University of São Paulo
"Luiz de Queiroz" College of Agriculture

Analysis of soybean crop data in the state of Mato Grosso, Brazil, in the period from 1990 to 2018

João Gabriel Ribeiro

Thesis presented to obtain the degree of Doctor in Science Area: Statistics and Agricultural Experimentation

Piracicaba
2021

João Gabriel Ribeiro
Licentiate in Mathematics

Analysis of soybean crop data in the state of Mato Grosso, Brazil, in the period from 1990 to 2018

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:

Profª.Drª: **Sônia Maria De Stefano Piedade**

Thesis presented to obtain the degree of Doctor in Science Area: Statistics and Agricultural Experimentation

Piracicaba
2021

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Ribeiro, João Gabriel
Analysis of soybean crop data in the state of Mato Grosso, Brazil, in the period from 1990 to 2018/ João Gabriel Ribeiro.- -versão revisada de acordo com a resolução CoPGr 6018 de 2011.- -Piracicaba, 2021.

103 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Soja 2. Imputação Univariada Múltipla 3. Análise de *Clusters* I. Título.

RESUMO

Análise dos dados da safra de soja no estado de Mato Grosso, Brasil, no período de 1990 a 2018

A produção de soja do Brasil possui um papel importante para o abastecimento dos mercados interno e externo. O Brasil ocupa o primeiro lugar na produção mundial de soja, impulsionado principalmente pelo estado de Mato Grosso, que lidera o complexo produtivo da soja no país. Neste contexto, foram coletados juntamente ao Instituto Brasileiro de Geografia e Estatistica (IBGE) os dados de produção de soja em grãos em mil toneladas, valor de produção de soja em grãos em mil reais e o valor de derivados de produção de soja em grãos em mil reais no período de 1990 a 2018, desse estado. Em seguida foram aplicados aos mesmos dados técnicas de imputação univariada via interpolação por *splines* cúbicas em dados faltantes de 46 municípios do estado, para as três variáveis, com a finalidade de completar este conjunto de dados, e revelar estimativas das mesmas variáveis. E por final ocorreram as aplicações de análises de agrupamentos (*clusters*), para os dados completos dos 141 municípios do estado durante 1990 a 2018, nas mesmas variáveis. E a partir dos 5, 5 e 4 grupos escalonados criados e validados estatisticamente para as variáveis de produção de soja em grão em mil toneladas, valor de produção de soja em grãos em mil reais e valor de derivados de produção de soja em grãos em mil reais, foi realizado um zoneamento da atividade da produtiva da soja no estado de Mato Grosso durante esse período, e este retrato produtivo da cultura pode ser uma contribuição para o estado na realização de políticas públicas desse segmento bem como um atrativo de investidores no estado interessados nesta cultura.

Palavras-chave: Produção; Imputação univariada múltipla; Análise de Aglomerados

ABSTRACT

Analysis of soybean crop data in the state of Mato Grosso from 1990 to 2018

Soybean production in Brazil plays a key role in supplying the domestic and foreign markets. Brazil ranks first in world soybean production, driven mainly by the state of Mato Grosso, which leads the soybean production complex in the country. In this context, data of soybean grain production in thousand tons, value of soybean grain production in thousand reais and the value of soybean derivatives in thousand reais for that state, in the period from 1990 to 2018, were collected from the Brazilian Institute of Geography and Statistics (IBGE). Then, univariate imputation techniques via cubic *spline* interpolation were applied to missing data from 46 municipalities in the state, for the three variables, in order to complete this data set, and reveal estimates for the same variables. Finally, there were applications of cluster analysis, for the complete data of the 141 municipalities in the state from 1990 to 2018, on the same variables. From the 5, 5 and 4 staggered groups created and statistically validated for the variables of soybean production in thousand tons, soybean production value in thousand reais and value of soybean production derivatives in thousand reais , a zoning of soybean production activity was generated out in the state of Mato Grosso during this period, and this productive overview of the crop can be a contribution to the state in developing public policies in this segment as well as an attraction of investors in the state interested in this culture.

Keywords: Production; Multiple univariate imputation; Cluster analysis

SUMMARY

1 INTRODUCTION

The soybean production chain in Brazil has a significant presence in the agroindustrial scenario of the country, closely linked to the foreign market through grain exports, and these are almost entirely destined for China. The main agglomerations specialized in the production of soybean and its derivatives in Brazil are located in the states of Goiás (GO), Mato Grosso (MT), Mato Grosso do Sul (MS), Paraná (PR) and Rio Grande do Sul (RS).

Based on data from IBGE (Brazilian Institute of Geography and Statistics), the state of Mato Grosso is the largest soybean producer in Brazil and accounts for 26.81% and 50.50% of the country's total production and Central-West region, respectively, in the year 2018. It is verified, in this period, that its estimated production reaches around 31,608,562 tons of grains, which represents a value of 29,976,533 thousand reais in 2018; in the same year, the state reached an amount of 20,189,266 thousand reais with the production of soybean derivatives. And, still in 2018, according to data from the MDIC (Ministry of Development, Industry and Foreign Trade) and SECEX (Secretariat of Foreign Trade), the soybean exported by the state of Mato Grosso came to represent 3.28% total of exports from Brazil.

Researchers such as CASTRO (2001), BATALHA and SILVA (2007), SAAB *et al.* (2009), HIRAKURI and LAZZAROTTO (2011), NAAS (2018) and TANCREDI *et al.* (2020) affirm the importance of Brazil and the state of Mato Grosso, and consequently its several producing municipalities, in the soybean production activity for the country's agribusiness. Based on this, the second chapter of this thesis revealed the history of the development of soybean crop in Brazil and in the world, as well as its evolutionary trajectory from 1990 to 2018, in the Center-West region, and in the state of Mato Grosso.

In a third chapter, and in possession of the data collected together with the IBGE on production of soybeans in thousand tons, production value of soybeans in thousand reais and soybean derivatives in thousand reais, referring to municipalities in the state of Mato Grosso from 1990 to 2018, there were missing data in the initial years in 46 municipalities in the state. Thus, the statistical technique of univariate imputation by interpolation by cubic splines was applied to each of these locations, for these three variables, in order to obtain estimates of complete data sets for the same period in all 141 municipalities in the state. Studies by DE BOOR (1978), GREEN and SILVERMAN (1993), RUGGIERO and LOPES (1997), KNOTT (2000), HASTIE *et al.* (2009) describe the cubic spline interpolation methodology (KOOPMAN *et al.*, 1999; FARIÑAS *et al.*, 2002; BALTAZAR and CLARIDGE, 2006; NADIR *et al.*, 2008; WONGSAI *et al.*, 2017; MORITZ and BARTZ-BEIELSTEIN, 2017; DEMIRHAN and RENWICK, 2018) and showed some applications of univariate imputations. The research by JUNNINEN *et al.* (2004) and NORAZIAN *et al.* (2008) compare simple univariate data imputation techniques related to air quality data, and the validation of the appropriate imputation method takes place through data simulations. KING *et al.* (2001) also make use of comparison imputation algorithms. MORITZ and BARTZ-BEIELSTEIN (2017) research indicates several types of univariate imputation for data in general. Twumasi-Ankrah *et al.* (2019) work demonstrates that MAR (Missing at Random) imputations, combined with interpolations, produce good results. This chapter of this thesis brings the univariate imputation by interpolation by cubic splines, in data linked to soybean production in 46 municipalities in the state of Mato Grosso, and the advantage of validating the imputed series by the Quenouille test that compares the functions of autocorrelation of the observed series with the observed series plus the imputed one, instead of using simulations as in other studies, and represents a gap in the current literature in data related to soybean crop.

In the fourth and last chapter, some authors were consulted, such as Everitt (1979), Johnson *et al.* (2002), Ferreira (2008), Everitt and Hothorn (2011) and Härdle and Simar (2015) among others, for the application of cluster analysis to the variables of soybean production in thousand tons, the value of soybean production in thousand reais and the value of soybeans in thousand reais in the state of Mato Grosso with each of the variables previously imputed data in 46 separate municipalities, thus composing 141 municipalities in the state between 1990 and 2018. Research by Broich and Palmer (1980) and Lee *et al.* (2008) use cluster analysis in soybean varieties. The research by Popović *et al.* (2011) uses this technique to build a cluster related to agribusiness in Serbia. The application of cluster analysis using the DTW distance and Ward method and validations by the cophenetic correlation and Pearson correlation test together with the Mantel test, revealed an estimated productive zoning of the productive economic activity of the soybean crop in the state, during this period, which represents yet another unprecedented contribution to scientific literature and can be a tool for generating public policies for the state and attracting investors.

1.1 References

Baltazar, J. C. and D. E. Claridge, 2006 Study of cubic splines and Fourier series as interpolation techniques for filling in short periods of missing building energy use and weather data. Journal of Solar Energy Engineering-Transactions of The ASME **128**: 226–230.

Batalha, M. O. and A. L. d. Silva, 2007 Gerenciamento de sistemas agroindustriais: definições e correntes metodológicas. Gestão agroindustrial **3**: 23–63.

Broich, S. L. and R. G. Palmer, 1980 A cluster analysis of wild and domesticated soybean phenotypes. Euphytica **29**: 23–32.

Castro, A. M. G. d. G., 2001 Prospecção de cadeias produtivas e gestão da informação. Transinformação **13**: 55–72.

De Boor, C., 1978 *A practical guide to splines*, volume 27. Springer-Verlag New York.

Demirhan, H. and Z. Renwick, 2018 Missing value imputation for short to mid-term horizontal solar irradiance data. Applied Energy **225**: 998–1012.

EVERITT, B. and T. HOTHORN, 2011 *An introduction to applied multivariate analysis with R*, volume 1. Springer Science & Business Media.

EVERITT, B. S., 1979 Unresolved problems in cluster analysis. Biometrics **35**: 169–181.

FARIÑAS, M. S., R. L. DE SOUSA, and R. C. SOUZA, 2002 Uma metodologia para a filtragem de séries temporais. Aplicação em séries de carga elétrica minuto a minuto. 34º.SBPO .

FERREIRA, D. F., 2008 *Estatística multivariada*, volume 1. EditoraaUfla.

GREEN, P. J. and B. W. SILVERMAN, 1993 *Nonparametric regression and generalized linear models: a roughness penalty approach*. Crc Press.

HASTIE, T., R. TIBSHIRANI, and J. FRIEDMAN, 2009 *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer series in statistics New York.

HIRAKURI, M. H. and J. J. LAZZAROTTO, 2011 Evolução e perspectivas de desempenho econômico associadas com a produção de soja nos contextos mundial e brasileiro. Londrina, PR: EMBRAPA pp. 1–47.

HÄRDLE, W. K. and L. SIMAR, 2015 *Applied multivariate statistical analysis*, volume 4. Springer-Verlag Berlin Heidelberg.

JOHNSON, R. A., D. W. W, and OTHERS., 2002 *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ.

JUNNINEN, H., H. NISKA, K. TUPPURAINEN, J. RUUSKANEN, and M. KOLEHMAINEN, 2004 Methods for imputation of missing values in air quality data sets. Atmospheric Environment **38**: 2895–2907.

KING, G., J. HONAKER, A. JOSEPH, and K. SCHEVE, 2001 Analyzing incomplete political science data: An alternative algorithm for multiple imputation. American political science review pp. 49–69.

KNOTT, G. D., 2000 *Interpolating cubic splines*, volume 18. Springer Science & Business Media.

KOOPMAN, S. J., N. SHEPHARD, and J. A. DOORNIK, 1999 Statistical algorithms for models in state space using SsfPack 2.2. The Econometrics Journal **2**: 107–160.

LEE, J. D., J. K. YU, Y. H. HWANG, S. BLAKE, Y. S. SO, G. J. LEE, H. T. NGUYEN, and J. G. S, 2008 Genetic diversity of wild soybean (Glycine soja Sieb. and Zucc.) accessions from South Korea and other countries. Crop Science **48**: 606–616.

Moritz, S. and T. Bartz-Beielstein, 2017 imputeTS: time series missing value imputation in R. R Journal. **9**: 207.

Naas, I. F., 2018 Introduçção ao Agronegócio, pp. 13–18 in *Engenharia de Produçção Aplicada ao Agronegócio (Vol. 1)*, edited by REIS, J. G. M. and P. L. O. C. NETO, Blucher: São Paulo.

Nadir, Z., N. Elfadhil, and F. Touati, 2008 Pathloss determination using Okumura-Hata model and spline interpolation for missing data for Oman. In *Proceedings of the world congress on Engineering*, volume 1, pp. 2–4, London, UK.

Norazian, M. N., Y. A. Shukri, R. N. Azam, and A. M. M. A. Bakri, 2008 Estimation of missing values in air pollution data using single imputation techniques. Science Society of Thailand **34**: 341–345.

Popović, B., R. Maletić, S. Ceranić, T. Paunović, and S. Janković-Šoja, 2011 Defining homogenous areas of Serbia based on development of SME in agribusiness using the cluster analysis. Technics technologies education management **6**: 811–818.

Ruggiero, M. A. G. and V. L. d. R. Lopes, 1997 *Cálculo numérico: aspectos teóricos e computacionais*. Makron Books do Brasil.

Saab, M. S. B. L., M. F. Neves, and L. D. G. Cláudio, 2009 O desafio da coordenação e seus impactos sobre a competitividade de cadeias e sistemas agroindustriais. Revista Brasileira de Zootecnia **38**: 412–422.

Tancredi, F. D., F. C. d. S. Silva, E. Matsuo, and S. T., 2020 Origem, distribuição geográfica e importância Econômica, pp. 14–24 in *Aplicações de técnicas biométricas no melhoramento genético da soja (Vol. 1)*, edited by Matsuo, E., C. D. Cruz, and T. Sediyama, Editora Mecenas: Londrina.

Twumasi-Ankrah, A. S., B. Odoi, A. P. W, and E. H. Gyamfi, 2019 Efficiency of imputation techniques im univariate time series. IJSET International Journal of Science, Environment and Technology **8**: 430–453.

Wongsai, N., S. Wongsai, and A. R. Huete, 2017 Annual seasonality extraction using the cubic spline function and decadal trend in temporal daytime MODIS LST data. Remote Sensing **9**: 1–17.

2 FINAL CONSIDERATIONS

The results of this work related to data imputation generate estimates for the variables linked to soybean production in the state of Mato Grosso, from the perspective of univariate imputation, therefore situations that are beyond the studied scenarios are a suggestion for future studies, as well as the investigations of other methods and methodologies for data imputation.

There are a number of missing data methods, both from the perspective of simple imputation and multiple imputation techniques, and the use of imputation should be done with great caution for each type of dataset and its different situations.

The results of the cluster analysis also generate estimates and help to portray the comportment of the soybean crop in the state of Mato Grosso, however there are also possibilities for the application of new analysis methods, as well as their validations, which may be addressed in future studies.