



Estimation of cropland prices in Rio Grande do Sul by multiple linear regression and principal component analysis

Cristiano Ziegler¹  Tobias da Paixão Fiegenbaum^{2*}  Renan Mitsuo Ueda¹ 
Valentina Wolff Lirio³  Adriano Mendonça Souza³ 

¹Departamento de Engenharia de Produção, Universidade Federal de Santa Maria (UFSM), Santa Maria, RS, Brasil.

²Departamento de Eletrônica e Computação, Universidade Federal de Santa Maria (UFSM), 97105-900, Santa Maria, RS, Brasil. E-mail: tobiasdpf@gmail.com. *Corresponding author.

³Departamento de Estatística, Universidade Federal de Santa Maria (UFSM), Santa Maria, RS, Brasil.

ABSTRACT: *This study aimed to price croplands in Rio Grande do Sul State (southern Brazil) and point which variables had the most significant impact on prices. The main purpose was achieved using multiple linear regression and principal component analysis. The variables used in this study were planted area, production, price, and yield of the commodities soybean, wheat, and corn. The period under analysis was from January 1994 to December 2017 (biannual observations). Multiple linear regression showed that five variables contributed to land pricing, being three related to soybean and two to wheat. Multivariate analysis grouped the investigated variables into clusters and indicated their influence, in addition to providing information on land prices and reducing variable dimensionality from fourteen original variables to three principal components to be analyzed. The two analyses complemented each other so that the croplands' price was explained by three variables, in which two corroborated in constructing the pricing model for croplands.*

Key words: *linear regression, principal component analysis, cropland pricing.*

Estimação do preço das lavouras no Rio Grande do Sul por meio de regressão linear múltipla e análise de componentes principais

RESUMO: *Este estudo teve como objetivo a precificação de terra para lavouras no Rio Grande do Sul e apresentar quais variáveis possuem maior impacto no preço. O objetivo foi alcançado por meio da aplicação da análise de regressão linear múltipla e de componentes principais. Variáveis relacionadas às commodities soja, trigo e milho, como a área plantada, produção, cotação e rendimento, formaram o banco amostral para as duas metodologias, compreendendo o período de janeiro de 1994 a dezembro de 2017, em observações bianuais. A regressão linear múltipla mostrou que três variáveis relacionadas à soja e duas ao trigo contribuem na precificação das terras. A análise multivariada agrupou as variáveis investigadas, indicando a influência entre as mesmas, fornecendo informações sobre o preço de terras e diminuindo a dimensionalidade do problema de 14 variáveis originais para três componentes a serem analisados. As duas análises se complementaram de forma que o preço de terras foi explicado por três variáveis e duas corroboraram na construção do modelo de precificação das lavouras.*

Palavras-chave: *regressão linear, análise de componentes principais, precificação de terra.*

INTRODUCTION

The rise in population requires increased food production and increasingly larger cropland areas (SILVEIRA et al.2017). Land is a useful speculation and value stockpile object (TELLES et al.2016); commodities such as soybean, corn, and wheat regulate cropland prices (HOLLAND et al., 2016), as well as, the production and physical infrastructural aspects (REYDON et al., 2014).

In southern Brazil, Rio Grande do Sul State (RS) has a vital role in the national food supply as it produces a wide variety of agricultural goods (MARCHESAN & SOUZA, 2010), with soybean being one of the main crops. What is more, the

growth of research, mechanization, and soybean use for animal feed has brought forth a booming environment for soybean in the global market (BONATO & BONATO, 1987). This crop is a prime ingredient in the food, swine, and poultry feed industries and biodiesel and oil production (HAAS et al., 2006). Wheat integrates the main grain production in RS and becomes flour, which is used in bread, doughs, and biscuits (MILOCA et al., 2007). Corn is an important commodity in the Brazilian economy for export and internal consumption (CALDARELLI & BACCHI, 2012). Moreover, Brazil produces significant quantities of corn, with RS being one of the leading corn producers of the country (VIEIRA et al., 2001).

The main contribution of this study lies in explaining and understanding how the price of land is influenced by the primary agricultural commodities produced in RS while considering the importance of the grain sector of this state. This study aimed to price croplands in RS utilizing the retail price of the main grains produced and land area planted, their production and yield by multiple linear regression and principal component analyses, and determine which variables have the greatest impact on cropland prices.

The use of statistical methods represents a way to measure how the variables under study affect the final price of crops. Moreover, the results reported herein provide information that can impact researchers, rural producers, and investors, given the magnitude of the topic, thus enabling interested parties to choose the most favorable moments to invest in purchasing agricultural land given the understanding of the factors that affect its price.

MATERIALS AND METHODS

Data

Variables that affect cropland prices were selected according to literature data (HOLLAND et al., 2016). Fourteen independent variables comprised the study: soybean planted area (ha), soybean yield (kg/ha), soybean production (ton), wheat planted area (ha), wheat yield (kg/ha), wheat production (ton), corn planted area (ha), corn yield (kg/ha), corn production (ton), dollar quota (R\$), soybean quota (R\$), wheat quota (R\$), and corn quota (R\$).

The sample period was from June 1994 to December 2017 (biannual observations). The retail price of land and crops was collected from AGRIANUAL (2017, 2019), ANUALPEC (2017), and Instituto Brasileiro de Economia da Fundação Getúlio Vargas (IBRE-FGV, 2014). Data regarding the planted area, production, and average yield of soybean, wheat, and corn were acquired from EMATER-RS (www.emater.tche.br); the dollar quotation was obtained from Instituto de Pesquisa Econômica Aplicada (IPEA) (www.ipeadata.gov.br); and soybean, wheat, and corn quota were collected from the New York Stock Exchange (br.investing.com).

Data of the prices of land and crops were not collected from a single source to make it possible to work with a larger and more up-to-date database, improving the effectiveness of the applied statistical

methods. For example, if a single source were used (e.g., FGV data), the data series would have ended in 2014. In the case of selecting only the Agrianual and Annual Pec data, the data would correspond to the 2013-2017 period. Both historical datasets were prepared by market professionals of each institution and provided similar market values.

All variables used were transformed into semiannual data. Although the variables have different measurement units, this was not a problem in relation to the statistical methods used (i.e., multiple linear regression and principal component analysis).

Methodological steps

The principal component analysis (PCA) was used to reduce the problem dimensionality. To avoid magnitude problems caused by the units of measurements, the variables will be standardized. The correlation matrix was used to obtain the eigenvalues and eigenvectors and the maximum variance explained by each eigenvector; only the eigenvectors greater than 1.00 were chosen, as described by VICINI et al. (2018). Upon selecting the principal components (PC), the reduction of dimensionality was achieved, and after selecting these components, the correlation of the PC with the original variables was performed, and only variables with significant correlation were selected. The PCA will be used to determine which variables have a major impact on price compositions.

Furthermore, multiple linear regression was applied to find regressors or explanatory elements for the response variable “cropland prices in Rio Grande do Sul.” Cropland prices were then associated with a minimum number of possible variables to anticipate the trend and real estate speculation of the objects’ value in the coming years. A test of significance was done for the regression coefficients through analysis of variance (ANOVA), and a significant regression model of independent variables was obtained (MILOCA et al., 2007). Afterward, the model was investigated regarding multicollinearity problems with variance inflation factor (VIF), and values exceeding 5 or 10 indicated strong multicollinearity (GÓMEZ et al., 2016). The best model was chosen by VIF, Akaike information criterion (AIC) (AKAIKE, 1974), residual sum of squares, Bayesian information criterion (BIC) (SCHWARZ, 1978), and determination coefficient R^2 .

Both criteria AIC and BIC helped select the model with fewer parameters (BRANCO et al., 2020), and the determination coefficient R^2

indicated the presence of multicollinearity when its value was elevated.

The assumptions of homoscedasticity and normality of residuals must be met for model validation. Homoscedastic residuals with their fixed maximum variance guarantee a linear relationship between dependent and independent variables, and normality removes the possibility of distortion between relations and tests of significance (OSBORNE, 2002). White's test and the chi-square test were applied to test homoscedasticity and normality of residuals, respectively. The values of regression coefficients were then obtained through the minimum least-squares method (THOLON & QUEIROZ, 2009).

RESULTS AND DISCUSSION

The cropland prices were explained by the variables soybean planted area, soybean quota, soybean production, average wheat yield, and wheat production and are modeled through multiple regression (Table 1). The p-value of the chosen variables indicates their significance for the model, and < 10 VIF indicates that multicollinearity does not interfere with the model. As shown in table 1, the regression equation is:

$$CPC = -17090.1 + 0.0089 SPA + 2.4937 AWY + 3.9279 SQ + 0.00036 SP - 0.0033 WP$$

Cropland prices (CPC) were adjusted to Brazilian reais (R\$), and; therefore, the coefficients have units according to table 2. Soybean planted area (SPA): soybean planting is a profitable activity because it is one of the most exported grains by Brazil (FREITAS & MENDONÇA, 2016). When soybean planted areas expand, more lands are occupied for planting, diminishing its offer and leading to higher cropland prices. The coefficient 0.0089 R\$/ha implies that an expansion of 10,000 ha of the planted area adds R\$ 89.00 to cropland prices.

Average wheat yield (AWY): if a small land area produces the same amount as a bigger land area, it implies that the smaller land is more valuable due to its higher yield. In this scenario, the cropland price increases since its value depends on the expected value of its production (REYDON et al, 2014). The coefficient 2.4937 R\$*ha/kg expresses an impact of R\$ 2.49 on cropland prices when the wheat yield of a harvest is equal to 1 kg/ha.

Soybean quota (SQ): when the soybean quota is high, producers tend to have higher profits after the harvest and thereby purchase more land and increase their farming area, which raises the cropland price. Farmers of some RS regions negotiate land in soybean quantity (bushel/hectare) (INCRA, 2017), making a higher soybean quota directly affects cropland prices. The coefficient related to soybean quota means that 1 dollar per bushel increases the cropland prices by R\$ 3.93 in RS.

Soybean production (SP): a land with excellent soybean production can become profitable when used to plant soybean, thereby increasing its price (FREITAS & MENDONÇA, 2016). A prolific land prompts farmers to seek more plantable land area, thus increasing the demand for land and increases its price. Soybean production has the largest magnitude among the model's variables, which is reflected by its small coefficient: a raise of 100,000 t in soybean production adds R\$ 36.00 to cropland prices.

Wheat production (WP): since wheat consumed in Brazil is generally imported from Argentina (BRUM & MÜLLER, 2008), when there is greater production with a constant yield — the wheat planted area grows — there is too much wheat in the national market, reducing the selling price. This lowers the land price because the farmers earn less while investing more money to buy enough wheat to plant on the bigger planted area. The coefficient -0.0033 R\$/ton reflects the opposite trend, indicating

Table 1 - Parameters of the regression model.

	Coefficients	Std. error	95% Confidence interval		P-value	VIF
Constant	-17090.10	1503.51	-20124.30	-14055.90	< 0.001	
Soybean planted area	0.0089	< 0.001	0.0070	0.0108	< 0.001	4.6170
Average wheat yield	2.4937	0.7266	1.0273	3.9601	< 0.002	5.4170
Soybean quota	3.9279	0.8479	2.2167	5.6391	< 0.001	2.1790
Soybean production	0.0036e-1	< 0.001	1.2599e-5	0.0007	0.0426	4.0610
Wheat prod.	-0.0033	< 0.002	-0.0058	-0.0008	0.0108	6.9150

Table 2 - Cropland prices adjusted to Brazilian Reais (R\$).

	Coefficients	Units
Constant	-17090.10	R\$
Soybean planted area (SPA)	0.0089	R\$/ha
Average wheat yield (AWY)	2.4937	R\$*ha/kg
Soybean quota (SQt)	3.9279	R\$*bushel/cents US\$
Soybean production (SPd)	0.0036e-1	R\$/ton
Wheat production (WPd)	-0.0033	R\$/ton

that if 10,000 fewer wheat tons were produced than the previous year, then cropland prices would rise by R\$ 33.00.

To check the homoscedasticity of the residuals, White's test was carried out with a null hypothesis H_0 "homoscedasticity exists"; if the p-value was above 5% (0.212), then the hypothesis was not rejected, thus confirming that the residuals were homoscedastic. For the normality test, the null hypothesis "normality exists" was tested; if the p-value was greater than 5% (0.339), then the hypothesis was not rejected, confirming that the residuals are normal.

The PCA was estimated with normalized data using the correlation matrix to find eigenvalues and eigenvectors, and varimax normalized rotation was used to interpret the graph. The eigenvalues and the percentage of explained variance are listed in table 3. Three eigenvalues were selected (Table 3), with eigenvalue using the criteria of eigenvalue above 1 and their explained accumulated variance being equal to 86.69%. The fourteen original variables were reduced to three PC, so the analysis reduced data dimensionality and simplified the results (RAMSER et al., 2019).

The four distinct sets are presented in figure 1. The variables of the most significant

influence in the X-axis are soybean yield (SoyYld) and corn production (CornProd), followed by corn yield (CornYld) and soybean production (SoyProd); they explain 58.43% and affect cropland prices. The X-axis is named commodities' quantity. The variables in the Y-axis (WheatQuota, CornQuota, and SoyQuota) are related to crop quota and explain 16.29% of the variance. This means that the Y-axis is named Commodities' Quota. Overall, the quality of the land has the most effect on cropland prices, followed by commodity prices.

Figure 2 has three distinct sets with four points outside them. The variables that influenced the X-axis the most are the same as in figure 1. Corn production and soybean yield are the relevant factors in the PC, thus being named Commodities' Quantity. The set of soybean area (SoyPltArea) and cropland prices (CroplandPrice) represent the PC3 on the Y-axis, which makes it relate to Soybean/Cropland Price.

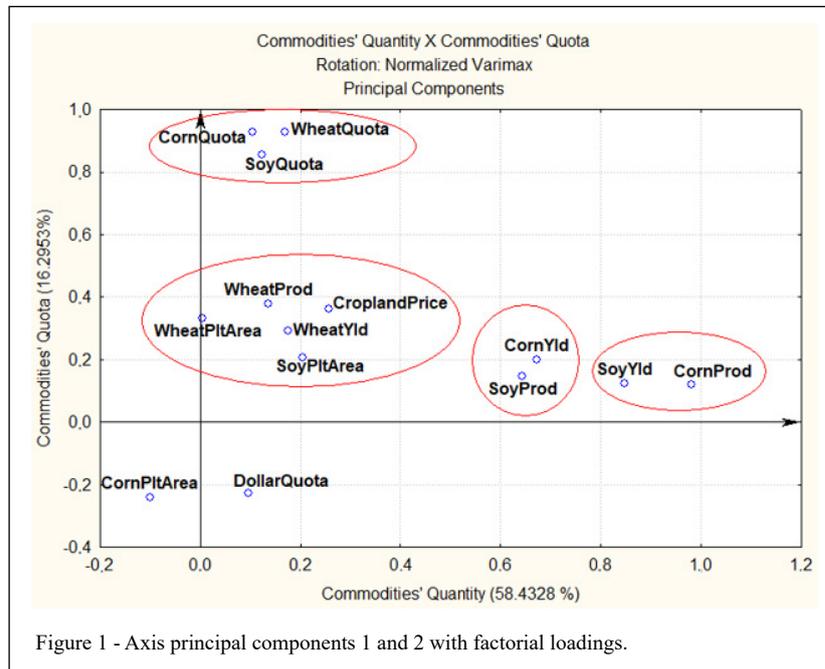
CONCLUSION

The principal component analysis reduced the dimension from fourteen variables to three principal components, namely Commodities Quantity, Commodities Quota, and Relation Soybean/Price,

Table 3 - Parameters of the regression model – principal components (PC).

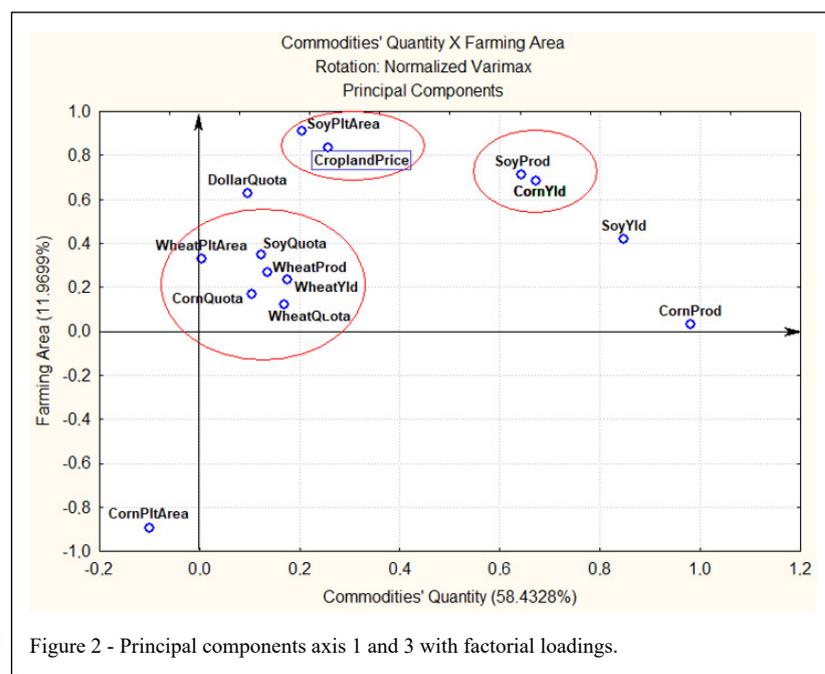
Principal components	Eigenvalues	Explained variance (%)	Accumulated eigenvalues	Accumulated variance (%)
PC1	8.1806	58.4328	8.1806	58.4328
PC2	2.2813	16.2953	10.4619	74.7281
PC3	1.6758	11.9699	12.1377	86.6981

Eigenvalues above 1 ($\lambda_i \geq 1$), percentage of the total variance (%), eigenvalue accumulation ($\sum \lambda_i$), and accumulated percentage (%).



which together accumulated 86.69% of the total variance. The clusters showed that the soybean area greatly influences cropland prices, and the principal components confirmed that the regression model has

relevant variables. Knowing which and how much these factors act on cropland prices in Rio Grande do Sul State is important to optimize planting and providing subsidies for speculation.



ACKNOWLEDGEMENTS

This study was carried out with the financial support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Financing Code 001. We would like to thank Laboratório de Análise e Modelagem Estatística (LAME) and the anonymous reviewers for their contributions to this paper.

DECLARATION OF CONFLICT OF INTEREST

The authors declare no conflict of interest. The founding sponsors had no role in the design of the study, collection, analyses, or interpretation of data, writing of the manuscript, and the decision to publish the results.

AUTHORS' CONTRIBUTIONS

All authors contributed equally to the conception and writing of the manuscript.

REFERENCES

- AGRIANUAL. **Anuário da Agricultura Brasileira**. São Paulo: IEG/FNP, 2017. p.432.
- AGRIANUAL. **Anuário da Agricultura Brasileira**. São Paulo: IEG/FNP, 2019. p.448.
- AKAIKE, H. A New Look at the Statistical Model Identification. **IEEE Transactions on Automatic Control**, v.19, n.6, p.716–723, 1974. Available from: <<https://ieeexplore.ieee.org/abstract/document/1100705>>. Accessed: Jun. 21, 2020. doi: 10.1109/tac.1974.1100705.
- ANUALPEC. **Anuário da Pecuária Brasileira**. São Paulo: IEG/FNP, 2017.
- BONATO, E. R.; BONATO, A. L. V. **A soja no Brasil: história e estatística**. Londrina, PR, Ed. EMBRAPA, 1987.
- BRANCO, A. N. et al. Analysis of the behavior of the supply of drugs for hypertension and diabetes in the state of Rio Grande do Sul between 2006 and 2017 using the ARIMA methodology. **Revista Gestão da Produção Operações e Sistemas**, [S.L.], v.15, n.2, p.91–110, jun. 2020. Available from: <<https://revista.feb.unesp.br/index.php/gepros/article/view/2435>>. Accessed: Jun. 21, 2020. doi: 10.15675/gepros.v15i2.2435.
- BRUM, A. L.; MÜLLER, P. K. A realidade da cadeia do trigo no Brasil: O elo produtores/cooperativas. **Revista de Economia e Sociologia Rural**, v.46, n.1, p.145–169, 2008. Available from: <https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-20032008000100007>. Accessed: Jun. 21, 2020. doi: 10.1590/S0103-20032008000100007.
- CALDARELLI, C. E.; BACCHI, M. R. P. Fatores de influência no preço do milho no Brasil. **Nova Economia**, jan-abril, 2012. Available from: <https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-63512012000100005>. Accessed: Jun. 21, 2020. doi: 10.1590/S0103-63512012000100005.
- EMATER/RS. **Informações Agropecuárias e Séries Históricas**. Disponível em: <http://www.emater.tche.br/site/info-agro/serie_historica.php#.XT7sqhKjDc> Accessed: Sept. 15, 2019.
- FREITAS, R. E.; MENDONÇA, M. A. A. Expansão Agrícola no Brasil e a Participação da Soja: 20 anos. **Revista de Economia e Sociologia Rural**, [S.L.], v.54, n.3, p.497–516, set. 2016. Available from: <https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-20032016000300497>. Accessed: Jun. 21, 2020. doi: 10.1590/1234-56781806-94790540306.
- GÓMEZ, R. S. et al. Collinearity diagnostic applied in ridge estimation through the variance inflation factor. **Journal Of Applied Statistics**, [S.L.], v.43, n.10, p.1831–1849, 25 fev. 2016. Available from: <<https://www.tandfonline.com/doi/abs/10.1080/02664763.2015.1120712>>. Accessed: Aug. 30, 2020. doi: 10.1080/02664763.2015.1120712.
- HAAS, M. J.; et al., A process model to estimate biodiesel production costs. **Bioresource Technology**, [S.L.], v.97, n.4, p.671–678, mar. 2006. Available from: <<https://www.sciencedirect.com/science/article/abs/pii/S0960852405001938>>. Accessed: Jun. 21, 2020. doi: 10.1016/j.biortech.2005.03.039.
- HOLLAND, T. G.; et al., Evolving frontier land markets and the opportunity cost of sparing forests in western Amazonia. **Land Use Policy**, v.58, p.456–471, 2016. Available from: <<https://www.sciencedirect.com/science/article/abs/pii/S0264837716308225>>. Accessed: Jun. 21, 2020. doi: 10.1016/j.landusepol.2016.08.015.
- IBRE - FGV. **FGVDados**. Available from: <<http://www14.fgv.br/fgvdados20/consulta.aspx>>. Accessed: Set. 15, 2019.
- INCRA. **Relatório de Análise de Mercado de Terras no Estado do Rio Grande do Sul - RAMT/RS**. Porto Alegre: INCRA, 2017.
- INVESTING. **Soja NY Futuros**. Available from: <<https://br.investing.com/commodities/us-soybeanbeans-historical-data>>. Accessed: Sept. 15, 2019.
- IPEA. **Instituto de Pesquisa Econômica Aplicada**. Available from: <<http://www.ipeadata.gov.br/Default.aspx>>. Accessed: Set 14, 2019.
- MARCHESAN, A; SOUZA, A. M. Forecasting the price of major grains produced in Rio Grande do Sul. **Ciência Rural**. v.40, n.11, p.2368–2374, 2010. Available from: <http://www.scielo.br/scielo.php?pid=S0103-84782010001100019&script=sci_abstract&tlng=pt>. Accessed: Oct. 16, 2020. doi: 10.1590/s0103-84782010001100019.
- MILOCA, S. A. et al. Relation between meteorological variables and the industrial quality of the wheat. **Ciência Rural**, Santa Maria, v.37, n.1, p.31–37, 2007. Available from: <https://www.scielo.br/scielo.php?pid=S0103-84782007000100006&script=sci_abstract&tlng=pt>. Accessed: Oct. 10, 2020. doi: 10.1590/s0103-84782007000100006.
- OSBORNE, W. J.; WATERS, E. Four assumptions of multiple regression that researchers should always test. **Practical Assessment, Research & Evaluation**, v.8, n.2, 2002. Available from: <<https://scholarworks.umass.edu/pare/vol8/iss1/2/>>. Accessed: Jun. 21, 2020. doi: 10.7275/r222-hv23.
- RAMSER, C. A. S. et al. The importance of principal components in studying mineral prices using vector autoregressive models: evidence from the Brazilian economy. **Resources Policy**, [S.L.], v.62, p.9–21, ago. 2019. Available from: <<https://www.sciencedirect.com/science/article/abs/pii/S0301420718305397>>. Accessed: Jun. 21, 2020. doi: 10.1016/j.resourpol.2019.03.001.

- REYDON, B. P. et al. Determination and forecast of agricultural land prices. *Nova Economia*, v.24, n.2, p.389-408, 2014. Available from: <https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-63512014000200389&lng=en&tlng=en>. Accessed: Mar. 18, 2020. doi: 10.1590/0103-6351/1304.
- SCHWARZ, G. Estimating the dimension of a model. *The Annals of Statistics*, v.6, n.2, p. 461-464, 1978. Available from: <<https://projecteuclid.org/euclid.aos/1176344136#info>>. Accessed: Nov. 05, 2020. doi: 10.1214/aos/1176344136.
- SILVEIRA, V. C. P.; GONZÁLEZ, J. A.; FONSECA, E. L. Land use changes after the period commodities rising price in the Rio Grande do Sul State, Brazil. *Ciência Rural*, Santa Maria, v.47, n.4, 2017. Available from: <https://www.scielo.br/scielo.php?script=sci_artext&pid=S0103-84782017000400931>. Accessed: Nov. 05, 2020. doi: 10.1590/0103-8478cr20160647.
- TELLES, T. S.; PALLUDETO, A. W. A.; REYDON, B. P. Price movement in the Brazilian land market (1994-2010): an analysis in the light of post-keynesian theory. *Revista de Economia Política*, [S.L.], v.36, n.1, p.109-129, 2016. Available from: <https://www.scielo.br/scielo.php?pid=S0101-31572016000100109&script=sci_abstract&tlng=pt>. Accessed: Oct. 20, 2020. doi: 10.1590/0101-31572016v36n01a07.
- THOLON, P.; QUEIROZ, S. A. Mathematic models applied to describe growth curves in poultry applied to animal breeding. *Ciência Rural*, Santa Maria, v.39, n.7, p.2261-2269, 2009. Available from: <https://www.scielo.br/scielo.php?script=sci_artext&pid=S0103-84782009000700050>. Accessed: Oct. 20, 2020. doi: 10.1590/s0103-84782009000700050.
- VICINI, L. et al. *Técnicas multivariadas exploratórias: teorias e aplicações no Software Statistica*. Santa Maria: Ed. UFSM, 2018.
- VIEIRA, R. C. M. T. et al. Cadeias produtivas no Brasil - Análise da competitividade. *Revista de Política Agrícola*. Ano X, n.4, out./nov./dez. 2001. Available from: <<https://seer.sede.embrapa.br/index.php/RPA/article/view/645>>. Accessed: Oct. 20, 2020.