Spanish Journal of Agricultural Research
14(3), e0207, 12 pages (2016)
eISSN: 2171-9292
http://dx.doi.org/10.5424/sjar/2016143-8635
Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA)

RESEARCH ARTICLE

OPEN ACCESS

Soybean yield modeling using bootstrap methods for small samples

Gustavo H. Dalposso[1], Miguel A. Uribe-Opazo[2] and Jerry A. Johann[2]

[1] *Federal Technological University of Paraná (UTFPR), 19 Cristo Rei Street, 85902-490, Toledo, PR, Brazil.* [2] *Western Paraná State University (UNIOESTE), 2069 Universitária Street, 85819-110, Cascavel, PR, Brazil.*

Abstract

One of the problems that occur when working with regression models is regarding the sample size; once the statistical methods used in inferential analyzes are asymptotic if the sample is small the analysis may be compromised because the estimates will be biased. An alternative is to use the bootstrap methodology, which in its non-parametric version does not need to guess or know the probability distribution that generated the original sample. In this work we used a set of soybean yield data and physical and chemical soil properties formed with fewer samples to determine a multiple linear regression model. Bootstrap methods were used for variable selection, identification of influential points and for determination of confidence intervals of the model parameters. The results showed that the bootstrap methods enabled us to select the physical and chemical soil properties, which were significant in the construction of the soybean yield regression model, construct the confidence intervals of the parameters and identify the points that had great influence on the estimated parameters.

Additional key words: multiple linear regression; model selection; bootstrap global influence diagnosis; bootstrap confidence intervals.

Abbreviations used: AIC (Akaike information criterion); BC (bias corrected); Ca (calcium, $cmol_c/dm^3$); Des (soil density, g/cm^3; Des_1, from 0 to 0.1 m; Des_2, from 0.1 to 0.2 m; Des_3, from 0.2 to 0.3 m depths; JaB (jackknife-after-bootstrap); K (potassium, mg/dm^3); Mg (magnesium, $cmol_c/dm^3$); Mn (manganese, mg/dm^3); OLS (ordinary least squares); P (phosphorus, mg/dm^3); Prod (soybean yield, t/ha); R^2_{Adj} (adjusted coefficient of determination); RMSE (root mean square error); SRP_1 (soil penetration resistance, MPa) from 0 to 0.1 m depth; SRP_2 (soil penetration resistance, MPa) from 0.1 to 0.2 m depth; SRP_3 (soil penetration resistance, MPa) from 0.2 to 0.3 m depth.

Authors' contributions: Conceptualized the paper, statistical analysis of data, final revision and discussion: GHD, MAUO and JAJ. Reviewing the literature and editing the working versions of the manuscript: GHD.

Citation: Dalposso, G. H.; Uribe-Opazo, M. A.; Johann, J. A. (2016). Soybean yield modeling using bootstrap methods for small samples. Spanish Journal of Agricultural Research, Volume 14, Issue 3, e0207. http://dx.doi.org/10.5424/sjar/2016143-8635.

Received: 13 Sep 2015. **Accepted**: 07 Jul 2016

Funding: Araucária Foundation; Council for Scientific and Technological Development (CNPq); Coordination for the Improvement of Higher Education Personnel (CAPES); Post-Graduate Program in Agricultural Engineering (PGEAGRI); Federal Technological University of Paraná (UTFPR).

Competing interests: The authors have declared that no competing interests exist.

Correspondence should be addressed to Gustavo H. Dalposso: gustavodalposso@utfpr.edu.br

Introduction

Soybean (*Glycine max* (L.) Merrill) is one of the most economically significant crops worldwide (Kulcheski *et al.*, 2016), and multiple linear regression models are constantly being developed to partially explain yield variations in that crop. When modeling soybean yield, variables concerning agricultural meteorology (Penalba *et al.*, 2007; Tao *et al.*, 2008), agriculture (Zheng *et al.*, 2009), management (Lobell *et al.*, 2005), vegetation index (Mercante *et al.*, 2010) and soil parameters (Garcia-Paredes *et al.*, 2000) are often utilized. Models differ according to the nature of the explanatory variables used in the modeling process, for which a review on model categories in use can be found in Vera-Diaz *et al.* (2008).

Considering that determining the values of certain variables is often a burdensome and arduous task, in some cases analyses are carried out on small samples[1]. This fact may call into question the inferences being

[1] There is no accepted definition of what constitutes a small sample, as such sample size depends on a number of factors, including the reliability of the estimate, and the relative variance of the variable under consideration (Levy & Lemeshow, 1980). Aiken & West (1991) regarded $n \leq 60$ as a small sample, whereas Ireland (2010) considered $n \leq 30$.

made, since traditional inference methods are asymptotic and standard errors and confidence intervals may be biased in small samples, as explained by Hao & Naiman (2010). Adopting more parsimonious models and determining influential points, are procedures that can also provide misleading results when working with small sample sets. Kamo *et al.* (2013) explain that the Akaike information criterion - AIC (Akaike, 1973) used for model selection presents a bias that cannot be ignored, especially with small samples, given that it is derived from asymptotic properties. Regarding the diagnostic measures of overall influences one problem is related to its cutoff points. According to Martin & Roberts (2010) they are based on large sample theory and therefore may not be suitable for small samples.

An alternative to traditional inference methods is the use of the bootstrap, a simulation method developed by Efron (1979) which uses resampling with replacement of the sample data set to perform statistical inferences such as hypothesis testing and determination of confidence intervals (Dubreuil *et al.*, 2014). The bootstrap method has applications in regression analysis (Rahman, 2014), model selection (Al-Marshadi, 2011) and definition of global influence diagnostics (Beyaztas & Alin, 2013).

By comparing the results obtained from bootstrap methods with results of asymptotic methods, Chaves-Neto & Faria (2015) conclude that bootstrap performed well in samples of all sizes and was higher than the asymptotic method in small samples. Although bootstrap is a well-known technique and is frequently employed in agricultural studies – as seen in works by Sabaghnia *et al.* (2010), García-Gallego *et al.* (2015), Losada *et al.* (2015) and Sutton *et al.* (2016) – the development of statistical and computing models has led to the study of new techniques based on the bootstrap method.

The objective of this work was to utilize bootstrap methods to select explanatory variables, investigate the existence of influential points through diagnostic analysis, and obtain confidence intervals for the parameters of a multiple linear regression model for soybean yield considering physical and chemical soil properties as explanatory variables.

Material and methods

Study area and data

The data used are from the agricultural year 2013/2014 and from a commercial farming area of 167.35 hectares located in the western region of Paraná, Brazil, near the city of Cascavel, with center coordinates latitude 24°57’18’’S and longitude 53°34’29’’W and average altitude of 714 m (Fig. 1). Climate in the region is mesothermal and super humid temperate, climate type Cfa (Koeppen) and soil is classified as a dystropheric

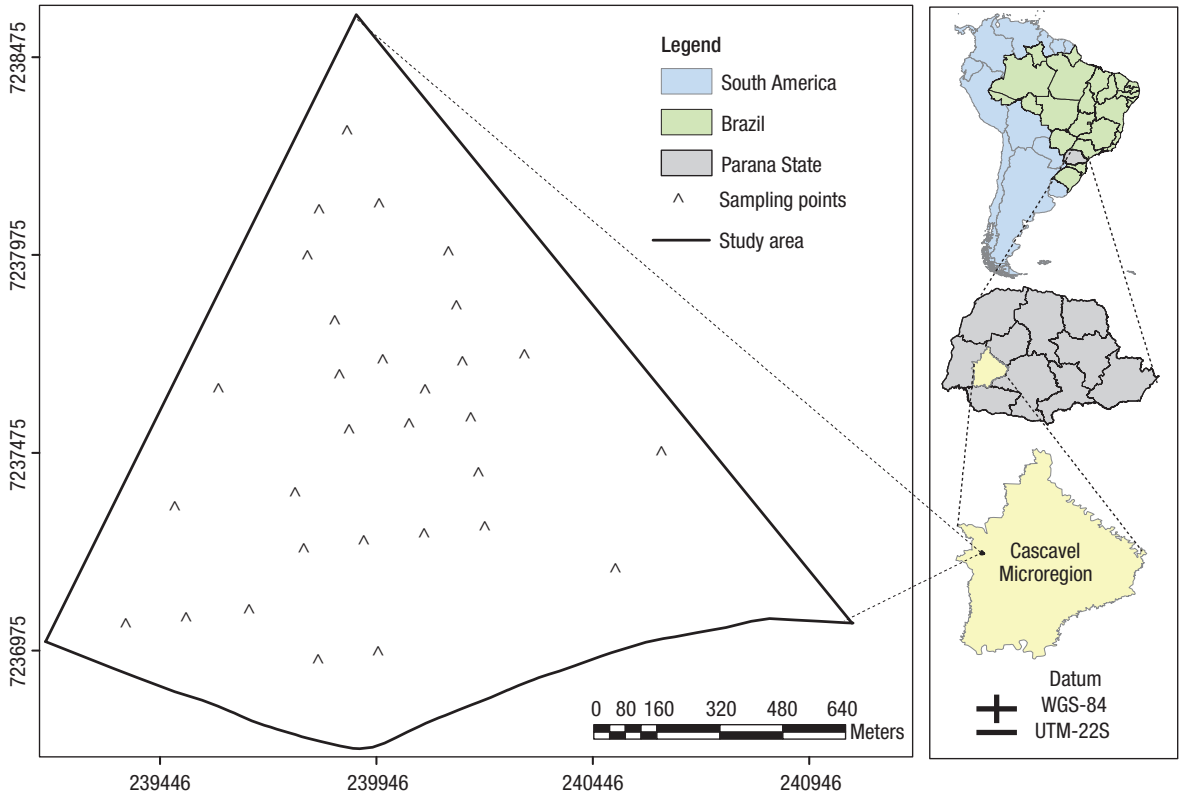

Figure 1. Location map of the study area.

red latosol with clay texture (EMBRAPA, 2013). Given a set of 30 Prod (soybean yield, t/ha) points uncorrelated the randomness was confirmed by the runs-test algorithm for randomness (Siegel, 1956). The respective values of the explanatory variables SRP_1, SRP_2 and SRP_3 (soil penetration resistances, MPa, from 0 to 0.1 m, 0.1 to 0.2 m and 0.2 to 0.3 m depths, respectively), Ca (calcium, $cmol_c/dm^3$), Mg (magnesium, $cmol_c/dm^3$), K (potassium, mg/dm^3), P (phosphorus, mg/dm^3), Mn (manganese, mg/dm^3), Des_1, Des_2 and Des_3 (soil densities, g/cm^3, from 0 to 0.1 m, 0.1 to 0.2 m and 0.2 to 0.3 m depths, respectively) have all been considered for each productivity value. The use of physical and chemical soil properties as explanatory variables is common practice in field surveys, as variations in soil properties account for most of crop yield variations, according to Khakural *et al.* (1999).

Exploratory analysis and modeling

Descriptive statistics of the variables under study were calculated and a multicollinearity[2] analysis of the explanatory variables was performed. A multiple linear regression model was built to describe the relationship between soybean yield and soil properties, with parameters estimated by the ordinary least squares (OLS) method.

Paired bootstrap

To determine the bootstrap replicates of the parameters of the regression model we used the paired bootstrap method (Freedman, 1981), presented in the following algorithm.

Algorithm 1: Paired bootstrap.

(a) Consider the matrix [**Y**,**X**] formed with the original data; (b) get a new matrix [$\mathbf{Y}^{*(1)}$,$\mathbf{X}^{*(1)}$] making a resampling with replacement of matrix rows [**Y**,**X**]; (c) find $\hat{\beta}^{*(1)}$ from [$\mathbf{Y}^{*(1)}$,$\mathbf{X}^{*(1)}$] in the same manner as $\hat{\beta}$ is calculated from the original data [**Y**,**X**]; (d) Get the bootstrap distribution calculating $\hat{\beta}^{*(b)}$, b = 1,…,B.

Confidence intervals using bootstrap

In order to determine the bootstrap intervals for the parameters of the regression models we used the percentile method (Efron, 1982) and BC (bias corrected) (Efron & Tibshirani, 1986). The Efron's percentile interval with confidence level $(1 - \alpha)\%$ was obtained by ordering the bootstrap replicates from the parameters $\hat{\theta}_i^*$, i = 1,…,B, and excluding $(\alpha/2)\%$ from the replicates situated in its ends. The technique employed to build the BC confidence interval utilizes a value known as constant-bias-correcting to fit the bootstrap distribution of $\hat{\theta}$; a roadmap to determining this interval can be found in Shasha & Wilson (2011).

Models selection using bootstrap

For the models selection the bootstrap method was used as proposed by Austin & Tu (2004), and presented in Algorithm 2, which combines bootstrap resampling with automated methods of variables selection.

Algorithm 2: Models selection method using bootstrap.

(a) Consider the matrix [**Y,X**] formed with the original data; (b) get B resamples of the previous matrix using the paired bootstrap method; (c) for each resample adjust one model and apply the backward method via AIC; (d) for each variable determine how often it was selected in the B models and the percentage of times in which estimated parameters presented positive and negative signs; (e) use the results of the previous step to determine candidate models and select the best model.

Global influence diagnostics using bootstrap in the response variable

In order to investigate the existence of influential points it was held the method proposed by Martin & Roberts (2010), bearing in mind the Cook's distance - D_i (Cook, 1977) as a measure of influence. Algorithm 3 shows the method proposed by Martin & Roberts, which is based on JaB (jackknife-after-bootstrap) technique developed by Efron (1992).

Algorithm 3: Determining cutting point D_i using JaB.

(a) Adjust the proposed model to the original dataset and estimate D_i, i = 1,…,n; (b) build B bootstrap samples using paired bootstrap method; (c) (JaB step) for each x_i sample of the original dataset consider the bootstrap samples set which do not contain the x_i

[2] Multicollinearity refers to high correlation among the independent variables and its existence tends to inflate the variances of the parameter estimates (Freud & Littell, 2000). Multicollinearity is often measured by diagnostics called variance inflation factors ($VIF = 1/1 - R^2$) where the R^2 is the coefficient of determination of the regression of independent variable x on all other independent variables in the postulated model. As a rule of thumb, when the VIF >10 we conclude that multicollinearity is a problem and that we should not base our decisions on the magnitude and sign of the regression coefficients (Hoerl & Snee, 2012).

sample (approximately B/e groups[3]) and for each sample of this group estimate the n values for Cook's distance; group all n·(B/e) values into a single vector; (d) the quantile 2.5% and 97.5% of the distribution generated by n·(B/e) values of Cook's distances are used as cutting points and if the D_i value is outside this interval then x_i is marked as an influential point.

Jackknife-after-Bootstrap graphic

The JaB technique provides another resource for establishing the effect of individual observations on the bootstrap distribution through development of the JaB plot (Efron, 1992). Based on the original [**Y**,**X**] dataset, consider the dataset $[\mathbf{Y}_{(i)},\mathbf{X}_{(i)}]$ obtained by deleting the i row in the original dataset and calculate the statistic of interest, denoted by $s_{(i)}$. The jackknife influence function for the statistic of interest is defined by:

$$u_i\{s\} = (n-1)(s_{(.)} - s_{(i)}), \quad [1]$$

where $s_{(.)} = \left[\sum_{i=1}^{n} s_{(i)}\right] / n$.

Intuitively, points with high positive or negative values of $u_i\{s\}$ have a high influence on the calculated statistic. To provide a clearer interpretation, the relative jackknife influence function shown in Eq. [2] is commonly used, being the number two the value established as the cutoff point (Efron, 1992). These values are ascending ordered and marked on the abscises axis.

$$u_i^{\uparrow}\{s\} = u_i\{s\} / \left[\sum_j u_j\{s\}^2 / (n-1)\right]^{(1/2)}. \quad [2]$$

After calculating the jackknife influence values for each point i, of the dataset, seven ordered pairs are determined, namely $(u_i^{\uparrow}\{s\}, P_k)$, k = {5,10,16,50,94,90,95} where P_k represents the k-th percentile of the bootstrap distribution formed with bootstrap replicates calculated from those bootstrap samples which do not have point i. For each percentile the neighboring ordered pairs are linked thus forming graphics, which are compared with dashed line segments perpendicular to the ordinate axis in points P_k, k = {5,10,16,50,84,90,95}, calculated from full bootstrap distribution formed by 3000 bootstrap replicates. The analysis is performed highlighting those points surpassing the cutoff point and comparing bootstrap distributions.

Computing resources

The analyses carried out in this work were developed in R statistical software (R Core Team, 2014). The bootstrap replicates used to determine the empirical distributions of model parameters were determined by the function lm.boot of package simpleboot (Peng, 2008), and confidence intervals were implemented manually. In order to determine the statistics related to the model selection method, function boot.stepAIC of package bootStepAIC was used (Rizopoulos, 2009). The algorithm utilized to determine the cutoff point for Cook's distance bootstrap was implemented with Cook's distance calculated by function cooks.distance of package stats, and JaB graphs were implemented by the authors.

Results

Descriptive statistics of the explanatory variables indicated homogeneous behavior of the variables, with no multicollinearity found. The multiple linear regression model of soybean yield, estimated through OLS considering all explanatory variables (Eq. [3]), showed an adjusted coefficient of determination (R^2_{Adj}) of 0.41 and root mean square error (RMSE) of 0.33.

$$\begin{aligned}\text{Prod} = {} & 8.858 - 0.271\text{SRP}_1 + 0.117\text{SRP}_2 - 0.003\text{SRP}_3 + 0.288\text{Ca} - 0.367\text{Mg} \\ & + 1.208\text{K} - 0.067\text{P} - 0.012\text{Mn} - 0.629\text{Des}_1 - 2.684\text{Des}_2 + 0.925\text{Des}_3. \end{aligned} \quad [3]$$

It could be observed that estimates for those parameters associated with SRP_1, SRP_3, Des_1 and Des_2 variables showed negative signs, indicating that an increase in the value of these variables implies a reduction in soybean yield (Eq. [3]). The parameters estimation associated to SRP_2 and Des_3 variables from the Eq. [3] showed different signals from the expected scenario, since it indicates a direct relation from such variables towards soybean productivity (Eq. [3]). The positive estimate signal, from the associated parameter of the variable K, indicates that, while maintaining other variables constant, an increase in one unit in the K variable produces an increase in soybean productivity, at a rate of 1.208 t/ha (Eq. [3]). The bootstrap intervals were determined with reliability of 95% for the parameters of the multiple linear regression model using the techniques of bootstrap percentile by Efron and BC bootstrap (Table 1).

It was observed that the vast majority of the confidence intervals, determined by the bootstrap technique, contained zero indicating, that with exception of the

[3] Given the sample set $\{y_1,\ldots,y_n\}$ the probability of y_j not being included in a bootstrap sample is $(1\text{-}n^{-1})^n = e^{-1}$, thus in B bootstrap samples the number of simulations that do not include y_j is approximately $B.e^{-1}$ (Davison & Hinkley, 1997). Thus, if we want to determine whether an individual data point is influential or not, and to obtain 1000 resamples without this individual data point, about $1000e \approx 3000$ resamples are required (Beyaztas & Alin, 2013).

Table 1. Nonparametric bootstrap 95% confidence intervals for the parameters of the multiple regression linear model of soybean yield considering all explanatory variables.

Parameters[1]	Efron's percentile			BC[3]		
	$\hat{\theta}_i$	$\hat{\theta}_u$	Amplitude[2]	$\hat{\theta}_i$	$\hat{\theta}_u$	Amplitude
β_{SRP_1}	-0.547	0.130	0.677	-0.530	0.197	0.727
β_{SRP_2}	-0.457	0.750	1.208	-0.423	0.798	1.221
β_{SRP_3}	-1.260	0.923	2.183	-1.336	0.888	2.224
β_{Ca}	-0.214	0.685	0.898	-0.268	0.643	0.911
β_{Mg}	-1.219	0.589	1.808	-1.168	0.654	1.821
β_K	-4.771	6.607	11.378	-5.463	5.928	11.390
β_P	-0.150	-0.007	0.143	-0.160	-0.012	0.147
β_{Mn}	-0.027	0.008	0.035	-0.025	0.009	0.035
β_{Des_1}	-4.412	5.076	9.487	-3.937	6.213	10.150
β_{Des_2}	-6.426	0.585	7.012	-6.505	0.524	7.029
β_{Des_3}	-2.052	2.872	4.925	-2.677	2.498	5.175
Intercept	2.903	13.952	11.049	2.965	14.279	11.314

[1]β_i: parameters associated with the variable i = {SRP_1, SRP_2, SRP_3, Ca, Mg, K, P, Mn, Des_1, Des_2, Des_3}; SRP_1, SRP_2 and SRP_3: soil penetration resistances, MPa, from 0 to 0.1 m, 0.1 to 0.2 m and 0.2 to 0.3 m depths, respectively; Ca: calcium, $cmol_c/dm^3$; Mg: magnesium, $cmol_c/dm^3$; K: potassium, mg/dm^3; P: phosphorus, mg/dm^3; Mn: manganese, mg/dm^3; Des_1, Des_2 and Des_3: soil densities, g/cm^3, from 0 to 0.1 m, 0.1 to 0.2 m and 0.2 to 0.3 m depths, respectively; [2]Amplitude: $\hat{\theta}_u - \hat{\theta}_i$; $\hat{\theta}_i$: lower limit; $\hat{\theta}_u$: upper limit; [3]BC: bias corrected.

variable P, the other explanatory variables may not be individually significant. In search for a more appropriate multiple linear regression model it was applied the model selection method using bootstrap considering 1000 resamples (Table 2). It was observed that, of the 1000 models for which bootstrap resamples had been adjusted, by applying the backward selection method with statistical Akaike – AIC to each of them, the result showed that in 91% of the models the predictor variable P was selected, indicating that phosphorus is an important soil attribute for soybean yield prediction. Furthermore, it was observed that in 100% of models in which phosphorus had been selected, its estimated parameter was negative, which ensures that when other variables are held constant an increase in phosphorus level implies reduction of soybean yield.

Other variables selected for most models were Des_2 with a selection percentage of 87%, Ca with 81% and SRP_1 with 79%. Analyzing the signs of the estimated parameters associated with these variables in the models in which they were selected it is highlighted that in 94% of models in which the Ca variable was selected the sign of its estimated parameter was positive, suggesting the increase in value of this variable contributes for increasing soybean yield. For those estimated parameters associated with SRP_1 and Des_2 variables, in 98% of models in which they were selected the signals were negative. It is clear that some variables may not be useful to explain soybean yield behavior. For example, among the 1000 models obtained, the Des_1 variable was selected in only 500 and additionally for 180 of those the estimated parameter sign was positive and for 320 of those the sign was negative, thus, this set of oscillations is a guarantee that this variable is not significant, therefore, can be deleted without causing damage to the modeling. A similar case occurs with SRP_3 variable, as well as being selected in only 460 models, the appropriate sign of its

Table 2. Selection percentage of variables and percentage of positive and negative signs of the estimated parameters obtained by applying the backward method via Akaike information criterion (AIC) in 1000 models generated by bootstrap.

Selection percentage		Signs of the estimated parameters		
Variables[1]	pct	Parameters[2]	pct +	pct –
P	91	β_P	0	100
Des_2	87	β_{Des_2}	2	98
Ca	81	β_{Ca}	94	6
SRP_1	79	β_{SRP_1}	2	98
Mn	75	β_{Mn}	6	94
Mg	71	β_{Mg}	9	91
Des_3	61	β_{Des_3}	80	20
Des_1	50	β_{Des_1}	36	64
K	48	β_K	84	16
SRP_3	46	β_{SRP_3}	50	50
SRP_2	42	β_{SRP_2}	78	22

[1]SRP_1, SRP_2 and SRP_3: soil penetration resistances, MPa, from 0 to 0.1 m, 0.1 to 0.2 m and 0.2 to 0.3 m depths, respectively; Ca: calcium, $cmol_c/dm^3$; Mg: magnesium, $cmol_c/dm^3$; K: potassium, mg/dm^3; P: phosphorus, mg/dm^3; Mn: manganese, mg/dm^3. Des_1, Des_2 and Des_3: soil densities, g/cm^3, from 0 to 0.1 m, 0.1 to 0.2 m and 0.2 to 0.3 m depths, respectively; [2]β_i: parameter associated with the variable i = {P, Des_2, Ca, Mn, Mg, Des_3, Des_1, K, SRP_3, SRP_2}; pct: selection percentage; pct+: percentage of positive signs; pct-: percentage of negative signs.

estimated parameter cannot be identified considering that 230 models had a positive sign and 230 had a negative sign. As per the parameters estimates associated to SRP_2 and Des_3 variables, it showed opposite signals from the expected scenario (Eq. [3]), it is desirable to verify the importance of such variables for modeling purposes. Although the positive signals from the associated estimated parameters to such variables appear in a great part of the models (80% and 78%, respectively), the selection percentages were not very elevated (61% and 42%, respectively) and, therefore, there were evidences that they were not significant and could be removed from modeling (Table 2). In view of these observations four models were set to be analyzed, namely M_{81}, M_{79}, M_{75} and M_{71} (Table 3), and each of them was determined according to a number of explanatory variables selected in accordance with how many times they had been selected in the bootstrap models.

Regressors present in the M_{81} model can explain only 37% of the soybean yield variation, a result lower than that obtained when considering the model containing all the explanatory variables (R^2_{Adj} = 0.41). The M_{75} (R^2_{Adj} = 0.42) and M_{71} (R^2_{Adj} = 0.49) models provided a greater degree of explanation between the explanatory variables and soybean yield than the full model, while the M_{79} (R^2_{Adj} = 0.41) model provided an equivalent level of explanation, however, these models had a higher RMSE compared to the complete model (RMSE = 0.33) and that difference is most evident in the M_{79} model (RMSE = 0.39). As the M_{71} model explained 49% of the soybean yield variation and RMSE of this model (RMSE = 0.34) is close to RMSE of the complete model (RMSE = 0.33) the M_{71} model was chosen as best adjusted model to soybean yield and analysis was performed using JaB to investigate the existence of influential points.

It is noteworthy to mention no points were detected as influential when value 1 is established as the cutoff point (Fig. 2). The same is true when considering the criteria that detects point i as influent if D_i is higher than the median of the distribution F of Snedecor with free-

Table 3. Parameters estimation and statistics for the multiple regression linear models of soybean yield.

Models[2]	Parameters[1]							Statistics	
	Intercept	β_P	β_{Des_2}	β_{Ca}	β_{SRP_1}	β_{Mn}	β_{Mg}	R^2_{Adj}	RMSE
M_{81}	7.827	-0.079	-1.994	0.099				0.37	0.41
M_{79}	8.482	-0.074	-2.185	0.103	-0.162			0.41	0.39
M_{75}	8.894	-0.067	-2.346	0.146	-0.179	-0.007		0.42	0.37
M_{71}	9.220	-0.076	-2.367	0.356	-0.221	-0.012	-0.479	0.49	0.34

[1] β_i: parameter associated with the variable i = {P, Des_2, Ca, SRP_1,Mn, Mg}; P: phosphorus, mg/dm^3; Des_2: soil density, g/cm^3, from 0.1 to 0.2 m depth; Ca: calcium, cmol$_c$/dm^3; SRP_1: soil penetration resistance, MPa, from 0 to 0.1 m depth; Mn: manganese, mg/dm^3; Mg: magnesium, cmol$_c$/dm^3; [2] M_i: model containing the variables selected in at least i = {81, 79, 75, 71}% of the bootstrap models; R^2_{Adj}: adjusted coefficient of determination; RMSE: root mean square error.

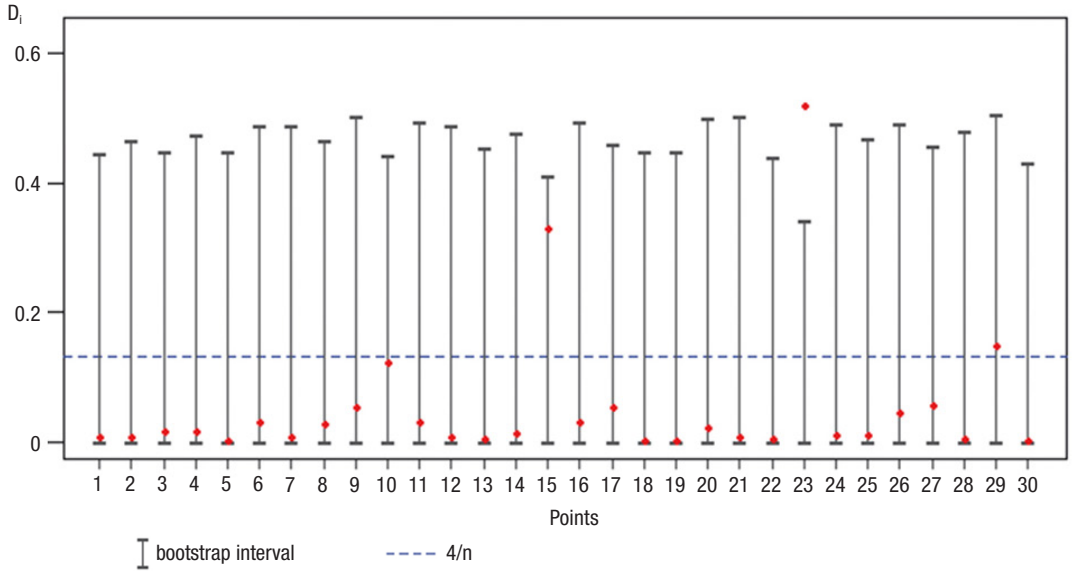

Figure 2. Determination of influential points using Cook's distance (D_i) with JaB methodology. Points that cross the dashed line are considered influent according to Cook's distance (D_i), while points outside the bootstrap interval are considered influent from jackknife-after-bootstrap (JaB) analysis.

dom degrees of p = 6 and n - p = 24 once the cutoff point is 2.50 to these, thus they were also not detected as influential points. Considering 4/n ≈ 0.13 as cutoff point, the points 15, 23 and 29 were detected as influential indicating these points can change the estimation of the parameters in the regression model, so it is important to investigate the model behavior without the use of these points. It should be emphasized that only point 23 was detected as being influential through analysis using Cook's distance (D_i) with JaB methodology.

JaB graphs were created to help identify influential points, they give a visual interpretation of how a particular point affects the bootstrap distribution for the estimation of parameters in M_{71} (Fig. 3). Observing the graphs in Fig. 3, it was noted that points 10, 15, 23 and 29 were detected as influent.

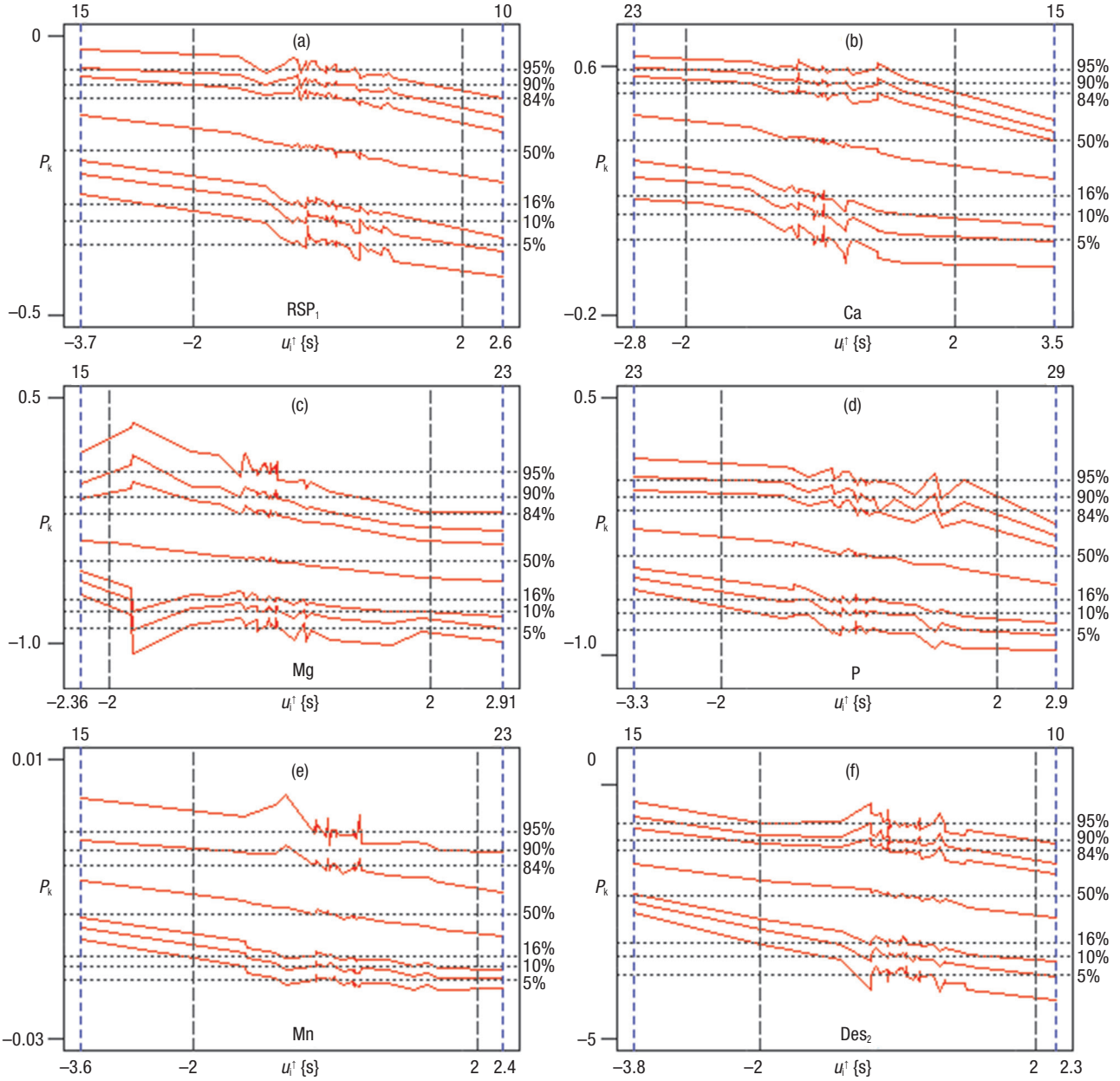

Figure 3. Jackknife-after-bootstrap (JaB) plots for the parameter estimates associated with the explicative variables SRP_1: soil penetration resistances from 0 to 0.1 m (a), Ca: calcium (b), Mg: magnesium (c), P: phosphorus (d), Mn: manganese (e) and Des_2: soil densities from 0.1 to 0.2 m (f) of the model containing the variables selected in at least 71% of the bootstrap models (M_{71}). $u_i^{\uparrow}$ {s}: Values of the relative jackknife influence function in growing order. P_k: *k*-th percentile of the bootstrap distribution, k = {5,10,16,50,84,90,95}. The vertical dashed lines that pass through values -2 and 2 represent the cutoff interval. The points allocated outside this interval are shown in the upper part of each graph and their relative jackknife influence values are highlighted with vertical dashed lines. The red lines represent the variations in these percentiles and the intersections with the blue dotted line representative of point *i* represent the percentiles obtained when the bootstrap distribution is formed with the bootstrap replicates calculated from the bootstrap samples that do not contain point *i*.

Two new models were adjusted to the variables P, Des_2, Ca, SRP_1, Mn, Mg as to measure the effect of influential points in modeling. The $M_{71\text{-}\{15,23,29\}}$ model was adjusted to the data set without points (15, 23, 29), as these were detected as influential by traditional Cook distance method with cutoff point of 4/n. The $M_{71\text{-}\{10,15,23,29\}}$ was also adjusted to the data set without points (10, 15, 23, 29) for these were considered as influential by analysis using JaB (Table 4).

The $M_{71\text{-}\{15,23,29\}}$ model adjusted to data set without sample elements 15, 23 and 29, which were identified as influential by the traditional method, is more explicative than M_{71} model attained from the complete set of points for after removal of these points the percentage of soybean yield variation that can be explained by the regressors increased from 49% to 63% (Table 4). When considering $M_{71\text{-}\{10,15,23,29\}}$ model prepared without points 10, 15, 23 and 29 it was observed that the adjusted coefficient of determination (0.65) was higher than the adjusted coefficient of determination obtained from $M_{71\text{-}\{15,23,29\}}$ model, resulting in a more explanatory model. When comparing RMSE of $M_{71\text{-}\{15,23,29\}}$ and $M_{71\text{-}\{10,15,23,29\}}$ models (Table 4) it was emphasized that the identification of point 10 as influential and its withdrawal from the data set as a result corroborated reduction of this statistic, thus resulting in a more accurate model for making predictions. In view of these results it was decided to choose the $M_{71\text{-}\{10,15,23,29\}}$ model as the best model suited to soybean yield and determine bootstrap confidence intervals for the parameters associated with the explanatory variables (Table 5).

By comparing the confidence intervals of parameters of the $M_{71\text{-}\{10,15,23,29\}}$ model with the respective intervals obtained from the multiple linear regression model generated with all the explanatory variables and all sampling points (Table 1) it can be observed that regardless of the bootstrap method used the confidence intervals of parameters for the $M_{71\text{-}\{10,15,23,29\}}$ model had lower amplitude, indicating estimates of this model was more accurate.

Discussion

The average soybean productivity in the monitored area (4.305 t/ha) is considered high compared with other regions, according to data from CONAB (2015) in the agricultural year 2013/2014 average productivity in Brazil was 2.854 t/ha and in Paraná was 2.950 t/ha.

The negative sign of estimates for parameters associated with SRP_1, SRP_3, Des_1 and Des_2 variables (Eq. [3]) are expected once soil density (Des) shows a direct relationship with SRP (Busscher *et al.*, 1997), and as SRP has great influence on plant growth, root growth

Table 4. Parameters estimation and statistics for the multiple regression linear models considering the exclusion of influential points.

Models[2]	Parameters[1]							Statistics	
	Intercept	β_P	β_{Des_2}	β_{Ca}	β_{SRP_1}	β_{Mn}	β_{Mg}	R^2_{Adj}	RMSE
$M_{71\text{-}\{15,23,29\}}$	7.453	-0.080	-1.335	0.405	-0.131	-0.013	-0.608	0.63	0.24
$M_{71\text{-}\{10,15,23,29\}}$	7.971	-0.080	-1.618	0.414	-0.169	-0.012	-0.635	0.65	0.23

[1]β_i: parameter associated with the variable i = {P, Des_2, Ca, SRP_1,Mn, Mg}; P: phosphorus, mg/dm^3; Des_2: soil density, g/cm^3, from 0.1 to 0.2 m depth; Ca: calcium, cmol$_c$/dm^3; SRP_1: soil penetration resistance, MPa, from 0 to 0.1 m depth; Mn: manganese, mg/dm^3; Mg: magnesium, cmol$_c$/dm^3; [2] $M_{71\text{-}\{15,23,29\}}$: adjusted model to the dataset without the points (15,23,29); $M_{71\text{-}\{10,15,23,29\}}$: adjusted model to the dataset without the points (10,15,23,29); R^2_{Adj}: adjusted coefficient of determination; RMSE: root mean square error.

Table 5. Nonparametric bootstrap 95% confidence intervals for the parameters of the $M_{71\{10,15,23,29\}}$ model.

Parameters[1]	Efron's percentile			BC[3]		
	$\hat{\theta}_i$	$\hat{\theta}_u$	Amplitude[2]	$\hat{\theta}_i$	$\hat{\theta}_u$	Amplitude
β_P	-0.119	-0.053	0.065	-0.122	-0.055	0.067
β_{Des_2}	-2.461	0.230	2.691	-2.580	-0.073	2.508
β_{Ca}	0.117	0.597	0.480	0.141	0.613	0.472
β_{SRP_1}	-0.326	-0.011	0.316	-0.319	0.016	0.335
β_{Mn}	-0.021	0.000	0.021	-0.021	0.000	0.021
β_{Mg}	-2.461	0.230	2.691	-2.580	-0.073	2.508
Intercept	5.088	9.601	4.512	5.149	9.727	4.579

[1]β_i: parameter associated with the variable i = {P, Des_2, Ca, SRP_1, Mn, Mg}; P: phosphorus, mg/dm^3; Des_2: soil density, g/cm^3, from 0.1 to 0.2 m depth; Ca: calcium, cmol$_c$/dm^3; SRP_1: soil penetration resistance, MPa, from 0 to 0.1 m depth; Mn: manganese, mg/dm^3; Mg: magnesium, cmol$_c$/dm^3; [2]Amplitude: $\hat{\theta}_u - \hat{\theta}_l$; $\hat{\theta}_l$: lower limit; $\hat{\theta}_u$:upper limit. [3]BC: bias corrected.

and crop yields vary inversely proportional to its value (Freddi *et al.*, 2006). The parameters estimation associated to SRP_2 and Des_3 variables show opposite signals from the expected scenario; however, since it is verified that multicollinearity was non-existent, it is also prudent to investigate the significance of such variables. The positive estimate signal from the associated parameter to K variable is expected, once and in accordance with Pettigrew (2008), potassium is one of major nutrients considered essential for crop growth and yield development.

The comparison of confidence intervals can be done in terms of their amplitudes according to Paes (1998), to whom a high amplitude interval indicates reduced accuracy of estimation as compared with a range of lower amplitude, thus comparing the two techniques bootstrap confidence intervals (Table 1) it is clear the intervals obtained by Efron percentile technique showed lower amplitude and therefore is the most accurate. Given the fact that zero is present in most of the confidence intervals (Table 1) it is prudent to investigate whether there are irrelevant variables and/or influential points in the data set for they cause an increase in the parameter variance (Rao, 1971; Meloun & Militký, 2001) and as a result confidence intervals tend to have a greater range and loss of accuracy.

The fact that predictor variable P is selected in a large share of models (Table 2) and that its signs of the estimated parameters are negative in all of them can be explain by the high phosphorus values found (on average 12 mg/dm^3) which according to Popp *et al.* (2002), may indirectly decrease yields due to micronutrients imbalance. The high percentage (94%) of times when the sign of the estimated parameter associated with variable Ca is positive is also expected, because calcium deficiency is among the main factors that inhibit root growth as reported by Oliveira *et al.* (2009), especially in latosols. Such a deficiency would have the plant vulnerable to biotic, biological and nutritional stresses and consequently would lead to reduced productivity (Dourado Neto *et al.*, 2014). As SRP_1 and Des_2 variables are used to assess the state of soil compaction, their effect on soybean yield is the opposite, for plants exhibit alterations in depth, branch and distribution of roots in response to soil compaction (Rosolem *et al.*, 2002), which undermines the efficient use of nutrients and water and limits crop yield (Alakukku & Elomen, 1995).

The model selection method using bootstrap is effective in determining the significant variables resulting in a more parsimonious model. Although the model determined by this method (M_{71}) has been the same selected by the conventional method using Akaike, the application of this methodology serve to attest the model selected by the Akaike criterion is not superparameterized, which can occur when the amount of samples is small.

Analyzing Fig. 3a, the graph of the bootstrap distribution of the parameter estimates associated with the variable SRP_1, it is seen that points 15 and 10 are detected as influential. Point 15 has a negative influence (-3.7) and its removal reduces bootstrap distribution amplitude, a fact that occurs mainly due to a shift in the initial percentiles if one considers the empirical distribution formed with 3000 replicates, $P_5 = -0.373$, $P_{10} = -0.330$, $P_{16} = -0.302$ and considers the empirical distribution formed only by bootstrap replicates with bootstrap samples not containing point 15 (1124 samples), $P_5 = -0.336$, $P_{10} = -0.295$, $P_{16} = -0.270$. The influence of point 10 is positive (2.6). It is observed that when considering the bootstrap distribution formed with those bootstrap samples that do not contain point 10 (1039 samples) the values considered percentile decrease, causing distribution displacement and reduction of its range from 0.865 to 0.727.

JaB graph in Fig. 3b for Ca variable indicated point 23 as negative influence (-2.8) and point 15 as positive influence (3.5), thus withdrawal of these points also causes changes in the empirical distribution of bootstrap estimates. After disregarding bootstrap replicates obtained from bootstrap samples which had point 23, the initial percentiles increased and the distribution range went from 1.214 to 0.999; and after disregarding those replicas obtained from samples containing point 15, there was a reduction in values of final percentiles, which also reduces the amplitude of the empirical distribution. Analyses of other graphs (Figs. 3c through 3f) are similar and indicate point 15 has a negative influence on bootstrap distributions of the parameters associated with the Mg, Mn and Des_2 variables; it also indicates point 23 has a positive influence on bootstrap distributions of the parameters associated with Mg and Mn variables and a negative influence on bootstrap distribution of the parameter associated with variable P. Point 29 has a positive influence on bootstrap distribution of the parameter associated with variable P and point 10 has a positive influence on bootstrap distribution of the parameter associated with variable Des_2.

Comparing all the points that were detected as influential in JaB graphs (Fig. 3), it is clear the sampling member 15 stands out due to its influence on most bootstrap distributions of estimated parameters, only the distribution of the parameter associated with the variable P is not influenced by excluding this element. The sample elements 23 and 10 also stood out as influential on various confidence intervals and the sample element 29 is the least influential of the four. By taking sample element 29 out of the empirical distribu-

tion of the bootstrap replicates obtained from bootstrap samples only the bootstrap distribution of parameter estimates associated with the variable P is influenced, seeing its range reduced from 0.163 to 0.112.

Regarding diagnostic analysis it is seen that the influential points determination method using JaB methodology together with Cook's distance (Fig. 2) do not identify some points clearly highlighted as influential by traditional methods and JaB graphics. Thus, JaB graphics prove to be a great alternative to identify influential points, as well as identifying the influential points with greater accuracy compared to traditional analysis they also provide information on bootstrap distributions of parameter estimates, making it possible to see what happens to confidence intervals when the influential samples are excluded.

Regarding the significance of the explanatory variables in $M_{71-\{10,15,23,29\}}$ model it is observed that only the parameter associated with Mn variable showed confidence intervals containing zero in both bootstrap techniques (Table 5) giving signs it may be irrelevant. This suspicion can be ruled out once there is evidence that the parameter signal associated with this variable is negative. It is necessary to simply notice that the zero has appeared at the high end of intervals and to also notice that in the variable selection method (Table 2) variable Mn is selected in 75% of models (750 models) and has a negative sign in 94% of them (705 models). By comparing techniques of confidence interval determination used it is observed they presented a similar behavior, though Efron percentile method stands out for providing intervals of lower amplitude. Cunha & Colosimo (2003) also highlights Efron percentile method to determine confidence intervals for regression models with measurement errors, since according to the authors this method is evidenced by its greater simplicity with equal performance compared to the others.

It is noteworthy that the bootstrap methods are fundamental to obtaining a more descriptive and more accurate model, as aside from the model $M_{71-\{10,15,23,29\}}$ to furnish a higher percentage of explanation of the soybean productivity (65%) than the initial model in Eq. [3] (41%), it furnishes a lower RMSE, being more accurate. It is important to highlight the explanatory influence of the model $M_{71-\{10,15,23,29\}}$ to be under satisfactory terms taking into account to be built only by physical and chemical features of the soil. The soybean productivity percentage variation not covered by such model (35%) is due to variables not considered, for example, the agricultural-meteorological, since climate has a significant impact upon the growth and development of crops (Hoogenboom, 2000). It is valuable to note the non-inclusion of the agricultural-meteorological variables in this essay to be due to the fact of the limitation on the spatial representation of the results to be obtained from the collected data in weather stations. (Junges & Fontana, 2011).

The results showed that the bootstrap methods enabled us to select the physical and chemical soil properties, which were significant in the construction of the soybean yield regression model, construct the confidence intervals of the parameters and identify the points that had great influence on the estimated parameters.

References

Aiken LS, West SG, 1991. Multiple regression: Testing and interpreting interactions. Sage Publications, Thousand Oaks, CA, USA. 224 pp.

Akaike H, 1973. Information theory and an extension of the maximum likelihood principle. Proc. 2nd Int. Symp. on Information Theory; Petrov BN, Csaki F (eds.). pp: 267–281. Akadémia Kiado, Budapest.

Alakukku L, Elomen P, 1995. Long-term effects of a single compaction by heavy field traffic on yield and nitrogen uptake of annual crops. Soil Till Res 36(3-4): 141-152. http://dx.doi.org/10.1016/0167-1987(95)00503-X.

Al-Marshadi AH, 2011. New weighted information criteria to select the true regression model. Aust J Basic Appl Sci 3(3): 317-312.

Austin P, Tu J, 2004. Bootstrap methods for developing predictive models. Am Stat 58(2): 131–137. http://dx.doi.org/10.1198/0003130043277.

Beyaztas U, Alin A, 2013. Jackknife-after-bootstrap method for detection of influential observations in linear regression models. Commun Stat Simulat C 42(6): 1256-1267. http://dx.doi.org/10.1080/03610918.2012.661908.

Busscher WJ, Bauer PJ, Camp CR, Sojka RE, 1997. Correction of cone index water content differences in a coastal plain soil. Soil Till Res 43(3-4): 205-217. http://dx.doi.org/10.1016/S0167-1987(97)00015-9.

Chaves-Neto A, Faria, TMB, 2015. Bootstrap for order identification in Arma(p,q) structures. Ind J Manag Prod 6(1): 169-181. http://dx.doi.org/10.14807/ijmp.v6i1.244.

CONAB, 2015. Soja – Brasil: Série histórica de produtividade. http://www.conab.gov.br. [24 March 2015].

Cook RD, 1977. Detection of influential observation in linear regression. Technometrics 19(1): 15-18. http://dx.doi.org/10.1080/00401706.1977.10489493.

Cunha WJ, Colosimo EA, 2003. Intervalos de confiança bootstrap para modelos de regressão com erros de medida. Rev Mat Estat 21(2): 25-41.

Davison AC, Hinkley DV, 1997. Bootstrap methods and their application. Press syndicate of the University of Cambridge, Cambridge, UK. 582 pp. http://dx.doi.org/10.1017/CBO9780511802843.

Dourado Neto D, Dario GJA, Barbieri APP, Martin TN, 2014. Biostimulant action on agronomic efficiency of corn and common beans. Biosci J 30(1): 371-379.

Dubreuil S, Berveiller M, Petitjean F, Salaün M, 2014. Construction of bootstrap confidence intervals on sensitivity indices computed by polynomial chaos expansion. Reliab Eng Syst Safe 121: 263-275. http://dx.doi.org/10.1016/j.ress.2013.09.011.

Efron B, 1979. Bootstrap methods: Another look at the jackknife. Ann Stat 7(1): 1-26. http://dx.doi.org/10.1214/aos/1176344552.

Efron B, 1982. The jackknife, the bootstrap and other resampling plans. SIAM, Philadelphia, PA, USA. 93 pp. http://dx.doi.org/10.1137/1.9781611970319.

Efron B, 1992. Jackknife-after-bootstrap standard errors and influence functions. J R Stat Soc 54: 83-127.

Efron B, Tibsshirani R, 1986. Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. Stat Sci 1(1): 54-75. http://dx.doi.org/10.1214/ss/1177013815.

EMBRAPA, 2013. Sistema brasileiro de classificação de solos, 3ª Ed. – Centro Nacional de Pesquisa de Solos, EMBRAPA – SPI, Rio de Janeiro. 412 pp.

Freddi OS, Carvalho MP, Veronesi-Jr V, Carvalho GJ, 2006. Relationship between maize yield and soil mechanical resistance to penetration under conventional tillage. Eng Agric 26(1): 113-121.

Freedman DA, 1981. Bootstrapping regression models. Ann Statist 9(6): 1218-1228. http://dx.doi.org/10.1214/aos/1176345638.

Freud RJ, Littell RC, 2000. SAS system for regression, SAS Inst., Cary, NC, USA. 264 pp.

García-Gallego JM, Chamorro-Mera A, García-Galán MM, 2015. The region-of-origin effect in the purchase of wine: The moderating role of familiarity. Span J Agric Res 13(3): e0103. http://dx.doi.org/10.5424/sjar/2015133-7581.

Garcia-Paredes JD, Olson KR, Lang JM, 2000. Predicting corn and soybean productivity for Illinois soils. Agric Syst 64(3): 151-170. http://dx.doi.org/10.1016/S0308-521X(00)00020-2.

Hao L, Naiman DQ, 2010. Assessing inequality. Sage, Thousand Oaks, CA, USA. 149 pp. http://dx.doi.org/10.4135/9781412993890.

Hoerl R, Snee RD, 2012. Statistical thinking: Improving business performance. John Wiley & Sons, Hoboken, USA. 544 pp. http://dx.doi.org/10.1002/9781119202721.

Hoogenboom G, 2000. Contribution of agrometeorology to the simulation of crop production and its applications. Agric For Meteorol 103: 137-157. http://dx.doi.org/10.1016/S0168-1923(00)00108-8.

Ireland CR, 2010. Experimental statistics for agriculture and horticulture. Cambridge University Press, Cambridge, UK. 384 pp.

Junges AH, Fontana DC, 2011. Agrometeorological-spectral model to estimate wheat yield in the state of Rio Grande do Sul, Brazil. Rev Ceres 58(1): 9-16. http://dx.doi.org/10.1590/S0034-737X2011000100002.

Kamo K, Yanagihara H, Satoh K, 2013. Bias-corrected AIC for selecting variables in poisson regression models. Commun Stat A – Theory 42(11): 1911-1921.

Khakural BR, Robert PC, Huggins DR, 1999. Variability of corn/soybean yield and soil/landscape properties across a southwestern Minnesota landscape. In: Precision Agriculture; Robert PC, Rust RH, Larson WE (eds.). pp: 573-579. Am. Soc. Agron., Madison, WI, USA.

Kulcheski FR, Molina LG, Fonseca GC, Morais GL, Oliveira LFV, Margis R, 2016. Novel and conserved microRNAs in soybean floral whorls. Gene 575(2): 213-223. http://dx.doi.org/10.1016/j.gene.2015.08.061.

Levy P, Lemeshow S, 1980. Sampling for health professionals. LLP, Belmont, CA, USA. 320 pp.

Lobell DB, Ortiz-Monasterio I, Asner GP, Naylor RL, Falcon WP, 2005. Combining field surveys, remote sensing, and regression trees to understand yield variations in an irrigated wheat landscape. Agron J 97: 241-249.

Losada B, Blas C, García-Rebollar P, Cachaldora P, Méndez J, Ibáñez M, 2015. Short communication: Prediction of apparent metabolisable energy content of cereal grains and by-products for poultry from its chemical composition. Span J Agric Res 13(2):06SC02. http://dx.doi.org/10.5424/sjar/2015132-6573.

Martin MA, Roberts S, 2010. Jackknife-after-bootstrap regression influence diagnostics. J Nonparametric Stat 22(2): 257-269. http://dx.doi.org/10.1080/10485250903287906.

Meloun M, Militký J, 2001. Detection of single influential points in OLS regression model building. Anal Chim Acta 439(2): 169-191. http://dx.doi.org/10.1016/S0003-2670(01)01040-6.

Mercante E, Lamparelli RAC, Uribe-Opazo MA, Rocha JV, 2010. Linear regression models to soybean yield estimate in the west region of the state of Paraná, Brazil, using spectral data. Eng Agríc 30(3): 504-517.

Oliveira IP, Costa KAP, Faquin V, Maciel GA, Neves BP, Machado EL, 2009. Effects of calcium sources on Grass growth in monoculture and intercropping. Ciênc Agrotec 33: 592-598. http://dx.doi.org/10.1590/S1413-70542009000200036.

Paes AT, 1998. Essential items in biostatistics. Arq Bras Cardiol 71(4): 575-580. http://dx.doi.org/10.1590/S0066-782X1998001000003.

Penalba OC, Bettolli ML, Vargas WM, 2007. The impact of climate variability on soybean yields in Argentina. Multivariate regression. Meteorol Appl 14: 3-14. http://dx.doi.org/10.1002/met.1.

Peng RD, 2008. Simpleboot: Simple bootstrap routines. R package version 1.1-3.

Pettigrew WT, 2008. Potassium influences on yield and quality production for maize, wheath, soybean and cotton. Physiol Plant 133: 670-681. http://dx.doi.org/10.1111/j.1399-3054.2008.01073.x.

Popp JS, Griffin TW, Popp MP, Baker WH, 2002. Profitability of variable rate phosphorus in a two crop rotation. J Ark cad Sci 56: 125-133.

R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rahman MS, 2014. Coefficient estimation of regression model and hypothesis testing by bootstrap method. Res Rew J Stat 3(2): 1-7.

Rao P, 1971. Some notes on misspecification in multiple regressions. Am Stat 25(5): 37-39.

Rizopoulos D, 2009. BootStepAIC: Bootstrap stepAIC. R package version 1.2-0.

Rosolem CA, Foloni JSS, Tiritan CS, 2002. Root growth and nutrient accumulation in cover crops as affected by soil compaction. Soil Till Res 65:109-115. http://dx.doi.org/10.1016/S0167-1987(01)00286-0.

Sabaghnia N, Dehghani H, Alizadeh B, Mohghaddam M, 2010. Interrelationships between seed yield and 20 related traits of 49 canola (Brassica napus L.) genotypes in non-stressed and water-stressed environments. Span J Agric Res 8(2): 356-370. http://dx.doi.org/10.5424/sjar/2010082-1195.

Shasha D, Wilson M, 2011. Statistic is easy. Morgan & Claypool Publishers, San Rafael, CA, USA. 162 pp.

Siegel S, 1956. Nonparametric statistics for the behavioral sciences. McGraw-Hill, New York. 312 pp.

Sutton NJ, Cho S, Armsworth PR, 2016. A reliance on agricultural land values in conservation planning alters the spatial distribution of priorities and overestimates the acquisition costs of protected areas. Biol Cons 194: 2-10. http://dx.doi.org/10.1016/j.biocon.2015.11.021.

Tao F, Yokozawa M, Liu J, Zhang Z, 2008. Climate-crop yield relationships at provincial scales in China and the impacts of recent climate trends. Clim Res 38: 83-94. http://dx.doi.org/10.3354/cr00771.

Vera-Diaz MC, Kaufmann RK, Nepstad DC, Schlesinger P, 2008. An interdisciplinary model of soybean yield in the Amazon Basin: The climatic, edaphic, and economic determinants. Ecol Econ 65(2): 420-431. http://dx.doi.org/10.1016/j.ecolecon.2007.07.015.

Zheng H, Chen L, Han X, Zhao X, Ma Y, 2009. Classification and regression tree (CART) for analysis of soybean yield variability among fields in northeast China: The importance of phosphorus application rates under drought conditions. Agric Ecosyst Environ 132: 98-105. http://dx.doi.org/10.1016/j.agee.2009.03.004.