

# ESTIMACIÓN DE PROTEÍNA EN SEMOLINA DE ARROZ, MEDIANTE APLICACIÓN DE REGRESIONES EN EL INFRARROJO CERCANO<sup>1</sup>

Oscar Centeno-Mora<sup>2</sup>

## RESUMEN

**Estimación de proteína en semolina de arroz, mediante aplicación de regresiones en el infrarrojo cercano.** El objetivo de este trabajo fue la comparación empírica de las técnicas de regresión por mínimos cuadrados parciales (MCP) y por componentes principales (RCP) en la predicción del porcentaje de proteína para la semolina de arroz. Las estimaciones se realizaron utilizando los valores de absorbancia en la zona del infrarrojo cercano. Se obtuvieron 135 muestras de semolina de arroz recolectadas entre el 2004 y 2012, procedentes de diversas fábricas para la elaboración de alimentos para animales en Costa Rica. Se validó la convergencia de los resultados mediante técnicas de simulación de Bootstrap. Las observaciones se dividieron en dos grupos: un conjunto de datos para la estimación del mejor modelo de regresión (n=120), y un conjunto de datos de validación (n=15). Para el conjunto de datos de estimación, los modelos presentaron dificultades a nivel de los valores extremos, lo cual produjo la eliminación de un valor para obtener el mejor modelo en el caso del MCP. En la validación de los modelos de regresión, los estadísticos de bondad y de ajuste del error estándar de predicción de medias (EEPM), la raíz del error estándar de predicción de medias (REEPM), el error estándar de predicción (EEP), la razón de desviación de predicción (RDP) y gráficos de valores observados contra predichos, confirmaron mejores ajustes para la regresión por MCP (EEP=0,304) respecto al RPC (EEP=0,312). El método de simulación mostró una mejor convergencia en los resultados de la regresión por MCP para predecir el porcentaje de proteína en la semolina de arroz.

**Palabras claves:** regresión por componentes principales, mínimos cuadrados parciales, espectroscopia del infrarrojo cercano, Bootstrap.

## ABSTRACT

**Estimate of protein content in rice semolina by applying regressions in the near infrared region.** The objective of this study was the empirically compare the partial least squares (PLS) regression model and the principal components regression (PCR) model to predict the protein percentage in rice semolina. The estimates were carried out using the absorbance values in the near infrared region. 135 samples of rice semolina were collected between 2004 and 2012 from several pet food plants in Costa Rica. The convergence of the results was validated through Bootstrapping techniques. The observations were split in two groups: one data set to estimate the best regression model (n=120) and a data set of validation (n=15). The models estimated in the data set showed difficulties with outliers, consequently an observation was removed to obtain the best PLS model. In the validation of the regression model, the goodness of fit referred to statistics of the mean standard error of prediction (MSEP), the root mean square error of prediction (RMSEP), the standard error of prediction (SEP), the ratio of performance to deviation (RPD), and the graphics of observed against predicted values confirmed better adjustments for the PLS regression (SEP=0.304) in comparison to the PCR model (SEP=0.312). The simulation method showed a better convergence in the results of the PLS regression technique, to predict the percentage of protein in rice semolina.

**Keywords:** principal components regression, partial least squares, near infrared spectroscopy, Bootstrap.

<sup>1</sup> Recibido: 2 de setiembre, 2015. Aceptado: 10 de noviembre, 2015. Este trabajo formó parte de la tesis de Master Scientiae del autor, en el Centro de Investigación en Nutrición Animal, Universidad de Costa Rica.

<sup>2</sup> Universidad de Costa Rica, Facultad de Ciencias Económicas, Escuela de Estadística. San José, Costa Rica. [oscarcenteno86@gmail.com](mailto:oscarcenteno86@gmail.com)



## INTRODUCCIÓN

La semolina de arroz es un subproducto procedente del grano de arroz entero, que se obtiene a partir de la fricción del grano de arroz, al ser sometido a una máquina “descascaradora” (Campabadal y Murillo, 1985a). Formada principalmente por las capas aleurónicas del grano, la semolina representa todos los pulimentos que se eliminan del grano después de que este es despojado de la cáscara (lema y palea, también conocido popularmente como *casquilla*). Al final de este proceso, se obtiene la semolina de arroz, una harina pajosa de color crema, con composiciones ligeramente grasosas (Campabadal et al., 1982; Campabadal y Murillo, 1985b).

Conocer los parámetros de nutrición o la composición química de la semolina de arroz es fundamental en la elaboración de las dietas destinadas a los animales. La optimización en el crecimiento o engorde de los animales, dependerá de los porcentajes promedio aplicados en las dietas de alimentación (Carmioli et al., 1982). La aplicación de dietas sin conocer la composición química, no solo limitaría los procesos de crecimiento en los animales, sino que también podría tener consecuencias en la digestión de los animales y ocasionar posibles perjuicios en las personas (Sauvant, 2005; Guyomard et al., 2006; FAO, 2010).

Al determinar la composición química de un producto, existen dos enfoques de la química analítica. El primero es el clásico, en el cual se utilizan herramientas tales como pipetas, buretas, matraces, balanzas, entre otros, para así conocer mediante determinados métodos la composición química. El segundo es el instrumental, el cual se basa en la interacción entre la energía y la materia, para lograr caracterizar esta última (Workman et al., 2003; Rouessac, 2004).

La aplicación de la química analítica instrumental necesita conceptos físicos en la interacción de la materia, la delimitación de una zona en el espectro electromagnético, la aplicación de una técnica en la extracción de la información, la utilización de algoritmos para la correcta lectura de la información y la utilización de equipos especiales para la obtención de la composición química de la materia (Bro, 1996; Büning-Pfaue, 2003). En este enfoque, una de las técnicas de extracción es la espectroscopia del infrarrojo cercano, la cual mide la interacción de la luz del infrarrojo cercano con la nube electrónica en los enlaces químicos (Rouessac y Rouessac, 1998; Blanco, 2002).

Esta se fundamenta en la absorción de la radiación infrarroja en un material de análisis, permitiendo, por medio de la detección de vibraciones, caracterizar ciertos enlaces químicos y efectuar los análisis de las funciones químicas expuestas en un producto (Dubernet y Dubernet, 2000; Otto, 2007).

El análisis efectuado por química analítica instrumental permite obtener resultados en un tiempo considerablemente menor con un grado de precisión aceptable. Para la determinación de la composición química, se necesita el análisis empírico de las observaciones extraídas de la interacción de la materia, o la información física (Tippens, 2001). Usualmente, se suelen presentar archivos de datos con más de 500 variables. En el caso de la espectroscopia del infrarrojo cercano, se poseen mediciones de absorbancia en las longitudes de onda, las cuales, por su naturaleza, son medidas que están altamente correlacionadas entre ellas al tratarse de observaciones consecutivas en el espectro electromagnético de la zona del infrarrojo cercano (Kubista et al., 1993; Blanco, 2002).

Los métodos de reducción de variables en las técnicas de regresión, tales como la regresión por componentes principales (RCP) y por mínimos cuadrados parciales (MCP), son dos opciones que permiten analizar datos de absorbancia provenientes de la zona del infrarrojo cercano (Ergon y Esbensen, 2002). Estos métodos brindan una adecuada reducción de la dimensión de las variables, además de crear componentes ortogonales, lo cual eliminaría las dificultades de alta multicolinealidad (Ergon, 2003; Ergon, 2006). El objetivo de este trabajo fue la comparación empírica de las técnicas de regresión por mínimos cuadrados parciales (MCP) y por componentes principales (RCP) en la predicción del porcentaje de proteína para la semolina de arroz.

## MATERIALES Y MÉTODOS

### Materiales

Para la aplicación de las técnicas de regresión, la variable dependiente correspondió a la medición del porcentaje de proteína en la semolina de arroz mediante química húmeda. Las variables independientes se obtuvieron mediante el método de espectroscopia en la zona del infrarrojo cercano; se extrajeron los valores de absorbancia para cada longitud de onda, utilizando un

espectrómetro infrarrojo. Los análisis de regresión se realizaron en SAS 9.4, mediante la PROC SQL, PROC MCP y macro programas para la sección de simulación de Bootstrap.

El presente estudio contó con 135 muestras de semolina de arroz recolectadas entre el 2004 hasta el 2012, procedentes de diversas fábricas para la elaboración de alimentos de nutrición animal en el territorio de Costa Rica, en diferentes épocas del año.

### Métodos de estimación

El método de regresión por componentes principales (RCP) se define como la descomposición de la matriz  $X$  en puntajes  $T$  y cargas factorial  $P$  de la siguiente forma:

$$T = \tilde{B} \times X_p$$

Se regresa la variable dependiente  $Y$  respecto a una matriz ortogonal  $X_j$ , produciendo una disminución de las variables que corresponden a las primeras  $j$  columnas de los puntajes  $T$ . La descomposición se expresa como:

$$X_j \approx X_j + \varepsilon_X = T_a P'_a + \varepsilon_X$$

En este caso,  $T$  representa los puntajes,  $P$  el vector de cargas factorial y  $\varepsilon$  el vector de error específico. Los valores de  $X_j$  logran describir lo mejor posible la matriz  $X_p$ , tal como en el análisis de componentes principales (Beebe, 1987).

Al regresar la  $Y$  en los puntajes  $T$  y el cambio de base  $B$ , la ecuación para los coeficientes de regresión se expresa como:

$$\tilde{B} = P(T'T)^{-1}(T'Y)$$

Para el método de regresión por MCP, se prosigue de la siguiente forma:

1. Se define el valor de la matriz  $X$ .
2. Se define el valor de la matriz  $Y$ .
3. Se define el valor de la matriz  $S$ .

$$S = X'Y$$

4. Se calculan los vectores singulares de la Matriz  $S$ ; el primer vector izquierdo y derecho,  $w$  y  $q$  (los pesos para la matriz  $X$  y vector  $Y$ ), son usados como ponderaciones para  $X$  y para  $Y$ , y de esta forma se obtienen los puntajes  $t$  y  $u$ , tal que:

$$t = Xw = Ew$$

$$u = Yq = Fq$$

$E$  y  $F$  son los valores inicializados de  $X$  y de  $Y$  respectivamente, expresados como:

$$X_0 = E$$

$$Y_0 = F$$

5. Los puntajes  $t$  de  $X$ , se normalizan, tales que:

$$t = t / \sqrt{(t't)}$$

6. Seguidamente se obtienen las cargas iniciales de  $X$  y de  $Y$  regresando esta contra el vector de  $t$ , para así obtener los nuevos vectores  $p$  y  $q$ :

$$p = E't$$

$$q = F't$$

La matriz de datos debe ser corregida: la información relacionada con las variables latentes que provienen de la forma de los productos  $tp'$  y  $tq'$ , es restada de los datos actuales de la matriz  $E$  y  $F$ , tal que:

$$E_{n+1} = E_n - tp'$$

$$F_{n+1} = F_n - tq'$$

7. Las estimaciones del próximo componente comienzan con la descomposición del valor singular del producto de la matriz  $E_{n+1}$  y  $F_{n+1}$ . Es decir:

$$S = X'_1 Y_1 = (E_{n+1})'(F_{n+1})$$

En cada iteración los vectores  $w$ ,  $t$ ,  $p$  y  $q$  son almacenados como columnas en matrices  $W$ ,  $T$ ,  $P$  y  $Q$ , respectivamente.

8. Las columnas de la matriz  $W$  no pueden ser comparadas directamente y deben ser derivadas de la corrección sucesiva de las matrices  $E$  y  $F$ . Se relacionan los valores de todas las columnas de la matriz  $X$ , y luego se establece la siguiente descomposición:

$$R = W(P'W)^{-1}$$

9. Se llega a los mismos resultados del RCP: en lugar de  $Y$  con  $X$ , se utilizan los puntajes  $T$  para calcular

los coeficientes de regresión, luego se convierten estos de vuelta al campo de las variables originales, pre multiplicando estos por la matriz  $R$  (dado que  $T = XR$ ). Finalmente, la estimación de los coeficientes estaría dada por la función (Tenenhaus, 1995; Mevik y Cederkvist, 2004):

$$\tilde{B} = R(T'T)^{-1}T'Y$$

La diferencia en la estimación de los betas es que en el RCP se utiliza la matriz  $P$ , mientras que en el MCP se utiliza la matriz  $R$ , la cual incluye la información de la variable dependiente  $Y$  en los procesos de estimación.

### Metodología de análisis

El análisis de regresión del porcentaje de proteína en la semolina de arroz se llevó a cabo de la siguiente forma. Primero se dividió el archivo de datos en dos partes: un conjunto para estimar el modelo de regresión<sup>3</sup>, y el otro conjunto para validar los modelos estimados<sup>4</sup> (Martens y Naes, 1989). En la estimación de los modelos de regresión se definió previamente el número de componentes a ser utilizados. Seguidamente, se analizaron los supuestos del modelo, para lo cual se realizaron los diagnósticos respectivos. Diversos modelos eran candidatos en ambas técnicas de regresión, por lo que se calcularon estadísticos de bondad y de ajuste para definir el mejor modelo de regresión según cada técnica. Posteriormente, se realizaron las predicciones del conjunto de validación. Se calcularon los estadísticos de bondad y de ajuste para cada uno de los conjuntos de validación con el fin verificar cuál modelo brindó los mejores resultados (Neter et al., 1996). Finalmente, al poseer una pequeña muestra de casos, se verificó la convergencia de los estadísticos de bondad y de ajuste. Para esto se realizaron simulaciones de Bootstrap modificando a su vez el número de componentes con el objetivo de comprobar la sensibilidad de los resultados (Efron y Gong, 1983; Efron et al., 2004).

### Estadísticos de bondad y de ajuste

Para la selección del número de componentes, se utilizaron los estadísticos de la raíz media del

<sup>3</sup> Denominado próximamente como conjunto de entrenamiento.

<sup>4</sup> Denominada próximamente como conjunto de validación.

PRESS, la  $T$  de Hotelling y el  $R^2$ . Estos se definen a continuación (Tenenhaus, 1998):

$$\text{Raíz media PRESS} = \sum_{i=1}^n (y_i - \hat{y}_{i,-i})^2$$

$$R^2 = 1 - \frac{SS_E}{SS_T} = \frac{SS_R}{SS_T}$$

$$T^2 = (\bar{X} - \bar{Y})' \left[ S \left( \frac{1}{n_x} + \frac{1}{n_y} \right) \right]^{-1} (\bar{X} - \bar{Y})$$

Por otra parte, los estadísticos de bondad y de ajuste utilizados son el error estándar de predicción media (EPPM<sup>5</sup>), la raíz del error estándar de predicción media (REPPM<sup>6</sup>), el error estándar de predicción (EEP<sup>7</sup>) y la razón de desviación de predicción (RDP<sup>8</sup>). Cada estadístico se define según la siguiente fórmula (Mevik y Cederkvist, 2004):

$$MSEP = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} = \frac{\sum_{i=1}^n e_i^2}{n}$$

$$RMSEP = \sqrt{MSEP} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n}}$$

$$SEP = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)}{n-1}} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-1}}$$

$$RPD = \frac{\text{Desviación estándar}}{SEP}$$

## RESULTADOS Y DISCUSIÓN

### Selección del número de componentes

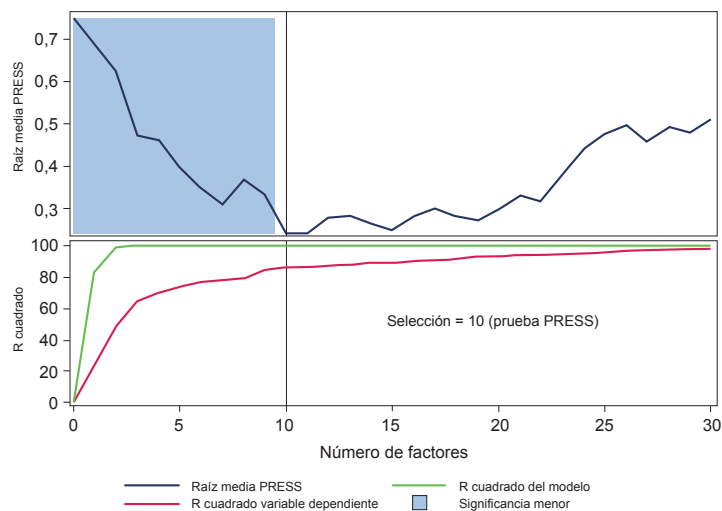
El estadístico de Hotelling y el menor valor de PRESS indicaron mejores soluciones para el modelo por MCP que se plantea para once componentes principales. Sin embargo, la Figura 1 muestra que la solución óptima argumentó a favor de diez componentes principales estimados; análisis

<sup>5</sup> Del inglés como Mean Standard Error Prediction.

<sup>6</sup> Del inglés como Root Mean Standard Error Prediction.

<sup>7</sup> Del inglés como Standard Error Prediction.

<sup>8</sup> Del inglés como Ratio Prediction Deviation.

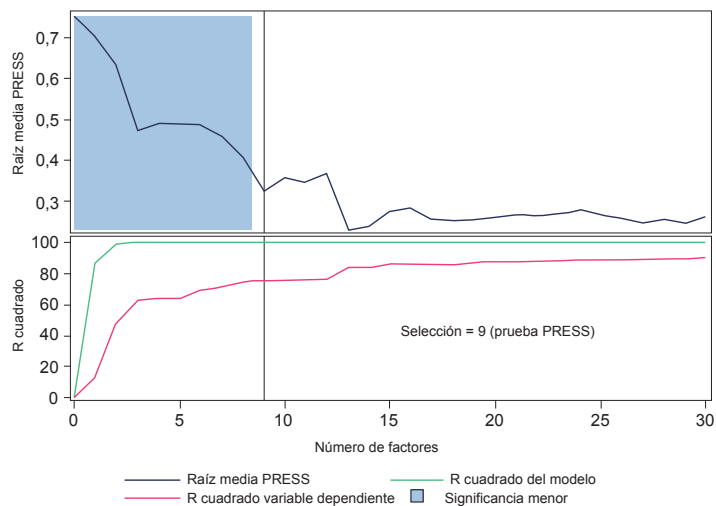


**Figura 1.** Selección del número de componentes en el modelo por mínimos cuadrados parciales (MCP) para la estimación del porcentaje de proteína en la semolina de arroz. Costa Rica. 2012.

**Figure 1.** Selection of the number components in the partial least squares (PLS) model to predict the protein percentage in rice semolina. Costa Rica. 2012.

posteriores (diagnósticos y estadísticos de bondad y de ajuste) apoyaron la solución de once. Una estimación del modelo con once componentes brindó un porcentaje de variancia explicada del 92,2%. Para el caso del RCP

el estadístico de Hotelling y el menor valor de PRESS, indicaron que las mejores soluciones se hallaron en trece componentes principales. La Figura 2 muestra que la solución óptima fue para nueve. Análogamente,



**Figura 2.** Selección del número de componentes en el modelo de regresión por componentes principales (RCP) para la estimación del porcentaje de proteína en la semolina de arroz. Costa Rica. 2012.

**Figure 2.** Selection of the number of components in the principal components regression model (PCR) to predict the protein percentage in rice semolina. Costa Rica. 2012.

análisis posteriores favorecieron la solución de dos cifras. Una solución de trece componentes brindó un porcentaje de variancia explicada del 85,1%. Por consiguiente, los modelos por MCP y RCP se estimaron para once y trece componentes, respectivamente.

### Diagnóstico de los modelos de regresión

En el caso de la regresión por MCP, los diagnósticos del modelo presentaron dificultades para el análisis de los valores extremos. El análisis de normalidad, heteroscedasticidad y linealidad fueron satisfactorios, al igual que todos los diagnósticos para el caso de la regresión por RCP. Sin embargo, se constató que una misma observación constituyó un posible punto de influencia en ambos modelos de regresión. No se utilizó ninguna medida remedial dado el cumplimiento satisfactorio de los supuestos. No obstante, se decidió estimar los modelos de regresión sin el valor extremo y con todos los casos en ambos métodos de regresión, para verificar la posible mejora a nivel de los estadísticos de bondad y de ajuste.

### Valoración de los primeros modelos estimados

En el caso del método por MCP, se observó cómo los estadísticos del EEPM, REEPM y el EEP atribuyeron los resultados óptimos al modelo con once componentes sin la observación extrema (Cuadro 1). Este modelo fue utilizado para predecir las observaciones del conjunto de validación. Por otra parte, de acuerdo con el método por RCP los mejores resultados, según los estadísticos de EEPM, EEPM y EEP, fueron para un modelo con trece componentes, con todas las observaciones (Cuadro 2). De esta forma, este modelo se utilizó para predecir los valores del conjunto de validación.

### Estadísticas de bondad y de ajuste en el conjunto de validación

De acuerdo con las Figuras 3 y 4 de los valores observados contra los valores predichos, se observó que ambos modelos de regresión estimaron correctamente los valores del porcentaje de proteína en el conjunto de validación. Todas las observaciones se encontraban dentro de los límites estadísticos; además, la totalidad de las predicciones fueron inferiores al EEP establecido como límite de confianza. Por otra parte,

**Cuadro 1.** Medidas de la bondad del ajuste por mínimos cuadrados parciales (MCP) para estimar el porcentaje de proteína en la semolina de arroz en el conjunto de entrenamiento. Costa Rica. 2012.

**Table 1.** Goodness of fit measures of partial least squares (PLS) to estimate the percentage of protein in rice semolina in the regression model estimation set. Costa Rica. 2012.

Estadístico	Todas las observaciones	Sin la observación extrema
n	110	109
EEPM	0,254	0,220
REEPM	0,504	0,469
EEP	0,506	0,471

n: número de muestras, EEPM: error estándar de predicción media, REEPM: raíz del error estándar de predicción media, EEP: error estándar de predicción / n: sample size, MSEPM: mean standard error of prediction, RMSEP root mean square error of prediction, SEP: standard error of prediction.

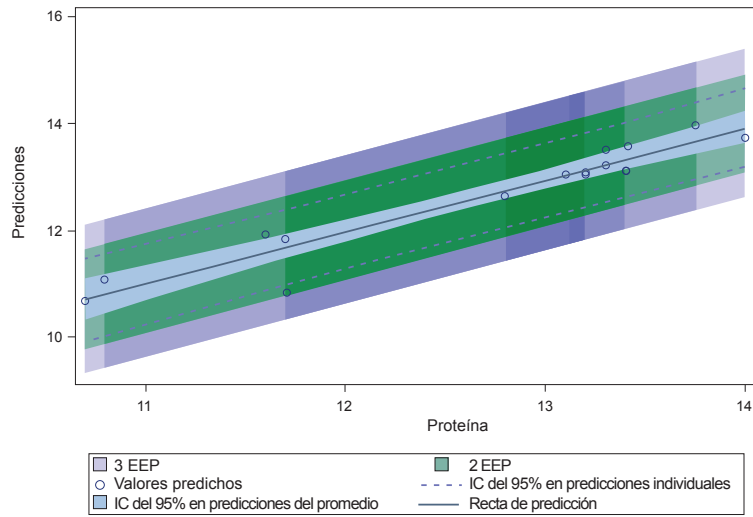
**Cuadro 2.** Medidas de bondad de ajuste en el conjunto de entrenamiento para determinar el mejor modelo en el método de regresión por componentes principales (RCP), para la estimación del porcentaje de proteína en la semolina de arroz. Costa Rica. 2012.

**Table 2.** Goodness of fit measures in regression model estimation set to determine the best model in principal components regression (PCR) model to predict the percentage of protein in rice semolina. Costa Rica. 2012.

Estadístico	Todas las observaciones	Sin la observación extrema
n	110	109
EEPM	0,269	0,302
REEPM	0,518	0,550
EEP	0,521	0,552

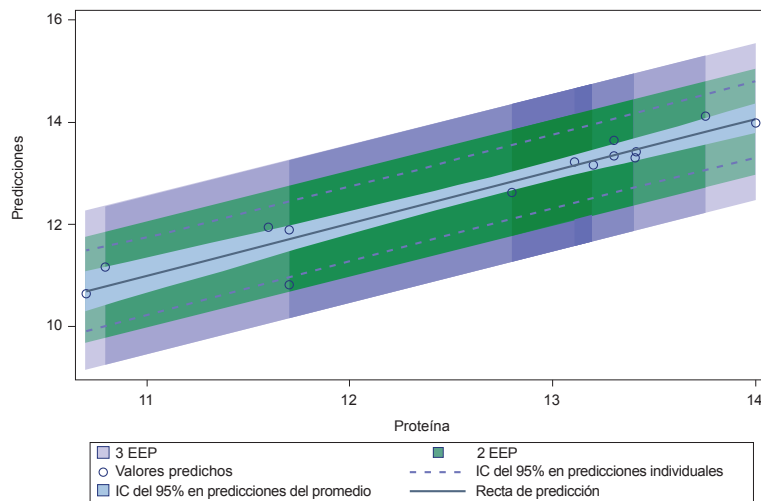
n: número de muestras, EEPM: error estándar de predicción media, REEPM: raíz del error estándar de predicción media, EEP: error estándar de predicción / n: sample size, MSEPM: mean standard error of prediction, RMSEP root mean square error of prediction, SEP: standard error of prediction.

los estadísticos de bondad y de ajuste para el conjunto de validación en la comparación de los métodos por MCP y RCP, constataron que los estadísticos de EEPM, REEPM y EEP presentaron mejores valores con el método de MCP (Cuadro 3). De igual forma, el



**Figura 3.** Valores observados contra valores predichos en el conjunto de validación para la predicción del porcentaje de proteína en la semolina de arroz por mínimos cuadrados parciales (MCP). Costa Rica. 2012.

**Figure 3.** Observed against predicted values in the validation of the estimated models for the prediction of protein percentage in rice semolina in the partial least squares (PLS) model. Costa Rica. 2012.



**Figura 4.** Valores observados contra valores predichos en el conjunto de validación para la predicción del porcentaje de proteína en la semolina de arroz para la regresión por componentes principales (RCP). Costa Rica. 2012.

**Figure 4.** Observed against predicted values in the validation of the estimated models for the prediction of protein percentage in rice semolina in the principal components regression (PCR) model. Costa Rica. 2012.



**Cuadro 3.** Medidas de bondad y de ajuste para la estimación del porcentaje de proteína de la semolina de arroz. en el conjunto de validación según los métodos mínimos cuadrados parciales (MCP) y regresión por componentes principales (RCP), Costa Rica. 2012.

**Table 3.** Goodness of fit measures to predict the percentage of protein in rice semolina on the validation of the estimated models according to the partial least squares (PLS) model and principal components regression (PCR) model. Costa Rica. 2012.

Estadístico	MCP	RCP
n	15	15
DS	1,068	1,068
EEPM	0,086	0,091
REEPM	0,294	0,302
EEP	0,304	0,312
RDP	3,503	3,416

n: número de muestras, DS: desviación estándar, EEPM: error estándar de predicción media, REEPM: raíz del error estándar de predicción media, EEP: error estándar de predicción, RPD: razón de predicción predicha / n: sample size, SD: standard deviation, MSEP: mean standard error prediction, RMSEP: root mean square error prediction, SEP: standard error of prediction, RPD: ratio of performance to deviation.

RDP argumentó a favor del MCP como mejor método de estimación en el conjunto de validación.

### Análisis de convergencia de los estadísticos: simulaciones por Bootstrap

Los resultados de convergencia de los dos métodos de regresión indicaron que, en el caso del MCP, a partir de un tamaño de 50 000 todos los estadísticos del EEP se estabilizaron (Cuadro 4). El valor del EEP simulado fue concordante con el valor estimado (EEP=0,471); en el caso del RCP, también a partir de 50 000 re-muestreos todos los estadísticos del EEP se estabilizaron (Cuadro 5). El valor del EEP simulado por el Bootstrap fue ligeramente sobrestimado de 0,02 con respecto al EEP estimado.

Dada la relación lineal de la variable dependiente y los componentes ortogonales creados a partir de las regresiones por MCP y RCP, la estimación y el arreglo de ciertas características en los modelos de regresión fue suficiente para obtener resultados satisfactorios en las predicciones de ambos métodos, corroborando desde diversos enfoques analíticos, para la presente aplicación, la mejor estimación del porcentaje de proteína en la semolina de arroz para el método por MCP. Las múltiples relaciones que se pueden dar

**Cuadro 4.** Análisis de convergencia para el estadístico del error estándar de predicción (EEP) en el modelo por mínimos cuadrados parciales (MCP). Costa Rica. 2012.

**Table 4.** Analysis of convergence for the statistic of the standard error of prediction (SEP) in the partial least squares (PLS) model. Costa Rica. 2012.

Tamaño de la muestra*	Número de componentes principales						
	6	8	10	11	12	14	16
1	0,661	0,625	0,485	0,444	0,417	0,421	0,413
5	0,599	0,606	0,486	0,491	0,450	0,415	0,412
10	0,630	0,606	0,491	0,461	0,459	0,414	0,389
20	0,617	0,584	0,477	0,462	0,425	0,399	0,380
50	0,637	0,597	0,472	0,468	0,454	0,421	0,399
100	0,643	0,596	0,478	0,470	0,445	0,409	0,396
1000	0,640	0,592	0,475	0,471	0,447	0,411	0,391
50000	0,639	0,592	0,475	0,471	0,447	0,411	0,392
100000	0,639	0,592	0,475	0,471	0,448	0,411	0,392
500000	0,639	0,592	0,475	0,471	0,448	0,411	0,392
1000000	0,639	0,592	0,475	0,471	0,448	0,411	0,392



**Cuadro 5.** Análisis de convergencia para el estadístico del error estándar de predicción (EEP) en el modelo de regresión por componentes principales (RCP). Costa Rica. 2012.

**Table 5.** Analysis of convergence for the statistic of standard error of prediction (SEP) in the principal components regression (PCR) model. Costa Rica. 2012.

Tamaño de la muestra	Número de componentes principales						
	8	10	12	13	14	16	18
1	0,664	0,615	0,651	0,492	0,556	0,533	0,521
5	0,667	0,681	0,639	0,523	0,501	0,512	0,500
10	0,680	0,674	0,647	0,542	0,519	0,464	0,487
20	0,694	0,675	0,657	0,524	0,506	0,492	0,490
50	0,660	0,676	0,658	0,516	0,511	0,479	0,475
100	0,678	0,658	0,649	0,518	0,526	0,490	0,480
1000	0,679	0,659	0,654	0,520	0,519	0,485	0,479
50000	0,679	0,659	0,655	0,519	0,518	0,484	0,479
100000	0,679	0,659	0,655	0,519	0,518	0,484	0,479
500000	0,679	0,659	0,655	0,519	0,518	0,484	0,479
1000000	0,679	0,659	0,655	0,519	0,518	0,484	0,479

en los estudios NIR han provocado el desarrollo constante de nuevas técnicas de estimación, aún más reciente que el MCP y RCP, así como otros métodos para poder asociar los datos químicos y físicos. Las técnicas más novedosas de regresión fueron capaces de determinar cualquier tipo de relación asimétrica en las variables, aunque no se pudo interpretar la relación obtenida por las ecuaciones (Dos Santos, 2009).

## LITERATURA CITADA

- Beebe, K. 1987. An introduction to multivariate calibration and analysis. *Anal. Chem.* 59:1007A-1017A.
- Blanco, M. 2002. NIR spectroscopy: a rapid-response analytical tool. *Anal. Chem.* 21:240-250.
- Bro, R. 1996. Multiway calibration - Multilinear PLS. *J. Chemom.* 10:47-61.
- Büning-Pfaue, H. 2003. Analysis of water in food by near infrared spectroscopy. *Food Chem.* 82:107-115.
- Campabadal, C., y M. Murillo. 1985a. Efecto de la adulteración de la semolina de arroz con carbonato de calcio en la alimentación de pollos de engorde. *Agron. Costarricense* 9(1):13-20.
- Campabadal, C., y M. Murillo. 1985b. Utilización de la semolina de arroz en la alimentación de gallinas en desarrollo de postura. *Agron. Costarricense* 9(1):13-20.
- Campabadal, C., M. Murillo, y J. Solís. 1982. Utilización de la semolina de arroz en dietas para pollos parrilleros con y sin suplementación de grasa. *Agron. Costarricense* 6(1/2):73-79.
- Carmiol, G., C. Campabadal, y M. Zumbado. 1982. Utilización de la semolina de arroz en la alimentación de pollos parrilleros. Adulteración con cascarilla de arroz. *Agron. Costarricense* 6(1/2):65-72.
- Dos Santos, R. 2009. Développement de nouvelles méthodes chimiométriques d'analyse. Ph.D. Tesis. Agro Paris Tech., Paris, FRA.
- Dubernet, M., et M. Dubernet. 2000. Utilisation de l'analyse infrarouge à transformée de Fourier pour l'analyse œnologique de routine. *Française d'Œnologie* 181:3914-3917.
- Ergon, R. 2003. Constrained numerical optimization of PCR/PLSR predictors. *Chemom. Intell. Lab. Syst.* 65:293-303.
- Ergon, R. 2006. Reduced PCR/PLSR models by subspace projections. *Chemom. Intell. Lab. Syst.* 81:68-73.
- Ergon, R., and K. Esbensen. 2002. PCR/PLSR optimization based on noise covariance estimation and Kalman filtering theory. *J. Chemom.* 16:401-407.
- Efron, B., and G. Gong. 1983. A leisurely look at the Bootstrap, the jackknife, and cross-validation. *Am. Stat.* 37:36-48.
- Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. Least angle regression (with discussion). *Ann. Stat.* 32:407-499.

- FAO. 2010. L'état de l'insécurité alimentaire dans le monde. Roma. <http://www.fao.org/docrep/013/i1683f/i1683f.pdf> (consulté 20 ago. 2015).
- Guyomard, H., D. Aubert, et S. Jumel. 2006. Entreprises et filières agro-alimentaires face à de nouveaux enjeux. Département Sciences Sociales, Agriculture et Alimentation, espace et environnement de l'Institut National de la Recherche Agronomique, FRA.
- Kubista, M., R. Sjoback, and B. Albinsson. 1993. Determination of equilibrium constants by chemometric analysis of spectroscopic Data. *Anal. Chem.* 65:994-998.
- Martens, H., and T. Naes. 1989. *Multivariate calibration*. Wiley, NY, USA.
- Mevik, B., and H. Cederkvist. 2004. Mean squared error of prediction (MSEP) estimates for principal component regression PCR and partial least squares regression (PLSR). *J. Chemom.* 18:422-429.
- Neter, J., M. Kutner, C. Nachtsheim, and W. Wasserman. 1996. *Applied linear statistical models*. 4<sup>th</sup> ed. McGraw-Hill, NY, USA.
- Otto, M. 2007. *Chemometrics: statistics and computer application in analytical chemistry*. 2<sup>nd</sup> ed. Wiley, NY, USA.
- Rouessac, F. 2004. *Analyse chimique: méthodes et techniques instrumentales modernes*. 6<sup>me</sup> éd. Dunod, Paris, FRA.
- Rouessac, F., et P. Rouessac. 1998. *Analyse chimique: méthodes et techniques instrumentales modernes. Cours et exercices résolus*. 4<sup>me</sup> éd. Dunod, Paris, FRA.
- Sauvant, D. 2005. *Principes généraux de l'alimentation animale*. Institut National Agronomique Paris-Grignon (INAP-G), Paris, FRA.
- Tenenhaus, M. 1995. A partial least squares approach to multiple regression, redundancy analysis, and canonical analysis. *Les cahiers de la recherche de HEC*, Paris, FRA.
- Tenenhaus, M. 1998. *La régression PLS: Théorie et pratique*. Editions Technip, Paris, FRA.
- Tippens, P. 2001. *Física. Conceptos y aplicaciones*. 6<sup>th</sup> ed. McGraw Hill, NY, USA.
- Workman, J., M. Koch, and D. Veltkamp. 2003. Process analytical chemistry. *Anal. Chem.* 75:2859-2876.