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.

DEDICATORY

Dedicated to my dear and beloved mother, **Marilda Aparecida Ribeiro**, for all the love and dedication from the womb to the present day, all her teachings are honored day after day.

Also, dedicated to my dear grandparents, grandma **Luzinete Cortez Balieiro** (*in memorian*) and grandpa **Ary Pedro Balieiro** (*in memorian*), for all the loving and joyful moments that we spent together, and all the teachings and values transmitted to me, that love lives and multiplies in my heart.

Dedicated to my uncle, **Wilson Carlos Ribeiro** (*in memorian*) and aunt **Sônia Marta Ribeiro Ferreira** (*in memorian*), who in many moments of my upbringing helped me and my mother, always with a lot of love and generosity.

ACKNOWLEDGMENTS

To God first, for having given me a lot of strength, health and joy to carry out this work. My aunt Marlene, who taught me to read near her sewing machine while my mother worked, to my uncle Aguinaldo, for all the love and help to this day, to my uncle Gilvan and aunt Fátima for their friendship and support, and a special thanks to my uncle Marcos and aunt Zoraíde for all the love and warmth and instruction to always choose the path of mental and spiritual balance, in these last 15 years of my journey through Brazil on business, I want to thank all my cousins in my family Ribeiro.

To my family Balieiro, for all their strength on the initial path, showing me, supporting and instructing my steps up to here, I love you very much: godmother (Bel) and godfather (Beto), uncle Ari and aunt Valéria, uncle Biel and aunt Marília and all my cousins and also Marinês, thanks for everything.

To all my great longtime friends, who always supported me, Rafa, Gui (big cousin), Otávio, Fabinho, Bruno, Saulo, Ricardo, Paulo, Rodrigo (Chelinha), Paulinho, Fred and Luciano (Korea).

To all doctoral and laboratory friends, who were part of this stage. Special thanks to Júlio (Julião), Welinton (Japa), Valdemiro (Mirão), Denise and Glória.

To the great fellow friends of the University of the State of Mato Grosso (UNEMAT) Sinop-MT campus, Giovane (and also Paula, his wife), Vlademir (and also Kátia, his wife) André (and also Karina, his wife) and Rubens da UFMT (and also Camila, his wife) for cheering for me and friendship. I couldn't also forget my friend Júlio, a professor at UNIFAP, for all his friendship and valuable advice and help in my professional journey since time where we lived in Roirama-RR. Special thanks to my great friend, Joaquim (and also Karina, his wife), a professor at UNEMAT on the campus of Nova Xavantina-MT, for friendship since the year 2000 in the first year of college until today.

I thank my girlfriend Daniele, for all the support and love and affection, in these difficult times and for believing in me and in my process.

To the Department of Exact Sciencies of ESALQ/USP, for all the lessons learned and countless contributions in my technical and personal training, all of whom were my professors and those who were not. I also thank all the employees, and especially the secretary of the program Solange Paes Sabadin, who from the very beginning when I was not even a student, helped me to understand all the bureaucratic procedures of ESALQ/USP, and for all the friendship, generosity and readiness in all moments of this walk.

Prof, PhD, Sônia Maria de Stefano Piedade, my dear advisor, for the affection, friendship, patience, understanding, motivation, generosity, readiness and commitment to me in this trajectory, believing in my work.

To the members of the qualification board, thesis defense committee, for having

accepted to participate in the evaluation of this work and for the valuable comments.

My psychotherapist Miriam Conceição de Oliveira, for all the help in my own understanding of the world in these 3 years.

Special thanks to the Coordination for the Improvement of Higher Education Personnel in Brazil (CAPES), for helping to support all of Brazil's graduate programs.

Finally, I would like to thank the UNEMAT campus in Sinop-MT and the government of the state of Mato Grosso, for the continued release of my obligations as a teacher to carry out this research.

SUMMARY

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

LIST OF FIGURES

LIST OF TABLES

LIST OF FRAMES

LIST OF ABBREVIATIONS AND ACRONYMS

GDP	Gross Domestic Product
IBGE	Brazilian Institute of Geography and Statistics
CEPEA	Center for Advanced Studies in Applied Economics
CNA	Brazilian Federation of Agriculture and Livestock
CONAB	National Supply Company
USDA	United States of Departament of Agriculture
MDIC	Ministry of Development, Industry and Foreign Trade
SECEX	Secretariat of Foreign Trade
FOB	Free On Board
US$ FOB	Values, in dollars, of a certain commodity, without considering the costs of insurance and free shipping
CPR	Rural Producer Note
GATT	General Agreement on Tariffs and Trade
EMBRAPA	Brazilian Agricultural Research Corporation
MCAR	Missing Completely at Random
MAR	Missing at Random
NMAR	No Missing at Random
DTW	Dinamic Time Warp

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 qual-

ity 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 HISTORY OF SOYBEAN PRODUCTION IN BRAZIL AND THE WORLD

Brazil, with its vast territory, has in agribusiness one of the most important and crucial pillars of its economy. In times of global crisis (health, food, sanitary and economic and others), this activity is necessary to supply countless nations, and has an important role in the balance of the country's trade balance.

In the last 40 years, the technological levels reached by Brazilian rural producers have reached expressive levels that can be measured by the increase in productivity in the field. This explains, for example, the fact that Brazil has managed to quintuple grain production to the current 260.08 millions tons, according to data from estimates by the National Supply Company (CONAB[1]) in 2020, in relation to the harvest of 50.8 million tons. tons obtained in the beginning of the 80's, with the same planted area. This performance in the field was only possible thanks to the use of inputs, basically seeds, fertilizers and pesticides, which are of the first line available for the sector.

The country has 388 million hectares of fertile agricultural land with high productivity potential, of which 90 million have not yet been exploited for commercial purposes, in addition to abundant solar energy, about 13% freshwater available in the world and diversified climate with regular rainfall distribution. This set of qualities, taken together, makes Brazil have a natural capacity for the development of agriculture and the businesses involved in its production chains. Agribusiness is now the major activity of the Brazilian economy and accounts for one in three reais generated in the country today.

Today agribusiness, understood as the sum of the productive sectors with those of processing of the final product and the manufacturing of inputs, accounts for almost a third of Brazil's GDP and by a similar amount of the country's total exports.

The agribusiness GDP (Gross Domestic Product) is of great importance for the Brazilian economy. Data from the IBGE[2] (Brazilian Institute of Geography and Statistics), CEPEA[3] (Center for Advanced Studies in Applied Economics) e CNA[4] (Brazilian Federation of Agriculture and Livestock) point out that, in 2019, the sector registered an increase of about 0.6% in relation to 2018, reaching a total of 21.4% total GDP, divided into approximate values, in agricultural and livestock inputs (5.1%), in agricultural and livestock production (23%), in agroindustry (30%) and in services (41.9%). The distribution of the agribusiness GDP in Brazil, in 2019 is presented in Figure 2.1.

[1] https://www.conab.gov.br/info-agro/safras/graos/boletim-da-safra-de-graos
[2] https://www.ibge.gov.br/estatisticas/economicas/contas-nacionais
[3] https://www.cepea.esalq.usp.br/br/pib-do-agronegocio-brasileiro.aspx
[4] https://www.cnabrasil.org.br/

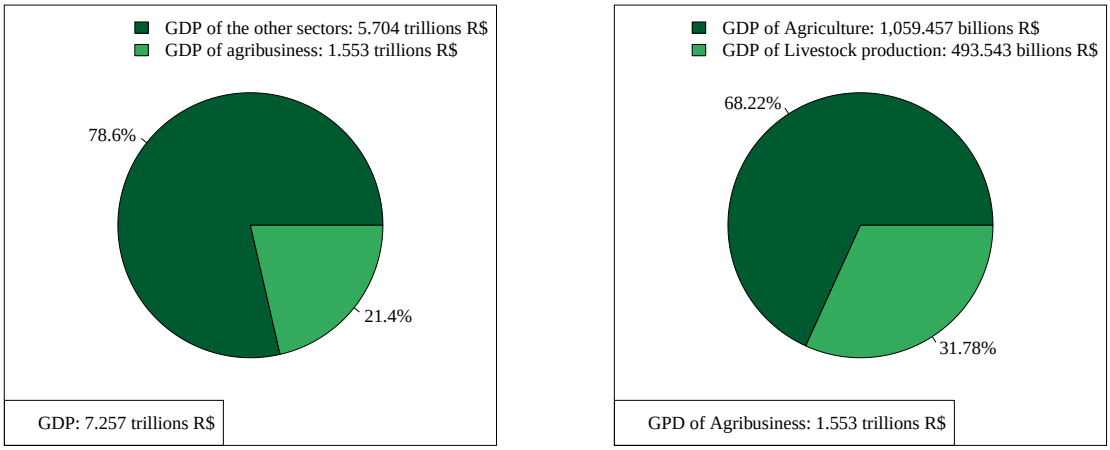


Figure 2.1: Gross Domestic Product (GDP) of agribusiness and the other sectors, in 2019.

Source: IBGE/CEPEA/CNA.

According to Roessing and Lazzarotto (2004), Brazilian agribusiness has been understood, in national and international environments, as one of the sectors with the greatest impact on the country's development. This is because it is the sector of the economy that, in addition to the greatest generation capacity of jobs, it is also the biggest source of stimuli for other activities. In the words of Rodrigues (2006), the Brazilian agribusiness sector is modern, efficient and competitive in international trade, as it has 22% of the arable land in the world, in addition to high technology used in the field.

The development of new technologies suitable for agricultural production, in tropical and subtropical climates, has led to a sharp growth in the Brazilian international trade, as well as scientific advances applied to improving domestic production. Agribusiness is presented as a way of looking at agriculture, as a system interconnected by several productive chains and their components and not as an isolated sector. It is a complex of actions that incorporate the production, processing, storage, distribution and commercialization of the products that were produced in the agricultural sector, until the arrival at the final consumer. Agribusiness is a larger system, which encompasses several other small and medium, interconnected systems (Castro, 2001; Batalha and Silva, 2007; Saab *et al.*, 2009; Naas, 2018).

The evolution of the GDP of Brazil and other sectors of the country is shown in Figure 2.2.

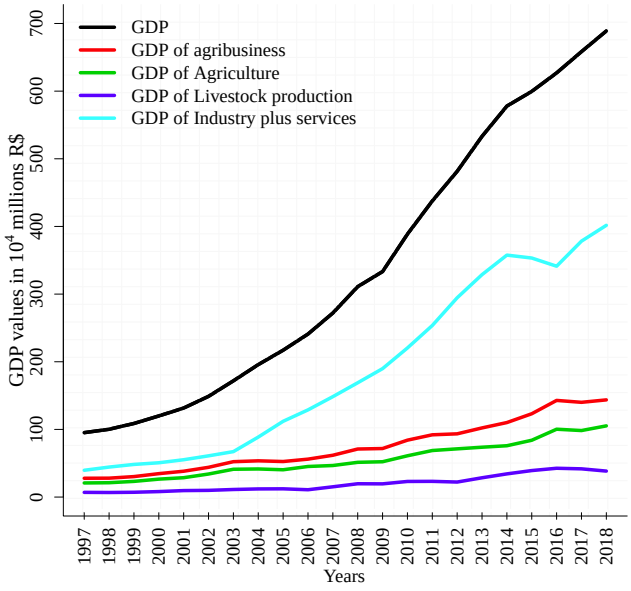


Figure 2.2: Gross Domestic Product (GDP) of Brazil and its sectors from 1997 to 2018.

Source: IBGE/CEPEA.

The agribusiness sector deserves prominence in the generation of wealth for Brazil, ranking third, behind only the Industry and Services sector in the movement of the country's GDP over the years 1997 to 2018.

The development of the Brazilian agribusiness enabled the evolution of agroindustrial concentration downstream and upstream of agricultural and livestock production, an important fact for improving global competitiveness and investments in quality for access to international markets. The sector is supplied mainly by the productive chains of cotton, beef cattle, dairy cattle, sugarcane and soybean, in the country, and these chains supply both the domestic and foreign markets, and they also have importance for job creation in the country.

The ascending comportment of these chains from 2001 to 2017 in Brazil is revealed in Figure 2.3.

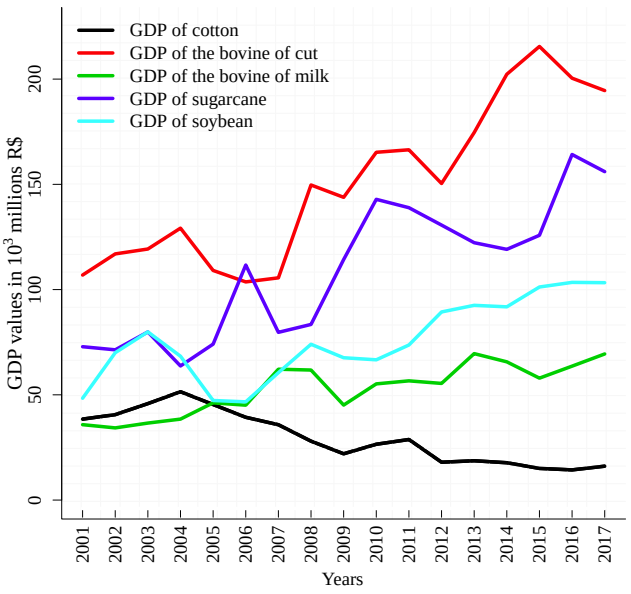


Figure 2.3: Gross Domestic Product (GDP) of Brazil and its sectors from 2001 to 2017.

Source: Barros *et al.* (2017)[5].

The cotton production chain had a decrease in its GDP over the years 2001 to 2017, while the other chains had a path of evolution in the same period.

Brazil is one of the world leaders in the production and export of various agricultural products. It occupies the first position in the production and export of coffee, sugar, alcohol and fruit juices. The country is also a leader in foreign sales of soybean, beef, chicken, tobacco, leather and leather shoes. Corn, rice, fresh fruit, cocoa, chestnuts and walnuts, in addition to pork and fisheries, are highlights of agribusiness in Brazil, which today employs around 20 million workers directly in the rural zone (Naas, 2018).

The origin center of the soybean crop is the Asian continent, more precisely, the region corresponding to Ancient China, in which soybean was the food base of the Chinese people for more than 5,000 years. As their domestication increases, their cultivation expands around the world. A little of the history of soybean cultivation in the world and in Brazil, is shown in Frames 2.1 and 2.2.

[5]https://www.cepea.esalq.usp.br/br/pib-de-cadeias-agropecuarias.aspx

Frame 2.1: Historical overview of soybean diffusion around the world.

Date	Event
2838 b.C	1st record at the Pen Ts'Ao Kang Mu Herbarium.
1500 a 1027 b.C	Cultivated in the Shang dynasty.
Centuries II b.C to III a.C	Introduced in Korea and Japan.
1739	1st experimental planting in Europe: (The Botanical Garden of Paris).
1790	Royal Botanical Garden in Kew (England).
1804	Pennsylvania-USA: Promising forage plant and grain producer.
1880	USA: grown as a forage plant.
1873	University of Vienna (Austria): Exhibits 19 varieties from Japan and China.
1876	Friedrich Haberlandt (Austria) sends seeds to Germany, Poland, Hungary, Switzerland the Netherlands.
1882	Introduced in Brazil (Bahia) by Gustavo D'Utra.
1909	Introduced in Argentina (Experimental Station of Córdoba).
1921	Introduced in Paraguay.
1928	Introduced in Colombia.

Source: Bonetti (1981), Bonato and Bonato (1987) and Tancredi *et al.* (2020).

Frame 2.2: Historical overview of soybean diffusion in Brazil.

Date	Event
1882	Introduced in the state of Bahia by Gustavo D'Utra.
1189	1st Reference written in Brazil: D'Utra G (1885) Culture of Chinese beans. Bulletin of the Agronomic Institute, Campinas 10 (3): 131-139.
1889	2nd Reference written in Brazil: D'Utra G (1889) New experimental soybean crop. Bulletin of the Agronomic Institute, Campinas 10(9/10):582-587.
1882	Introduction of cultivars in São Paulo - Agronomic Institute of Campinas (IAC) (Daffert).
1908	São Paulo: Japanese immigration reveals the use of soybean in human food.
1914	1st Introduction in the state of Rio Grande do Sul by Professor Craig: School of Agronomy and Veterinary at the Technical University (current UFRS).
1921	Henrique Lobbe: Experimental Field of Seeds of São Simão at the IAC (Introduction of 5 varieties from China).
1926	Henrique Lobbe (IAC): introduction of 48 cultivars from the USA.
1930	USA: beginning of soybean cultivation for grain production.

Frame 2.2: Continued.

Date	Event
1931-33	Studies on 23 varieties at the Sugarcane and Oilseed Experimental Station in Piracicaba, State of São Paulo.
1941	RS: soybean appears for the first time in the official statistics of that state.
1941	RS: the first soybean processing industry was built.
1945	São Paulo: soybean appears for the first time in the official statistics of that state.
1949	China: Mao Tse Tung Cultural Revolution (opening space in the world oilseed and protein market).
1949	Brazil (RS): 1st export = 18,000 tons of grain (appears for the first time in international statistics).
1954	State of Paraná: occurrence of frost in coffee crops that stimulates the production of soybeans in the summer.
1960-70	Big boost in production due to the soybean x wheat binomium.
1970-80	Expansion of soybean to Central Brazil (low latitude regions).
1980 Years	Japan: encourages its milling agribusiness and buys soybeans. EU: also encourage its milling agribusiness by buying grains on the international, market, in addition to subsidizing the production of other production of other "proteoleaginosas" species (peas, rapeseed and sunflower). Expansion of soybean crops to the west of the state of Bahia (Barreiras). Consolidation of Barreiras-BA as the main Northeastern pole of soybean production. Expansion of soybean crops to the north of Brazil, where the national regions of lower latitudes are found (MA, TO, PI, PA). The need to establish waterways and river ports, with warehouses was revealed, as part of the implementation of a logistics for the flow of crops produced in the Central-West and North regions of Brazil.
1990 Years	The practice of financing production based on green soybeans and CPR (Rural Producer Certificate) grows, based on attracting credit from multinational agricultural input companies, resellers companies, companies tradings. End of the decade: the area cultivated with soybean grows in the state of Roraima (Northern Hemisphere). End of the decade: trade round of GATT (General Agreement on Tariffs and Trade) in Uruguay: developing countries increase the resistance against subsidies applied to agricultural activities in the European Union.

Frame 2.2: Continued.

Date	Event
2001 to 2010	USA: - Reinforcement of agricultural subsidies policy. - Clean fuel policy (corn ethanol). - Origin of the world economic crisis that started in October 2008. China: - 1,400,000,000 Chinese (social inclusion of 50%). - Economic growth at annual rates from 8 to 12%. - The world's largest consumer of commodities. European Union: - Entry of new members from Eastern Europe. - Suffers drastic effects from the October 2008 global crisis. - Maintains the agricultural subsidies policy. Brazil: - Hurricane Catarina near the port of Paranaguá, state of Paraná. - 2003/04 and 2004/05: crisis in the grain sector.
	- 2004/05 to 2007/08: expansion of the sugar and alcohol sector. - 2005 to 2008: piggyback on the world's growth. - Great exporter of commodities.
Today	Brazil has a thriving agribusiness even with the world pandemic of COVID-19, and occupies the place of largest soybean producer and exporter of the , product to China, obtaining soybean superharvests year after year.

Source: Adapted from Bonetti (1981), Bonato and Bonato (1987), Câmara (2011) and Tancredi *et al.* (2020).

In the world agribusiness, soybean production is among the economic activities that, in the last decades, presented more expressive growth. And the main factors that made this development possible were: development and structuring of a solid international market related to the trade of products from the agroindustrial complex linked to soybeans; consolidation of this oilseed as an important source of vegetable protein, especially to meet the growing demands of sectors linked to the production of animal origin products; and generation and supply of technologies, which have propelled the rise of soybean exploitation to different regions of the world (Hirakuri and Lazzarotto, 2011).

The soybean agroindustrial complex has significant and growing importance for the Brazilian agriculture. It is one of the country's main crops in terms of volume and income generation, which has been playing a key role in the development of various regions of the country.

The distribution of production, pressing, import, export and final stock of the 2019 world soybean harvest is presented in Figures 2.4, 2.5 and 2.6.

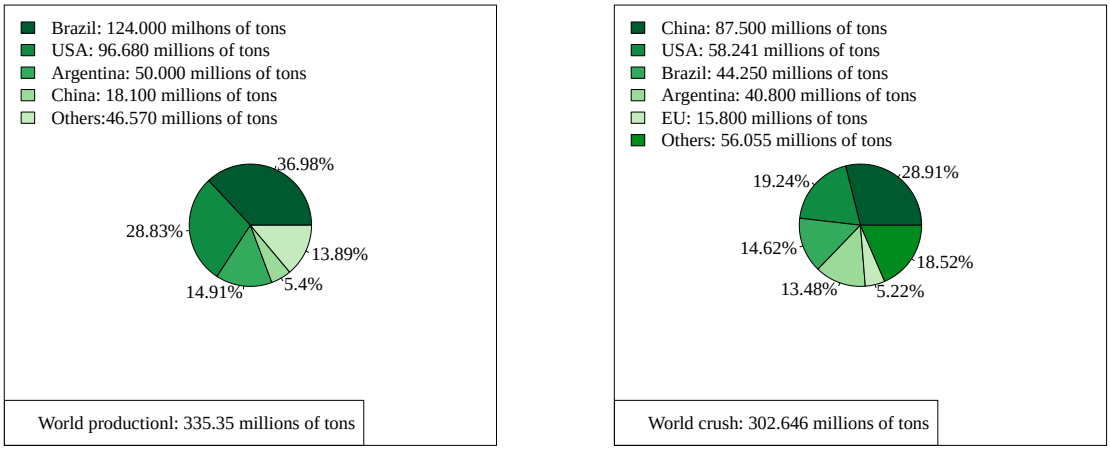


Figure 2.4: Distribution of soybean production and crush in the world in 2019.

Source: USDA (United States Departament of Agriculture).

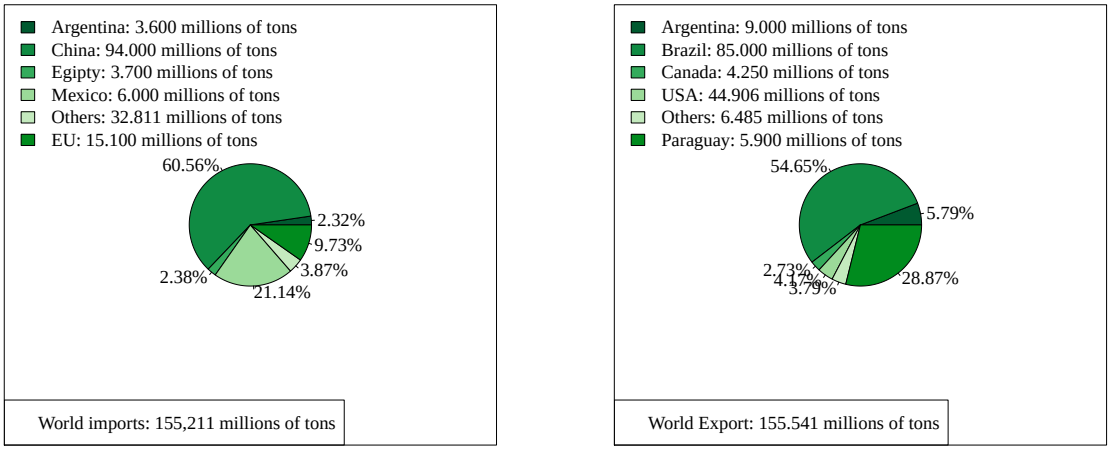


Figure 2.5: Distribution of soybean imports and exports in the world in 2019.

Source: USDA (United States Departament of Agriculture).

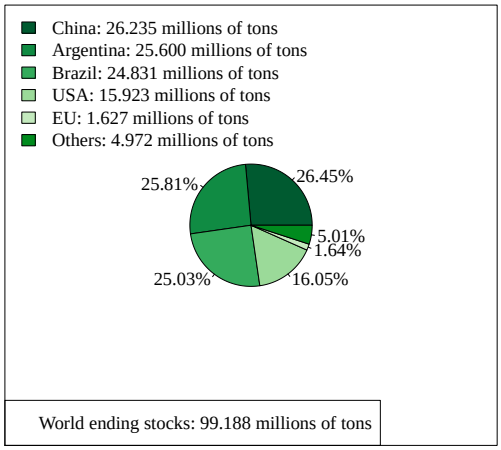


Figure 2.6: Distribution of ending stocks in the world in 2019.

Source: USDA (United States Departament of Agriculture)[6].

Figures 2.4, 2.5 and 2.6 illustrates the importance of the soybean agro-industrial production complex for the Brazilian economy, in which Brazil, the United States and Argentina 2019 produced 80.72% world production in 2019, with Brazil being the largest producer. The biggest beneficiary country of the world's soybean production is China, with 28.91% and Brazil ranks third with 14.62%. Thus, China also ranks first in world soybean imports. And with regard to exports, Brazil and the United States are responsible for 83.52% world exports, Brazil being the largest world exporter, with 54.65%. In relation to the product's final world stocks, China, Argentina and Brazil correspond to 77.29%, and Brazil occupies third place in this regard.

The progress in world demand for soybeans is due, in particular, to the fact that soybeans are being used in animal feed, mainly for poultry, cattle and pigs, due to their high protein content. Another reason that partly explains this evolution of production, exports and demand for soybeans around the world is mainly due to changes in the consumption habits of the populations of developing countries (Naas, 2018).

However, from the soybean meal used in animal feed, crude oil is obtained, and refined oil and lecithin from the grain are produced from it. The list of the main by-products from soybean processing is listed in Frame 2.3.

[6]https://usda.library.cornell.edu/concern/publications/6m311p28w?locale=en&page=3#release-items

Frame 2.3: Mains subproducts from soybean processing in grains.

Refined oil		**Soybean lecithin**	
Edible use	**Technical use**	**Edible use**	**Technical use**
Antibiotics	Desinfetantes	Medicinal Products	Cosmetics
Cooking oil	Glazing putty	Production of sweets	Alcohol production
Margarine	Thermal Insulation	Margarine production	Production of powdered metals
Pharmaceuticals	Waterproof cement	Bakery products	Paint manufacturing
Medicinal Products	Soap	Pharmaceuticals	Textiles
Spices	Plasticizers	Chocolates	Gasoline production
Vegetable fat	Insecticides	Fats	Pigments

Source: Soybean Embrapa[7].

In addition to moving billions of dollars in exports of soybean, meal and oil, the Brazilian soybean production generates 1.5 million jobs in 17 states of the country. For each job generated by soybean, the number rises to 12.66 workers, in the whole soybean production complex, and the sector drives the growth of sectors involved in the productive development of soybean crop through investments in technologies, research, and expansion of new agricultural and industrial areas for grain processing and oil refining have promoted positive results not only in volumes operated, but also in improving the quality of life of the population (Tancredi *et al.*, 2020).

There are several other uses for soybean, which can be consumed in the form of milk or added to fruit juices, for example. However, the most common derivatives are obtained from its oil, such as cooking oil, the salad dressings, vegetable shortenings, mayonnaise and margarines.

Another application that has also gained prominence in large countries producers is the use of soybean oil for the production of biodiesel. The renewable fuel that reduces the emission of polluting gases can be used pure or mixed with petroleum diesel in different proportions (Lemos *et al.*, 2017).

The comportment of Brazil's trade balance, as well as the evolutionary representation of soybean exports from the state of Mato Grosso from 1997 to 2018, can be found in Figure 2.7 and Table 2.1.

[7]https://www.embrapa.br/soja

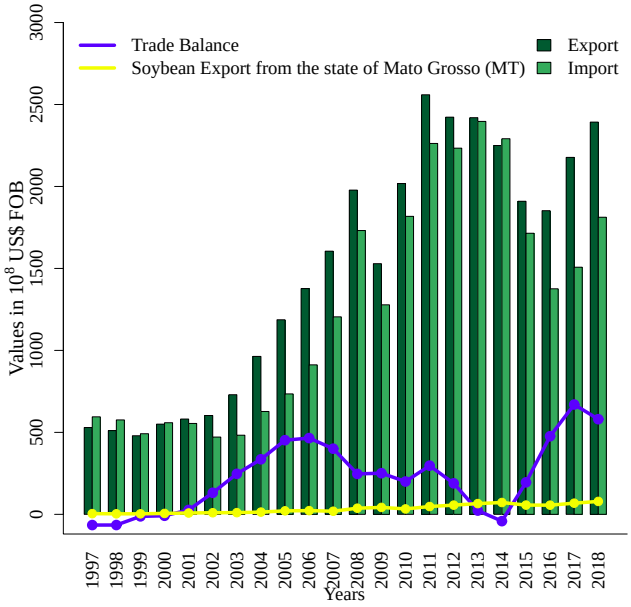


Figure 2.7: Brazil's trade balance from 1997 to 2018.

Source: MDIC (Ministry of Development, Industry, Foreign Trade)/SECEX (Secretariat of Foreign Trade)[8].

Table 2.1: Participation of the main export sectors of the Brazilian economy, from 2015 to 2018.

Sectors/Billions of US$	**2015**	**2016**	**2017**	**2018**
Soybean	27.958	25.422	31.722	40.704
	(14.64%)	(13.72%)	(14.57%)	(17.01%)
Petroleum and Derivatives	16.511	13.477	21.180	31.637
	(8.65%)	(7.28%)	(9.73%)	(13.22%)
Transport Material and Compo-	21.514	25.401	26.384	29.518
nents	(11.27%)	(13.71%)	(12.12%)	(12.34%)
Metallurgical Ore	16.654	15.816	22.397	23.670
	(8.72%)	(8.54%)	(10.29%)	(9.89%)
Metallurgical Products	13.409	11.608	14.631	15.893
	(7.02%)	(6.27%)	(6.72%)	(6.64%)
Soybean from MT	5.636	5.605	6.807	7.879
	(2.95%)	(3.03%)	(3.13%)	(3.29%)
Total	190.971	185.232	217.739	239.263
	(100.00%)	(100.00%)	(100.00%)	(100.00%)

Source: MDIC/SECEX.

Brazil leads both production and world exports, which generates a direct positive

[8]http://www.mdic.gov.br/index.php/comercio-exterior/estatisticas-de-comercio-exterior/series-historicas

impact on the country's trade balance of exports. In 2019, the entire complex involved with soybeans moved 40.91% country's total exports.

The sector ranks first in the participation of the country's exports, just behind the sectors of oil and derivatives, transport material and components, metallurgical ores and products, and meat. It is also worth noting that the state of Mato Grosso occupies a prominent position in the soybean productive sector, and presents an evolutionary path in exports, reaching 3.29% exports at the end of 2018. The evolution of monetary values referring to soybean exports and imports in Brazil, added to the share of soybean exports from the state of Mato Grosso, in this period, is represented by the Figure 2.8.

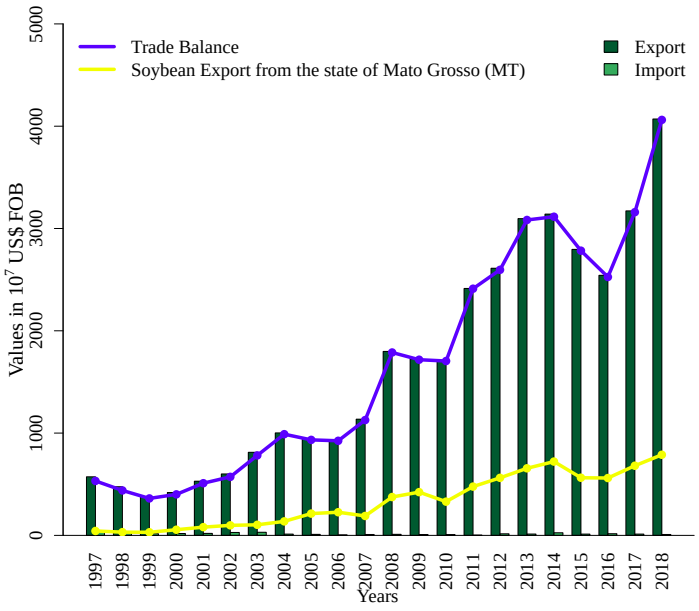


Figure 2.8: Soybean trade balance in Brazil from 1997 to 2018.

Source: MDIC/SECEX.

In the period from 1997 to 2018, soybean exports were greater than imports, and a large part of these exports came from the state of Mato Grosso. At the end of 2018, the state represented approximately 19.36% country's total soybean exports.

The trade balance of the state of Mato Grosso together with representation of soybean exports, is illustrated in Figure 2.9.

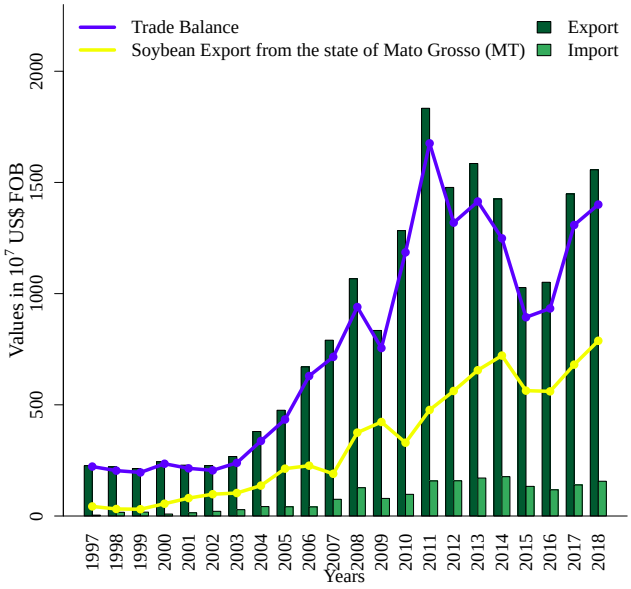


Figure 2.9: Trade balance of the state of Mato Grosso from 1997 to 2018.

Source: MDIC/SECEX.

Soybean exports represented 50.61% total exports in 2018, and in the period from 1997 to 2018, there was an increase of approximately 276.73%. The main soybean producing regions and states in the country are discriminated in Table 2.2.

Table 2.2: Distribution of Brazilian soy production in its regions and main producing states in the 2018 harvest.

Rank	**Regions in descending order**	**2018 harvest (10^6 thousand tons)**
1º	Central-West (MT, GO, MS, DF)	53.126 (45.07%)
2º	South (PR, RS, SC)	38.911 (33.00%)
3º	Southeast (MG, SP, RJ, ES)	8.848 (7.51%)
4º	Northeast (BA, MA, PI, AL, CE, RN, PB, PE, SE)	11.534 (9.78%)
5º	North (TO, PA, RO, RR, AP, AC, AM)	5.468 (4.64%)
Total	Brazil	117.888 (100.00%)

Source: IBGE[9].

The state of Mato Grosso is the most prominent in the Central-West region and ranks first in soybean production, producing around 26.81% country's total production in 2018, and about 50.50% total production in the Central-West region.

The spatial distribution of soybean production in 1990 and 2018 for the state of Mato Grosso, is represented in Figure 2.10.

[9]The naming of the regions and states of Brazil is listed in Frames 5.1 e 5.2 in the appendix.

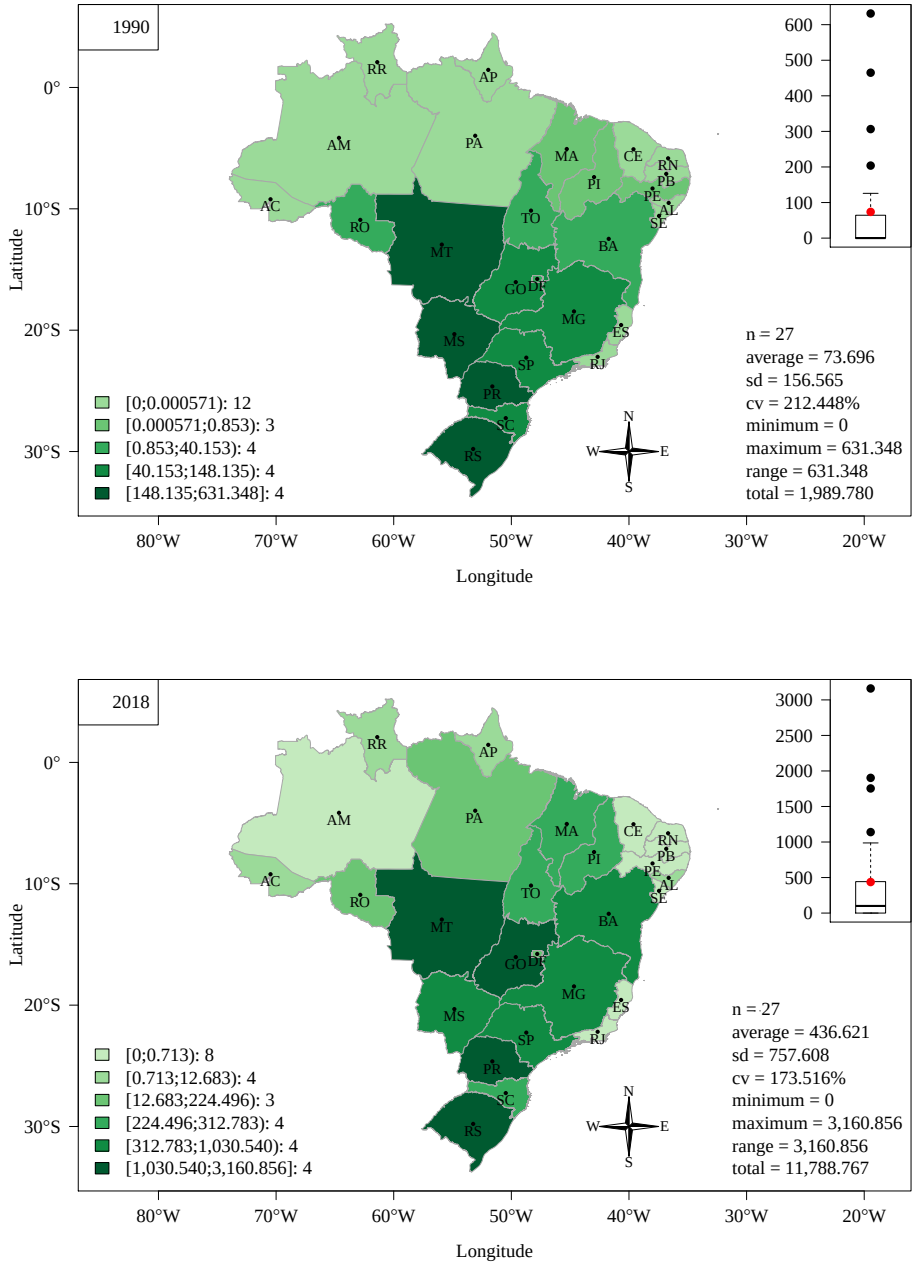


Figure 2.10: Spatial distribution of soybean production from temporary and permanent crops in Brazil, in 10^4 thousand tons, from 1990 to 2018.

Source: IBGE.

From 1990 to 2018, there was a marked development in the number of soybean-producing regions by the states of the country.

The state of Mato Grosso in 1990 was among the major soybean producers in the country, reaching the year 2018 as the largest producer. It is worth highlighting the development of the region known by the acronym MA-TO-PI-BA, composed of the states of its acronym, and is considered as the last productive frontier of soybean and other

crops.

The spatial distribution of soybean production in 1990 and 2018 in Brazil, is revealed by Figure 2.11.

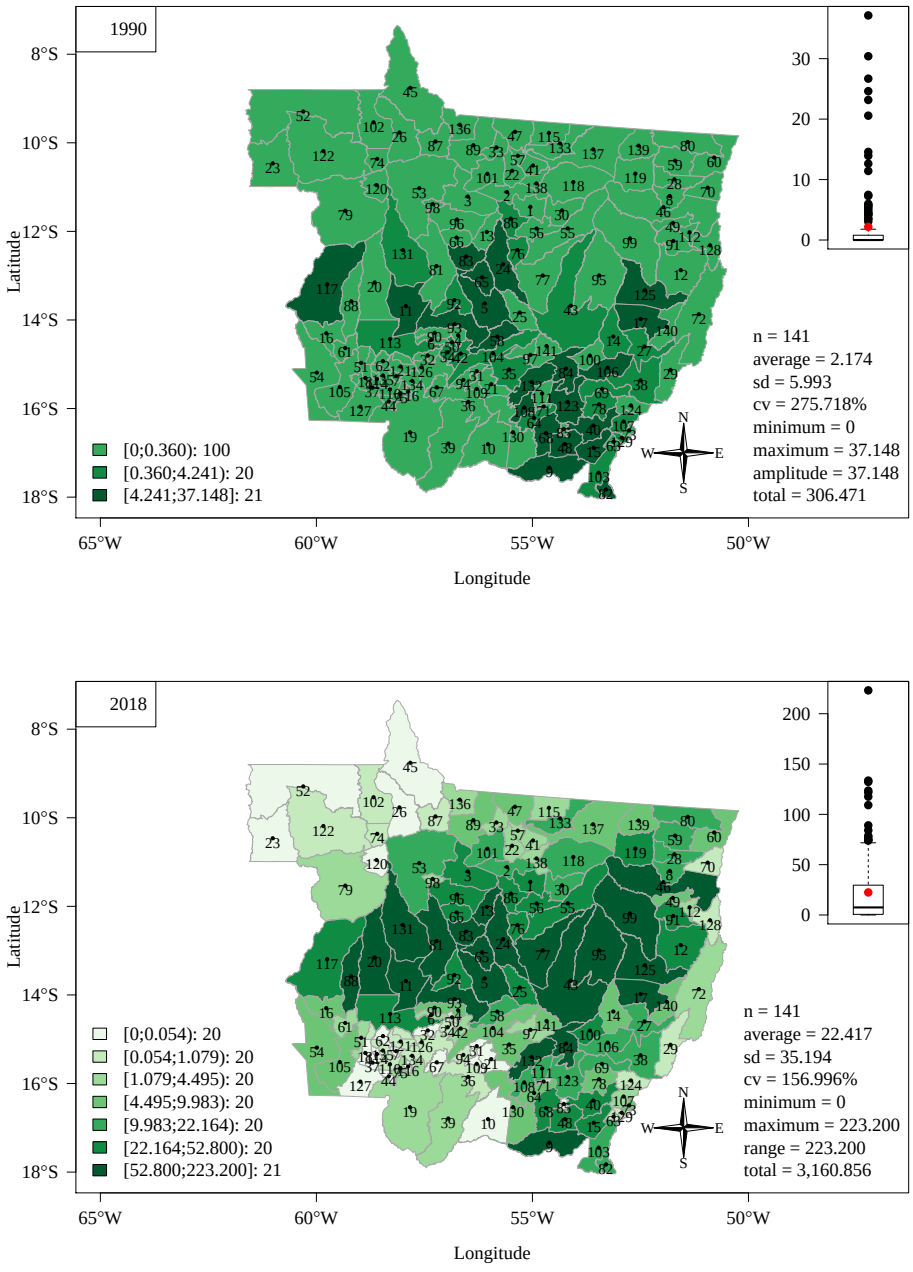


Figure 2.11: Spatial distribution of soybean production from temporary and permanent crops in the state Mato Grosso, in 10^4 thousand tons, from 1990 to 2018.

Source: IBGE[10].

[10]The identification of the numbering of the 141 municipalities in the state of Mato Grosso is listed in Frame 5.3 in the appendix.

Between 1990 and 2018, the state showed an increase in soybean production of approximately 9.70%, supplying the domestic and foreign market with the human and animal nutrition. The main soybean producers in the state are related in Table 2.3.

Table 2.3: Main soybean-producing municipalities in the state of Mato Grosso in the 2018 harvest.

Rank	Identification in the Map	Municipalities	2018 harvest (thousand tons)
1º	24	Sorriso (MT)	2,232,000
2º	5	Nova Mutum (MT)	1,335,600
3º	11	Campo Novo do Parecis (MT)	1,322,400
4º	20	Sapezal (MT)	1,235,400
5º	77	Nova Ubiratã (MT)	1,218,000
6º	99	Querência (MT)	1,176,000
7º	93	Diamantino (MT)	1,091,880
8º	84	Primavera do Leste (MT)	890,400
9º	125	Canarana (MT)	841,500
10º	131	Brasnorte (MT)	786,480

Source: IBGE.

The growth behavior of soybean production from 1990 to 2018, in Brazil, in the Central-West region and in the state of Mato Grosso is illustrated in Figure 2.12.

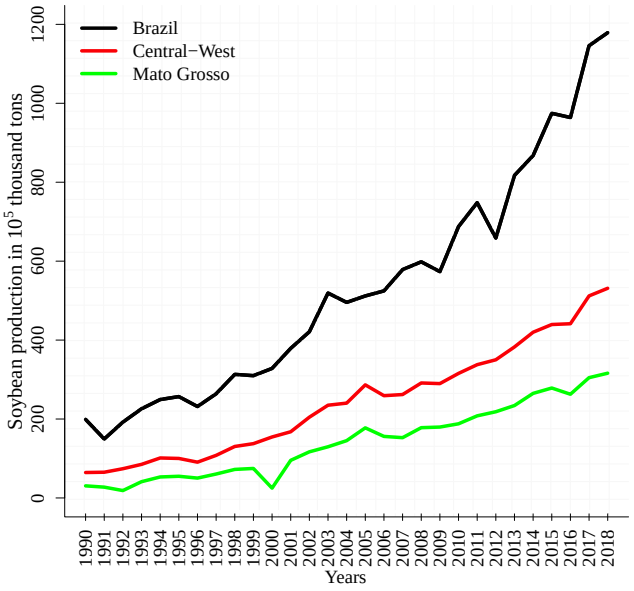


Figure 2.12: Distribution of soybean production from temporary and permanent crops, in Brazil, Central-West region, and the state of Mato Grosso, from 1990 to 2018.

Source: IBGE.

There was an increase in the Brazilian soybean production in this period, and the Central West region and the state of Mato Grosso were responsible for a good part of this growth, together with the other producing regions of the country. The spatial movement of soybean grain production values in Brazil at that time can be seen in figure 2.13.

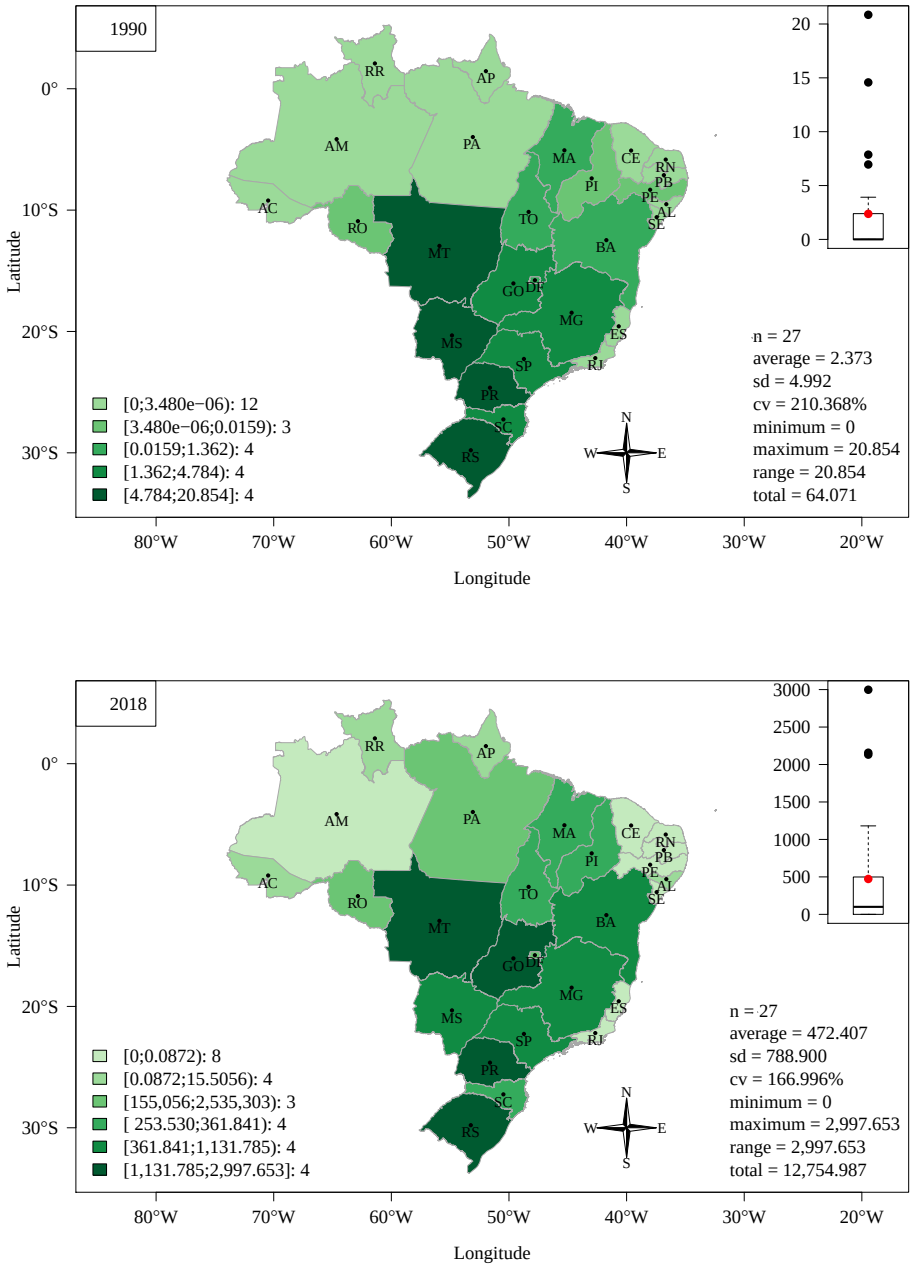


Figure 2.13: (1990) Spatial distribution of soybean production from temporary and permanent crops, in Brazil, in thousand reais (R$), in 1990. (2018) Spatial distribution of soybean production from temporary and permanent crops, in Brazil, in 10^4 thousand reais (R$), in 2018.

Source: IBGE[11].

The distribution of values generated for soybean production in the state of Mato Grosso, is illustrated in Figure 2.14.

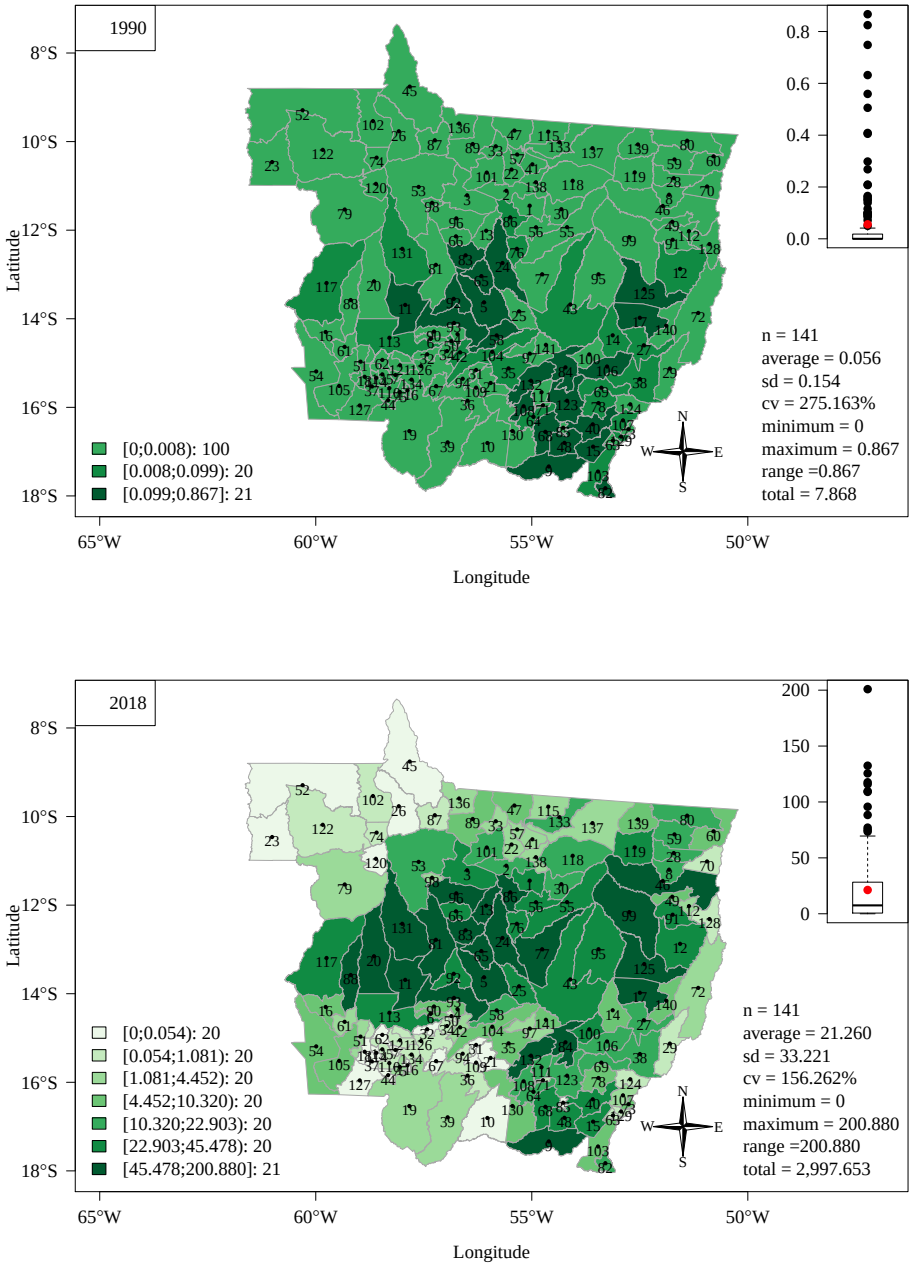


Figure 2.14: (1990) Spatial distribution of soybean production from temporary and permanent crops, in the state of Mato Grosso, in thousand reais (R$), in 1990. (2018) Spatial distribution of soybean production from temporary and permanent crops, in the state of Mato Grosso, in 10^4 thousand reais (R$), in 2018.

Source: IBGE[12].

[11]Values for 1990 were converted from one thousand Cruzeiros (Cr$) to one thousand Reais (R$). It is worth remembering that this conversion is nominal, so it does not take into account the effects of inflation.

In this way, in 2018, the state of Mato Grosso represented 23.50% soybean production value in Brazil. This detailed development between 1900 and 2018 can be seen in Figure 2.15.

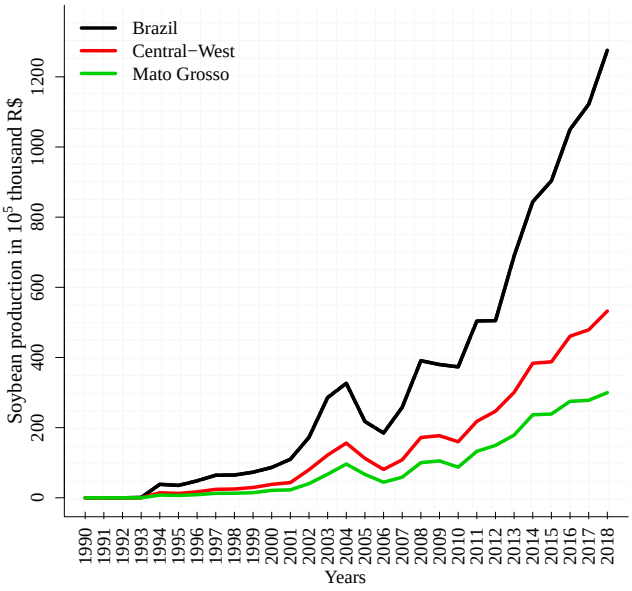


Figure 2.15: Distribution of soybean production values, in thousand reais, of temporary and permanent crops in Brazil, Central-West region and state of Mato Grasso, from 1990 to 2018.

Source: IBGE[13].

The evolution of soybean production values in this period is evident in Brazil, Central-West region and mainly in the state of Mato Grosso. At the end of 2018, the state accounts for 52.28% and 14.60% value of soybeans in grains from the Central-West region and Brazil, respectively.

Soybean processing produces derivatives, which also generate profits for the country in this period, as illustrated in Figure 2.16.

[12]Values for 1990 were converted from one thousand Cruzeiros (Cr$) to one thousand Reais (R$). It is worth remembering that this conversion is nominal, so it does not take into account the effects of inflation

[13]Values for the years 1990, 1991 and 1992 were converted from one thousand Cruzeiros (Cr$) to one thousand Reais (R$). And the values for the year 1993 were converted from one thousand Cruzeiros Reais (Cr$) to one thousand Reais (R$). It is worth remembering that these conversions are nominal, therefore, they do not take into account the effects of inflation.

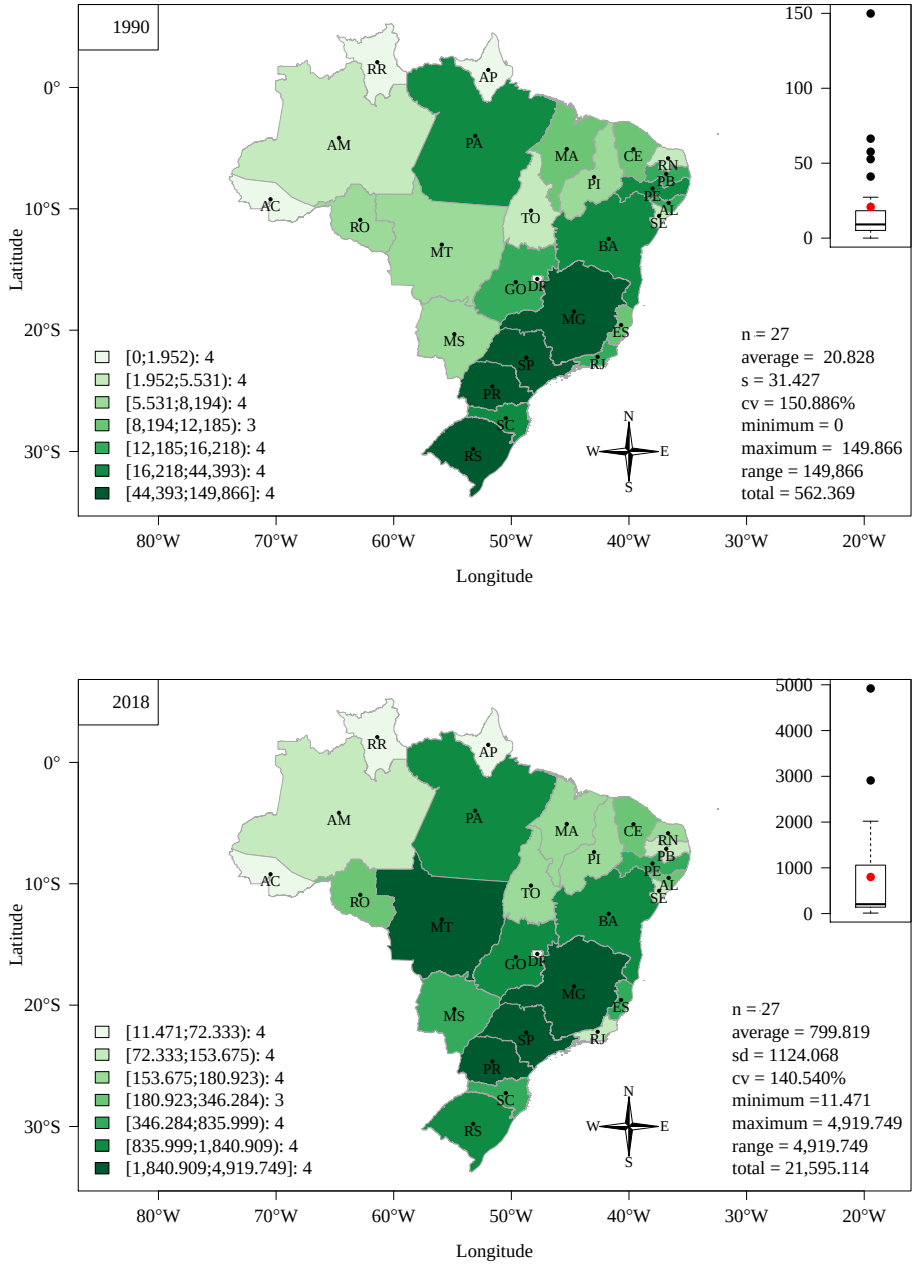


Figure 2.16: (1990) Spatial distribution of soybean derivative values from temporary and permanent crops in Brazil, in thousand reais (R$), in 1990. (2018) Spatial distribution of soybean derivative values from temporary and permanent crops in Brazil, in 10^4 thousand reais (R$), in 2018.

Source: IBGE[14].

The state of Mato Grosso stands out in generating profits from the production of soybean products, together with the states of Paraná, São Paulo and Minas Gerais. The

[14]Values for 1990 were converted from one thousand Cruzeiros (Cr$) to one thousand Reais (R$). It is worth remembering that this conversion is nominal, so it does not take into account the effects of inflation.

comportment of the distribution of values of derived from soybeans obtained by the state of Mato Grosso over these years, is presented in the Figure 2.17.

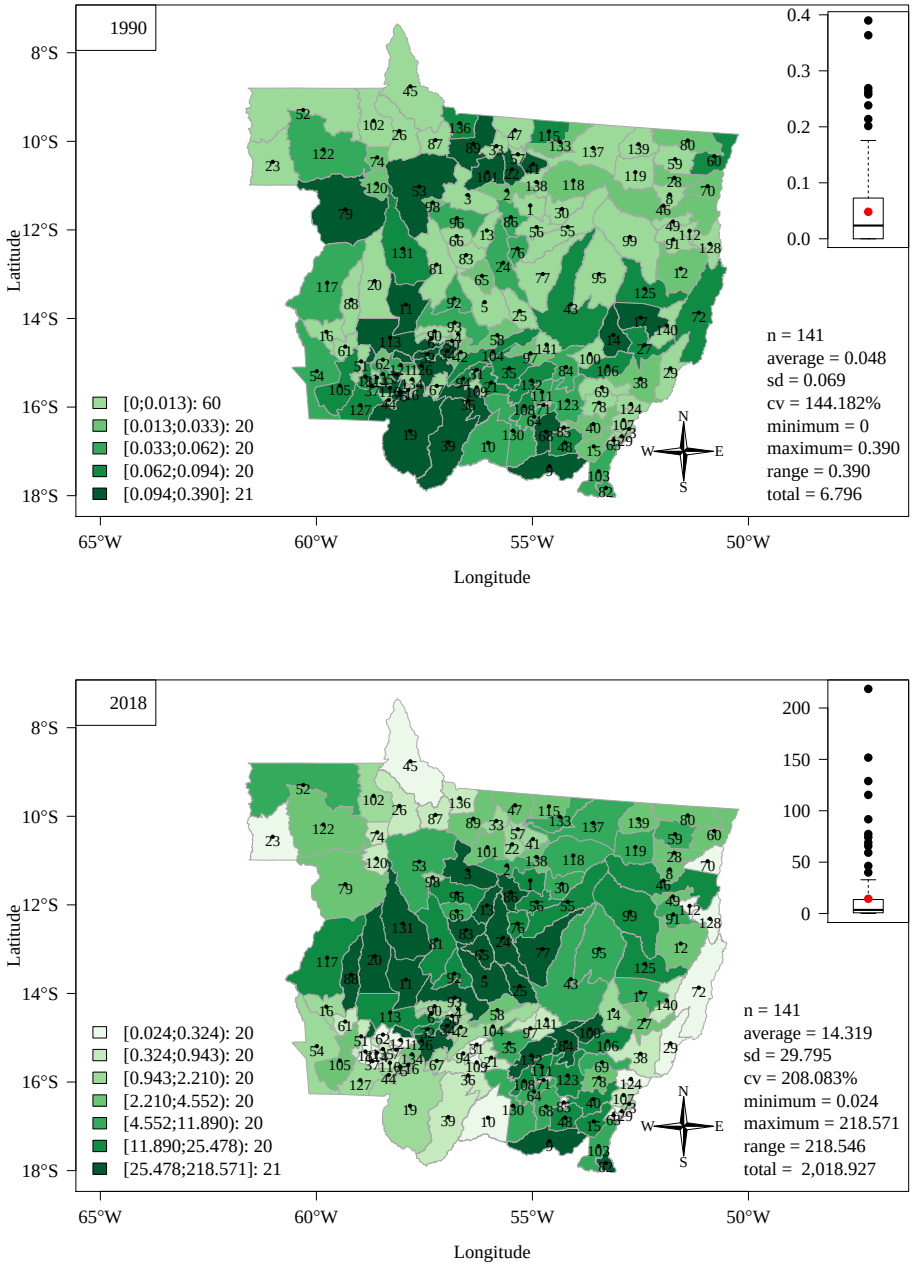


Figure 2.17: (1990) Spatial distribution of soybean derivative values from temporary and permanent crops in the state of Mato Grosso, in thousand reais (R$), in 1990. (2018) Spatial distribution of soybean derivative values from temporary and permanent crops in the state of Mato Grosso, in 10^4 thousand reais (R$), in 2018.

Source: IBGE[15].

[15]Values for 1990 were converted from one thousand Cruzeiros (Cr$) to one thousand Reais (R$). It is worth remembering that this conversion is nominal, so it does not take into account the effects of inflation.

The evolution in soybean derivative production in this period is detailed in Figure 2.18.

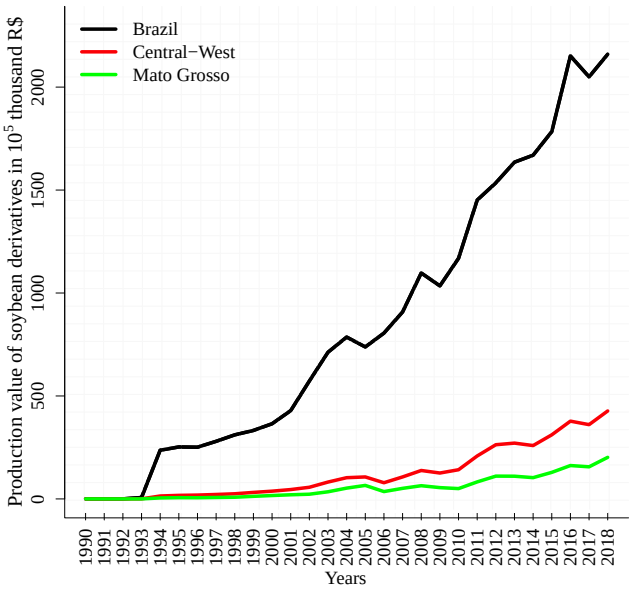


Figure 2.18: Distribution of the values in thousand reais of the production of soybean derivatives in grains from temporary and permanent crops in Brazil, Central-West and Mato Grosso from 1990 to 2018.

Source: IBGE[16].

In 2018, the state of Mato Grosso was responsible for 47.27% and 9.34% soybean derivative production values of the Central-West region and Brazil, respectively. Despite the growth in derivative production values from Brazil, Central-West region and Mato Grosso, production is still small in this segment, since soybean grains are practically all exported to China, which has numerous crushing and processing companies spread all over that country. In this context, the importance of the production and marketing of soybean is revealed, as well as for the generation of jobs in Brazil and in the world, mainly in the state of Mato Grosso.

2.1 References

Barros, G. S. d. C., A. F. Silva, A. L. Fachinello, N. R. Castro, L. Gilio, and G. F. Giachini, 2017 PIB Cadeias do Agronegócio: 1º Semestre de 2017. Piracicaba: CEPEA/ESALQ/USP **16**: 1–15.

[16]Values for the years 1990, 1991 and 1992 were converted from one thousand Cruzeiros (Cr$) to one thousand Reais (R$). And the values for the year 1993 were converted from one thousand Cruzeiros Reais (Cr$) to one thousand Reais (R$). It is worth remembering that these conversions are nominal, therefore, they do not take into account the effects of inflation.

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

Bonato, E. R. and A. L. V. Bonato, 1987 A soja no Brasil: história e estatística. Londrina, PR: EMBRAPA pp. 1–61.

Bonetti, L. P., 1981 Distribuição da soja no mundo, pp. 1–16 in *A soja no Brasil*, edited by Miyasaka, S. and J. C. Medina, Campinas: Instituto de Tecnologia de Alimentos.

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

Câmara, G. d. S., 2011 Introdução ao agroneócio soja. Texto básico da disciplina essencial LPV 584: Cana-de-açúcar, mandioca e soja do curso de graduação em Egenharia Agronômica da USP/ESALQ .

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.

Lemos, M. L. F., D. D. Guimarães, G. B. S. Maia, and G. F. Amaral, 2017 Agregação de valor na cadeia de soja. BNDES Setorial **46**: 167–217.

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

Rodrigues, R., 2006 O céu é o limite para o agronegócio brasileiro. Revista Conjuntura Econômica **60**: 14–15.

Roessing, A. C. and J. J. Lazzarotto, 2004 Criação de empregos pelo complexo agroindustrial da soja. Embrapa Soja-Documentos (INFOTECA-E) .

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.

3 ESTIMATION OF MISSING VALUES BY APPLYING SPLINE INTERPOLATION TECHNIQUES TO DATA ON VARIABLES RELATED TO SOYBEAN PRODUCTION IN A MUNICIPALITIES IN THE STATE OF MATO GROSSO, BRAZIL, FROM 1990 TO 2018

3.1 Resumo

Os dados de produção de grãos de soja em mil toneladas, valores de produção de grãos de soja em mil reais e valores de produção derivados de soja em mil reais do Brasil são coletados anualmente pelo IBGE (Instituto Brasileiro de Geografia e Estatística) e, geralmente, são encontrados dados ausentes em estados e municípios do país, o que pode causar viés de estimativas nestas variáveis, com relação aos dados observados e não observados. O estado de Mato Grosso figura como maior produtor e exportador de soja do país, assim, utilizou-se da ferramenta de imputação univariada de séries temporais, por meio da aplicação da interpolação *spline*, em 46 de seus municípios nestas três variáveis atreladas à produção de soja, com o intuito de avaliar as diferenças entre as séries observadas e as com valores imputados, em cada um destes 46 municípios. Observou-se que após a imputação todas séries foram comparadas com as observadas e validadas estatisticamente pelo teste de Queinouille nos 46 municípios, e no total acumulado das três variáveis envolvidas com a produção de soja também pelo teste de Wilcoxon, e que não houveram mudanças no padrão geográfico destas variáveis no estado de Mato Grosso, e ocorreram aumentos de 5,92%, 3,58% e 2,84% para a produção de soja, valor de produção de soja e valor de derivados de soja no estado após a imputação dos 46 municípios do estado, e que esta metodologia facilita o monitoramento da cultura de soja no estado de Mato Grosso e seus municípios de 1990 a 2018.

Palavras-chave: Produção; imputação; comparação das estimativas.

3.2 Abstract

Production data of soybeans in thousand tons, production values of soybeans in thousand reais and production values derived from soybeans in thousand reais in Brazil are collected yearly by IBGE (Brazilian Institute of Geography and Statistics) and, generally, missing data are found in states and cities in the country, which can cause estimation bias in these variables, with respect to observed and unobserved data. State of Mato Grosso represents the largest producer and exporter of soybeans in the country, so it used the univariate imputation tool of time series, through application of spline interpolation, in 46 of its municipalities in these three variables linked to soybean production, in order to assess the differences between observed series and those with imputed values, in each of

these 46 municipalities. It was observed that after imputation, all series were compared with those observed and statistically validated by the Queinouille test in 46 municipalities. In the cumulative total of the three variables involved with soybean production, also by Wilcoxon test, there were no changes in geographic pattern of these variables in state of Mato Grosso, and there were increases of 5.92%, 3.58% and 2.84% for soy production, value of soybean production and value of soybean derivatives in the state after imputation of 46 municipalities in the state. This methodology facilitates the process of estimating and monitoring soybean production in Mato Grosso and its municipalities from 1990 to 2018.

Keywords: Production; imputation; comparison of estimates.

3.3 Introduction

The monitoring and data collection of production, production value and production value of soybean derivatives is carried out to map the potential and weaknesses of this economic activity of great importance for Brazil, and it is done by IBGE. However, this process is complicated by the frequent presence of large data sets with missing values since, both in the collection and in the updating of data, there is loss of information due to the country´s territorial extension and the difficulty of access to many regions, lack of access to economic information considered confidential together with farms and producing companies in the country's cities, small number of people to collect, low financial investments for this task, human failures, among others. The problem with incomplete data sets is that they can lead to different results from those that would be obtained from a complete data set.

There are three main problems that can arise when dealing with incomplete data, the first is that there may be a decrease in information, as a consequence, there is a loss of efficiency in the estimates. Second, there are several complications related to data handling, computing and analysis, due to irregularities in data structure and the impossibility of using different methodologies to optimize data analysis. Third, and more importantly, the results can be biased due to systematic differences between observed and unobserved data (Hawthorne and Elliott, 2005; Norazian *et al.*, 2008).

There are two types of missing data: the non-ignorable data, which is the case where the probability of losing a data depends only its value, there is no interference of external factors to generate the presence or absence of data, and the ignorable missing data, which happens when the probability of absence of a data does not depend on its value, but on several external factors, not measured/identified in the data collection. There are three mechanisms associated with the ignored missing data: Missing Completely at Random (MCAR), Missing at Random (MAR) and No Missing at Random (NMAR) (Rubin, 1976, 1978).

MCAR is the lack of completely random data, which occurs when the absence of data is random in all the data in the working set, that is, when missing data have no relationship with another variable measured in the study, nor with the own variable that presents the absence; the process happened due to unforeseen and random causes linked to the collection (Little and Rubin, 2002; Enders, 2010; Buuren, 2018).

In MAR, the absence of data appears in certain subsets in a random way in the data of a given variable, and indicates that the observation gap occurs because of certain information contained in the data set. In other words, the lost data depend directly on observed and controlled information, available for collection and directly correlated with the variable that has missing data. This type of situation occurs when the probability of a variable presenting lack of data depends directly on available and prior information to data collection (Little and Rubin, 2002; Enders, 2010; Buuren, 2018).

NMAR happens when the two previous conditions are present; soon, the probability of data loss may vary and is not characterized by an available predictor. In this case, there may be a problem in the study's sample planning (Little and Rubin, 2002; Enders, 2010; Buuren, 2018).

This research focuses on application of MAR mechanism, with recommendation of univariate imputation by means of interpolation of cubic splines method, to solve incomplete data problem of each of the 46, of a total of 141 cities in the state of Mato Grosso/Brazil, from 1990 to 2018, taking into account particularity and variability of each of the production variables of soybeans in thousand tons, production values of soybeans in thousand reais (R$) and production values of soybeans derivatives in thousand reais (R$).

In the field of multiple imputation, some studies by authors such as Little (1988), Rubin and Schenker (1991), Rubin (2004), Buuren and Groothuis-Oudshoorn (2010), Spratt *et al.* (2010), Junger and De Leon (2015), focusing on longitudinal data and also on multivariate time series data.

The use of univariate data imputation has been used in several researches and in different data sets. In Junninen *et al.* (2004), there is a comparison of simple univariate imputation methods by means of missing data simulations, in a data set that involves air quality. Norazian *et al.* (2008) also allude to the comparison of simple univariate imputation techniques in air quality data set, by monitoring concentrations of air PM_{10} (inhalable particles with a diameter less than 10 microns) and their ideal concentrations. Honaker and King (2010) propose the comparison of univariate imputation algorithms in data related to social sciences. The works of Twisk and Vente (2002), Engels and Diehr (2003), discuss imputations of clinical longitudinal data and e Perneger and Burnand (2005) also make use of clinical data for application of simple univariate imputation.

In the context of univariate time series, nowadays, numerous imputation tech-

niques are available such as: various interpolations, Kalman filter, Persistence, media Weighted Furniture, Classic Imputation, Random Sample, Seasonal Decomposition and Seasonality by parts. Some of several time series univariate imputation algorithms are implemented for use in imputeTS package of R software (Moritz and Bartz-Beielstein, 2017; Demirhan and Renwick, 2018; R Core Team, 2021).

In this way, to solve the problem of missing data in variables related to soybean production in municipalities in the state of Mato Grosso/Brazil, and aiming to have better estimates and complete monitoring of soybean grain production in thousand tons, with a value of soybean production in one thousand reais (R$) and the value of soybean derivatives in one thousand reais (R$) in the state of Mato Grosso and its municipalities in the period from 1990 to 2018, we propose the application of time series univariate imputation by means of interpolation of cubic splines in 46 municipalities of the state, a research unprecedented in the current scenario scientific. Thus, this research presents the following sections: Introduction, Material and Methods, Results and Discussion and, finally, the Conclusions.

3.4 Material and Methods

3.4.1 Data related to soybean production in the state of Mato Grosso/Brazil

Each year, data on soybean production in Brazil and its states are published and made available by IBGE. In this way, data from production of soybeans in thousand tons, production values of soybeans in thousand reais (R$) and production derivatives of soybean in thousands reais (R$), for each of the 141 municipalities in the state of Mato Grosso, were collected together with IBGE. In the period from 1990 to 2018, for the 141 municipalities, a total of 4,089 observations were evaluated in the total set, with 3,785 (92.57%) valid observations and 304 (7.43%) missing observations in each of the three variables above. Authors such as Buuren (2018) indicate multivariate imputation to complete these observations. The total results accumulated over this period can be better evaluated in Table 3.1.

Table 3.1: Values of accumulated data related to soy produced in the state of Mato Grosso from 1990 to 2018

Soybean	**Accumulated**	**Observed (%)**	**Missing (%)**
Grain production (thousand tons)	419,602,745	92.57	7.43
Grain production (R$ thousand reais)	256,253,655,60	92.57	7.43
Derivates (R$ thousand reais)	156,815,568,20	92.57	7.43

Source: IBGE.

The three variables indicate that the accumulated values are not complete in their calculations, because there is absence of information in the collected data contained in missing values. These values are unavailable in 46 muncipalities in the state of Mato Grosso/Brazil, as shown in Figure 3.1 and, later, the detailed amount of data missing in each of these municipalities, is presented in Table 3.2.

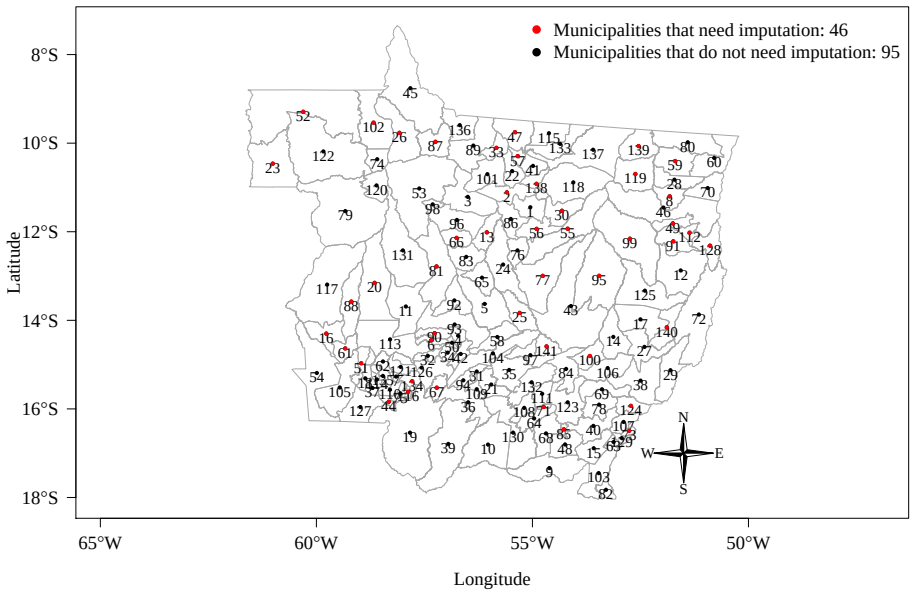


Figure 3.1: Map of Mato Grosso discriminating municipalities that need and those that do not need imputation.

Source: IBGE.

Table 3.2: Municipalities that necessitate imputation in their variables of Soybean production (thousand tons), Soybean production (thousand reais (R$)), Soybean derivatives (thousand reais (R$))

ID	Municipality	Missing Data	%	Absence Period
13	Ipiranga do Norte	15	51.72	1990 to 2004
66	Itanhangá	15	51.72	1990 to 2004
23	Rondolândia	11	37.93	1990 to 2000
25	Santa Rita do Trivelato	11	37.93	1990 to 2000
51	Vale de São Domingos	11	37.93	1990 to 2000
52	Colniza	11	37.93	1990 to 2000

Table 3.2: Continued

ID	Municipality	Missing Data	%	Absence Period
61	Conquista D'Oeste	11	37.93	1990 to 2000
91	Bom Jesus do Araguaia	11	37.93	1990 to 2000
100	Santo Antônio do Leste	11	37.93	1990 to 2000
112	Serra Nova Dourada	11	37.93	1990 to 2000
116	Curvelândia	11	37.93	1990 to 2000
128	Novo Santo Antônio	11	37.93	1990 to 2000
138	Nova Santa Helena	11	37.93	1990 to 2000
139	Santa Cruz do Xingu	11	37.93	1990 to 2000
140	Nova Nazaré	11	37.93	1990 to 2000
95	Gaúcha do Norte	8	27.58	1990 to 1996; and 2002
16	Nova Lacerda	7	27.14	1990 to 1996
20	Sapezal	7	27.14	1990 to 1996
30	União do Sul	7	27.14	1990 to 1996
33	Carlinda	7	27.14	1990 to 1996
47	Novo Mundo	7	27.14	1990 to 1996
55	Feliz Natal	7	27.14	1990 to 1996
77	Nova Ubiratã	7	27.14	1990 to 1996
88	Campos de Júlio	7	27.14	1990 to 1996
3	Tabaporã	3	10.34	1990 to 1992
6	Santo Afonso	3	10.34	1990 to 1992
8	Cana Brava do Norte	3	10.34	1990 to 1992
26	Nova Bandeirantes	3	10.34	1990 to 1992
44	Glória D'Oeste	3	10.34	1990 to 1992
49	Alto Boa Vista	3	10.34	1990 to 1992
56	Santa Carmem	3	10.34	1990 to 1992
57	Nova Guarita	3	10.34	1990 to 1992
59	Confresa	3	10.34	1990 to 1992
67	Porto Estrela	3	10.34	1990 to 1992
71	São Pedro da Cipa	3	10.34	1990 to 1992
73	Ribeirãozinho	3	10.34	1990 to 1992
81	Nova Maringá	3	10.34	1990 to 1992
85	São José do Povo	3	10.34	1990 to 1992
87	Nova Monte Verde	3	10.34	1990 to 1992
90	Nova Marilândia	3	10.34	1990 to 1992
99	Querência	3	10.34	1990 to 1992
102	Cotriguaçu	3	10.34	1990 to 1992
119	São José do Xingu	3	10.34	1990 to 1992
124	Pontal do Araguaia	3	10.34	1990 to 1992
134	Lambari D'Oeste	3	10.34	1990 to 1992
141	Planalto da Serra	3	10.34	1990 to 1992

Source: IBGE.

Table 3.2 indicated that the 46 municipalities have 15 (51.72%), 11 (37.93%), 8 (27.58%), 7 (27.14%) and 3 (10.34%) missing data sections in their time series with 29 data, contained in the period from 1990 to 2018.

It is worth mentioning that studies by Ragel (2000), Batista and Monard (2003), Acurna and Rodriguez (2004), Farhangfar *et al.* (2007), Harrell Jr (2015)identified the use of multiple imputation resource in data with a percentage of missing data from 5% to 50%, and in the work of Schafer and Graham (2002) there is the use of the same resource with up to 70% of lack. However, in studies of political science data, King *et al.* (2001) mostraram show data imputations that exceeded 50% of missing data, reaching up to 90%. The research by Twumasi-Ankrah *et al.* (2019) has the use of univariate imputation of time series with missing data from 10% to 90%, and this study still indicates that the use of imputations by interpolations generates good results in MAR imputation procedures.

3.4.2 Procedure for losses at random MAR

A time series with n observations can be written as $\{Y_t; t = 1, \ldots, n\}$ and divided into two other series $\{Y_t^{observed}; t = 1, \ldots, n-k\}$, and in this series there will be $k < n$ missing data, arranged at random, denoted in its inside as follows $\{Y_t^{missing}; t = 1, \ldots, k\}$. This time series can be expressed by matrices $\mathbf{Y}^{observed} = [Y_1 \ldots Y_{n-k}]$ and $\mathbf{Y}^{missing} = [Y_1 \ldots Y_k]$, where $\mathbf{Y}$ is formed by two subsets of these two matrices, $\mathbf{Y} = \left\{\mathbf{Y}^{observed}, \mathbf{Y}^{missing}\right\}$. There is a matrix $\mathbf{R}$, of the same dimension as $\mathbf{Y}$, with elements $r_i j$, in which, if $r_{ij} = 1$ there will be the presence of the values of $\mathbf{Y}^{observed}$, and if $r_{ij} = 0$ will occur the presence of $\mathbf{Y}^{missing}$ (Buuren, 2018).

The distribution of $\mathbf{R}$ is linked to $\mathbf{Y}$, either by sampling design or simply by chance. Therefore, the general identity (3.1) defines a missing data probability model:

$$P(\mathbf{R}|\mathbf{Y}^{observed}, \mathbf{Y}^{missing}, \boldsymbol{\psi}) \tag{3.1}$$

Where the term $\boldsymbol{\psi}$ reserves parameters not known to the probability model. Equation (3.2) reveals a MAR procedure.

$$P(\mathbf{R}{=}\mathbf{0}|\mathbf{Y}^{observed}, \mathbf{Y}^{missing}, \boldsymbol{\psi}) = P(\mathbf{R}{=}\mathbf{0}|\mathbf{Y}^{observed}, \boldsymbol{\psi}) \tag{3.2}$$

The missing data depends only on the observed information and external factors, available for analysis and correlated with the variable that has missing data (Gandolfi, 2016; Buuren, 2018).

3.4.3 Imputation through Interpolation by cubic Spline

Several situations involving mathematical problems do not have an exact solution. Therefore, it uses methods that indicate an approximate solution. Interpolation is widely

used when a function $f(x)$ is known in its domain $[a, b]$, and it has the objective of discovering its value at a certain point that belongs at this same interval. It is worth to say that the expression of $f(x)$ can be very complex and reveal a degree of difficulty in the operations of differentiation and integration. Therefore, one always looks for a single function $g(x)$ with similar characteristics to exchange it with $f(x)$ (Ruggiero and Lopes, 1997; Atkinson, 2008; Dahlquist and Björck, 2008; Conte and De Boor, 2017).

A function is called Spline when approximations are made in a given subinterval $[x_i, x_{i+1}]$, with $\{i = 0, 1, \ldots, n-1\}$, by a polynomial of degree p, with some assumptions about Spline function Spline. Given a function $f(x)$ and at points x_i,a function $S_p(x)$ is called interpolating Spline of degree n with its defined nodes as the points x_i, respecting the following requirements:

1. In each subinterval $[x_i, x_{i+1}]$, $S_p(x)$is a polynomial of degree n, given by $s_p(x)$;
2. $S_p(x)$ is continuous and presents a continuous derivative up to the order $(p-1)$ in $[x_0, x_n]$;
3. $S_p(x_i) = f(x_i)$.

A function named by cubic Spline uses a third degree polynomial mechanism to interpolate values between each pair of observed points. In this situation, a distinct polynomial of degree 3 is used at each interval, and it is constructed in such a way that it always presents a trajectory through the original points included in the study interval, and it has continuous derivatives in the connections of each subinterval. This process ensures that there are no sudden transitions in the slopes or curvatures between successive subintervals and, when reaching the end of the interval under analysis, a smooth curve can be guaranteed, under respective points of interest (De Boor, 1978; Green and Silverman, 1993; Knott, 2000).

In a cubic Spline function, the first and second derivatives must be continuous, which ensures that there are no peaks or abrupt inversion of curvature in their nodes, therefore, it is one of the most used, given by the composition of polynomials of degree 3 named $s_k(x)$ in the intervals $[x_{k-1}, x_k]$ with $k = \{1, 2, \ldots, n\}$. Establishing a $f(x)$, and the points x_i and $i = \{1, 2, \ldots, n\}$ of function $S_3(x)$, which will be defined as interpolating cubic spline of $f(x)$ in nodes x_i, if there are n polynomials of degree 3, named as $s_k(x)$ and $k = \{1, 2, \ldots, n\}$, which follow the following conditions:

1. $S_3(x) = s_k(x)$ with $x \in [x_{k-1}, x_k]$;
2. $S_3(x) = f(x_i)$ and $i = 1, 2, \ldots, n$;
3. $s_k(x_k) = s_{k+1}(x_k)$ and $k = 1, 2, \ldots, (n-1)$;

4. $s'_k(x_k) = s'_{k+1}(x_k)$ and $k = 1, 2, \ldots, (n-1)$;

5. $s''_k(x_k) = s''_{k=1}(x_k)$ and $k = 1, 2, \ldots, (n-1)$.

Thus, authors such as Schultz (1973), Press *et al.* (1986), Ruggiero and Lopes (1997), Prenter (2008) and Hastie *et al.* (2009) demonstrate that $S_3(x)$ composed by a system of cubic polynomials by parts, expressed as follows:

$$S_3(x) = \begin{cases} s_1(x),\ x_0 \leq x \leq x_1 \\ \quad \vdots \\ s_k(x),\ x_{k-1} \leq x \leq x_k \\ \quad \vdots \\ s_n(x),\ x_{n-1} \leq x \leq x_n \end{cases} \tag{3.3}$$

Where $s_k(x)$, is polynomial of the third degree under the interval of $[x_{k-1}, x_k]$ e $k = \{1, 2, \ldots, n\}$ described by equation (3.4).

$$s_k(x) = a_k(x - x_k)^3 + b_k(x - x_k)^2 + c_k(x - x_k) + d_k \tag{3.4}$$

The continuity conditions of $s_k(x)$ in its first derivatives indicate several relationships between the coefficients.

In calculating $S_3(x)$, it is necessary to determine four coefficients for each k, that is, a total of $4n$ coefficients as follows: $a_1, b_1, c_1, d_1, a_2, b_2, c_2, d_2, \ldots, a_n, b_n, c_n, d_n$. Since $S_3(x)$, $S'_3(x)$ and $S''_3(x)$ are continuous, we have for all $k = \{1, 2, \ldots, n-1\}$:

1. $s_k(x_k) = s_{k+1}(x_{k+1})$;

2. $s'_k(x_k) = s'_{k+1}(x_{k+1})$;

3. $s''_k(x_k) = s''_{k+1}(x_{k+1})$.

The applications of cubic Splines can be incorporated in the context of adjustments of mixed linear models (White *et al.*, 1999; Verbyla *et al.*, 1999). In a more current work, authors like Souza and Villegas (2020) present cubic Splines technique in partially symmetrical linear models.

In this research, interpolation technique by cubic splines was used by the fgm method developed by the authors Forsythe *et al.* (1977), with nodes being the subintervals of the time series used to fill in missing data from 46 cities in the state of Mato Grosso/Brazil, in soybean production variables, in the value of soybean production and in the value of production derivatives, from 1990 to 2018.

Several studies have made use of imputation techniques for different types of data through application of techniques that involve Splines. The smoothed cubic spline has been used in the imputation of time series by Koopman *et al.* (1999) and in the work of

Fariñas *et al.* (2002), in which this application occurs in missing data in high frequency time series of hourly electric charges in a concessionaire located in the Southeast region of Brazil. Baltazar and Claridge (2006), make use of cubic splines interpolation in their research to fill gaps in meteorological data and hourly energy generation. In Nadir *et al.* (2008), research, the cubic Spline interpolation was used to compose missing data, in data from mobile phone networks, due to eventual losses in the transmission signal of the mobile phones. Já Wongsai *et al.* (2017) makes use of a semi-parametric method, that is, of cubic Splines functions coupled to the ordinary least squares method for filling non observed data of ground temperature, which are captured through satellite, failures can occur due to the harmful effect of heavy rains and clouds on the adhesion of this information.

The application of this imputation procedure by Splines interpolation method in the present research was done with the help of software R, through the use of the imputeTS package, built Moritz and Bartz-Beielstein (2017), to compose the missing data in municipal time series of each of these variables.

3.4.4 Assessment of Imputation Performance

In order to evaluate the performance of the imputations, the observed values were compared with the observed values added to the imputed values in each variable of each of the 46 cities, through the test of equality of the autocorrelation functions proposed by Quenouille (1958) in each of the variables. We opted for this test because we are dealing with small samples and of different sizes, which do not meet the assumptions of normality, and their distributions are considered free.

A time series with $n-k$ observations $\{Y_t^{observed}; t = 1, \ldots, n-k\}$, and in this series $k < n$ values will be imputed randomly inside it denoted in the following way $\{Y_t^{imputed}; t = 1, \ldots, k\}$. Thus, after the procedure, there is a new time series $Z_t^{imputed}$= $\{Y_t^{observed}; t = 1, \ldots, n-k\} \cup \{Y_t^{imputed}; t = 1, \ldots, k\}$ with n values. In this way, one can admit that $Y_t^{observed}$ e $Z_t^{imputed}$ have their autocorrelation functions $\rho_{observed}(j)$ and $\rho_{imputed}(j)$, to $j = \{\pm 0, \pm 1, \pm 2, \ldots\}$. Soon, the following hypotheses will be tested:

$$H_0 : \rho_{observed}(j) = \rho_{imputed}(j)$$
$$H_1 : \rho_{observed}(j) \neq \rho_{imputed}(j)$$

The methodological procedures of Quenouille (1958) test were also detailed later in the works of Echeverry and Toloi (2000) and Costa and Sáfadi (2010), by the following steps:

1. Acquire autocorrelation functions of $\rho_{observed}(j)$ and $\rho_{imputed}(j)$, for the series $Y_t^{observed}$ e $Z_t^{imputed}$, com $j = \{0, 1, 2, \ldots, J\}$;

2. Obtain the autocorrelation function common to the series $\widehat{\rho}(j)$, by calculating the expression:

$$\widehat{\rho}(j) = \frac{n_{observed} \cdot \widehat{\rho}_{observed}(j) + n_{imputed} \cdot \widehat{\rho}_{imputed}(j)}{n_{observed} + n_{imputed}} \tag{3.5}$$

in which $n_{observed}$ and $n_{imputed}$ are the observation numbers of each of the series;

3. Calculate the estimated common autocorrelation function $\widehat{\Phi}(k)$ from $\widehat{\rho}(j)$;

4. From $\widehat{\Phi}(k)$, identify the self-regressive order p;

5. Estimate the p coefficients of the autoregressive model $AR(p)$, by solving the Yule-Walker equations;

6. Adjust an autoregressive model for each of the series $Y_t^{observed}$ e $Z_t^{imputed}$ a autoregressive model $AR(p)$ with the coefficients obtained in step 5, thus obtaining the respective residual series $\widehat{a}_t^{observed}$ e $\widehat{a}_t^{imputed}$;

7. Calculate the partial autocorrelation functions (facp) $v_t^{observed}$ and $v_t^{imputed}$ to the residual series $\widehat{a}_t^{observed}$ and $\widehat{a}_t^{imputed}$;

8. Find the test statistic:

$$SQ = \sum_{j=1}^{J} \frac{(v_j^{observed} - v_j^{imputed})^2}{\frac{1}{n_{observed}-1} + \frac{1}{n_{imputed}-1}}; \tag{3.6}$$

9. If $SQ > C_\alpha$, in C_α is such that $P(\chi_j^2 > C_\alpha) = \alpha$, H_0 is rejected at the level of significance α. In other words, accepting H_0 indicates that there is strong evidence of equality of autocorrelation functions of $Y_t^{observed}$ e $Z_t^{imputed}$, variables, which indicates that both variables come from the same training process, that is, they are equivalent by the test. Otherwise, there will not be this indicator and it is assumed that the variables have differences, and thus, the data imputation did not produce good results.

These tests were carried out for the variables soybean production, soybean production value and soybean production value in the 46 municipalities of the state of Mato Grosso from 1990 to 2018. For the accumulated values of these same variables in 141 municipalities, their average and distributions were compared simultaneously, through application of the classic Wilcoxon test described in Wilcoxon (1945), before and after the imputation of the missing values. The main descriptive measures were also calculated and compared in these same described variables (number of observations: n; arithmetic average: average; standard deviation: sd; coefficient of variation in %: cv; minimum value: minimum; maximum value: maximum; total range: range; total accumulated from 1990 to 2018: total), and their differences between compared series. A regional comparison was also carried out before and after the imputation, through construction of maps for the three accumulated variables.

3.5 Results and Discussion

A city was chosen, from among 46 analyzed in the state of Mato Grosso, to demonstrate the whole process of imputation and validation of imputed data by interpolation by cubic splines, from 1990 to 2018, for variables related to soybean production in the state of Mato Grosso. The chosen city was Ipiranga do Norte, and it has 51.72% lack of data in this period. IBGE data referring to the number of estimated inhabitants in 2020[1] and relative to the total and per capita GDP (Gross Domestic Product), and the agricultural representation of 2017[2] were used to demonstrate the importance of agriculture and agribusiness, as well as soybean production, in the cities of the state, and the dates of foundation of each city were collected together with the respective city halls and in the National Confederation of Cities (CNM[3]).

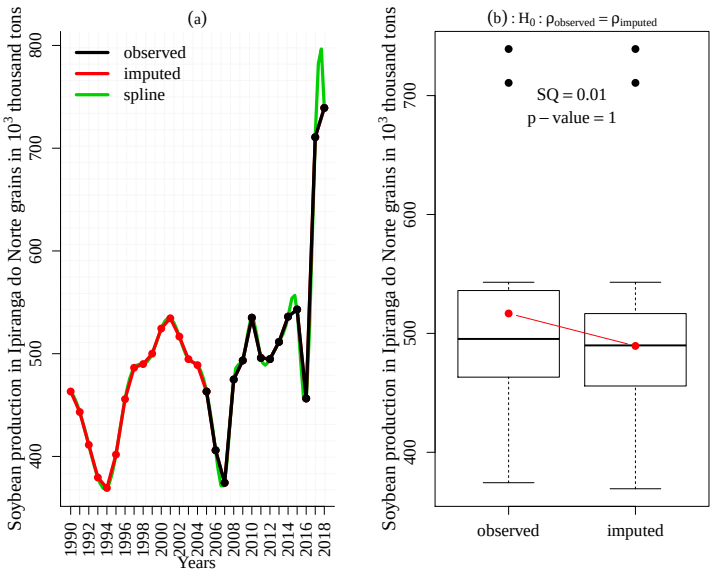
Ipiranga do Norte is located in the intermediate region of Sinop, it is close to the Sorriso, and it has an estimated population of 7,920 inhabitants, according to data from IBGE in 2020. Based on data from IBGE, referring to 2017, the city presents a total GDP of 691,756.65 thousand reais (R$) and excluding taxes, it is verified that 52.79% of this amount comes from agricultural area, which helps to raise per capita GDP of this city to values equivalent to 96,465.86 thousand reais (R$), and the production of local soybean has an important role in this process.

In the period from 1990 to 2018, Ipiranga do Norte accumulated a total production of 7,234,407 thousand tons of soybeans, which makes the city the 17th and sixth producer of soybeans in this period when compared to all 141 cities and to the 46 municipalities that were imputed, respectively. In 2018, the city occupied the 14th position in the whole state, producing 739,200 thousand tons. This production, in that time interval, generated 4,884,823 thousand reais (R$), which ranks this city in the 16th place in the state of Mato Grosso in this regard. The production of soybean derivatives from Ipiranga do Norte, in turn, generated profits of around 2.263,938 thousand reais (R$) in this period, and occupied the 19th place in this segment. The performance of imputation of missing values for production variable, for production value and for soybean derivatives in the city of Ipiranga do Norte can be demonstrated in Figures 3.2, 3.3 and 3.4.

[1] https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html?=&t=resultados

[2] https://www.ibge.gov.br/estatisticas/economicas/contas-nacionais/9088-produto-interno-bruto-dos-municipios.html?t=pib-por-municipio&c=5104526

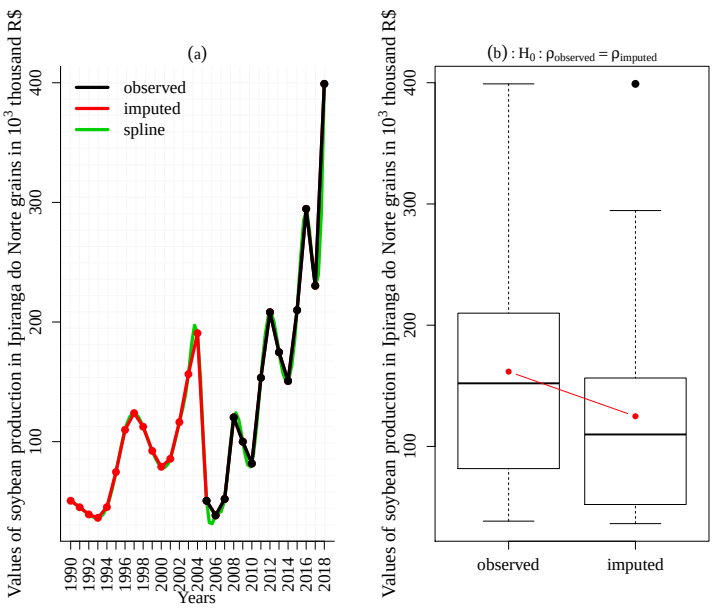
[3] www.cnm.org.br



(c)	OBSERVED	IMPUTED	DIFFERENCE
n	14	29	15
average	516.74	489.44	27.3
sd	100.3	82.29	18.01
cv	19.41	16.81	2.6
minimum	374.4	369.3	5.1
maximum	739.2	739.2	0
range	364.8	369.9	5.1
total	7234.41	14193.9	6959.5

Figure 3.2: (a) Graph of Imputation by Cubic Spline for soybean production in Ipiranga do Norte. (b) Boxplot for the comparison between observed and imputed values. (c) Table of the main descriptive measures for observed and imputed values.

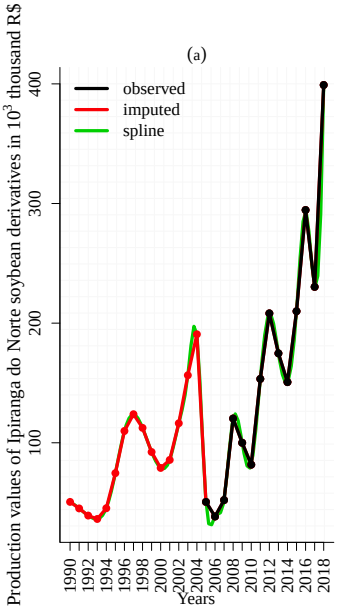
Source: Results of Research.

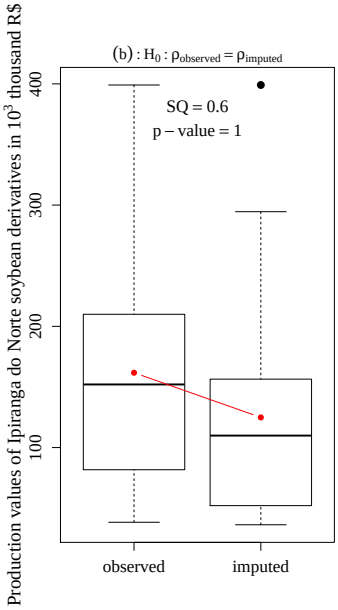


(c)	OBSERVED	IMPUTED	DIFFERENCE
n	14	29	15
average	161.71	124.88	36.83
sd	101.92	84.61	17.31
cv	63.02	67.75	4.73
minimum	38.38	36.31	2.06
maximum	399.06	399.06	0
range	360.68	362.74	2.06
total	2263.94	3621.62	1357.68

Figure 3.3: (a) Graph of Imputation by Cubic Spline for soybean production values in Ipiranga do Norte. (b) Boxplot for the comparison between observed and imputed values. (c) Table of the main descriptive measures of observed and imputed values.

Source: Results of Research.





(c)	OBSERVED	IMPUTED	DIFFERENCE
n	14	29	15
average	161.71	124.88	36.83
sd	101.92	84.61	17.31
cv	63.02	67.75	4.73
minimum	38.38	36.31	2.06
maximum	399.06	399.06	0
range	360.68	362.74	2.06
total	2263.94	3621.62	1357.68

Figure 3.4: (a) Graph of Imputation by Cubic Spline for values of derivatives of soybean production in Ipiranga do Norte. (b) Boxplot for the comparison between observed and imputed values. (c) Table of the main descriptive measures of observed and imputed values.

Source: Results of Research.

It is observed that when imputing 15 initial values from 1990 to 2004, the performance of these series is similar to that of the series from 2005 to 2018, in which the comparison of observed series with the observed series with the values imputed by Quenouille test revealed equivalence of the same series, and at the end of the process there was an increase of 49.03%, 40.48% and 37.49% in grain production, in production value and in value of soybean derivatives in the city, respectively. This fact can be expected, as the city acquires its emancipation only in 2000 and only in 2005 has its first individual management, which apart it from Tapurah. This event may have generated the absence of data in the initial periods, and data may have been collected and accumulated by the city or by several other cities around Ipiranga do Norte, in that time interval. Thus, for the 46 analyzed cities in the three related variables, the growth trend of the total values was observed after the imputation and the presence of missing data in the initial years.

Next, Table 3.3 summarizes the results of the comparison tests of observed data against observed data added to imputed data, for the variable soybean production in the state of Mato Grosso/Brazil.

Table 3.3: Comparison tests of the 46 municipalities in the state of Mato Grosso that need imputation, for Grain Production variable (thousand tons) during 1990 to 2018

ID	μ_o	μ_{o+i}	T_o	T_{o+i}	$\%A_{o+i}$	SQ	$p-value$	F
13	516,743	489,445	7,234,407	14,193,903	50.97	0.01	1.00	2000
66	185,646	159,048	2,599,039	4,612,389	77.47	0.00	0.99	2000
23	0.00	0.00	0.00	0.00	0.00	0.00	0.99	1998
25	415,878	379,943	7,485,801	11,018,335	47.19	0.00	0.99	1999
51	4,060	2,873	73,083	83,307	13.99	0.15	0.99	1999
52	0.00	0.00	0.00	0.00	0.00	0.00	0.99	1998
61	3,503	3,015	63,049	87,431	38.67	0.00	0.99	1999
91	152,321	108,818	2,741,776	3,155,708	15.10	0.00	0.99	1999
100	360,907	346,886	6,496,326	10,059,696	54.85	0.00	0.99	1998
112	5,365	4,664	96,576	135,246	40.04	0.00	0.99	1999
116	90	150	1,620	4,336	167.70	1.12	0.99	1998
128	681	422,90	12,264	12,264,004	3.26e-05	0.00	0.99	1999
138	12,818	8,854	230,727	256,780	11.29	0.00	0.99	1998
139	29,271	209,7811	526,886	608,376	15.47	0.00	0.99	1999
140	15,692	11,749	282,463	340,712	20.62	0.00	0.99	1999
95	220,313	165,594	4,626,566	4,802,223	3.80	0.00	0.99	1997
16	18,399	14,221	404,771	412,407	1.89	0.00	0.99	1995
20	973,801	891,945	21,423,620	25,866,393	20.74	0.00	0.99	1994
30	45,633	34,626	1,003,924	1,004,161	0.02	0.00	0.99	1995
33	4,247	3,238	93,440	93,905	0.50	0.00	0.99	1994
47	22,891	17,606	503,610	510,571	1.38	0.00	0.99	1995
55	144,022	109,612	3,168,491	3,178,755	0.32	0.00	0.99	1995
77	673,071	549,089	14,807,565	15,923,572	7.54	0.00	0.99	1995
88	493,949	450,479	10,866,887	13,063,889	20.22	0.00	0.99	1994
3	195,291	175,088	5,077,571	5,077,554	3.34e-04	0.02	0.99	1991
6	10,386	9,464	270,035	274,444	1.63	0.00	0.99	1991
8	26,050.15	23,355.31	677,304	677,303.97	4.43E-8	0.00	0.99	1991
26	44	39.31	1,140	1,140	0.00	0.00	0.99	1991
44	36.88	45.31	959	1,314.12	37.03	0.10	0.99	1991
49	14,320	12,839	372,331	372,331.01	3.00e-08	0.00	0.99	1991

Table 3.3: Continued

ID	μ_o	μ_{o+i}	T_o	T_{o+i}	$\%A_{o+i}$	SQ	$p-value$	F
56	129,307	116,175	3,361,983	3,369,065	0.21	0.05	0.99	1991
57	7,212.73	6,467.13	187,505	187,547	0.02	0.01	0.99	1991
59	24,642	22,093	640,697	640,697	0.00	0.00	0.99	1991
67	687	616.21	17,870	17,870.04	2.24e-04	0.00	0.99	1991
71	957.08	858.07	24,884	24,884	0.00	0.00	0.99	1991
73	31,222	28,091	811,778	814,651	24.39	0.06	0.99	1991
81	224,947	202,185	5,848,631	5,863,360	0.25	0.06	0.99	1991
85	0.00	0.00	0.00	0.00	0.00	0.00	0.99	1993
87	295.77	264.28	7,664	7,664	0.00	0.00	0.99	1991
90	38,306	36,652	995,943	1,062,898	6.72	0.00	0.99	1991
99	447,082	402,102	11,624,141	11,660,967	0.30	0.08	0.99	1991
102	43	38	1,080	1,080	0.00	0.00	0.99	1991
119	542,731	48,658.30	1,411,091	1,411,090.70	2.13e-05	0.01	0.99	1991
124	397	356	10,326	10,326	0.00	0.00	0.99	1991
134	1,278	1,145	33,219	33,219	0.00	0.00	0.99	1991
141	15,762	14,167	409,819	410,840	0.25	0.00	0.99	1991

Source: Results of Research.
Notes:
ID: municipalities identification; μ_o: average of observed values;
μ_{o+i}: average of observed added imputed values; T_o: observed total;
T_{o+i}: total of observed added imputed; $\%A_{o+i}$: percentage of total increase;
SQ: statistic of Quenouille test; $p-value$: probability of the test;
F: foundation year of the municipalities.

It is noticed that observed series, when compared with observed series added to imputed values for each of the 46 cities, indicate equivalence in their training processes; it is verified, also, after values imputation, a decrease of averages and an increase of total values, in the majority of the cities, for soybean production. Figure 3.5, reveals the spatial distribution of accumulated production from 1990 to 2018 for observed values and for observed values added the imputed ones.

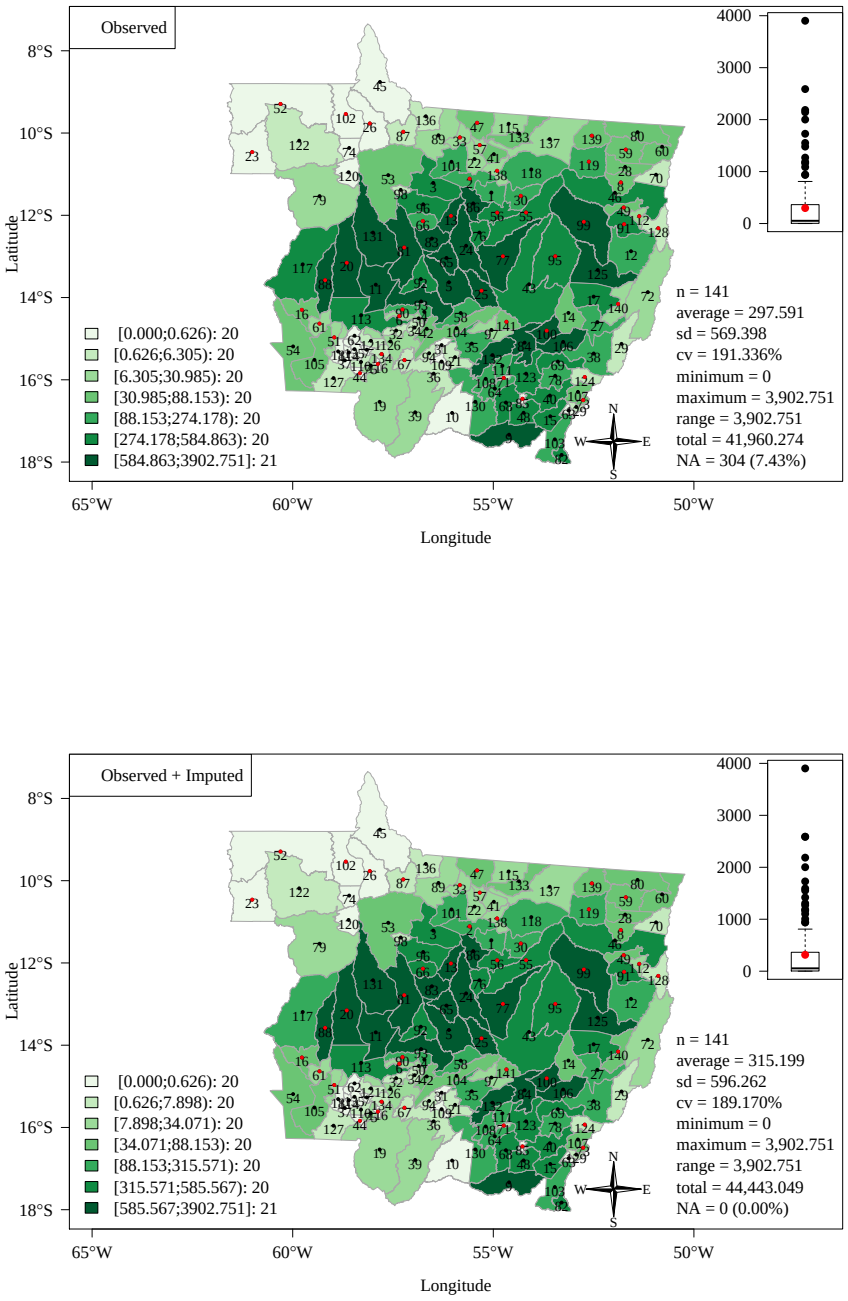


Figure 3.5: Spatial distribution of observed and observed added the imputed values from 1990 to 2018 in Mato Grosso for soybean production in grain accumulated in a thousand tons.

Source: Results of Research.

It is observed that spatial distribution of accumulated soybean production in the state of Mato Grosso, during the period from 1990 to 2018, remains the same. However, an increase of 5.92% is observed in the average and in the total accumulated production value of the 141 cities. It is also identified that all cities were founded during the period of missing data on soybean production; another important fact was that after the foundation

of the city, its installation took place, which took three to five years to occur in most locations.

The performance of observed data comparison tests against observed data added to those imputed for the variable value of soybean production in the state of Mato Grosso, is described in Table 3.4.

Table 3.4: Comparison tests of the 46 municipalities in the state of Mato Grosso that need imputation, for the variable value of soybean production (thousand reais (R$)) during 1990 to 2018

ID	μ_o	μ_{o+i}	T_o	T_{o+i}	$\%A_{o+i}$	SQ	$p-value$	F
13	348,916	283,022	4,884,823	8,207,647	68.02	0.31	0.99	2000
66	127,955	94,291	1,791,371	20,734,431	52.64	0.00	0.99	2000
23	0.00	0.00	0.00	0.00	0.00	0.00	0.99	1998
25	261,126	209,351	4,700,262	6,071,174	29.17	0.00	0.99	1999
51	3,278	2,171	59,003	62,945	6.68	0.10	0.99	1999
52	0.00	0.00	0.00	0.00	0.00	0.00	0.99	1998
61	2,860	2,141	51,485	62,093	20.60	0.00	0.99	1999
91	115,019	76,894	207,034	2,229,926	7.71	0.00	0.99	1999
100	237,795	199,050	4,280,317	5,772,462	34.86	0.00	0.99	1998
112	4,444	3,437	80,000	99,670	24.59	0.00	0.99	1999
116	34	57	615	1,642	166.99	0.98	0.99	1998
128	620	384.83	11,160	11,160.004	4.00e-05	0.00	0.99	1999
138	10,560	6,828	190,082	198,005	4.17	0.00	0.99	1998
139	25,912	17,119	466,409	496,447	6.44	0.00	0.99	1999
140	13,225	9,142	238,045	265,109	11.37	0.00	0.99	1999
95	153,294	112,476	3,219,167	3,261,813	1.32	0.00	0.99	1997
16	15,451	11,770	339,932	341,329	0.41	0.00	0.99	1995
20	595,138	484,384	13,093,039	14,047,143	7.29	0.00	0.99	1994
30	35,554	26,975	782,186	782,272	0.01	0.00	0.99	1995
33	4,037	3,066	88,812	88,924	0.13	0.00	0.99	1994
47	18,181	13,873	399,979	402,305	0.58	0.00	0.99	1995
55	104,394	79,296	2,296,672	2,299,582	0.13	0.00	0.99	1995

Table 3.4: Continued

ID	μ_o	μ_{o+i}	T_o	T_{o+i}	$\%A_{o+i}$	SQ	$p-value$	F
77	448,812	349,409	9,873,869	10,132,873	2.62	0.00	0.99	1995
88	294,462	240,994	6,478,162	6,988,814	7.88	0.00	0.99	1994
3	133,237	119,454	3,464,162	3,464,159	8.66e-05	0.02	0.99	1991
6	8.930	8.024	232.181	232.703	0.22	0.00	0.99	1991
8	21,958	19,686.59	570,911	570,911	0.00	0.00	0.99	1991
26	32	29	828	828	0.00	0.00	0.99	1991
44	17	17	452	501	10.91	0.00	0.99	1991
49	12,456	11,168	323,861	323,861	0.00	0.00	0.99	1991
56	86,704	77,758	2,254,307	2,254,996	0.03	0.03	0.99	1991
57	5,652	5,067.20	146,944	146,948.90	3.34e-03	0.01	0.99	1991
59	23,836	21,370	619,727	619,727	0.00	0.00	0.99	1991
67	352	316	9,161	9,161	0.00	0.00	0.99	1991
71	961	861.31	24,978	24,978	0.00	0.00	0.99	1991
73	23,117.80	20,738.20	601,063.60	601,407.20	0.06	0.01	0.99	1991
81	157,691	141,418	4,099,960	4,101,131	0.03	0.01	0.99	1991
85	0.00	0.00	0.00	0.00	0.00	0.00	0.99	1993
87	283	254	7,370	7,370	0.00	0.00	0.99	1991
90	22,923	20,734	595,999	601,280	0.89	0.00	0.99	1991
99	321,710	885,562	8,364,449	8,368,132	0.04	0.01	0.99	1991
102	40	36	1,036	1,036	0.00	0.00	0.99	1991
119	63,181	56,645	1,642,699	1,642,699	0.00	0.01	0.99	1991
124	385	345	10,326	10,003	1.003	0.00	0.99	1991
134	654	586	17,008	17,008	0.00	0.00	0.99	1991
141	21,658	19,694	563,118	571,131	1.00	0.01	0.99	1991

Source: Results of Search.
Notes:
ID: municipalities identification; μ_o: average of observed values;
μ_{o+i}: average of observed added imputed values; T_o: observed total;
T_{o+i}: total of observed added imputed; $\%A_{o+i}$: percentage of total increase;
SQ: statistic of Quenouille test; $p-value$: probability of the test;
F: foundation year of the municipalities.

It is noticed that observed series, when compared with observed series added to imputed values for each of the 46 cities, also point to equivalence in their training processes. It is verified, in the same way, after imputation of values a decrease of averages and an increase of the total values for the value of soybean production in most cities. Figure 3.6 reveals spatial distribution of accumulated soybean production value from 1990 to 2018 for observed values and for observed values added to imputed ones.

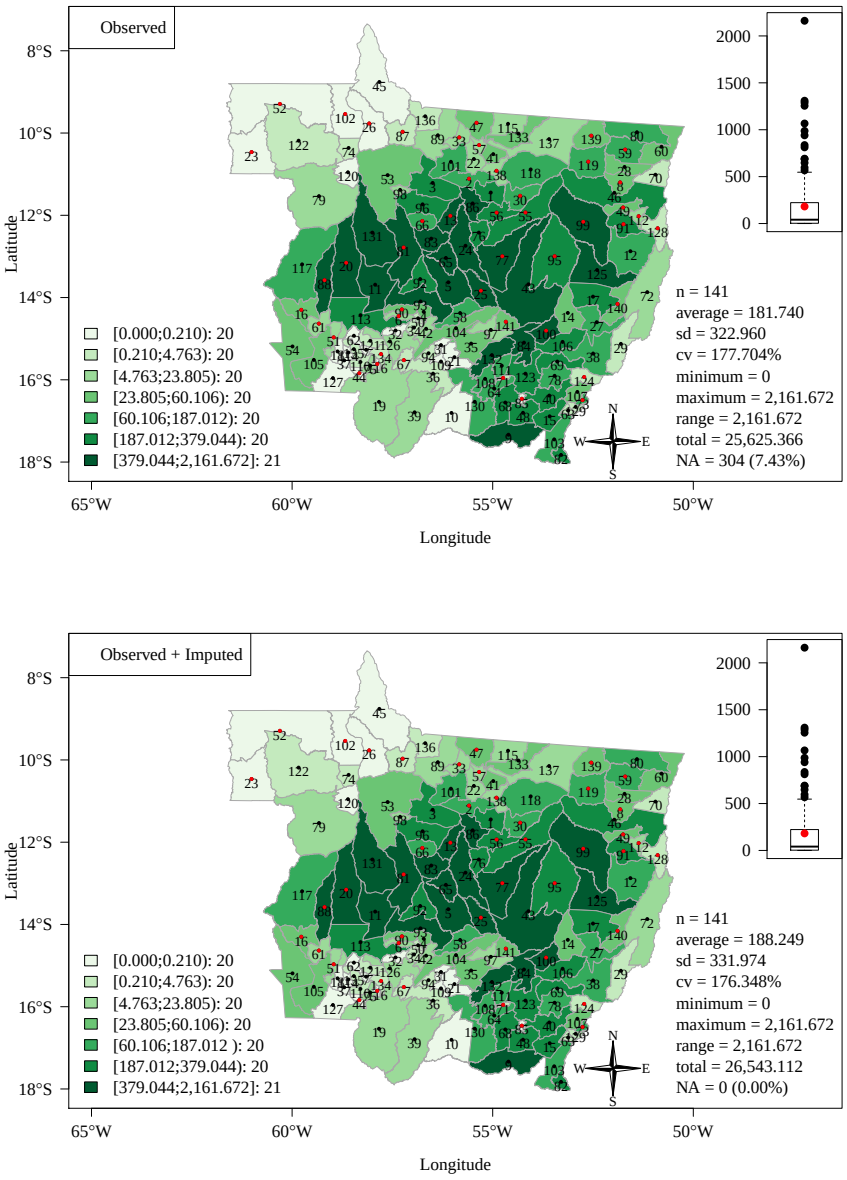


Figure 3.6: Spatial distribution of observed values and observed values added to imputed ones from 1990 to 2018 in Mato Grosso for the value of soybean production in grain accumulated in thousand reais (R$).

Source: Results of Research.

The spatial distribution of accumulated production value of soybean in the state of Mato Grosso, from 1990 to 2018, remains the same after missing data imputation. However, there is an increase of 3.58% in the average and in the total accumulated production value of the 141 municipalities. Table 3.5 describes the performance of comparison tests of observed data against observed data added to those imputed for the variable production value of soybean derivatives in the state of Mato Grosso, is described in Table 3.5. .

Table 3.5: Comparison tests of the 46 municipalities in the state of Mato Grosso that need imputation, for the variable production value of soybean derivatives (thousand reais (R$)) during 1990 to 2018

ID	μ_o	μ_{o+i}	T_o	T_{o+i}	$\%A_{o+i}$	SQ	$p-value$	F
13	161,710	124,884	2,263,938	3,621,623	59.97	0.63	0.99	2000
66	32,867	28,723	460,138	832,977	81.03	0.00	0.99	2000
23	2,249	2,181	40,477	63,238	56.23	0.02	0.99	1998
25	115,436	86,648	2,077,851	2,512,797	20.93	0.00	0.99	1999
51	1,378	1,081	24,804	31,360	26.43	0.00	0.99	1999
52	35,154	30,945	632,775	897,399	41.82	0.00	0.99	1998
61	1,879	2,038	33,815	59,101	74.78	1.29	0.99	1999
91	32,052	23,164	576,937	671,756	16.43	0.00	0.99	1999
100	136,302	112,379	2,453,427	3,258,991	32.83	0.00	0.99	1998
112	1,958	1,875	35,242	54,386	54.32	1.43	0.99	1999
116	10,592	7,202	190,649	208,849	9.55	0.00	0.99	1998
128	1,411	1,300	25,403	37,707	48.44	0.00	0.99	1999
138	8,508	6,308	153,142	182,934	19.45	0.00	0.99	1998
139	7,706	6,404	138,705	185,722	33.90	0.00	0.99	1999
140	3,097	2,808	55,747	81,429	46.07	0.00	0.99	1999
95	20,677	16,173	434,209	469,029	8.02	0.00	0.99	1997
16	4,565	3,687	100,431	106,910	6.45	0.00	0.99	1995
20	578,269	454,857	12,721,908	13,190,866	3.67	0.00	0.99	1994
30	14,638	11,306	322,041	327,880	1.81	0.00	0.99	1995
33	7,069	6,414	155,522	186,012	19.60	0.00	0.99	1994
47	14,911	12,975	328,048	376,276	14.70	0.00	0.99	1995
55	40,796	31,306	897,520	907,882	1.16	0.00	0.99	1995
77	174,459	135,885	3,838,102	3,940,651	2.67	0.00	0.99	1995
88	228,054	176,594	5,017,187	5,121,233	2.07	0.00	0.99	1994
3	59,011	52,942	1,534,297	1,535,330	0.07	0.01	0.99	1991
6	7,060	6,371	183,561	184,758	0.65	0.08	0.99	1991
8	11,501	10,704	299,021	310,415	3.81	0.00	0.99	1991
26	7,878	7,397	204,837	214,502	4.72	0.01	0.99	1991
44	4,165	4,062	108,295	117,787	8.76	0.28	0.99	1991
49	8,034	7,309	208,880	211,974	1.48	0.02	0.99	1991
56	38,756	34,791	1,007,669	1,008,936	0.13	0.01	0.99	1991
57	4,825	4,441	125,453	128,778	2.65	0.01	0.99	1991
59	27,256	24,906	708,652	722,279	1.92	0.14	0.99	1991
67	5,016	4,594	130,406	133,234	2.17	0.00	0.99	1991
71	3,698	3,456	96,140	100,214	4.24	0.00	0.99	1991
73	5,406	4,874	140,569	141,332	0.54	0.00	0.99	1991
81	32,095	28,864	834,478	837,068	0.31	0.00	0.99	1991

Table 3.5: Continued

ID	μ_o	μ_{o+i}	T_o	T_{o+i}	$\%A_{o+i}$	SQ	$p-value$	F
85	1,982	1,919	51,535	55,639	7.96	0.00	0.99	1993
26	7,878	7,397	204,837	214,502	4.72	0.01	0.99	1991
44	4,165	4,062	108,295	117,787	8.76	0.28	0.99	1991
49	8,034	7,309	208,880	211,974	1.48	0.02	0.99	1991
56	38,756	34,791	1,007,669	1,008,936	0.13	0.01	0.99	1991
57	4,825	4,441	125,453	128,778	2.65	0.01	0.99	1991
59	27,256	24,906	708,652	722,279	1.92	0.14	0.99	1991
67	5,016	4,594	130,406	133,234	2.17	0.00	0.99	1991
71	3,698	3,456	96,140	100,214	4.24	0.00	0.99	1991
73	5,406	4,874	140,569	141,332	0.54	0.00	0.99	1991
81	32,095	28,864	834,478	837,068	0.31	0.00	0.99	1991
85	1,982	1,919	51,535	55,639	7.96	0.00	0.99	1993
87	4,051	3,746	105,325	108,645	3.15	0.07	0.99	1991
90	6,679	6,001	173,642	174,025	0.22	0.01	0.99	1991
99	61,820	55,480	1,607,324	1,608,916	0.10	0.02	0.99	1991
102	7,746	7,103	201,401	205,997	2.28	0.00	0.99	1991
119	21,183	19,069	550,758	553,011	0.41	0.01	0.99	1991
124	1,688	1,541	43,891	44,690	1.82	0.00	0.99	1991
134	26,071	23,666	677,842	686,325	1.25	0.00	0.99	1991
141	6,663	6,221	173,248	180,414	4.14	0.02	0.99	1991

Source: Results of Search.

Notes:

ID: municipalities identification; μ_o: average of observed values;
μ_{o+i}: average of observed added imputed values; T_o: observed total;
T_{o+i}: total of observed added imputed; $\%A_{o+i}$: percentage of total increase;
SQ: statistic of Quenouille test; $p-value$: probability of the test;
F: foundation year of the municipalities.

It is observed that observed series, when compared with observed series added to imputed values for each of the 46 cities, reveal signs of equivalence in their training processes; and it is also verified, after imputation of values, a decrease of averages and an increase of total values for production value of soybean derivatives in most cities. Figure 3.7 shows the spatial distribution of production value of soybean derivatives accumulated from 1990 to 2018 for observed and observed values added the imputed ones.

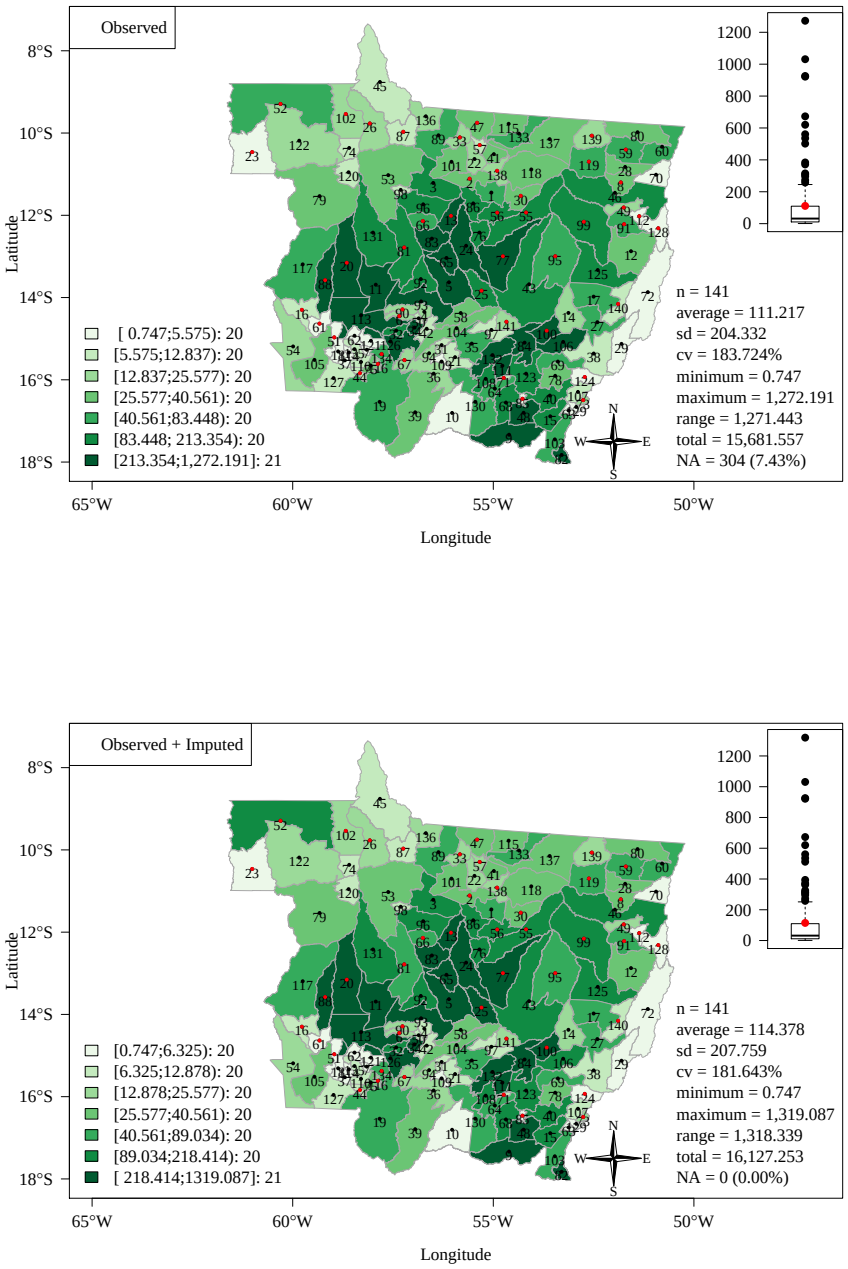


Figure 3.7: Spatial distribution of observed and observed added imputed values from 1990 to 2018 in Mato Grosso for production value of soybean derivatives in grain accumulated in thousand reais (R$).

Source: Results of Research.

It is observed that the spatial distribution of accumulated production value of soybean derivatives in the state of Mato Grosso, from 1990 to 2018, remains the same after missing data imputation. However, there is an increase of 2.84% in average and in total value of accumulated production of derivatives in Mato Grosso.

In addition, comparison tests were made between the three accumulated variables before and after imputation in the 141 municipalities in the state, as shown in Table 3.6.

Table 3.6: Tests comparing the variables accumulated in the 141 municipalities during 1990 to 2018, before and after imputation: 1) grain production (thousand tons), 2) production value of soybeans (thousand reais (R$)) and 3) value of production of soybean derivatives (thousand reais (R$))

V	μ_o	μ_{o+i}	md_o	md_{o+i}	T_o	T_{o+i}	$\%A_{o+i}$	W	$p-value$
1	2,975,906	3,151,989	526,886	546,099	419,602,745	444,430,491	5.92	9,868	0.92
2	1,817,402	1,882,490	409,819	410,839,90	256,253,656	265,431,117	3.58	9,881	0.93
3	1,112,167	1,143,777	526,886	546,099	156,815,568	161,272,532	2.84	9,784.50	0.82

Source: Results of Research.

Notes:

V: compared variables; μ_o: average of observed values;

μ_{o+i}: average of observed added imputed values; md_o: median of observed values;

md_{o+i}: median of observed values added imputed values; T_o: observed total;

T_{o+i}: total of observed added imputed; $\%A_{o+i}$: percentage of total increase;

W: statistic of Wilcoxon; $p-value$: probability of the test;

It is proved that observed variables accumulated in the 141 cities, when compared with observed variables added accumulated imputed values, also seem to be equivalent both in their averages and in their distributions.

3.6 Conclusions

The analysis of observed data, compared to observed values added to imputed ones, for production variables, production value and value of soybean derivatives for each of the municipalities, was effective, a fact demonstrated by Quenouille test that indicated the series are the same in their entirety for each of the 46 municipalities with no data. Wilcoxon test also showed evidences of equivalence in the same accumulated variables in the 141 municipalities, and after imputation there was an increase in the average and in the total values of 5.92%, 3.58%, 2.84% for the three series, respectively; this fact did not generate a sudden change in spatial distribution of each of the accumulated variables after imputation process was applied.

It was observed that most of imputed series imitate the working of previous series with a smaller scale distribution. Some impute zero, which is an expected occurrence since part of the municipalities did not have skills for soybean production in the early years. It was also found that none of imputed values was greater than observed values, for the three variables in all 46 municipalities.

It is worth noting that missing data occurred in the initial periods from 1990 to 2018 for all variables in the same periods, and most cities did not have the status of municipality, and were still to be set up with a period of three to five years, after its public foundation, which may have caused difficulties in the collection of data linked to soybean

production by IBGE, because a certain municipality is subordinate to some other in this period.

In addition to complete a set of data for 141 municipalities in the state of Mato Grosso, from 1990 to 2018, for variables related to soybean productive area, which are so important for the economic development of the country, this work presents itself as an innovation/recommendation for missing data in these variables for soybean production and for other crops in the state of Mato Grosso, and also in other locations in the country, because there is an estimate of a complete productive representation of this crop, in a certain period of interest.

3.7 References

Acurna, E. and C. Rodriguez, 2004 The treatment of missing values and its effect in the classifier accuracy, classification, clustering, and data mining applications. In *Proceedings of the Meeting of the International Federation of Classification Societies (IFCS)*, pp. 639–647.

Atkinson, K. E., 2008 *An introduction to numerical analysis*. John wiley & sons.

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.

Batista, G. E. A. P. A. and M. C. Monard, 2003 An analysis of four missing data treatment methods for supervised learning. Applied artificial intelligence **17**: 519–533.

Buuren, S. V. and K. Groothuis-Oudshoorn, 2010 mice: Multivariate imputation by chained equations in R. Journal of statistical software pp. 1–68.

Buuren, V. S., 2018 *Flexible imputation of missing data*. CRC press.

Conte, S. D. and C. De Boor, 2017 *Elementary numerical analysis: an algorithmic approach*. SIAM.

Costa, F. M. and T. Sáfadi, 2010 Comparação Estatística de duas séries de material particulado (MP10) na cidade de São Paulo. Rev. Bras. Biom **28**: 23–38.

Dahlquist, G. and Â. Björck, 2008 *Numerical methods in scientific computing, volume I.*. SIAM.

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.

Echeverry, G. and C. d. C. Toloi, 2000 Testes para comparação de séries temporais: uma aplicação a séries de temperatura e salinidade da água, medidas em profundidades diferentes. Rev. Bras. Estat pp. 51–80.

Enders, C. K., 2010 *Applied missing data analysis*. Guilford press.

Engels, J. M. and P. Diehr, 2003 Imputation of missing longitudinal data: a comparison of methods. Journal of clinical epidemiology **56**: 968–976.

Farhangfar, A., L. A. Kurgan, and W. Pedrycz, 2007 A novel framework for imputation of missing values in databases. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans **37**: 692–709.

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 .

Forsythe, G. E., M. A. Malcolm, and C. B. Moler, 1977 *Computer methods for mathematical computations*, volume 11. Englewood Cliffs, New Jersey. Prentice Hall, Inc.

Gandolfi, M., 2016 *Imputação múltipla via algorítmo MICE e método IMLD*. Master's thesis, Universidade Estadual de Maringá.

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

Harrell Jr, F. E., 2015 *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

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.

Hawthorne, G. and P. Elliott, 2005 Imputing cross-sectional missing data: comparison of common techniques. Australian & New Zealand Journal of Psychiatry **39**: 583–590.

Honaker, J. and G. King, 2010 What to do about missing values in time-series cross-section data. American journal of political science **54**: 561–581.

Junger, W. and A. P. De Leon, 2015 Imputation of missing data in time series for air pollutants. Atmospheric Environment **102**: 96–104.

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.

Little, R. J. A., 1988 A test of missing completely at random for multivariate data with missing values. Journal of the American statistical Association **83**: 1198–1202.

Little, R. J. A. and D. B. Rubin, 2002 *Statistical analysis with missing data*. John Wiley & Sons.

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

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.

Perneger, T. V. and B. Burnand, 2005 A simple imputation algorithm reduced missing data in SF-12 health surveys. Journal of clinical epidemiology **58**: 142–149.

Prenter, P. M., 2008 *Splines and variational methods*. Courier Corporation.

Press, W. H., B. Flannery, S. Teukolsky, and W. Vetterling, 1986 *Numerical Recipices, the Art of Scientific Computing*.

Quenouille, M., 1958 The comparison of correlations in time-series. Journal of the Royal Statistical Society: Series B (Methodological) **20**: 158–164.

R Core Team, 2021 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ragel, A., 2000 A preprocessing method to treat missing values in knowledge discovery in databases. Computing and Information Systems **7**: 66–72.

Rubin, D. B., 1976 Inference and missing data. Biometrika **63**: 581–562.

Rubin, D. B., 1978 Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, volume 1, pp. 20–34, American Statistical Association.

Rubin, D. B., 2004 *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Rubin, D. B. and N. Schenker, 1991 Multiple imputation in health-are databases: An overview and some applications. Statistics in medicine **10**: 585–598.

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

Schafer, J. L. and J. W. Graham, 2002 Missing data: our view of the state of the art. Psychological methods **7**: 147.

Schultz, M. H., 1973 *Spline analysis*. Prentice-Hall.

Souza, V. J. C. and C. Villegas, 2020 Generalized symmetrical partial linear model. Journal of Applied Statistics pp. 1–16.

Spratt, M., J. Carpenter, J. A. C. Sterne, J. B. Carlin, J. Heron, J. Henderson, and K. Tilling, 2010 Strategies for multiple imputation in longitudinal studies. American journal of epidemiology **172**: 478–487.

Twisk, J. and W. Vente, 2002 Attrition in longitudinal studies: how to deal with missing data. Journal of clinical epidemiology **55**: 329–337.

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.

Verbyla, A. P., B. R. Cullis, M. G. Kenward, and S. J. Welham, 1999 The analysis of designed experiments and longitudinal data by using smoothing splines. Journal of the Royal Statistical Society: Series C (Applied Statistics) **48**: 269–311.

White, I. M. S., R. Thompson, and S. Brotherstone, 1999 Genetic and environmental smoothing of lactation curves with cubic splines. Journal of Dairy Science **82**: 632–638.

Wilcoxon, F., 1945 Individual comparisons by ranking methods. Biometrics **1**: 80–83.

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.

4 ZONING OF SOYBEAN PRODUCTION IN THE STATE OF MATO GROSSO, BRAZIL, FROM 1990 TO 2018, VIA CLUSTER ANALYSIS

4.1 Resumo

O estado de Mato Grosso se apresenta como maior produtor e exportador de soja do Brasil. Desta forma este trabalho tem por objetivo utilizar a técnica análise de *clusters* para criar grupos e ranqueá-los nos 141 municípios do estado de Mato Grosso, de 1990 a 2018, para as variáveis de produção de soja em mil toneladas, valor de produção em mil reais (R$) e valor de derivados em mil reais (R$), para isto utilizou-se em cada destas variáveis a aplicação da distância DTW (*Dinamic Time Warp*), aliada ao método de agrupamento de Ward para criação dos grupos, criando-se três estruturas de dendrogramas para as mesmas, e desta maneira ocorreu a validação do número adequado de grupos pelo método de Mojena nas estruturas de dendrogramas. Assim, a partir dos dendrogramas construídos, foram criados 5, 5 e 4 grupos de forma escalonada para cada variável, e respectivamente averiguou-se que os mesmos possuem tendência de crescimento em cada uma destas variáveis. Observou-se que o desenvolvimento dessas variáveis no estado se fez dos municípios centrais para os mais periféricos e margeando as principais rodovias que cortam o estado. E por final com intuito de melhorar as estimativas da análise proposta, estes grupos criados também foram validados estatisticamente pelos testes de Pearson e de Mantel.

Palavras-chave: Produção; municípios similares; avaliação das estimativas.

4.2 Abstract

The state of Mato Grosso is the largest producer and exporter of soybean in Brazil. Thus, this work aimed to use the *clusters* analysis technique to create groups and rank them in the 141 municipalities of the state of Mato Grosso, from 1990 to 2018, for the soybean production variables in thousand tons, value of production in thousand reais(R$) and value of derivatives in thousand reais(R$), for this the application of the DTW distance (Dynamic Time Warp), together with the Ward's grouping method, was used to create the groups, creating three structures of dendrograms for them, and in this way the validation of the adequate number of groups by the method of Mojena in the structures of dendrograms occurred. Thus, from the dendrograms constructed, 5, 5 and 4 groups were created in a staggered manner for each variable, and respectively it was found that they have a trend of increase in each of these variables. It was observed that the development of these variables in the state was made from the central municipalities to the more peripheral ones and bordering the main highways that cross the state. Finally,

in order to improve the estimates of the proposed analysis, these groups created were also statistically validated by Pearson and Mantel tests.

Keywords: Production; similar municipalities; estimate evaluations.

4.3 Introduction

Brazil is the leading producer and exporter of soybeans in the world, with super harvests year after year in this activity, it is worth mentioning the state of Mato Grosso, which has been reaching record levels since the year 1990, occupying the first position in the cultivation and export of this crop in the country. Most of this production goes to China, which often buys this large volume even before its harvest, to feed its huge population and also for animal nutrition.

Given the importance of the entire complex involved in soybean production for the country and for the state of Mato Grosso, in the financial movement and development of agribusiness, maintenance of internal and external markets and generation of jobs in all its productive structure and others, the main objective of this study is to create and sort from the highest to the lowest relevance groups and to analyze them statistically and validate them, through the application of grouping analysis techniques, in the variables soybean production in thousand tons, soybean production value in thousand reais (R$) and value of soybean derivatives in thousand reais (R$) in the 141 municipalities of the state from 1990 to 2018, and also aims to generate an estimate of the comportment in each of these variables, in this period. This research, in addition to being a scientific innovation in the current research scenario, may also serve as an instrument to be used by the state in the construction of new public policies linked to the development of the crop and may also be another decision making tool for the old and and an attraction for new investors in this sector.

In an analysis of clusters /conglomerates, the aim is to identify and classify the existence of objects, items, or individuals according to their similarity, and from this common characteristic to determine homogeneous groups within a given variable of interest (Bussab *et al.*, 1990; Everitt, 1993; Michaud, 1997; Johnson *et al.*, 2002; Ferreira, 2008; Hair *et al.*, 2009).

The research by Broich and Palmer (1980) make use of clustering techniques to examine phenotypic characteristics of soybean varieties from a USDA (United States Department of Agriculture) germplasm database (United States Department of Agriculture).

Lee *et al.* (2008) use these techniques to evaluate differences in the variety of wild soybeans Glycine, Glycine soja Sieb. and Zucc., produced in South Korea, compared to different varieties produced in other countries.

The investigation of Popović *et al.* (2011) allude to cluster analyses for the

formation of municipal groups linked to Serbia's agribusiness, and the characteristics of each group created were also assessed.

A more recent research by Rong-zhen *et al.* (2020), in addition to using the cluster analysis methodology, other multivariate techniques are used to create groups and compare the photosynthetic capacity of Glycine soybean and other varieties from crossing Glycine soybean x Glycine max.

In this context, section 4.4 was reserved for data presentation and methodological details, and 4.5 and 4.6 are intended for results and conclusions, respectively.

4.4 Material and Methods

4.4.1 Data related to soybean production in the state of Mato Grosso

Data on production, production value and value of soybean derivatives in the state of Mato Grosso were from IBGE (Brazilian Institute of Geography and Statistics), from 1990 to 2018, for 141 municipalities. During the collection, there was an absence of data in 46 municipalities for each of the variables, as illustrated in Figure 4.1

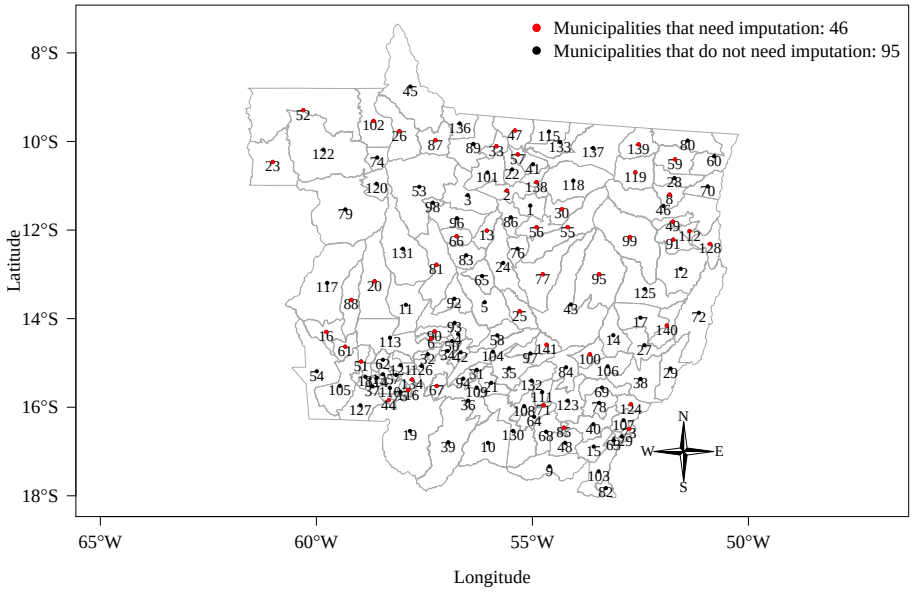


Figure 4.1: Map of the state of Mato Grosso detailing the municipalities that require or not imputation.

Source: IBGE.

To complete these data gaps, the univariate imputation method by cubic splines was used with the fgm procedure developed by Forsythe *et al.* (1977) for each of the 46 municipalities for each of the variables, which made it possible to obtain an estimate of a complete data set for the state of Mato Grosso in this study period. With data in

hand, three input matrices $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ were construced, where x_{ij} are elements with $i = 1990, \ldots, 2018$ years and $j = 1, \ldots, 141$ municipalities of the same, and these represent respectively the observations of each of the variables of soybean production (in thousand tons), value of soybean production (in thousand reais (R$)) and value of production of soybean derivatives (in thousand reais (R$)). The three matrices are expressed by matrix $\mathbf{X}$, described in sentence 4.1 for each of its variables.

$$\mathbf{X} = \begin{bmatrix} x_{1990,1} & x_{1990,2} & \cdots & x_{1990,141} \\ x_{1991,1} & x_{1991,2} & \cdots & x_{1991,141} \\ \vdots & \vdots & \ddots & \vdots \\ x_{2018,1} & x_{2018,2} & \cdots & x_{2018,141} \end{bmatrix} \tag{4.1}$$

The application of all the procedures in the later sections of this study was done with the aid of the R software (R Core Team, 2021).

4.4.2 Procedure for using the DTW (Dinamic Time Warp) distance

In studies using clusters analysis containing time series, the DTW dissimilarity measure has been applied, which has a higher performance than usual, as this metric has the advantage of being corrected by relatively superior invariance methods and more complex than the others, in this case (Berndt and Clifford, 1994; Müller, 2007; Batista *et al.*, 2011; Giusti and Batista, 2013; Rakthanmanon *et al.*, 2013). In this study, we used 29 values of the time series of 141 municipalities in the state of Mato Grosso as arranged in matrix $\mathbf{X}$, in order to capture similarities among them. First, the dissimilarities were calculated for each pair $(\gamma_{ij}, \gamma_{i'j'})$, with $j = j' = 1, ..., 141$ and municipalities and $i = i' = 1990, ..., 2018$ values of the time series, and it functions as a dynamic synchronizer for the distances in the variables of production, production value and value of soybean derivatives, in which S is the set of all possible sequences $s = \{(\gamma_{1990j}, \gamma_{1990j'}) \ldots (\gamma_{ij}, \gamma_{i'j'})\}$. This distance is represented by equation 4.2.

$$d(\gamma_{ij}, \gamma_{i'j'}) = \min_{s \in S} \left(\sum_{t=1}^{s} |\gamma_{tj} - \gamma_{t'j}| \right) \tag{4.2}$$

The use of the DTW distance has the advantage of identifying similar shapes between two time series, even when displaced or with different number of observations.

4.4.3 Ward's grouping method

The clustering method proposed by Ward Jr (1963) is based on the internal variation of the groups that are being formed in each step of the clustering. In this procedure, the variance within the groups should be kept to a minimum, that is, the groups are formed by maximizing the internal homogeneity of the groups, in other words,

two groups R and S are joined, which minimize the sum of the squares of errors (SQE). According to Ferreira (2008), in a cluster analysis, given $k \leq n$ groups and if the $j-th$ object of the $l-th$ group is represented by x_j^l, with $l = 1, \ldots, k$ and $j = 1, \ldots, n_l$, the sum of squares of errors (SQE_l), is defined, according to sentence 4.3.

$$SQE_l = \left(\sum_{j=1}^{n_l} (x_j^l - \bar{x}_\cdot^l)'.(x_j^l - \bar{x}_\cdot^l) \right) \tag{4.3}$$

where n_l is the number of elements of the $l-th$ group and $n = \sum_{l=1}^{k} n_l$. In this way, the groups R and S can be combined in a single group in order to minimize SQE_{RS}, and this process can be represented by $\Delta\, SQE_{RS}$, this junction can be explained by:

$$\Delta\, SQE_{RS} = SQE_{RS} - SQRE_R - SQE_S$$

In which:

1. $$SQE_{RS} = \left(\sum_{j=1}^{n_{RS}} (x_j^{RS} - \bar{x}_\cdot^{RS})'.(x_j^{RS} - \bar{x}_\cdot^{RS}) \right);$$

2. $n_{RS} = n_R + n_S$: size of the RS group;

3. $\bar{x}_\cdot^{RS} = \frac{(n_R.\bar{x}_\cdot^R + n_S.\bar{x}_\cdot^S)}{n_R + n_S}$: centroid of the new RS group.

Similarly, Ward Jr (1963) demonstrates that the $\Delta\, SQE_{RS}$ is agglomeration that can also be expressed by equation 4.4.

$$\Delta\, SQE_{RS} = \left(\frac{n_R.n_S}{n_R + n_S} \right) (\bar{x}_\cdot^R - \bar{x}_\cdot^S)'.(\bar{x}_\cdot^R - \bar{x}_\cdot^S) \tag{4.4}$$

Under certain conditions, there is a relationship between Ward Jr (1963) method and the maximum likelihood method when the distribution of the analyzed variables is normal multivariate in each step of the cluster (Scott and Symons, 1971). It is worth noting that for the application of the Ward Jr (1963) method, there is no need for the data to present a multivariate normal distribution.

4.4.4 Optimal number of groups by the Mojena method

The method of Mojena (1977) indicates that the number of k groups optimizes the quality of the fit of a given conglomerate to the data. Thus, given a defined dendrogram, we calculate:

$$\alpha_j = \bar{\alpha} + \phi.S_\alpha \tag{4.5}$$

where $j = 1, 2, \ldots, 141$ municipalities and α_j refer to the distance value for the fusion stage corresponding to $141 - j + 1$ groups, $\bar{\alpha}$ and S_α represent the mean and the standard deviation of $\alpha' s$ and ϕ is a constant suggested by MILLIGAN and COOPER (1985) as a stop rule based on data simulations as being 1.25. Thus, from obtaining the cutoff point α_j, a horizontal line is drawn at its value perpendicular to the distance axis in the obtained dendrogram, and the formed municipal groups can be visualized.

4.4.5 Cluster structure assessment

From the structure of the dendrogram and its groups, the r_{cof} cophenetic correlation coefficient is computed, which quantifies the similarity of the $c_{jj'}$ (which is the distance in the dendrogram between two municipalities that was clustered, that is, they are the nodes of the dendrogram structure) with the original distance matrix using the DTW metric named $d_{jj'}$. Some authors such as SOKAL and ROHLF (1962) and SILVA and DIAS (2013) show that this coefficient can be calculated by the following sentence 4.6.

$$r_{cof} = \frac{\sum_{j=1}^{141} \sum_{j'<j}^{141-1} (d_{jj'} - \bar{d}).(c_{jj'} - \bar{c})}{\sqrt{\sum_{j=1}^{141} \sum_{j'<j}^{141-1} (d_{jj'} - \bar{d})^2 . \sum_{j=1}^{n} \sum_{j'<j}^{141-1} (c_{jj'} - \bar{c})^2}} \tag{4.6}$$

where n is the dimension of matrices $d_{jj'}$ and $c_{jj'}$, and that

$$\bar{d} = \frac{2}{141(141-1)} \sum_{j=1}^{141} \sum_{j'<j}^{141-1} d_{jj'};$$

$$\bar{c} = \frac{2}{141(141-1)} \sum_{j=1}^{141} \sum_{j'<j}^{141-1} c_{jj'}.$$

These r_{cof} coefficients were calculated for the three variables under analysis in the present study, and we estimated that the higher their value, the less distortions caused by the generated clusters. And later, in order to evaluate the results found, we computed $\rho_{cof} = \sum_{j=1}^{141} \sum_{j'<j}^{141-1} (d_{jj'} - \bar{d})(c_{jj'} - \bar{c})$, then, unilateral statistical tests of correlation of Pearson, described in PEARSON (1904) were applied, to test the hypothesis $H_0 : \rho_{cof} = 0$, which indicates that there is no relationship between the distances $d_{jj'}$ and $c_{jj'}$, at a certain level of significance α, against $H_1 : \rho_{cof} > 0$, which indicates that the distances $d_{jj'}$ and $c_{jj'}$, have a positive relationship and, consequently, the formation of clusters without distortions, producing good results. Finally, the tests of Mantel described in the survey of MANTEL (1967) was applied to the variables, in order to analyze the validity of the same previous hypotheses as the Pearson correlation test, through 10,000 permutations randomized of the two matrices, $d_{jj'}$ and $c_{jj'}$ involved, in this way an empirical distribution of these results is removed, and the unilateral test provides z_{calc} statistics that allows H_0 to be accepted or not.

4.5 Results and Discussion

In possession of the matrices $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ respectively, obtained from each of the variables - production, production value and value of soybean derivatives - the DTW distance was calculated, and then the Ward method was applied and finally the cutoff point was computed by the Mojena method, to determine the number of groups formed. All of these procedures gave rise to three dendrogram structures, which generate similar groups for each of the respective variables, as illustrated in Figures 4.2, 4.3 and 4.4.

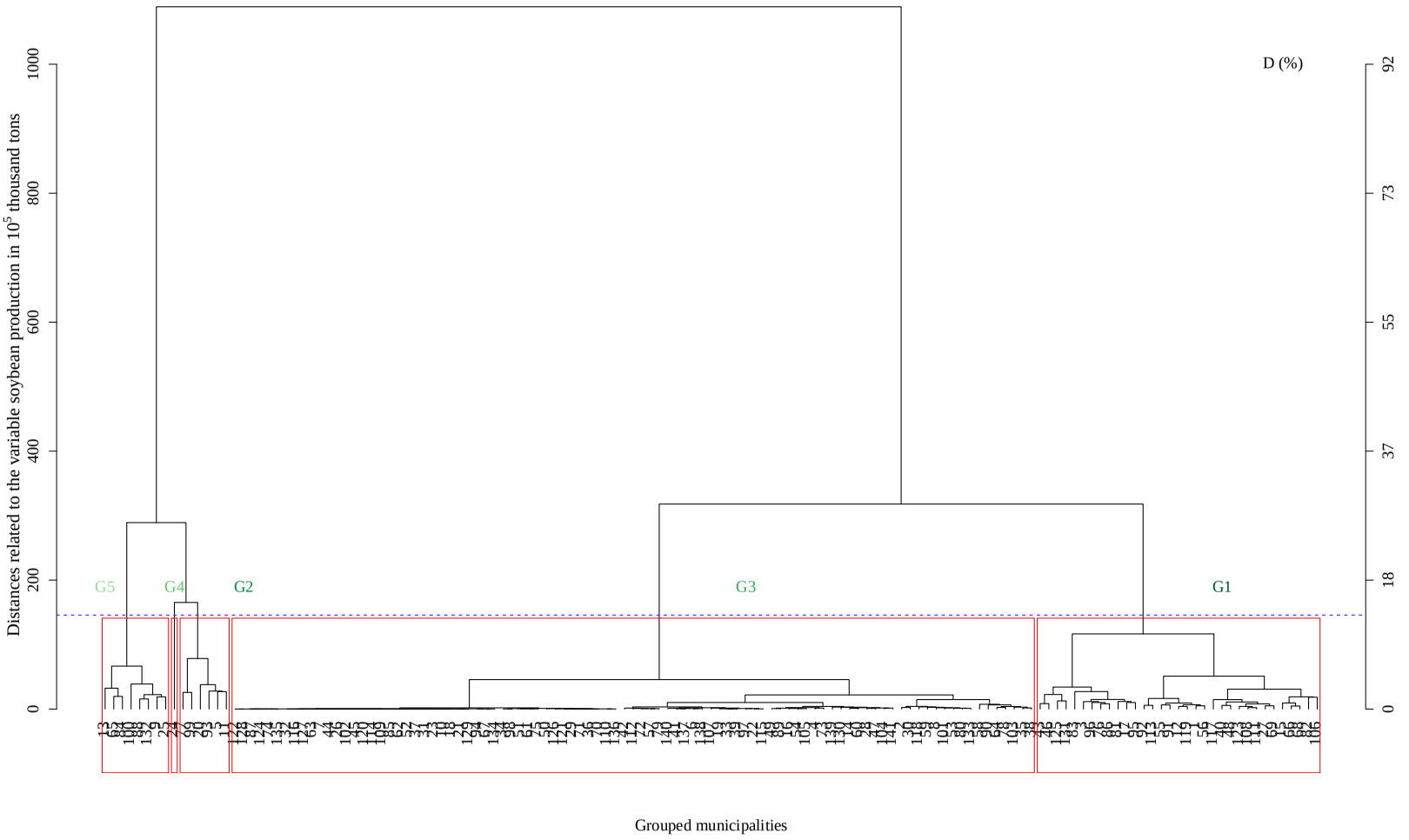


Figure 4.2: Dendrogram constructed for soybean production in 10^5 thousand tons, and its groups formed.

Source: Results of Research.

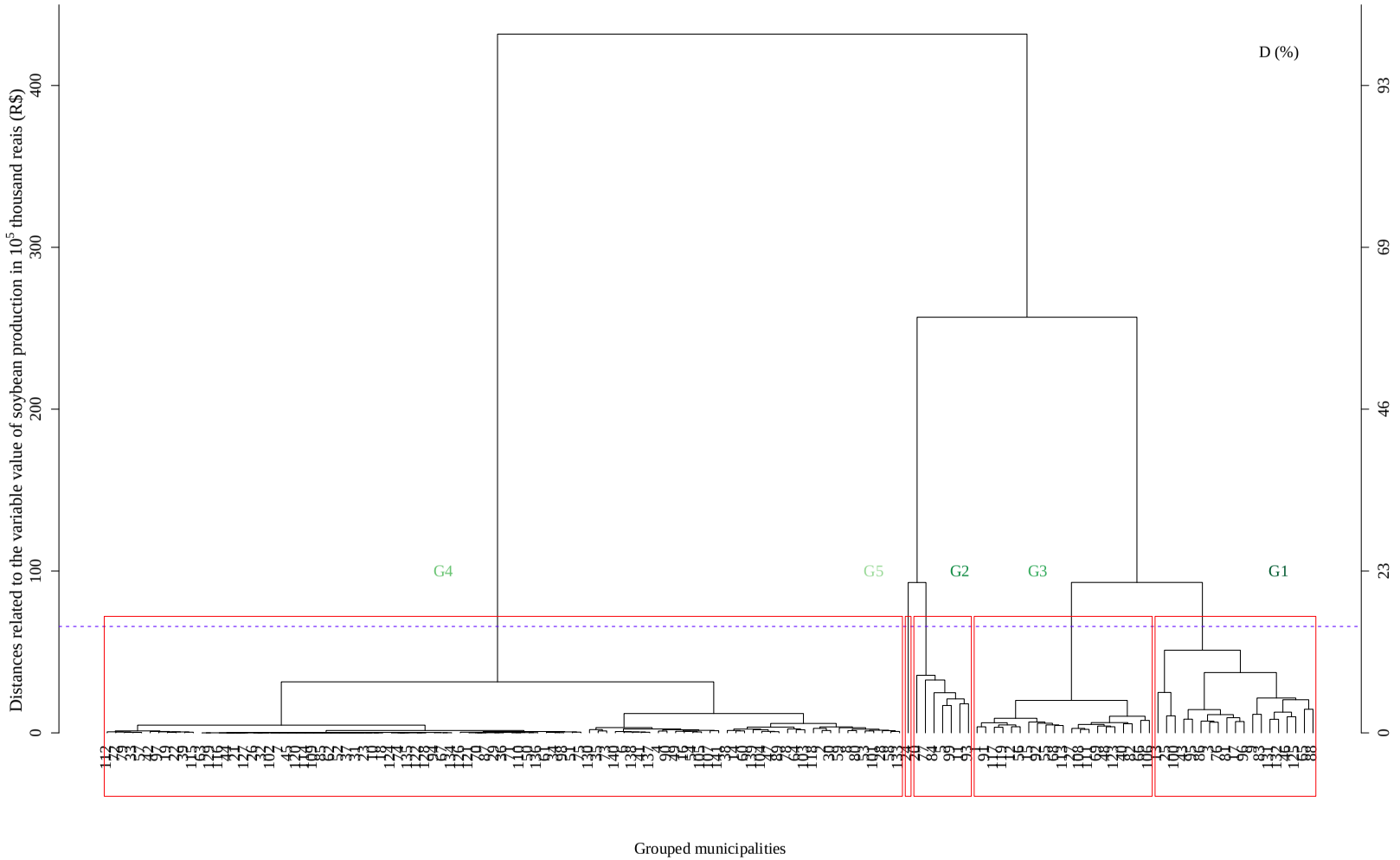


Figure 4.3: Dendrogram constructed for variable soybean grain production value in 10^5 thousand reais (R$), and its groups formed.

Source: Results of Research.

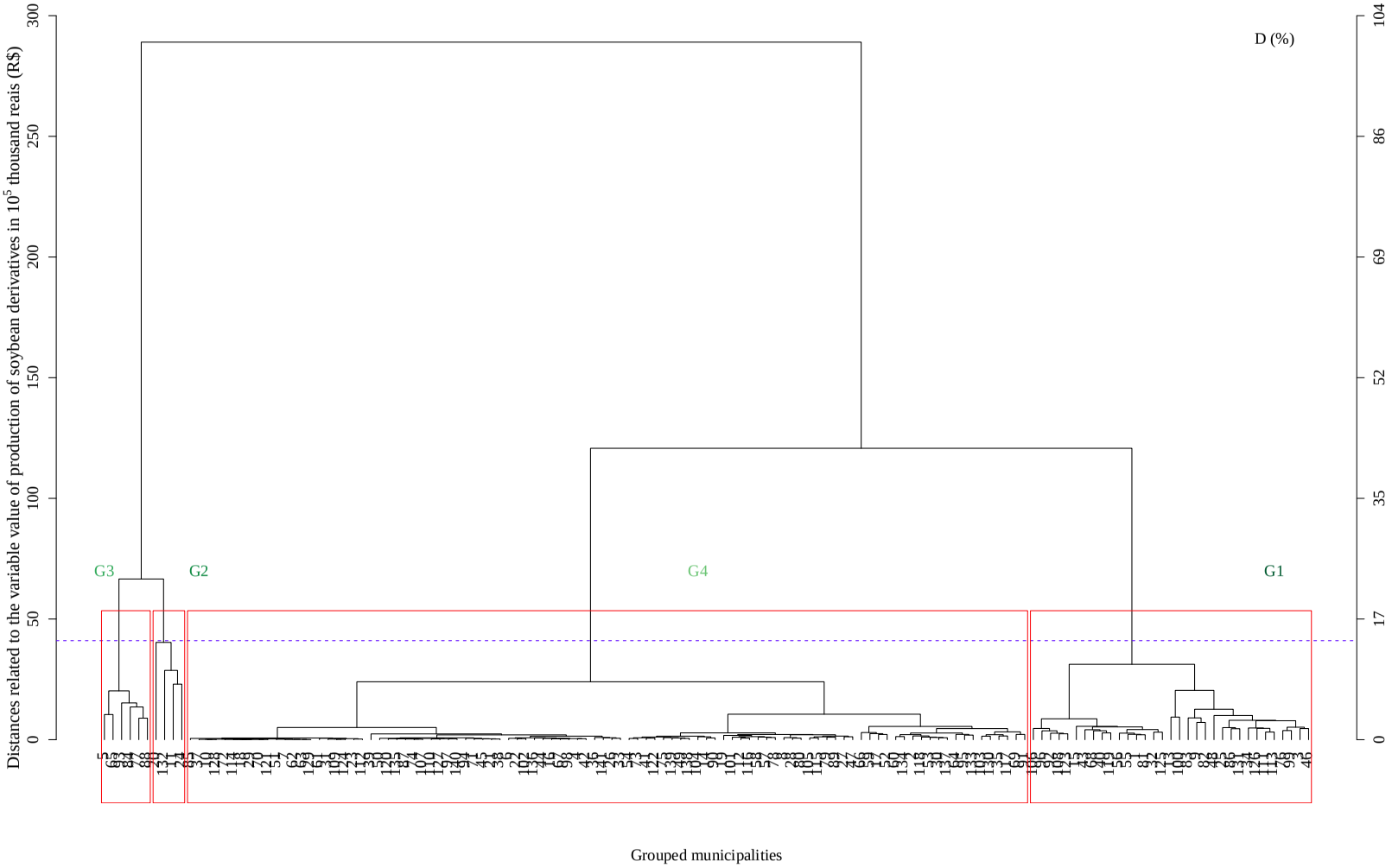


Figure 4.4: Dendrogram constructed for variable value of production of soy derivatives in 10^5 thousand reais (R$), and its groups.

Source: Results of Research.

The structures of dendrograms formed in the three Figures 4.2, 4.3 and 4.4 draw a clear picture of the groups formed by the combinations of cluster analysis techniques. The cutoff points of the Mojena method are observed on the blue dashed line on the vertical distance axis of each of the dendrograms relative to the three variables, and they were respectively 146 , 66 and 41, so it can be verified later on the formation of 5, 5 and 4 groups for each of the variables, and their percentages of dissimilarities D (%) were 13.13%, 15.12% and 6.98%, close values. The groups with their respective municipalities are summarized in Tables 4.1, 4.2 and 4.3.

Table 4.1: Municipal groups formed from the soybean grain production variable

Groups	**Municipalities**
G1	1, 3, 12, 15,17, 27, 40, 43, 46, 48, 55, 56, 66, 68, 69, 76, 81, 82, 83, 86, 91, 92, 95, 96, 106, 108 111, 113, 117, 119, 123, 125, 131
G2	5, 11, 20, 77, 93, 99
G3	2, 4 , 6, 7, 8, 10, 14, 16, 18, 19, 21, 22, 23, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41, 42, 44, 45, 47, 49, 50, 51, 52, 53, 54, 57, 58, 59, 60, 61, 62, 63, 64, 67, 70, 71, 72, 73, 74, 75, 78, 79, 80, 85, 87, 89, 90, 94, 97, 98, 101, 102, 103, 104, 105, 107, 109, 110, 112, 114, 115, 116, 118, 120, 121, 122, 124, 126,127, 128, 129, 130, 133, 134, 135, 136, 137, 138, 139, 140, 141
G4	24
G5	9, 13, 25, 65, 84, 88, 100, 132

Source: Results of Research.

Table 4.2: Municipal groups formed from the variable soybean grain production value

Groups	**Municipalities**
G1	3, 9, 13, 17, 25, 43, 46, 65, 76, 81, 83, 86, 88, 95, 96, 100, 125, 131, 132
G2	5, 11, 20, 77, 84,93, 99
G3	1, 12, 15, 27, 40, 48, 55, 56, 66, 68, 69, 82, 91, 92, 106, 108, 111, 113, 117, 119
G4	2, 4, 6, 7, 8, 10, 14, 16, 18, 19, 21, 22, 23, 26, 28, 29, 30, 31, 32, 33, 35, 36, 37, 38, 39, 41, 42, 44, 45, 47, 49, 50, 51, 52, 53, 54, 57, 58, 59, 61, 62, 63, 64, 67, 70, 71, 72, 73, 74, 75, 78, 79, 80, 85, 87, 89, 90, 94, 98, 101, 102, 103, 104, 105, 107,109, 110, 112, 114, 115, 116, 118, 120, 121, 122, 124, 126, 127, 128, 129, 130, 133, 134, 135, 136, 137, 138, 139, 140, 141
G5	24

Source: Results of Research.

Table 4.3: Municipal groups formed from the variable value of soybean derivatives

Groups	Municipalities
G1	1, 3, 9, 13, 15, 25, 32, 34, 40, 43, 46, 48, 55, 56, 68, 76, 81, 82, 83, 86, 92, 96, 99, 100, 106, 108, 111, 113, 119, 123, 125, 126, 131
G2	11, 20, 24, 132
G3	5, 65, 77, 84, 88, 93
G4	2, 4, 6, 7, 8, 10, 12, 14, 16, 17, 18, 19, 21, 22, 23, 26, 27, 28, 29, 30, 31, 33, 35, 36, 37, 38, 39, 41, 42, 44, 45, 47, 49, 50, 51, 52, 53, 54, 57, 58, 59, 60, 61, 62, 63, 64, 66, 67, 69, 70, 71, 72, 73, 74, 75, 78, 79, 80, 85, 87, 89, 90, 91, 94, 95, 97, 98, 101, 102, 103, 104, 105, 107, 109, 110, 112, 114, 115 116, 117, 118, 120,121 122, 124, 127, 128, 129, 130, 133, 134, 135 135, 136, 137, 138, 139, 140, 141

Source: Results of Research.

It should be noted that the first three groups (G1, G2, G3) have the highest concentration in the three variables, and group G4 and G5 in the production and production value variables contains only municipality 24 (Sorriso-MT), which has a prominent role in this sector in this period, and also appears in the G2 group of derivatives that has only 4 municipalities.

The ten most relevant municipalities in each of these variables are listed in Tables 4.4, 4.5 and 4.6.

Table 4.4: Main soybean producing municipalities in the state of Mato Grosso from 1990 to 2018.

Rank	Identification in the Map	Municipalities	10^6 (thousand tons)
1º	24	Sorriso (MT)	39.027
2º	11	Campo Novo do Parecis (MT)	25.879
3º	5	Nova Mutum (MT)	21.855
4º	20	Sapezal (MT)	21.423
5º	93	Diamantino (MT)	20.058

Source: IBGE.

Table 4.5: Main soybean producing municipalities in the state of Mato Grosso from 1990 to 2018, in production value.

Rank	Identification in the Map	Municipalities	10^6 (thousand reais)
1º	24	Sorriso (MT)	21.616
2º	20	Sapezal (MT)	13.093
3º	11	Campo Novo do Parecis (MT)	12.881
4º	50	Nova Mutum (MT)	12.551
5º	93	Diamantino (MT)	10.664

Source: IBGE.

Table 4.6: Main municipalities producing soybean derivatives in the state of Mato Grosso from 1990 to 2018, in production value.

Rank	Identification in the Map	Municipalities	10^6 (thousand reais)
1º	20	Sapezal (MT)	12.721
2º	132	Campo Verde (MT)	10.311
3º	11	Campo Novo do Parecis (MT)	9.259
4º	24	Sorriso (MT)	9.217
5º	84	Primavera do Leste (MT)	6.728

Source: IBGE.

The municipality of Sorriso occupies the first position in the production and production value variables and the fifth place for the production of soybean derivatives. It is observed that groups G3, G4 and G4 are the ones with more municipalities 93, 93 and 98, for the variables production, production value and soybean derivatives. Note that groups G3 and G4 contain the same municipalities in the production and production value variables, 90 municipalities are common to these two groups, which also appear in the G4 group of the soybean derivatives variable, with only eight municipalities that are out of the G4, namely 12, 77, 27, 66, 69, 91, 95, 117, in other words these three groups of each of the variables are similar.

Thus, for the same variables, an analysis of the comportment profiles of these groups was constructed, as illustrated in Figures 4.5, 4.6 and 4.7.

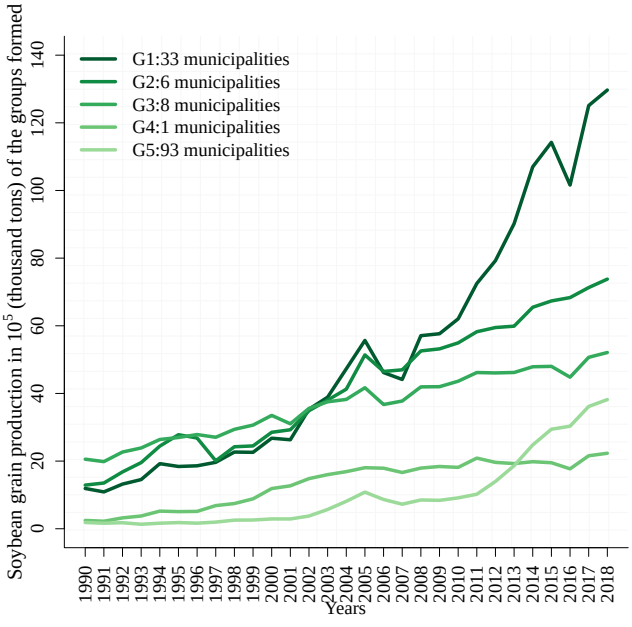


Figure 4.5: Profiles of the groups of the soybean production variable in grains.

Source: Research results.

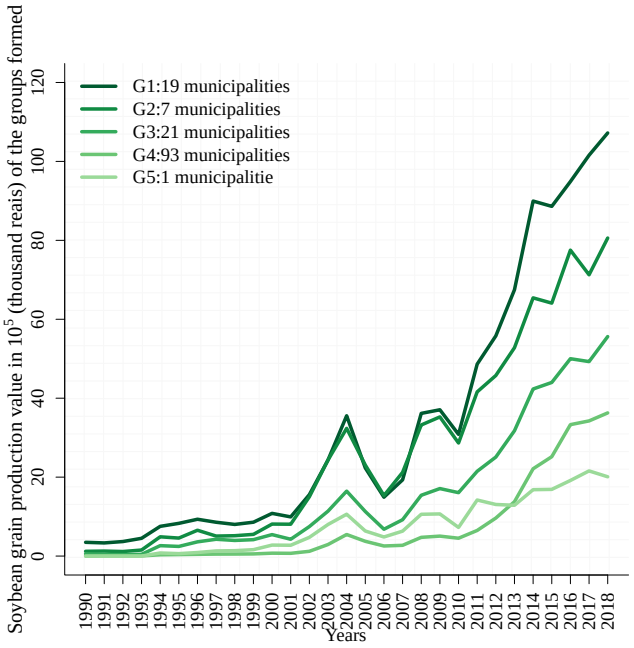


Figure 4.6: Profiles of the groups of the soybean grain production value variable.

Source: Research results.

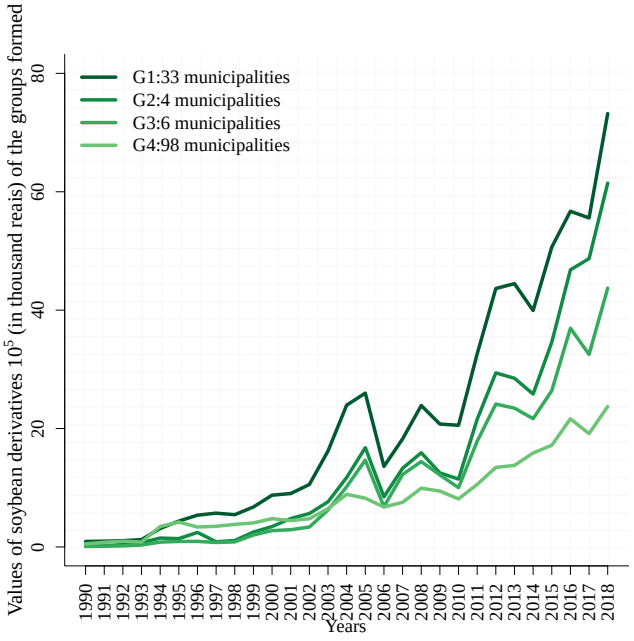


Figure 4.7: Profiles of the groups of the value of soybean derivatives variable.

Source: Results of Research.

Figures 4.5, 4.6 and 4.7 show, in decreasing order, that the highest productive concentrations and financial movements of soybean are in the first 3 groups, which represent only 33.33%, 33.33% and 30.50% of the 141 municipalities in the state of Mato Grosso, for the variables of production, production value and value of soybean derivatives. And in

all groups created, there is a growing trend in the activity of soybean production, which indicates that the groups with lower concentrations are evolving in these activities related to the three variables. The structures of the generated dendrograms were spatialized as shown in Figures 4.8, 4.9 and 4.10.

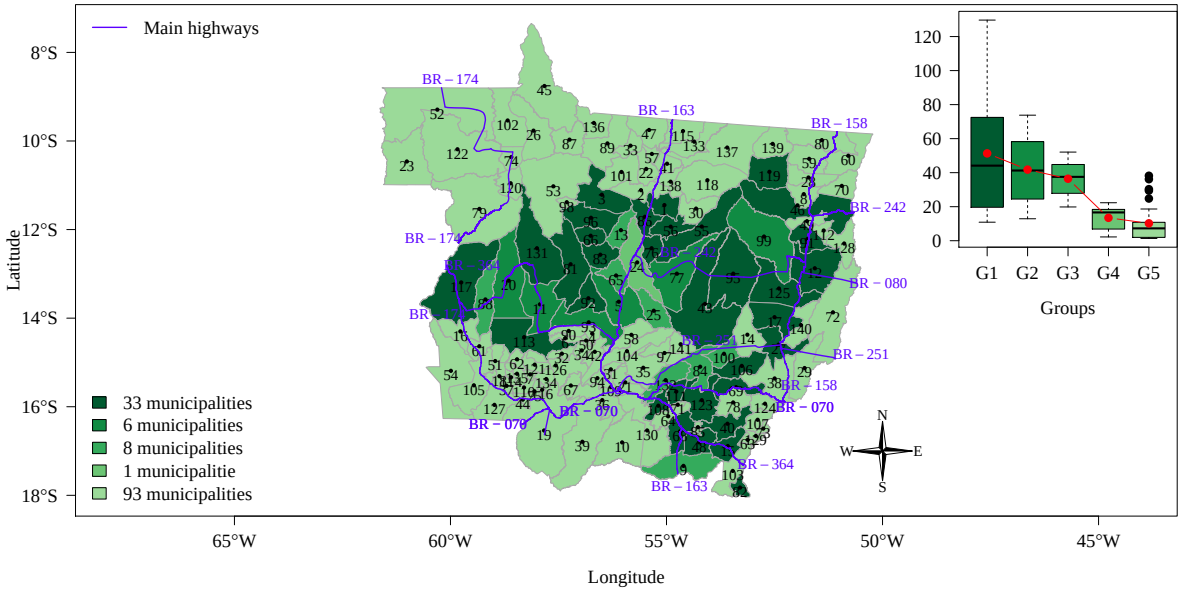


Figure 4.8: Map of the groups for the variable soybean production in 10^5 thousand tons.

Source: Results of Research.

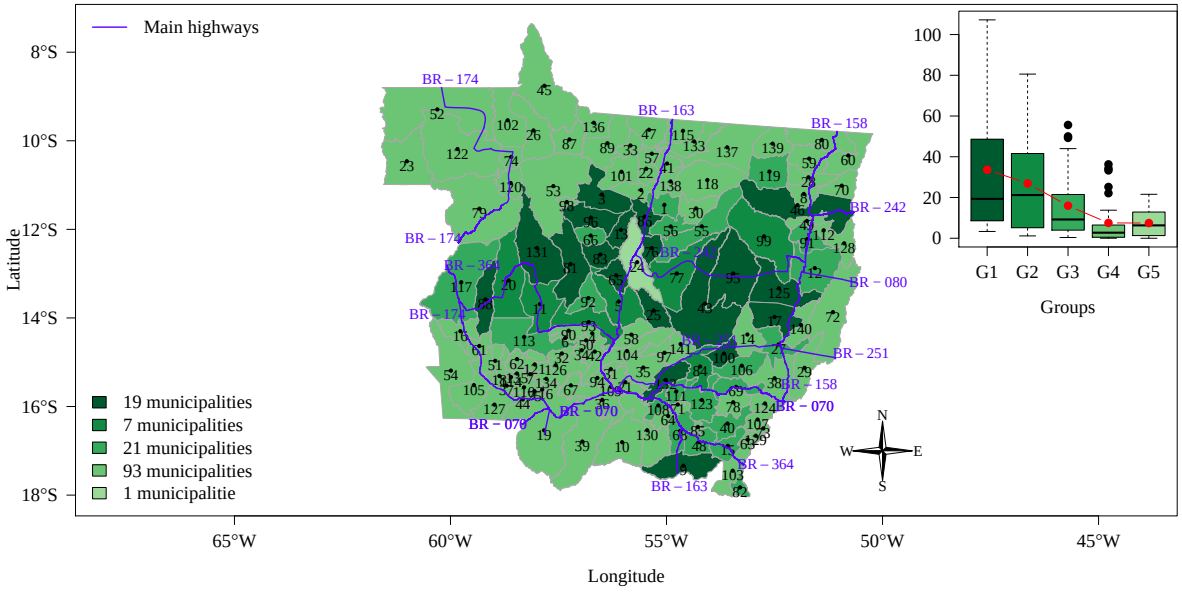


Figure 4.9: Map of the groups for the variable soybean production value in 10^5 thousand reais (R$).

Source: Results of Research.

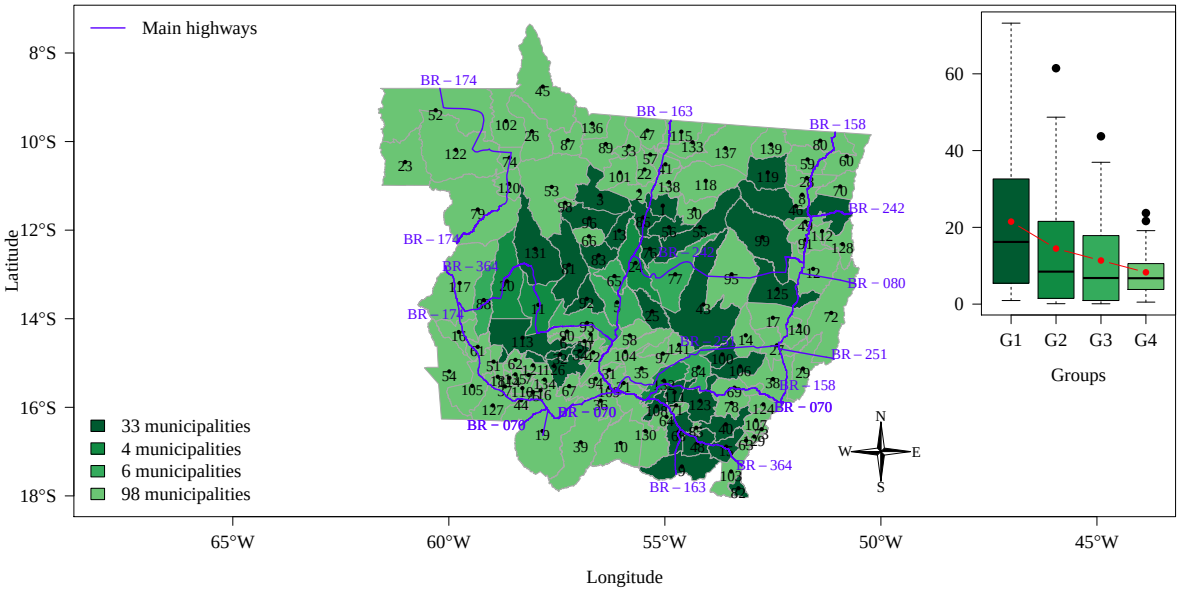


Figure 4.10: Map of the groups for the variable value of soybean derivatives in 10^5 thousand reais (R$).

Source: Results of Research.

The municipal development of production, production value and value of soybean derivatives, occurs from the interior to the outside of the state of Mato Grosso, from the group with the highest concentration to the one with the lowest, and the maps for each of the variables have similarity. And this phenomenon occurs around the main highways crossing the state, and that allow the flow of the production of soybean and its derivatives, and it is the most used modal in the country, to supply the internal and external grain markets.

The statistical validation of the structures of the dendrograms and their groups formed for these three variables occurred through the application of the cophenetic correlation, Pearson correlation and Mantel test with 10,000 randomized permutations of the matrices of distances simultaneously, as shown in Figures 4.11, 4.12 and 4.13.

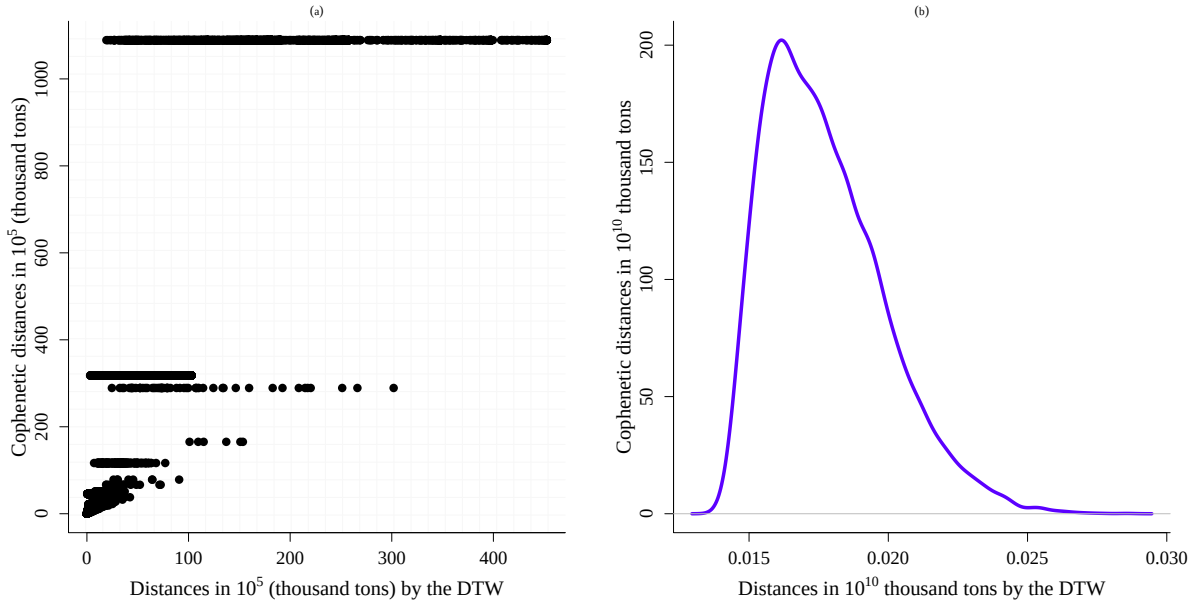


Figure 4.11: (a) Cophenetic correlation of $r_{cof} = 0.83$ related to production of soybean, with $p - value = 2.22e - 16$. (b) Mantel test related to production of soybean for 10,000 permutations and with $z_{calc} = 0.045$ and $p - value = 9.99e - 5$.

Source: Results of Research.

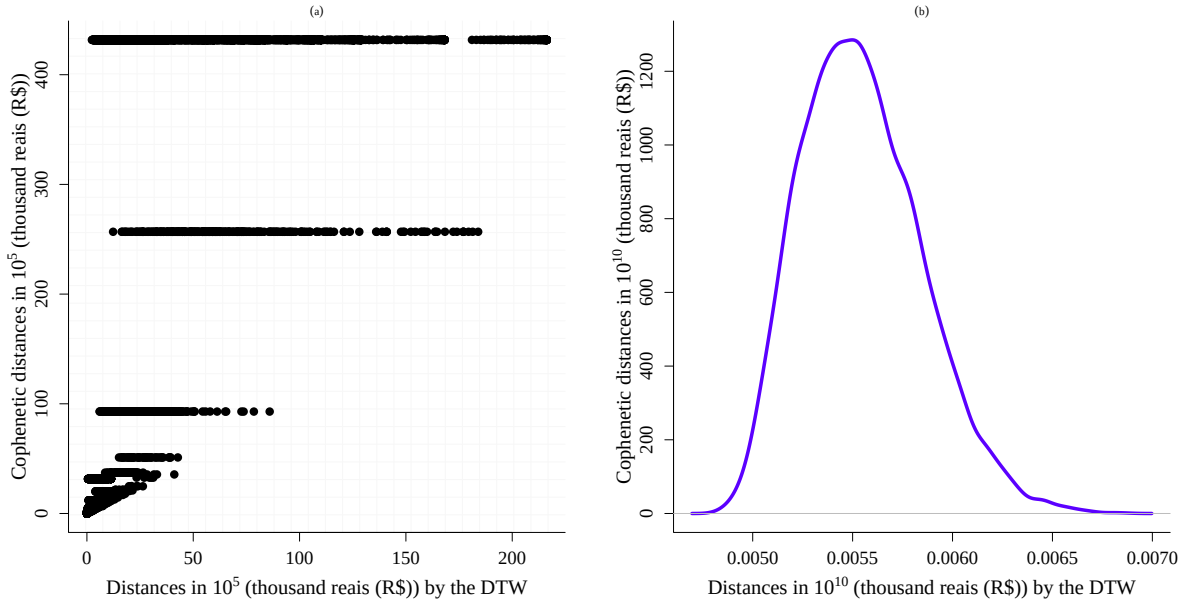


Figure 4.12: (a) Cophenetic correlation of $r_{cof} = 0.56$ related to production value of soybean, with $p - value = 2.2e - 16$. (b) Mantel test related to production value of soybean for 10,000 permutations and with $z_{calc} = 0.001$ and $p - value = 9.99e - 5$.

Source: Results of Research.

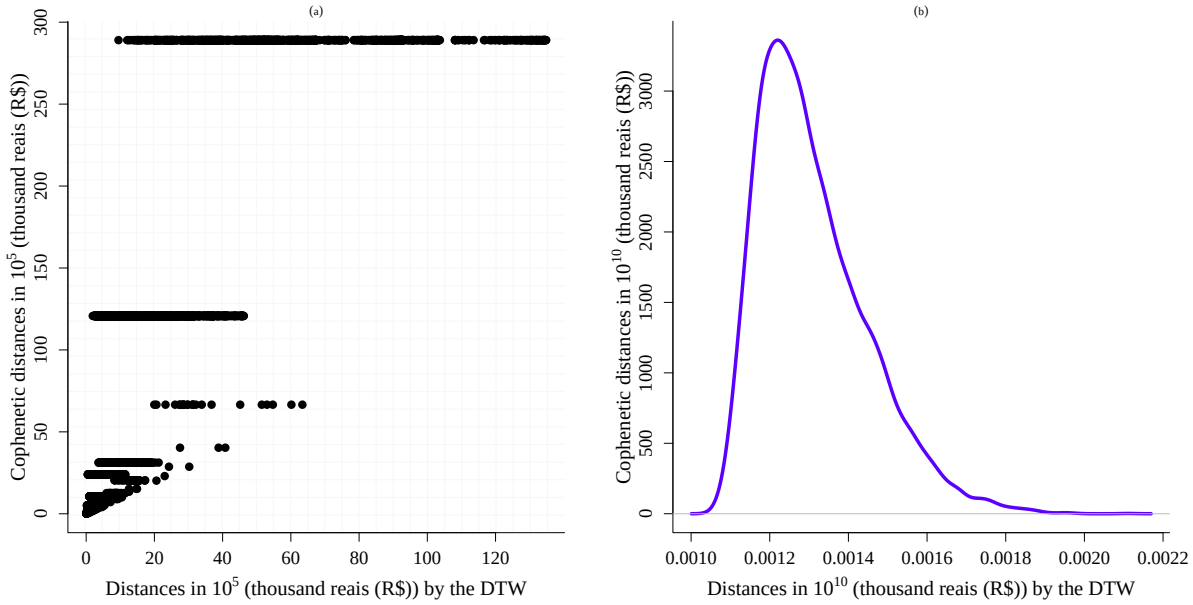


Figure 4.13: (a) Cophenetic correlation of $r_{cof} = 0.83$ related to value of soybean derivatives, with $p - value = 2.22e - 6$. (b) Mantel test related to production value of soybean for 10,000 permutations and with $z_{calc} = 0.003$ and $p - value = 9.99e - 5$.

Source: Results of Research.

The correlations between the distance axes in each of the variables were 0.83, 0.56 and 0.83, the same being statistically validated by the Pearson correlation test, added to the Mantel test with 10,000 randomized permutations of the two axes of distances, which also show positive results for the formation of the generated group structures, in each of the dendrograms initially exposed.

4.6 Conclusions

The DTW distance combined with Ward's clustering method for the variables of production, production value, value of soybean derivatives, proved to be effective and validated with the proposed statistical tests. It was observed the creation of 5, 5 and 4 clusters in the same variables, and they were ordinated from the highest to the lowest concentration.

All of these groups have a growing trend, which reveals the heating up of activity in the state in this period. From the creation of these spatialized groups, it can be seen that the development of the three variables occurs from the municipalities located in the center (larger groups) of the state to those more distant from the center (smaller groups), and this development occurs around the main highways that cross the state, a modal that drives the development of the activity through the flow to supply the domestic and foreign markets.

The present study allowed to have a picture of the production activity, production value and value of soybean derivatives in the state of Mato Grosso, from 1990 to 2018, and also enabled the formation of scaled clusters or zones of these same variables, and this result can contribute to the generation of public policies in the state and also makes it possible to map possible new areas of investment for producers interested in the coordinated expansion of the crop in this region, since there is a portrait of the comportment of the groups, revealing those with greater and lesser influence, as well as their municipalities for each of three variables in this period.

The research methodology can also be updated year after year and can also serve as a basis for analyses of other crops in the state of Mato Grosso (corn, cotton and other crops of interest) and other states and even for Brazil as a whole.

4.7 References

Batista, G. E. A. P. A., X. Wang, and E. J. Keogh, 2011 A complexity-invariant distance measure for time series. In *Proceedings of the 2011 SIAM international conference on data mining*, pp. 699–710, SIAM.

Berndt, D. J. and J. Clifford, 1994 Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pp. 359–370, Seattle, WA, USA:.

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

Bussab, W., É. Miazaki, and D. Andrade, 1990 Introdução à análise de agrupamentos: In: Simpósio Brasileiro de Probabilidade e Estatística, 9., 1990, São Paulo. ABE .

Everitt, B., 1993 *Cluster analysis*. London: Edward Arnold.

Ferreira, D. F., 2008 *Estatística multivariada*. Editora Ufla Lavras.

Forsythe, G. E., M. A. Malcolm, and C. B. Moler, 1977 *Computer methods for mathematical computations*, volume 11. Englewood Cliffs, New Jersey. Prentice Hall, Inc.

Giusti, R. and G. E. A. P. A. Batista, 2013 An empirical comparison of dissimilarity measures for time series classification. In *2013 Brazilian Conference on Intelligent Systems*, pp. 82–88, IEEE.

Hair, J. F., W. C. Black, B. J. Babin, R. E. Anderson, and R. L. Tatham, 2009 *Análise multivariada de dados*. Bookman editora.

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

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.

Mantel, N., 1967 The detection of disease clustering and a generalized regression approach. Cancer research **27**: 209–220.

Michaud, P., 1997 Clustering techniques. Future Generation Computer Systems **13**: 135–147.

Milligan, G. W. and M. C. Cooper, 1985 An examination of procedures for determining the number of clusters in a data set **50**: 159–179.

Mojena, R., 1977 Hierarchical grouping methods and stopping rules: an evaluation. The Computer Journal **20**: 359–363.

Müller, M., 2007 Dynamic time warping. Information retrieval for music and motion pp. 69–84.

Pearson, K., 1904 *On the theory of contingency and its relation to association and normal correlation*. Dulau and Company.

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.

R Core Team, 2021 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rakthanmanon, T., B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, 2013 Addressing big data time series: Mining trillions of time series subsequences under dynamic time warping. ACM Transactions on Knowledge Discovery from Data (TKDD) **7**: 1–31.

Rong-zhen, S., W. Ming-jiu, Z. Tian-qi, *et al.*, 2020 Comparison of photosynthetic characteristics and cluster analysis in Glycine soja and strains from Glycine soja× Glycine max cross. Chinese Journal of Oil Crop Sciences **42**: 255.

Scott, A. J. and M. J. Symons, 1971 Clustering methods based on likelihood ratio criteria. Biometrics pp. 387–397.

Silva, A. R. and C. T. S. Dias, 2013 A cophenetic correlation coefficient for Tocher's method. Pesquisa Agropecuaria Brasileira **48**: 589–596.

Sokal, R. R. and F. J. Rohlf, 1962 The comparison of dendrograms by objective methods. Taxon **11**: 33–40.

Ward Jr, J. H., 1963 Hierarchical grouping to optimize an objective function. Journal of the American statistical association **58**: 236–244.

5 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.

APPENDICES

Apêndice I

Frame 5.1: Identification of regions in the state of Brazil.

Regions	**States**
Central-Westt	MT, GO, MS, DF
South	PR, RS, SC
Southeast	MG, SP, RJ, ES
North -East	BA, MA, PI, AL, CE, RN, PB, PE, SE
North	TO, PA, RO, RR, AP, AC, AM

Source: IBGE.

Frame 5.2: Identification of states in the Brazil.

State	**Name**
MT	Mato Grosso
GO	Goiás
MS	Mato Grosso do Sul
DF	Distrito Federal
PR	Paraná
RS	Rio Grande do Sul
SC	Santa Catarina
MG	Minas Gerais
SP	São Paulo
RJ	Rio de Janeiro
ES	Espirito Santo
BA	Bahia
MA	Maranhão
PI	Piauí
AL	Alagoas
CE	Ceará
RN	Rio Grande do Norte
PB	Paraíba
PE	Pernambuco
SE	Sergipe
TO	Tocantins
PA	Pará
RO	Rondônia
RR	Roraima
AP	Amapá

Frame 4.2: Continued.

State	Name
AC	Acre
AM	Amazonas

Source: IBGE.

Frame 5.3: Identification of municipalities in the state of Mato Grosso.

ID	Municipalities
1	Cláudia (MT)
2	Itaúba (MT)
3	Tabaporã (MT)
4	Nortelândia (MT)
5	Nova Mutum (MT)
6	Santo Afonso (MT)
7	Rio Branco (MT)
8	Canabrava do Norte (MT)
9	Itiquira (MT)
10	Barão de Melgaço (MT)
11	Campo Novo do Parecis (MT)
12	Ribeirão Cascalheira (MT)
13	Ipiranga do Norte (MT)
14	Campinápolis (MT)
15	Alto Garças (MT)
16	Nova Lacerda (MT)
17	Água Boa (MT)
18	Jauru (MT)
19	Cáceres (MT)
20	Sapezal (MT)
21	Cuiabá (MT)
22	Colíder (MT)
23	Rondolândia (MT)
24	Sorriso (MT)
25	Santa Rita do Trivelato (MT)
26	Nova Bandeirantes (MT)
27	Nova Xavantina (MT)
28	Porto Alegre do Norte (MT)
29	Araguaiana (MT)

Frame 4.3: Continued.

ID	Municipalities
30	União do Sul (MT)
31	Acorizal (MT)
32	Nova Olímpia (MT)
33	Carlinda (MT)
34	Denise (MT)
35	Chapada dos Guimarães (MT)
36	Nossa Senhora do Livramento (MT)
37	Figueirópolis D'Oeste (MT)
38	Barra do Garças (MT)
39	Poconé (MT)
40	Guiratinga (MT)
41	Terra Nova do Norte (MT)
42	Alto Paraguai (MT)
43	Paranatinga (MT)
44	Glória D'Oeste (MT)
45	Apiacás (MT)
46	São Félix do Araguaia (MT)
47	Novo Mundo (MT)
48	Pedra Preta (MT)
49	Alto Boa Vista (MT)
50	Arenápolis (MT)
51	Vale de São Domingos (MT)
52	Colniza (MT)
53	Juara (MT)
54	Vila Bela da Santíssima Trindade (MT)
55	Feliz Natal (MT)
56	Santa Carmem (MT)
57	Nova Guarita (MT)
58	Nobres (MT)
59	Confresa (MT)
60	Santa Terezinha (MT)
61	Conquista D'Oeste (MT)
62	Reserva do Cabaçal (MT)
63	Araguainha (MT)
64	Juscimeira (MT)
65	Lucas do Rio Verde (MT)
66	Itanhangá (MT)
67	Porto Estrela (MT)
68	Rondonópolis (MT)
69	General Carneiro (MT)
70	Luciara (MT)
71	São Pedro da Cipa (MT)
72	Cocalinho (MT)

Frame 4.2: Continued.

ID	Municipalities
73	Ribeirãozinho (MT)
74	Juruena (MT)
75	Mirassol d'Oeste (MT)
76	Vera (MT)
77	Nova Ubiratã (MT)
78	Tesouro (MT)
79	Juína (MT)
80	Vila Rica (MT)
81	Nova Maringá (MT)
82	Alto Taquari (MT)
83	Tapurah (MT)
84	Primavera do Leste (MT)
85	São José do Povo (MT)
86	Sinop (MT)
87	Nova Monte Verde (MT)
88	Campos de Júlio (MT)
89	Alta Floresta (MT)
90	Nova Marilândia (MT)
91	Bom Jesus do Araguaia (MT)
92	São José do Rio Claro (MT)
93	Diamantino (MT)
94	Jangada (MT)
95	Gaúcha do Norte (MT)
96	Porto dos Gaúchos (MT)
97	Nova Brasilândia (MT)
98	Novo Horizonte do Norte (MT)
99	Querência (MT)
100	Santo Antônio do Leste (MT)
101	Nova Canaã do Norte (MT)
102	Cotriguaçu (MT)
103	Alto Araguaia (MT)
104	Rosário Oeste (MT)
105	Pontes e Lacerda (MT)
106	Novo São Joaquim (MT)
107	Torixoréu (MT)
108	Jaciara (MT)
109	Várzea Grande (MT)
110	São José dos Quatro Marcos (MT)
111	Dom Aquino (MT)
112	Serra Nova Dourada (MT)
113	Tangará da Serra (MT)
114	Indiavaí (MT)
115	Guarantã do Norte (MT)

Frame 4.2: Continued.

ID	Municipalities
116	Curvelândia (MT)
117	Comodoro (MT)
118	Marcelândia (MT)
119	São José do Xingu (MT)
120	Castanheira (MT)
121	Salto do Céu (MT)
122	Aripuanã (MT)
123	Poxoréu (MT)
124	Pontal do Araguaia (MT)
125	Canarana (MT)
126	Barra do Bugres (MT)
127	Porto Esperidião (MT)
128	Novo Santo Antônio (MT)
129	Ponte Branca (MT)
130	Santo Antônio do Leverger (MT)
131	Brasnorte (MT)
132	Campo Verde (MT)
133	Matupá (MT)
134	Lambari D'Oeste (MT)
135	Araputanga (MT)
136	Paranaíta (MT)
137	Peixoto de Azevedo (MT)
138	Nova Santa Helena (MT)
139	Santa Cruz do Xingu (MT)
140	Nova Nazaré (MT)
141	Planalto da Serra (MT)

Source: IBGE.
Note: ID is identification on the map.