

## Validation of models with proportional bias

### Validación de modelos con sesgo proporcional

Salvador Medina-Peralta<sup>1</sup> M.Sc, Luis Vargas-Villamil<sup>2\*</sup> Ph.D,  
Luis Colorado-Martínez<sup>1</sup> M.Sc, Jorge Navarro-Alberto<sup>3</sup> Ph.D.

<sup>1</sup>Universidad Autónoma de Yucatán, Facultad de Matemáticas, Apartado Postal 172, C.P. 97119, Mérida, Yucatán, México. <sup>2</sup>Colegio de Postgraduados, Campus Tabasco, Periférico Carlos A. Molina Km. 3.5, Apartado Postal 24, C.P. 86500, Cárdenas, Tabasco, México. <sup>3</sup>Universidad Autónoma de Yucatán, Facultad de Medicina Veterinaria y Zootecnia, Apartado Postal 4-116 Itzimná, C.P. 97100, Mérida, Yucatán, México. \*Correspondence: [luis@avanzavet.com](mailto:luis@avanzavet.com)

Received: May 2016; Accepted: November 2016.

#### ABSTRACT

**Objective.** This paper presents extensions to Freese's statistical method for model-validation when proportional bias (PB) is present in the predictions. The method is illustrated with data from a model that simulates grassland growth. **Materials and methods.** The extensions to validate models with PB were: the maximum anticipated error for the original proposal, hypothesis testing, and the maximum anticipated error for the alternative proposal, and the confidence interval for a quantile of error distribution. **Results.** The tested model had PB, which once removed, and with a confidence level of 95%, the magnitude of error does not surpass 1225.564 kg ha<sup>-1</sup>. Therefore, the validated model can be used to predict grassland growth. However, it would require a fit of its structure based on the presence of PB. **Conclusions.** The extensions presented to validate models with PB are applied without modification in the model structure. Once PB is corrected, the confidence interval for the quantile 1- $\alpha$  of the error distribution enables a higher bound for the magnitude of the prediction error and it can be used to evaluate the evolution of the model for a system prediction.

**Keywords:** Bias, confidence intervals, errors, models, validity (*Source: CAB thesaurus, MeSH*).

#### RESUMEN

**Objetivo.** En este trabajo se presentan extensiones al método estadístico de Freese para validar modelos con sesgo proporcional (SP) en sus predicciones y se ilustra el método con datos provenientes de un modelo de simulación de crecimiento de praderas. **Materiales y métodos.** Las extensiones para validar un modelo con SP fueron: el error máximo anticipado para el planteamiento original, la prueba de hipótesis y error máximo anticipado para el planteamiento alternativo, y el intervalo de confianza para un cuantil de la distribución de los errores. **Resultados.** El modelo evaluado presentó SP, una vez removido y con un nivel de confianza del 95% la magnitud del error no sobrepasa 1225.564 kg ha<sup>-1</sup>. Por lo que el modelo validado podría usarse para predecir el crecimiento de praderas, sin embargo, requerirá un ajuste en su estructura con base a la presencia de SP. **Conclusiones.** Las extensiones presentadas para validar modelos en presencia de SP se aplican sin que el modelo sea modificado en su estructura. El intervalo de confianza para el cuantil 1- $\alpha$  de la distribución de los errores una vez que se corrige el SP, permite determinar una cota superior para la magnitud del error de predicción y usarla para evaluar la evolución del modelo en predicción del sistema.

**Palabras clave:** Errores, intervalos de confianza, modelos, sesgo, validez (*Fuentes: CAB thesaurus, MeSH*).

## INTRODUCTION

Validation of a system prediction model consists of comparing the predictions of the model with observed values of the actual system to determine its predictive capacity by means of some method. In this stage of the mathematical modeling process, accuracy and precision of the model are assessed. Accuracy refers to the proximity of the predictions ( $z$ ) to observed values ( $y$ ), for example, their differences ( $d=y-z$ ) from zero. Precision refers to dispersion of the points ( $z,y$ ). However, in the presence of accuracy, precision is measured by quantifying the dispersion of the points with respect to a reference, for example, the deterministic line  $y=z$ , or by evaluating the variances of the differences ( $\sigma_D^2$ ) around zero ( $\mu_D=0$ ).

In the literature, different techniques for validating models designed for purposes of prediction have been proposed; see Tedeschi (1): linear regression analysis, fitted error analysis, concordance correlation coefficient, diverse measures of deviation, mean square error of prediction, non-parametric analyses, and comparing the distribution of observed and predicted data. Medina-Peralta et al (2) point out that some deviation measures to validate models contradict graphic methods when the model predictions are biased. They recommend the joint use of deviation measurement and graphic methods for model validation.

Among the statistical inference techniques dealing with validation of models with no bias (NB), and with constant (CB), or proportional bias (PB) in their predictions, the procedure given by Freese (3) stands out. This method consists in determining whether accuracy and precision of a prediction model, or technique meets the requirements of the model developer or user. Medina et al (4) extended the Freese method when the model has CB in its predictions: hypothesis tests and maximum anticipated error for the alternative proposal, and the confidence interval for a quantile of the error distribution.

This paper presents extensions to the statistical method for model validation proposed by Freese (3), Rennie and Wiant (5), Reynolds (6), Barrales et al (7), and Medina et al (4) when the model has PB in its predictions: 1) maximum anticipated error for the original proposal, 2) hypothesis testing, and maximum anticipated error for the alternative proposal, and 3) the interval of confidence for a quantile of error distribution. The method is illustrated with published data of Barrales et al (7) corresponding to a model that simulates grassland growth.

## INTRODUCCIÓN

La validación de un modelo en predicción del sistema es la comparación por medio de algún método de las predicciones del modelo con los valores observados del sistema real para determinar su capacidad predictiva. En esta etapa del proceso de modelación matemática, se evalúan la exactitud y precisión del modelo; la primera se refiere a la proximidad de las predicciones ( $z$ ) con los valores observados ( $y$ ), por ejemplo, sus diferencias ( $d=y-z$ ) del cero y la segunda a la dispersión de los puntos ( $z,y$ ). Sin embargo, en presencia de exactitud la precisión se mide cuantificando la dispersión de dichos puntos respecto a una referencia, por ejemplo, la recta determinística  $y=z$ , o bien, evaluar la varianza de las diferencias ( $\sigma_D^2$ ) alrededor del cero ( $\mu_D=0$ ).

En la literatura se han expuesto diferentes técnicas para validar modelos diseñados con propósitos predictivos; Tedeschi (1), presenta: análisis de regresión lineal, análisis de los errores ajustados, coeficiente de correlación de concordancia, diversas medidas de desviación, cuadrado medio del error de predicción, análisis no paramétricos y la comparación de la distribución de los datos observados y predichos. Medina-Peralta et al (2), señalan que algunas medidas de desviación para validar modelos se contradicen con los métodos gráficos cuando el modelo presenta sesgo en sus pronósticos; recomiendan utilizar conjuntamente medidas de desviación y métodos gráficos para validar modelos.

Entre las técnicas estadísticas inferenciales que aborda la validación de modelos sin sesgo (SS) y con sesgo constante (SC) o proporcional (SP) en sus pronósticos, se encuentra la propuesta de Freese (3), la cual consiste fundamentalmente en determinar si la exactitud y precisión de un modelo o técnica de estimación satisface los requerimientos del modelador o usuario del modelo. Medina et al (4) extendieron el método de Freese cuando el modelo presenta SC en sus predicciones: prueba de hipótesis y error máximo anticipado para el planteamiento alternativo, y el intervalo de confianza para un cuantil de la distribución de los errores.

En este trabajo se presentan extensiones al método estadístico para validar modelos planteados por Freese (3), Rennie y Wiant (5), Reynolds (6), Barrales et al (7) y Medina et al (4) cuando el modelo presenta SP en sus predicciones: 1) el error máximo anticipado para el planteamiento original, 2) la prueba de hipótesis y error máximo anticipado para el planteamiento alternativo y 3) el intervalo de confianza para un cuantil de la distribución de los errores. Se ilustra el método

## MATERIALS AND METHODS

**Basic concepts.** Assume that we have "n" pairs to compare ( $y_i, z_i$ )  $i=1,2,\dots,n$ , where for the  $i^{\text{th}}$  pair,  $y_i$  is the observed value,  $z_i$  the corresponding predicted value for the deterministic model to be validated, and  $d_i=y_i-z_i$  the difference between the two values. In the development of extensions of the method, the observed values are considered as realizations of random variables  $Y_i$ , and the predicted values are deterministic, thus  $D_i=Y_i-z_i$ .

In Freese's (3) approach applied to determining the required accuracy and precision, the "e" values (maximum admitted error of the deviations  $|y_i-z_i|=|d_i|$ ) and  $\alpha$  ( $1-\alpha$  represents the required certainty) specified by the model developer or user, it is needed that D be normally distributed with zero mean and  $P(|D|\leq e)\geq 1-\alpha$ , in order to accept the model and consider it as sufficiently reliable for system prediction. Therefore, a model is exact, or with no bias (NB), when the differences ( $d_i$ ) fit a normal distribution with mean zero.

Medina, Vargas-Villamil (4) indicate that CB is recognized by an average value of the differences ( $\bar{d}$ ) distant from zero and that the graph of the points ( $z_i, d_i=y_i-z_i$ ) forms a horizontal band centered at  $\bar{d}$  with systematic distribution either positive or negative (points above and below the line  $d=\bar{d}$ ). When the samples are related, a t-test applied to determine if the mean of the differences is significantly different from zero, would prove the presence of CB in the model predictions.

**Determination of proportional bias.** PB is recognized in the graph of points ( $z_i, d_i=y_i-z_i$ ) whenever a positive or negative linear trend is found; the magnitude of the bias ( $d_i=y_i-z_i$ ) increases or decreases in direct proportion to the predicted values ( $z_i$ ). A simple linear regression analysis of the bias vs predicted ( $D_i=\beta_0+\beta_1z_i+\varepsilon_i$ ) contributes to the detection of PB more objectively (2,8).

**Validation of models with proportional bias.** Given that we have PB, the points ( $z_i, D_i$ )  $i=1,2,\dots,n$  are related by  $E(D_i)=\beta_0+\beta_1z_i$  and  $Var(D_i)=\sigma_D^2=\sigma^2$  with  $D_i\sim NI(\beta_0+\beta_1z_i, \sigma^2)$ . That is,  $D_i=\beta_0+\beta_1z_i+\varepsilon_i$ , where  $E(\varepsilon_i)=0$ ,  $Var(\varepsilon_i)=\sigma^2$  and  $\varepsilon_i=D_i-(\beta_0+\beta_1z_i)=[D_i-E(D_i)]\sim NI(0, \sigma^2)$  and thus, the model will be accurate when carrying out the correction  $\varepsilon_i=D_i-(\beta_0+\beta_1z_i)$ . In practice, a regression model is fit and the errors ( $\varepsilon_i$ ):  $d_i-\hat{d}_i=d_i-(a+bz_i)$  are estimated, where a and b are the least squares estimates of the intercept ( $\beta_0$ ) and the slope ( $\beta_1$ ), respectively. Thus, we

con datos previamente publicados por Barrales et al (7), correspondientes a un modelo de simulación de crecimiento de praderas.

## MATERIALES Y MÉTODOS

**Conceptos básicos.** Suponga que se tienen "n" pares para comparar ( $y_i, z_i$ )  $i=1,2,\dots,n$  donde para el  $i$ -ésimo par,  $y_i$  es el valor observado,  $z_i$  el valor predicho correspondiente por el modelo determinístico que será validado y  $d_i=y_i-z_i$  la diferencia entre ellos. En el desarrollo de las extensiones al método, los valores observados son considerados realizaciones de variables aleatorias  $Y_i$  y los valores predichos como determinísticos, así  $D_i=Y_i-z_i$ .

En el planteamiento de Freese (3) para la determinación de la exactitud y precisión requerida, se necesita que los valores "e" (máximo error admitido de las desviaciones  $|y_i-z_i|=|d_i|$ ) y  $\alpha$  ( $1-\alpha$  representa el nivel de certidumbre requerida) especificados por el modelador o usuario del modelo satisfagan, respectivamente, que D se distribuya normal con media cero y  $P(|D|\leq e)\geq 1-\alpha$ , para que el modelo sea aceptable y éste sea considerado suficientemente confiable para predicción del sistema. Por lo que un modelo es exacto o sin sesgo (SS) cuando las diferencias ( $d_i$ ) se ajustan a una distribución normal con media cero.

Medina, Vargas-Villamil et al (4) indican que el SC es reconocido por un valor del promedio de las diferencias ( $\bar{d}$ ) alejado del cero, y que el gráfico de los puntos ( $z_i, d_i=y_i-z_i$ ) forme una banda horizontal centrada alrededor de  $\bar{d}$  con una distribución sistemática a ser positivos o negativos (puntos arriba y debajo de la recta  $d=\bar{d}$ ); una prueba T cuando las muestras son relacionadas para determinar si la media de las diferencias es significativamente diferente de cero, comprobaría la presencia de SC en las predicciones del modelo.

**Determinación de sesgo proporcional.** El SP es reconocido cuando en el gráfico de los puntos ( $z_i, d_i=y_i-z_i$ ) se forme una tendencia lineal positiva o negativa, es decir, la magnitud del sesgo ( $d_i=y_i-z_i$ ) crece o decrece directamente con los valores predichos ( $z_i$ ). Un análisis de regresión lineal simple del sesgo sobre los predichos ( $D_i=\beta_0+\beta_1z_i+\varepsilon_i$ ) contribuye a determinar el SP de manera más objetiva (2,8).

**Validación de modelos con sesgo proporcional.** Dado que se tiene SP el conjunto de los puntos ( $z_i, D_i$ )  $i=1,2,\dots,n$  se encuentran relacionados por  $E(D_i)=\beta_0+\beta_1z_i$  y  $Var(D_i)=\sigma_D^2=\sigma^2$  con  $D_i\sim NI(\beta_0+\beta_1z_i, \sigma^2)$ , es decir,  $D_i=\beta_0+\beta_1z_i+\varepsilon_i$  donde  $E(\varepsilon_i)=0$ ,  $Var(\varepsilon_i)=\sigma^2$  y  $\varepsilon_i=D_i-(\beta_0+\beta_1z_i)=[D_i-E(D_i)]\sim NI(0, \sigma^2)$ , por lo que el modelo será exacto

would need only a statistical test for the required precision  $P(|D| \leq e) \geq 1 - \alpha$ , which is translated to  $\sigma^2 \leq e^2 / \chi_{1,1-\alpha}^2$ , once PB is removed by the referred correction.

If  $\varepsilon \sim N(0, \sigma^2)$ , then  $P\left(\frac{\varepsilon^2}{\sigma^2} \leq \chi_{1,1-\alpha}^2\right) = P(\chi_1^2 \leq \chi_{1,1-\alpha}^2) = 1 - \alpha$ .

Moreover, for e and  $\alpha$  satisfying  $P(|\varepsilon| \leq e) \geq 1 - \alpha$  it follows that  $P\left(\frac{\varepsilon^2}{\sigma^2} \leq \frac{e^2}{\sigma^2}\right) \geq 1 - \alpha$ ; thus,  $\frac{e^2}{\sigma^2} \geq \chi_{1,1-\alpha}^2$  and  $\sigma^2 \leq \frac{e^2}{\chi_{1,1-\alpha}^2}$  where, in general,  $\chi_{k,\gamma}^2$  represents the quantile  $\gamma$  of the chi-square distribution with "k" degrees of freedom ( $\chi_k^2$ ); that is,  $P(\chi_k^2 \leq \chi_{k,\gamma}^2) = \gamma$ .

The indicated correction results in an accurate model, and only the statistical test for precision is needed. Thus, the following step is to test the hypotheses with the original proposal (OP)  $H_0 : \sigma_D^2 \leq e^2 / \chi_{1,1-\alpha}^2$  vs  $H_1 : \sigma_D^2 > e^2 / \chi_{1,1-\alpha}^2$  (3) or the alternative proposal (AP)  $H_0^{AP} : \sigma_D^2 > e^2 / \chi_{1,1-\alpha}^2$  vs  $H_1^{AP} : \sigma_D^2 \leq e^2 / \chi_{1,1-\alpha}^2$  (8). The test statistic with  $\varepsilon \sim NI(0, \sigma^2 = \sigma_D^2)$  and under true null hypotheses  $H_0$  or  $H_0^{AP}$  is:  $V = \left(\frac{\chi_{1,1-\alpha}^2}{e^2}\right)(n-2)(CME)_D \sim \chi_{n-2}^2$ .  $H_0$  or  $H_0^{AP}$  is rejected with a significance level  $\alpha'$  if  $V_c > \chi_{n-2,1-\alpha'}^2$  or  $V_c \leq \chi_{n-2,\alpha'}^2$ , where  $V_c$  corresponds to the calculated value of the test statistic (V), and  $(CME)_D$  to the mean square of the error for the estimated model  $d_i = a + bz_i$ . Therefore, if  $H_0$  is not rejected or  $H_0^{AP}$  is rejected, then the model is considered acceptable for prediction under OP and AP, respectively.

Another approach for evaluating precision once PB is corrected, is to use confidence intervals (CI), similar to the procedure presented by Medina et al (4) when the model has CB in its predictions. This approach is motivated by Freese's proposal: different users of the model may have different requirements of precision, leading to different values of the maximum admitted error of the deviations (e).

From the rejection region for  $H_0$ ,  $V_c = \left(\frac{\chi_{1,1-\alpha}^2}{e^2}\right)(n-2)(CME)_D > \chi_{n-2,1-\alpha'}^2$  and solving for "e", the critical error is found (8):

$$e < \left(\frac{\chi_{1,1-\alpha}^2(n-2)(CME)_D}{\chi_{n-2,1-\alpha'}^2}\right)^{1/2} = e_{SP}^*$$

Therefore,  $H_0$  will be rejected if  $e < e_{PB}^*$ , and it will not be rejected if  $e \geq e_{PB}^*$ ; that is, if the model user specifies a value for "e" such that  $e \geq e_{PB}^*$ , then the model is considered acceptable in the prediction of the system under OP.

al efectuar la corrección  $\varepsilon_i = D_i - (\beta_0 + \beta_1 z_i)$ . En la práctica se ajusta el modelo de regresión y se estiman los errores ( $\varepsilon_i$ ):  $d_i - \hat{d}_i = d_i - (a + bz_i)$ , donde a y b son las estimaciones por mínimos cuadrados del intercepto ( $\beta_0$ ) y la pendiente ( $\beta_1$ ) respectivamente. Así, sólo faltaría una prueba estadística para la precisión requerida  $P(|D| \leq e) \geq 1 - \alpha$ , que se traduce en  $\sigma^2 \leq e^2 / \chi_{1,1-\alpha}^2$  una vez que es removido el SP mediante dicha corrección.

Si  $\varepsilon \sim N(0, \sigma^2)$  entonces  $P\left(\frac{\varepsilon^2}{\sigma^2} \leq \chi_{1,1-\alpha}^2\right) = P(\chi_1^2 \leq \chi_{1,1-\alpha}^2) = 1 - \alpha$ .

Además para e y  $\alpha$  tales que satisfacen  $P(|\varepsilon| \leq e) \geq 1 - \alpha$  se sigue que  $P\left(\frac{\varepsilon^2}{\sigma^2} \leq \frac{e^2}{\sigma^2}\right) \geq 1 - \alpha$ , así,  $\frac{e^2}{\sigma^2} \geq \chi_{1,1-\alpha}^2$  y  $\sigma^2 \leq \frac{e^2}{\chi_{1,1-\alpha}^2}$  donde, en general,  $\chi_{k,\gamma}^2$  representa el cuantil  $\gamma$  de la distribución ji-cuadrada con "k" grados de libertad ( $\chi_k^2$ ), es decir,  $P(\chi_k^2 \leq \chi_{k,\gamma}^2) = \gamma$ .

La corrección indicada lleva a que el modelo sea exacto y sólo faltaría una prueba estadística para la precisión. Así, el siguiente paso es probar las hipótesis con el planteamiento original (PO)  $H_0 : \sigma_D^2 \leq e^2 / \chi_{1,1-\alpha}^2$  vs  $H_1 : \sigma_D^2 > e^2 / \chi_{1,1-\alpha}^2$  (3) o el planteamiento alternativo (PA)  $H_0^{PA} : \sigma_D^2 > e^2 / \chi_{1,1-\alpha}^2$  vs  $H_1^{PA} : \sigma_D^2 \leq e^2 / \chi_{1,1-\alpha}^2$  (8). La estadística de prueba con  $\varepsilon \sim NI(0, \sigma^2 = \sigma_D^2)$  y bajo  $H_0$  o  $H_0^{PA}$  verdadera es:

$$V = \left(\frac{\chi_{1,1-\alpha}^2}{e^2}\right)(n-2)(CME)_D \sim \chi_{n-2}^2$$

La  $H_0$  o  $H_0^{PA}$  se rechazará con un nivel de significación  $\alpha'$  si  $V_c > \chi_{n-2,1-\alpha'}^2$  o  $V_c \leq \chi_{n-2,\alpha'}^2$ , donde  $V_c$  corresponde al valor calculado de la estadística de prueba (V) y  $(CME)_D$  al cuadrado medio del error del modelo estimado  $d_i = a + bz_i$ . Por lo tanto, si  $H_0$  no se rechaza o  $H_0^{PA}$  se rechaza entonces el modelo es considerado aceptable para predicción bajo el PO y el PA respectivamente.

Otro enfoque para evaluar la precisión una vez corregido el SP es utilizar intervalos de confianza (IC), similarmente al procedimiento presentado en Medina et al (4) cuando el modelo presenta SC en sus predicciones. Este enfoque es motivado a que en el planteamiento de Freese, diferentes usuarios del modelo, pueden tener distintos requerimientos de precisión, la cual conduce a diferentes valores del máximo error admitido de las desviaciones (e).

De la región de rechazo para  $H_0$ ,  $V_c = \left(\frac{\chi_{1,1-\alpha}^2}{e^2}\right)(n-2)(CME)_D > \chi_{n-2,1-\alpha'}^2$ , y despejando "e", se obtuvo el error crítico (8):  $e < \left(\frac{\chi_{1,1-\alpha}^2(n-2)(CME)_D}{\chi_{n-2,1-\alpha'}^2}\right)^{1/2} = e_{SP}^*$

Por lo tanto  $H_0$  se rechazará si  $e < e_{SP}^*$  y no se rechazará si  $e \geq e_{SP}^*$ , es decir, si el usuario del modelo especifica un valor de "e" tal que  $e \geq e_{SP}^*$

Using an analogous procedure for AP, the maximum anticipated error, or critical error was obtained (8):

$$e_{PB}^{**} = \left( \frac{\chi_{1,1-\alpha}^2 (n-2) (CME)_D}{\chi_{n-2, \alpha'}^2} \right)^{\frac{1}{2}}$$

Thus, if the model user specifies a value of "e" such that  $e \geq e_{PB}^{**}$ , then the model is considered acceptable in predicting the system under AP. In a previous study (9), it can be checked that hypothesis tests and confidence intervals are related; this relationship enables to construct one from the other.

Other published works from Reynolds (6) and Barrales et al (7) do not present the confidence interval (CI) for  $\varepsilon_{PB} = (\sigma^2 \chi_{1,1-\alpha}^2)^{1/2}$ , the quantile 1- $\alpha$  of the distribution of  $|\varepsilon|$  (in practice, the errors (e) once PB is corrected)

or equivalently  $\varepsilon_{PB}^2$ , is the quantile 1- $\alpha$  of the distribution of  $\varepsilon^2$  since, if  $\varepsilon \sim N(0, \sigma^2)$ , then  $1-\alpha = P\left(\frac{\varepsilon^2}{\sigma^2} \leq \chi_{1,1-\alpha}^2\right) = P(\varepsilon^2 \leq \sigma^2 \chi_{1,1-\alpha}^2) = P(\varepsilon^2 \leq \varepsilon_{PB}^2)$ .

A (1- $\alpha'$ )100% CI for  $\varepsilon_{PB}$  (8) based on CI (1- $\alpha'$ )100% for  $\sigma^2$  (where  $\tau(\sigma^2) = (\sigma^2 \chi_{1,1-\alpha}^2)^{1/2} = \varepsilon_{PB}$  is an increasing monotonic function) is given by

$$\tau\left(\frac{(n-2)(CME)_D}{\chi_{n-2, 1-\frac{\alpha'}{2}}^2}\right) < \tau(\sigma^2) < \tau\left(\frac{(n-2)(CME)_D}{\chi_{n-2, \frac{\alpha'}{2}}^2}\right)$$

$$\left(\frac{\chi_{1,1-\alpha}^2 (n-2) (CME)_D}{\chi_{n-2, 1-\frac{\alpha'}{2}}^2}\right)^{1/2} < \varepsilon_{PB} < \left(\frac{\chi_{1,1-\alpha}^2 (n-2) (CME)_D}{\chi_{n-2, \frac{\alpha'}{2}}^2}\right)^{1/2}$$

$e_{IPB}^* < \varepsilon_{PB} < e_{IPB}^{**}$  where  $e_{IPB}^*$  and  $e_{IPB}^{**}$  correspond to the critical errors  $e_{PB}^*$  and  $e_{PB}^{**}$ , with the difference that  $\alpha'$  is substituted by  $\alpha'/2$ . The estimated CI ( $e_{IPB}^*, e_{IPB}^{**}$ ) means that we have a confidence of 100(1- $\alpha'$ )% that the value of the distribution for  $|\varepsilon| = |D_i - (\beta_0 + \beta_1 z_i)|$  having below the 100(1- $\alpha$ ) percent of the absolute errors is located in some location of the referred interval. This enables us to determine with certain probability an upper bound for the magnitude of the prediction error (e), and to use it in the evaluation of the evolution of the model for system prediction.

**RESULTS**

To illustrate the application of the methodology, the CI approach (4) was used with a published data set from Barrales et al (7), which corresponds

entonces el modelo es considerado aceptable en pronóstico del sistema bajo el PO.

Procediendo análogamente para el PA se obtuvo el error máximo anticipado o error crítico (8):

$$e_{SP}^{**} = (\chi_{1,1-\alpha}^2 (n-2) (CME)_D / \chi_{n-2, \alpha'}^2)^{1/2}$$

Así, si el usuario del modelo especifica un valor de "e" tal que  $e \geq e_{SP}^{**}$  entonces el modelo es considerado aceptable en pronóstico del sistema bajo el PA. En un trabajo previo (9) puede consultarse la relación existente entre pruebas de hipótesis e intervalos de confianza, la cual permite construir uno a partir del otro.

Otros trabajos publicados por Reynolds (6) y Barrales et al (7), no presentaron un intervalo de confianza (IC) para  $\varepsilon_{SP} = (\sigma^2 \chi_{1,1-\alpha}^2)^{1/2}$ , el cuantil 1- $\alpha$  de la distribución de  $|\varepsilon|$  (en la práctica, los errores (e) una vez que se corrige el SP) o equivalentemente  $\varepsilon_{SP}^2$  es el cuantil 1- $\alpha$  de la distribución de  $\varepsilon^2$ , ya que si  $\varepsilon \sim N(0, \sigma^2)$  entonces .

$$1-\alpha = P\left(\frac{\varepsilon^2}{\sigma^2} \leq \chi_{1,1-\alpha}^2\right) = P(\varepsilon^2 \leq \sigma^2 \chi_{1,1-\alpha}^2) = P(\varepsilon^2 \leq \varepsilon_{SP}^2)$$

Un IC del (1- $\alpha'$ )100% para  $\varepsilon_{SP}$  (8) a partir del IC del (1- $\alpha'$ )100% para  $\sigma^2$  (con  $\tau(\sigma^2) = (\sigma^2 \chi_{1,1-\alpha}^2)^{1/2} = \varepsilon_{SP}$  función monótona creciente) está dado por

$$\tau\left(\frac{(n-2)(CME)_D}{\chi_{n-2, 1-\frac{\alpha'}{2}}^2}\right) < \tau(\sigma^2) < \tau\left(\frac{(n-2)(CME)_D}{\chi_{n-2, \frac{\alpha'}{2}}^2}\right)$$

$$\left(\frac{\chi_{1,1-\alpha}^2 (n-2) (CME)_D}{\chi_{n-2, 1-\frac{\alpha'}{2}}^2}\right)^{1/2} < \varepsilon_{SP} < \left(\frac{\chi_{1,1-\alpha}^2 (n-2) (CME)_D}{\chi_{n-2, \frac{\alpha'}{2}}^2}\right)^{1/2}$$

$e_{ISP}^* < \varepsilon_{SP} < e_{ISP}^{**}$  donde  $e_{ISP}^*$  y  $e_{ISP}^{**}$  corresponden a los errores críticos  $e_{SP}^*$  y  $e_{SP}^{**}$  con la diferencia de que  $\alpha'$  se sustituye por  $\alpha'/2$ . El IC estimado ( $e_{ISP}^*, e_{ISP}^{**}$ ) significa que se tiene confianza en un 100(1- $\alpha'$ )% que el punto de la distribución  $|\varepsilon| = |D_i - (\beta_0 + \beta_1 z_i)|$  que tiene debajo el 100(1- $\alpha$ )% de los errores absolutos está localizado en alguna parte de dicho intervalo. Permite determinar con cierta probabilidad una cota superior para la magnitud del error de predicción (e) y usarla para evaluar la evolución del modelo en predicción del sistema.

**RESULTADOS**

Para ilustrar la aplicación de la metodología se utilizó el enfoque de IC (4), a un conjunto de datos previamente publicados por Barrales et al (7), los cuales corresponden a un modelo

to a model of grassland growth simulation. In this example, we applied the method of calculating the maximum anticipated errors or critical errors of the OP and AP, when the model has PB. In addition, 95% CI was calculated for  $\varepsilon_{PB}$ , the 1- $\alpha$  quantile of the error (e) distribution, once PB is corrected with  $d_i - (a + bz_i)$ .

The distribution of the points in Figure 1 indicates that the model has PB in its predictions; the magnitude of the bias ( $d_i = y_i - z_i$ ) decreases directly with the predicted values ( $z_i$ ) and the linear relationship was significant ( $d = 912.838 - 0.396z$ :  $F = 13.665$ ,  $p = 0.001$ ). With  $\alpha = \alpha' = 0.05$ , the critical errors were  $e_{PB}^* = 697.311 \text{ kg ha}^{-1}$  and  $e_{PB}^{**} = 1156.273 \text{ kg ha}^{-1}$ . Therefore, if the modeler or the model user specifies a value "e" such that  $e \geq e_{PB}^*$  or  $e \geq e_{PB}^{**}$ , then the model is considered sufficiently reliable in predicting the system based on OP and AP, respectively. For example, if  $e = 1157 \text{ kg ha}^{-1}$ , the model is considered acceptable for predicting with both proposals (OP:  $e > 697.311 \text{ kg ha}^{-1}$ ; AP:  $e > 1156.273 \text{ kg ha}^{-1}$ ).

The 1- $\alpha' = 95\%$  CI for  $\varepsilon_{PB}$  is  $669.689 \text{ kg ha}^{-1} < \varepsilon_{PB} < 1225.564 \text{ kg ha}^{-1}$ , and thus, we are 95% confident that the distribution value  $|\varepsilon| = |D_i - (\beta_0 + \beta_1 z_i)|$  found below 1- $\alpha = 95\%$  of the absolute errors is located in some part of the interval ( $669.689 \text{ kg ha}^{-1} - 1225.564 \text{ kg ha}^{-1}$ ); that is, there is a 95% of probability that the magnitude of the prediction error does not surpass  $1225.564 \text{ kg ha}^{-1}$ .

With 95% of confidence, the critical error with AP and CI for  $\varepsilon_{PB}$  would have to allow  $1156.273 \text{ kg ha}^{-1}$  as the minimum prediction error, which should not surpass  $1225.564 \text{ kg ha}^{-1}$ . In this way, the validated model could be used to predict grassland growth. However, it would require a fit in its structure based on the presence PB in its predictions.

## DISCUSSION

Identifying the type of bias may allow to improve the structure of the model and its evaluation, the data and methods used in all the processes of its construction, and validation (4). McCarthy et al have pointed out that testing a model helps to identify its weaknesses so that its predictive performance can be improved (10), since identification and acceptance of inaccuracies of a model is a step towards the evolution of a more accurate, more reliable model (1).

The extensions we present for validating deterministic model possessing PB in their

de simulación de crecimiento de praderas. En este ejemplo se aplicó el método de calcular los errores máximos anticipados o errores críticos del PO y PA cuando el modelo presenta SP. Adicionalmente se calculó el IC del 95% para  $\varepsilon_{SP}$ , el cuantil 1- $\alpha$  de la distribución de los errores (e) una vez que se corrige el SP mediante  $d_i - (a + bz_i)$ .

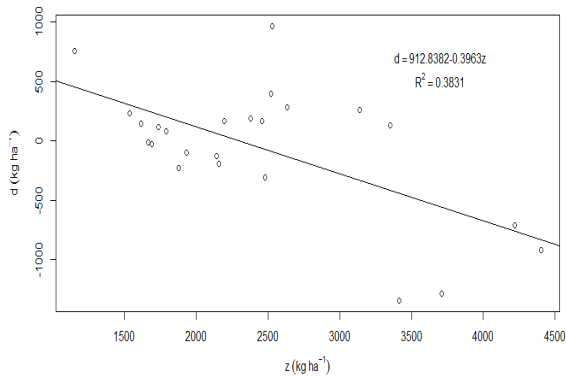
La distribución de los puntos en la figura 1 indica que el modelo presenta SP en sus pronósticos, la magnitud del sesgo ( $d_i = y_i - z_i$ ) decrece directamente con los valores predichos ( $z_i$ ) y la relación lineal resultó significativa ( $d = 912.838 - 0.396z$ :  $F = 13.665$ ,  $p = 0.001$ ). Con  $\alpha = \alpha' = 0.05$  los errores críticos resultaron  $e_{SP}^* = 697.3114 \text{ kg ha}^{-1}$  y  $e_{SP}^{**} = 1156.2730 \text{ kg ha}^{-1}$ . Por lo tanto, si el modelador o usuario del modelo especifica un valor "e" tal que  $e \geq e_{SP}^*$  o  $e \geq e_{SP}^{**}$ , entonces el modelo será considerado suficientemente confiable en predicción del sistema con base en el PO o PA respectivamente. Por ejemplo, si  $e = 1157 \text{ kg ha}^{-1}$  el modelo es considerado aceptable para predicción con ambos planteamientos (PO:  $e > 697.311 \text{ kg ha}^{-1}$ ; PA:  $e > 1156.273 \text{ kg ha}^{-1}$ ).

El IC del 1- $\alpha' = 95\%$  para  $\varepsilon_{SP}$  es  $669.689 \text{ kg ha}^{-1} < \varepsilon_{SP} < 1225.564 \text{ kg ha}^{-1}$ , por lo que se tiene confianza en un 95% de que el punto de la distribución  $|\varepsilon| = |D_i - (\beta_0 + \beta_1 z_i)|$  que tiene debajo el 1- $\alpha = 95\%$  de los errores absolutos está localizado en alguna parte en el intervalo ( $669.689 \text{ kg ha}^{-1} - 1225.564 \text{ kg ha}^{-1}$ ), es decir, con probabilidad del 95% la magnitud del error de predicción no sobrepasa  $1225.564 \text{ kg ha}^{-1}$ .

Con 95% de confianza, del error crítico con el PA y del IC para  $\varepsilon_{SP}$ , tendría que permitirse como mínimo  $1156.273 \text{ kg ha}^{-1}$  de error de predicción y éste no sobrepasaría  $1225.564 \text{ kg ha}^{-1}$ . Así, el modelo validado podría usarse para predecir el crecimiento de praderas, sin embargo, requerirá un ajuste en su estructura con base a la presencia de SP en sus pronósticos.

## DISCUSIÓN

Identificar el tipo de sesgo permitirá mejorar la estructura y evaluación del modelo, los datos y métodos empleados en todos los procesos de construcción y validación (4). McCarthy et al (10) señalan que probar un modelo ayuda a identificar sus debilidades con el fin de mejorar su desempeño predictivo, ya que la identificación y aceptación de inexactitudes de un modelo es un paso hacia la evolución de un



**Figure 1.** Relationship between bias ( $d_i$ ) and simulated values ( $z_i$ ) for the grassland growth data.

predictions are applied with no modifications in the model structure because this type of bias is removed from the available data ( $z_i, y_i$ ) by means of the bias values ( $d_i$ ), namely, by the correction  $d_i - (a + bz_i)$ .

The test statistic and critical error with AP require a larger maximum admissible error ( $e$ ) value than that with OP, in order to infer whether the model is acceptable for predicting the system or not. A previous study by Medina et al (4) points out that the inconvenience of applying the OP is the ambiguity occurring when the null hypothesis is not rejected ( $H_0$ ) since what it can be inferred is that the data do not provide sufficient evidence to reject it, and that the affirmation stated in  $H_0$  is not accepted. Moreover, as the research hypothesis is posed in the alternative hypothesis, we recommend using the AP in validating a system prediction model.

Validation by means of critical errors or by the confidence limits approach (the latter being a procedure equivalent to the hypothesis testing approach, as mentioned above), is reduced to calculating the maximum anticipated error, or critical error, in which the model developer or the user decide whether the model is acceptable in predicting the system. This is carried out by comparing the critical error with the required accuracy ( $e$ ) under the values  $\alpha$  and  $\alpha'$  specified beforehand. This requires that the model developer or the user have good understanding of the system to establish, *a priori*, the maximum admissible error ( $e$ ). Medina et al (4) recommend the CI approach to validate a model for several reasons. It provides the range of possible values of the parameter under consideration thus, it is more informative than hypothesis testing in the sense that it permits determining the *a posteriori* maximum admissible prediction error.

modelo más exacto y de más confianza (1).

Las extensiones presentadas para validar modelos determinísticos en presencia de SP en sus pronósticos se aplican sin que el modelo sea modificado en su estructura, ya que se remueve el tipo de sesgo a los datos disponibles ( $z_i, y_i$ ) a través de los valores del sesgo ( $d_i$ ), es decir,  $d_i - (a + bz_i)$  es la corrección por SP.

La prueba estadística y error crítico con el PA requieren un valor más grande del máximo error admisible ( $e$ ) que con el PO para inferir que el modelo es aceptable en predicción del sistema. En un trabajo previo de Medina et al (4), se señala que el inconveniente de aplicar el PO es la ambigüedad que se presenta al no rechazar a la hipótesis nula ( $H_0$ ), ya que lo que se puede inferir es que los datos no proporcionan suficiente evidencia para rechazarla y no que se acepte la declaración establecida en  $H_0$ ; además de que la hipótesis de investigación se plantea en la hipótesis alternativa. Por lo que se recomienda utilizar el PA en la validación de un modelo en predicción del sistema.

La validación por medio de los errores críticos o de enfoque de límites de confianza, la cual es equivalente al enfoque de pruebas de hipótesis, se reduce a calcular el error máximo anticipado o error crítico, en donde el modelador o usuario decide si el modelo es aceptable en predicción del sistema al comparar el error crítico con la exactitud requerida ( $e$ ) bajo los valores  $\alpha$  y  $\alpha'$  especificados con anticipación. Lo anterior implica una buena comprensión del sistema por parte del modelador o usuario del modelo para establecer *a priori* el máximo error admisible ( $e$ ). Medina et al (4) recomiendan el enfoque de IC para validar un modelo por las siguientes razones: proporciona la amplitud de los posibles valores del parámetro en consideración y es más informativo que las pruebas de hipótesis, en el sentido de que permite *a posteriori* conocer el máximo error admisible de predicción.

Cabe señalar que el error crítico o máximo error admisible *a posteriori*, puede utilizarse para comparar varios modelos para un mismo sistema, de modo que el mejor modelo sería el que tenga el menor error crítico. Ahora, si lo que se quiere es evaluar la mejora de un modelo en predicción del sistema, sería observar si en cada proceso de mejora, el máximo error admisible *a posteriori* decrece.

It should be pointed out that the critical error, or a *posteriori* maximum admissible error, can be used to compare several models for a single system, in such a way that the best model would be the one with the lowest critical error. To assess the improvement of a model in system prediction it is necessary to observe whether a *posteriori* maximum admissible error decreases at each process of improvement.

The estimated CI for the quantile  $1-\alpha$  of the error distribution, once PB is corrected, allows us to determine an upper bound for the magnitude of the prediction error with a certain probability, and to use it in the evaluation of the evolving model whenever an improvement is required. This is similar to the case of validating a model with CB in its predictions (4). Thus, based on the critical error with AP and the aforementioned CI, we can establish the minimum permitted prediction error, and what amount of error of prediction would not surpass.

In conclusion, the extensions to Freese's statistical method presented here to validate models in the presence of proportional bias are applied without modification in the model structure.

In validating a model, we recommend using the confidence interval approach under the alternative proposal.

The confidence interval for the  $1-\alpha$  quantile of error distribution, once the proportional bias is corrected, allows determination of an upper bound for the magnitude of the prediction error.

Both a *posteriori* maximum admissible error, and the upper bound for the magnitude of the prediction error can be used to evaluate the evolution of a model in predicting the system in the face of a modification of the modeling process.

### Acknowledgements

We thank the Subsecretaría de Educación Superior e Investigación Científica, Programa de Mejoramiento del Profesorado (PROMEP) Secretaría de Educación Pública of México for funding this study.

El IC estimado para el cuantil  $1-\alpha$  de la distribución de los errores una vez que se corrige el SP permite determinar, con cierta probabilidad, una cota superior para la magnitud del error de predicción y usarla para evaluar la evolución del modelo ante una mejora; similar al caso de la validación de un modelo con SC en sus predicciones (4). Así, con base en el error crítico con el PA y el IC antes mencionado, puede establecerse cuál sería el mínimo error de predicción permitido y cuál cantidad de error no sobrepasaría.

En conclusión, las extensiones presentadas al método estadístico de Freese para validar modelos en presencia de sesgo proporcional, se aplican sin que el modelo sea modificado en su estructura.

En la validación de un modelo se recomienda utilizar el enfoque de intervalo de confianza bajo el planteamiento alternativo.

El intervalo de confianza para el cuantil  $1-\alpha$  de la distribución de los errores una vez que se corrige el sesgo proporcional, permite determinar una cota superior para la magnitud del error de predicción.

Tanto el máximo error admisible *a posteriori* como la cota superior para la magnitud del error de predicción, pueden utilizarse para evaluar la evolución de un modelo en predicción del sistema ante una modificación en el proceso de modelado.

### Agradecimientos

A la Subsecretaría de Educación Superior e Investigación Científica, Programa de Mejoramiento del Profesorado (PROMEP), Secretaría de Educación Pública de México por el financiamiento a esta investigación.

## REFERENCES

1. Tedeschi LO. Assessment of the adequacy of mathematical models. *Agr Syst* 2006; 89(2-39):225-247.
2. Medina-Peralta S, Vargas-Villamil L, Navarro-Alberto J, Canul-Pech C, Peraza-Romero S. Comparación de medidas de desviación para validar modelos sin sesgo, sesgo constante o proporcional. *Univ Cienc.* 2010; 26(3):255-263.
3. Freese F. Testing accuracy. *For Sci* 1960; 6:139-45.
4. Medina PS, Vargas-Villamil L, Navarro AJ, Avendaño L, Colorado L, Arjona-Suarez E, Mendoza-Martínez G. Validación de modelos con sesgo constante: un enfoque aplicado. *Rev MVZ Córdoba* 2014; 19(2):4099-08.



5. Rennie JC, Wiant HV. Modification of Freese's chi-square test of accuracy. USDI Bureau of Land Management, Denver Colorado. Resource Inventory 1978; 14:1-3.
6. Reynolds MR. Estimating the error in model predictions. For Sci 1984; 30(2):454-69.
7. Barrales VL, Peña RI, Fernández RB. Model validation: an applied approach. Agric Tech 2004; 64:66-73.
8. Medina PS. Validación de modelos mecanísticos basada en la prueba ji-cuadrada de Freese, su modificación y extensión. [Master's thesis]. Montecillo, México: Colegio de Postgraduados; 2006.
9. Bickel PJ, Doksum KA. Mathematical Statistic: basic ideas and selected topics Vol I. Second edition. New Jersey: Pearson Prentice Hall; 2007.
10. McCarthy MA, Possingham HP, Day JR, Tyre AJ. Testing the accuracy of population viability analysis. Conserv Biol 2001; 15:1030-38.