


STATISTICAL MODEL FOR ESTIMATING ELECTRICITY CONSUMPTION BY THE AGRICULTURAL SECTOR IN SÃO PAULO

 <https://doi.org/10.56238/arev6n4-377>

Submitted on: 23/11/2024

Publication date: 23/12/2024

Monclar Nogueira Christovão¹ and Mario Mollo Neto²

ABSTRACT

Electricity is a fundamental and essential input for the socioeconomic development of a region. Until 2016, the main energy resource used in the agricultural sector of São Paulo was diesel oil. However, in the last six years (2017 to 2022), electricity has become the most widely used energy source for this sector. In this context, it is vital to make a future projection of load demand in order to support deliberations in the planning and expansion of the state's electricity supply system. The objectives of this study are to create a statistical model capable of predicting the consumption of electricity in the rural sector of São Paulo, as well as to recognize the variables that influence the projection of this consumption. Based on quantitative information collected on the website of the Department of Environment, Infrastructure and Logistics of the state of São Paulo, the multiple linear regression model was used. For the construction of the model, six independent variables were selected, three of which came from agribusiness in São Paulo: number of consumers, electricity tariff and consumption of diesel oil; two related to the state: installed power of biomass thermal plants and photovoltaic plants and one of national agribusiness: GDP. Due to the lack of statistical significance, two variables were rejected: power of thermal biomass plants and GDP. Therefore, with the proposed mathematical model, built with four significant variables (with different levels of influence), it is possible to make future predictions of the energy consumption of the researched sector. The installed power of photovoltaic plants had a significant impact, while the number of consumers, the energy tariff and the consumption of diesel oil showed a strong influence on the forecast of energy consumption in the sector. The highlight was the confirmation, by the adjusted statistical model, of the transition from the consumption of diesel oil to electricity as the main source of energy used by the agricultural sector in São Paulo.

Keywords: Mathematical Model. Forecast. Tendency. Rural Sector. Linear regression.

¹ Master's degree in Agribusiness and Development from São Paulo State University (UNESP) Federal Institute of São Paulo (IFSP), Tupã
E-mail: monclar.christovao@ifsp.edu.br
ORCID: <https://orcid.org/0000-0001-5509-0574>
LATTES: <http://lattes.cnpq.br/2038445224049133>

² Dr. in Agricultural Engineering from the State University of Campinas
São Paulo State University, Faculty of Science and Engineering, Tupã
E-mail: mario.mollo@unesp.br
ORCID: <https://orcid.org/0000-0002-8341-4190>
LATTES: <http://lattes.cnpq.br/6037463340047597>

INTRODUCTION

In Brazil, the presence of oil and its derivatives in the domestic energy supply grew from 34% in 1970 to 46% in 2000 and is estimated to decrease to 31% in 2030 (BRASIL-MME, 2007).

In the 1970s, firewood was the dominant energy source, representing about 46% of total consumption, but it quickly gave way to the use of petroleum products, which, in the aforementioned period, reached a share of 37.85%, while hydroelectricity reached the mark of 5.5% (BRASIL-MME, 2007).

According to Brasil-MME (2007), between the 80s and 2000s, the presence of oil and its derivatives was established in the range between 40 and 50%, while firewood had a reduction to 8% and hydroelectricity was consolidated around 16%.

Electricity contributes to the progress and prosperity of a society. Electricity is a basic and fundamental input and, as such, socially excludes the portion of the population that has restrictions on its access (BISOGNIN; WERNER, 2020).

According to Brasil-EPE (2024), the total installed electricity power in Brazil in 2023 reached the mark of 226 Giga watt (GW). Allocating this amount among the four main generating sources in the country, it can be seen that hydroelectric plants (HPP) were responsible for 48.6% of this power, followed by thermoelectric plants (UTE) with 20.0%, solar plants with 16.7% and wind farms with 12.7%. The highlight was the growth of close to 55% in the capacity of photovoltaic plants, compared to 2022, with an installed capacity of 37,843 GW.

As for the production of electricity at the national level, contrasting the years 2022 and 2023, according to Brasil-EPE (2024), there were two highlights. The first was the 68.1% increase in photovoltaic generation, from 30,126 GWh to 50,633 GWh and the second was the 19.3% reduction in the generation of UTEs, which use petroleum products, from 7,485 GWh to 6,041 GWh.

During the campaign for the presidency of Brazil, the government then elected for the 2023-2026 quadrennium made available a document called "Letter for the Brazil of Tomorrow" containing proposals and commitments in which the resumption of the "Light for All" program is included (MELLO, 2022). The objective of this program is to provide access to the public electricity distribution service to residents of distant and needy rural areas.

The total consumption of electricity by all areas in the state of São Paulo in 2021, also including self-producers, was 151,729 Giga watt-hours (GWh), an increase of 4.32%

compared to the previous year, which was equal to 145,451 GWh. The industrial (9.2%) and commercial (3.8%) areas showed the largest increases in electricity use for the period. The residential (0.5%) and agricultural (0.2%) sectors showed very low growth, almost stability, in the consumption of this energy source (SÃO PAULO-SIMA, 2022).

As reported by Bisognin and Werner (2020), accurate forecasts of electricity demand and consumption are essential for correct decisions to be made in the allocation of resources, construction, and improvement of a region's electrical infrastructure. The need to plan ahead results from the fact that electricity is an input that cannot be stocked.

For the purposes of estimation and forecasting, one of the most used techniques in the area of management and in universities is regression analysis, whose main point is the presence of a statistical relationship between a variable called dependent, or explained, or predicted, with one or more variables designated independent, or predictors, or explanatory (CUNHA; COELHO, 2014).

According to Cunha and Coelho (2014), the purpose of regression analysis is to predict the amounts of the dependent variable based on the known amounts of the independent variables.

Also according to Cunha and Coelho (2014), regression is understood as the formation of a functional relationship between variables involved in the explanation of a phenomenon.

As for the classification of energy types, according to Brasil-MME (2024), primary and secondary energies are categories within the general classification of energy sources.

Primary energy is the energetic result of resources that are extracted from nature in their original form, such as: firewood, oil, natural gas, coal, and others (BRASIL-MME, 2024). Some agro-industry waste, such as coffee grounds and rice straw, are primary sources used in steam generation (SÃO PAULO-SEMIL, 2024).

Secondary energy is the result of industrialized products that undergo modifications in one or more transformations. Secondary energy sources are diesel oil, fuel oil, liquefied petroleum gas (LPG), kerosene, electricity, and others (BRASIL-MME, 2024).

For the above, the research of statistical prediction of electricity consumption in the agricultural sector of São Paulo is justified, since, in addition to being a diagnostic tool, it is also capable of generating a mathematical model to estimate the energy consumption of this sector, contributing with regional information to meet the needs of improvements and expansion of the entire infrastructure of transmission and distribution of electricity due to

the growing trend in the use of this type of energy. It also contributes to the creation of public policies for the agribusiness sector in favor of the social well-being of the rural population and the development of the economy.

The general objective of this research was to generate a statistical model that estimates the electricity consumption of the agricultural sector in São Paulo. The specific objective was to identify the variables that contribute to the prediction of this consumption.

METHODOLOGY

The tool used in this work is regression analysis, which, according to Martins and Domingues (2014), is used mainly for the purpose of prediction. The objective is to construct a statistical model for the prediction of responses of a dependent variable (Y) depending on the value of a variable (X) or more variables (X1, X2, ... and Xn) independent(s), establishing a mathematical relationship between these two types of variables (MARTINS; DOMINGUES, 2014).

The response variable (dependent) must be quantitative, while the predictor variable(s) can be quantitative or qualitative.

In the survey, the dependent variable is the annual forecast of electricity consumption by the agricultural sector in São Paulo, based on a statistical model.

The regression model is called simple, that is, when there is only one independent (explanatory) variable. In the existence of two or more of these variables, the regression model is called multiple (MARTINS; DOMINGUES, 2014).

BUILDING A SIMPLE LINEAR REGRESSION MODEL

The relationship between the two types of variables can take several forms, among them, a simplified model, which is the simple linear relationship containing only one independent variable, such as the equation of the line, according to the Equation (1).

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon \quad (1)$$

where:

- β_0 : linear coefficient (line intercept);
- β_1 : slope coefficient or regression coefficient (slope of the line);
- ε : regression error or residue.

Based on sample data, we have \hat{Y} , which is the result of the prediction Y for an observation X, according to the adjusted model of the simple linear regression enunciated by the Equation (2).

$$\hat{Y} = b_0 + b_1.X \quad (2)$$

where:

- b_0 : the β_0 estimator; and
- B_1 : or β_1 estimator.

The values of b_0 and b_1 will be collected in the "Coefficients" column of the Data Analysis tool of the Microsoft Excel® spreadsheet editor from samples of ordered pairs of two variables.

CONSTRUCTING A MULTIPLE LINEAR REGRESSION MODEL

The relationship between the two types of variables can take several forms, among them, a multiple model that aims to better predict and explain the behavior of the researched dependent variable (Y). In this way, other explanatory variables (X_1, X_2, \dots, X_n) can be incorporated into the model, according to the Equation (3).

$$Y = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \dots + \beta_n.X_n + \varepsilon \quad (3)$$

where:

- β_0 : linear coefficient (intercept/intersection of the line);
- $\beta_1, 2, \dots, n$: slope or regression coefficient (slope of the line);
- ε : experimental error of Y (pontos fora da reta).

Cunha and Coelho (2014) clarify that the coefficient β_0 symbolizes the value of the intersection of the regression graph with the ordinate axis; therefore, β_0 portrays the value of Y when X is zero.

Multiple regression analysis is used to study the relationship between a single dependent variable, chosen by the researcher, with numerous independent variables of known values (GIL, 2021).

Based on sample data, we have \hat{Y} , which is the result of prediction Y for X observations, according to the fitted model provided by the Equation (4).

$$\hat{Y} = b_0 + b_1.X_1 + b_2.X_2 + \dots + b_n.X_n \quad (4)$$

where:

- b_0 : β_0 estimator;
- $b_1, 2, \dots, n$: respective estimators of $\beta_1, 2, \dots, n$.

Using the tool called "Regression", available in "Data Analysis" of the Microsoft Excel® spreadsheet editor, from samples of ordered pairs of three or more variables, the values of b_0, b_1, b_2, \dots , and b_n will be collected in the "Coefficients" column.

TESTING THE SIGNIFICANCE OF THE REGRESSION COEFFICIENTS

In the spreadsheet called "ANOVA", created by the "Regression" tool of Microsoft Excel® Data Analysis, the individual amounts of "P-value" of the coefficients (parameters) are present.

The importance of one or more independent variables (individually) is tested. The information contained in the individual "P-values" allows us to determine whether each explanatory variable is statistically relevant or not.

Chance:

- $H_0: \beta_i = 0$, consists of checking whether each parameter β_i of the regression is equal to zero;
- $H_1: \beta_i \neq 0$, if $P\text{-value} \leq \alpha$, hypothesis H_0 is rejected, concluding that $\beta_i \neq 0$ for a risk (significance level) α .

TESTING THE EXISTENCE OF LINEAR REGRESSION

Also in the "ANOVA" spreadsheet, described in the previous item, is the value of "F for meaning". This index indicates whether the linear regression model as a whole is statistically relevant or not. It indicates whether the combination of explanatory variables has a statistically significant association with the response variable.

For the condition of descriptive level or probability of significance ($p\text{-value}$: F for significance) $\leq \alpha$ (level of significance) adopted in the research, it can be concluded that there is regression, that is, the model can explain and predict Y (response variable).

TOOLS FOR ANALYZING RESULTS

Quality of Fit

Linear correlation coefficient

Applying the "Regression" tool contained in the Excel® Data Analysis, from samples of ordered pairs of two or more variables, the so-called Linear Correlation Coefficient (multiple R) present in the table called "Regression Statistics" is obtained. This coefficient symbolizes the level of connection (association) between the autonomous and dependent variables (CUNHA; COELHO, 2014).

Coefficient of Determination or Explanation

In the table "Regression Statistics", there is also the so-called Coefficient of Determination or Explanation (R^2).

R^2 is a measure of the proportion of variation in Y (response), which is explained by X (explanatory) by the fit of the linear model, thus being a descriptive measure of the quality of the fit. The value of this coefficient is between $0 < R^2 < 1$, and the closer to the unit value, the better the quality of the fit of this linear model.

Adjusted coefficient of determination

The Adjusted Coefficient of Determination (adjusted R^2) is listed in the "Regression Statistics" table, and is used in the comparison between regression equations that contain different amounts of independent variables or sample sizes (WEDGE; COELHO, 2014).

The regression equation that best explains the dependent variable studied is the one whose amount of the adjusted R^2 coefficient is the closest to the unit value.

Standard Error

The "Standard Error" is the average value, in the unit of the dependent variable, that the adjusted equation is missing more or less; In other words, it is the average distance between the observed amounts and the linear regression line.

Also present in the table mentioned above, the "Standard Error" is an indicator of regression accuracy, and the lower its value, the better the model estimate (CUNHA; COELHO, 2014).

Comparing regression equations

Regression equations can be compared, using the Adjusted Coefficient of Determination (adjusted R²) of both, and the regression that best explains the dependent variable is the one whose amount of the adjusted R² coefficient is the closest to the unit value.

Another factor of comparison, between regressions that explain the same dependent variable, is the "Standard Error" which, the lower its value, the more accurate the model estimate will be.

Residual analysis (ϵ)

It is suggested to verify some attributes of the distribution of waste, such as:

- a) To ascertain whether the mean probability distribution of the variable ϵ is zero;
- b) Test for homoscedasticity. In analysis of variance (ANOVA), errors must have common variance;
- c) To verify the normality of the probability distribution of the variable ϵ using the Shapiro-Wilk test.

Chance:

- H₀: the variable under study is normally distributed;
- H₁: the variable under study is not normally distributed.

If the p-value $\leq \alpha$, hypothesis H₀ is rejected for a risk (significance level) α .

If the p-value $> \alpha$, hypothesis H₀ is not rejected for a risk (significance level) α .

- d) Absence of serial autocorrelation. According to Cunha and Coelho (2014), the model should assume that the correlation between the residuals along the entire spectrum of independent variables is null; this means that the impact of an observation on a given variable X does not affect subsequent observations. Therefore, there is no causal relationship between the residuals and the variable X and, consequently, the variable Y is affected only by the variable X in question, and not by lagged effects of X₁ on X₂ and, consequently, on Y.

Multicollinearity between independent variables

Multicollinearity occurs when two or more independent variables, present in the model and that try to explain the same phenomenon, have similar information.

Consequently, the high correlation between two or more independent variables can make it

difficult to distinguish their individual effects on the dependent variable, causing an overlap of information in the explanation and estimation, which can lead to the loss of significance of one of them in the elucidation of the behavior of the phenomenon (CUNHA; COELHO, 2014).

According to Cunha and Coelho (2014), multicollinearity has the effect of distorting the calculated angular coefficients of the affected variables, impairing the predictive accuracy of the model and the understanding of the true impact of the independent variable on the behavior of the dependent variable.

The implications of the above are presumed and listed below:

- Increase in standard errors;
- Reduction in the efficiency of estimators;
- Less accurate estimates;
- Increased sensitivity to small fluctuations in data.

These inferences described above make it difficult to separate the effects of each variable.

The multicollinearity between the independent variables will be investigated, using Pearson's Correlation Test in order not to use redundant variables.

Chance:

- $H_0: p = 0$ (there is no linear correlation between the variables under study);
- $H_1: p \neq 0$ (there is a linear correlation between the variables under study).

If the $p\text{-value} \leq \alpha$, hypothesis H_0 is rejected for a risk (significance level) α .

If the $p\text{-value} > \alpha$, hypothesis H_0 is not rejected for a risk (significance level) α .

Next, the application of the proposed method is presented, as well as the exploration and discussion of the results.

RESULTS AND DISCUSSION

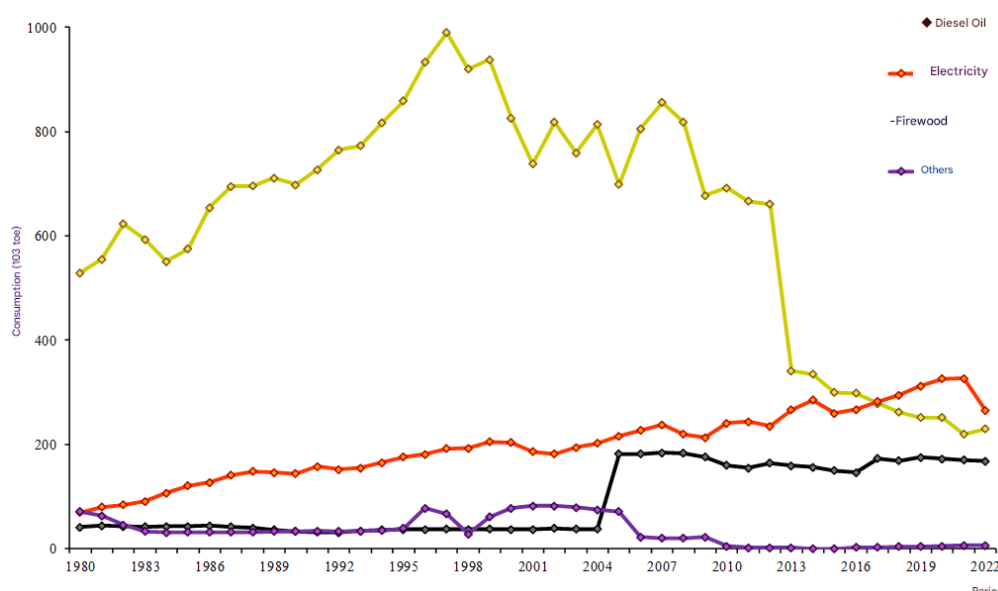
The data collection was mainly based on the annual digital report entitled Energy Balance of the State of São Paulo (BEESP) made available by the Secretariat of Environment, Infrastructure and Logistics (SEMIL) of the São Paulo government, in which the agricultural sector is included in the class called "other" (along with public services, public authorities, public lighting and own consumption) in the tables of this document.

In this research, the approach is quantitative due to the nature of all the variables involved. And these variables – dependent variable Y and independent variables X1 to X6 – are related and detailed below.

- The dependent variable Y – electricity consumption, in Megawatt-hour (MWh), of São Paulo agribusiness – had its historical series (2000 to 2023) collected from the SEMIL website (<https://dadosenergeticos.energia.sp.gov.br/portalsev2/intranet/Eletricidade/index.html>) in "Energy Data/Distribution System/Annual Consumption".
- Independent variable X1 (number of consumers in São Paulo agribusiness). The information was gathered from the same SEMIL website mentioned above (<https://dadosenergeticos.energia.sp.gov.br/portalsev2/intranet/Eletricidade/index.html>) in "Energy Data/Distribution System/Annual Consumption".
- Independent variable X2 – electricity tariff, in reais per Megawatt-hour (R\$/MWh), of the rural sector of São Paulo, with the values of conventional electricity tariffs [Distribution System Use Tariff (TUSD) + Energy Tariff (TE)], between the years 2000 and 2009, was determined by the arithmetic average, year by year, of three subgroups, which are: B2-Rural, B2-Rural Electrification Cooperative and B2-Public Irrigation Service, contained in the "History prior to 2010/History of Homologatory Resolutions", published by the National Electric Energy Agency (ANEEL) on the *website* (<https://portalrelatorios.aneel.gov.br/luznatarifa/basestarifas>) of the following energy distributors: Companhia Paulista de Força e Luz (CPFL Paulista), Companhia Piratininga de Força e Luz (CPFL Piratininga), Companhia Luz and Força Santa Cruz (CLFSC) and Elektro – Eletricidade e Serviços S/A. As for the amounts of energy tariffs (TUSD + TE), as of 2010, in the *link*: <https://portalrelatorios.aneel.gov.br/luznatarifa/basestarifas> in "TUSD and TE/Approved Tariff Data", they were also calculated by the arithmetic average, year by year, of the three B2 subgroups mentioned above of the subsequent distributors: CPFL Paulista, CPFL Piratininga, CPFL Santa Cruz, Energias de Portugal (EDP) São Paulo, Enel Distribuição São Paulo, Energisa Sul Sudeste (ESS) and Neoenergia Elektro.
- Independent variable X3 (diesel fuel consumption). From BEESP 2019 (base year 2018), associated with BEESP 2023 (base year 2022), information was

obtained on the energy sources consumed by the aforementioned sector, from which Graph 1 was plotted in order to evaluate the evolution of the consumption of each energy source and its dependence by the sector. The highlight was electricity, diesel oil and firewood. The other energy sources consumed (denominated in the graph as "others") are composed of the sum of three energy sources: LPG, kerosene and fuel oil. In Graph 1, it is possible to observe the decrease in the consumption of diesel oil from 2007, which, until 2016, is the most used energy source by the rural sector of São Paulo. On the other hand, there is the growing use of electricity, being the most consumed energy in the last six years.

Graph 1 - Evolution of energy drink consumption, in toe³, by the agricultural sector of São Paulo



Source: Prepared by the author from São Paulo-SIMA (2019) and São Paulo-SEMIL (2023).

In view of the above, the independent variable X3 was assigned to the consumption of diesel oil, in one thousand (103) cubic meters (m3), of São Paulo agribusiness, and its volumes were obtained in two stages. The first, between the years 2000 and 2012, of BEESP 2019 (base year 2018) available at the [link](https://smastr16.blob.core.windows.net/2001/2023/12/BEESP2019ab2018.pdf):

<https://smastr16.blob.core.windows.net/2001/2023/12/BEESP2019ab2018.pdf>. The second, from 2013 to 2022, was collected from BEESP 2023: Base Year 2022

³ According to São Paulo-SEMIL (2023), the unit of energy used in some energy balance reports is the ton of oil equivalent (toe), with one toe being equivalent to 10x10⁹ calories or 10 Giga calories (10 Gcal).

(<https://smastr16.blob.core.windows.net/2001/2024/02/BEESP2023ab2022-2a-edicao-2.pdf>).

- Independent variable X4 [Biomass TPP⁴, in kilowatt (kW), from the state of São Paulo]. It is the sum of the installed electrical power of the thermoelectric generating plants in São Paulo that use the following energy compounds: sugarcane bagasse; biogas; biogas-AGR⁵; biogas-UK⁶; coal-UK; rice husks; firewood; black liquor (lye⁷); wood waste and forest residues. This information was collected from the digital files "Executive Summary: Energy production data from renewable sources" from the years 2013 to 2023 (https://dadosenergeticos.energia.sp.gov.br/portalcev2/intranet/BiblioVirtual/renovaveis/resumo_executivoRE.pdf).
- Independent variable X5 (Photovoltaic Generation, in kW, in the state of São Paulo): the same source cited in the previous variable (X4) was used to gather the information on the sum of the installed electrical power of the solar plants in São Paulo.
- Independent variable X6 [Gross Domestic Product (GDP) of Brazilian agribusiness]: information on the GDP of Agribusiness in the State of São Paulo is not yet available in the *site*; for this reason, the values of the Brazilian Agribusiness GDP were collected in the *link*: [https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fcepea.esalq.usp.br%2Fupload%2Fkceditor%2Ffiles%2FPlanilha_PIB_Cepea_Portugues_Site%2520\(1\)\(4\).xlsx&wdOrigin=BROWSELINK](https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fcepea.esalq.usp.br%2Fupload%2Fkceditor%2Ffiles%2FPlanilha_PIB_Cepea_Portugues_Site%2520(1)(4).xlsx&wdOrigin=BROWSELINK) in the "GDP" tab of the spreadsheet Center for Advanced Studies in Applied Economics (CEPEA). The result of this GDP is the sum, in millions of current reais, of four factors, which are:
 - Inputs used in agriculture;
 - Basic agricultural production;
 - Agro-industrial production;
 - Agricultural services.

⁴ Generating units that use renewable sources for the production of electricity.

⁵ AGR: agricultural waste.

⁶ RU: municipal solid waste.

⁷ Liquid waste from the pulp & paper sector.

The data collected produced historical series for the target sector of the study, between 2000 and 2022. The amounts related to energy expenditures in the years 2020 and 2021 may have been affected by the COVID-19 pandemic.

The dependent variable Y (electricity consumption), as well as the independent variables X1 (number of consumers), X2 (electricity tariff) and X3 (diesel oil consumption) are pertinent to São Paulo agribusiness, while the independent variables X4 (electricity generation by thermoelectric plants that use biomass) and X5 (electricity generation by photovoltaic plants) refer to the state of São Paulo; finally, the independent variable X6 (gross domestic product) is allusive to national agribusiness.

TESTING THE SIGNIFICANCE OF THE COEFFICIENTS

In "Data" (ribbon tab) within the Microsoft Excel® spreadsheet editor, the "Data Analysis" add-in is chosen and, subsequently, the "Regression" analysis tool is selected, where the variable Y is inserted, as well as from X1 to X6, thus obtaining Table 1.

Table 1, in red, we have the independent variable X4 with a high amount for a P-value of approximately 81.81%, much higher than the adopted significance level (α) of 5%. In addition, in the 2nd part of Table 1, in "Coefficients", the lower and upper limits of the confidence interval are presented in the columns "lower 95%" and "upper 95%", which can be expressed as: $P(-0.192 \leq \beta_4 \leq +0.240)$. The test described in item 2.3 indicates the non-rejection of hypothesis $H_0: \beta_4 = 0$. The non-rejection of this hypothesis is confirmed by the confidence interval that covers the value zero.

Considering the above, it is admitted that there is no regression in the electricity consumption of the São Paulo agribusiness sector on the installed capacity in the Biomass TPPs in the state of São Paulo. Therefore, X4 is not statistically significant for this regression model, and is discarded.

Table 1 - Evaluation of the P-values of the six independent variables

| | Coefficientes | valor-P | 95% inferiores | 95% superiores |
|---|----------------------|-----------------|----------------|----------------|
| Interseção | 358.982,066 | 0,692462 | -1.530.548,003 | 2.248.512,135 |
| X1: Quantidade de Consumidores | 13,674 | 0,003160 | 5,320 | 22,029 |
| X2: Tarifa de Energia Elétrica (R\$/MWh) | -5.017,753 | 0,011741 | -8.758,809 | -1.276,696 |
| X3: Consumo de Óleo Diesel (10^3 m^3) | -807,850 | 0,222354 | -2.156,661 | 540,961 |
| X4: UTE Biomassa (kW) | 0,024 | 0,818148 | -0,192 | 0,240 |
| X5: Usina Fotovoltaica (kW) | 1,783 | 0,043915 | 0,055 | 3,511 |
| X6: PIB (milhões R\$) | 0,536 | 0,270721 | -0,460 | 1,531 |

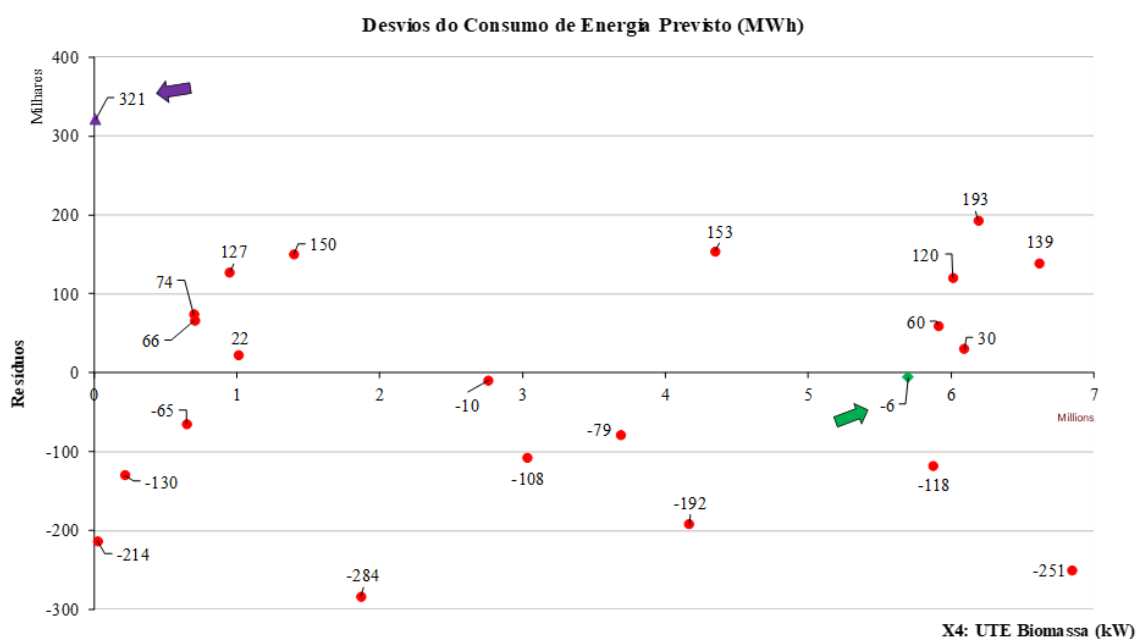
Source: Prepared by the author.

Graph 2 plots the Waste (MWh) in relation to the installed capacity (kW) of the São Paulo TPPs that use biomass as fuel. Residuals are the differences between the actual values of energy consumption and the values estimated by the adjusted model.

There was an alternation between the positive points (12) and the negative points (11), in a total of 23 points (years collected), signaling the absence of a pattern, which allows them to be considered random.

The farthest point (in purple), shown in Graph 2, has an estimated consumption of approximately 321,000MWh lower than the actual one, while the closest point (in green) between the actual and estimated consumption is negative 6,000MWh, that is, the expected consumption, in this case, is higher than the actual one.

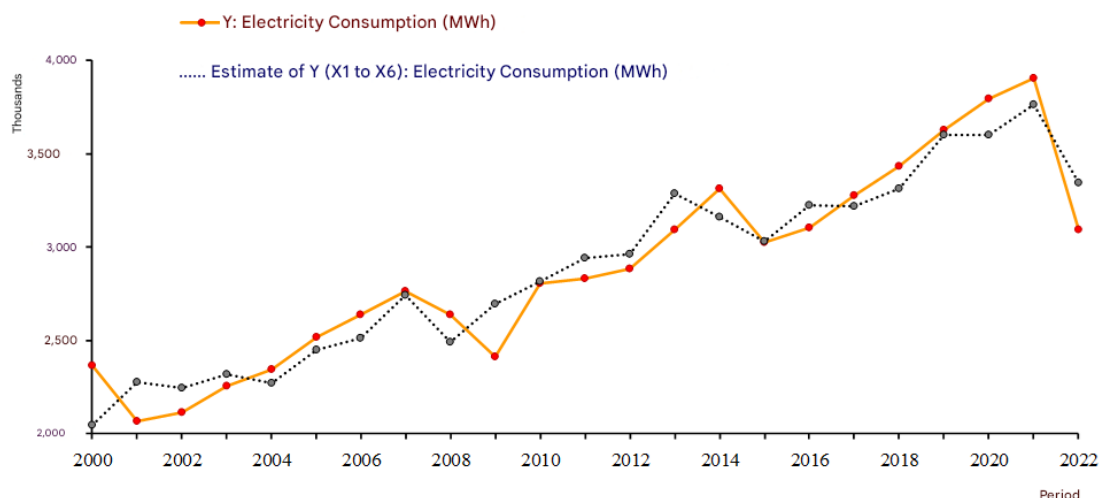
Graph 2 - Dispersion of waste from energy consumption with the power of the Biomass TPPs



Source: Prepared by the author.

Graph 3 shows two graphic lines for the same time interval of 23 years, with the continuous graphic line (in orange) being the actual or observed energy consumption by the rural sector of São Paulo, in MWh, in contrast to the dotted line (in black), which is an estimate of consumption in this sector, as a function of the six independent variables surveyed (X1 to X6), using the multiple linear regression technique. You can observe errors between the points of the actual line and the estimated line.

Graph 3 - Graphs of actual and estimated energy consumption for six variables



Source: Prepared by the author.

The "Regression" analysis tool was again performed for variable Y, as well as all independent variables, with the exception of X4, which was discarded, generating Table 2. In it, described in red, is the variable X6 with an approximate value for a P-value of 11.48%, which is higher than the significance level of 5%.

Also observing the limits of the "bottom 95%" and "top 95%" columns for the confidence interval of X6 equal to $P(-0.162 \leq \beta_6 \leq +1.367)$, this means that β_6 passes through zero, indicating the non-rejection of hypothesis $H_0: \beta_0 = 0$, described in item 2.3.

For the reasons listed above, it is conceivable that there is no regression of the electricity consumption of São Paulo agribusiness on the GDP of Brazilian agribusiness.

Table 2 - Evaluation of the coefficients for five independent variables

| Coefficientes | | valor-P | 95% inferiores | 95% superiores |
|---|-------------|----------|----------------|----------------|
| Interseção | 353.073,910 | 0,688524 | -1.473.692,129 | 2.179.839,948 |
| X1: Quantidade de Consumidores | 14,102 | 0,000660 | 6,945 | 21,259 |
| X2: Tarifa de Energia Elétrica (R\$/MWh) | -5.053,089 | 0,008820 | -8.658,034 | -1.448,145 |
| X3: Consumo de Óleo Diesel (10^3 m^3) | -921,762 | 0,033122 | -1.760,465 | -83,058 |
| X5: Usina Fotovoltaica (kW) | 1,696 | 0,027905 | 0,208 | 3,185 |
| X6: PIB (milhões R\$) | 0,602 | 0,114887 | -0,162 | 1,367 |

Source: Prepared by the author.

The model was reconstructed for the last time, this time, inserting variable Y, as well as all variables X, with the exception of variables X4 and X6, which were excluded because they were not statistically significant.

The program generated Table 3, whose P-value amounts are highlighted in orange, with approximate values equal to: 0.014% for X1, 3.124% for X2, 0.264% for X3 and

1.169% for X5. These four variables were statistically accepted, considering a margin of error (significance level) of 5%, recognizing the existence of regression of energy consumption on these four variables.

Table 3 - Data analysis with four independent variables
RESUMO DOS RESULTADOS

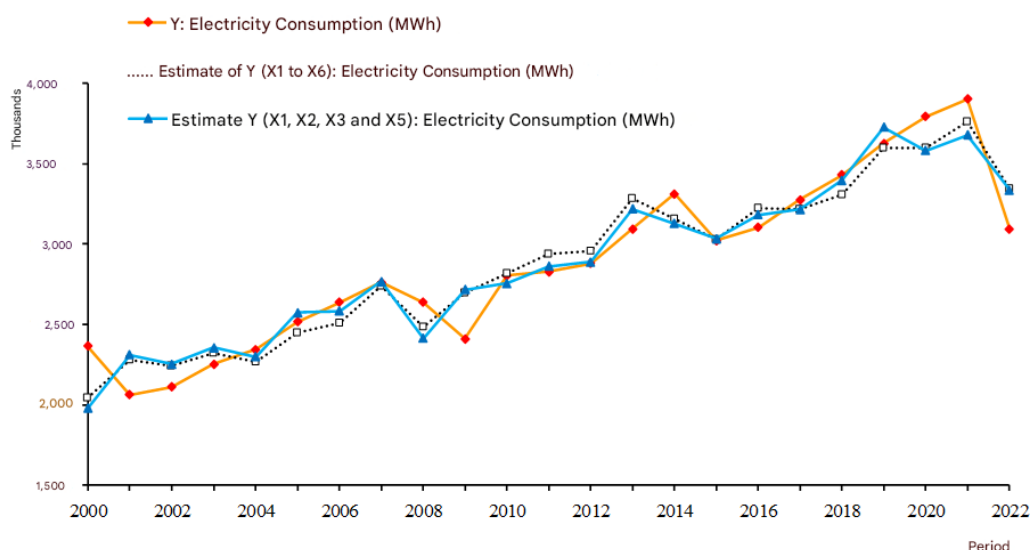
| Estatística de regressão | | |
|--|-------------|-------------------|
| R múltiplo | | 0,9465 |
| R-Quadrado | | 0,8958 |
| R-quadrado ajustado | | 0,8726 |
| Erro padrão | | 185.374,5251 |
| Observações | | 23 |
| ANOVA | | |
| | gl | F de significação |
| Regressão | 4 | 1,31415E-08 |
| Resíduo | 18 | |
| Total | 22 | |
| Coeficientes | | valor-P |
| Interseção | 319.305,420 | 0,728883 |
| X1: Quantidade de Consumidores | 16,044 | 0,000141 |
| X2: Tarifa de Energia Elétrica (R\$/MWh) | -3.529,563 | 0,031241 |
| X3: Consumo de Óleo Diesel (10 ³ m ³) | -1.254,294 | 0,002643 |
| X5: Usina Fotovoltaica (kW) | 2,002 | 0,011692 |

Source: Prepared by the author.

In Graph 4, three graphic lines are plotted for the period from 2000 to 2022 (23 years). The graphic line in orange reveals the variation in the actual or observed electricity consumption (MWh) of the agricultural sector in São Paulo, while the dotted line in black is a forecast of the sector's consumption, in MWh, as a function of all the independent variables selected (X1 to X6).

The blue line in the graph is the forecast of electricity consumption (MWh) of the agricultural sector in São Paulo, for the same period in question, considering four independent variables (X1, X2, X3 and X5). The plotted points, both on the dotted line in black and on the solid line in blue, were obtained using the statistical technique of Multiple Linear Regression.

Graph 4 - Graphs of actual and projected energy consumption



Source: Prepared by the author.

GENERATION OF THE MULTIPLE LINEAR REGRESSION MODEL

The intersection coefficients (b_0) and the independent variables X_1 (b_1), X_2 (b_2), X_3 (b_3) and X_5 (b_5) were collected from Table 3 (in blue), as follows:

- $b_0 = + 319,305.420 = >$ estimated energy consumption, not associated with the independent variables;
- $b_1 = + 16.044 \Rightarrow$ estimated increase in energy consumption for each additional consumer, keeping the other independent variables constant;
- $b_2 = - 3,529.563 = >$ estimated decrease in energy consumption for each increase in the electricity tariff, in R\$/MWh), keeping the other variables unchanged;
- $b_3 = - 1,254.294 \Rightarrow$ estimated decrease in energy consumption for each additional 103 m3 of diesel oil, keeping the other independent variables constant;
- $b_5 = + 2.002 \Rightarrow$ estimated increase in energy consumption for each kilowatt (kW) of additional power in the photovoltaic plants in São Paulo, keeping the other independent variables constant.

Substituting the above five coefficients into the Equation (4), we have the Equation (5):

$$\hat{Y} = 319305,42 + 16,044 \cdot X_1 - 3529,563 \cdot X_2 - 1254,294 \cdot X_3 + 2,002 \cdot X_5 \quad (5)$$

The Equation (5) is the adjusted statistical model of multiple linear regression for the prediction of electricity consumption in the agricultural sector of São Paulo, considering the variables selected in this research.

This multiple linear regression model predicts for the sector researched:

- Approximate increase of 16.044MWh in electricity consumption for each increase of 1 (one) consumer;
- Reduction of about 3,529.563MWh in energy consumption for each increase of R\$ 1.00 per MWh in the electricity tariff;
- Reduction of around 1,254.294MWh in energy consumption for each increment of 1 (one) 103 m³ of diesel oil used;
- An increase of close to 2.002MWh in energy consumption for each unit of kW, plus power in the photovoltaic plants in São Paulo.

TESTING THE EXISTENCE OF REGRESSION

Table 3 (in green) shows the approximate value of "F for significance" of 1.314×10^{-6} %. As "F for meaning" is less than α (significance level adopted at 5%), it is concluded that there is regression, that is, the explanatory variables (X1, X2, X3 and X5), together, have a statistically relevant association that can explain and predict Y (electricity consumption), according to the description of item 2.4.

ANALYSIS OF THE RESULTS OF THE MULTIPLE LINEAR REGRESSION MODEL

Quality of Fit

Linear correlation coefficient

The Correlation Coefficient (R multiple) is highlighted in pink in Table 3, with a value equal to 0.9465. This amount suggests a great similarity in behavior between the variables, pointing in the same direction. In other words, there is a considerable positive correlation between the variables (MARTINS; DOMINGUES, 2014).

Coefficient of Determination or Explanation

Extracting from Table 3 the Coefficient of Determination or Explanatory Power (R-Squared), in violet, with a value equal to 0.8958, and using the rounding to one decimal place, the amount of R² becomes 0.9.

It is found that the proportion in Y (electricity consumption), which is explained, together, by X1 (number of consumers), X2 (electricity tariff), X3 (diesel oil consumption) and X5 (electricity generation by photovoltaic plants) through the linear regression model obtained, has a high explanatory power ($0.9 \leq R^2 < 1.0$), according to Martins and Domingues (2014).

Therefore, the variables X1, X2, X3 and X5, together, through the model, explain 89.58% of the changes or modifications that occurred in Y. The remainder (10.42%) can be attributed to other eventual variables (not included in this model).

Standard Estimate Error

Table 3 shows the value of the "Standard Error" close to 185,374MWh (highlighted in gold). In this research, the actual values (observed) are on average 185,374MWh away from the linear regression line of this model.

Table 4 describes the percentage errors of consumption estimated by the Equation (5) of the adjusted model compared to the actual (observed) energy consumption values in the sector.

Table 4 - Errors between actual consumption and those estimated by the model

| Observation | Error (%) |
|-------------|-----------|
| 1 | -16,26 |
| 2 | 12,04 |
| 3 | 6,73 |
| 4 | 4,61 |
| 5 | -1,83 |
| 6 | 2,34 |
| 7 | -2,03 |
| 8 | 0,15 |
| 9 | -8,50 |
| 10 | 12,63 |
| 11 | -1,81 |
| 12 | 1,17 |
| 13 | 0,37 |
| 14 | 4,05 |
| 15 | -5,55 |
| 16 | 0,40 |
| 17 | 2,56 |
| 18 | -1,80 |
| 19 | -0,96 |
| 20 | 2,77 |
| 21 | -5,56 |
| 22 | -5,69 |
| 23 | 7,82 |

Source: Prepared by the author.

The standard error is another way to measure the accuracy of the calculation, being a kind of standard deviation with respect to the regression line. The lower the standard error of the estimate, the more accurate the forecast model will be.

And finally, in the "Observations" line, the number 23 is verified, which refers to the size of the sample surveyed (23 years).

Residual analysis (ϵ)

The period covered by the research is 23 years (23 observations), and the information in the "Residuals" column, highlighted in brown (Table 5), was taken from the "Regression" tool of Microsoft Excel®.

Table 5 - Teste de Shapiro-Wilk

| Observation | Waste | Teste de Shapiro-Wilk | |
|-------------|--------------|-----------------------|-------------|
| 1 | 384.528,969 | Sample size | 23 |
| 2 | -248.358,280 | Average | 0,000 |
| 3 | -142.160,421 | Standard deviation | 167.677,567 |
| 4 | -103.973,389 | ln | 0,9725 |
| 5 | 42.852,549 | p | 0,7206 |
| 6 | -58.843,459 | | |
| 7 | 53.570,456 | | |
| 8 | -4.245,696 | | |
| 9 | 224.231,042 | | |
| 10 | -304.571,083 | | |
| 11 | 50.691,540 | | |
| 12 | -32.989,903 | | |
| 13 | -10.712,205 | | |
| 14 | -125.362,612 | | |
| 15 | 183.952,766 | | |
| 16 | -12.102,422 | | |
| 17 | -79.460,857 | | |
| 18 | 59.001,669 | | |
| 19 | 32.973,947 | | |
| 20 | -100.334,068 | | |
| 21 | 210.936,122 | | |
| 22 | 222.182,276 | | |
| 23 | -241.806,940 | | |

Source: Adapted by the author of the BioStat® software.

The Shapiro-Wilk test, using the BioEstat® software, version 5.3⁸, is performed to calculate the mean probability distribution of the variable \mathcal{E} (item 2.5.3, part a), as well as to verify its normality (item 2.5.3, part c).

The result of the mean probability distribution of the variable \mathcal{E} is equal to zero (highlighted in red) in Table 5.

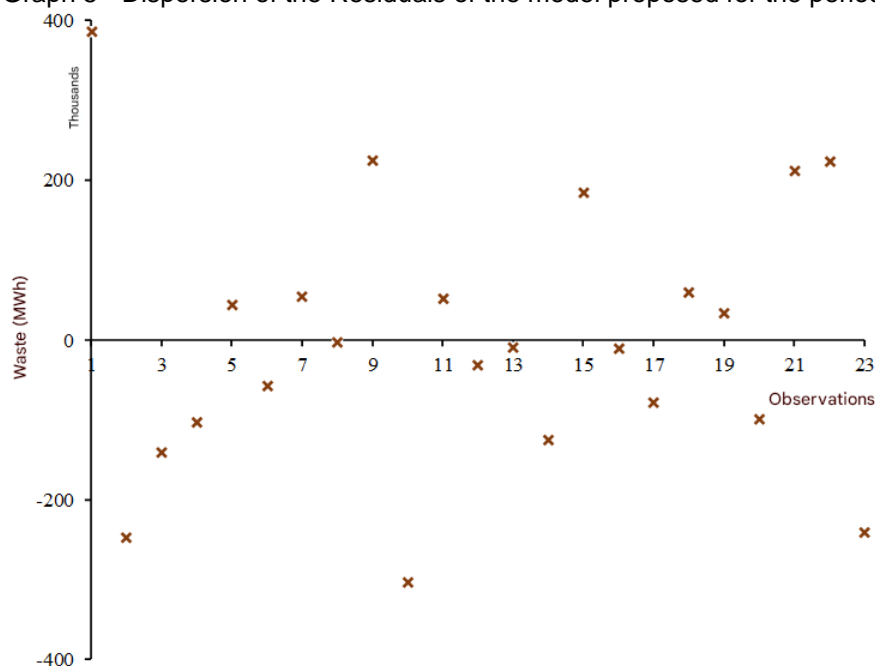
Regarding the verification of the normality of the probability distribution of the variable \mathcal{E} , according to Table 5, the amount of p-value is 0.7206 (highlighted in green), i.e., 72.06%. As a $> \alpha$ -value for a significance level (α) of 5%, hypothesis H_0 is not rejected. Therefore, there are indications that the distribution of errors or residues is normally distributed.

It is also noticed, according to Table 5 (in brown), the alternation of positive (ten) and negative (thirteen) signs of the "Residuals", confirming a good fit of the model.

⁸ <https://www.mamiraua.org.br/downloads/programas/>

In Graph 5, the residual values are randomly plotted around the Observations axis (23 years), without showing any type of pattern. Therefore, it is deduced that the residuals are independent of each other, noting only the effect of the independent variables on the dependent variable, which means the absence of residual autocorrelation.

Graph 5 - Dispersion of the Residuals of the model proposed for the period



Source: Prepared by the author.

Ascertaining the existence of multicollinearity between the dependent variable and the selected independent variables

The information of variables Y, X1, X2, X3 and X5 were transported to the Microsoft Excel® spreadsheet editor, where the configuration of the following menu was used: Data => Data Analysis => Correlation => OK, creating Table 6.

Analyzing the table, it is noted that the dependent variable (Y) has, in decreasing order of correlation intensity (in module), a strong correlation with the variable X3 of 87.28%, followed by the variable X1, with 80.87%, X2, with 75.63% and, finally, the lowest correlation, with a percentage of 60.89%, with the variable X5.

According to Martins and Domingues (2014), the ordering of the independent variables described in the previous paragraph will be the same in terms of the level of importance of contribution to a good response of the regression analysis under study, i.e., the variable X3 is the most significant, followed by X1 and X2, determining X5 as the lowest participation in the regression.

According to Gil (2021), the correlations in Table 6 can be described by phrases, such as:

- The correlation between Y and X3 of 87.28% (negative) is characterized as a very strong negative correlation because it is in the range between -70% and -99%;
- The positive correlations between Y and X1 of 80.87% and between Y and X2 equal to 75.63% are portrayed as very strong positive correlations because they are inserted between 70% and 99%;
- The correlation between Y and X5 of 60.89% (positive) is specified as a substantial positive correlation because it is between 50% and 69%.

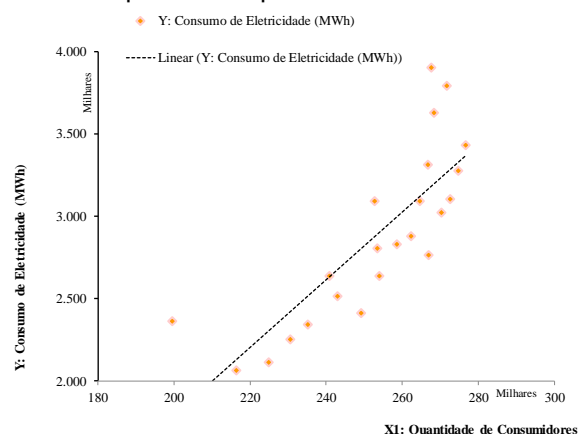
Table 6 - Correlation between the dependent variable in relation to the independent variables

| Correlation | Y: |
|---------------------------------|-------------------------------|
| | Electricity Consumption (MWh) |
| Y: | 1 |
| Electricity Consumption (MWh) | |
| X1: | + 0,8087 |
| Number of Consumers | |
| X2: | + 0,7563 |
| Electricity Tariff (R\$/MWh) | |
| X3: | - 0,8728 |
| Diesel Oil Consumption (103 m3) | |
| X5: | + 0,6089 |
| Photovoltaic Power Plant (kW) | |

Source: Prepared by the author.

Graph 6, which represents the energy consumption of the São Paulo agribusiness sector relative to the number of its consumers, can be seen as a positive correlation, i.e., the variables are positively related and grow in the same direction (MARTINS; DOMINGUES, 2014), with a few scattered points around the black dotted linear trend line.

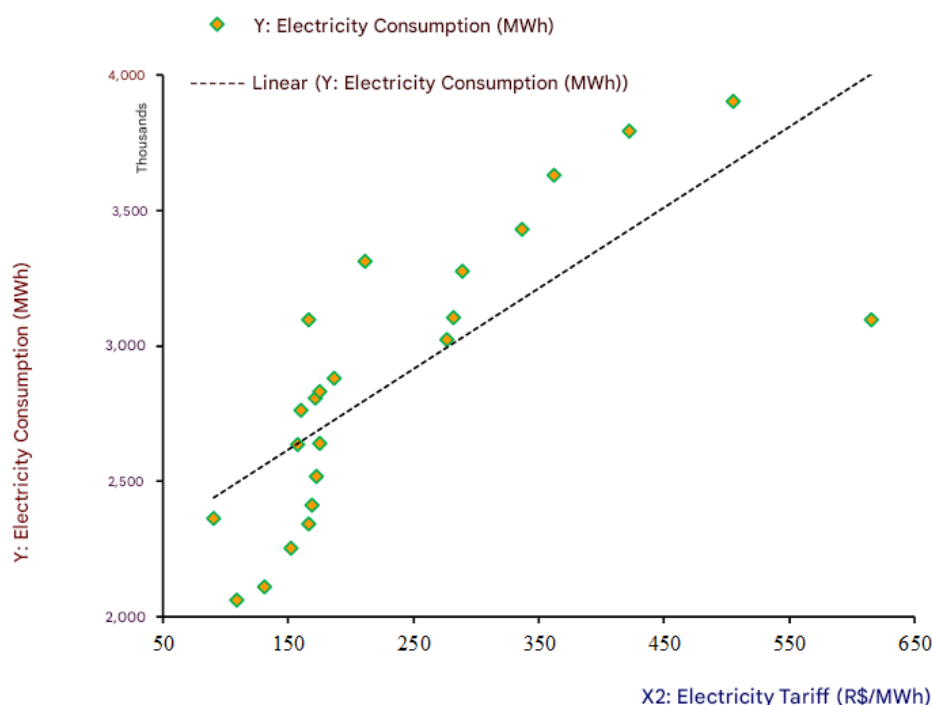
Graph 6 - Y dispersion relative to X1



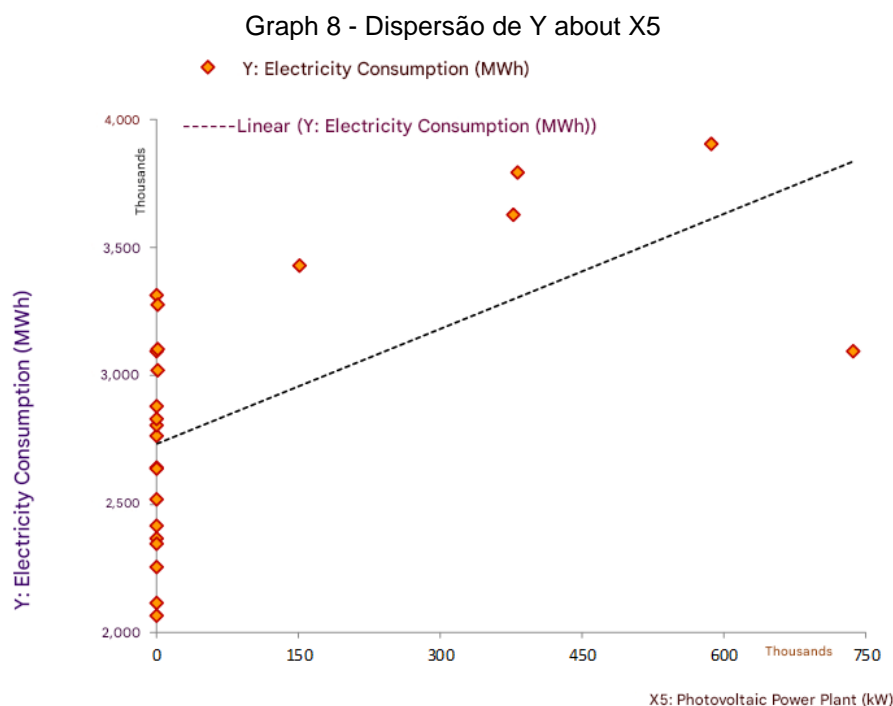
Source: Prepared by the author.

As for Graph 7, which illustrates the energy consumption of the rural sector in São Paulo in relation to the electricity tariff, and Graph 8 on the energy consumption of this sector in relation to the installed capacity of photovoltaic plants, both have a positive correlation, but with many scattered points in reference to the dotted linear trend lines (SPIEGEL; STEPHENS, 2009).

Graph 7 - Dispersion of Y with respect to X2



Source: Prepared by the author.

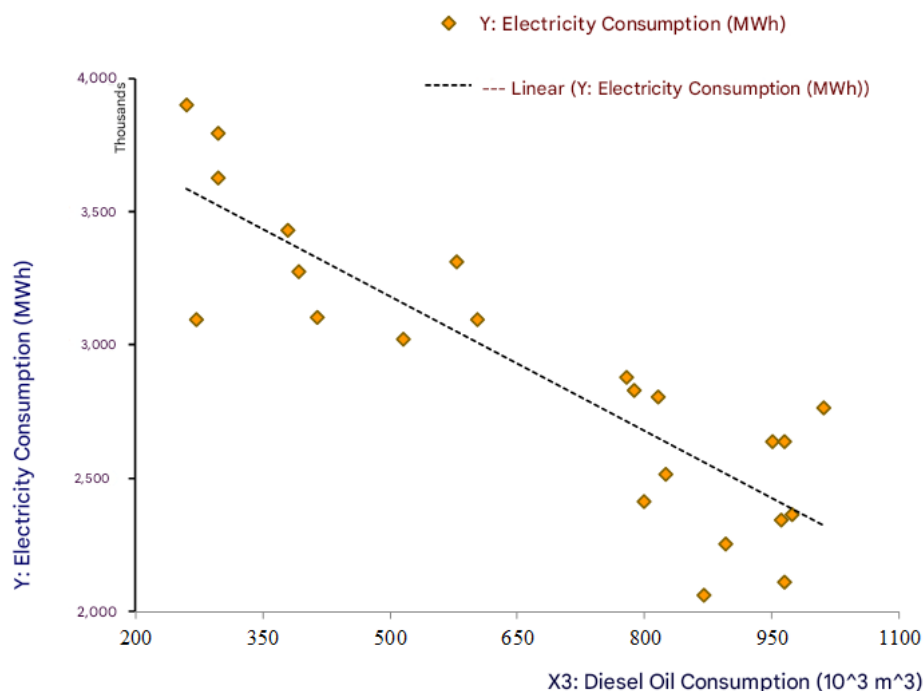


Source: Prepared by the author.

In the case of Graph 9, which deals with the consumption of electricity by the São Paulo agribusiness sector related to the consumption of diesel oil in this sector, the correlation is negative. This means that the growth of one variable occurs with the decrease of the other (SPIEGEL; STEPHENS, 2009).

There is dispersion of the dots in reference to the dotted black linear trend line.

Graph 9 - Y dispersion in reference to X3



Source: Prepared by the author.

Verifying the existence of multicollinearity between the independent variables

Running the "Correlation" analysis tool again, on this occasion, with only the information of the independent variables X1, X2, X3 and X5, Table 7 was reproduced.

According to Gil (2021), the variable X2 has a very strong positive correlation with X5 equal to 92.34%, as it is inserted between 70% and 99%, while the values of the correlations between X2 and X3 (-87%) and between X3 and X5 (-70.53%) are characterized as very strong negative correlations because they are located between -70% and -99%.

From the above, it can be seen that X2 is strongly correlated with X3 (87%) and, mainly, with X5 (92.34%). This flaw is called Multicollinearity, which, according to Martins and Domingues (2014), presents an extra obstacle in the analysis, since the final results are masked, especially the hypothesis tests due to the influence of one independent variable on the other. In addition, this distortion also leads to the violation of the premises of the model.

Variable X1 has a substantial negative correlation with X3 (-67.90%) and a substantial positive correlation with X2 (53.20%), as observed in Table 7, as its values are in the range of -50% to -69% and 50% to 69%, respectively (GIL, 2021).

Regarding the 26.07% correlation between X1 and X5 (Table 7), Gil (2021) reports that it is a low correlation because it is between 10% and 29%.

Table 7 - Correlation between independent variables

| Correlation | X1: | X2: | X3: | X5: |
|---------------------------------|----------|----------|----------|-----|
| X1: | 1 | | | |
| Number of Consumers | | | | |
| X2: | + 0,5320 | 1 | | |
| Electricity Tariff (R\$/MWh) | | | | |
| X3: | - 0,6790 | - 0,8700 | 1 | |
| Diesel Oil Consumption (103 m3) | | | | |
| X5: | + 0,2607 | + 0,9234 | - 0,7053 | 1 |
| Photovoltaic Power Plant (kW) | | | | |

Source: Prepared by the author.

Pearson's Correlation Test, using the BioStat® software, was also used to investigate the correlation between the independent variables.

Evaluating the correlation between variables X1 and X2, Pearson's correlation coefficient, r (Pearson), with a value equal to 0.5320, was highlighted in green (Table 8). Since the coefficient is in the range $0 < r < 1$, it means that it is a positive linear correlation, in which the variables change in the same direction.

Since the value of r (Pearson) is within the range of $0.50 < r < 0.69$, it is possible to conclude that this is a substantial correlation. Regarding the p-value (in light blue in Table 8), it is as follows: $0.0089 < 0.05$, rejecting hypothesis H_0 for a risk (significance level) of 5%, as described in item 2.5.4; therefore, there are indications that the variables X1 and X2 are correlated.

Table 8 - Correlation between the number of consumers and the electricity tariff

| X1 | X2 (R\$/MWh) | Linear Correlation Test | |
|---------|--------------|-------------------------|----------|
| 199.539 | 90,30 | n (parents) | 23 |
| 216.269 | 109,45 | r (Pearson) | + 0,5320 |
| 224.846 | 130,92 | (p) | 0,0089 |
| 230.601 | 152,55 | | |
| 235.127 | 166,19 | | |
| 243.022 | 172,67 | | |
| 253.998 | 175,00 | | |
| 266.882 | 160,58 | | |
| 240.722 | 158,03 | | |
| 249.125 | 169,32 | | |
| 253.394 | 171,71 | | |
| 258.543 | 174,96 | | |
| 262.243 | 186,66 | | |
| 264.528 | 166,73 | | |
| 266.735 | 211,42 | | |
| 270.180 | 276,62 | | |
| 272.649 | 281,68 | | |
| 274.610 | 288,71 | | |
| 276.707 | 336,77 | | |
| 268.339 | 362,69 | | |
| 271.668 | 421,97 | | |
| 267.680 | 505,28 | | |
| 252.614 | 615,37 | | |

Source: Adapted by the author of the *BioStat®* software.

Regarding the correlation between variables X1 and X3, according to the purple highlight in Table 9, the correlation coefficient r (Pearson) is 0.6790 (negative). In this case, the coefficient is in the range $-1 < r < 0$, showing a negative linear correlation, in which the variables change in the opposite direction.

The amount of r (Pearson) is between $-0.50 < r < -0.69$, showing that this is a substantial correlation. Considering the p-value (in light pink in Table 9) equal to 0.0004, where $0.0004 < 0.05$, hypothesis H0 is rejected for a risk of 5%, as specified in item 2.5.4; therefore, there are indications that the variables X1 and X3 are correlated.

Table 9 - Correlation between the number of consumers and the consumption of diesel fuel

| X1 | X3 (103 m3) | Linear Correlation Test | |
|---------|-------------|-------------------------|----------|
| 199.539 | 974 | n (parents) | 23 |
| 216.269 | 870 | r (Pearson) | - 0,6790 |
| 224.846 | 965 | (p) | 0,0004 |
| 230.601 | 895 | | |
| 235.127 | 960 | | |
| 243.022 | 824 | | |
| 253.998 | 950 | | |
| 266.882 | 1.010 | | |
| 240.722 | 965 | | |
| 249.125 | 799 | | |
| 253.394 | 816 | | |
| 258.543 | 787 | | |
| 262.243 | 779 | | |
| 264.528 | 602 | | |
| 266.735 | 578 | | |
| 270.180 | 515 | | |
| 272.649 | 414 | | |
| 274.610 | 392 | | |
| 276.707 | 379 | | |
| 268.339 | 297 | | |
| 271.668 | 297 | | |
| 267.680 | 260 | | |
| 252.614 | 271 | | |

Source: Adapted by the author of the *BioStat®* software.

Checking the correlation between variables X1 and X5, the correlation coefficient r (Pearson) of 0.2607 is highlighted in orange (Table 10). As the coefficient is in the range of $0 < r < 1$, there is a positive linear correlation and, consequently, the variables convert to the same path.

The value of the coefficient is in the range of $0.10 < r < 0.29$; Therefore, this is a low correlation. Table 10 highlights the p-value of 0.2294, i.e., $0.2294 > 0.05$; therefore, hypothesis H0 is not rejected for a risk (significance level) of 5%, indicating that the variables X2 and X3 are not correlated.

Table 10 - Correlation between the number of consumers and the power of photovoltaic plants

| X1 | X5 (kW) | Linear Correlation Test | |
|---------|---------|-------------------------|----------|
| 199.539 | 0 | n (parents) | 23 |
| 216.269 | 0 | r (Pearson) | + 0,2607 |
| 224.846 | 0 | (p) | 0,2294 |
| 230.601 | 0 | | |
| 235.127 | 0 | | |
| 243.022 | 0 | | |
| 253.998 | 0 | | |
| 266.882 | 0 | | |
| 240.722 | 0 | | |
| 249.125 | 0 | | |
| 253.394 | 0 | | |
| 258.543 | 0 | | |
| 262.243 | 0 | | |
| 264.528 | 0 | | |
| 266.735 | 0 | | |
| 270.180 | 1.100 | | |
| 272.649 | 1.100 | | |
| 274.610 | 1.100 | | |
| 276.707 | 151.217 | | |
| 268.339 | 377.426 | | |
| 271.668 | 382.426 | | |
| 267.680 | 587.064 | | |
| 252.614 | 736.887 | | |

Source: Adapted by the author of the *BioStat®* software.

The amount of the correlation coefficient r (Pearson), highlighted in marine in Table 11, is 0.8700 (negative) and is in the range $-0.50 < r < -0.69$, proving a very strong negative linear correlation (the variables move in opposite directions).

With a p-value (in green in Table 11) lower than 0.0001 and, therefore, less than 0.05, hypothesis H0 is rejected for a risk of 5% (detailed in item 2.5.4); That said, there is evidence that the variables X2 and X3 are correlated.

Table 11 - Correlation between electricity tariff and diesel fuel consumption

| X2 (R\$/MWh) | X3 (103 m3) | Linear Correlation Test | |
|--------------|-------------|-------------------------|----------|
| 90,30 | 974 | n (parents) | 23 |
| 109,45 | 870 | r (Pearson) | - 0,8700 |
| 130,92 | 965 | (p) | < 0,0001 |
| 152,55 | 895 | | |
| 166,19 | 960 | | |
| 172,67 | 824 | | |
| 175,00 | 950 | | |
| 160,58 | 1.010 | | |
| 158,03 | 965 | | |
| 169,32 | 799 | | |
| 171,71 | 816 | | |
| 174,96 | 787 | | |
| 186,66 | 779 | | |
| 166,73 | 602 | | |
| 211,42 | 578 | | |
| 276,62 | 515 | | |
| 281,68 | 414 | | |
| 288,71 | 392 | | |
| 336,77 | 379 | | |
| 362,69 | 297 | | |
| 421,97 | 297 | | |
| 505,28 | 260 | | |
| 615,37 | 271 | | |

Source: Adapted by the author of the *BioStat®* software.

According to Table 12, Pearson's correlation coefficient, r (Pearson), has a value equal to 0.9234 (highlighted in red). As this value is within the range $0 < r < 1$, it is concluded that it is a very strong positive linear correlation, as the coefficient is in the range between 0.70 and 0.99.

If p-value < 0.0001 (in yellow in Table 12) is less than 5%, the null hypothesis is rejected for the level of significance adopted; therefore, there are indications that the variables X2 and X5 are correlated.

Table 12 - Correlation between electricity tariff and power of photovoltaic plants

| X2 (R\$/MWh) | X5 (kW) | Linear Correlation Test | |
|--------------|---------|-------------------------|----------|
| 90,30 | 0 | n (parents) | 23 |
| 109,45 | 0 | r (Pearson) | + 0,9234 |
| 130,92 | 0 | (p) | < 0,0001 |
| 152,55 | 0 | | |
| 166,19 | 0 | | |
| 172,67 | 0 | | |
| 175,00 | 0 | | |
| 160,58 | 0 | | |
| 158,03 | 0 | | |
| 169,32 | 0 | | |
| 171,71 | 0 | | |
| 174,96 | 0 | | |
| 186,66 | 0 | | |
| 166,73 | 0 | | |
| 211,42 | 0 | | |
| 276,62 | 1.100 | | |
| 281,68 | 1.100 | | |
| 288,71 | 1.100 | | |
| 336,77 | 151.217 | | |
| 362,69 | 377.426 | | |
| 421,97 | 382.426 | | |
| 505,28 | 587.064 | | |
| 615,37 | 736.887 | | |

Source: Adapted by the author of the *BioStat®* software.

Table 13, highlighted in red, shows the correlation coefficient r (Pearson) of X3 and X5 equal to 0.7053 (negative) inserted in the range: $-0.70 < r < -0.99$, presenting a very strong negative linear correlation (they vary in the opposite direction).

As a p-value $< \alpha$, i.e., $0.02\% < 5\%$ (in blue in Table 13), the null hypothesis is rejected for the level of significance adopted (detailed in item 2.5.4); therefore, there is evidence that there is a linear correlation between them.

Table 13 - Correlation between diesel fuel consumption and the power of photovoltaic plants

| X3 (103 m3) | X5 (kW) | Linear Correlation Test | |
|-------------|---------|-------------------------|----------|
| 974 | 0 | n (parents) | 23 |
| 870 | 0 | r (Pearson) | - 0,7053 |
| 965 | 0 | (p) | 0,0002 |
| 895 | 0 | | |
| 960 | 0 | | |
| 824 | 0 | | |
| 950 | 0 | | |
| 1.010 | 0 | | |
| 965 | 0 | | |
| 799 | 0 | | |
| 816 | 0 | | |
| 787 | 0 | | |
| 779 | 0 | | |
| 602 | 0 | | |
| 578 | 0 | | |
| 515 | 1.100 | | |
| 414 | 1.100 | | |
| 392 | 1.100 | | |
| 379 | 151.217 | | |
| 297 | 377.426 | | |
| 297 | 382.426 | | |
| 260 | 587.064 | | |
| 271 | 736.887 | | |

Source: Adapted by the author of the *BioStat®* software.

The variable X1 (number of consumers) was the one that presented the lowest correlation (association) with the other variables, with the following correlation coefficients r (Pearson):

- +53.20% (Table 8) with X2 (electricity tariff, in R\$/MWh);
- -67.90% (Table 9) with X3 (diesel fuel consumption, in 103m3);
- +26.07% (Table 10) with X5 (Photovoltaic Plant, in kW).

On the other hand, it was pointed out that the variable X2 had the highest correlations (associations) with the variables studied, especially with X3 and X5, according to the correlation coefficients r (Pearson) shown below:

- -87.00% (Table 11) with X3 (diesel fuel consumption, in 103m3);
- +92.34% (Table 12) with X5 (Photovoltaic Plant, in kW).

According to Cunha and Coelho (2014), the problem of multicollinearity refers mainly to the degree, and not to nature. The presence of correlation between the independent variables is inevitable, and it is preferable to select those with a lower degree to avoid complications in the interpretation of the results.

The ideal scenario would be to have multiple independent variables that are highly or perfectly correlated with the dependent variable, but with minimal or no correlation between them.

Therefore, it can be considered to discard the variable X2 due to its strong association with the other independent variables.

CONCLUSION

The development of this research provided considerable contributions to the achievement of the objectives proposed here.

The generation of a statistical model to estimate electricity consumption by the agricultural sector in São Paulo, which is the general objective, was achieved, as well as the specific objective, which was to identify the variables that contribute to this estimate.

Six variables were chosen to build the model, three of which were from the São Paulo agribusiness sector: the number of consumers, the electricity tariff and the consumption of diesel oil; two referring to the state of São Paulo: installed power of biomass thermal plants and photovoltaic plants and one of Brazilian agribusiness: GDP.

Of these, two variables were discarded: the installed power of the biomass thermal plants in São Paulo and the GDP of Brazilian agribusiness, because the linear regression model considered them statistically non-significant. A possible cause for the exclusion of this variable is the fact that the national GDP, and not the São Paulo GDP, of a continental country such as Brazil, which is a major exporter of *agricultural and livestock* commodities, was considered.

The other variables were part of the statistical model, but with different levels of influence. The total installed power of the photovoltaic plants in São Paulo showed a substantial influence, while the others – number of consumers, electricity tariff and consumption of diesel oil in the rural area of São Paulo – showed very strong influences on the estimate of electricity consumption in the sector studied.

It was found that the consumption of diesel fuel was the most significant variable studied, but with inverse action (negative sign in the adjusted statistical model) in the multiple linear regression, thus proving its replacement by electricity as the main energy source of the rural sector of São Paulo.

In future studies, the methodology applied here could employ the stepwise variable selection technique, also known as the stepwise method, which is a frequent approach in the sequential search. It is possible, by this method, to examine the additional impact of each independent variable on the model, since the inclusion of each variable is evaluated before the formulation of the mathematical equation.

The present research can contribute to the understanding of elements that stimulate the growth of energy consumption in the agricultural sector of São Paulo, providing tools of analysis and future estimation to those responsible for the state's electricity sector for its permanent maintenance and expansion, thus enabling the continuity of the supply of electricity, a primordial public service, to carry out daily tasks of the rural community of São Paulo.

ACKNOWLEDGMENTS

The present work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Financing Code 001.

I thank the Federal Institute of São Paulo (IFSP) for granting me paid leave for *Stricto Sensu* qualification.

REFERENCES

1. Bisognin, C., & Werner, L. (2020). Análise do consumo mensal de energia elétrica no Estado de São Paulo. REP – Revista de Engenharia de Produção, 2, 59–72. <https://periodicos.ufms.br/index.php/REP/article/view/9397>
2. Brasil. Empresa de Pesquisa Energética (EPE). (2024). Anuário estatístico de energia elétrica 2024: Ano base 2023 (6 p.). Rio de Janeiro: EPE. <https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-160/topico-168/anuario-factsheet-2024.pdf>
3. Brasil. Ministério de Minas e Energia (MME). (2024). Balanço Energético Nacional 2024: Ano base 2023 (274 p.). Rio de Janeiro: EPE. <https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-819/topico-723/BEN2024.pdf>
4. Brasil. Ministério de Minas e Energia (MME). (2007). Plano Nacional de Energia 2030: Combustíveis líquidos (Vol. 12, 98 p.). Brasília: MME. <https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-165/topico-173/PNE%202030%20-%20Combust%C3%ADveis%20L%C3%ADquidos.pdf>
5. Cunha, J. V. A. da, & Coelho, A. C. (2014). Regressão linear múltipla. In L. J. Corrar, E. Paulo & J. M. Dias Filho (Coords.), Análise multivariada: para os cursos de administração, ciências contábeis e economia (1ª ed., 7ª reimpr., pp. 131–231). São Paulo: Atlas.
6. Gil, A. C. (2021). Métodos e técnicas de pesquisa social (7ª ed., 3ª reimpr., 230 p.). São Paulo: Atlas.
7. Martins, G. de A., & Domingues, O. (2014). Estatística geral e aplicada (5ª ed., rev. e ampl., 416 p.). São Paulo: Atlas.
8. Mello, D. (2022, outubro). Lula diz que é preciso melhorar salários de professores. Agência Brasil. <https://agenciabrasil.ebc.com.br/politica/noticia/2022-10/lula-diz-que-e-preciso-melhorar-salarios-de-professores>
9. São Paulo (Estado). Secretaria de Infraestrutura e Meio Ambiente (SIMA). (2019). Balanço Energético do Estado de São Paulo (BEESP) 2019: Ano base 2018 (274 p.). São Paulo: SIMA. <https://smastr16.blob.core.windows.net/2001/2023/12/BEESP2019ab2018.pdf>
10. São Paulo (Estado). Secretaria de Infraestrutura e Meio Ambiente (SIMA). (2022). Balanço Energético do Estado de São Paulo (BEESP) 2022: Ano base 2021 (270 p.). São Paulo: SIMA. <https://dadosenergeticos.energia.sp.gov.br/portalcev2/intranet/BiblioVirtual/diversos/BalancoEnergetico.pdf>

11. São Paulo (Estado). Secretaria de Meio Ambiente, Infraestrutura e Logística (SEMIL). (2024). Balanço Energético do Estado de São Paulo (BEESP) 2023: Ano base 2022 (2ª ed., 160 p.). São Paulo: SEMIL. <https://smastr16.blob.core.windows.net/2001/2024/02/BEESP2023ab2022-2a-edicao-2.pdf>
12. Spiegel, M. R., & Stephens, L. J. (2009). Estatística (4ª ed.) [J. L. do Nascimento, Trad.]. Porto Alegre: Bookman. (Obra original publicada em inglês como Schaum's Outlines: Theory and Problems of Statistics, 4th ed., ISBN 9780071485845)