COMPARISON OF LINEAR AND NONLINEAR MODELS: THE CASE OF THE ROSE QUALITY STEMS

COMPARACIÓN DE MODELOS LINEALES Y NO LINEALES: EL CASO DE LA CALIDAD DE TALLOS DE ROSA

Martha Elva **Ramírez-Guzmán**[1], Ma. de Lourdes **Arévalo-Galarza**[2], Gumercindo **de la Cruz-Guzman**[3]

[1]Estadística. [2]Fruticultura. Campus Montecillo. Colegio de Postgraduados. 56230. Montecillo, Estado de México. (martharg@colpos.mx), (larevalo@colpos.mx). [3]Unidad de Morfología y Función, Facultad de Estudios Superiores Iztacala, Universidad Nacional Autónoma de México, Avenida de los Barrios Núm. 1, Los Reyes Iztacala. 54090, Tlalnepantla, Estado de México; México. (moashi@unam.mx).

ABSTRACT

In agronomic design of experiments, researchers frequently assume linear associations between explanatory and response variables, also they assume a gaussian distribution or transform it. Inherent variability of the experiment units is not considered. Besides, it is frequent to analyze the time as a factor of a factorial design when repeated measures are taken through time, missing the opportunity to analyze the plant growth. New statistical models present the opportunity to analyze the raw data without transforming, they also can consider the variability of the experiment units as random components and they can identify the relationship between explanatory and response variables as linear or nonlinear. In cut roses production, one of the main quality factors is stem length, but is not clear how temperature, relative humidity and light affect it. The aim of this research was to analyze new statistical models to better understanding the impact of environmental variables on stem length of two rose cultivars (Samurai and Blush) during two peak production periods. The models used were: generalized linear models (GLM), generalized linear mixed models (GLMM), generalized additive models (GAM), generalized additive mixed models (GAMM), vector generalized lineal models (VGLM) and repeated measures model. The results showed that GAM model with a gamma distribution and identity link function was the best model with minimum BIC value and minimum error variance. It identified a nonlinear effect of relative humidity and linear effects of heat and light. The best conditions to produce plants between 50 and 70 cm, were 650 to 830 heat units and 82.5 to 85% of relative humidity. GAMM identified the period of April to May with

* Author for correspondence ❖ Autor para correspondencia.
Received: September, 2019. Approved: August, 2020.
Published as ARTICLE in Agrociencia 54: 939-953. 2020.

RESUMEN

En el diseño agronómico de experimentos, los investigadores con frecuencia asumen asociaciones lineales entre variables explicativas y de respuesta, también asumen una distribución gaussiana de ésta, o la transforman. La variabilidad inherente de las unidades experimentales no es considerada. Además, es frecuente analizar el tiempo como factor de un diseño factorial cuando se toman medidas repetidas a lo largo del tiempo, perdiendo la oportunidad de analizar el crecimiento de la planta. Los nuevos modelos estadísticos presentan la oportunidad de analizar los datos sin transformar; también pueden considerar la variabilidad de las unidades experimentales como componentes aleatorios y pueden identificar la relación entre las variables explicativas y de respuesta como lineales o no lineales. En la producción de rosas cortadas, uno de los principales factores de calidad es la longitud del tallo, pero no está claro cómo la temperatura, la humedad relativa y la luz la afectan. El objetivo de esta investigación fue analizar nuevos modelos estadísticos para comprender mejor el impacto de las variables ambientales en la longitud del tallo de dos cultivares de rosas (Samurai y Blush), durante dos períodos de mayor producción. Los modelos utilizados fueron: modelos lineales generalizados (GLM), modelos lineales mixtos generalizados (GLMM), modelos aditivos generalizados (GAM), modelos mixtos aditivos generalizados (GAMM), modelos lineales generalizados vectoriales (VGLM) y modelo de medidas repetidas. Los resultados mostraron que el modelo GAM con una distribución gamma y una función liga de identidad fue el mejor modelo con un valor BIC mínimo y una mínima varianza de error. Se identificó un efecto no lineal de humedad relativa y efectos lineales de calor y luz. Las mejores condiciones para producir plantas entre 50 y 70 cm de longitud consistieron en la aplicación de 650 a 830 unidades de calor y 82,5 a 85% de humedad relativa. GAMM identificó el período de abril a mayo con cultivar Blush

cultivar Blush as the best conditions to produced roses. Repeated measures analysis confirmed that plants initiated high (low), remained so through time.

Key words: Generalized Linear Mixed Model, Generalized Additive Mixed Model, Vector Generalized Lineal Models, quality, flower production.

INTRODUCTION

In the design of experiments in agronomic analysis, researchers frequently assume gaussian response distribution and homocedasticity of treatments (De Felipe *et al.* 2016). New statistical models present an opportunity to incorporate asymmetrical distributions, heterocedasticity and nonlinear associations. For example, Vector Generalized Linear Models incorporated new non exponential distributions like Weibull distribution (Cattaneo *et al*., 2018; Yee, 2019). Generalized Linear Mixed Models incorporated random effects (Breslow and Clayton, 1993). Hastie and Tibshirani (1986) and Hastie (2018) included nonlinear models through Generalized Additive Models. Wood and Scheip (2017) developed a Generalized Additive Mixed Models which incorporated random effects to nonlinear relationships. These models present the opportunity to analyze the raw data without transforming, and another pitfall is the lack of incorporation in the model of repeated observations. In agronomic experiments when the plant growth is evaluated, it is frequent to analyze the time as a factor in a factorial design, instead of considering it as a chronological measurement, missing information of the plant growth.

In rose cut flower industry one of the requirements is the stem length. Mexico has a floricultural potential due to natural resources (water, soil, climate), as well as to social conditions and its proximity to the United States, one of the main flower consumers in the world (Mordor Int. 2020). The production peaks of roses are in February, May, November and December; in these months there are high variability of temperatures, RH and radiation inside the greenhouse, that affect the growth rate of the rose plants. Nevertheless, the growers are unaware of the influence of these factors that affect the rose stem quality through the year. The objective of this research was to analyze these new statistical models

como las mejores condiciones para producir rosas. El análisis de medidas repetidas confirmó que las plantas tuvieron un inicio alto (bajo), y permanecieron así a lo largo del tiempo.

Palabras clave: modelo lineal mixto generalizado, modelo mixto aditivo generalizado, modelos lineales generalizados vectoriales, calidad, producción de flores.

INTRODUCCIÓN

En el diseño de experimentos en análisis agronómico, con frecuencia los investigadores asumen una distribución gaussiana de la respuesta y homocedasticidad de los tratamientos (De Felipe *et al*., 2016). Los nuevos modelos estadísticos presentan una oportunidad para incorporar distribuciones asimétricas, heterocedasticidad y asociaciones no lineales. Por ejemplo, los modelos lineales generalizados vectoriales incorporaron nuevas distribuciones no exponenciales como la distribución de Weibull (Cattaneo *et al*., 2018; Yee, 2019). Los modelos lineales mixtos generalizados incorporaron efectos aleatorios (Breslow y Clayton, 1993). Hastie y Tibshirani (1986) y Hastie (2018) incluyeron modelos no lineales a través de Modelos Aditivos Generalizados. Wood y Scheip (2017) desarrollaron generalizados modelos mixtos aditivos que incorporaron efectos aleatorios a relaciones no lineales. Estos modelos presentan la oportunidad de analizar los datos sin transformar. Otro obstáculo es la falta de incorporación de observaciones repetidas en el modelo. En los experimentos agronómicos, cuando se evalúa el crecimiento de la planta, es frecuente analizar el tiempo como factor en un diseño factorial, en lugar de considerarlo como una medida cronológica, perdiendo así información sobre el crecimiento de la planta.

En la industria de las rosas cortadas, uno de los requisitos es la longitud del tallo. México tiene un potencial floricultural debido a los recursos naturales (agua, suelo, clima), así como a las condiciones sociales y su cercanía a Estados Unidos, uno de los principales consumidores de flores del mundo (Mordor Int., 2020). Los picos en la producción de rosas se dan en febrero, mayo, noviembre y diciembre; en estos meses hay una gran variabilidad de temperaturas, RH y radiación dentro del invernadero, que afectan la tasa de crecimiento de las plantas de rosas. Sin embargo, los cultivadores de rosas desconocen la influencia de estos factores que afectan la calidad del

to better understanding the impact of environmental variables on stem length of two rose cultivars.

MATERIALS Y METHODS

Generalized Linear Model (GLM)

GLM describe the dependence of an observable variable y_i (i=1,…, n) on a vector of regressors, x_j (j=1,…, p), where xi = (x_{i1}, x_{i2},…, x_{ip}), and x_{i0} = 1 serves as an intercept coefficient. The conditional distribution of $y_i|x_i$ is a linear exponential family with density function:

$$f(y_i;\theta_i,\phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i,\phi)\}$$

where θ_i is the natural parameter that depends on the explanatory variables via a linear predictor: $\eta_i = \sum_{j=1}^{p}\beta_j x_{ij}$; and ϕ is a dispersion parameter, usually a(ϕ) = 1. The functions $a(\cdot)$, $b(\cdot)$ y $c(\cdot)$, determine which member of the exponential family is used (e.g. normal, gamma, inverse Gaussian, binomial and Poisson). The mean and the conditional variance of y_i are $E[y_i|x_i] = \mu_i = b'(\theta_i)$ and Var $[y_i|x_i] = \phi \cdot b''(\theta_i)$, respectively. The mean variance is $V(\mu_i)=b''(\theta_i)$. Dependence of the conditional mean $E[y_i|x_i] = \mu_i$ on x_j, is specified via a regression model:

$$g(\mu_i) = x_i^T\beta \quad (1)$$

where g is a link function and β is a vector of regression coefficients that is estimated by maximum likelihood, using a weighted least squares iterative algorithm. The estimable functions of the regression model are estimated according to the member of the exponential family.

Generalized Linear Mixed Model (GLMM)

GLMM is obtained from the GLM with the incorporation of random effects in the linear predictor, and this idea was developed by Breslow and Clayton (1993). The structure of a GLMM is given by the expression: $g(\mu_i) = x_{ij}\beta + z_{ij}u_i$; i=1,…, n; j=1,…p, where β is the vector that contains the fixed effects of the explanatory variables x_{ij}; z_{ij} are the variables associated to the random effects and $\{u_i\}$ are the random effects that are assumed to have a normal probability distribution. In matrix form, this model can be written as:

$$g(\mu) = X\beta + Z\gamma$$

where X and Z are matrices with fixed and random effects, respectively, β is a vector of regression coefficients, γ is a random effects vector with normal distribution, $\gamma \sim N(o, \psi)$, g is a link

tallo de la rosa a lo largo del año. El objetivo de esta investigación fue analizar estos nuevos modelos estadísticos para comprender mejor el impacto de las variables ambientales en la longitud del tallo de dos cultivares de rosas.

MATERIALES Y MÉTODOS

Modelo lineal generalizado (GLM)

El GLM describe la dependencia de una variable observable y_i (i=1,…, n) sobre un vector de regresores, x_j (j=1,…, p), donde xi = (x_{i1}, x_{i2},…, x_{ip}), y x_{i0} = 1 sirve como coeficiente de intercepción. La distribución condicional de $y_i|x_i$ es una familia exponencial lineal con función de densidad:

$$f(y_i;\theta_i,\phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i,\phi)\}$$

donde θ_i es el parámetro natural que depende de las variables explicativas mediante un predictor lineal: $\eta_i = \sum_{j=1}^{p}\beta_j x_{ij}$; y ϕ es un parámetro de dispersión, usualmente a(ϕ) = 1. Las funciones $a(\cdot)$, $b(\cdot)$ y $c(\cdot)$ determinan qué miembro de la familia exponencial se utiliza (por ejemplo, normal, gamma, gaussiana inversa, binomial y Poisson). La media y la varianza condicional de y_i son $E[y_i|x_i] = \mu_i = b'(\theta_i)$ y Var $[y_i|x_i] = \phi \cdot b''(\theta_i)$ respectivamente. La varianza media es $V(\mu_i)=b''(\theta_i)$. La dependencia de la media condicional $E[y_i|x_i] = \mu_i$ en x_j se especifica mediante un modelo de regresión:

$$g(\mu_i) = x_i^T\beta \quad (1)$$

donde g es una función de enlace y β es un vector de coeficientes de regresión que se estima por máxima verosimilitud, utilizando un algoritmo iterativo de mínimos cuadrados ponderados. Las funciones estimables del modelo de regresión se estiman según el miembro de la familia exponencial.

Modelo mixto lineal generalizado (GLMM)

El GLMM se obtiene del GLM con la incorporación de efectos aleatorios en el predictor lineal, y esta idea fue desarrollada por Breslow y Clayton (1993). La estructura de un GLMM es dada por la expresión: $g(\mu_i) = x_{ij}\beta + z_{ij}u_i$; i=1,…, n; j=1,…p, donde β es el vector que contiene los efectos fijos de las variables explicativas x_{ij}; z_{ij} son las variables asociadas a los efectos aleatorios y $\{u_i\}$ son efectos aleatorios y se asume que tienen una distribución de probabilidad normal. En forma matricial, este modelo se puede escribir como:

$$g(\mu) = X\beta + Z\gamma$$

function defined as in (1) and ψ is a covariance matrix of random effects.

Generalized Additive Model (GAM)

GAM, proposed by Hastie and Tibshirani (1986) and Hastie (2018), is an extension of the GLM. This model is a non-parametric regression that relaxes both classical assumptions of normality and linearity; besides, it allows non-linear relationships between a response and explanatory variables. The great advantage of this model is that the user does not have to suggest the type of function between the variables, since it is the model that defines the relationship form. That is, instead of having to choose a single parameter β_i that best fits all the range of values of the corresponding explanatory variable (x_j), GAM establish the form of the relationship by a spline curve, which joins two or more polynomial curves. The locations of the joins are known as "knots". A cubic spline is a curve constructed as a sum of sections of cubic polynomials, joined at the ends in such a way that a continuous function is generated up to the second derivative. The form of the function will be determined by the available data and by a smoothing parameter (λ) that establishes how close the function has to adjust to the data points. Distributions allowed by GAM models include exponential family, negative binomial distribution and tweedy distribution. The last distribution includes a response variance equal to the average raised to the power p. The linear predictor can be written as a function of fixed effects $X\beta$ and smoothing function $f_k(x_k)$ with a specified parametric form for each p explanatory variable (Wood, 2017).

$$g(\mu) = X\beta + \sum_{j=1}^{p} f_k(x_k)$$

GAM models are preferred over lowess (locally weighted least squares regression models) (Cleveland, 1979), because the latter can produce predictions less than zero or greater than 1 when the data comes from a binomial distribution (Agresti, 2015).

Generalized Additive Mixed Model (GAMM)

This model is an extension of the GAM model. Here, random effects and correlation structures are included as in GLMM. The main interest focuses on the fixed effects of the model:

$$g(\mu) = X\beta + \sum_{j=1}^{p} f_j(x_j) + z_\gamma$$

where X and Z are defined as in GLMM and the functions $f_j(x_j)$ are defined as in GAM.

donde X y Z son matrices con efectos fijos y aleatorios, respectivamente, β es un vector de coeficientes de regresión, γ es un vector de efectos aleatorios con distribución normal, $\gamma \sim N(o, \psi)$, g es una función de enlace definida como en (1), y ψ es una matriz de covarianza de efectos aleatorios.

Modelo aditivo generalizado (GAM)

El GAM, propuesto por Hastie y Tibshirani (1986) y Hastie (2018), es una extensión del GLM. Este modelo es una regresión no paramétrica que relaja los supuestos clásicos de normalidad y linealidad; además, permite relaciones no lineales entre una variable de respuesta y variables explicativas. La gran ventaja de este modelo es que el usuario no tiene que sugerir el tipo de función entre las variables, ya que el modelo define la forma de la relación. Es decir, en lugar de tener que elegir un solo parámetro β_i que mejor se ajuste a todo el rango de valores de la variable explicativa correspondiente (x_j), GAM establece la forma de la relación mediante una curva spline, que une dos o más curvas polinomiales. Las ubicaciones de las uniones se conocen como "nodos". Un spline cúbico es una curva construida como una suma de secciones de polinomios cúbicos, unidos en los extremos de tal manera que se genera una función continua hasta la segunda derivada. La forma de la función estará determinada por los datos disponibles y por un parámetro de suavizado (l) que establece qué tan cerca la función a de ajustarse a los datos. Las distribuciones permitidas por los modelos GAM incluyen a la familia exponencial, la distribución binomial negativa y a la distribución tweedy. La última distribución incluye una varianza de la respuesta igual a la media elevada a la potencia p. El predictor lineal puede escribirse como una función de efectos fijos $X\beta$ y una función de suavizado $f_k(x_k)$ con una forma paramétrica específica para cada variable explicativa p (Wood, 2017).

$$g(\mu) = X\beta + \sum_{j=1}^{p} f_k(x_k)$$

Los modelos GAM son preferidos sobre los lowess (modelos de regresión de mínimos cuadrados ponderados localmente) (Cleveland, 1979), porque estos últimos pueden producir predicciones menores que cero o mayores que 1 cuando los datos provienen de una distribución binomial (Agresti, 2015).

Modelos mixtos aditivos generalizados (GAMM)

Este modelo es una extensión del modelo GAM. Aquí los efectos aleatorios y las estructuras de correlación se incluyen como en GLMM. El principal interés se centra en los efectos fijos del modelo:

Vector Generalized Linear Model (VGLM)

A variant of the GLM is the VGLM, which includes distributions that do not belong to the exponential distribution (Yee, 2019). Some of these distributions are: beta-binomial, Generalized Extreme Value Distribution, Gumbel and Inverse Gaussian. VGLM are a family of at least one hundred models which result of the combination of several distributions and link functions. One important feature of these models is that they can model variance, skewness and kurtosis as functions of explanatory variables. This approach can produce multivariate models with different distribution and link functions for each response variable (Yee, 2019).

Repeated Measures Model

Repeated Measures Model is very useful when there is interest in analyzing a variable through time (Hox *et al*. 2017). It controls for non-independence among the repeated observations for each individual, and it actually adds one or more random effects for individuals to the model. Examples of random effects are random intercept and random slope models. Random intercept model controls the variability between individuals, whereas random slope model assumes that individuals (e.g. plants) could have different slopes. Both take the form of additional residual terms. The GAM with intercept and slope random effects model can be written as:

$$g(\mu) = X\beta + \sum_{j=1}^{p} f_j(x_j) + u_{1i} + u_{2i}\,time \qquad (2)$$

where u_{1i} is the intercept random effect and u_{2i} is the slope random effect associated to time, where $u_{1i} \sim N(o,\ \tau^2{}_1)$, $u_{2i} \sim N(o,\ \tau^2{}_2)$, and $cov\ (u_{1i},\ u_{1i}) = \tau_{12}$. Thus model (2) controls the effects of individuals in order to emphasize the linear and no linear effects of regressors on response variable.

Where the variance-covariance matrix of nxn is:

$$\sum = \begin{bmatrix} \sigma^2 + \tau_2^2 & \cdots & \tau_2^2 \\ \vdots & \ddots & \vdots \\ \tau_2^2 & \cdots & \sigma^2 + \tau_2^2 \end{bmatrix}$$

Estimation of parameters

Restricted maximum likelihood method (REML) estimates the parameters of GLMM and GAMM. Maximum likelihood (ML) method estimates the parameters of GLM, GAM and VGLM. This implies that an information criterion statistic for selecting the best model like the Bayesian (BIC) cannot be

$$g(\mu) = X\beta + \sum_{j=1}^{p} f_j(x_j) + z_\gamma$$

donde X y Z se definen como en GLMM y las funciones $f_j(x_j)$ se definen como en GAM.

Modelo lineal generalizado vectorial (VGLM)

Una variante del GLM es el VGLM, que incluye distribuciones que no pertenecen a la distribución exponencial (Yee, 2019). Algunas de estas distribuciones son: beta-binomial, distribución de valor extremo generalizada, Gumbel e inversa gaussiana. VGLM son una familia de al menos cien modelos que resultan de la combinación de varias distribuciones y funciones de enlace. Una característica importante de estos modelos es que pueden modelar la varianza, la asimetría y la curtosis como funciones de variables explicativas. Este enfoque puede producir modelos multivariados con diferentes funciones de distribución y funciones de enlace para cada variable de respuesta (Yee, 2019).

Modelo de medidas repetidas

El modelo de medidas repetidas es muy útil cuando existe interés en analizar una variable a través del tiempo (Hox *et al*., 2017). Éste controla la dependencia entre las observaciones repetidas para cada individuo y, de hecho, adiciona uno o más efectos aleatorios para los individuos en el modelo. Ejemplos de efectos aleatorios son los modelos de intercepto aleatorio y de pendiente aleatoria. El modelo de intercepto aleatorio controla la variabilidad entre individuos, mientras que el modelo de pendiente aleatoria asume que los individuos (por ejemplo, las plantas) podrían tener diferentes pendientes. Ambos toman la forma de términos residuales adicionales. El modelo de efectos aleatorios GAM con intercepto y pendiente se puede escribir como:

$$g(\mu) = X\beta + \sum_{j=1}^{p} f_j(x_j) + u_{1i} + u_{2i}\,time \qquad (2)$$

donde u_{1i} es el efecto aleatorio del intercepto y u_{2i} es el efecto aleatorio de la pendiente asociado al tiempo, donde $u_{1i} \sim N(o,\ \tau^2{}_1)$, $u_{2i} \sim N(o,\ \tau^2{}_2)$, and $cov\ (u_{1i},\ u_{1i}) = \tau_{12}$. Por lo tanto, el modelo (2) controla los efectos de los individuos para enfatizar los efectos no lineales de los regresores sobre la variable de respuesta.

La matriz de varianza-covarianza de nxn es :

$$\sum = \begin{bmatrix} \sigma^2 + \tau_2^2 & \cdots & \tau_2^2 \\ \vdots & \ddots & \vdots \\ \tau_2^2 & \cdots & \sigma^2 + \tau_2^2 \end{bmatrix}$$

estimated for GLMM and GAMM models (Schwarz, 1978; Wood, 2006). A BIC statistic performs better than an AIC one because the first one introduces a penalty term for the number of parameters in the model. Therefore, for model comparison, the best model will be that one which presents a white noise behavior and a minimum variance.

The experiment

The rose flowers were produced in a Sawtooth type greenhouse, located in Tequexquinahuac, state of Mexico, Mexico. The climate in the area is temperate semi-dry, with an annual average temperature of 15.9 °C, with infrequent frosts and an average annual rainfall of 686 mm and altitude of 2450 m.

The commercial greenhouse has 5000 m^2, the soil is vertisol at pH 5.8, 4% of organic matter and controlled drip irrigation. Flower producers were interested in identifying which of two cultivars (Samurai (1) and Blush (2)) in peak seasons (period 1: April-May; period 2: September-October) was the best to accomplish the required specifications of the stem length (between 50 to 70 cm) required by the market. They wanted to know how heat units (heat), relative humidity (RH) and light (μmol m^{-2} s^{-1}) influence the stem length of roses. In order to answer these questions, an experiment with two factors, cultivars and periods, was conducted.

Environmental variables light (lux), RH (%) and heat units [(Tmax-Tmin)-5.3] (where 5.3 was the base temperature for roses growth), were monitored inside the commercial greenhouse with data logger sensors (HOBO®). Repeated measures of stem length (10 plants per 5 dates) for each cultivar and period were measured at harvest with a metric scale. The fixed effects were periods and cultivars and the random effects were plants. The interaction was of interest.

RESULTS AND DISCUSSION

Stem length (y), heat units (x_1), relative humidity (x_2) and light units (x_3) descriptive statistics showed high variability of stem length and some positive skewness of RH (Table 1). Negative kurtosis indicates that the distribution has lighter tails than the normal distribution.

Stem length density function by kernel (Epanechnikov, 1969), Anderson-Darling test (Anderson and Darling, 1954; Marsaglia and Marsaglia, 2004) and qq plots of normal, gamma, Weibull and lognormal distributions (Figure 1), showed that stem length data comes from a normal distribution. However, after including the

Estimación de parámetros

El método de máxima verosimilitud restringida (REML) estima los parámetros de GLMM y GAMM. El método de máxima verosimilitud (ML) estima los parámetros de GLM, GAM y VGLM. Esto implica que una estadística de criterio de información para seleccionar el mejor modelo como el Bayesiano (BIC) no puede estimarse para los modelos GLMM y GAMM (Schwarz, 1978; Wood, 2006). Una estadística BIC funciona mejor que una AIC porque la primera introduce un término de penalización para el número de parámetros en el modelo. Por lo tanto, en la comparación de modelos, el mejor modelo será aquel que presente un comportamiento de ruido blanco y una mínima varianza en los residuales.

El experimento

Las rosas se produjeron en un invernadero tipo Sawtooth, ubicado en Tequexquinahuac, Estado de México, México. El clima de la zona fue templado semiseco, con una temperatura media anual de 15.9 °C, con heladas poco frecuentes, una precipitación media anual de 686 mm y una altitud de 2.450 m.

El invernadero comercial tenía 5000 m^2, el suelo era de vertisol con un pH 5.8, 4% de materia orgánica y riego por goteo controlado. Los productores de flores estaban interesados en identificar cuál de los dos cultivares (Samurai (1) y Blush (2)) en las temporadas pico (período 1: abril-mayo; período 2: septiembre-octubre) era el mejor para cumplir con las especificaciones sobre la longitud del tallo (entre 50 y 70 cm) requeridas por el mercado. Querían saber cómo las unidades de calor (calor), la humedad relativa (HR) y la luz (μmol $m^{-2}s^{-1}$) influían en la longitud del tallo de las rosas. Para responder a estas preguntas se realizó un experimento con dos factores: cultivares y períodos.

Variables ambientales como: luz (lux), HR (%) y unidades de calor [(Tmax-Tmin) -5.3] (donde 5.3 fue la temperatura base para el crecimiento de rosas), se tomaron dentro del invernadero comercial con sensores de registro de datos (HOBO®). Medidas repetidas de la longitud del tallo (10 plantas por 5 fechas) para cada cultivar y período se tomaron en la cosecha con una escala métrica. Los efectos fijos fueron los períodos y cultivares y los efectos aleatorios fueron las plantas. La interacción era de interés.

RESULTADOS Y DISCUSIÓN

Las estadísticas descriptivas de longitud del tallo (y), unidades de calor (x_1), humedad relativa (x_2) y unidades de luz (x_3) mostraron una alta variabilidad de longitud del tallo y cierta asimetría positiva de la RH (Cuadro 1). La curtosis negativa indica que la

Table 1. Descriptive statistics.
Cuadro 1. Estadística descriptiva.

Variable	Mean	Median	SD	Skewness	Kurtosis	Max	Min
Stem length (cm)	45.11	45.05	25	0.1	-0.73	108.3	2.9
Heat Units	686.6	694.71	135.78	0.05	-1.03	917.38	476.72
RH (%)	82.03	81.86	1.05	0.83	-0.08	84.38	80.85
Light (mmoles $m^{-2}s^{-1}$)	448.01	446.83	20.46	-0.36	-1.04	474.09	408.75

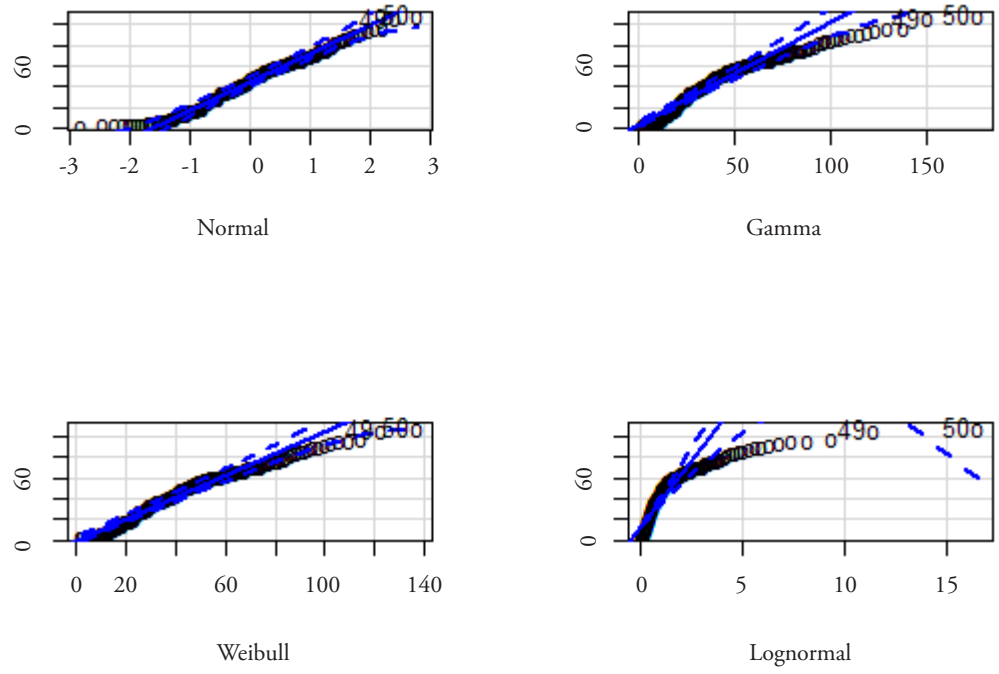

Figure 1. Normal, gamma, Weibull and lognormal qq plots of stem length (*y*).
Figura 1. Gráficas qq de las distribuciones normal, gamma, Weibull y lognormal de longitud del tallo(*y*).

explanatory variables in the model, residuals showed a gamma distribution, which was included in several models (Table 2). All the models were fitted with R (R Core Team, 2013) using the libraries lme4 (Bates *et al.*, 2015), MASS (Venables and Ripley, 2002), mgcv (Wood, 2011), gamm4 (Wood and Scheipl, 2017) and VGAM (Yee, 2019).

Models for explaining stem length as a function of environmental variables

Table 2 shows the models adjusted to length of the stem (Y), as a function of heat units (x_1), relative humidity (x_2) and light (x_3). Gaussian model (model

distribución tiene colas más ligeras que la distribución normal.

La función de densidad de la longitud del tallo por kernel (Epanechnikov, 1969), la prueba de Anderson-Darling (Anderson y Darling, 1954; Marsaglia y Marsaglia, 2004) y las gráficas qq de las distribuciones normal, gamma, Weibull y lognormal (Figura 1) mostraron que los datos de longitud del tallo proceden de una distribución normal. Sin embargo, luego de incluir las variables explicativas en el modelo, los residuos mostraron una distribución gamma, la cual fue incluida en varios modelos (Cuadro 2). Todos los modelos fueron ajustados en R (R Core Team, 2013), con las bibliotecas lme4 (Bates *et*

Table 2. Adjusted models.
Cuadro 2. Modelos ajustados.

Model	Type	Distribution	Link	BIC
A	VGLM	Normal	Indentity	1681
B	VGLM	Weibull	Identity	1696
C	GLM	Gamma	Inverse	1775
D	GLM	Gamma	Log	1696
E	GLM	Gamma	Identity	1642
F	GLMM	Gamma	Identity	.
G	GLMM	Gamma	Inverse	.
H	GLMM	Gamma	Log	.
I	GAM	Gamma	Log	1629
J	**GAM**	**Gamma**	**Identity**	**1619**
K	GAM	Normal	Identity	1647
L	GAMM	Gamma	Log	.
M	GAMM	Gamma	Identity	.

K of Table 2) produced Pearson residuals between -38 to 38, while the rest was between -3 to 3 (Figure 2). An explanation of the great variability of this model was the wrong assumption of linear relationship between the response variable and the explanatory variables. GAM and GAMM models showed residuals with minimum variability (Figure 2). The best model was the GAM (model J) with identity link function and deviance explained of 75.1%.

Model J identified significant nonlinear effect of RH ($p \leq 0.05$) and linear effects of heat units ($p \leq 0.05$) and light ($p \leq 0.10$). Model 3 presented the explicit form of Model J. Non-linearity relationship

al., 2015), MASS (Venables y Ripley, 2002), mgcv (Wood, 2011), gamm4 (Wood y Scheipl, 2017) y VGAM (Yee, 2019).

Modelos para explicar la longitud del tallo en función de variables ambientales

El Cuadro 2 muestra los modelos ajustados a la longitud del tallo (Y), en función de las unidades de calor (x_1), la humedad relativa (x_2) y la luz (x_3). El modelo gaussiano (modelo K del Cuadro 2) produjo residuales de Pearson entre -38 a 38, mientras que para el resto entre -3 a 3 (Figura 2). Una explicación

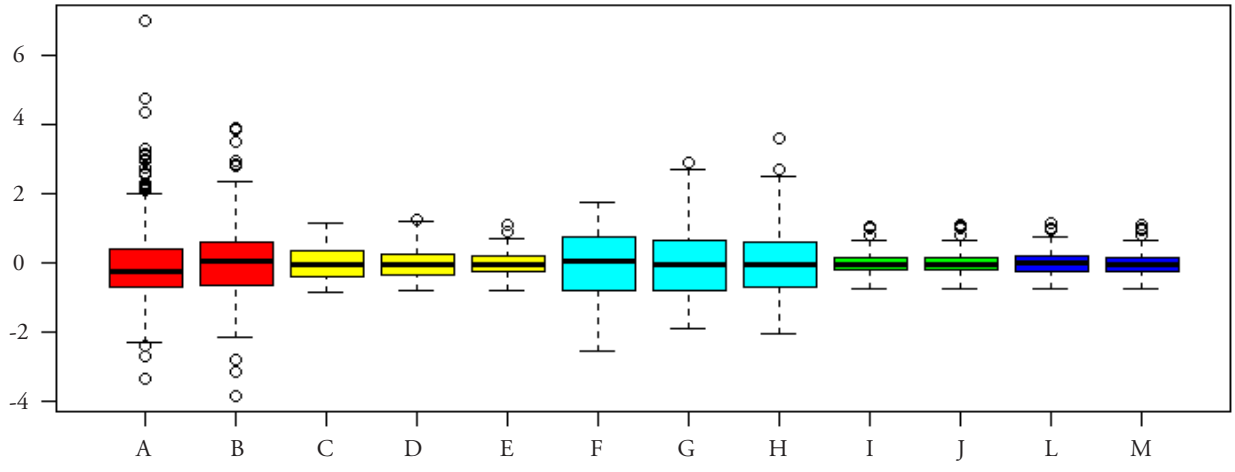

Figure 2. Residual box plots of the best 12 models of Table 1.
Figura 2. Gráficas de cajas de residuales de los 12 s mejores modelos del Cuadro 1.

between RH was approximately of degree 4 with a smooth function *s* (this function used ten knots), which was evident in Figure 3.

$$\hat{y} = 45.045 + s(heat, 1) + s(RH, 3.83) + s(light, 1) + \in; \quad (3)$$

where $y \sim$ Gamma (shape = 2.17, rate = 0.04); $\varepsilon \sim N(0, 0.11)$.

According to model (3), the best conditions to produce plants between 50 and 70 cm are 82.5 to 85% RH and 650 to 830 heat units (Figure 3). Stem length and predicted stem length, predicted by model 3, shows the quality of the model (Figure 4).

Models for detecting best production period of roses and cultivar

The stem length averages of period 1 and 2 were 63 cm and 25 cm, respectively (Figure 5). Regarding cultivars, the averages were 63 and 67 cm for cultivars 1 and 2, respectively. Box plots showed that cultivar 2 during period 1 produced the larger roses stem (Figure 5). Density functions by kernel showed

de la gran variabilidad de este modelo fue el supuesto erróneo de una relación lineal entre la variable de respuesta y las variables explicativas. Los modelos GAM y GAMM mostraron residuales con mínima variabilidad (Figura 2). El mejor modelo fue el GAM (modelo J) con función liga de identidad y devianza explicada de 75.1%.

El modelo J identificó un efecto no lineal significativo de RH ($p \leq 0.05$) y efectos lineales de unidades de calor ($p \leq 0.05$) y luz ($p \leq 0.10$). El Modelo 3 presenta de forma explícita al Modelo J. La relación de no linealidad de RH fue aproximadamente de cuarto grado, con una función suave *s* (esta función usó diez nodos), lo cual fue evidente en la Figura 3.

$$\hat{y} = 45.045 + s(heat, 1) + s(RH, 3.83) + s(light, 1) + \in; \quad (3)$$

donde $y \sim$ Gamma (shape = 2.17, rate = 0.04); $\varepsilon \sim N(0, 0.11)$.

De acuerdo al modelo (3), las mejores condiciones para producir plantas entre 50 y 70 cm, son de 82.5 a 85% de RH y de 650 a 830 unidades de calor (Figura 3). La longitud del tallo y la longitud

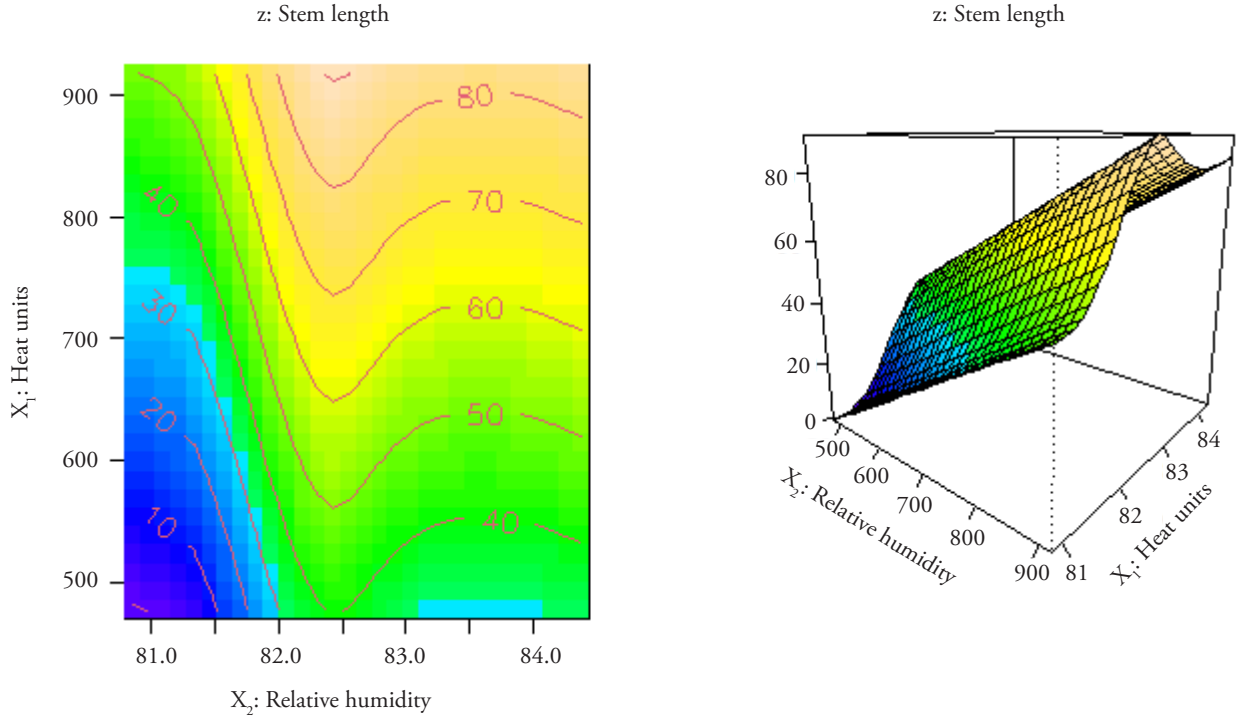

Figure 3. Heat units and Relative humidity (X_2) effects on length of the stem.
Figure 3. Efectos de unidades de calor (X_1) y humedad relativa (X_2) sobre la longitud del tallo.

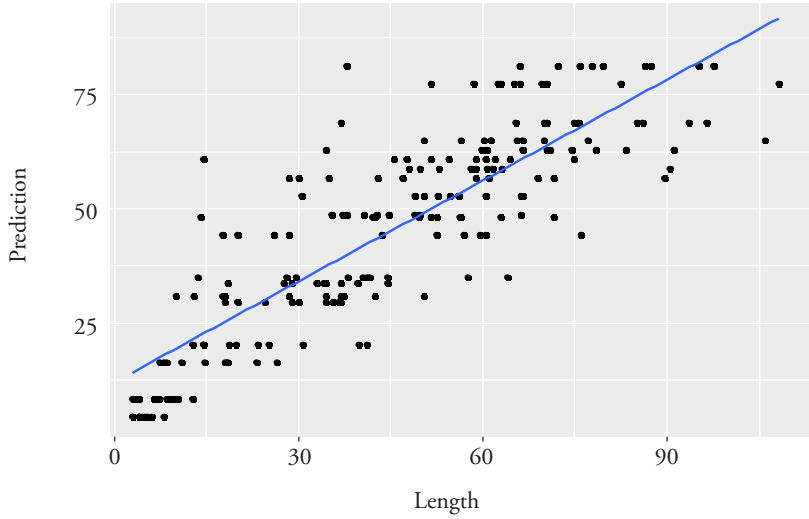

Figure 4. Stem length vs predicted stem length by GAM (model 3). Shadow color represents 95% confidence interval.
Figura 4. Longitud del tallo versus la longitud del tallo pronosticada por GAM (modelo 3).

skewed distributions of roses stem length (Figure 6).

To identify the period and cultivar of roses of best production, model 3 was extended to include cultivar and period of the year. To achieve this a GAMM model was fitted (model 4), with plant

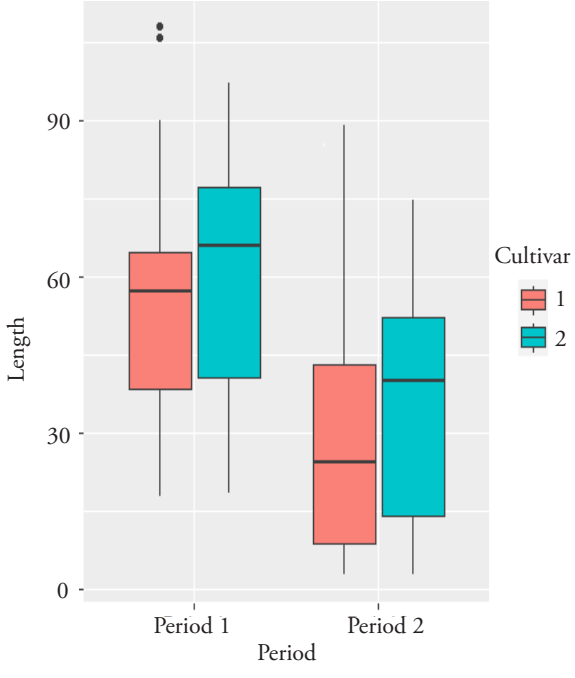

Figure 5. Box plot of stem length by period and cultivar.
Figura 5. Gráficas de caja de longitud del tallo por período y cultivar.

pronosticada por el modelo 3 muestran la calidad del modelo (Figura 4).

Modelos para detectar el mejor período de producción de rosas y cultivares

Los promedios de longitud del tallo del período 1 y 2 fueron 63 cm y 25 cm, respectivamente (Figura 5). En cuanto a los cultivares, los promedios fueron de 63 y 67 cm para los cultivares 1 y 2, respectivamente. Las gráficas de caja mostraron que el cultivar 2 durante el período 1 produjo el tallo de rosas más grande (Figura 5). Las funciones de densidad por kernel mostraron distribuciones sesgadas de la longitud del tallo de las rosas (Figura 6).

Para identificar el período y el cultivar de rosas de mejor producción, el modelo 3 fue extendido para incluir el cultivar y el período del año. Para lograr este propósito, se ajustó un modelo GAMM (modelo 4), con planta dentro de la fecha como efecto aleatorio (γ_p). La interacción no fue significativa.

$$\hat{y} = 63.06 - 38.61\ P_2 + 4.07\ C_2 + s(x_1, 1) + s(x_2, 3.83) + s(x_3, 1) + \gamma_p + \in; \quad (4)$$

donde $y \sim$ Gamma (shape = 2.17, rate = 0.04); $\varepsilon \sim N(0, 0.11)$; $\gamma_p \sim N(0, 0.12)$.

El análisis ANOVA del modelo 4 mostró que las características de calidad se pueden lograr con el cultivar 2

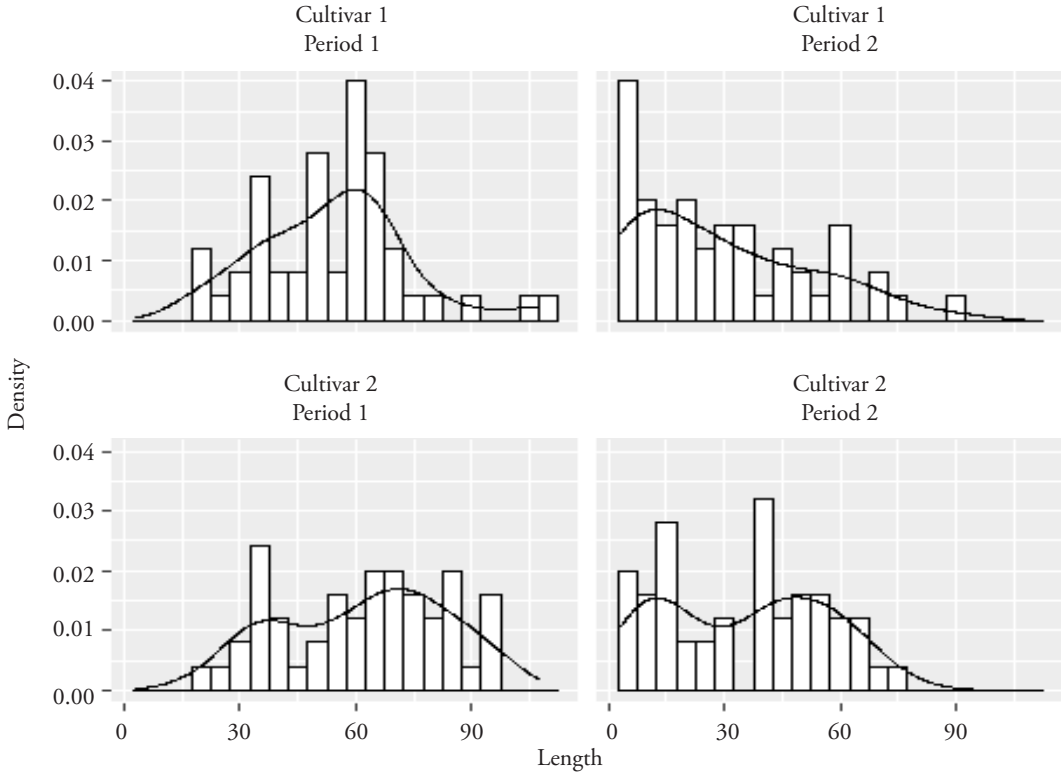

Figure 6. Density functions by kernel of stem length by cultivar and period.
Figura 6. Funciones de densidad por kernel de longitud del tallo por cultivar y período.

within date as random effect (γ_p). The interaction was not significant.

$$\hat{y}= 63.06 - 38.61\ P_2 + 4.07\ C_2 + s(x_1, 1) + s(x_2, 3.83) + s(x_3, 1) + \gamma_p + \in; \quad (4)$$

where $y \sim$ Gamma (shape = 2.17, rate = 0.04); $\varepsilon \sim N(0, 0.11)$; $\gamma_p \sim N(0, 0.12)$.

ANOVA analysis of model 4 showed that quality characteristics can be achieved with cultivar 2 (Intercept = 63.0578, p value = 2e-16; cultivar estimate = 4.0750, p value < 1.02e-07) during period 1 (period 2 estimate = -38.0612, p value < 2e-16) (Figure 7). Because there were two levels for cultivar and two for periods, it was unnecessary to use any multiple comparison test; therefore, the highest level for period and cultivar were the best one. Some point predictions as function of specific values of period, cultivar, date, heat units, RH and stem length show the quality of the model (Table 3).

(Intercepto = 63.0578, valor de p = 2e-16; estimación de cultivar = 4.0750, valor de p <1.02e-07) durante el período 1 (estimación del período 2 = -38.0612, valor de p <2e-16) (Figura 7). Debido a que había dos niveles para cultivar y dos para períodos, no fue necesario utilizar ninguna prueba de comparación múltiple; por lo tanto, el nivel más alto para período y cultivar fue el mejor. Algunas predicciones puntuales en función de valores específicos de período, cultivar, fecha, unidades de calor, RH y longitud del tallo muestran la calidad del modelo (Cuadro 3).

Análisis de medidas repetidas

El tamaño del tallo de las plantas puede ser estudiada con un modelo de medidas repetidas (5) con curvas de crecimiento de intercepto y pendiente como efectos aleatorios ($u_{1i} + u_{2i}$ time) y una estructura simétrica de la matriz de varianza-covarianza para efectos aleatorios $\hat{\sum}$ de 10x10.

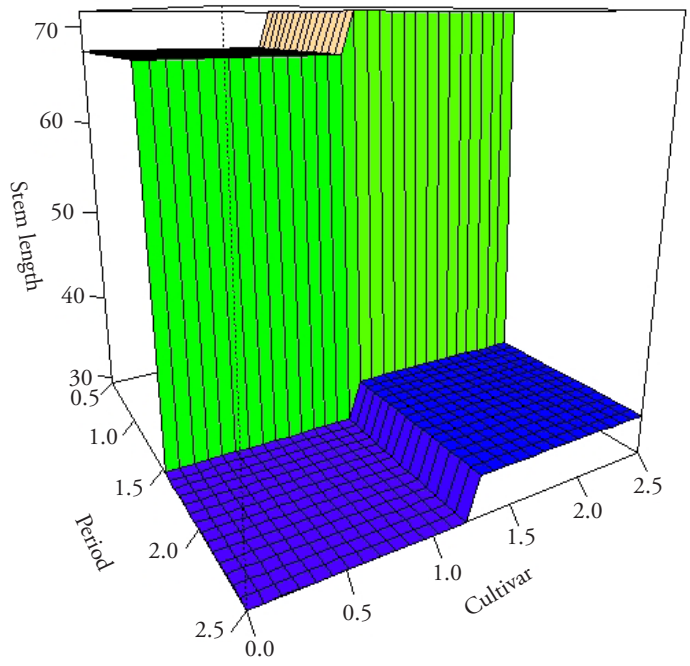

Figure 7. Predicted stem length of Model 4.
Figura 7. Longitud del tallo predicha por el Modelo 4.

Repeated measures analysis

The stem size of the plants can be studied with a repeated measures model (5) with intercept and slope growth curves as random effects (u_{1i} + u_{2i} time) and a 10x10 compound symmetry structure of the variance-covariance matrix for random effects $\hat{\Sigma}$.

$$\hat{y}= 30.98 - 27.66\ P_2 + 4.16\ C_2 + u_{1i}, + u_{2i}\ \text{time} + \in;\quad \varepsilon \sim N(0, 0.11) \qquad (5)$$

$$\hat{\sum}\begin{bmatrix} 0.11 + 0.90\hat{}2 & \cdots & 0.90\hat{}2 \\ \vdots & \ddots & \vdots \\ 0.90\hat{}2 & \cdots & 0.11 + 0.90\hat{}2 \end{bmatrix}$$

Éste incluye medidas repetidas de la longitud del tallo durante los días 0, 7, 14, 21 y 28 (Figura 8). El

Table 3. Some point predictions as a function of cultivar, period, heat units, RH and light units based on model 4.
Cuadro 3. Algunas predicciones puntuales en función de cultivar, período, unidades de calor, RH y unidades de luz basadas en el Modelo 4

Period	Cultivar	Date	Heat Units	RH	Light	Length (y)	Prediction (ŷ)
Period 1	Cultivar 1	28	816.4	82.6	433.49	66	79.89
Period 1	Cultivar 2	28	816.4	82.6	433.49	97.5	83.96
Period 1	Cultivar 1	21	718.97	82.54	433.87	65.5	67.21
Period 1	Cultivar 2	21	718.97	82.54	433.87	96.5	71.28

$$\hat{\Sigma}\begin{bmatrix} 0.11 + 0.90\hat{}2 & \cdots & 0.90\hat{}2 \\ \vdots & \ddots & \vdots \\ 0.90\hat{}2 & \cdots & 0.11 + 0.90\hat{}2 \end{bmatrix}$$

This includes repeated measures of stem length during days 0, 7, 14, 21 and 28 (Figure 8). Cultivar 1 presented more variability in slope than cultivar 2, which showed almost parallel slopes (Figure 8). This characteristic indicated a uniform increase through time of cultivar 2. The ANOVA table with intercept and slopes as sources of variation, indicated that there was an intercept difference (estimate = 30.814, std error = 0.374, t value = 82.247, p-value = $5.33e^{-13}$) and no difference.

Besides, it was noticed greater variability within cultivars (Figure 9), which led to incorporate plants as random effects, as shown in model 5.

This observation was validated with estimated random components, which were 0.70 and 0.90 for intercepts and slopes, respectively. The estimated scale parameter 0.08 indicated a good precision. Correlation between intercepts and the slopes of the cultivar 1 presentó más variabilidad en pendiente que el cultivar 2, el cual mostró pendientes casi paralelas (Figura 8). Esta característica indicó un aumento uniforme a lo largo del tiempo del cultivar 2. El cuadro del ANOVA, con intercepto y pendientes como fuentes de variación, indicó que hubo una diferencia de interceptos (estimación = 30.814, error estándar = 0.374, valor t = 82.247, valor p = $5.33e^{-13}$) y sin diferencia.

Además, se observó una mayor variabilidad dentro de los cultivares (Figura 9), lo que llevó a incorporar plantas como efectos aleatorios, como se muestra en el modelo 5.

Esta observación se validó con los componentes aleatorios estimados, los cuales fueron 0.70 y 0.90 para interceptos y pendientes, respectivamente. El parámetro estimado de escala de 0.08 indicó que la precisión fue adecuada. La correlación entre los interceptos y las pendientes de las curvas de crecimiento de las plantas fue casi una. Esto significa que las plantas que comenzaron "altas" permanecieron así a lo largo del tiempo, así como las plantas "bajas" lo siguieron siendo a través del tiempo (Figura 10).

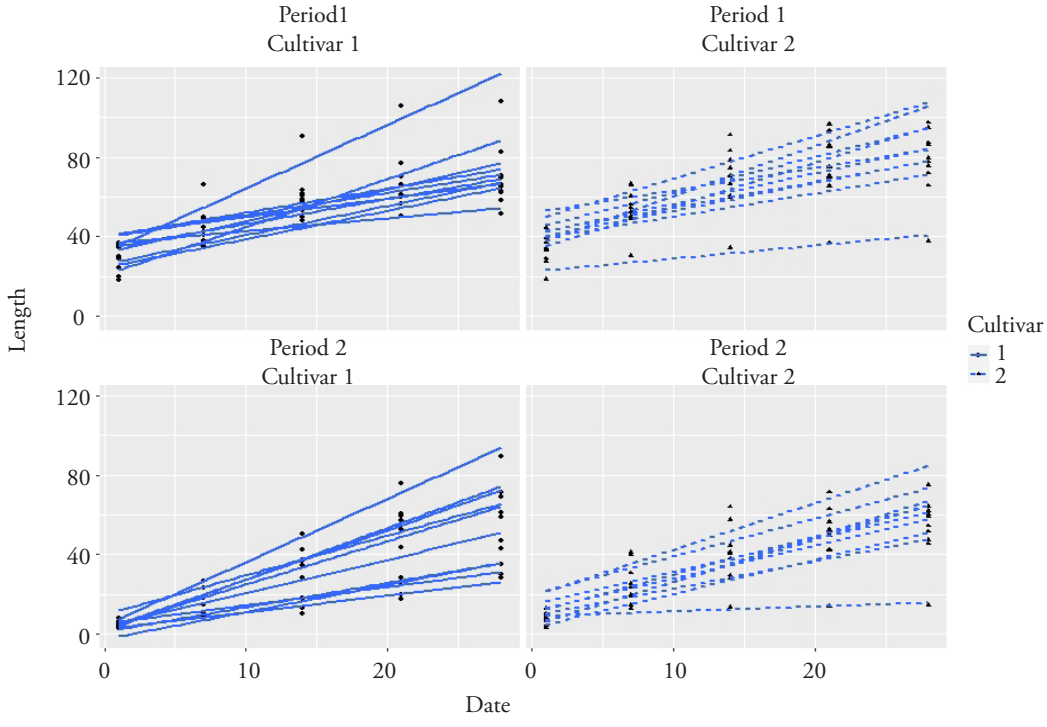

Figure 8. Linear models adjusted to the repeated measurements at 0, 7, 14, 21 and 28 days.
Figura 8. Modelos lineales ajustados a las medidas repetidas a los 0, 7, 14, 21 y 28 días.

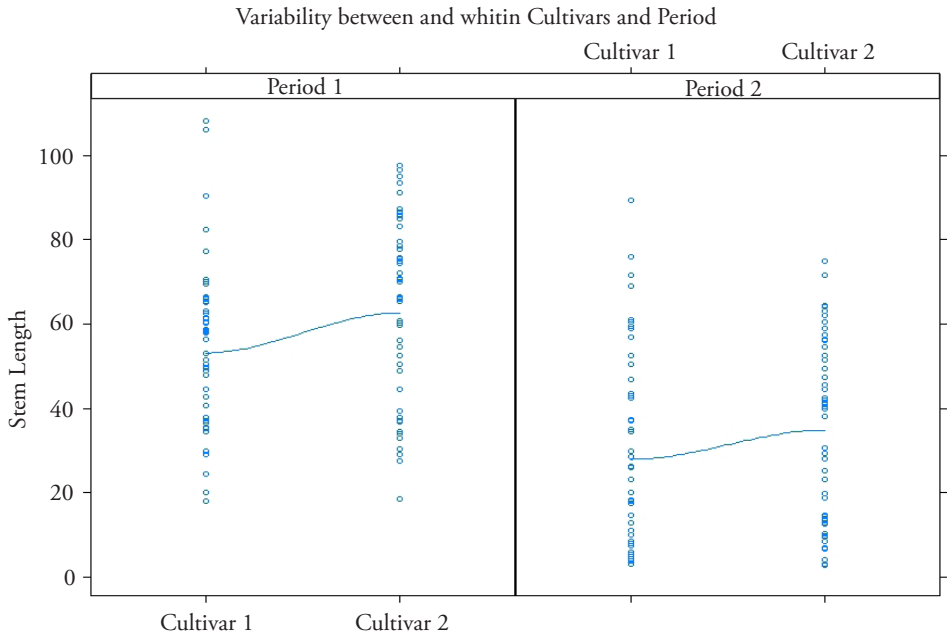

Figure 9. Variability inside cultivars and variability inside periods.
Figura 9. Variabilidad dentro de los cultivares y variabilidad dentro de los períodos.

growth curves of the plants was almost one. It means that plants which initiate "high" remained so over time, as well as "low" plants remained in this way over time (Figure 10).

CONCLUSIONS

This research showed that it is essential to know the distribution of the data, as well as the type of relationship between the explanatory and response variables, in order to correctly model it. After comparing several linear and nonlinear models, a GAM model was the best to identify the cultivar and time for producing roses between 50 to 70 cm of stems length, an appreciated quality characteristic. It identified a non-linear effect of relative humidity of a curve of approximately four degrees of freedom and linear effects of heat units and light. The best conditions to produce plants between 50 and 70 cm, were 650 to 830 heat units and between 82.5 and 85% of relative humidity. A GAMM model with plants variability as the random component, identified the April to May period with cultivar Blush as the best conditions to produced roses with the required characteristics.

CONCLUSIONES

Este estudio mostró que es fundamental conocer la distribución de los datos, así como el tipo de relación entre las variables explicativas y de respuesta, para poder modelarlas correctamente. Después de comparar varios modelos lineales y no lineales, el modelo GAM fue el mejor para identificar el cultivar y el tiempo para producir rosas con una longitud de tallo entre 50 y 70 cm, característica muy apreciada de calidad. Este modelo identificó un efecto no lineal de humedad relativa con una curva de cuatro grados de libertad aproximadamente y efectos lineales de unidades de calor y luz. Las mejores condiciones para producir plantas entre 50 y 70 cm de longitud fueron de 650 a 830 unidades de calor y entre 82.5 y 85% de humedad relativa. Un modelo GAMM, con la variabilidad de plantas como componente aleatorio, permitió identificar el período de abril a mayo para el cultivar Blush como las mejores condiciones para producir rosas con las características requeridas.

—Fin de la versión en Español—

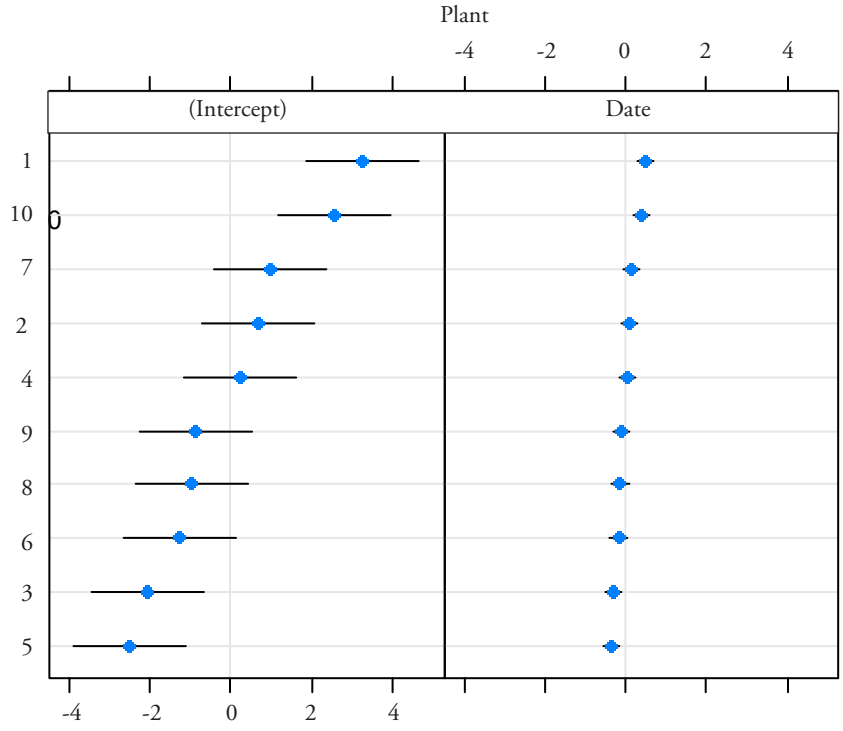

Figure 10. Intercepts and slopes centered random effects of plants.
Figura 10. Efectos aleatorios centrados de interceptos y pendientes de plantas.

LITERATURE CITED

Agresti, A. 2015. Foundations of Linear and Generalized Linear Models. Hoboken, NJ: John Wiley &Sons, Inc. 444 p.

Anderson, T. W., and D. A. Darling. 1954. A test of goodness of fit. J. Am. Statist. Assoc. 49: 765-769.

Bates, D., M. Maechler, B. Bolker, and S. Walker. 2015. Fitting Linear Mixed-Effects models using lme4. J. Stat. Software. 67: 1-48. doi:10.18637/jss.v067.i01.

Breslow, N. E., and D. G. Clayton. 1993. Approximate inference in Generalized Linear Mixed models. J. Am. Stat. Assoc. 88: 9–25.

Cattaneo M. D., M. Jansson M., and W. K. Newey. 2018. Inference in Linear Regression Models with many covariates and heteroscedasticity. J. Am. Stat. Assoc. 113: 523, 1350-1361, DOI: 10.1080/01621459.2017.1328360.

Cleveland W. S. 1979. Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc. 74: 829–936.

De Felipe, M., J. A. Gerde, and J. L. Rotundo. 2016. Soybean genetic gain in maturity groups III to V in Argentina from 1980 to 2015. Crop Sci. 56: 3066-3077.

Hastie, T. 2018. Package gam. R package version 1.16.

Hastie T., and R. Tibshirani. 1986. Generalized Additive Models (with discussion). Statistical Science. Institute of Mathematical Statistics. 1: 297-318.

Hox, J. J., M. Moerbeek, and R. Van de Schoot. 2017. Multilevel Analysis: Techniques and Applications. Routledge. 348 p.

Marsaglia, G., and J. Marsaglia. 2004. Evaluating the Anderson-darling distribution. J. Stat. Softw. 9: 1–5.

Mordor Intelligence LLP. 2020. United States Floriculture Market - Growth, Trends, and Forecast (2020 - 2025). 60 p. ID: 5865892.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Schwarz G. 1978. Estimating the dimension of the model. The Ann. Stat. 6: 461-464.

Venables W. N., and B. D. Ripley. 2002. Modern Applied Statistics with S. Fourth edition. Springer. 446 p.

Wood, S. N. 2006. Generalized Additive Models: An Introduction with R. Second edition. CRC/Chapman & Hall, Boca Raton, Florida. 496 p.

Wood, S. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. Royal Stat. Soc. Series B (Statistical Methodology). 73: 3-36.

Wood, S., and F. Scheip. 2017. gamm4: Generalized Additive Mixed Models using 'mgcv' and 'lme4'. https://CRAN.R-project.org/package=gamm4.

Yee T. W. 2019. VGAM: Vector Generalized Linear and Additive. Models. R package version 1.1-1. URL: https://CRAN.R-project.org/package=VGAM.