entre *Y(x)* e *Y'(x)*. Todos los puntos de *Y(x)*, que difieran en más de, la diferencia máxima divida entre 2, de sus correspondientes puntos *Y'(x)*, se eliminan de *Y(x)*, lo que genera un subconjunto del espectro con la parte faltante de los picos principales.

Esta iteración, se repite 20 veces, sucesivamente sobre la iteración anterior. Como resultado los puntos de *Y'(x)* resultantes, se restan de los puntos *Y(x)* originales, para obtener la línea de base corregida del espectro.

El espectro final se obtiene compensado este con su punto mínimo, siendo el mínimo el nuevo punto cero. En la Figura 1, se muestra el espectro de la fructosa, con su respectiva corrección de línea de base. La gráfica roja corresponde al espectro original, y la gráfica azul al espectro corregido su línea de base.

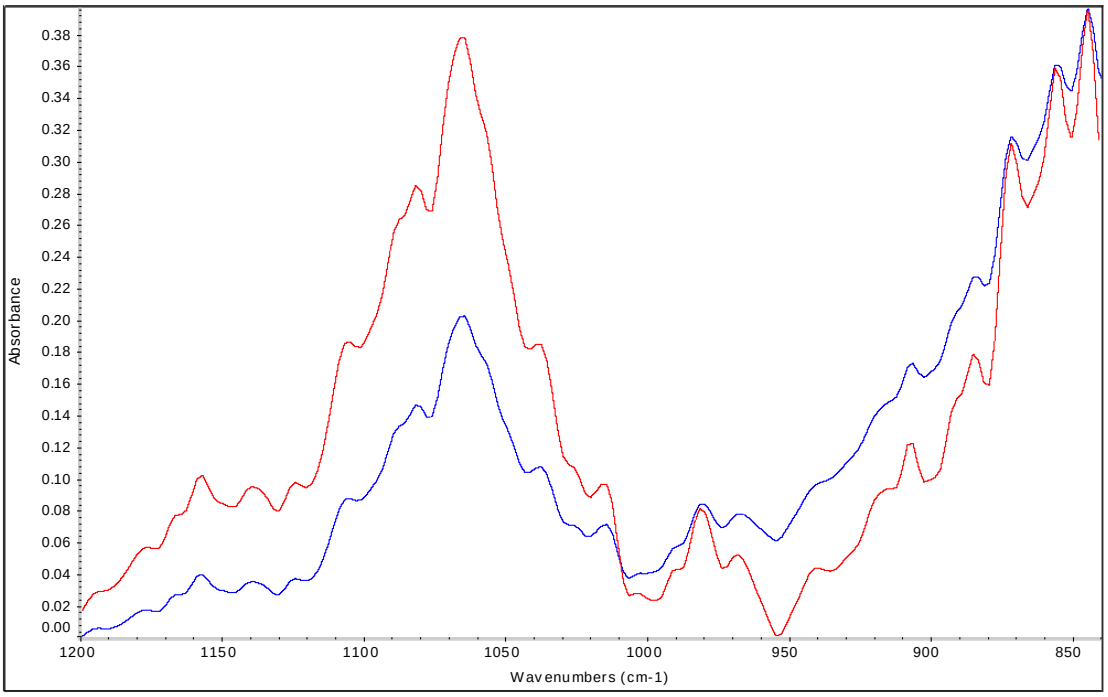


Figura 1. Corrección de la Línea de base del espectro infrarrojo de la fructosa

#### 2.1.1.2. SUAVIZACION DEL ESPECTRO

La función Automatic Smooth, del software OMNIC, (Thermo Electron Corporation), elimina los datos de alta frecuencia del espectro.

Para ello se utiliza el método propuesto por Savistzky-Golay para suavizar señales. El cual consiste en calcular una regresión polinomial local, de grado n, utilizando por lo menos n+1 puntos del espectro, con la cual se calculan los nuevos puntos del espectro.

La función Automatic Smooth, trabaja primero utilizando el filtro de Savistky-Golay con 15 puntos, generando un espectro Y', y posteriormente calculando dos factores de escala (SF) ecuaciones 2 y 3.

$$SF_1 = \frac{K}{\frac{\%T * 2.302585}{100}} \qquad (2)$$

Siendo *K* la ecuación de frecuencia, que depende del rango en el cual se hace la suavización. Las ecuaciones de *K* se puede observar en la tabla 1.

Tabla 1 Ecuaciones de frecuencia

| **Rango** | **K** |
|---|---|
| **> 4000** | 3.5 + (Freq-4000)*0.0115 |
| **4000 - 2000** | 1.0 + (Freq-2000)*0.00125 |
| **2000 - 1200** | 1.0 |
| **1200 - 600** | 1.0 + (1200-Freq)*0.00417 |
| **< 600** | 3.5 + (600-Freq)*0.0825 |

Fuente: OMNIC, (Thermo Electron Corporation)

$$SF_2 = Ruido * SF_1 \quad (3)$$

El ruido en la ecuación 3, es el ruido estimado del espectro. Se calcula la diferencia entre el espectro original y el espectro suavizado y un factor de adaptabilidad (*AF*), que depende de la comparación, del valor del cuadrado de la diferencia entre el espectro original y espectro suavizado con el filtro de Savistky-Golay (*Diff²*), y el valor del factor $SF_2$ entonces:

Si $SF_2 > Diff^2$ entonces $AF = 1.0 - \frac{SF_2}{Diff^2}$

Si $SF_2 < Diff^2$ entonces $AF = 0$

El espectro suavizado definitivo *Y"* se obtiene como e muestra a continuación

$$Y'' = Y' + Af * Diff \quad (4)$$

En la figura 2, se muestra un espectro al cual, se ha corregido su línea de base y se ha suavizado. La grafica roja corresponde al espectro de la fructosa, y la gráfica verde a la corrección de línea de base y al suavizado de la señal.

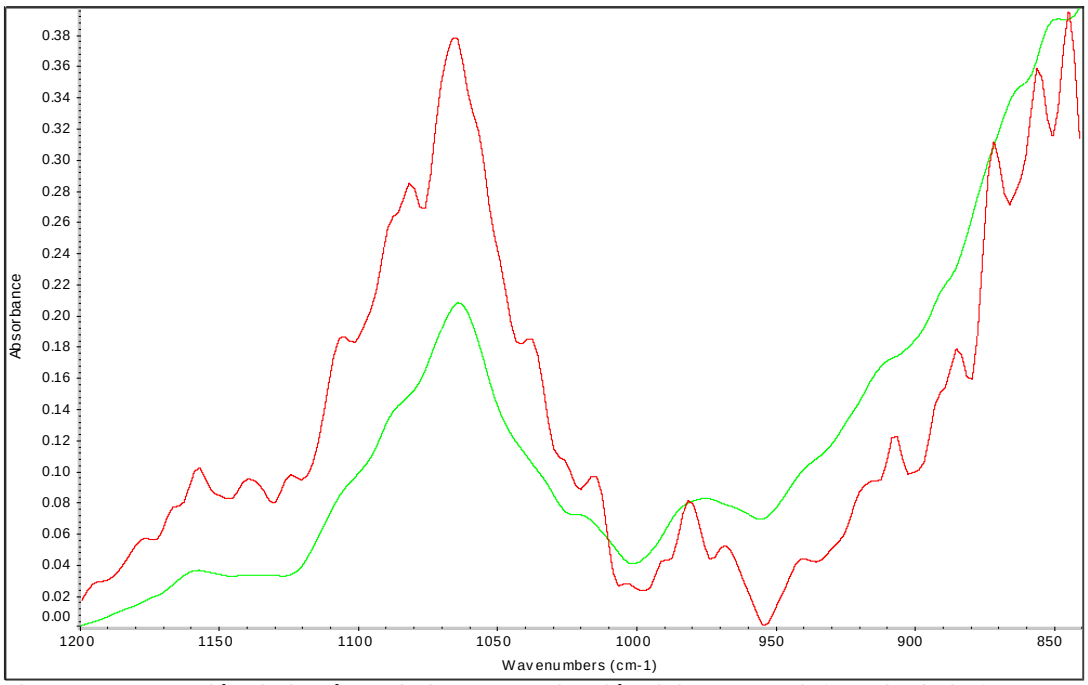

Figura 2. Corrección de la Línea de base, suavización del espectro infrarrojo de la fructosa

#### 2.1.1.3. NORMALIZACION DE LA ESCALA

La normalización de la escala busca tener los datos de absorbancia, de un espectro, dentro de un rango que va de 0 a 1, siendo 0 el punto más bajo y 1 el punto más alto. Para esto, se

debe relacionar la absorbancia a una longitud de onda determinada, con los valores máximo y mínimo de esta.

$$Normalizado = \frac{(Abs - AbsMin)}{(AbsMax - AbsMin)} \quad (5)$$

Luego de haber aplicado estos pre-tratamientos generales, se optó por calcular las derivadas de los espectros, debido a que tanto la primera como la segunda derivada transforman el espectro original, evidenciando información del espectro que no puede verse en el espectro original. (Fearn, A look at some standard pre-treatments for spectra). En la figura 3, se observa un espectro al cual se han aplicado, corrección de la línea de base, suavización y normalización de la escala. La gráfica de color violeta corresponde al espectro original y la gráfica roja al espectro ya tratado.

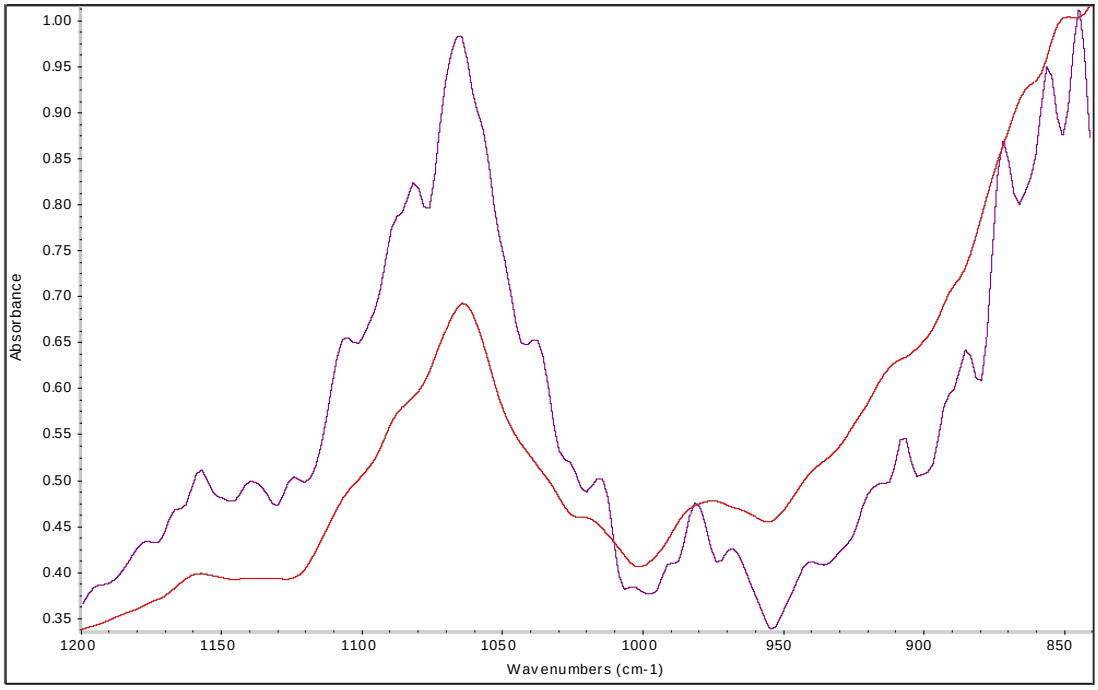


Figura 3. Corrección de la Línea de base, suavización y normalización de la escala del espectro infrarrojo de la fructosa

#### 2.1.1.4. DERIVADAS

Tanto la primera como la segunda derivada transforman el espectro original. Evidenciando información del espectro que no puede verse en el espectro original. La primera derivada al ser la pendiente de la curva, permite tener un máximo donde la pendiente es más grande y cero, donde existe un pico o un valle. Mientras la segunda derivada, al ser la pendiente de la primera derivada, se observa como inversión del espectro original, lo que ayuda a evidenciar los picos que no están bien definidos o la información complementaria en picos complejos (Fearn, A look at some standard pre-treatments for spectra).

#### 2.1.1.4.1. PRIMERA DERIVADA

Para obtener la primera derivada de un espectro se deriva la absorbancia en función del número de onda.

$$D_i = \frac{Y_{i-1} - Y_{i+1}}{X_{i-1} - X_{i+1}} \quad (6)$$

Donde $D_i$ corresponde a la derivada en cada punto *i,* $Y_i$ es la absorbancia en cada punto *i* y $X_i$ es la ubicación de cada punto i. (Thermo Electron Corporation)

#### 2.1.1.4.2. SEGUNDA DERIVADA

La segunda derivada se obtiene aplicando la ecuación 7

$$D'_i = \frac{Y_{i-2} - 2*Y_i + Y_{i+2}}{\left(X_{i-1} + X_{i+1}\right)^2} \quad (7)$$

Donde $D'_i$ es la segunda derivada en cada punto *i,* $Y_i$ es la absorbancia en cada punto *i* y $X_i$ es la ubicación de cada punto *i*. (Thermo Electron Corporation)

En la figura 1 se muestra el espectro de una solución acuosa de fructosa, junto con su primera y segunda derivadas.

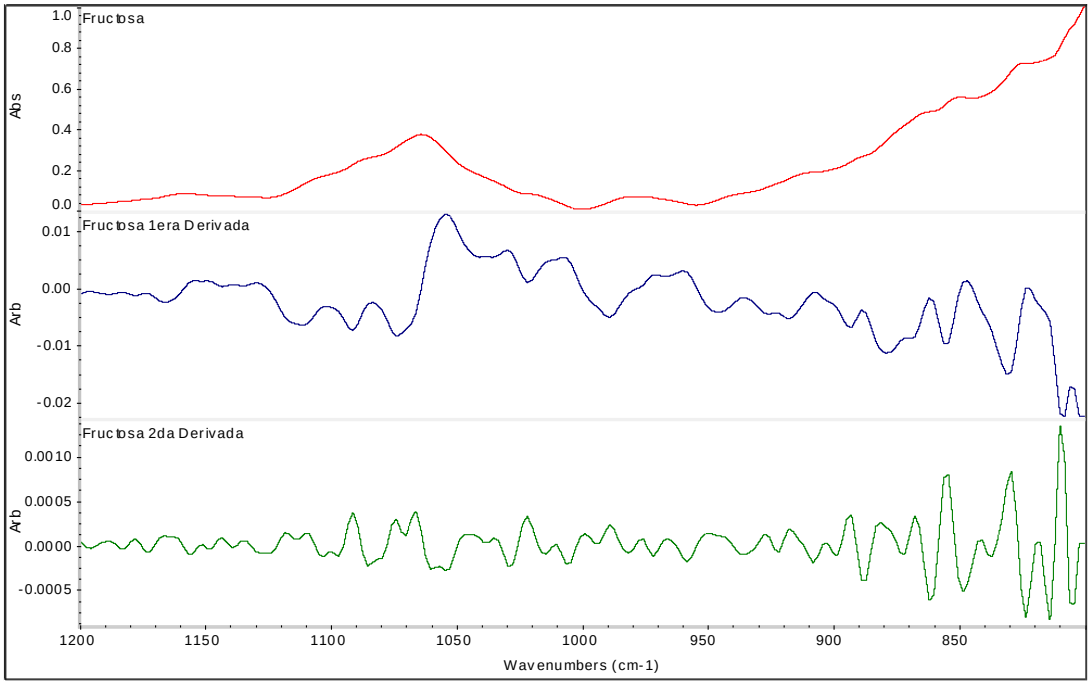


Figura 4 Espectro de la Fructosa junto a su primera y segunda derivadas

#### 2.1.1.4.3. FILTRO DE NORRIS

El filtro de Norris, consiste en derivar datos continuos en intervalos de segmentación, utilizando filtros móviles. Este filtro puede aplicarse a derivadas de 1er a 4to orden. Estos

filtros móviles consisten en ceros y unos, normalizados por un factor *nf*, que es un factor de normalización El filtro se determina por el orden de la derivada que se va aplicar, el tamaño del segmento y el intervalo entre segmentos. (K. Norris, Instrument Design. Interactions among instrument band pass, instrument noise, sample-absober bandwidth and calibration erro) (K. Norris, Applying Norris Derivatives. Understanding and correcting the factors which affect diffuse transmittance spectra) (Hopkins)

El filtro F, se define como:

$$F = \frac{1}{nf * F'} \tag{8}$$

Donde *F´* tiene la siguiente forma para las derivadas 1era y 2da:

1era: [- unos (1,*s*) ceros (1,*g*) unos (1,*s*) ]

2da: [unos (1,*s*) ceros (1,*g*) – 2* unos (1,*s*) ceros (1,*g*) unos (1,*s*)]

Siendo,

*s* = el tamaño del segmento compuesto de la cantidad de unos presentes, este valor debe ser impar cuando se aplica 2da y 4ta derivada.

*g* = el intervalo de los segmentos, compuesto por la cantidad de ceros presentes, este valor debe ser impar para la 1era y 3era derivada.

*len* = es el número de puntos del filtro, debe ser menor al número total de los datos.

*nf* = el factor de normalización.

El factor de normalización, para las derivadas 1era y 2da corresponde a:

$$nf1 = \sum (j * F'_{i+j}) \tag{9}$$

$$nf2 = \frac{1}{2} \sum (j^2 * F'_{i+j}) \tag{10}$$

Donde $j = -m,\ldots,m$ y $m = \frac{(len-1)}{2}$

Para la aplicación del filtro de Norris se utilizó un valor de *s* = 5 y *g* = 5, que es el valor predeterminado del software OMNIC.

#### 2.1.1.4.4. FILTRO DE SAVITSKY GOLAY

El algoritmo de Savitzky-Golay, se utiliza para suavizar y para derivar espectros. El filtro de Savitsky-Golay se basa en la realización de una regresión lineal de mínimos cuadrados, en

forma de, polinomio denotado como $p_i$ (x), de grado M, el cual se aplicará sobre $n_L + n_R + 1$ puntos. De donde el punto suavizado corresponde a la expresión, $g_i=p_i(x_i)$ y es el punto medio entre $n_L + n_R + 1$ (Gander y von Matt).

$$p_i(x) := \sum_{k=0}^{M} b_k \left( \frac{x - x_i}{\Delta x} \right)^k \tag{11}$$

La ecuación 8, muestra la forma del polinomio $p_i(x)$, donde asumimos que las abscisas $x_i$, están espaciadas de forma uniforme donde $x_{i+1} - x_i = \Delta x$. La determinación de los coeficientes $b_k$ responde a la siguiente estructura:

$$\sum_{j=i-n_R}^{i+n_R} (p_i(x_j) - f_j)^2 = \min \tag{12}$$

La derivada es entonces, la derivada del polinomio ajustada en cada punto.

Los parámetros de aplicación del filtro de Savitsky Golay fueron: $\Delta x = 7$ y $M = 3$.

### 2.1.2.SELECCION DE VARIABLES

Los métodos de selección de variables, se han desarrollado con el fin de conseguir la mayor información posible, de un sistema compuesto por muchas variables, utilizando pocas variables del sistema. Estos métodos se pueden aplicar a casos de regresión y clasificación. En el caso de este trabajo se indicarán las características de estos métodos aplicados a modelos de regresión. Para los modelos de regresión es muy importante que, se hayan eliminado aquellas variables que tengan información irrelevante para el modelo y/o aquellas variables que presenten ruido que conlleven a obtener falsa información en la predicción.

#### 2.1.2.1. ESCALADO DE VARIABLES

El escalado de variables consiste en realizar una transformación simple de los elementos de una matriz, que tiene como objetivo obtener datos cerrados. Estos elementos transformados, son variables que no tienen dependencia de una unidad de medida. El tipo de escalado que se utilice tendrá influencia en los estimadores que se calculen, que se verá reflejado en los desarrollos posteriores, (Frank y Todeschini), en este caso el cálculo de modelos de regresión.

##### 2.1.2.1.1. AUTOSCALING

Es uno de los métodos de escalado de columna más utilizados, está compuesto de una columna centrada ( $\overline{x}_j$ ) y una columna escalada ( $x_{ij}$ ) así:

$$x'_{ij} = \frac{(x_{ij} - \overline{x}_j)}{s_j} \tag{13}$$

La media de una variable autoescalada es 0 y la varianza es 1, una variable autoescalda también recibe el nombre de variable estandarizada (Frank y Todeschini).

#### 2.1.2.1.2. CENTERING

Este método de escalado, se obtiene restando a cada punto una constante, la media, teniendo como resultado en la media un valor de 0 (Frank y Todeschini). El centrado en columnas se expresa de la siguiente forma:

$$x'_{ij} = x_{ij} - \bar{x}_j \quad \text{donde} \quad \bar{x}_j = \frac{\sum_i x_{ij}}{n} \tag{14}$$

### 2.1.2.2. ALGORITMOS GENÉTICOS (GA)

Se basan en las reglas de la evolución biológica de los organismos vivientes, este método consiste en crear una población de individuos (regresiones lineales), los que luego se reproducirán apareándose, mutando o transfiriendo directamente la información, lo que provoca una evolución a través de generaciones sucesivas hasta alcanzar una solución final óptima, en este caso son las variables a utilizar (Niazi y Leardi).

Leardi (Leardi, Boggia y Terrile), plantea que para el desarrollo de GA, se deben seguir 5 pasos básicos que son: codificación de variables, creación de la población de partida, evaluación de las respuesta, reproducción y mutación. Los pasos 3 al 5, se repiten las veces que sean necesarias en función de tres consideraciones: llegar a la respuesta más óptima posible, cumplir un determinado número de generaciones o cumplir con un tiempo determinado inicialmente

#### 2.1.2.2.1. CODIFICACIÓN DE VARIABLES

Los cromosomas, son codificados en forma binaria, de esta forma cada cromosoma está representado por una serie de ceros y unos. Cada cromosoma tiene un número de genes específico y cada gen está constituido por un solo bit, lo que significa que cuando el valor es 0 significa ausencia de la variable, y 1 significa la presencia de la variable.

#### 2.1.2.2.2. CREACIÓN DE LA POBLACIÓN DE PARTIDA

La población original se compone de N número de cromosomas (generalmente entre 50 y 500 dependiendo de la complejidad del problema). En esta población cada bit de cada cromosoma, toma un valor aleatorio. La probabilidad de que una variable esté presente en los cromosomas de la población de partida se define como:

$$p = \frac{n}{v} \tag{13}$$

Donde *n* es el número de unos (1) que queremos en cada cromosoma y *v,* es el número total de variables. Por tanto, *p* puede ser considerado un vector de cuyos elementos *v* tienen el mismo valor.

#### 2.1.2.2.3. EVALUACIÓN DE LAS RESPUESTAS

Se evalúa cada cromosoma, con su respuesta experimental correspondiente. Cuando las condiciones experimentales, no se encuentran en el dominio experimental o corresponden a un experimento no ejecutable, se pueden considerar como respuestas nulas. De esta forma los cromosomas no pasarán a la siguiente generación. La evaluación se hace utilizando el coeficiente de correlación cross validado, simbolizado como $R^2_{cv}$ o $Q^2$, que se detallarán en la sección 2.1.4.

#### 2.1.2.2.4. REPRODUCCIÓN

Este paso tiene como objetivo la creación de una población de cromosomas, la cual sería la generación producto de la población de partida. Este proceso tiene dos sub-pasos que son: selección-copia y entrecruzamiento.

##### 2.1.2.2.4.1. SELECCIÓN-COPIA

En la selección copia el operador selecciona N veces un cromosoma de la población, teniendo los mejores cromosomas la mayor probabilidad de ser elegidos, sobre los peores. Esta probabilidad de selección se basa en una función de respuesta asociada de la forma:

$$p(i) = \frac{response\ (i)}{\sum responses} \tag{14}$$

Leardi (R. Leardi), considera que, cuando los GA, son aplicados a datos espectroscópicos, la frecuencia con que las variables han sido seleccionadas en una generación, permite modificar el vector p de la siguiente generación, de tal forma que, los valores de los elementos, que corresponden a las variables que han sido seleccionadas con más frecuencia sea mayor que las que tienen menos frecuencia de ser seleccionadas de la siguiente forma:

$$p_i = n * \frac{sel_i}{\sum_{j=i}^{v} sel_j} \tag{15}$$

Donde $sel_j$ es el número de selecciones de la variable *j* en las generaciones previas.

Para empezar una nueva generación, en la creación de la población inicial, se selecciona un número aleatorio para cada una de las variables *v* y se compara con el valor correspondiente del vector *p*. Si el valor es menor entonces el bit recibe el valor de 1, caso contrario se adjudica el valor de 0. Por consiguiente si el valor de $p_i$ es alto, la probabilidad de que la variable *i*, esté presente en el cromosoma es mayor.

Es necesario considerar que esta solución tiene dos problemas principales que son:

- No considera la auto-correlación que existe entre las longitudes de onda adyacentes
- Las variables que no fueron seleccionadas en generaciones anteriores, tendrán un vector *p*=0.

Debido a que las variables espectroscópicas tienen una auto-correlación elevada, las variables adyacentes a una variable *v* relevante, tienen que ser relevantes también, siendo lógico que esto aumente también su probabilidad de ser seleccionadas. La autocorrelación entre las longitudes de onda, puede resolverse suavizando el espectro aplicando una media móvil, lo que daría por resultado un nuevo vector *ps.*

El segundo problema, donde la frecuencia de la selección de variables, depende del número de generaciones evolucionadas; se resuelve aplicando una media ponderada entre el vector *p* original y él vector $p_i$, que se calcula de la siguiente forma:

$$pf_i = \frac{\frac{n}{v} * (R - r) + ps_i * r}{R} \tag{16}$$

Donde $pf_i$, es la probabilidad final de que la variable *i*, esté presente en un cromosoma de la población inicial, *R* es el número total de generaciones que se realizarán, *r* es el número de generaciones ya transcurridas y $ps_i$ es la probabilidad de *i* este presente después del suavizado.

#### 2.1.2.2.4.2. ENTRECRUZAMIENTO

Con los N cromosomas que forman la población original se conforma una nueva población pareada aleatoriamente de tamaño N/2. De esta población de pares (padres) se generan nuevos individuos (descendientes), estos descendientes mantendrán las características comunes de sus padres y las características no comunes se mezclaran en función de la probabilidad de entrecruzamiento.

Padre 1: 1 0 1 1 0 0 1 1

Padre 2: 1 0 0 1 1 0 1 1

Cada descendiente de este entrecruzamiento tendrá un cromosoma base así:

Comosoma: 1 0 ? 1 ? 0 1 1

La generación de descendientes se crea utilizando, la información de un padre por vez, la generación de números aleatorios y la probabilidad de entrecruce. Cuando el valor del número aleatorio es menor a la probabilidad de entrecruce la variable presente en el padre será excluida, considerando al padre 1, el gen 3 será 0. Cuando el valor del número aleatorio es mayor a la probabilidad de entrecruce, el gen 3 será 1 (Todeschini, Consonni y Mauri, MobyDigs: Software for Regression and Classification Models by Genetic Algorithms).

#### 2.1.2.2.5. MUTACIÓN

En este último paso, algunos genes serán seleccionados aleatoriamente y cambiarán de 1 a 0 y viceversa. De forma similar al entrecruzamiento, los genes del cromosoma, que aleatoriamente, se han escogido que pueden mutar, se evalúan con números aleatorios generados al azar y la probabilidad de mutación. La probabilidad de mutación es muy baja, por lo general es del 1%. Este paso produce genes que no habían sido probados antes, en un espacio experimental diferente. (Leardi, Boggia y Terrile)

Los parámetros utilizados para el desarrollo de los Algoritmos Genéticos fueron:

- Método de análisis: OLS o PLS
- Escalado de las variables: autoscaling o centering
- Número de Generaciones: 100
- Grupos para la validación cruzada: 10
- Tipo de validación cruzada: ventanas venecianas
- Número de cromosomas de la población de partida: 30
- Variables por cromosoma en la población inicial: 5
- Cantidad máxima de variables: 30
- Probabilidad de mutación: 0.01 (1%)
- Probabilidad de cruzamiento: 0.5
- Número de corridas: 100

### 2.1.2.3. ALGORITMO DE REEMPLAZO SECUENCIAL RENOVADO (RSR)

El algoritmo de Remplazo Secuencial Renovado (RSR), fue desarrollado por Cassotti et al. (Cassotti, Grisoni y Todeschini), utilizando como base el algoritmo de Reemplazo Secuencial (SR) propuesto por Miller (Miller), al cual se han adicionado 4 funcionalidades que son:

a) Disminución del tiempo de cálculo.
b) Introducción de herramientas de validación, sobre la capacidad predictiva de los modelos.
c) Aumento de la probabilidad de converger al mejor modelo.
d) Identificar modelos con patologías (superposición, correlación casual, redundancia de variables y colinealidad de predictores).

El esquema propuesto por Cassotti et al. (Cassotti, Grisoni y Todeschini), para el desarrollo del algoritmo RSR se muestra en la figura 1.

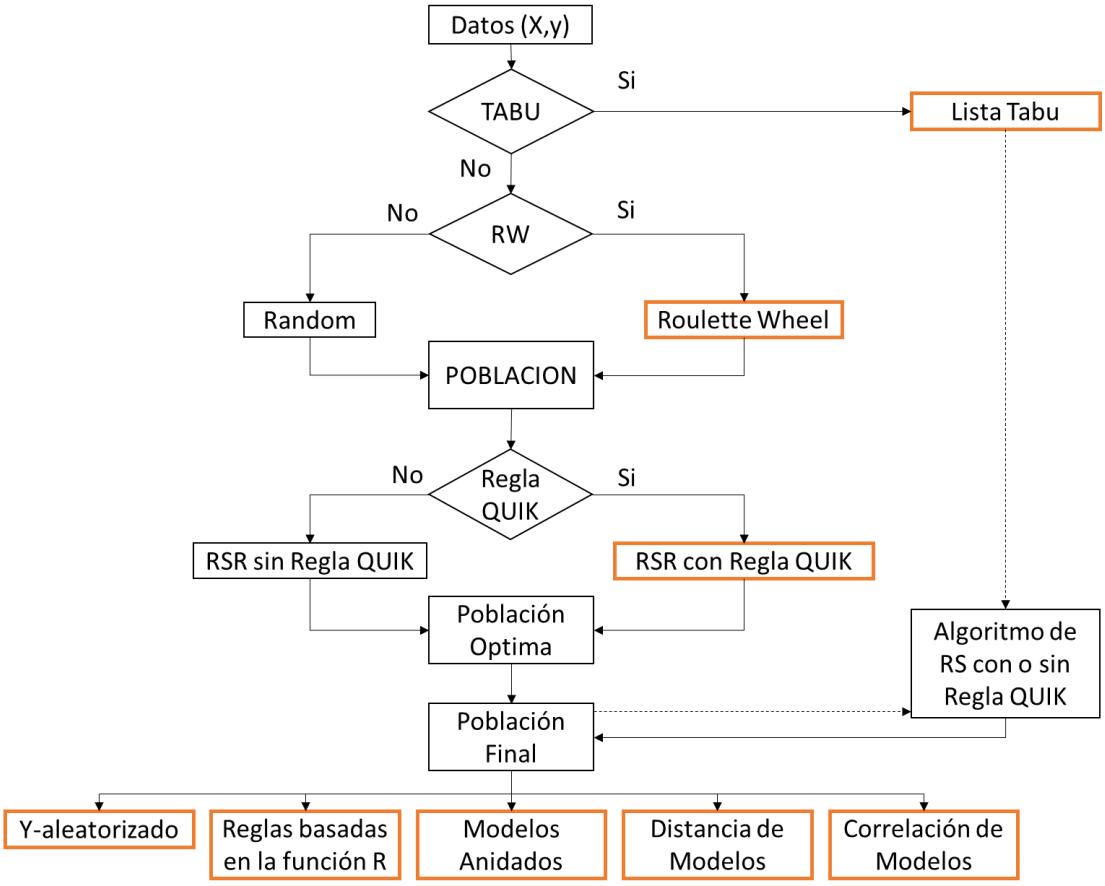


Figura 5. Secuencia del desarrollo del algoritmo de Remplazo Secuencial Renovado.

Fuente: (Cassotti, Grisoni y Todeschini)

#### 2.1.2.3.1. FUNCION DE EVALUACION (FITNESS FUNCTION)

La función de evaluación que utiliza el algoritmo RSR, es el el coeficiente de correlación cuadrático predictivo ($Q^2_{cv}$), que presenta ventaja en comparación a la suma de los cuadrados de los residuos, que es la función utilizada por el algoritmo SR. El coeficiente $Q^2_{cv}$ se explicará en la sección 2.1.4.

#### 2.1.2.2.2. LISTA TABU (LT)

Cuando existe una gran cantidad de variables en un sistema, puede que varias de ellas no tengan información relevante que lleve al desarrollo de un buen modelo. La LT, contiene estas variables que no aportarían positivamente en la calidad del modelo. Para que una variable sea considerada irrelevante, y forme parte de la LT, el modelo univariante de ésta, tendría un valor de $Q^2_{cv}$ negativo, así

$$Q^2(y,x) < 0 \quad \Rightarrow \quad x \in \mathrm{LT}$$

Las variables de la LT, se pueden utilizar, en el modelo final, con la posibilidad de mejorar el modelo.

#### 2.1.2.2.3. REGLA DE LA RULETA (ROULETTE WHEEL (RW))

La regla de la ruleta, es un algoritmo de selección sesgado, que obtiene soluciones de alta calidad. Este algoritmo se utiliza para la selección de la población inicial de semillas, donde las variables que tienen los mejores modelos univariantes tienen mayor posibilidad de estar presentes en esta población inicial. La calidad de los modelos está dada por el valor de $Q^2_{cv}$, solo las variables con $Q^2_{cv}$ positivo son consideradas dentro de la población inicial, y aquellas con un valor alto de $Q^2_{cv}$ tienen mayor probabilidad de ser seleccionadas al inicio.

#### 2.1.2.2.4. REGLA QUIK

La regla QUIK es una prueba estadística que facilita rechazar los modelos, en los que sus predictores tienen alta colinealidad. Se basa en la correlación multivariante *K* que mide la correlación general de un conjunto de variables. El índice *K* se define como:

$$K = \frac{\sum_{j=1}^{p} \left| \left( \lambda_j / \sum_j \lambda_j \right) - (1/p) \right|}{2*(p-1)/p} \qquad (17)$$

Donde $\lambda_j$ son los valores propios de la matriz de correlación y *p* es el número de variables.

Se compara la correlación total de las variables *X* ($K_X$) y la correlación entre el bloque *X* y la respueta ***y*** ($K_{XY}$).

Si $K_{XY} - K_X < \delta K \rightarrow$ se rechaza el modelo.

Donde $\delta K$ es definido por el usuario generalmente entre 0.01 y 0.05. Mientras más grande sea este valor, más estricto será el criterio, por lo tanto la cantidad de modelos rechazados será mayor.

#### 2.1.2.2.5. Y-ALEATORIZADO (Y-SCRAMBLING)

Es un test estadístico que se utiliza, en general, para identificar la posible presencia de correlación casual entre la respuesta ***y***, y los predictores *X*. Para realizarla se aleatoriza el vector ***y***, con el objetivo de que los objetos de la matriz X, ya no se encuentren asociados con su correcta respuesta en el vector ***y***. Cada modelo se ajusta y se valida con este vector aletorizado y se calculan los parámetros estadísticos. Este proceso se repite muchas veces y se promedian los parámetros estadísticos calculados.

Al tener un vector ***y***, alterado se espera obtener “malos” modelos calculados, en el caso de que se obtenga algún modelo de calidad comparable con los modelos reales calculados, el modelo real debe de ser rechazado debido a la probable presencia de una correlación casual.

#### 2.1.2.2.6. REGLAS BASADAS EN LA FUNCION R

La función R, refiere a la función de correlación estadística. El estudio de esta función ha llevado al desarrollado varias funciones de evaluación, que evalúan los modelos en relación de la correlación. En este caso, el objetivo de las reglas basadas en la función *R*, es encontrar "malos" modelos provocados por dos situaciones:

a) Modelos que pueden presentar redundancia entre las variables explicativas. Considerado como un exceso de "buenos" predictores
b) Modelos que contienen variables ruidosas. Considerado como un exceso de "malos" predictores.

Las reglas utilizadas para detectar estos modelos se introdujeron basadas en las funciones $R^P$ y $R^N$, (Todeschini, Consonni y Mauri, Detecting "bad" regression model: multicriteria fitness functions in regression analysis) estos índices están definidos en términos de la cantidad $M_j$, que determina el rol de cada variable y está definido como:

$$M_j = \frac{R_{jy}}{R} - \frac{1}{p} \qquad -\frac{1}{p} \le M_j \le \frac{p-1}{p} \tag{18}$$

Donde $R_{jy}$ es el valor absoluto del coeficiente de correlación entre la jotaesima variable y la respuesta y. *R* es el coeficiente de correlación del modelo, *p* es el número de variables del modelo.

El coeficiente $R^P$, se calcula de la siguiente forma:

$$R^P = \prod_{j=1}^{p^+} \left(1 - M_j\left(\frac{p}{p-1}\right)\right) \quad \forall M_j > 0 \quad \text{y} \quad 0 \le R^P \le 1 \tag{19}$$

Donde el producto en $R^P$, en las corridas $p^+$, dan un valor positivo de $M_j$. $R^P$, identifica la redundancia de variables explicativas de la siguiente forma

Si $R^P < t^P \rightarrow$ se rechaza el modelo

Donde $t^P$ es un umbral definida por el usuario entre 0.01 y 0.1, dependiendo de los datos. Mientras más bajo es el umbral más estricta es la prueba.

Cuando los valores de $M_j$ son negativos, Todeschini et al (Todeschini, Consonni y Mauri, Detecting "bad" regression model: multicriteria fitness functions in regression analysis), proponen que, l la suma sea realizada de la siguiente forma:

$$R^N = \sum_{j=1}^{p^-} M_j \qquad \forall M_j < 0 \quad \text{y} \quad -1 \le R^N \le 0 \tag{20}$$

Entonces $R^N < t^N \rightarrow$ se rechaza el modelo.

Donde $t^N$ es un umbral definida por el usuario entre -0.01 y -0.1, siendo una prueba muy estricta Cassotti et al, proponen realizar la prueba sobre cada valor negativo, no como suma sino como valor así:

$\forall M_j < 0, \quad \exists M_j < t^N \quad \rightarrow$ se rechaza el modelo.

El umbral $t^N$, depende de un parámetro ajustable $\varepsilon$ definido por el nivel de ruido de la respuesta ***y,*** altos valores de $\varepsilon$ corresponden a altos valores de $t^N$.

Las reglas basadas en la función R, se calculan únicamente en la población final de los modelos de regresión (cuando se ha utilizado mínimos cuadrados ordinales).

#### 2.1.2.2.7. MODELOS ANIDADOS

Un modelo *F* se puede considerar anidado si un modelo *G,* con mayor número de variables, tiene las mismas variables que F y su rendimiento es muy similar, es decir, que la diferencia de predicción del modelo es menor a un umbral (*thr*) definido anteriormente:

$$\left|Q^2(G) - Q^2(F)\right| \leq thr \tag{21}$$

En consecuencia, el modelo *G* se rechaza pues su mayor complejidad no justifica su rendimiento. El umbral que se sugiere para modelos anidados es de 0.05.

#### 2.1.2.2.8. DISTANCIA Y CORRELACION DE LOS MODELOS

Estas dos condiciones, se aplican a la población de modelos finales, para compararlos entre sí, puesto que cuando se calculan modelos utilizando métodos de selección de variables, se obtienen modelos con buena capacidad de predicción y diferentes variables. Con esto se puede identificar si los modelos obtenidos son realmente diferentes o no, a pesar de tener diferentes variables.

Esta comparación se hace, calculando los índices Canónicos de Medida de Distancia (CMD) y de Correlación (CMC) (Todeschini, Ballabio y Consonni). Estos índices se calculan utilizando las siguientes fórmulas:

$$CMD_{AB} = p_A + p_B - 2 \cdot \sum_{j=1}^{M} \sqrt{\lambda_j} \qquad 0 \leq CMD_{AB} \leq (p_A + p_B) \tag{22}$$

$$CMC_{AB} = \frac{\sum_{j=1}^{M} \sqrt{\lambda_j}}{\sqrt{p_A \cdot p_B}} \qquad 0 \leq CMC_{AB} \leq 1 \tag{23}$$

Donde *A* y *B* son los dos conjuntos de variables que se van a comparar, $p_A$ y $p_B$ son el número de variables de grupo *A* y *B* respectivamente, $\lambda$ son los valores propios de la matriz simétrica de correlación cruzada y *M* es el número de valores propios diferentes de 0.

La matriz de correlación cruzada, es una matriz rectangular anti-simétrica de tamaño ( $p_A$ x $p_B$ ) y recoge las parejas correlacionadas de las variables de los dos conjuntos, definido por:

$$\mathbf{C_{AB}} = \left[ \boldsymbol{\rho}(\mathbf{x_i^A x_j^B}) \right], \qquad i = 1,..., p_A, \quad j = 1,..., p_B \tag{24}$$

Donde $\rho$ indica las parejas correlacionas entre las variables de los conjuntos *A* y *B*. En forma alternativa la matriz de correlación cruzada, puede definirse intercambiando filas por columnas, esto se puede hacer solo entre dos matriz correlacionadas cruzadas simétricas, tienen los mismos valores propios, diferentes de cero.

$$\mathbf{Q}_A = \mathbf{C_{AB}} \cdot \mathbf{C_{BA}} \qquad \text{o} \qquad \mathbf{Q_B} = \mathbf{C_{BA}} \cdot \mathbf{C_{AB}} \tag{25}$$

### 2.1.3.MODELOS MULTIVARIANTES DE REGRESION

#### 2.1.3.1. MÍNIMOS CUADRADOS ORDINALES (OLS)

Es el método de regresión más simple, y por ello de gran difusión a nivel general, debido a que relaciona de forma lineal parámetros y predictores. Este modelo se representa con la siguiente ecuación:

$$\boldsymbol{y} = \mathbf{Xb} + \mathbf{e} \tag{26}$$

Para calcular los coeficientes de regresión **b,** se debe resolver la ecuación normal

$$\hat{\boldsymbol{b}} = \left( \mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \boldsymbol{y} \tag{27}$$

la cual genera el estimador de mínimos cuadrados. Los coeficientes se minimizan con la suma de los cuadrados de los residuos, como se aprecia en la ecuación 12, que es equivalente a maximizar la correlación entre la combinación lineal de los predictores con la respuesta indicado en la ecuación 13 (Frank y Todeschini).

$$\min_b \left[ \sum_i e_i^2 \right] = \min_b \left[ \sum_i \left( y_i - \boldsymbol{b} x_i^T \right)^2 \right] \tag{28}$$

$$\max_b \left[ \mathrm{corr}^2(\mathrm{y}, \mathrm{bx_i^T}) \right] \tag{29}$$

#### 2.1.3.2. MÍNIMOS CUADRADOS PARCIALES (PLS)

Este es el método de regresión más aplicado en química. Es un método de regresión sesgado que relaciona, un conjunto de, *p*, variables predictoras $\mathbf{X}$ (n, p), con un conjunto de, r, variables respuesta $\mathbf{Y}$ (n, r) (Wold, Sjöström y Eriksson). PLS calcula de forma iterativa pares de variables latentes, denominadas t y u, a partir de los conjuntos de variables $\mathbf{X}$ e $\mathbf{Y}$, buscando siempre que la correlación entre estas sea la máxima posible.

Donde

$$t_m = \mathbf{X}w_m^T \qquad \text{y} \qquad u_m = \mathbf{Y}q_m^T \tag{30}$$

$$\max_m = [\text{corr}^2(u_m, t_m), \quad \text{var}(u_m)\,\text{var}(t_m)] \qquad \text{con} \qquad m= 1,M \tag{31}$$

El número M de las variables latentes es estimado mediante validación.

La forma de los modelos PLS es la siguiente:

$$\mathbf{Y} = \mathbf{TBQ} + \mathbf{E} \tag{32}$$

Siendo $\mathbb{T}$(n, M), el conjunto de variables latentes, $\mathbf{B}$ (M, M) la matriz diagonal que contiene los coeficientes mínimos cuadrados de las variables latentes, $\mathbf{Q}$ (M, r) contiene los pesos de las respuestas y E es la matriz de error (Frank y Todeschini).

Tanto el sesgo como la varianza son determinados en función del número de variables latentes M, de tal forma que cuando mayor sea el número de variables latentes, mayor será la varianza y menor el sesgo. Cuando el número de M es igual al número de variables predictoras p, el modelo PLS converge en un modelo OLS (Frank y Todeschini).

#### 2.1.3.3. REGRESIÓN DE COMPONENTES PRINCIPALES (PCR)

Es un método de regresión sesgado, que se basa en utilizar un número reducido de componentes principales, ***M***, que representan el espacio de los predictores, ***p***. Eliminando de esta forma la información que tenga mínima varianza, que se puede representar de la siguiente forma:

$$\boldsymbol{y} = \mathbf{X}\mathbf{L}_H\mathbf{L}_H^T\boldsymbol{b}^T + \boldsymbol{e} \tag{33}$$

Donde ***X*** es la matiz de datos original y $\boldsymbol{L_H}$ se denomina “loading matrix”, y es la matriz producto de los vectores propios y las variables originales.

Las componentes principales son combinaciones lineales ortogonales entre los predictores y los vectores propios, se calculan considerando el criterio de máxima varianza entre los vectores propios y los predictores cumpliendo que:

$$\max_m \left[\text{var}(l_m x)\right] \quad \text{con} \quad m=1,M \tag{34}$$

Los coeficientes del modelo PCR, se estiman utilizando la siguiente función

$$\hat{b} = L_H a = L_H \Lambda_H^{-1} L_H^T X^T y \quad (35)$$

Siendo ***a*** el coeficiente de la regresión de mínimos cuadrados sobre las componentes principales, ***Λ$_H$*** es el vector del valor propio el subespacio de alta varianza ( $\mathbf{X}^T\mathbf{X}$ ). De igual forma que en PLS, el sesgo y la varianza son determinados en función de M (Frank y Todeschini).

### 2.1.4. VALIDACIÓN DE MODELOS

En el desarrollo de trabajos estadísticos y quimiométricos, se ha generado la necesidad de evaluar la capacidad predictiva de los modelos. Esto se debe a que esta capacidad de predicción es diferente a la capacidad de evaluación cuando se desarrolla un modelo. El parámetro que se utilizará en este trabajo es el coeficiente de correlación cuadrático predictivo $Q^2$.

#### 2.1.4.1. COEFICIENTE DE CORRELACIÓN CUADRÁTICO PREDICTIVO Q²

El coeficiente de $Q^2$, se define como:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{PRESS}{TSS} \quad (36)$$

Donde n es el número total de objetos en todo el conjunto de datos, *TSS*, es la suma total de los cuadrados y *PRESS*, es la suma de los cuadrados de los errores de predicción. Este método es aplicable, utilizando los datos, sacando un dato a la vez (leave one out, LOO) o sacando varios datos a la vez (leave more out, LMO) en validación cruzada (Consonni, Ballabio y Todeschini).

Cuando se utilizan datos externos para la validación del modelo, estos datos no deben haber sido utilizados en el desarrollo del modelo, y deben ser independientes entre sí. Por consecuencia el valor de $Q^2$ obtenido de estos datos ($Q^2_{ext}$) y la media de los valores de $Q^2$ obtenidos tomado cada uno de los datos externos a la vez deberían coincidir, cumpliendo la propiedad ergódica (Consonni, Ballabio y Todeschini).

Así $$Q^2_{ext} = \frac{\sum_{i=1}^{n_{ext}} Q_i^2}{n_{ext}} \quad (37)$$

Donde $Q^2_{ext}$ es el $Q^2$ externo calculado teniendo en cuenta solo el i-ésimo dato del conjunto externo de evaluación, $n_{ext}$ es el número total de objetos que conforman el conjunto externo de evaluación.

## 2.2. BASE DE DATOS

La base de datos para este trabajo se obtuvo de la medición de diferentes muestras acuosas, compuestas por glucosa, fructosa, sacarosa y lactosa, en un espectrofotómetro infrarrojo THERMO Scientific Nicolet IR100 FT-IR, que trabaja en la región del MIR, del espectro infrarrojo, entre 7800 – 350 $cm^{-1}$, con una resolución espectral de 2 $cm^{-1}$, el ruido de la señal es menor a $4.3x10^{-5}$ unidades de absorbancia, utiliza una fuente de emisión MIR cerámica. El detector es de sulfato de triglecerina deuterado.

Estas muestras se prepararon siguiendo un diseño de mezclas, simplex-lattice con 4 componentes. El número de retículo utilizado es 3, con el cual se generan 19 experimentos. El tipo de punto (TipoPt), corresponde a la siguiente nomenclatura:

- 1 – vértice
- 2 – combinación doble
- 3 – combinación triple
- 0 – punto central
- -1 – punto axial

La concentración de cada azúcar presente en cada mezcla, está representada por la fracción de ésta, así en los vértices existirá la presencia de solo un azúcar y su fracción será 1, en una combinación doble existirán 2 azúcares con su fracción 0.5 respectivamente, en una combinación triple existirán 3 azúcares con su fracción 0.33 respectivamente, el punto central tendrá una combinación de los 4 azúcares y la fracción será 0.25, y en los puntos axiales estarán presentes los 4 azúcares, donde uno de ellos tendrá una presencia mayoritaria correspondiente a la fracción de 0.625, mientras que los 3 azúcares restantes estarán presente en una concentración de 0.125 cada uno. Los experimentos se pueden observar en la tabla 2

Tabla 2 Diseño de Experimentos

| Mezcla | TipoPt | Glucosa | Fructosa | Sacarosa | Lactosa |
|---|---|---|---|---|---|
| **1** | 1 | 1.000 | 0.000 | 0.000 | 0.000 |
| **2** | 1 | 0.000 | 1.000 | 0.000 | 0.000 |
| **3** | 1 | 0.000 | 0.000 | 1.000 | 0.000 |
| **4** | 1 | 0.000 | 0.000 | 0.000 | 1.000 |
| **5** | 2 | 0.500 | 0.500 | 0.000 | 0.000 |
| **6** | 2 | 0.500 | 0.000 | 0.500 | 0.000 |
| **7** | 2 | 0.500 | 0.000 | 0.000 | 0.500 |
| **8** | 2 | 0.000 | 0.500 | 0.500 | 0.000 |
| **9** | 2 | 0.000 | 0.500 | 0.000 | 0.500 |
| **10** | 2 | 0.000 | 0.000 | 0.500 | 0.500 |
| **11** | 3 | 0.333 | 0.333 | 0.333 | 0.000 |
| **12** | 3 | 0.333 | 0.333 | 0.000 | 0.333 |
| **13** | 3 | 0.333 | 0.000 | 0.333 | 0.333 |
| **14** | 3 | 0.000 | 0.333 | 0.333 | 0.333 |
| **15** | 0 | 0.250 | 0.250 | 0.250 | 0.250 |
| **16** | -1 | 0.625 | 0.125 | 0.125 | 0.125 |
| **17** | -1 | 0.125 | 0.625 | 0.125 | 0.125 |
| **18** | -1 | 0.125 | 0.125 | 0.625 | 0.125 |
| **19** | -1 | 0.125 | 0.125 | 0.125 | 0.625 |

Se utilizarán 4 niveles, para obtener 76 lecturas experimentales y de esta forma tener datos para hacer la validación externa del modelo. Las concentraciones de los azúcares correspondientes a 1 en diseño experimental, para cada uno de los niveles, se muestran en la tabla 3.

Tabla 3 Concentración de Azúcares a diferentes niveles

| Nivel | Concentración (g/100ml) | | | |
|---|---|---|---|---|
| | Glucosa | Fructosa | Sacarosa | Lactosa |
| **1** | 15.037 | 15.752 | 14.425 | 5.982 |
| **2** | 9.986 | 10.136 | 10.234 | 4.020 |
| **3** | 4.984 | 5.102 | 5.070 | 2.036 |
| **4** | 1.174 | 1.150 | 1.132 | 0.373 |

La concentración de Lactosa difiere del resto de azúcares, pues la máxima solubilidad es 10 veces menor a la de la sacarosa (Alais).

### 2.3. SOFTWARE UTILIZADO

Los softwares utilizados, para las diferentes etapas del trabajo fueron:

- Pre-tratamiento de Datos: OMNIC 7.3 desarrollado por Thermo Electron Corporation.
- Selección de variables y cálculo de modelos de regresión: Se utilizaron los toolbox: Reshaped Sequential Replacement (RSR), GA-toolbox y regression toolbox ver 1.0 para Matlab desarrolado por Milano Chemometrics and QSAR Research Group de la Universidad de la Bicocca de Milan Italia. Estos toolbox se ejecutaron en el programa Matlab R2013a desarrollado por MathWorks.
- Para organizar datos y el resto de cálculos necesarios: Microsoft Excel 2013.

## 3. RESULTADOS Y DISCUSIÓN

El esquema de trabajo ejecutado se puede observar en la figura 2. Donde se puede observar, que el trabajo se desarrolló en tres etapas generales, a) Pre-tratamiento de datos, b) Selección de Variables y c) Desarrollo de modelos. Luego del pre-tratamiento de datos se generan 7 bases de datos por cada azúcar, por cada nivel. Los procesos de corrección de línea de base, normalizado y suavización se aplicaron a los espectros originales, para tener espectros procesados, siendo a estos a los cuales se les aplica el esquema de trabajo.

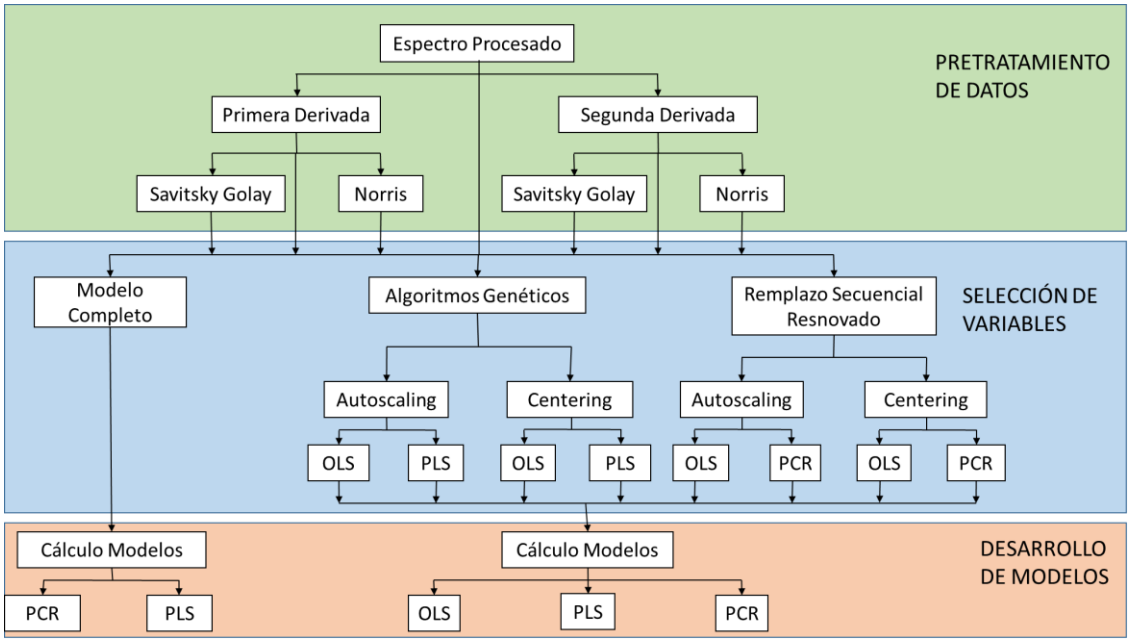


Figura 6. Esquema de trabajo

Para el trabajo, cada base de datos se dividió aleatoriamente en 2 partes la primera parte compuesta por el 70% de los datos, que se utilizaron como training set, para desarrollar la etapa de selección de variables y desarrollo de modelos. Mientras que el 30% restante se utilizó como test set, en la etapa de desarrollo de modelos, para validar, los modelos resultantes del training set.

Este trabajo de dividir las bases de datos se realizó por triplicado, para comparar los resultados obtenidos y seleccionar los mejores. De esta forma después del pre-tratamiento

de datos y las divisiones de bases de datos se obtuvieron 21 bases de datos por cada azúcar. Luego del proceso de selección de variables se tuvo 189 conjuntos de datos con los que calcularían los modelos resultantes, 2 para el Modelo Completo, sin selección de variables, y 3 por cada selección de variables, lo que da un total de 546 modelos por cada azúcar. Dando un total de 2184 modelos calculados.

Los modelos calculados, después de ser validados se escogieron, utilizando como valor de selección el $Q^2_{ext}$. Para los modelos en los que se utilizó selección de variables, se escogió el mejor de cada tipo escalado utilizado.

Todos los modelos calculados para la lactosa, presenta valores de $Q^2$ y $Q^2_{ext}$, muy bajos, por lo cual ninguno de los modelos es aceptable, para predecir la cantidad de lactosa en muestras acuosas compuestas por varios azúcares.

Los mejores modelos calculados para cada método de regresión OLS, PCR y PLS de cada azúcar se presentan en la tabla 4, 5 y 6 respectivamente. En la que se indica el pre-tratamiento utilizado, siendo Espectro procesado EP, Primera derivada filtro Norris 1DN y Primera derivada filtro Savitsky Golay 1DSG. El tipo de escalado también se indica siendo Autoscaling Auto y Centerig Cent. La selección de variables utilizada, donde Algoritmos Genéticos es GA y Remplazo Secuencial Renovado es RSR, y los coeficientes de validación $Q^2$ y $Q^2_{ext}$.

El mejor modelo de cada azúcar, en los que existe selección de variables, en función del escalado, se marca con rojo en todas las tablas.

En la tabla 4, se resumenn las abreviaciones y nomenclaturas que se utilizarán en las posteriores tablas.

Tabla 4. Abreviaciones utilizadas para la presentación de resultados

| Abreviación | Significado |
|---|---|
| **EP** | Espectro Procesado |
| **1D** | Primera Derivada |
| **1DN** | Primera Derivada Filtro de Norris |
| **1DSG** | Primera Derivada Filtro de Savitsky Golay |
| **Auto** | Autoscaling |
| **Cent** | Centering |
| **GA** | Algorítmos Genéticos |
| **RSR** | Remplazo Secuencial Renovado |
| **OLS** | Mínimos Cuadrados Ordinales |
| **PLS** | Mínimos Cuadrados Parciales |
| **PCR** | Regresión de Componentes Principales |
| **Q2** | Coeficiente de Correlacción Cross Validado "Leave one out" |
| **Q2ext** | Coeficiente de Correlacción Cross Validado "Leave one out" externo |

Tabla 5 Mejores resultados OLS

| Azúcar | Fructosa | | Glucosa | | Sacarosa | |
|---|---|---|---|---|---|---|
| **Tratamiento** | 1DN | EP | EP | 1DSG | 1D | 1D |
| **Escalado** | Auto | Cent | Auto | Cent | Auto | Cent |
| **Selección de Variables (SV)** | GA | RSR | RSR | RSR | RSR | RSR |
| **Método de SV** | PLS | OLS | PCR | PCR | OLS | PCR |
| **Q2** | 0.904 | 0.908 | 0.933 | 0.921 | 0.971 | 0.970 |
| **Q2ext** | 0.921 | 0.896 | 0.857 | 0.860 | 0.937 | 0.943 |

Tabla 6 Mejores modelos PCR

| Azúcar | Fructosa | | Glucosa | | Sacarosa | |
|---|---|---|---|---|---|---|
| **Tratamiento** | 1DN | 1DN | 1DSG | 1DSG | 1DN | 1DN |
| **Escalado** | Auto | Cent | Auto | Cent | Auto | Cent |
| **Selección de Variables (SV)** | GA | GA | GA | GA | RSR | GA |
| **Método de SV** | PLS | PLS | PLS | PLS | PCR | PLS |
| **Q2** | 0.928 | 0.936 | 0.889 | 0.925 | 0.974 | 0.969 |
| **Q2ext** | 0.921 | 0.914 | 0.865 | 0.880 | 0.939 | 0.960 |

Tabla 7 Mejores modelos PLS

| Azúcar | Fructosa | | | Glucosa | | Sacarosa | |
|---|---|---|---|---|---|---|---|
| **Tratamiento** | 1DN | 1DN | 1DSG | 1DSG | 1DSG | 1DN | 1DN |
| **Escalado** | Auto | Cent | Cent | Auto | Cent | Auto | Cent |
| **Selección de Variables (SV)** | GA | GA | GA | GA | GA | RSR | GA |
| **Método de SV** | PLS | PLS | PLS | PLS | PLS | PCR | PLS |
| **Q2** | 0.920 | 0.938 | 0.938 | 0.888 | 0.922 | 0.973 | 0.969 |
| **Q2ext** | 0.921 | 0.915 | 0.915 | 0.863 | 0.877 | 0.939 | 0.959 |

De similar forma los mejores resultados obtenidos en cada pre-tratamiento de datos, en modelos completos, sin selección de variables, se presentan en las tablas 7. Los mejores modelos utilizando selección de variables, se presentan en la tabla 8.

Tabla 8 Modelos completos de los pretratamientos de datos

| **Pretratamientos** | Parámetros | Fructosa | Glucosa | Sacarosa |
|---|---|---|---|---|
| **Espectro Procesado** | Método | PLS | PLS | PLS |
| | Q2 | 0.807 | 0.817 | 0.875 |
| | Q2ext | 0.799 | 0.759 | 0.907 |
| **Primera Derivada** | Método | PLS | PLS | PLS |
| | Q2 | 0.846 | 0.829 | 0.927 |
| | Q2ext | 0.859 | 0.801 | 0.939 |
| **Primera Derivada Filtro de Norris** | Método | PLS | PLS | PLS |
| | Q2 | 0.857 | 0.867 | 0.950 |
| | Q2ext | 0.869 | 0.829 | 0.950 |
| **Primera Derivada Filtro de Savitsky Golay** | Método | PLS | PLS | PLS |
| | Q2 | 0.872 | 0.860 | 0.930 |
| | Q2ext | 0.867 | 0.884 | 0.946 |
| **Segunda Derivada** | Método | PLS | PCR | PLS |
| | Q2 | 0.641 | 0.781 | 0.830 |
| | Q2ext | 0.615 | 0.810 | 0.842 |
| **Segunda Derivada Filtro de Norris** | Método | PLS | PCR | PLS |
| | Q2 | 0.747 | 0.674 | 0.934 |
| | Q2ext | 0.743 | 0.759 | 0.936 |
| **Segunda Derivada Filtro de Savitsky Golay** | Método | PLS | PLS | PLS |
| | Q2 | 0.672 | 0.810 | 0.849 |
| | Q2ext | 0.690 | 0.779 | 0.802 |

Tabla 9 Mejores modelos de los pretratamientos de datos utilizando selección de variables

| Pretratamientos | Parámetros | Fructosa | | Glucosa | | Sacarosa | |
|---|---|---|---|---|---|---|---|
| **Espectro Procesado** | Escalado | Auto | Cent | Auto | Cent | Auto | Cent |
| | Selección de Variables (SV) | GA | GA | GA | GA | RSR | RSR |
| | Método de SV | PLS | PLS | PLS | PLS | PCR | PCR |
| | Q2 | 0.886 | 0.926 | 0.896 | 0.920 | 0.966 | 0.837 |
| | Q2ext | 0.873 | 0.902 | 0.857 | 0.834 | 0.925 | 0.937 |
| **Primera Derivada** | Escalado | Auto | Cent | Auto | Cent | Auto | Cent |
| | Selección de Variables (SV) | GA | GA | GA | GA | RSR | RSR |
| | Método de SV | PLS | PLS | PLS | PLS | PCR | PCR |
| | Q2 | 0.903 | 0.940 | 0.891 | 0.917 | 0.971 | 0.967 |
| | Q2ext | 0.885 | 0.901 | 0.757 | 0.804 | 0.937 | 0.953 |
| **Primera Derivada Filtro Norris** | Escalado | Auto | Cent | Auto | Cent | Auto | Cent |
| | Selección de Variables (SV) | GA | GA | GA | GA | RSR | RSR |
| | Método de SV | PCR | PLS | PLS | PLS | PCR | PCR |
| | Q2 | 0.928 | 0.938 | 0.867 | 0.902 | 0.950 | 0.969 |
| | Q2ext | 0.921 | 0.915 | 0.829 | 0.803 | 0.950 | 0.959 |
| **Primera Derivada Filtro Savitsky Golay** | Escalado | Auto | Cent | Auto | Cent | Auto | Cent |
| | Selección de Variables (SV) | GA | GA | GA | GA | RSR | RSR |
| | Método de SV | PLS | PLS | PLS | PLS | PCR | PCR |
| | Q2 | 0.869 | 0.938 | 0.889 | 0.925 | 0.950 | 0.970 |
| | Q2ext | 0.874 | 0.915 | 0.865 | 0.880 | 0.919 | 0.923 |
| **Segunda Derivada** | Escalado | Auto | Cent | Auto | Cent | Auto | Cent |
| | Selección de Variables (SV) | GA | GA | GA | GA | RSR | RSR |
| | Método de SV | PLS | PLS | PLS | PLS | PCR | PCR |
| | Q2 | 0.815 | 0.882 | 0.818 | 0.807 | 0.892 | 0.934 |
| | Q2ext | 0.736 | 0.785 | 0.798 | 0.715 | 0.897 | 0.928 |
| **Segunda Derivada Filtro Norris** | Escalado | Auto | Cent | Auto | Cent | Auto | Cent |
| | Selección de Variables (SV) | GA | GA | GA | GA | RSR | RSR |
| | Método de SV | PLS | PLS | PLS | PLS | PCR | PCR |
| | Q2 | 0.853 | 0.9185 | 0.895 | 0.887 | 0.965 | 0.971 |
| | Q2ext | 0.828 | 0.785 | 0.824 | 0.773 | 0.924 | 0.935 |

En las tablas 9, 10 y 11 se presentan los mejores modelos obtenidos para cada tipo de selección de variables, de la fructosa, glucosa y sacarosa respectivamente

Tabla 10 Mejores modelos para cada tipo de selección de variables de la fructosa

| Parámetros | Tipo de Selección de variables | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Selección de Variables (SV)** | Modelo Completo | GA | GA | GA | GA | RSR | RSR | RSR | RSR |
| **Método de SV** | N/A | OLS | OLS | PLS | PLS | OLS | OLS | PCR | PCR |
| **Escalado** | N/A | Auto | Cent | Auto | Cent | Auto | Cent | Auto | Cent |
| **Pre-Tratamiento** | 1DN | 1DN | 1DN | 1DN | 1DN | 1DN | EP | 1DN | 1DN |
| **Método de Regresión** | PCR | PLS | PCR | PLS | PLS | PLS | PLS | PLS | OLS |
| **Q2** | 0.822 | 0.857 | 0.909 | 0.928 | 0.938 | 0.954 | 0.909 | 0.948 | 0.954 |
| **Q2ext** | 0.913 | 0.947 | 0.903 | 0.921 | 0.915 | 0.904 | 0.898 | 0.878 | 0.844 |

Tabla 11 Mejores modelos para cada tipo de selección de variables de la glucosa

| Parámetros | Tipo de Selección de variables | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Selección de Variables (SV)** | Modelo Completo | GA | GA | GA | GA | RSR | RSR | RSR | RSR |
| **Método de SV** | N/A | OLS | OLS | PLS | PLS | OLS | OLS | PCR | PCR |
| **Escalado** | N/A | Auto | Cent | Auto | Cent | Auto | Cent | Auto | Cent |
| **Pre-Tratamiento** | 1DSG | 1DSG | 1DSG | 1DSG | 1DSG | 2DN | 1DSG | EP | 1DSG |
| **Método de Regresión** | PLS | PCR | OLS | PCR | PCR | PLS | OLS | OLS | OLS |
| **Q2** | 0.860 | 0.883 | 0.844 | 0.889 | 0.925 | 0.895 | 0.928 | 0.933 | 0.921 |
| **Q2ext** | 0.884 | 0.864 | 0.793 | 0.865 | 0.880 | 0.824 | 0.846 | 0.857 | 0.860 |

Tabla 12 Mejores modelos para cada tipo de selección de variables de la sacarosa

| Parámetros | Tipo de Selección de variables | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Selección de Variables (SV)** | Modelo Completo | GA | GA | GA | GA | RSR | RSR | RSR | RSR |
| **Método de SV** | N/A | OLS | OLS | PLS | PLS | OLS | OLS | PCR | PCR |
| **Escalado** | N/A | Auto | Cent | Auto | Cent | Auto | Cent | Auto | Cent |
| **Pre-Tratamiento** | 1DN | 1DN | 1DN | 1DN | 1DN | 1D | 1DN | 1DN | 1DN |
| **Método de Regresión** | PLS | OLS | PCR | PLS | PCR | OLS | OLS | PLS | OLS |
| **Q2** | 0.950 | 0.932 | 0.956 | 0.953 | 0.969 | 0.971 | 0.977 | 0.973 | 0.970 |
| **Q2ext** | 0.950 | 0.924 | 0.942 | 0.935 | 0.960 | 0.937 | 0.942 | 0.939 | 0.943 |

Los resultados observados en los modelos seleccionados, nos muestran las siguientes características comunes:

- No se observa que ninguno de los escalados de variables utilizados, ofrezcan una mejora en calidad de los modelos, en comparación con el otro.
- En este caso específico la aplicación de la segunda derivada, no genera modelos de mejor calidad que los obtenidos con la primera derivada, utilizando cualquier de los filtros propuestos.
- La mayoría de los mejores modelos corresponden al Pretratamiento 1DN, Primera Derivada con el Filtro Norris y 1DSG, Primera Derivada con el Filtro Savitsky Golay.

Los mejores modelos para cada azúcar, con las variables que componen cada modelo, se muestran en la tabla 12.

Tabla 13 Mejores modelos de cada azúcar

| Azúcar | Pretratamiento | Selección de Variables | Escalado | Método de Regresión | $Q^2$ | $Q^2_{ext}$ | Número de Componentes | Variables (Longitudes de Onda) $cm^{-1}$ |
|---|---|---|---|---|---|---|---|---|
| **Fructosa** | 1DN | GA-PLS | Auto | PLS | 0.953 | 0.935 | 7 | **989.3, 991.2, 993.1, 1039.4, 1047.1, 1049.1, 1051.0, 1052.9, 1054.9, 1056.8, 1112.7, 1114.6, 1116.6, 1118.5, 1122.4, 1278.6** |
| **Glucosa** | 1DSG | GA-PLS | Cent | PCR | 0.925 | 0.880 | 7 | **1002.8, 1004.7, 1006.7, 1016.3, 1018.2, 1027.9, 1039.4, 1041.4, 1043.3, 1045.2, 1047.2, 1049.1, 1085.7, 1087.7, 1089.6, 1091.5, 1101.2, 1120.4, 1122.4,** |
| **Sacarosa** | 1DN | GA-PLS | Cent | PCR | 0.969 | 0.960 | 7 | **973.9, 979.7, 981.6, 983.5, 985.4, 987.4, 989.3, 991.2, 993.2, 995.1, 1022.1, 1024.0, 1049.1, 1062.6, 1064.5, 1066.4, 1068.4, 1070.3, 1232.3** |

## 4. CONCLUSIONES

El algoritmo de selección de variables que da mejor resultado, aplicado en señales infrarrojas de soluciones acuosas de mezclas de azúcares, son los algoritmos genéticos, utilizando el método PLS. En el caso del escalado de variables la fructosa presenta mejores resultados utilizando autoscaling, mientras que la glucosa y sacarosa centering. El pretratamiento de datos mejor para la glucosa es la Primera Derivada con el filtro de Savitsky Golay, mientras que para la fructosa y la sacarosa es la Primera Derivada con el filtro de Norris. El método de regresión que mejor se adapta a los datos de la fructosa son la Mínimos Cuadrados Parciales (PLS), mientras que para la Glucosa y Sacarosa es la Regresión de Componentes Principales (PCR).

Siendo que los azúcares en los alimentos, siempre se presentan como una mezcla de ellos y comparten la misma zona del espectro infrarrojo, la interferencia que estos pueden provocar en su cuantificación es evidente, ejemplo claro de esto fue la determinación de la lactosa, azúcar que no fue posible modelar. También hay que tener en cuenta, en el caso de la lactosa, que la presencia de la misma en las muestras medidas es inferior al del resto de azúcares, debido a su baja solubilidad. La cuantificación de la lactosa de forma individual, utilizando espectroscopía infrarroja es posible como lo indica Fairbrother et al (Faitbrother, George y Williams).

Los mejores modelos obtenidos, son los correspondientes a la primera derivada, indiferente del filtro que se haya utilizado. Se puede concluir que la información más útil de los espectros, al tener tantas posibles interferencias, se la obtiene al maximizar las pendientes del espectro y llevar a cero los picos y valles, realizando la derivada por segmentos, como sugieren los filtros de Norris y Savitsky Golay, y de esta forma disminuir el efecto de las variables ruidosas que se pueda contener el espectro.

Los métodos de regresión que funcionan mejor son los que métodos de regresión sesgada, ya que al no utilizar la información de presente en las variables, obtiene modelos más estables y con mejor predicción, debido a que las variables originales están muy correlacionadas entre sí, información que no permite un buen desarrollo de un modelo de Mínimos Cuadrados Ordinarios. Por coincidencia los modelos obtenidos tienen 7 componentes. La correlación entre las variables se puede observar en la tabla 13.

Tabla 14 Correlación entre variables

| Fructosa | | Glucosa | | Sacarosa | |
|---|---|---|---|---|---|
| **Variables** | $R^2$ | Variables | $R^2$ | Variables | $R^2$ |
| **989.3** | 0.992 | **1002.8** | 0.906 | **973.9** | 0.931 |
| **991.2** | 0.985 | **1004.7** | 0.936 | **979.7** | 0.995 |
| **993.2** | | **1006.7** | | **981.6** | 0.996 |
| **1047.2** | 0.990 | **1016.3** | 0.969 | **983.5** | 0.996 |
| **1049.1** | 0.985 | **1018.2** | | **985.4** | 0.995 |
| **1051.0** | 0.978 | **1039.4** | 0.963 | **987.4** | 0.994 |
| **1052.9** | 0.965 | **1041.4** | 0.964 | **989.3** | 0.992 |
| **1054.9** | | **1043.3** | 0.955 | **991.2** | 0.985 |
| **1112.7** | 0.979 | **1045.2** | 0.958 | **993.2** | 0.962 |
| **1114.7** | 0.982 | **1047.2** | 0.960 | **995.1** | |
| **1116.6** | | **1049.1** | | **1022.1** | 0.989 |
| | | **1085.7** | 0.963 | **1024.0** | |
| | | **1087.7** | 0.949 | **1062.6** | 0.897 |
| | | **1089.6** | 0.909 | **1064.5** | 0.941 |
| | | **1091.5** | | **1066.4** | 0.968 |
| | | **1120.4** | 0.814 | **1068.4** | 0.982 |
| | | **1122.4** | | **1070.3** | |

## 5. REFERENCIAS BIBLIOGRÁFICAS


Alais, Charles. *Ciencia de la Leche. Principios de ténica Lechera*. Sevilla: Reverté S.A., 2003.

Blanco Romía, Marcelo y Manuel Alcalà Bernàrdez. «Multivariate Calibration for Quantitative Analysis.» Sun, Da-Wen. *Infrared Spectroscopy for Food Quality Analysis and Control*. Elsevier, 2009. 51-80.

Cassotti, Matteo, Francesca Grisoni y Roberto Todeschini. «Reshaped Sequencial Replacement algorithm: An efficient approach to variable selection.» *Chemometrics and Intelligent Laboratory Systems* (2014): 136-148.

Consonni, Viviana, Davide Ballabio y Roberto Todeschini. «Evaluation of model predictive ability by external validation techniques.» *Journal of Chemometrics* (2009): 194-201.

De Jong, Kenneth, William Spears y Diana Gordon. «Using Genetic Algorithms for Concept Learning.» Grefenstette, John. *Genetic Algorithms for Machine Learning*. Vol. 13. Springer, 1993. 2-3 vols. 5-32.

Duchowicz, Pablo, Eduardo Castro y Francisco Fernández. «Alternative Algorithm for the Search of an Optimal Set of Descriptors in QSAR-QSPR Studies.» *MATCH Communications in Mathematical and in Computer Chemistry* (2006): 179-192.

Dufour, Éric. «Principles of Infrared Spectroscopy.» Sun, Da-Wen. *Infrared Spectroscopy for Food Quality Analysis and Control*. Elsevier, 2009. 3-25.

Faitbrother, Paul, William George y Jill Williams. «Whey fermentation: on-line analysis of lactose and lactic acid by FTIR spectroscopy.» *Applied Microbiology and Biotechnology* (1991): 301-305.

Fearn, Tom. «A look at some standard pre-treatments for spectra.» *NIR news* (1999): 10-11.

—. «The effect of spectral pre-treatments on interpretation.» *NIRnews* (2009): 15-16.

Fearn, Tom, y otros. «On the geometry of SNV and MSC.» *Chemometrics and Intelligent Laboratory Systems* (2009): 22-26.

Fehling, Hernan von. « Die quantitative Bestimmung von Zucker und Stärkmehl mittelst Kupfervitriol.» *Justus Liebigs Annalen der Chemie* (1849): 109-113.

Frank, Ildikó y Roberto Todeschini. *The data analysis handbook*. Amsterdam: Elsevier, 1994.

Gabriel, Luciana, y otros. «Multivariate Analysis of the Spectroscopic Profile of the Sugar Fraction of Apple Pomace.» *Brazilian Archives of Biology and Technology* (2013): 439-446.

Gander, W y U von Matt. «Chapter 9. Smoothing Filters.» Gander, W y Jiří Hřebíček. *Solving Problems in Scientific Computing Using MAPLE and MATLAB*. Berlin Heidelberg: Springer, 1993. 121-139.

Gramatica, Paola. «Principles of QSAR models validation:internal y external.» *QSAR & Comtinatorial Science* (2007): 694-701.

Grisoni, F, M Cassotti y R Todeschini. «Reshaped Sequential Replacement for variable selection in QSPR: comparasion with other reference methods.» *Journal of Chemometrics* (2014): 249-259.

Hopkins, David. «What is a Norris derivative?» *NIR news* (2011): 3.

Leardi, R, R Boggia y M Terrile. «Genetic Algorithms as a strategy for feature selection.» *Journal of Chemometrics* (1992): 267-281.

Leardi, Riccardo. «Application of genetic algorithm-PLS for feature selection in spectral data set.» *Journal of Chemometrics* (2000): 643-655.

Miller, Alan. «Selection of Subsets of Regression Variables.» *Journal of the Royal Statistical Society. Series A (General)* (1984): 389-425.

Niazi, Ali y Riccardo Leardi. «Genetic algorithms in chemometrics.» *Journal of Chemometrics* (2012): 345-351.

Norris, H K. «Extraction information from spectrophotometric curves. Predicting chimical conposition from visible and near-infrared spectra.» *Food Research and Data Analysis* (1983): 95-114.

Norris, Karl. «Applying Norris Derivatives. Understanding and correcting the factors which affect diffuse transmittance spectra.» *NIR news* (2001): 6.

—. «Instrument Design. Interactions among instrument band pass, instrument noise, sample-absober bandwidth and calibration erro.» *NIR news* (1998): 3.

Rinnan, Åsmund, y otros. «Data Pre-processing.» Sun, Da-Wen. *Infrared Spectroscopy for Food Quality Analysis and Control*. Elsevier, 2009. 29-50.

Savitzky, Abraham y Marcel Golay. «Smoothing and Differentiation of Data by Simplified Least Squares Procedures.» *Analytical Chemistry* (1964): 1627-1639.

Schüürmann, Gerrit, y otros. «External Validation and Prediction Employing the Predictive Squared Correlation Coefficient s Test Set Activity Mean vs Training Set Activity Mean.» *J. Chem. Inf. Model* (2008): 2140-2145.

Sorol, Natalia, y otros. «Visible/near infrared-partial least-squares analysis of Brix in sugar cane juice. A test field for variable selction methods.» *Chemometrics and Intelligent Laboratory Systems* (2010): 100-109.

STATGRAPHICS. « StatPoint, Inc.» 14 de Septiembre de 2006. *STATGRAPHICS.* Documento. 16 de Abril de 2016.

Subramanian , Anand y Luis Rodriguez-Saona. «Fourier Transform Infrared (FTIR) Spectroscopy.» Da-Wen, Sun. *Infrared Spectroscopy for Food Quality Analysis and Control*. New York, USA: Elsevier, 2009. 119-141.

Thermo Electron Corporation. *Using OMNIC Algorithms*. Madison: Thermo Electron Corporation, s.f.

Thermo Scientific. «Advantages of a Fourier Transform Infrared Spectrometer.» Technical Note 50674. 2015.

Todeschini, Roberto, y otros. «Canonical Measure of Correlation (CMC) and Canonical Measure of Distance (CMD) between sets of data. Part I. Theory and simple chemomerics applications.» *Analytica Chimica Acta* (2009): 45-51.

—. «Detecting "bad" regression model: multicriteria fitness functions in regression analysis.» *Analytica Chimica Acta* (2004): 199-208.

Todeschini, Roberto, y otros. «MobyDigs: Software for Regression and Classification Models by Genetic Algorithms.» Leardi, R. *Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks*. Amsterdam: Elsevier, 2003. 141 - 167.

Wang , Jun, y otros. «Rapid Analysis of Glucose, Fructose, Sucrose, and Maltose in Honeys from Different Geographic Regions using Fourier Transform Infrared Spectroscopy andMultivariate Analysis.» *Food Chemistry* (2010): 208-214.

Wang, Jun, y otros. «Rapid Analysis of Gluvose, Fructose, Sucrose and Maltose in Honeys from Different Geographic Regions using Fourier Transform Infrared Spectroscopy and Multivariante Analysis.» *Food Chemistry* (2010): 208-2014.

Wehrens, Ron, Hein Putter y Lutgarde Buydens. «The bootstrap: a tutorial.» *Chemometrics and Intelligent laboratory systems* (2000): 35-52.

Wold, Svante, Michael Sjöström y Lennart Eriksson. «PLS-regression: a basic tool of chemometrics.» *Chemometrics and intelligent laboratory sistems* (2001): 109-130.