Diagnóstico de la operación de un Espectrómetro NIR montado en línea mediante Análisis Multivariante

Moya-González, A[1)], Barreiro, P[1)], Ortiz-Cañavate, J[1)]

[1)] Universidad Politécnica de Madrid, Laboratorio de Propiedades Físicas y Tecnologías Avanzadas en Agroalimentación. Avda. Complutense s/n, 28040 Madrid, Spain, Tel. 34 91 336 5862. E-mail: adolfo.moya@upm.es

Resumen

El presente trabajo presenta un análisis no supervisado para el diagnóstico de operación de un espectrómetro NIR montado en línea y funcionando en la industria desde 2004 para la selección de bulbos de cebolla. Mediante este análisis multivariante se propone y estudia el empleo de determinados estadísticos de control de procesos para la identificación de individuos fuera de control durante las campañas analizadas (2004-2007) empleando los datos espectrales con y sin la realización de un pre-procesado. Los resultados obtenidos muestran que el empleo del pre-procesado resulta de gran utilidad en la eliminación de la varianza interferente y con ello en la reducción de los individuos fuera de control. Las nuevas fuentes de varianza interferente se incrementan a lo largo de las campañas y hacen necesaria la realización de un estudio pormenorizado para la eliminación de sus efectos.

Abstract

This study presents a non-supervised analysis for the diagnosis of an on-line NIR spectrometer under industrial use for onion quality determination since 2004. Process control statistics are used for a multivariate supervision of the onion bulb classification under breeding strategy during four seasons (2004-2007) comparing the use of pre-processed and non pre-processed spectral data. The results shows that pre-process algorithms are very useful for the elimination of interference variance and thus to reduce the quantity of out of control individuals. The increase of interference variance sources trough seasons, points the need of further studies for an appropriate control.

Palabras Clave (Keywords)

Robustez del análisis (robustness analysis), aplicación NIR (NIR application), cebolla (onion), clasificación (classification).

1. Introducción

El empleo de la Espectrometría en el Infra Rojo Cercano (NIR) para la determinación de la calidad interna en frutas y hortalizas está ampliamente aceptado incluso en aplicaciones en tiempo real. Todavía existe un elevado número de fuentes de variación no controladas que modifican la información espectral y reducen la precisión de las estimaciones. Algunas de estas fuentes de variación son: la temperatura interna del producto y del espectrómetro [9], el grosor de la piel, y la presencia de capas o huecos que actúan como barrera en el interior del producto [7].

En muchas aplicaciones NIR, existe una limitación en la transferencia de los modelos predictivos durante la campaña y también entre distintas campañas debida a una pobre validación, aunque estos modelos pueden resultar sumamente útiles para la clasificación de un reducido número de categorías [8]. La precisión obtenida con los modelos se ve incrementada al incluir una mayor variabilidad en la muestra de calibración, aunque la inclusión de datos atípicos produce el efecto contrario [11]. La aparición de nuevas fuentes de variación debidas tanto al equipo, al material empleado o a las condiciones ambientales genera problemas en el uso de los modelos desarrollados previamente [6]. Existen también estudios acerca de la importancia de asegurar el correcto uso del análisis IR en línea por medio de procedimientos multivariantes de robustez [14].

El LPFTAG, conjuntamente con la empresa AGROTÉCNICA EXTREMEÑA S.L. ha llevado a cado un proyecto de innovación y transferencia de tecnología (BULBONIR) que ha dado como resultado una línea de clasificación de cebollas que emplea espectroscopía NIR en interactancia. El procedimiento de análisis y clasificación está en vías de patente desde mayo de 2007 (P2007011214). La interactancia, a pesar de suponer la situación más difícil para su implementación en línea, ha mostrado resultados alentadores de cara a la obtención de buenos modelos de predicción [12].

Figura 1. Línea de clasificación de cebollas en las instalaciones de la industria

Desde el año 2004, se han clasificado 1.036.001 bulbos de cebolla mediante la línea automatizada. Los bulbos seleccionados forman parte del programa de mejora de SS en cebolla para deshidratado llevado a cabo por la empresa.

La línea automatizada emplea un modelo de estimación lineal para la clasificación de los bulbos en cuatro categorías según su contenido en SS. La base de datos para la calibración del modelo fue generada fuera de línea durante la campaña 2002. El sistema dispone de un software propio capaz de identificar y descartar espectros anómalos en tiempo real, mediante su proyección en un espacio de componentes principales (PC) generado a partir de la base de datos de calibración [5].

El estudio del rendimiento del modelo una vez implementado para la medición en línea y las medidas de control adoptadas se detallan en las referencias [2], [3], [4] y [5].

El porcentaje de MS (materia seca) de las líneas seleccionadas se ha incrementado consistentemente en 0.2 puntos porcentuales de media por año desde el comienzo del programa de mejora en el año 2002 hasta la campaña 2006 última en la que se dispone de este dato. Adicionalmente se dispone de las distribuciones de SS correspondientes al material seleccionado en las que se aprecia un claro desplazamiento hacia un mayor contenido en SS desde la campaña 2005 hasta la 2008. El techo de la selección masal aplicada es un parámetro desconocido y su determinación deberá basarse en el estudio de la evolución del contenido en MS a lo largo de futuras campañas.

Si bien el funcionamiento cualitativo de la línea automatizada ha resultado aceptable según los resultados obtenidos, se aprecia un incremento de sesgos no explicados en las variables de control establecidas inicialmente, lo que hace necesaria la implementación de un nuevo sistema de control más eficiente.

Objetivo

El objetivo de este trabajo es el diagnóstico de la operación del sistema de clasificación en línea a lo largo de las campañas 2004-2007 y la propuesta de nuevos procedimientos de supervisión que garanticen el control mediante técnicas de análisis multivariante susceptibles de ser implementadas en línea.

2. Materiales y métodos

Material disponible

Se dispone de todos los espectros (media de 5 ó 3 repeticiones) para los 772.069 bulbos analizados durante las campañas 2004 a 2007. Para la realización del trabajo, y dada la enorme cantidad de información disponible se ha llevado a cabo un muestreo representativo de los espectros disponibles para cada una de las campañas analizadas. La Tabla 1 muestra el número de espectros analizados correspondientes a cada campaña de selección. Los espectros analizados incluyen aquellos identificados como anómalos en tiempo real durante el funcionamiento de la línea mediante su proyección sobre el espacio de PC

definido a partir de la base de datos de calibración. Los mencionados espectros anómalos no fueron empleados en su momento para la clasificación de bulbos.

Tabla 1. Espectros analizados por campaña

CAMPAÑA	ESPECTROS ANALIZADOS
2004	54.005
2005	49.993
2006	30.539
2007	65.447

Algoritmos de pre-procesado de los espectros

Con el fin de eliminar la variación interferente como puede ser la producida por la dispersión de la luz, se han aplicado distintos procedimientos para el pre-tratamiento de los espectros (244 longitudes de onda de 894 a 1649 nm). El pre-procesado de los datos que se ha llevado a cabo incluye la aplicación del algoritmo de Savitsky-Golay para el suavizado de los espectros [10], la *varianza* normal estándar (SNV) [1], y el algoritmo De-Trend para la corrección de la línea base mediante un polinomio de orden 2 basado en el algoritmo definido por Barnes y cols. en [1].

Control de procesos basado en análisis multivariante con y sin pre-procesado de espectros

Mediante la realización de un análisis de componentes principales (PCA) sobre los espectros de la campaña 2004, definimos un nuevo espacio en el que son proyectadas posteriormente las distintas campañas. Adicionalmente se determinan los estadísticos Q y T^2 según se especifica en [13].

El estadístico T^2 es una medida de la distancia de Mahalanobis en el espacio reducido entre la posición de una muestra (su valor para los distintos PC) y el origen que define aquellas muestras de variación mínima. Una señal fuera de control se identifica porque su valor de T^2 supera el valor límite.

El estadístico Q se define como la forma cuadrática de los residuos, lo cual es el cuadrado de la diferencia entre los valores observados y los predichos por el modelo PCA. Suponiendo que el modelo lineal de PCA es válido, la distribución de los residuos estará bien aproximada mediante la forma cuadrática de una distribución normal. El estadístico Q define la distancia Euclídea a la posición de una observación desde el hiperplano formado por la representación de PCA.

3. Resultados

Se ha llevado a cabo un PCA a partir de los espectros sin procesar y otro a partir de los espectros pre-procesados, ambos para la campaña 2004. Ambos PCA se han realizado estableciendo 50 PC o variables latentes con objeto de maximizar las fuentes de variación contenidas en los datos originales, es decir, la reducción de la dimensionalidad minimizando la pérdida de información.

Como resultado de estos PCA se han obtenido las proyecciones de las observaciones sobre sus respectivos espacios reducidos, los valores de los estadísticos Q y T^2 para cada observación, y los valores límite para cada estadístico con una significación del 95%. Mediante la proyección de los espectros de las restantes campañas (pre-procesados y sin pre-procesar) obtenemos sus valores sobre los respectivos espacios reducidos y los valores de Q y T^2 para cada observación.

Las Tablas 2 y 3, muestran el número total de observaciones, el número de observaciones cuya Q se encuentra por debajo del límite establecido, el número de observaciones cuya T^2 se encuentra por debajo del límite establecido y el número de observaciones para las que ambos parámetros se encuentran dentro del límite de control.

En el caso de los espectros sin pre-procesar (Tabla 2), los valores límite establecidos para los estadísticos Q y T^2 son: Q_{lim}= 0.0012; límite T^2_{lim}= 69.7384.

La tabla 2 demuestra que el número de individuos dentro del rango de control definido por ambos estadísticos sufre un importante descenso ya desde la campaña 2005.

Tabla 2. Número de individuos que presentan valores de Q y T^2 inferiores a los límites establecidos (espectros sin pre-procesar)

	Número de individuos			
ESPECTROS	Total	Q < lim Q	T2 < lim T2	Q & T2 < lim
2004R_sx	54005	50254	50958	48774
2005_sx	49993	178	21155	147
2006_sx	30539	220	14896	209
2007_sx	65447	242	14135	170

En el caso de los espectros pre-procesados (tabla 3), los valores límite establecidos para los estadísticos Q y T^2 son: límite $Q_{lim} = 1.6529*10^{-6}$; $T^2_{lim} = 69.7384$.

La Tabla 3 demuestra que el número de individuos dentro del rango de control definido por ambos estadísticos desciende a lo largo de las distintas campañas de forma menos pronunciada que en el caso de los espectros sin pre-procesar.

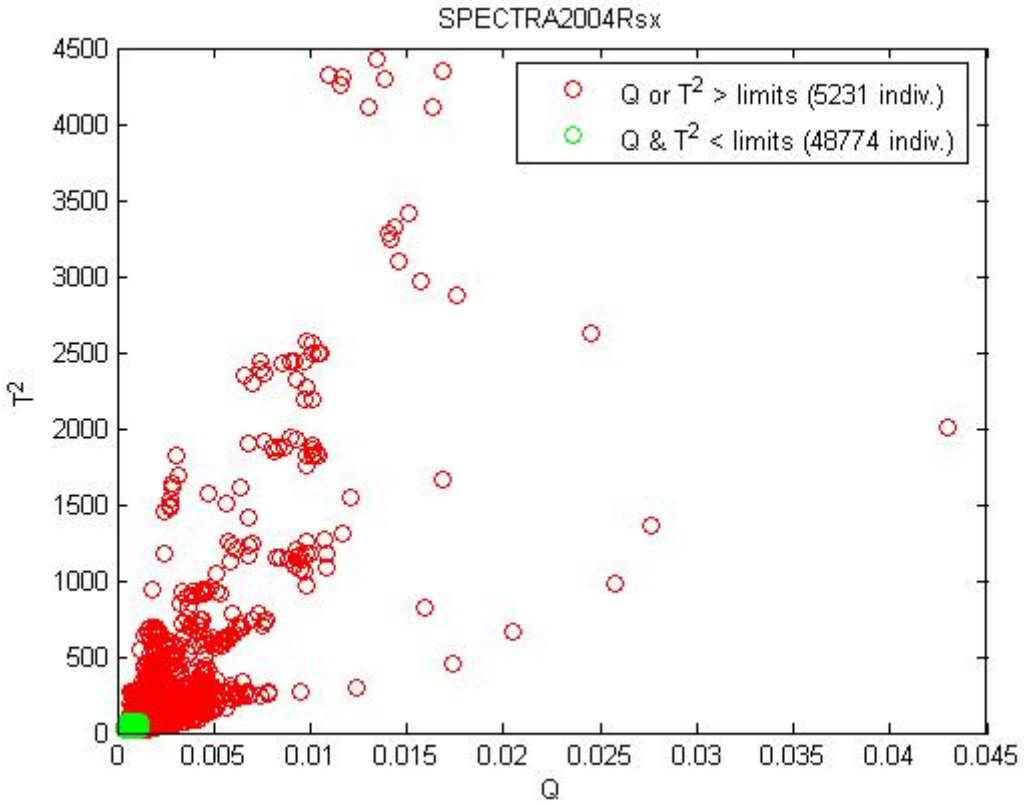
Tabla 3. Número de individuos que presentan valores de Q y T^2 inferiores a los límites establecidos (espectros pre-procesados)

	Número de individuos			
ESPECTROS	Total	Q < lim Q	T2 < lim T2	Q & T2 < lim
2004RxsnvD	54005	51992	52700	51681
2005xsnvD	49993	15259	29027	13304
2006xsnvD	30539	2368	11604	2198
2007xsnvD	65447	1424	16760	664

Las Figuras 2 y 3 representan los valores de Q contra T^2 para los individuos de la campaña 2004, se indican en verde las observaciones para las que tanto Q como T^2 se encuentran dentro de los límites establecidos y en rojo aquellas en las que al menos uno de los mencionados estadísticos es superior al límite para él establecido.

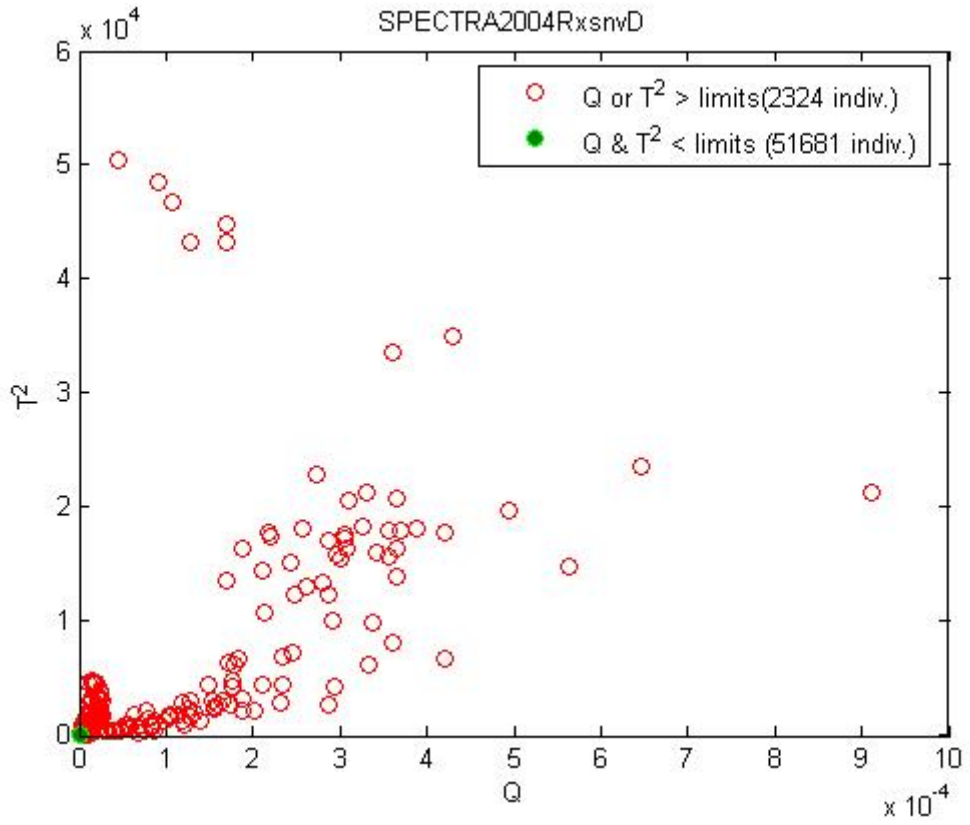
La Figura 2 muestra la distribución de los valores de Q y T^2 para los espectros sin pre-procesar. Los individuos con alguno de los estadísticos mencionados fuera de rango parecen alejarse de la nube de puntos según trayectorias definidas.

Figura 2. Valores de Q vs T^2 de *Hotelling* para el PCA desarrollado a partir de los espectros sin pre-procesar



La Figura 3 muestra la distribución de los valores de Q y T^2 para los espectros pre-procesados. El número de individuos con ambos estadísticos dentro de los límites de control es mayor que para los espectros sin pre-procesar. Al igual que en caso anterior los individuos con alguno de los estadísticos mencionados fuera de rango parecen alejarse de la nube de puntos según trayectorias definidas

Figura 3. Valores de Q vs T^2 de *Hotelling* para el PCA desarrollado a partir de los espectros pre-procesados



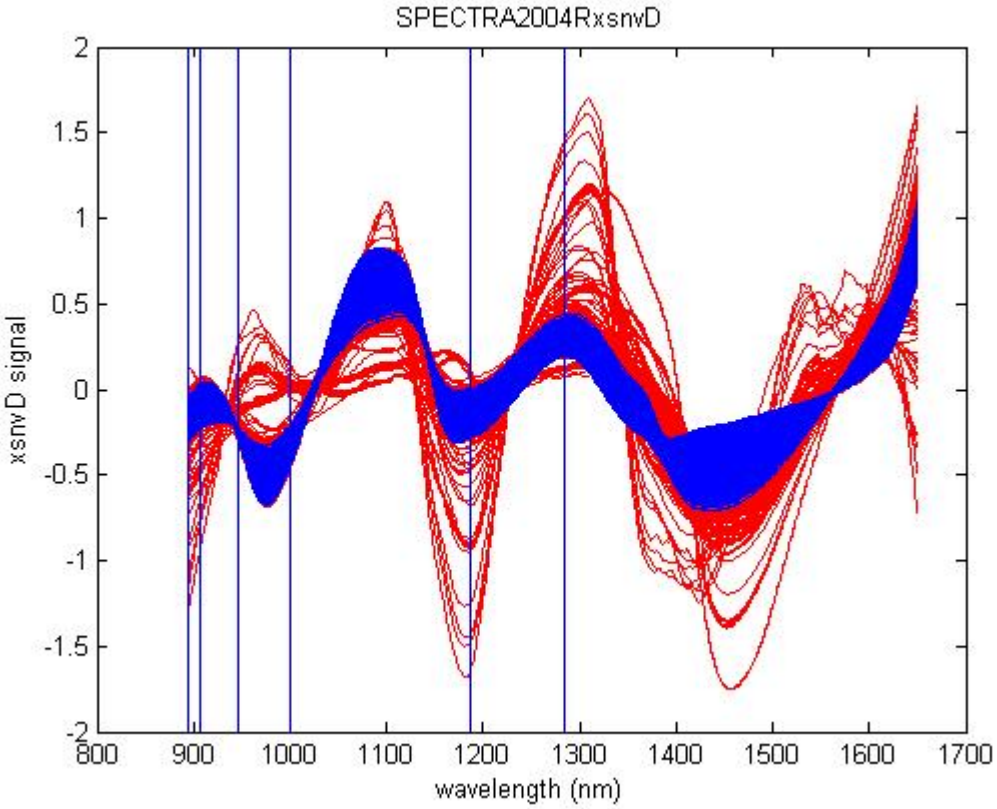
La Tabla 4 compara los valores máximos de Q y T^2 en relación a sus límites de control para la campaña 2004, con y sin pre-procesado de espectros. A mayor valor de estos parámetros mayor es la anormalidad detectada.

Tabla 4. Relaciones entre los valores máximos detectados y los límites de control para los parámetros Q y T^2. Para la campaña 2004 (espectros pre-procesados y sin pre-procesar)

	CAMPAÑA 2004	
	espectros pre-procesados	espectros sin pre-procesar
Q_{max}/Q_{lim}	551.90	35.83
T^2_{max}/T^2_{lim}	724.06	63.48

La Figura 4, muestra los espectros pre-procesados, en azul los correspondientes a individuos cuyos estadísticos Q y T^2 están dentro de los límites de control y en rojo los correspondientes a aquellos individuos con alguno de los mencionados estadísticos fuera de límites. Las líneas verticales marcan las longitudes de onda utilizadas por el modelo de regresión lineal empleado para las estimaciones de SS. Las mencionadas longitudes de onda se encuentran en zonas de máxima variabilidad de los espectros pre-procesados dentro de los límites de control, excepto λ=947 nm, que está en una zona de variabilidad mínima.

Figura 4. Espectros pre-procesados de la campaña 2004. En azul se muestran aquellos correspondientes a individuos con Q y T2 dentro de los límites definidos, en rojo si alguno de los dos estadísticos está fuera de control



Discusión

El estadístico T^2 define la medida en que una muestra se aparta de los valores que presentan la mínima variación, que representan el funcionamiento medio del proceso. La evolución del T^2 a lo largo de las distintas campañas puede estar en parte explicada por la evolución del material vegetal.

El incremento de los valores de Q que superan los límites de definidos a lo largo de las campañas está directamente relacionado con la aparición de fuentes de variación

interferentes no recogidas en el PCA realizado. Mediante el pre-procesado de los datos logramos eliminar una gran parte de esta nueva variación interferente que permite un gran incremento en el número de individuos con valores de Q dentro de los límites. Sin embargo aún en éste último caso, los valores de Q fuera de rango crecen a lo largo de las campañas hasta suponer cerca de un 98% para la campaña 2007 lo que indica que el proceso ha incorporado fuentes de variación no incorporadas en el set de datos original y que no son eliminadas totalmente mediante el pre-procesado. De donde se deduce que es necesario establecer un proceso de transferencia de calibración de los modelos de estimación entre campañas.

La determinación de individuos dentro de los límites para un nivel de significación más elevado (99%) reduce en cierta medida el número de individuos fuera de control pero supone una corrección insuficiente para el caso de los individuos con Q fuera de límites.

El estudio de las características de los distintos individuos situados fuera de control, en relación a su posición en el gráfico $QvsT^2$ podrá ayudarnos a definir tipologías de individuos fuera de control. El estudio de estos grupos, que podrían estar relacionados con distintas causas de interferencia, facilitaría la identificación de éstas últimas.

Mediante el cálculo de las relaciones Q_{max}/Q_{lim} y T^2_{max}/T^2_{lim} observamos que el pre-procesado establece una mayor diferencia entre los individuos dentro y fuera de control, produciendo en estos últimos una mayor dispersión que resultará útil de cara a su clasificación.

El hecho de que las longitudes de onda utilizadas por el modelo lineal empleado para las estimaciones de SS (a partir de espectros sin pre-procesar) se encuentren en zonas de máxima variabilidad de los espectros pre-procesados dentro de los límites de control indica que estas zonas podrían resultar informativas para el contenido en SS. λ=947 nm, que está en una zona de variabilidad mínima, podría estar siendo empleada por el modelo lineal para eliminar fuentes de varianza interferente.

4. Conclusiones

El procedimiento de clasificación, si bien ha demostrado su eficacia para la clasificación cualitativa a lo largo de varias campañas, se está viendo afectado por nuevas fuentes interferentes desconocidas.

Para asegurar la viabilidad del sistema de clasificación en el futuro es imprescindible neutralizar los efectos de las nuevas y futuras fuentes de varianza interferente.

El pre-procesado llevado a cabo reduce sensiblemente los efectos interferentes sobre los espectros, aunque no parece que sea suficiente de cara al mantenimiento de la estabilidad de las medidas.

La caracterización de los individuos fuera de control y su clasificación puede resultar de gran utilidad en la identificación de las causas que producen las interferencias.

El control de las causas identificadas, en el caso de que sea posible, y la adaptación de los algoritmos de pre-procesado para la eliminación de la varianza interferente, conocida o no, es necesario para asegurar la viabilidad de operación del sistema.

Las técnicas de pre-procesado pueden ser aplicadas a la generación de bases de datos de calibración más idóneas.

Referencias

[1] R.J. Barnes, M.S. Danoha, S.J. Lister. 1989. Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. Appl. Spectrosc. Vol. (43), 772-777.

[2] P. Barreiro, A. Moya-González, J.I. Robla, M. Ruiz-Altisent. 2005. Analysis of the Effect of Product Temperature on the Segregation of Onions by Means of online NIR Spectrometry. Frutic, 12-16 septiembre 2005. Montpellier (Francia).

[3] P. Barreiro, M. Ruiz-Altisent, C. Bielza, A. Moya-González. 2005. Multivariate Analysis of an On-line NIR Spectrometer under Industrial Use . ISHS Acta Horticulturae 674: 513-519. III International Symposium on Applications of Modelling as an Innovative Technology in the Agri-Food Chain; MODEL-IT, 29 May - 2 June 2005. Leuven, (Bélgica).

[4] P. Barreiro, F. Chauchard, J.M. Roger, A. Moya-González, V. Bellon-Maurel. 2005. Robust modeling for at-line on-line calibration transfer in a NIR industrial application. Chimiométrie 2005. 30 Noviembre a 1 Diciembre. Villeneuve d'Ascq (Francia)

[5] P. Barreiro, E.L. Henche, M. Ruiz-Altisent, N. Hernández, A. Moya-González. 2004. Multivariate diagnosis of the variability of NIR spectrometers under industrial applications. Spanish Journal of Agricultural Research Vol. (2), 485-492.

[6] R. N. Feudale, N. A. Woody, H. Tan, A. J. Myles, S. D. Brown, J. Ferré. 2002. Transfer of multivariate calibration models: a review. Chemometrics and Intelligent Laboratory Systems, Vol (64), Issue 2, 181-192

[7] D.G. Fraser, R.B. Jordan, R. Künnemeyer, V.A. McGlone. 2003. Light distribution inside mandarin fruit during internal quality assessment by NIR spectroscopy. Postharvest Biol Tec Vol. (27), 185-196.

[8] J. Guthrie, B. Wedding, K. Walsh. 1998. Robustness of NIR calibrations for soluble solids in intact melon and pineapple. J Near Infrared Spec Vol. (6), 259-265.

[9] N. Hernández-Sanchez, S. Luron, J.M. Roger, V. Bellon-Maurel. 2003. Robustness of models based on NIR spectra for sugar content prediction in apples. J Near Infrared Spec Vol. (11), 97-107.

[10] B. Nicolai, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K.I. Theron, J. Lammertyn. 2007. Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. Postharvest Biol Tec Vol (46), 99-118.

[11] A. Peirs, J. Tirry, B. Verlinden, P. Darius, B. Nicolai. 2003. Effect of biological variability on the robustness of NIR models for soluble solids content of apples. Postharvest Biol Tec Vol (28), 269-280.

[12] P.N. Schaare, D.G. Fraser. 2000. Comparison of reflectance, interactance and transmission modes of visible-near infrared spectroscopy for measuring internal properties of kiwifruit. Postharvest Biol Tec Vol. (20), 175-184.

[13] A. Simoglou, E. B. Martin, A. J. Morris. 2000. Multivariate statistical process control of an industrial fluidised-bed reactor. Control Engineering Practice, Vol. (8), Issue 8, 893-909

[14] M. Zeatier, J.M. Roger, V. Bellon-Maurel, D.N. Rutledge. 2004. Robustness of models developed by multivariate calibration. Part I. The assessment of robustness. Trends in Analytical Chemistry Vol. (23), 157-170.