energies

MDPI

Article

Definition of Regulatory Targets for Electricity Non-Technical Losses: Proposition of an Automatic Model-Selection Technique for Panel Data Regressions

Eduardo Correia [1], Rodrigo Calili [1,*], José Francisco Pessanha [2] and Maria Fatima Almeida [1]

[1] Postgraduate Programme in Metrology, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro 22453-900, Brazil
[2] Institute of Mathematics and Statistics, Rio de Janeiro State University, Rio de Janeiro 20550-000, Brazil
* Correspondence: calili@puc-rio.br

Abstract: Non-technical losses (NTLs) are one of the main problems that electricity distribution utilities face in developing regions such as Latin America, the Caribbean, sub-Saharan Africa, and South Asia. Particularly in Brazil, based on the socioeconomic and market variables concerning all the distribution utilities, the National Electric Energy Agency (ANEEL) has formulated several specifications of econometric models for panel data with random effects, all aimed at determining an index that reflects the difficulty of combating NTLs according to the intrinsic characteristics of each distribution area. Nevertheless, given the exhaustive search for combinations of explanatory variables and the complexity inherent to defining regulatory NTL targets, this process still requires the evaluation of many models through hypothesis and goodness-of-fit tests. In this regard, this article proposes an automatic model-selection technique for panel data regressions to better assist the Agency in establishing NTL regulatory targets for the distribution of utilities in this country. The proposed technique was applied to panel data containing annual observations from 62 Brazilian electricity distribution utilities from 2007 to 2017, thus generating 1,097,789 models associated with the regression types in the panel data. The main results are three selected models that showed more adherence to the actual capacity of Brazilian distribution utilities to reduce their NTLs.

Keywords: non-technical losses; automatic model-selection technique; economic regulation; panel data regression; electricity distribution utilities

Citation: Correia, E.; Calili, R.; Pessanha, J.F.; Almeida, M.F. Definition of Regulatory Targets for Electricity Non-Technical Losses: Proposition of an Automatic Model-Selection Technique for Panel Data Regressions. *Energies* **2023**, *16*, 2519. https://doi.org/10.3390/en16062519

Academic Editors: Alan Brent, Katarzyna Widera, Marcin Rabe and Katarzyna Chudy-Laskowska

Received: 29 December 2022
Revised: 28 January 2023
Accepted: 9 February 2023
Published: 7 March 2023

1. Introduction

The quality of electricity distribution is a crucial factor for industrial competitiveness and society's well-being [1,2]. However, in developing regions such as Latin America and the Caribbean, sub-Saharan Africa, and South Asia, the distribution network infrastructure tends to be more risky, leading to economic and regulatory challenges associated with the payback of investments.

Electricity losses in distribution networks are due to both technical and non-technical factors. The technical losses (TLs) are associated with the physical characteristics of the electrical equipment used in the distribution and are mainly to do with the energy dissipated in the conductors.

Non-technical losses (NTLs), the focus of this study, are losses caused by electricity theft, metering, or billing errors or even by consumer units without metering equipment and are therefore associated with the commercial management of the distribution utilities [3–6].

The growing electricity demand, combined with the complex socioeconomic conditions in countries of continental dimensions, such as Brazil and India, leads to economic and regulatory challenges directly associated with the sustainability of distribution systems and, particularly, with the combating of NTLs. In these countries, the current levels of NTLs result in risks to the economic solvency, the limited investment capacity of the

distribution utilities, increased electricity tariffs for regular customers, and more carbon emissions [7–10]. Because of these negative impacts on the distribution sector, combating non-technical losses has deserved close attention from regulators and distribution companies over time. From the regulatory perspective, the central question concerning loss treatment is the implementation of measures and incentives that reward distribution utilities for decisions that economically limit or reduce the NTL volume and costs.

Thus, in order to broaden the debate on the NTL phenomenon in developing regions, Miranda et al. [11] conducted a comparison of the regulatory experiences in countries from different continents that reveals six major innovative approaches, as follows: (i) regulatory segmentation in countries with large territorial extensions (India, Canada, Australia, and the USA); (ii) complementation of the regulatory recognition of losses in a reference company efficient model; (iii) differentiated regulatory treatment for each group of companies with similar characteristics; (iv) use of regulatory adaptation strategies in periods of economic crisis; (v) differentiated regulatory treatment according to the company's performance; and (vi) specific regulatory treatment for areas of high criminality.

The observed trend in large countries (e.g., India, Canada, Australia, and the USA) is the adoption of regulatory segmentation and individualized mechanisms by concession areas or state regions. The regional or state regulatory agencies in these countries have the autonomy to define their appropriate methodological strategies to reduce NTLs based on general guidelines established by a central institution to ensure regulatory consistency across states [12–14].

The second approach adopted by countries such as Chile, Guatemala, and Peru [15–19] refers to complementing the regulatory recognition of losses in a reference company efficient model. So, the calculation methodologies follow equations for determining the technical losses (TLs). Regarding NTLs, for the low-voltage network, an additional loss percentage is considered for a residual value whose exclusion is not economically reasonable. Based on a reference company model, according to which an efficient loss value is represented as a function of the costs associated with combating the losses, the final values can be calculated. The difference between its applications in the mentioned countries is the maximum percentage allowed for NTLs.

Concerning the third approach—differentiated regulatory treatment for each group of companies with similar characteristics—the concept of 'typical areas' has been employed with the purpose of clustering companies with similar characteristics according to their limitations in combating the electricity losses in their areas. Thus, the regulatory agencies recognize differences in the performance complexity between regions and propose a specific calculation for each group of comparable companies. By way of illustration, the regulation of electricity losses in Italy considers each concession area's reality. Accordingly, specific regulatory treatments have been established, taking into account the socioeconomic differences between the north, center, and south, with higher factors in the south [20].

Some European countries, namely Portugal, Spain, and Italy, have adopted regulatory adaptation strategies in periods of economic crisis. These countries have a fully liberalized electricity market for the final consumer, with the distinction between network operators (Distribution System Operators or DSOs) and energy suppliers (suppliers). According to the regulatory approach implemented in Portugal and Spain, the cost of losses is the responsibility of the suppliers. Therefore, the loss coverage must be considered when negotiating energy contracts. In the suppliers' responsibility model, these agents estimate the losses to be adopted in the contracts and acquire additional energy to cover them. The amount of energy corresponding to the difference between the estimated loss and the current loss is traded at the clearing market equilibrium price. As there is no final regulated tariff, there is no separate recognition approach for NTLs being negotiated directly in the contracts, as occurs in Italy. In turn, in Spain, there is a specific incentive mechanism for the reduction in fraud. The distribution utility will receive, as additional remuneration in the year 'n', the amount of 20% of the frauds regularized with regard to the year 'n − 2', with a limit of 1.5% of the distribution company's remuneration [21].

According to the distribution utilities' performance, a differentiated regulatory treatment strategy has been employed in Colombia since 2018. The new proposal by the Colombian Energy and Gas Regulation Commission consisted of determining a target value at the national level and classifying companies into two categories, i.e., companies in the position above or below this new benchmark. The companies below the target have been classified as 'maintaining losses', and the companies above the fixed value have been called 'loss reduction'. For those companies classified in the 'loss reduction' category, the Regulation Commission began to recognize the percentage of actual losses in the energy tariff relative to that obtained in 2017, which is the starting point for a reduction trajectory for the following years. So, they must present loss reduction plans, which the regulatory body must approve for recognition in the tariff, not only of the level of losses but also of the operational costs of combating theft. If it is verified that the company is not succeeding in achieving its goals, the recognized remuneration of the plan is suspended and may even be cancelled [22].

The last mechanism that makes regulatory compliance more flexible concerns the recognition of the incidence of criminal violence in NTLs, as observed in countries such as Panama [23] and El Salvador. Consequently, this results in the allowance of losses to deal with crime-related factors that place NTLs outside the control of distribution companies [24].

Focusing on the Brazilian context, NTLs are responsible for a significant economic imbalance. The high levels of unbilled electricity impact the tariff and influence the concessionaires' investment capacity in the service and distribution network. In addition, the tariff pressure and the low quality of the distribution system reduce the competitiveness of the industrial sectors, generating an increase in default and electricity theft rates [25–27]. This vicious cycle creates a knock-on effect and weakens economic development, the reliable energy supply, and regulatory policies. In fact, NTLs in the country represented a cost of approximately BRL 8.6 billion in 2020 [28]; so, this phenomenon cannot be treated as a trivial regulatory problem. Understanding the different cultural, social, economic, and geographical realities in a country of continental size such as Brazil is the starting point for developing intelligent actions to mitigate NTLs in the medium and long term.

Based on the socioeconomic and market variables concerning all distribution utilities in the country, in 2015 the National Electric Energy Agency (ANEEL) formulated several specifications of econometric models for panel data regressions, all aimed at determining an index that reflects the difficulty of combating NTLs according to the intrinsic characteristics of each distribution area [29].

Nevertheless, more recently, the Agency understood that it would be necessary to review the currently adopted methodology to determine regulatory targets for NTLs due to a series of methodological problems pointed out by agents of the Brazilian electricity sector and academic experts [30,31]. Among the questions brought by those agents, we can highlight the following: (i) failure to capture the complexity of risk areas, places where there is a parallel power governed by drug trafficking gangs or ex-police militias and where the public power is not allowed access; (ii) the need for a control variable (operational efficiency); (iii) the need to define a more objective criterion for selecting variables (by the statistical method called Least Absolute Shrinkage and Selection Operator or LASSO [31–34]); and (iv) the need to not consider variables related to criminality, for example, by local distribution utilities, such as ENEL, Light, and EDP, which seek to portray risk areas [31].

The Agency used a database of more than 20 variables, already used and tested in previous discussions, and variables suggested by agents of the Brazilian electricity sector, mainly related to violence and criminality. As a result, a new regulation proposal was published in Technical Note No. 46/2020 [34]. Although the methodology used by the ANEEL represents a notable regulatory advance for defining NTL regulatory targets, the correlation problems between the omitted variables and the variables that make up the index have persisted. In this regard, Castro et al. [31] argued that the option for the random

effects model allows the isolation of effects related to the distribution area that do not vary over time. However, it is essential to verify the existence of a correlation between the individual effects and the other variables present in the model. Therefore, the random effects model should only be considered if this correlation does not exist.

In the case of the model currently used by the Agency, this correlation is evidenced by the Hausman test [35,36], and the fixed effects model should be chosen under the penalty of obtaining inconsistent estimators. Therefore, it is understood that the ANEEL should apply the Hausman test to decide between the fixed effects and the random effects models or to present additional arguments that justify such a choice from a statistical point of view [30].

In addition, as Leite et al. [37] argued, the model also fails because it does not clearly define the efficiency frontier, which is the main component of benchmarking frameworks. To overcome this deficiency, they proposed a stochastic frontier cost model for panel data whose equation could provide tolerable limits for the percentage of NTLs [37]. The model was applied to a panel of data containing annual observations of 41 distribution utilities in the Brazilian electrical system over ten years. The option for the stochastic frontier analysis (SFA) model [38–40] in this study maintained the econometric framework initially adopted by the ANEEL (i.e., the same dependent and explanatory variables in a panel data model with random effects). The SFA formulation allowed the authors to estimate the efficiency frontier and provided the NTL target values more transparently when compared to the current methodology adopted by the ANEEL. The results revealed that the proposed SFA cost model could provide feasible NTL regulatory targets, i.e., those that could be achieved by the distribution utilities and satisfy a range of economic, social, and political constraints while focusing on reducing non-technical losses.

Aiming to contribute to the continuous improvement of the methodology adopted by the ANEEL, Simões et al. [30] estimated, analyzed, and predicted the short-term NTLs of the electric power of Brazilian distribution utilities based on different assumptions for the covariance structure of the errors and controlling for the socioeconomic confounding variables. Although the correlation among the repeated responses is not usually of intrinsic interest, the authors consider it to be an essential aspect of the data that must adequately be accounted for to produce valid inferences in longitudinal or panel data analysis. In the extended linear mixed effects model, the response vector's covariance matrix comprised two subcomponents, i.e., a random effect component representing the between-group variation and an intraclass or within-group component. So, to adequately treat the longitudinal character of the NTL data, they used the data from 59 Brazilian distribution utilities from 2004 to 2012 that fit a conditionally independent errors model and three other models with autoregressive-moving average parameterization to the intraclass disturbances. Finally, they compared models using the mean absolute deviation (MAD) [41] and the mean absolute percentage error (MAPE) metrics [42] in the prediction of NTLs for the year 2013. The findings suggest that the approach can be satisfactorily implemented in future statistical analyses of NTLs.

As can be evidenced from previous works on this topic, considerable efforts have been devoted to using econometric models for panel data regressions to define NTL regulatory targets in Brazil. For example, the ANEEL conducted an extensive survey of official sources of secondary data to identify a large set of possible explanatory variables for NTLs and, more recently, evaluated the application of LASSO in selecting explanatory variables [31,34]. Despite that, a literature review covering the last two decades on this subject showed the nonexistence of previous studies exploring the application of automatic model-selection techniques for panel data regressions in the regulatory context of combating NTLs, as could be evidenced by the search histories presented in Appendix A.

However, the ANEEL made some mistakes in applying LASSO, particularly in not considering the panel data structure [31]. More specifically, the ANEEL applied LASSO to make a first selection of variables, reducing the set of explanatory variables to just seven (five selected by LASSO plus two entered manually) and then analyzing all 127 models

resulting from combinations of the seven variables selected [31,34]. The second phase of variable selection after the selection made by the LASSO approach needs to be clarified.

In addition, Castro et al. [31], David and Desboulets [43], Hastie et al. [44], and Bertsimas et al. [45] point out that LASSO-based inference methods tend to reduce the coefficients too much and can generate biases which put the reliability of the results in doubt. According to Hastie et al. [44], the problem can be mitigated using a variant called relaxed LASSO.

Considering that:

- NTLs are one of the main problems that electricity distribution utilities face in developing regions such as Latin America and the Caribbean, sub-Saharan Africa, and South Asia;
- The growing electricity demand, combined with complex socioeconomic conditions in countries of continental dimensions, such as Brazil and India, leads to economic and regulatory challenges directly related to the sustainability of distribution systems and the combating of NTLs;
- In Brazil, the ANEEL has formulated several specifications of econometric models for panel data with random effects, all aimed at determining an index that reflects the difficulty of distribution utilities combating NTLs according to the intrinsic characteristics of each distribution area;
- The exhaustive search for combinations of explanatory variables and the complexity inherent in defining NTL regulatory targets in Brazil still require evaluating a considerable number of models through hypothesis and goodness-of-fit tests;
- The literature review covering the last two decades on the use of econometric models using panel data to define regulatory targets for electricity NTLs revealed few studies concerning this issue and exploring automatic model-selection techniques for panel data regressions to define regulatory targets for electricity NTLs.

This paper addresses the research gaps by investigating the following research questions:

- RQ1: What are the main limitations of current econometric models for defining the NTL regulatory targets in Brazil?
- RQ2: How should the explanatory variables that have the most significant impact on the NTL phenomenon and that define panel data regressions that can better assist the ANEEL in establishing NTL regulatory targets for the distribution utilities in this country be selected?
- RQ3: To what extent can an automatic model-selection technique for panel data regressions support the definition of regulatory targets for electricity non-technical losses?
- RQ4: Is it possible to demonstrate the applicability of the proposed technique for defining regulatory targets for NTLs in the context of the electricity distribution sector in Brazil, highlighting its differentials compared with the current econometric models adopted by the ANEEL?

From a regulation-oriented perspective, this paper aims to propose an automatic model-selection technique for panel data regressions to better assist the ANEEL in establishing NTL regulatory targets for the electricity distribution segment in Brazil. Inspired by the glmulti package [46], the proposed technique generates all possible specifications for the panel data model from the list of explanatory variables and identifies the best model. In the context of the Brazilian electricity sector, the exhaustive search approach was used in selecting the explanatory variables of a logistic regression model for panel data to predict the insolvencies of the distribution utilities [47], and the NTLs appear among the selected variables.

Therefore, the proposed approach consists of an exhaustive search for models, i.e., a best subset selection algorithm. Although it demands a more significant computational effort, this search allows excellent flexibility in specifying the models to be searched for. For example, all the evaluated models consider the panel data structure with random or fixed

effects. Theoretically, the proposed technique avoids the biases that can be introduced by the LASSO approach. However, Hastie et al. [44] argue that neither the best subset selection nor the LASSO uniformly dominates the other.

The article is structured in five sections. Following the introduction, the second section briefly presents the research design and the adopted methodology. Section 3 introduces an automatic model-selection technique for panel data regressions to define NTL regulatory targets for the distribution utilities in Brazil. Section 4 presents and discusses the research results, highlighting the differentials of this methodological approach compared with the current econometric models for panel data adopted in Brazil for defining NTL regulatory targets. Lastly, Section 5 synthesizes the concluding remarks and future developments of this research for those interested in advancing the knowledge on the regulatory measures concerning reducing NTLs based on robust methodological approaches for defining NTL regulatory targets in developing countries.

2. Research Design and Methodology

Following a procedural model based on that of Martins et al. [48] to provide an underlying structure and an approved course of action for this research, its design encompasses three phases and six stages, as synthesized in Table 1. Accordingly, the research phases are: (i) motivation; (ii) conceptualization and development; and (iii) validation.

Table 1. Research design.

Phase	Stage	Research Question [Section]
Motivation (Why?)	1. Problem definition and the rationale for the research	Why should we propose an automatic model-selection technique for panel data regressions to better assist the ANEEL in establishing NTL regulatory targets for Brazilian distribution utilities? [Section 1].
Conceptualization and development (What and how?)	2. State of research on the central themes and identification of research gaps and unsolved problems	What is the state of research on econometric models for establishing NTL regulatory targets? And on the automatic model-selection techniques for panel data regressions? What are the main limitations of the current models adopted by the ANEEL for defining NTL regulatory targets? [Section 1].
	3. Definition of the research design and methodology	How could an automatic model-selection technique for panel data regressions aiming to establish regulatory NTL targets be developed and validated in the context of the electricity distribution sector in Brazil? [Section 2].
	4. Development of an automatic model-selection technique for panel data regressions to better assist the ANEEL in establishing regulatory NTL targets for distribution utilities	How should the explanatory variables that have the most significant impact on the NTL phenomenon and define panel data regressions that can better assist the ANEEL in defining NTL regulatory targets for Brazilian distribution utilities be selected? [Section 3].
Validation (How can the applicability of the proposed technique be demonstrated?)	5. Demonstration of the applicability of the proposed automatic model-selection technique for panel data regressions in the context of the electricity distribution sector in Brazil	Is it feasible to demonstrate the applicability of the proposed methodological approach to establish regulatory NTL targets for Brazilian distribution utilities? [Section 4]. Could the research results demonstrate the applicability of the proposed approach in the context of the electricity distribution sector in Brazil? [Section 4].
	6. Discussion of the research results and managerial implications	What are the differentials of the automatic model-selection techniques for the panel data regressions compared with the state of research on econometric models for establishing NTL regulatory targets and the current models adopted by the ANEEL for this purpose? [Section 5].

The six stages described in Table 1 refer to the problem definition and the rationale for the research (first stage); the identification of research gaps based on the literature review (second stage); the definition of the research design and methodology (third stage); the development of a new methodological approach to defining NTL regulatory targets by selecting automatic models for panel data regressions (fourth stage); the application of the methodological approach to proposing models of the electricity distribution in Brazil that can better assist the distribution utilities in establishing regulatory targets to reduce their NTLs (fifth stage); and the discussion of the research results, highlighting the differentials of the proposed approach compared with the current models adopted in Brazil for defining NTL regulatory targets (sixth stage).

In the first two stages, a literature review covering the period from 1992 to 2022 was conducted by systematic searches in the leading scientific production databases (e.g., Scopus and Web of Science), as detailed in Appendix A. This review focused on: (i) electricity non-technical losses and regulatory issues concerning reducing NTLs; (ii) the model-selection techniques for panel data regressions; and (iii) the interplays between these two themes.

Based on the state of research on the econometric models for establishing NTL regulatory targets, the research design was defined in the third stage, encompassing the development of an automatic model-selection technique for panel data regressions to better assist the ANEEL in establishing NTL regulatory targets for distribution utilities in Brazil and the demonstration of the applicability of the proposed methodological approach to panel data containing annual observations from 62 Brazilian electricity distribution utilities from 2007 to 2017.

The automatic model-selection technique for panel data regressions proposed in this work was inspired by the glmulti package [46] and aimed to select regression models automatically. From a set of explanatory variables, this package generates a list of all the possible combinations involving those variables and, optionally, their paired interactions [46]. However, the use of the package does not consider panel data regressions. Nowadays, the modelling for panel data regressions is performed with the support of the plm package [36]. With the plm package, it is possible to consider regression models with least squares, random effects, or fixed effects. In addition, the package includes hypothesis tests to choose the most parsimonious model. In this way, all the possible variables tested were listed, and the possible combinations that could be tested were calculated. Then, the plm package was applied to the panel data regressions in the three types of panel data—least squares, random effect, and fixed effect. In addition, the Hausman and the Breusch–Pagan (BP) tests [35,49] were used to verify the most parsimonious model.

Finally, the proposed automatic model-selection approach could be demonstrated from the panel data containing annual observations from 62 Brazilian electricity distribution utilities from 2007 to 2017, generating 1,097,789 models (combinations of explanatory variables). The analyzed combinations include models with up to nine explanatory variables, given that models with ten or more explanatory variables do not show statistical significance in all regression coefficients. After applying the selection criteria, 18 panel data models with random effect and 42 panel data models with fixed effect were obtained, and among them, only three models showed greater adherence to the data.

3. Proposition of an Automatic Model-Selection Technique for Panel Data Regressions

This section proposes an automatic model-selection technique for panel data regressions to help the ANEEL establish NTL regulatory targets. Figure 1 shows a general view of the proposed technique.

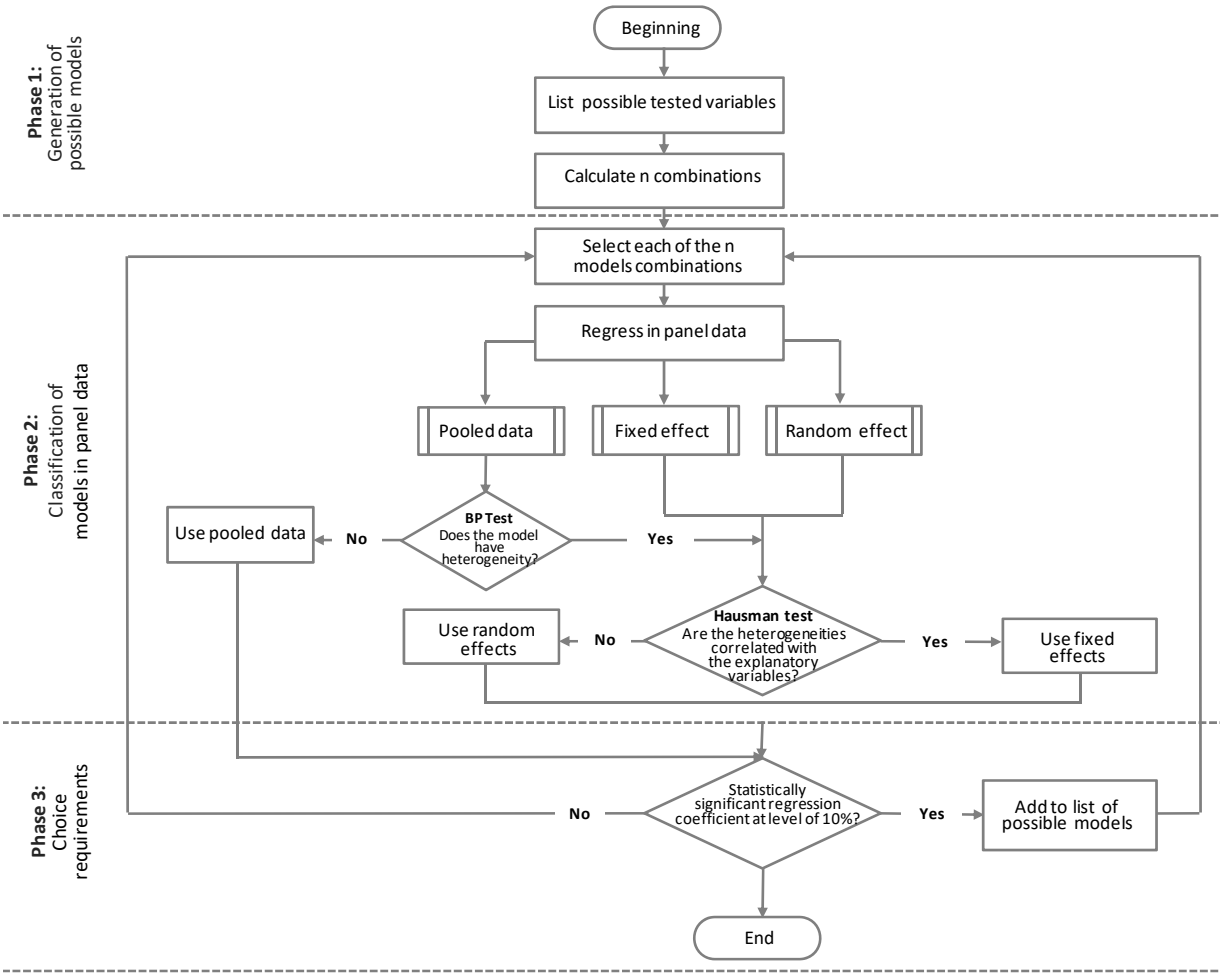

Figure 1. General view of the proposed automatic model-selection technique for panel data regressions.

3.1. Phase 1: Generating Possible Models

All possible combinations of variables are listed from the database with k explanatory variables, totaling $2^k - 1$ combinations, each corresponding to a regression model.

3.2. Phase 2: Classification of Models in Panel Data

In the panel data field, the econometric approach emphasizes model specification and testing, tackling several issues arising from the particular statistical problems associated with the economic data [50,51].

There are many ways to estimate the response of a dependent variable concerning an explanatory variable, each depending on the availability and structure of the data. When information is available by cross-section and by the evolution of these units over time (year, month, day), it is said that this information is organized in the panel data. The advantages of using panel data are the larger sample size and the capacity to better estimate the relationships among the variables, even in the face of heterogeneity among the distribution utilities. This heterogeneity is due to a set of unique characteristics of the concession area, such as the relief, the vegetation, the climatic factors, and the population culture, among others, which are considered to be invariant over time.

According to Hsiao [51], panel data allow more complex behavioral aspects to be modelled. For example, by investigating the effects of public policies before and after, controlling the effect of omitted variables, and unveiling dynamic relationships among variables. Thus, it is possible to analyze how past causes can influence the present through lags.

A critical point in estimating panel data is that the observations cannot be assumed to be independently distributed over time. Omitted variables or "unobservable" factors that

affect, for example, the global losses of a given distribution utility in 2012 possibly affect them in 2013. In Equation (1), we present the generic equation of the panel data with N individual units (in this case, the electricity distribution utilities) over T years.

The heterogeneity among the electricity distribution utilities is due to a set of unique characteristics of the concession area, such as the vegetation, the climatic factors, and the population culture, among others, which are considered to be invariant over time.

$$y_{it} = \mu + \beta x_{it} + \alpha_i + \varepsilon_{it},\ i = 1, ..., \text{N and } t = 1, ..., \text{T} \quad (1)$$

where

y_{it} = dependent variable from unit i at year t;
α_i = effect of the *i-th* unit on the overall parameter;
β = vector with regression coefficients;
x_{it} = explanatory variables in unit i at year t;
μ = general intercept, time-invariant;
$\varepsilon_{it} \sim$ NIID (0, $\sigma\varepsilon$) = idiosyncratic deviations.

Using variables in a panel data format, three main types of models can be specified: (i) pooled data; (ii) fixed effects; and (iii) random effects. The last type was chosen by the ANEEL.

The model for pooled—or stacked—data assumes that the heterogeneity (α_i) of each unit i does not exist. In contrast, the fixed and random effects models assume the existence of heterogeneity (α_i). Additionally, the difference between the two models is that while the random effects model assumes that there is no correlation of heterogeneity with any explanatory variable ($E(\alpha_i x) = 0$), the fixed effects model assumes such a correlation exists ($E(\alpha_i x) \neq 0$) [52,53].

The characteristics of the three mentioned models are presented below:

- The use of the pooled model assumes that the intercept and response parameters do not differ among individuals and are constant over time. In Equation (2), the resulting model is presented.

$$y_{it} = \mu + \beta x_{it} + \varepsilon_{it} \quad (2)$$

- The fixed effects model intends to control the effects of the omitted variables, which vary among individuals and remain constant over time. Thus, it assumes that the intercept varies from one individual to another but is constant over time. The fixed effect model is presented in Equation (3), where N-1 dummy variables represent the fixed effects (α). So, the fixed effect model can be fitted by an ordinary least square, and the estimator of β is referred to as the Least Square Dummy Variable estimator (LSDV) [52,53].

$$y_{it} = \mu + \alpha_i + \beta x_{it} + \varepsilon_{it} \quad (3)$$

- The random effects models consider that the individuals on whom the data are available are random samples from a larger population of individuals. In this case, $\alpha_i \sim$ NIID (0, $\sigma\alpha$) and cov(α_i, ε_{it}) = 0. The sum $\alpha_i + \varepsilon_{it}$ is a composite error with two components: the white noise ε_{it} and the individual specific component α_i that does not vary over time. Then, the composite error presents autocorrelation, and in this case, the model fitting should be carried out by the Generalized Least Square (GLS) [50,51].

$$y_{it} = \mu + \alpha_i + \beta x_{it} + \varepsilon_{it} \quad (4)$$

The choice of the most appropriate specification of the model is carried out using the Breusch–Pagan (BP) and Hausman tests. Initially, the Breusch–Pagan (BP) test should be applied, whose hypotheses are presented below:

$$\begin{cases} \text{H}_0 :\ \alpha_i = 0,\ \text{use pooled} \\ \text{H}_\text{a} :\ \alpha_i \neq 0,\ \text{use fixed or random effects} \end{cases} \quad (5)$$

The null hypothesis in (5) corresponds to the model for pooled data. Based on a Lagrange multiplier test, the statistics test has a chi-square distribution, which can be accessed in [48]. Having verified the presence of heterogeneity (by the rejection of the null hypothesis of the Breush–Pagan test), the Hausman test [35] is applied, whose hypotheses are described below:

$$\begin{cases} H_0: \ E(\alpha_i, x) = 0; \text{use random effect models} \\ H_a: \ E(\alpha_i, x) \neq 0; \text{use fixed effects models} \end{cases} \quad (6)$$

Additionally, possible violations of the basic assumptions assumed in the adjustment of the model should be checked. More specifically, heteroscedasticity, contemporary correlation, and serial correlation of errors should be assessed. Although such violations do not result in bias in estimating the regression coefficients, their presence implies the biased estimates of standard errors. Thus, they compromise the statistical inference of the regression coefficients.

In a regression, there are basic assumptions for the ordinary least squares estimator to be BLUE (Best Linear Unbiased Estimator), i.e., the best non-biased linear estimator. Among the assumptions that stand out are the assumptions that the variance of errors is constant (homoscedasticity) and that there is no serial correlation (correlation of values over time) or contemporaneity among the errors. Thus, verifying whether these assumptions comply is essential for a BLUE estimator. Such assumptions are essential for the estimator's efficiency and for not affecting the consistency. Inefficient estimators have the characteristic of changing the confidence interval of the parameters, making the t-tests misleading, thus making the decision to "accept" the significance of the estimated coefficients imprecise.

The evaluation of the possible violations is carried out through the following hypothesis tests: the Breusch–Pagan LM (BPLM), Breusch–Godfrey/Wooldridge (BGW), and Breusch–Pagan tests for contemporary autocorrelation, serial autocorrelation, and heteroscedasticity, respectively. Briefly, the three tests, in sequence, have the following hypotheses presented in Equations (7)–(9) (for more details, see references [52–55]).

$$\text{BPLM} = \begin{cases} H_0: \text{A Absence of contemporary autocorrelation} \\ H_a: \ \text{P Presence of contemporary autocorrelation} \end{cases} \quad (7)$$

$$\text{BGW} = \begin{cases} H_0: \text{Absence of serial autocorrelation} \\ H_a: \ \text{Presence of serial autocorrelation} \end{cases} \quad (8)$$

$$\text{BP} = \begin{cases} H_0: \ \text{Homoscedasticity} \\ H_a: \ \text{Heteroscedasticity} \end{cases} \quad (9)$$

3.3. Phase 3: Requirements for Choosing Models

In this last phase, we seek to select the best models. In each model, the *p*-values of the regression coefficients are evaluated (an alpha significance level of 10% is considered). Thus, only the models with all their significant variables are selected at the end of the exhaustive search. The same control strategy was used by the ANEEL [34]. Finally, after the model has gone through all the steps, a list with the estimated coefficients, *p*-values, AIC, and adjusted R^2 of all the selected models is set [56,57].

More than just selecting the models according to the hypothesis testing criteria and evaluating the significance of the variables is required to find the best specification. In addition, it is necessary to assess the signs resulting from the regression coefficients in each model. In addition, when analyzing the models at the end of the process described in the previous items of this section, some variables will generally be more recurrent than others, indicating greater relevance in the NTL modelling.

Thus, the Akaike Information Criterion (AIC) was used to avoid overfitting in the model selection. The AIC is based on information theory. When a statistical model is used to represent a particular process, the representation will never be accurate; that is,

the model will never be perfect, and indeed, some information will be lost. This criterion estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model and the lower the AIC score [58–60].

A direct solution to the problem would be to select the model with the most significant number of variables; however, more complex models are prone to overfitting the training data. The best model should provide an adequate data report using a minimum number of parameters [59], according to the principle of Ockham's Razor.

It is worth noting that the ANEEL uses the R^2 statistic to rank the models and selects the one with the highest R^2 [34]. This control strategy results in less parsimonious models with a high probability of overfitting. So, we considered the Akaike weights according to [58–60], aiming to analyze the models in this study. Wagenmakers and Farrell [60] demonstrated that AIC values could be easily transformed into so-called Akaike weights, which can be indirectly interpreted as conditional probabilities for each model, which can significantly facilitate the interpretation of the results of the AIC model comparison procedures.

In Equation (10), for each model, the differences in the AIC concerning the AIC of the best candidate model, the one with the lowest AIC, are calculated.

$$\Delta_i(\mathrm{AIC}) = \mathrm{AIC}_i - \mathrm{minAIC} \tag{10}$$

By the differences in the AIC, we can obtain an estimate of the relative likelihood L of the model—by the simple transformation presented in Equation (11), in which $\propto$ means "is proportional to".

$$L(M_i \mid data) \propto \exp\left\{-\frac{1}{2}\Delta_i(\mathrm{AIC})\right\} \tag{11}$$

In the last step, the relative likelihoods of the model are normalized (that is, divided by the sum of the likelihoods of all models) to obtain the Akaike weight, according to Equation (12).

$$w_i(\mathrm{AIC}) = \frac{\exp\left\{-\frac{1}{2}\Delta_i(\mathrm{AIC})\right\}}{\sum_{k=1}^{k}\exp\left\{-\frac{1}{2}\Delta_k(\mathrm{AIC})\right\}} \tag{12}$$

where $\sum w_i(\mathrm{AIC}) = 1$.

The weight w (AIC) can be interpreted as the probability of model i being the best model (in the sense of AIC, which minimizes the Kullback–Leibler Discrepancy), as mentioned in [60]. Note that the Akaike weights are subject to sampling variability and that a different sample will likely generate a different set of weights for the models in the joint candidate.

4. Results and Discussion

In this section, the results of the first stage of the modelling are presented and further discussed. In addition, the main result of the second stage is exposed, that is, the predictive power of the model. Some results of a descriptive nature are also presented to help visualize and compare the evolution of some of the selected indicators.

4.1. Results of Applying the Automatic Model-Selection Technique for Panel Data Regressions

Figure 2 shows the results of applying the automatic model-selection technique in the electricity distribution sector in Brazil.

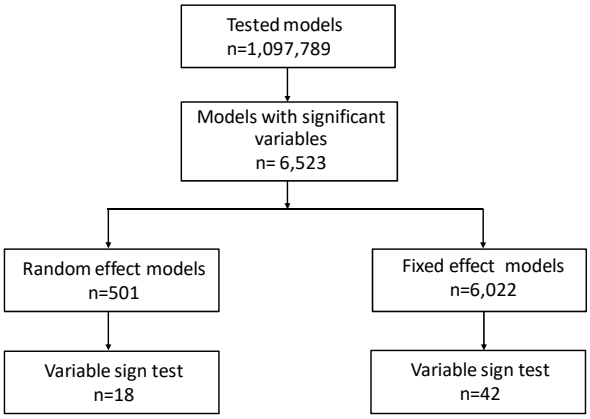

Figure 2. Results of applying the automatic model-selection technique for panel data regressions.

It is observed that 1,097,789 models were tested with different combinations of variables. As a result of the significant variables, 6523 remained, which can be divided into 501 by random effects and 6022 by fixed effects. Finally, the coefficients' signs were evaluated to be consistent with our variable of interest. Thus, as a result, 18 panel data models with random effects and 42 panel data models with fixed effects were obtained.

Table 2 shows the expected signs of the coefficients of the candidate explanatory variables on the screen. In addition, this table shows the correlations between the candidate variables and the dependent variable, i.e., the NTL index [37], defined by the ratio below between the non-technical losses and the low-voltage market, both given in MWh:

$$NTL_{i,t} = \frac{\text{Non} - \text{technical losses of the utility } i \text{ in year } t(\text{MWh})}{\text{Low voltage market the utility } i \text{ in year } t \text{ (MWh)}} \quad (13)$$

Table 2. Summary of variables and expected signs of the coefficients.

Variable	Code	Expected Sign	Correlation
Density of residents per dormitory	Admd	+	70%
High school dropout rate	Eem	+	−18%
Percentage of people below the poverty line	Pob3	+	48%
Percentage of households with general piped water network	Rga	−	−4%
Percentage of urban households with garbage collection	Lixo.u	−	−13%
Social inequality index	Gini2	+	13%
Percentage of people living in subnormal households	Sub2	+	67%
General default in the SFN credit sector	Inad	+	40%
Default by private individuals in the SFN credit sector	Inad.pf	+	42%
Intentional homicide	Homi_dolo	+	26%
Police interventions and war operations	Int_pol	+	20%
Vehicle theft	Furto_v	+	−18%
Vehicle robbery	Roubo_v	+	16%
Robbery	Latro	+	13%
Assault by firearm	Agraf	+	32%
Homicide (deaths due to aggression)	Vio	+	40%
Homicides (violent deaths)	H.mvci	+	37%
Low-income consumer units/BT consumer units	Ucbr.Mb1UCbr	−	22%
Low-income market/B1 total market	Mbr.Mb1Mbr	−	11%
Low-income market/Total BT market	Mbr.Mbt	−	17%
Concession area GDP per capita	PIB.PC	−	−19%
Time in years	Time	Not applicable	−6%

It is noteworthy that some regression coefficients show opposite signs to the expected, as shown in Table 3. In such cases, one hypothesis is that the variables are poorly correlated with the phenomenon of non-technical losses.

Table 3. Correlation with the inconsistent expected sign.

Negative Correlation	Positive Correlation
Furto_v	Eem, Ucbr.Mb1UCbr, Mbr.Mb1Mbr and Mbr.Mbt

Then, of the 22 used variables, only 11 were selected in some of the panel data models, as shown in Table 4. In addition, verifying the recurrence in which this same variable was used is possible. Thus, it is possible to verify that some attributes are better predictors of NTLs than others, such as default (Inad and Inad.pf) and sanitation (Lixo.u), which are recurrent in the significant models.

Table 4. Results of the set of explanatory variables.

#	Code	Variable Title	Model Amounts
1	Inad	General default in the SFN credit sector	28
2	Inad.pf	Private individual defaults in the SFN credit sector	24
3	Lixo.u	Percentage of urban households with garbage collection	19
4	Pob3	Percentage of people below the poverty line	17
5	Gini	Social inequality index	15
6	Time	Time in years	14
7	Admd	Density of residents per dormitory	11
8	Furto_V	Vehicle theft	7
9	PIB.PC	Concession area GDP per capita	4
10	Vio	Homicide (deaths due to aggression)	2
11	Mbr.Mb1Mbr	Low-income market/B1 total market	1

A priori, the percentage of people living in subnormal households (precarious housing situations) or the sub2 variable is a good proxy for complexity. However, it did not present a statistically significant coefficient in any of the 60 selected models.

The importance of an explanatory variable can be assessed by adding the Akaike weights of all the models that include the variable [58–60]. The results associated with the importance of the explanatory variables are presented in Figure 3, applying this concept to the 60 selected models. It is observed that the defaults (inad and inad.pf), the Gini index (Gini2), the poverty index (Pob3), garbage collection (Lixo.u), and the density of residents (Admd) are the variables that concentrate the most importance.

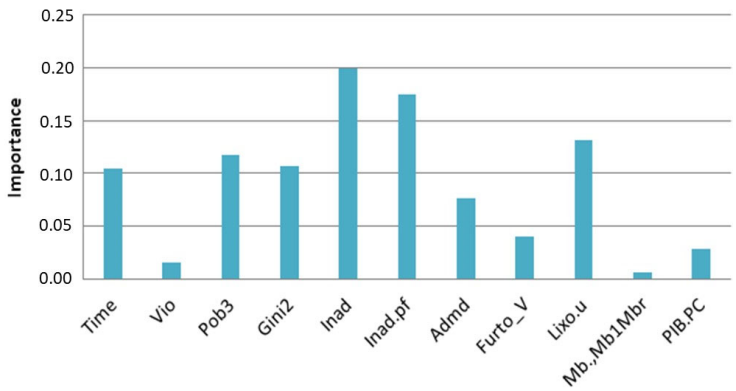

Figure 3. Importance of explanatory variables.

4.2. Use of Akaike Weights

For the application of the Akaike weights, we decided to do it separately, respecting the group of random effects and fixed effects. Table 5 shows the results of the random effect models that had the most significant impact. One can observe that the two models concentrate 100% of the information.

Table 5. Results of panel data models with random effects.

#	(Intercept)	Pob3	Inad	Inad.pf	AIC	R^2-Adjusted	DELTA AIC	AKAIKE Weight	*p*-Value BP Test
1	0.01	0.38	1.33		−2749.87	0.17	0.57	0.43	0
2	0.01	0.33	1	0.53	−2750.44	0.17	-	0.57	0

Table 6 shows the results of the fixed effects model. By Akaike weight, only one model was enough to contain the information from all the others.

Table 6. Results of panel data models with fixed effects.

Time	Pob3	Inad.pf	AIC	R^2-Adjusted	DELTA AIC	AKAIKE Weight	*p*-Value BP Test
0.0	0.34	1.21	−2817.83	0.1	-	1	0

According to Tables 5 and 6, the only variable which appears in all the models is Pob3. In our approach, the model with the highest R2 (0.22) includes five explanatory variables (Pob3, Gini2, Inad, Inad.pf, and Furto_V); it has fixed effects and does not belong to the set of 60 models selected; the corresponding AIC statistic is equal to −1927.79 (the smallest AIC is equal to −2817.83).

This result shows that the control strategy based on the R^2 statistic produces less parsimonious models. Note that this model includes the three explanatory variables present in the selected models, as illustrated in Tables 5 and 6.

4.3. Models Predictions

Table 7 presents the estimates obtained by the regression models for the electricity distribution utility loss index for the year 2017, accompanied by the respective verified NTLs. Given that the explanatory variables of the three selected models are basically the same, the expected values estimated by the three selected models (targets) are close. Note that, in general, the verified NTLs came close to the target defined by the models. In addition, this table shows the size and region of Brazil served by each utility.

Table 7. Predictive result of electricity distribution utilities.

Distribution Utility	Region *	Random Effect Model (#1)	Random Effect Model (#2)	Fixed Effect Model	NTLs	Size **
AES-SUL	S	8%	8%	8%	11%	L
AMAZONAS	N	125%	124%	124%	124%	L
AMPLA	SE	32%	31%	31%	30%	L
BANDEIRANTE	SE	19%	19%	19%	13%	L
CAIUA	SE	2%	2%	2%	2%	S
CEAL	NE	55%	54%	53%	36%	L
CEB	MD	8%	8%	8%	9%	L
CEEE	S	28%	28%	28%	26%	L
CELESC	S	4%	4%	4%	9%	L
CELG	MD	8%	8%	8%	10%	L
CELPA	N	46%	46%	45%	39%	L
CELPE	NE	22%	21%	21%	18%	L

Table 7. *Cont.*

Distribution Utility	Region *	Random Effect Model (#1)	Random Effect Model (#2)	Fixed Effect Model	NTLs	Size **
CELTINS	N	7%	6%	7%	4%	L
CEMAR	NE	22%	21%	22%	11%	L
CEMAT	MD	12%	12%	12%	10%	L
CEMIG	SE	12%	12%	11%	14%	L
CEPISA	NE	43%	42%	43%	30%	L
CERON	N	43%	42%	43%	50%	L
CFLO	S	1%	0%	0%	0%	S
CHESP	MD	3%	3%	3%	7%	S
CJE	SE	3%	2%	2%	0%	S
MOCOCA	SE	3%	2%	2%	7%	S
SANTA CRUZ	SE	3%	2%	2%	5%	S
NACIONAL	SE	1%	1%	1%	1%	S
COCEL	S	4%	3%	3%	5%	S
COELBA	N	12%	11%	11%	12%	L
COELCE	NE	8%	8%	7%	14%	L
COOPERALIA	S	6%	6%	6%	6%	S
COPEL	S	4%	4%	4%	4%	L
COSERN	NE	7%	6%	7%	3%	L
CPEE	SE	5%	5%	5%	6%	S
PIRATININGA	SE	6%	6%	6%	9%	L
CPFL PAULISTA	SE	7%	6%	6%	10%	L
CSPE	SE	5%	4%	4%	11%	S
DEMEI	S	7%	6%	6%	3%	S
DMED	SE	5%	4%	4%	3%	S
EBO	NE	8%	8%	9%	3%	S
EVP	SE	1%	0%	0%	1%	S
BRAGANTINA	SE	2%	2%	2%	2%	S
JOAO CESA	SE	2%	1%	2%	2%	S
EFLUL	S	3%	3%	3%	3%	S
ELEKTRO	SE	5%	4%	4%	8%	L
ELETROACRE	N	26%	25%	26%	27%	L
ELETROCAR	S	5%	5%	5%	6%	S
ELETROPAULO	SE	13%	12%	13%	10%	L
SANTA MARIA	SE	10%	9%	9%	3%	S
EMG	SE	4%	4%	3%	4%	L
ENERSUL	MD	15%	15%	16%	7%	L
ENF	SE	4%	4%	3%	0%	S
EPB	NE	16%	16%	17%	8%	L
ESCELSA	SE	23%	23%	22%	17%	L
ESE	NE	14%	14%	14%	8%	L
FORCEL	S	1%	1%	1%	5%	S
HIDROPAN	S	1%	0%	0%	4%	S
IENERGIA	S	8%	8%	8%	9%	S
LIGHT	SE	49%	48%	48%	51%	L
MUXFELDT	S	2%	2%	2%	0%	S
RGE	S	6%	6%	6%	7%	L
SULGIPE	NE	12%	11%	11%	9%	S
UHENPAL	S	4%	4%	3%	0%	S

Notes: * Region: MD—Midwest; N—North; NE—Northeast; S—South; SE—Southeast. ** Size: L—Large; S—Small.

For that, the concept of size is defined in [29] as follows:

"Being considered larger (Group 1) those that have a market greater than 1000 GWh/year and serve more than 500,000 consumer units or that have more than 15,000 km of electricity (. . .). The other distribution utilities are considered to be in Group 2".

As illustrated by the boxplots in Figure 4, the smaller distribution utilities tend to have a lower level of NTLs. The reason for this is that the complexity of their concession areas is also lower.

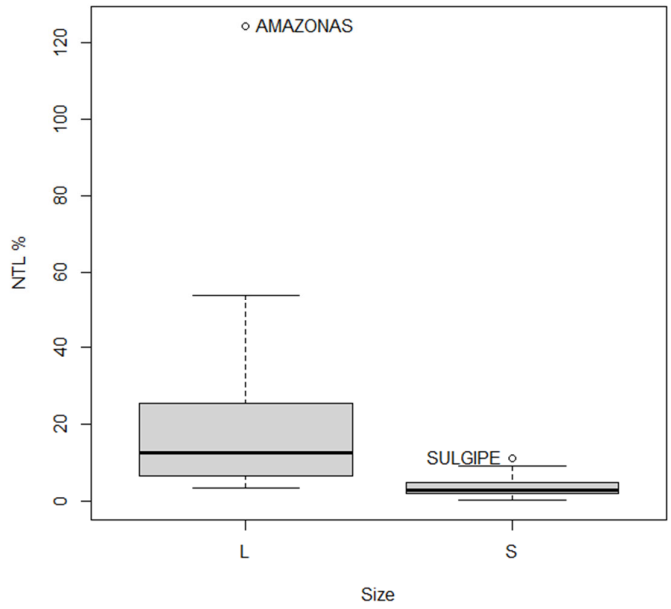

Figure 4. NTLs by size of concession area.

In addition, Figure 5 shows that there are higher levels of NTLs in the north and northeast regions, as they are poorer states and of greater socioeconomic complexity. Despite this, the main utilities (AMPLA and LIGHT) in Rio de Janeiro state, in the southeast region, draw attention due to its high NTL rates. Rio de Janeiro has great social and cultural complexity, boosted by the large population living in shanty towns (subnormal households) and high default rates.

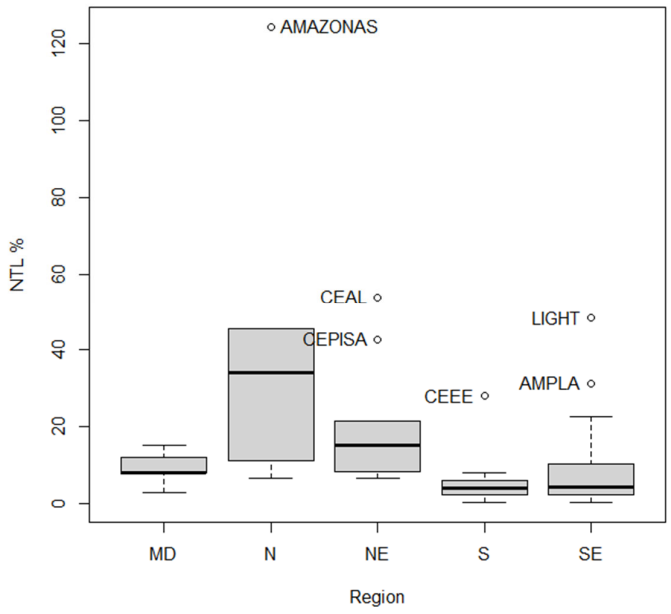

Figure 5. NTLs by region.

The scatterplot in Figure 6 shows the averages of the expected values estimated by the three models (NTL regulatory targets) and the respective verified NTL values shown in Table 7. Initially, we can observe that the estimates (targets) are proportional to the verified values of the NTL index, attesting to the goodness of fit.

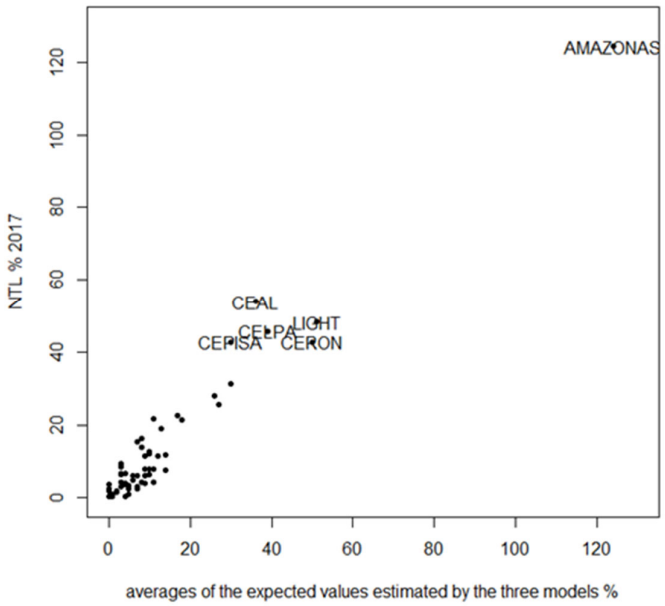

Figure 6. Scatterplot of the averages of the expected values estimated by the three models and the respective verified NTLs.

In addition, when comparing the real NTLs of each company with the targets, it is possible to infer which utilities are inefficient in combating their losses. However, some companies have targets as high as the NTL levels. For example, attention is drawn to the losses of the larger distribution utilities, mainly AMAZONAS, LIGHT, and AMPLA.

Thus, taking the NTL average as a reference, the scatterplot in Figure 6 was divided into four quadrants, as illustrated in Figure 7.

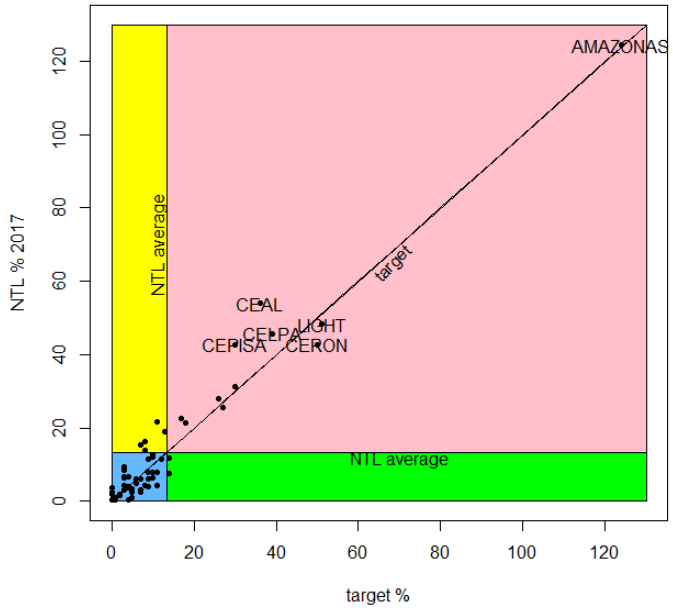

Figure 7. Scatterplot of the averages of the expected values estimated divided into four quadrants.

In Figure 7, the lower left quadrant (blue) contains the concessionaires with the losses under control (losses less than the average NTL) and with feasible targets. In contrast, the companies with high losses (losses more significant than the average NTL) are in the upper right quadrant (pink). Note that LIGHT and CERON have NTL levels below the target, but the respective targets are benevolent and do not encourage reducing NTL levels. Therefore, the regulator must set aggressive targets for the companies in the pink quadrant, such as

by taking half of the target proposed by the regression models or the lower bound of the expected value confidence interval.

The companies in the lower right quadrant (green) have NTLs below average and the regression model set benevolent targets. In this case, the regulator can also adopt aggressive targets. On the other hand, in the upper left quadrant (yellow), the regulator must be cautious, given that the NTL levels exceed the average, and the targets defined by the regression models may not be feasible.

4.4. Discussion

As described in the previous section, exhaustive simulations were made, combining all the data provided by the ANEEL. As a filter, we considered only the models in which the regression coefficients were statistically significant at the level of significance of 10% and with signs consistent with the expected. The variables related to the default had a more significant correlation with the phenomenon of NTLs (Figure 3).

The resulting models were classified into two groups: one with a random effect and another with a fixed effect. Furthermore, Akaike weights were used to select the best models. Two models were selected for random effects, while only one was shown to have all the information for fixed effects.

It is important to mention that the ANEEL has considered ad hoc criteria for choosing models to define regulatory NTL regulatory targets for electricity distribution utilities in Brazil [34]. In contrast, the proposed model allows the automatic selection of explanatory variables, using more objective and easy-to-understand criteria and avoiding the problem of overfitting.

Despite this methodological contribution, there is a risk in the automatic model selection for the panel data regressions that significant variables representing particular phenomena can be "left out". Thus, it is up to the analyst to look for other variables to explain these phenomena. Therefore, it is essential that the analyst carefully evaluates the results after running the model, in addition to evaluating whether the weighting of the variables is adequate and giving the appropriate signal.

The automatic selection of variables also depends on the choice of input variables. From this perspective, the analyst must carefully choose these variables, representing a particular phenomenon that impacts the increase or decrease in NTLs. Therefore, although automation can contribute significantly to the definition of regulatory targets for NTLs, a critical analysis of the models selected at the end of the process needs to be considered. It is up to the analysts involved to decide whether or not to consider certain variables or selected models.

In comparison with the previous studies reviewed in the introductory section, the main feature of the proposed methodological approach lies in the possibility of using the plm package [36]. Unfortunately, the main R package with routines for fitting panel data regressions with random or fixed effects does not have those for the automatic selection of explanatory variables. Thus, the proposed approach could be readily employed by Simões et al. [30] and Leite et al. [37] to improve their panel data regressions for defining the regulatory targets for electricity NTLs. It is worth noting that Leite et al. [37] used the frontier R package [38] to fit a stochastic frontier model with the panel data, and even in that case, the approach proposed here would be readily applicable. It could also be applied in the automatic selection of explanatory variables of generalized linear models, as shown in the work of Silva and Pessanha [47].

The proposed methodology is a best subset selection algorithm, but the biggest problem of this approach resides in the computational effort to evaluate a large number of models. However, this problem can be mitigated by parallelizing the computational implementation. Furthermore, the methodological approach allows the screening of models based on the signs of the regression coefficients and the statistical significance levels, among other constraints [45].

As mentioned before, it can also avoid the biases that can be introduced by the LASSO approach [31,34,44,45]. Additionally, it can provide an ensemble of alternative models, which may contribute to reducing the generalization error.

Although the proposed model focused on a regional problem, some countries may benefit from using it to minimize the adverse effects of NTLs (e.g., Colombia), considering the methodological advantages and limitations discussed above. In addition, panel data regressions can be employed in other areas of knowledge, and the automatic model-selection technique can help analysts choose variables and models in different application contexts.

5. Conclusions

This paper attempted to present an automatic model-selection technique for panel data regressions to effectively support the definition of the NTL regulatory targets for electricity distribution utilities. The proposed technique was applied to the panel data containing annual observations from 62 Brazilian electricity distribution utilities from 2007 to 2017, thus generating 1,097,789 models associated with the types of panel data regressions. The main results are three models that showed more adherence to the actual capacity of Brazilian distribution companies to reduce their NTLs.

The main conclusions associated with the research questions defined in the introductory section are presented here.

Concerning the main limitations of current the econometric models for defining NTL regulatory targets in Brazil, from the review of recent works on this issue, it was possible to identify the following shortcomings: (i) the failure to capture the complexity of the risk areas; (ii) the need to define a more objective criterion for selecting variables; and (iii) the need to not consider variables related to criminality.

With respect to the second and third research questions, we conclude that the proposed automatic model-selection technique for panel data regressions allowed the selection of the explanatory variables that have the most significant impact on the NTL phenomenon, aiming to define panel data regressions that can better assist the ANEEL in establishing NTL regulatory targets for the distribution utilities in Brazil. An exhaustive search with combinations of econometric and market variables was conducted, expanding the number of possible models with greater statistical robustness and considering panel data with the random effects and the fixed or least squares.

Finally, regarding the fourth research question, we can affirm that it was possible to demonstrate the applicability of the proposed automatic model-selection technique for defining the regulatory targets for NTLs in the context of the electricity distribution sector in Brazil, highlighting its differentials compared with the current econometric models adopted by the ANEEL. As a result, 18 panel data models with random effects and 42 panel data models with fixed effects were obtained. Additionally, 11 variables were selected in some of the panel data models, making it possible to verify that some attributes are better predictors of NTLs than others.

Despite the promising results presented in this paper, further studies should be carried out to apply other panel data models using more infrastructure variables that have shown synergy with the NTL phenomenon.

Author Contributions: E.C., R.C. and J.F.P. conceived and designed the research; E.C. performed the literature review and descriptive analysis and wrote Section 3 and part of Section 4; M.F.A. wrote Sections 1 and 2; E.C., R.C. and J.F.P. wrote Section 4; E.C., R.C., J.F.P. and M.F.A. jointly wrote the conclusions (Section 5). All authors commented on all the sections and reviewed the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by three Brazilian Funding Agencies: Coordination for the Improvement of Higher Education Personnel (acronym in Portuguese, Capes), National Council for Scientific and Technological Development (acronym in Portuguese, CNPq), and Foundation for Supporting Research in the State of Rio de Janeiro (acronym in Portuguese, Faperj).

Data Availability Statement: Not applicable.

Acknowledgments: The authors are thankful for the financial support provided by three Brazilian Funding Agencies (CNPq, Capes, and Faperj). The authors also thank the R&D program of the Brazilian Electricity Regulatory Agency (ANEEL) for the financial support (R&D project PD 00383-0062/2017). Finally, special thanks go to the anonymous reviewers for carefully reading the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Search Histories in the Web of Science and Scopus Databases

Table A1. Search strategy in the Web of Science database.

Number	Keyword Search	Retrieved Documents (n)
#1	TS= (“non-technical loss*” OR “non-technical electricity loss*” OR NTL)	1364
#2	TS= (electricity OR energy OR “power system*”)	4,086,556
#3	TS= (“model selection”)	25,522
#4	TS= (“panel data”)	49,202
#5	#1 AND #2	403
#6	#3 AND #4	135
#7	#5 AND #3	0
#8	#5 AND #4	9
#9	#5 AND #6	0

Note: Search strategy and assessment on 17 December 2022.

Table A2. Search strategy in the Scopus database.

Number	Keyword Search	Retrieved Documents (n)
#1	TITLE-ABS-KEY (“non-technical loss*” OR “non-technical electricity loss*” OR NTL)	2283
#2	TITLE-ABS-KEY (electricity OR energy OR “power system*”)	5,819,182
#3	TITLE-ABS-KEY (“model selection”)	25,160
#4	TITLE-ABS-KEY (“panel data”)	51,155
#5	#1 AND #2	587
#6	#3 AND #4	130
#7	#5 AND #3	1
#8	#5 AND #4	7
#9	#5 AND #6	0

Note: Search strategy and assessment on 17 December 2022.

References

1. Luqman, M.; Haq, M.; Ahmad, I. Power outages and technical efficiency of manufacturing firms: Evidence from selected South Asian countries. *Int. J. Energy Econ. Policy* **2021**, *11*, 133–140. [CrossRef]
2. Tehero, R.; Aka, E.B.; Çokgezen, M. Drivers of the quality of electricity supply. *Int. J. Energy Econ. Policy* **2020**, *10*, 183–195. [CrossRef]
3. Smith, T.B. Electricity theft: A comparative analysis. *Energy Policy* **2004**, *32*, 2067–2076. [CrossRef]
4. Savian, F.S.; Siluk, J.C.M.; Garlet, T.B.; Nascimento, F.M.; Pinheiro, J.R.; Vale, Z. Non-technical losses: A systematic contemporary article review. *Renew. Sustain. Energy Rev.* **2021**, *147*, 111205. [CrossRef]
5. Savian, F.S.; Siluk, J.C.M.; Garlet, T.B.; Nascimento, F.M.; Pinheiro, J.R. Non-technical losses in electricity distribution: A bibliometric analysis. *IEEE Lat. Am. Trans.* **2021**, *19*, 359–368. [CrossRef]
6. Carr, D.; Thomson, M. Non-Technical Electricity Losses. *Energies* **2022**, *15*, 2218. [CrossRef]
7. Price Water House Coopers (PWC). *Best Practices and Strategies for Distribution Loss Reduction-Final Report Forum of Regulators*; Price Water House Coopers (PWC): London, UK, 2016.
8. Lewis, F.B. Costly “throw-ups”: Electricity theft and power disruptions. *Electr. J.* **2015**, *28*, 118–135. [CrossRef]
9. Zanetti, M.; Jamhour, E.; Pellenz, M.; Penna, M.; Zambenedetti, V.; Chueiri, I. A tunable fraud detection system for advanced metering infrastructure using short-lived patterns. *IEEE Trans. Smart Grid* **2019**, *10*, 830–840. [CrossRef]

10. dos Santos Benso Maciel, L.; Bonatto, B.D.; Arango, H.; Arango, L.G. Evaluating public policies for fair social tariffs of electricity in Brazil by using an economic market model. *Energies* **2020**, *13*, 4811. [CrossRef]
11. Miranda, M.; Ruffin, C.; Dantas, G.; Pedreira, N.; Guerra, M.; Zamboni, L.; Mendes, P.; Alvares, J. International review of regulatory aspects related to electricity loss in distribution sector. In Proceedings of the International Conference on Applied Energy 2019, Västerås, Sweden, 12–15 August 2019.
12. Australian Energy Regulator (AER). *2020 Electricity Network Performance Report*; Australian Energy Regulator: Melbourne, Australia, 2020.
13. Ontario Energy Board (OEB). *Filing Requirements for Electricity Transmission Applications*; Ontario Energy Board: Toronto, ON, Canada, 2022.
14. Tuttle, D.P.; Gülen, G.; Hebner, R.; King, C.W.; Spence, D.B.; Andrade, J.; Wible, J.A.; Baldick, R.; Duncan, D.; The History and Evolution of the U.S. Electricity Industry; White Paper UTEI/2016-05-2. 2016. Available online: http://energy.utexas.edu/the-full-cost-of-electricity-fce/ (accessed on 8 February 2023).
15. Comisión Nacional de Energía (CNE). *Resolución Exenta n° 699*; CNE: Santiago, Chile, 2015.
16. Dammert, A.; Carpio, R.G.; Molinelli, F. *Regulación y Supervisión del Sector Eléctrico*; Fondo Editorial de la Pontificia Universidad Católica del Perú: Lima, Peru, 2008.
17. Droguett, L.A.L. *Rentabilidad de las Empresas de Distribución y su Relación con las Fijaciones Tarifarias*; Pontifícia Universidad Catolica de Chile: Santiago, Chile, 2004.
18. Organismo Supervisor de la Inversión de Energía y Mineria (OSINERGMIN). *Fundamentos Técnicos y Económicos del Sector Eléctrico Peruano*; OSINERGMIN: Lima, Peru, 2011.
19. Presidencia de la República de Guatemala. *Acuerdo Gubernativo n°256 de 1997-Reglamento de la Ley General de Electricidad*; Presidencia de la República de Guatemala: Guatemala City, Guatemala, 1997.
20. Autorità di Regolazione per Energia Reti e Ambiente (ARERA). *Revisione dei Fattori Percentuali Convenzionali di Perdita e del Meccanismo di Perequazione delle Perdite sulle Reti di Distribuzione-Orientamenti Finali*; Documento per la Consultazione 202/2015/R/EEL; ARERA: Milan, Italy, 2015.
21. Marín, J.C. Un nuevo modelo de retribución para la distribución eléctrica. *Cuad. Energía* **2016**, *47*, 141–147.
22. Comision de Regulacion de Energia y Gas (CREG). *Resolución n°15 de 2018. Ministerio de Minas y Energía*; Comision de Regulacion de Energia y Gas: Bogota, Colombia, 2018.
23. Autoridad Nacional de los Servicios Públicos (ASEP). *Reglamento de Distribución y Comercialización de Energía Eléctrica Titulo V: Régimen de Suministro*; ASEP: Panama City, Panama, 2020.
24. Jiménez, R.; Serebrisky, T.; Mercado, J. *Power Lost: Sizing Electricity Losses in Transmission and Distribution Systems in Latin America and the Caribbean*; Inter-American Development Bank (IDB): Washington, DC, USA, 2014.
25. Savian, F.S.; Siluk, J.C.M.; Garlet, T.B.; Nascimento, F.M.; Pinheiro, J.R.; Vale, Z. Non-technical losses in Brazil: Overview, challenges, and directions for identification and mitigation. *Int. J. Energy Econ. Policy* **2022**, *12*, 93–107. [CrossRef]
26. Chaves, A.C.; Tavares, A.; Ferreira, D.; Tommaso, F.; Dantas, G.; de Barros Alvares, J.E.; Takeuchi, J.T.; Câmara, L.; Mendes, P.F.; Maestrini, M.; et al. *As Perdas Não Técnicas no Setor de Distribuição Brasileiro: Uma Abordagem Regulatória*; GESEL/UFRJ/CPFL/ANEEL: Rio de Janeiro, Brazil, 2020.
27. Zanardo, R.P.; Siluk, J.C.M.; Savian, F.S.; Schneider, P.S. Energy audit model based on a performance evaluation system. *Energy* **2018**, *154*, 544–552. [CrossRef]
28. Agência Brasileira de Energia Elétrica (ANEEL). *Perdas de Energia Elétrica na Distribuição (01/2021)*; Agência Brasileira de Energia Elétrica: Brasília, Brazil, 2021.
29. Agência Nacional de Energia Elétrica (ANEEL). *Nota Técnica No. 106/2015–Metodologia de Tratamento Regulatório Para Perdas não Técnicas de Energia Elétrica*; SGT/SRM/ANEEL: Brasília, Brazil, 2015.
30. Simões, P.F.M.; Souza, R.C.; Calili, R.F.; Pessanha, J.F.M. Analysis and short-term predictions of non-technical loss of electric power based on mixed effects models. *Socio-Economic Plan. Sci.* **2020**, *71*, 100804. [CrossRef]
31. Castro, N.; Chaves, A.C.; Ferreira, D.V.; Tommaso, F.; Ozorio, L.; Maestrini, M.; de Miranda, M.; Brandão, R.; Eduardo, J.; Mendes, P.; et al. Análise das Propostas de Alterações Metodológica para Determinação das Metas Regulatórias das Perdas não Técnicas, na Distribuição de Energia Elétrica–NT46/2020. In *TDSE Texto de Discussão do Setor Elétrico N° 94 outubro de 2020*; UFRJ—Grupo de Estudos do Setor Elétrico: Rio de Janeiro, Brazil, 2020.
32. Ahrens, A.; Hansen, C.B.; Schaffer, M.E. Lassopack: Model selection and prediction with regularised regression in Stata. *Stata J.* **2020**, *20*, 176–235. [CrossRef]
33. Zou, H. The adaptive LASSO and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
34. Agência Nacional de Energia Elétrica (ANEEL). *Nota Técnica No. 46/2020–Proposta de Consulta Pública Para Revisão da Metodologia e Atualização dos Parâmetros dos Submódulos 2.2/2.2A (Receitas Irrecuperáveis) e 2.6 (Perdas de Energia) dos Procedimentos de Revisão Tarifária-PRORET*; Agência Nacional de Energia Elétrica (ANEEL): Brasília, Brazil, 2020.
35. Hausman, J.A. Specification Tests in Econometrics. *Econometrica* **1978**, *46*, 1251–1271. [CrossRef]
36. Croissant, Y.; Millo, G. Panel-Data Econometrics in R: The plm Package. *J. Stat. Softw.* **2008**, *27*, 1–43. [CrossRef]
37. Leite, D.; Pessanha, J.; Simões, P.; Calili, R.; Souza, R. A stochastic frontier model for definition of non-technical loss targets. *Energies* **2020**, *13*, 3227. [CrossRef]

38. Aigner, D.J.; Lovell, C.A.K.; Schmidt, P. Formulation and estimation of stochastic frontier production functions. *J. Econom.* **1977**, *6*, 21–37. [CrossRef]
39. Kumbhakar, S.C.; Wang, H.J.; Horncastle, A.P. *A Practitioner's Guide to Stochastic Frontier Analysis Using Stata*; Cambridge University Press: Cambridge, UK, 2015.
40. Behr, A. *Production and Efficiency Analysis with R*; Springer: Berlin, Germany, 2015.
41. Konno, H.; Yamazaki, H. Mean-absolute deviation portfolio optimisation model and its applications to Tokyo stock market. *Manag. Sci.* **1991**, *37*, 519–531. [CrossRef]
42. Makridakis, S. Accuracy measures: Theoretical and practical concerns. *Int. J. Forecast.* **1993**, *9*, 527–529. [CrossRef]
43. David, L.; Desboulets, D. A review on variable selection in regression analysis. *Econometrics* **2018**, *6*, 45.
44. Hastie, T.; Tibshirani, R.; Tibishirani, R. Best Subset, Forward Stepwise, or LASSO? Analysis and Recommendations Based on Extensive Comparisons. *Stat. Sci.* **2020**, *35*, 579–592. [CrossRef]
45. Berstsimas, D.; King, A.; Mazumder, R. Best subset selection via a modern optimisation lens. *Ann. Stat.* **2016**, *44*, 813–852.
46. Calcagno, V.; Mazancourt, C. glmulti: An R package for easy automated model selection with generalised linear models. *J. Stat. Softw.* **2010**, *34*, 1–29. [CrossRef]
47. Silva, S.F.P.; Pessanha, J.F.M. Identificação de indicadores para previsão de insolvência das distribuidoras de energia elétrica por meio de regressão logística para dados em painel. *Contabilometria. Braz. J. Quant. Methods Appl. Account.* **2022**, *9*, 73–91.
48. Martins, F.; Almeida, M.F.; Calili, R.; Oliveira, A. Design Thinking applied to smart home projects: A user-centric and sustainable perspective. *Sustainability* **2020**, *12*, 10031. [CrossRef]
49. Breusch, T.S.; Pagan, A.R. A Simple Test for Heteroskedasticity and Random Coefficient Variation. *Econometrica* **1979**, *47*, 1287–1294. [CrossRef]
50. Hsiao, C.; Appelbe, T.W.; Dineen, C.R. A General Framework for Panel Data Analysis—With an Application to Canadian Customer Dialed Long Distance Service. *J. Econom.* **1993**, *59*, 63–86. [CrossRef]
51. Hsiao, C. *Analysis of Panel Data*, 2nd ed.; Econometric Society Monograph 36; Cambridge University Press: New York, NY, USA, 2003.
52. Greene, W.H. *Econometric Analysis*, 7th ed.; Prentice Hall: Hoboken, NJ, USA, 2012.
53. Baltagi, B. *Econometric Analysis of Panel Data*, 2nd ed.; Wiley: New York, NY, USA, 2001.
54. Breusch, T.S.; Godfrey, L.G. *A Review of Recent Work on Testing for Autocorrelation in Dynamic Economic Models*; Discussion Paper n. 8017; University of Southampton: Southampton, UK, 1980.
55. Wooldridge, J. *Econometric Analysis of Cross Section and Panel Data*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2010.
56. Myung, I.J. The importance of complexity in model selection. *J. Math. Psychol.* **2000**, *44*, 190–204. [CrossRef]
57. Burnham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed.; Springer: New York, NY, USA, 2002.
58. Bozdogan, H. Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions. *Psychometrika* **1987**, *52*, 345–370. [CrossRef]
59. Bozdogan, H. Akaike's Information Criterion and Recent Developments in Information Complexity. *J. Math. Psychol.* **2000**, *44*, 62–91. [CrossRef] [PubMed]
60. Wagenmakers, E.-J.; Farrell, S. AIC model selection using Akaike weights. *Psychon. Bull. Rev.* **2004**, *11*, 192–196. [CrossRef] [PubMed]