Brazilian Commerce and Distribution Companies:

a bayesian semiparametric stochastic frontier analysis

Vitor H. Tavares da Silva[1] and Guilherme Valle Moura[2]

[1]Universidade Federal de Santa Catarina (UFSC))

[2]Universidade Federal de Santa Catarina (UFSC)

March 19, 2025

Abstract

Commerce and distribution firms face dynamic challenges accelerated by the COVID-19 pandemic, including shifting consumer preferences, rapid digitalization, and competition from global e-commerce. This study evaluates the technical efficiency of Brazilian retail and wholesale publicly traded firms using a Bayesian semiparametric stochastic frontier analysis (SFA) with regression trees (BART). Compared to traditional parametric SFA models, the BART approach demonstrates greater flexibility in capturing nonlinear labor-capital-revenue relationships and robust uncertainty quantification. Results indicate the BART model achieves superior fit compared to parametric methods. Efficiency segmentation is significant: food and home appliance firms present the highest efficiency, while textiles are the lowest. The sector shows labor-intensive technology, decreasing returns to scale, and stable efficiency over time. The study highlights the risk of overestimating inefficiency with restrictive models and advocates for integrating Bayesian nonparametric methods in efficiency analyses. These insights inform managerial and policy decisions on resource allocation and competitiveness in Brazil's critical commerce sector.

Keywords: commerce and distribution; stochastic frontier analysis; nonparametric models; Bayesian inference.

1 Introduction

Companies engaged in activities related to the trade and distribution of goods have frequently been the focus of discussions and analyses by investors, banks, customers and the government. Following the COVID-19 pandemic, new challenges emerged for these companies — from the direct impacts of the sanitary and humanitarian crisis, to changes in consumer habits and market trends. In emerging economies, the effects were even more severe (Hevia et al. [2020]).

In Brazil, national retail and wholesale companies had to deal with the strong penetration of foreign players in e-commerce and the greater integration between various online and offline sales channels (Costa et al. [2020], Cruz [2021], Delardas et al. [2022]). Significant changes in consumption patterns are also notable, bringing new characteristics such as a reduction in the number of visits to physical stores, increased demand for food, and a strong preference for digital sales channels (Gupta and Mukherjee [2022]). Although related to the pandemic period, many of these changes began before the crisis, were accelerated by the harsh context, and represent a new pattern with noticeable medium and long-term effects.

In a context of increased competitiveness, new challenges, and economic recovery, the ability to rigorously analyze the level of efficiency of these companies can shed light on how the sector has faced this scenario, the effects felt, and, most importantly, which firms are more capable of making better use of resources. This constitutes a synthesis of various pertinent information for agents operating in this market. This is the issue addressed by this paper.

As a general objective, this research seeks to analyze the technical efficiency level of Brazilian trade and distribution companies listed on B3 - the brazilian stock market. To achieve this, it aims to estimate a nonparametric stochastic frontier analysis (SFA) model based on Bayesian regression trees, comparing its results with the estimates made through convential SFA models, developed by Aigner et al. [1977] and Battese and Coelli [1992], considering both Cobb-Douglas and translog production frontiers.

On the one hand, this proposal is justified by the relevance of the sector's companies in the Brazilian economy: companies listed by the B3 in the subsectors of "Commerce" and "Commerce and Distribution" account for 4.8% of the IBOVESPA portfolio — the main index of the Brazilian stock market — and approximately R$ 200 billion in market cap. Furthermore, in 2021, the Brazilian retail sector generated revenues amounting to R$ 4,404 billion, considering only companies with at least 20 employees (IBGE [2021]) and, in the second quarter of

2024, responded to 18.9% of the total occupied workers in Brazil, also playing a significant role in the reduction of informal employment (IPEA [2024]).

Beyond that, the adoption of nonparametric modeling constitutes a relevant contribution to the literature, bringing greater flexibility to traditional SFA models and avoiding overly restrictive assumptions. Furthermore, the small number of studies focused on the efficiency levels of Brazilian companies represents a wide space for adopting this type of analysis in the construction of new empirical evidence.

This paper is structure as follows: after the current introduction, the theoretical framework is presented, detailing the context of the Brazilian companies to be evaluated and theoretical and empirical contributions to the chosen approach; subsequently, the methodology to be used is introduced, detailing the traditional SFA approaches and the Bayesian nonparametric model, as well as the dataset. Following this, the results of both models are presented, highlighting it's fit performance and some of the insights we can get from the estimations. The last section summarizes the results in face to the previous works and expectations, also building a comparison between the models and suggesting new extensions to it.

2 Literature Review

The sample of companies analyzed in this study includes those classified by B3 in the subsectors of "Commerce" and "Commerce and Distribution". This nomenclature identifies firms belonging to four distinct sectors - Industrial Goods, Cyclical Consumption, Non-Cyclical Consumption, and Health - but whose operations have similarities. Regarding the taxonomic criteria used, the Brazilian stock market defines:

> For the classification of companies, the products or services that contribute the most to the formation of the companies' revenue were analyzed, considering also the revenues generated within the scope of invested companies proportionally to the shareholdings held. (B3 [2023])

Thus, by adopting the classification used by B3, we expect to ensure comparability among the analyzed companies without neglecting their respective idiosyncrasies.

Although it does not exhaust the fields of operation of the selected firms, retail - defined as the set of "activities and steps necessary to place a product in the hands of a consumer or provide a service to the consume" (Dunne et al. [2011]) - constitutes a fundamental part of the business. Therefore, the broad range of activities encompassed by companies operating in this segment is fertile ground for various studies in different fields of knowledge.

In an effort to synthesize the various forces acting over the sector at the time, Santos and Costa [1997] highlight the diversity of activities performed (product selection, acquisition, distribution, commercialization, and delivery), the high absorption of low-skilled labor, and the high sensitivity to economic policy. In addition, they emphasize that greater access to information places customers at the center of more intense competition between companies, leading to the conclusion that "there is, therefore, no ideal retail format, being more important the search for the efficiency of the chosen business and the definition of consistent strategic options" (Santos and Costa [1997]).

More recently, de Barcellos et al. [2015] sought to outline a new panorama for the sector, based on the context faced by the four largest supermarket chains operating in the country. Although restricted, the work highlights the difficulty in competing with local and small companies - a fact also addressed by Farina et al. [2005] - and meeting new consumer demands, increasingly concerned with sustainability and variety in the mix of products. Moreover, they

highlight the need to integrate online and offline shopping experiences, designing sales models called "omnichannel" that are simultaneously efficient and aligned with customer expectations.

The adoption of a digital culture aligned with the trends of an increasingly online market is the focus of the work of Pinto et al. [2023]. Conducting a cluster analysis, the authors identified a high level of digitalization in Brazilian retail, but pointed out the need for greater investments in small and medium-sized enterprises, which face higher barriers to digital transformation.

Turning attention to topics related to technical efficiency, de Melo et al. [2018] analyze the Brazilian supermarket segment between 2005 and 2012 using the Data Envelopment Analysis (DEA) methodology, a nonparametric alternative to stochastic frontier models. The authors point to greater efficiency in large companies, but faster growth in smaller firms - while the first group presented technology with decreasing returns to scale, the latter demonstrated increasing gains. Besides the size effect, the weight of technological progress on efficiency gains is also highlighted, as well as a notable weight of geographical factors. Finally, the role of mergers and acquisitions is also pointed out as a channel for efficiency gains among the major players. In turn, using a SFA model, Vries [2010] analyze the Brazilian retail sector, comparing the performance of formal and informal companies - indicating greater efficiency in the first group.

The research by Lee and Tyler [1978] was globally pioneering in using microaccounting and financial data to estimate an SFA model. The work was based on data collected from 850 industrial companies in Brazil and highlighted the need, at the time, for further investigations into the impact that the original assumptions of the Aigner et al. [1977] model might have on the results. Taylor and Shonkwiler [1986] also used Brazilian data, this time observing the performance of rural producers in the Southeast region under a rural credit program. The authors employed both the SFA and a more primmitive approach called the Full Frontier Model, nowadays obsolete.

Other works focused on Brazil include: Tovar et al. [2011], observing electric energy distribution companies and concluding that mergers and acquisitions can represent efficiency leaps in large companies; Leite et al. [2020], who propose, from an SFA model adjusted for non-technical efficiency losses, a new approach to the Brazilian energy regulatory agencies; Schmidt et al. [2008], who analyze the productivity of farms located in the Brazilian Midwest and point to spatial effects on technical efficiency; and Schneider et al. [2020], who use the

approach to evaluate the efficiency of Brazilian pension funds.

The international literature using SFA models is vast, particularly covering the financial and industrial sectors: Diaz and Sanchez [2007] analyze small and medium-sized Spanish industries, concluding that larger companies are less efficient. The authors relate this phenomenon to the greater difficulty in managing resources and people in large organizations, as well as the market selection hypothesis, in which small and inefficient companies tend not to prosper and, therefore, go bankrupt quickly; Fenn et al. [2008] observe the European insurance market and reach similar conclusions - larger companies with high market share tend to be more inefficient; Alshammari et al. [2019] also analyze the insurance market but focus on the *takaful* segment, present in countries influenced by Islamic culture and characterized by a less competitive model based on mutual cooperation principles. In this case, the less competitive scenario is related to the lower efficiency of managers.

The work from Breivik et al. [2023] analyses the inventory efficiency of Norwegian commercial firms through an SFA model, based on the premisse that inventory efficiency leads to a better financial performance. The study found a positive relationship between the number of employees and efficiency, also highlighting the role of environmental variables on efficiency. In the Peruvian market, Alvarez et al. [2020] indicates a positive relationship between e-commerce adoption and efficiency in retail and wholesale firms, with evidences that suggest labor-intense technologies and decreasing returns to scale.

The SFA approach continues to receive notable contributions: Kumbhakar and Lovell [2000] access heteroskedascity in SFA models with single or multiple outputs, while also bringing a deep review on the framework's foundations; Robaina-Alves et al. [2015] innovates by using entropy-based estimators to combine characteristics of DEA and SFA approaches, applying the technique to analyze the eco-efficiency of European economies; Simar and Wilson [2022] develop a nonparametric approach to SFA, allowing the estimation of models with multiple outputs and using less restrictive assumptions.

Futhermore, the combination of Bayesian methods to SFA is also a vast field, adressing more complex inefficiency structures and model uncertainty: Tsionas [2005] makes a detailed introduction, while Koop and Steel [2003] provides a broader perspective - also pointing out the possibility of estimating nonparametric frontiers; Tsionas [2003] propose and apply to American airlines firms a method that takes DEA efficiency results as priors for a Bayesian SFA estimation using the Gibbs sampler algorithm with data augmentation - a framework that

innovates on providing data-oriented priors for Bayesian inference of SFA models; Koop et al. [1999] use noninformative priors and Gibbs sampling to estimate Bayesian SFAs for analysing economic growth among OECD countries, highlighting the benefits of uncertainty quatification for small datasets and the capability to identify efficiency changes' determinants. Other recent applied works include Tsionas et al. [2019], who develop methods that allows for transitions and switches between technologies and Carvalho and Marques [2016], analysing the water sector in Portugal and also highlighting the benefits of Bayesian SFA methods on small samples.

In general terms, as highlighted by Parmeter and Kumbhakar [2023], the further developments from the seminal work of Aigner et al. [1977] and the popular panel-data formulation of Battese and Coelli [1992] are mainly focused on robustness, looking for methods that do not rely on the parametric assumptions about inefficiency and error terms distributions, and the functional form of the frontier. In other words, more robust frameworks for estimating SFA models would lead to more general results, where prior specifications (in both frequentist or Bayesian approaches) do not add any sort of bias to the results - an aspect that, in the Bayesian framework, could be partially addressed through prior sensitivity analysis.

In this sense, the works of Tsionas [2022] and Wei et al. [2024] are notorious recent contributions, combining Bayesian inference methods to nonparametrical production functions - estimated via Bayesian additive regression trees (BART). Those approaches benefit from the flexibility and precision of tree-based models, as well as the better uncertainty quantification that Bayesian proccedures allow. In this paper, we follow a similar approach, bringing further details in the following section.

Thus, our study tackle two major gaps in the existent literature: first, the absence of applied studies that analyse the efficiency measures of the largest Brazilian retail and wholesale firms, making use of publicly available microaccounting data and considering listed companies, which tipically are object of rigourous analysis by market agents; and second, the implementation of new approachs to the well-established SFA method, estimating a nonparametric frontier and implementing Bayesian inference procedures. With those contributions, we aim to provide robust and trustful evidence about the sector's efficiency levels, as well as estimulate further researches, applying those methods for different markets, countries or even investigating other aspects of the findings we present.

3 Methodology

In this section, we present in details the standard formulation of the stochastic frontier models and the novel - and more flexible - structure implemented in this paper.

3.1 The traditional stochastic frontier model

The Stochastic Frontier Analysis (SFA) models were initially developed by Aigner et al. [1977], in an attempt to fill the gap between conventional economic theory and econometric work.

Assume the following production function:

$$Y_{it} = f\left(\mathbf{X_{it}}; \beta\right) \xi_{it} \tau_{it} \tag{3.1.1}$$

where Y_{it} represents the production obtained through the technology described in $f(.)$, from the vector of inputs X_{it} and the vector of unknown parameters β for firm i in period t. Additionally, the stochastic terms ξ_{it}, representing a random shock and $0 < \tau_{it} < 1$, representing a measure of technical efficiency of the firm in the period, also affect the output.

Assuming a Cobb-Douglas production function, from Equation 3.1.1 one can obtain

$$y_{it} = \sum_{j=1}^{K} \left(x_{it\,j} \cdot \beta_j\right) + v_{it} - u_{it} \tag{3.1.2}$$

where $y_{it} = \ln Y_{it}$; $x_{it\,j} = \ln X_{it\,j}$; $v_{it} = \ln \xi_{it}$; $u_{it} = \ln \tau_{it}$; and $j = 1, \ldots, K$ denotes the number of inputs used in the production process. The most common formulation in the literature defines $v_{it} \overset{iid}{\sim} N(0, \sigma_v^2)$ and $u_{it} \overset{iid}{\sim} N^+(\mu, \sigma_u^2)$ and, as specified by Battese and Coelli [1992], the estimator for the technical efficiency measure of firm i in period t is obtained by:

$$E\{\exp(u_{it}) \,|\varepsilon_{it}\} = \left[\frac{1 - \phi\{\eta_{it}\tilde{\sigma}_i - \left(\tilde{\mu}_i/\tilde{\sigma}_i\right)\}}{1 - \phi\left(-\tilde{\mu}_i/\tilde{\sigma}_i\right)}\right] \exp\left(\eta_{it}\tilde{\mu}_i + \frac{1}{2}\eta_{it}^2\tilde{\sigma}_i^2\right) \tag{3.1.3}$$

where

$$\tilde{\mu}_i = \frac{\mu\sigma_v^2 - \sum_{t=1}^{T_i} \eta_{it}\varepsilon_{it}\sigma_u^2}{\sigma_v^2 + \sum_{t=1}^{T_i} \eta_{it}^2\sigma_u^2} \tag{3.1.4}$$

$$\tilde{\sigma}_i^2 = \frac{\sigma_v^2 \sigma_u^2}{\sigma_v^2 + \sum_{t=1}^{T_i} \eta_{it}^2 \sigma_u^2} \tag{3.1.5}$$

$$\varepsilon_{it} = y_{it} - \sum_{j=1}^{K} \left(x_{it\,j} \cdot \beta_j \right) \tag{3.1.6}$$

Beyond that,

$$\eta_{it} = \exp\left[-\eta \left(t - T \right) \right] \tag{3.1.7}$$

where η is an unknown parameter and $t \in \mathscr{T}i$, the set of Ti periods for which there are observations for firm i. In other words, the model allows the efficiency term to vary over time. For models where the relative inefficiency does not vary over time, $\eta = 0$.

One common assumption made when dealing with SFA models it to estimate a translog production function (SFA-T), instead of a Cobb-Douglas (SFA-CD), such that

$$y_{it} = \sum_{j=1}^{K} \left(x_{it\,j} \cdot \beta_j \right) + 0.5 \sum_{j=1}^{K} \left(x_{it\,j}^2 \cdot \gamma_j \right) + \sum_{j=1}^{K} \sum_{l=1}^{K} \left(x_{it\,j} x_{itl} \theta_{jl} \right) + v_{it} - u_{it} \tag{3.1.8}$$

This is a more flexible specification, allowing for inputs' elasticities to vary across firms and time, as well as capturing nonlinear interactions between inputs and output Pavelescu [2011]. In this research, both models are estimated and the results are compared with the nonparametric frontier, presented in the following section.

For both approaches, the estimation is made through maximum likelihood.

3.2 The tree-based semiparametric approach

Although it provides consistent and easily interpretable results, the "classic" SFA model has the main weakness of using excessively restrictive assumptions: it requires the specification of a production function, as well as the distributions that describe the behavior of u and v. This paper propose a more flexible approach, replacing some of those assumptions with a more computationally demanding solution that can generate results with greater empirical adherence, without losing interpretability or grounding in economic theory.

From Equation 3.1.2, we will replace the somatory term with a sum-of-trees Bayesian Additive Regression Trees (BART) model, assuming the form

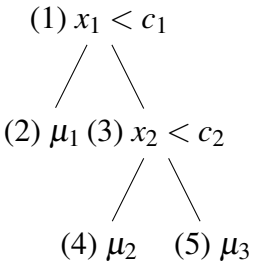
$$y_{it} = g(X_{it\,j}, \theta) + v_{it} - u_{it} \tag{3.2.1}$$

where θ represents the set of parameters that compose the BART model. The estimation proccedure relays on a Markov-Chain Monte Carlo (MCMC) approach and further details are present along this section.

The concept of regression trees was originally proposed by Breiman et al. [1984], presenting a nonparametric approach to regression and classification problems. The central idea is to segment the sample based on the observed values of $\mathbf{X_i}$ - the vector of explanatory variables. For each observation within these segments - referred to as "leaves" or "terminal nodes" - the same value for the dependent variable Y_i is estimated (typically the mean is used, but some approaches suggest using the median or even performing a linear adjustment). This segmentation is defined based on splits, binary decisions made by comparing the observed value of x_{ik} (observation of variable k for individual i) with a cutoff rule such as the mean $\bar{x}_k$.

Figure 3.2.1 illustrates the decision-making behavior for estimating a regression tree. The first split is based on the covariate x_1, comparing the observed values in each observation x_{1i} with a constant c_1. If $x_{1i} < c_1$, it is estimated that the dependent variable is μ_1 (usually, the mean of the dependent variable in the partition that meets the criterion). Otherwise, the model analyzes the covariate x_2 in a way analogous to the first.

Figure 3.2.1: Regression tree example



However, the literature points to a recurring problem in this type of modeling: there is a strong tendency for the model to overfit the training data. As a result, the estimators lose their generalization ability, becoming biased towards the specific way the variables interact in the sample.

To mitigate this undesirable behavior, several strategies have been developed. Algorithms like boosting, random forests and bagging (a combination of bootstrap and adding) are based on the premise that the concatenation of several small and weakly explanatory trees (weak

learners) is more efficient than adopting large, individual trees. These approaches are part of the ensemble models framework.

To address this same problem, Chipman et al. [2010] proposed the BART models. To ensure that each of the m trees is a weak learner, it uses nondeterministic hyperparameters. That is, it is defined that these will be treated as random variables that follow probability density functions defined *a priori* and tend to generate small trees. According to the authors, "BART can be viewed as a nonparametric Bayesian approach that fits a rich model using a strongly influential *a priori* distribution." (Chipman et al. [2010]).

Formalizing the presented concepts, consider that there is an interest in estimating the random variable Y given a matrix of covariates x, such that

$$Y = f(x) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

(3.2.2)

The goal is to approximate $f(x) = E(Y|x)$ using a sum-of-trees model $h(x) = \sum_{j=1}^{m} g_j(x)$ where each g_j represents a tree.

Let T be a binary tree consisting of a set of decision nodes, with binary split rules (for example, nodes 1 and 2 in Figure 3.2.1) and b terminal nodes, with estimates for Y (nodes 2, 4 and 5 in Figure 3.2.1). Furthermore, $M = \{\mu_1, \mu_2, \ldots, \mu_b\}$ represents the set of parameters associated with each terminal node. By "binary," it is understood that in this tree, each decision node generates only two new nodes, with a split criterion in the format $x_k \leq c_k$ vs $x_k > c_k$ for continuous variables. For dummy variables, $c = 0$.

In this structure, each observation i of the covariate matrix x is associated with one and only one, terminal node of T, from the chain of binary decisions that originate them. Thus, $g(x;T;M)$ denotes the function that associates each observation of covariates x_i with a $\mu_i \in M$. In this way, the sum-of-trees model is constructed.

$$Y = \sum_{j=1}^{m} g(x;T_j;M_j) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2) \tag{3.2.3}$$

where $E(Y|x)$ is the sum of each terminal node μ.

Finally, it remains to specify the shrinkage prior that allows the model to generate

weak learner trees. For simplification purposes, the authors focus only on the parameters $(T_1, M_1) \dots (T_m, M_m)$ and σ, which are independent of each other.

For $p(T_j)$, it is defined that the probability of a node of depth $d = 0, 1, 2, \dots$ not being terminal (i.e., the probability of performing the split at a node of depth d) is given by

$$\alpha(1+d)^{-\beta}, \qquad \alpha \in (0,1), \beta \in [0, \text{inf}) \tag{3.2.4}$$

In other words, by α and β, it is possible to probabilistically limit the size of the trees so that at each new depth level, the probability of performing a new split is smaller. Cross-validation is commonly used in the literature to test different configurations for these hyperparameters.

Additionally, a Uniform distribution is used to select which covariate will be used in a node, as well as to define the selection criterion adopted in each node.

Regarding $p(\mu_i|T_j)$, the conjugate Normal distribution $N(\mu_\mu, \sigma_\mu^2)$ is used, which is recurrent in various formulations of the Bayesian approach due to the computational gains obtained with its use. The authors argue that, in empirical applications of the model, it is highly probable that the *a priori* values of μ_{ij} will be located in the interval between the minimum and maximum points of the dependent variable, y_{min} and y_{max}. Therefore, it is coherent to opt for specifications of μ_μ and σ_μ^2 that concentrate the probability mass of $p(\mu_i|T_j)$ within this interval. The authors provide these specifications, but in terms of applicability, it is sufficient to centralize Y between -0.5 and 0.5, adopting the transformed version of Y as the dependent variable of the model and defining a value of σ_μ^2 that respects the identity $h\sqrt{m\sigma_\mu} = 0.5$ for a given value of the constant h. That is,

$$\mu_i|T_j \sim N(0, \sigma_\mu^2) \qquad \sigma_\mu^2 = \frac{0.5}{h\sqrt{m}} \tag{3.2.5}$$

It is worth noting that this definition of $p(\mu_i|T_j)$ induces the model to generate weak explanatory trees by shrinking the parameters μ_{ij} to zero and, consequently, reducing the weight of the tree T_j in $E(Y|x)$.

Finally, $p(\sigma)$ is defined as a conjugate Inverse Chi-Square distribution, such that

$$\sigma^2 \sim \frac{\nu\lambda}{\chi_\nu^2} \tag{3.2.6}$$

Again, the hyperparameters can be chosen through a cross-validation procedure. How-

ever, the authors map out combinations that generate more or less conservative estimates.

Finally, the number of trees that make up the model, m, must also be defined beforehand. Like the others, this hyperparameter can be adjusted through cross-validation, but the authors suggest adopting $m = 200$ and then testing if small variations in this value generate substantial changes in the model's performance. As a rule of thumb, values above 50 are sufficient to generate satisfactory performance metrics.

The sampling and estimation procedures using Markov Chain Monte Carlo (MCMC) follow the procedure detailed by Chipman et al. [2010] and Kapelner and Bleich [2016].

However, notice that in Equation 3.2.1 the BART model explains only part of the observed outcomes Y_{it}; the residuals are, then, separated in two terms, that - just as in the parametric SFA approach - aims to explain the deviations based in a random error v_{it} and a inefficiency term u_{it}. Thus,

$$Y_{it} = \sum_{j=1}^{m} g(x; T_j; M_j) + v_{it} - u_{it} \tag{3.2.7}$$

Using Bayesian inference techniques, instead of assuming the probability functions that u and v follow, we estimate 3.2.7 via MCMC, in a hierarchical structure where the priors for those parameters are inspired by the assumptions that the usual SFA approach make. Formally,

$$u_0 \sim N^{+}(\mu, \sigma_u^2) \qquad \sigma_u^2 \sim IG(\alpha_{\sigma_u^2}, \beta_{\sigma_u^2}) \tag{3.2.8}$$

$$v_0 \sim N(0, \sigma_v^2) \qquad \sigma_v^2 \sim IG(\alpha_{\sigma_v^2}, \beta_{\sigma_v^2}) \tag{3.2.9}$$

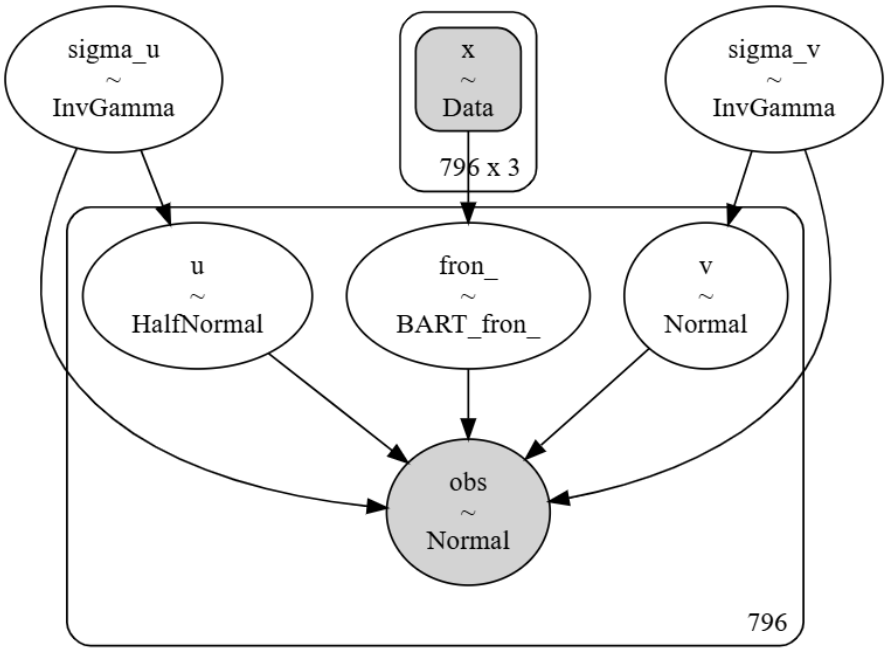
The hierarchical structure is resumed in the graph visualization in Figure 3.2.2:

The MCMC estimation makes use of the NUTS algorithm, made available through the PyMC framework. The variables `v_hat` and `u_hat` are deterministic PyMC variables used to store samples of u_{it} and v_{it} for each observation.

3.3 Data

This paper analyzes Brazilian companies traded on the B3 stock market that are part of the "Commerce" and "Commerce and Distribution" segments. Two inputs were considered: labor and capital. As a proxy for the first, we considered the total expenditure with social and labor obligations; for the latter, the total stock of fixed assets. The output is the net revenue,

Figure 3.2.2: Graph - tree-based semiparametric SFA



both measured quarterly and publicly available. All variables were adjusted according to the official Brazilian inflation, the "IPCA" index.

The time period under analysis extended from the first quarter of 2016 until the third quarter of 2024, resulting in a database with 796 observations (the amount of observations available for each company may vary). In addition, the observations of "americanas s.a" were discarted, as the company faces a bankruptcy proccess due to investigations of fraud and masking of financial reports. "Rodobens" and "Profarma" were also disconsidered, as the companies did not share their labor expenses, leading to a sample with 28 companies, segmented in 5 industries: *Food*, *Home appliances*, *Transport materials*, *Medicines and other products*, *Fabrics, clothing and other products* and *Miscellaneous products*. This segmentation is also made by B3. Descriptive statistics for the considered variables are presented in Table 3.3.1.

For the BART model, it was also generated a discrete numerical variable to identify the year-quarter of each observation. This solution aims to identify if exists any relationship between firms' efficiency and time - a result that, using the parametric SFA approach, may be visualized through the η parameter. The final dataset is available in the link provided at the Appendix.

The descriptive statistics illustrate how diverse and heterogeneous the sector is, including companies of different sizes, market segments and origins, advocating for rigourous methods to evaluate operational efficiency and capability of generating revenues in such a com-

Table 3.3.1: Descriptive Statistics (Millions R$)

Company	**Avg. Net Revenue.**	**Avg. Labor Expend.**	**Avg. Fixed Assets**	**Obs.**
Carrefour BR	55270.7494	1085.5932	21332.0586	32
P.Acucar-Cbd	35367.7034	904.8036	16166.0781	35
Assai	32616.5541	675.6075	15358.4013	8
Casas Bahia	20115.669	605.6292	4190.716	35
Magaz Luiza	17681.9974	393.7454	3632.1136	35
RaiaDrogasil	15596.881	466.806	4705.7407	35
Grupo Mateus	12667.1246	303.5777	3746.8324	22
Pague Menos	7132.3748	210.9163	1006.2682	9
Lojas Renner	6786.4173	343.0366	4409.833	35
Viveo	5537.9892	106.5973	413.2499	18
Guararapes	5241.3981	295.445	2929.7701	35
Cea Modas	3969.7543	186.9258	2458.3869	26
Hypera	3674.4878	282.3859	2066.0531	35
Allied	3591.5381	57.3506	143.5722	22
Grupo Sbf	3030.9153	176.9574	1854.9327	26
Dimed	2277.317	63.4338	727.1332	35
Petz	1894.323	91.3838	1719.904	13
Lojas Marisa	1854.9194	99.3389	336.6979	35
Azzas 2154	1783.5076	84.6972	454.395	35
Quero-Quero	1421.2345	84.8526	558.8489	22
Wlm Ind Com	913.4482	16.9059	380.0917	35
Blau	842.4406	38.3125	411.19	29
Veste	787.2396	63.9168	406.89	35
Le Biscuit	739.4376	25.065	625.7339	22
Espacolaser	545.6905	50.7613	348.4171	22
Minasmaquina	413.487	5.7033	39.8778	35
Grazziotin	378.5302	15.1745	302.7242	35
Embpar S/A	335.69	12.9054	25.7863	35

petitive environment.

4 Results

In this section we present the results of both approaches: the frequentist parametric SFA on both specifications (Cobb-Douglas and translog), presented in Section 3, and the bayesian semiparametric approach. The quality of adjustment is compared by their root mean squared errors (RMSE) and mean absolute percentage errors (MAPE).

4.1 Parametric SFA with Cobb-Douglas frontier

The stochastic frontier based on a Cobb-Douglas production function showed the weakest fit performance, with RMSE equal to 1.3662 and MAPE equal to 8.25%. The estimated parameters are presented in Table 4.1.1.

Table 4.1.1: Results - SFA-CD

	Coef.	Std. Error	z	P $\|z\|$
ln_fixed_assets*	0.1869	0.0436	4.28	0.000
ln_labor_exp_*	0.5570	0.0519	10.74	0.000
const.*	6.8604	0.7257	9.45	0.000
η	0.0011	0.0021	0.55	0.584
μ	0.3945	0.2255		
σ_u^2	0.4394	0.1957		
σ_v^2	0.3532	0.0181		

* Significant at 99% level

The SFA-CD results suggest a highly labor-intense technology, which is alligned to some of the studies previously mentioned (Vries [2010]; Alvarez et al. [2020]). The central role that physical stores still plays in the sector, as well as the growing vertical supply chain integration in the biggest companies, may help to explain the relevance of labor factor in the industry - even with the continuous trend to automatization and digitalization. It also dialogue with the relevance the sector plays in the Brazilian job market, as previously mentioned.

About the inefficiency related parameters, the results indicate that most of the error's variance is explained due to inneficiency ($\gamma = \sigma_u^2/(\sigma_u^2 + \sigma_v^2) = 0.5543$). The parameter η, that measures the evolution of the inneficiency measures throughout the time, is stastiscally insignificant, suggesting that the technical efficiency of the analysed companies are stable along the whole period under analysis. In addition, the linear combination of the input parameters is lower than 1 with a 99% confidence, evidencing that the technology of these companies shows decreasing returns to scale.

The relative inefficiency's stability throughout the time may indicate two distinct phonemenons: the first hypothesis is that there were no relevant technical improvements in the sampled firms along the period under analysis, which could generate a concern about lack of innovation and investment; otherwise, it could simply evidence that technical improvements happened, but affected all the firms homogeneously and did not impact the relative inefficiency measure. In this sense, further investigations must be implemented to evaluate which of the scenarios better describe the industry behavior in the last years.

Table 4.1.2 presents the technical efficiency measures estimated through Equation 3.1.3. Food companies - which are, essentially, supermarket chains that operate both in retail and wholesale - are among the most efficient. On the other hand, firms in the clothing sector are the most inefficient. A notorious segmentation effect is highlighted, evidencing that although all the sample consists of commercial firms, the market in which they operate defines distinct levels of inefficiency.

It calls attention how the size effect over efficiency shows a direct relationship: the biggest firms (in terms of market share) are clearly among the most efficient ones, in contradiction to what most of the previous studies that analyse other economic sectors found and not corroborating the market-selection hypothesis. On the other hand, those evidences are alligned to the relationship of M&A in big companies and efficiency gains. Some examples are: *Carrefour BR* has made many acquisitions in the 1990's, focusing on local supermarket chains, and since 2020 it acquired stores and a whole brand from the global retail player Walmart; *Assai* and *P.Acucar-Cbd* were part of the same company for several years and split into two different companies in 2020, also counting on a notorious path of mergers and acquisitions of regional supermarkets; *Magaz Luiza* is one of the biggest retail companies in Brazil, acquiring several competitors in the 2000's and many tech and logistics firms since 2019, positioning itself as one of the major e-commerce players in the local market; although smaller operations, *Viveo*, *RaiaDrogasil* and *Pague Menos*, the biggest players in the segment of drugstores and medicines retail, also made several M&As, specially in the last 15 years.

4.2 Parametric SFA with translog frontier

The stochastic frontier model with translog production frontier (Equation 3.1.8) performed better in both adjustment metrics, with a RMSE equal to 1.0951 and MAPE equal to 6.40%. The more flexible structure allows the model to capture nonlinearities between the in-

Table 4.1.2: Technical Effiency - SFA-CD

Company	Avg. Efficiency	Segment (B3)	Market Share.[1]
Carrefour BR	0.8971	Food	23.86%
Allied	0.8044	Home appliances	1.17%
Assai	0.7475	Food	15.18%
Magaz Luiza	0.7328	Home appliances	7.71%
Viveo	0.7211	Medicines and other products	2.45%
P.Acucar-Cbd	0.6773	Food	3.84%
Casas Bahia	0.6553	Home appliances	5.44%
Grupo Mateus	0.6279	Food	6.61%
RaiaDrogasil	0.5819	Medicines and other products	8.14%
Pague Menos	0.5483	Medicines and other products	2.63%
Minasmaquina	0.4054	Transport materials	0.28%
Dimed	0.3748	Medicines and other products	1.02%
Wlm Ind Com	0.3076	Transport materials	0.67%
Lojas Renner	0.2941	Fabrics. clothing and other products	2.78%
Azzas 2154	0.2814	Fabrics. clothing and other products	1.41%
Lojas Marisa	0.271	Fabrics. clothing and other products	0.26%
Guararapes	0.2681	Fabrics. clothing and other products	1.87%
Cea Modas	0.2669	Fabrics. clothing and other products	1.44%
Embpar S/A	0.2309	Transport materials	0.01%
Petz	0.2204	Miscellaneous	0.69%
Grupo Sbf	0.2195	Miscellaneous	1.41%
Quero-Quero	0.2141	Miscellaneous	0.55%
Blau	0.2139	Medicines and other products	0.37%
Hypera	0.2084	Medicines and other products	1.68%
Le Biscuit	0.1962	Miscellaneous	0.44%
Grazziotin	0.186	Fabrics. clothing and other products	0.14%
Veste	0.1372	Fabrics. clothing and other products	0.23%
Espacolaser	0.1253	Miscellaneous	0.22%

[1] Note: Measured as the company's share in the sum of revenues of the last available quarter (2024q3).

puts and the outputs, as well as a more detailed view on returns to scale. The results are shown in Table 4.2.1.

The γ parameter is equal to 0.7709, significantly larger then the estimation from SFA-CD and advocating for an even greater impact of inneficiencies over the deviations from the frontier. Alligned to the previous results, the translog frontier does not indicate any time-effect over efficiency measures. Furthermore, no significant changes are observed in the efficiency ranking (Appendix A.2.2) and the apparent positive correlation between efficiency and firm size.

A characteristic peculiar to the SFA-T model are the less interpretable parameters,

Table 4.2.1: Results - SFA-T

	Coef.	**Std. Error**	**z**	P $\|z\|$
const ***	10.2829	2.3399	4.3945	0.0001
ln_labor_exp ***	-0.9441	0.2750	-3.4327	0.0006
ln_fixed_assets *	0.7838	0.3385	2.3154	0.0206
I(0.5 * ln_labor_exp^2)*	0.1365	0.0644	2.1196	0.0340
I(0.5 * ln_fixed_assets^2)	-0.0753	0.0545	-1.3809	0.1673
I(ln_fixed_assets * ln_labor_exp)	0.0210	0.0589	0.3556	0.7221
η	-0.0002	0.0026	-0.0717	0.9428
μ	0.4954	0.2162		
σ_u^2	0.8285	0.2810		
σ_v^2	0.3411	0.0174		

* Significant at 99.99% level; ** Significant at 99.9% level; *** Significant at 99% level

once the intensity of usage of each input has a less straightforward interpretation due to the quadratic and interaction terms. In this sense, to properly evaluate the aspects of this technology, we present the calculated elasticities for each input and it's sum (that measures the returns to scale), for each firm in the sample.

Since all variables are logged, the elasticity of the input j in relation to the output can be approximate as its partial derivative. Applying this to the translog production function in Equation 3.1.8 leads to

$$e_{it\,j} = \frac{\partial Y_{it}}{\partial x_{it\,j}} = \beta_j + \gamma_j x_{it\,j} + \theta x_{it\,l}{}^{1} \tag{4.2.1}$$

Table 4.2.2 presents the average elasticities and returns to scale by each market segment. The results clarify the segmentation effect on inefficiency previously mentioned, bringing evidences that those effects could be related to the different technologies those firms are based on. The results for each firm are available in the Appendix A.2.

The general result is alligned with the labor-intense and decreasing returns to scale technology suggested by the SFA-CD model (Table 4.1.1). However, it calls attention how the most efficient segments (i.e, Food and Home Appliances) are the ones with largest returns to scale and labor intense technologies. In this sense, the results suggest that those companies, although being the largest in the sample, still have potentital to achieve higher revenue levels. On the other hand, as the social debate about salaries, labor rights and the so called "6x1 work

[1] For the sake of simplicity, the notation presents only one θ parameter since the model we propose has only one interaction term, between labor and capital; in a more general definition, it should be replaced by a sommatory term, regarding all the interactions between inputs.

Table 4.2.2: Average Elasticities and Returns to Scale by Market Segment

Segment (B3)	Avg. e(Labor)	Avg. e(Capital)	Avg. Returns
Food	1.2191	-0.1551	1.064
Home appliances	1.036	-0.0039	1.0321
Transport materials	0.5297	0.1378	0.6675
Medicines and other products	0.9549	-0.0037	0.9512
Miscellaneous	0.859	-0.004	0.8549
Fabrics. clothing and other products	0.9089	0.0002	0.9091
Total	0.9272	-0.0093	0.9179

scale" (a regime particularly common in retail activities in Brazil, where the worker attend to work 6 days in a row, with 1 day to rest and no scheduled breakes beyond that day-off) gain traction between politicians and the media, those results indicate how the sector efficiently obtains operational results from a low-skilled workforce.

Beyond that, those results don't fully corroborate the findings from de Melo et al. [2018]: the study, focused on supermarket companies, concludes that bigger firms are more efficient, but present decreasing returns to scale. Although the size-effect over efficiency is identified in our study - demanding a more focused and controlled analysis to actively measure this effect - the evidences from the SFA-T model suggest increasing returns to scale in that segment. However, the comparisson must be done cautiously, once the samples are distinct.

Furthermore, the segregated results allow us to identify that the Transport Material segment presents the lowest returns to scale, as well as firms with significantly lower revenues and below-average efficiency measures, defining a segment with particular issues and characteristics that distinguishes from the rest of the sample.

4.3 Bayesian Stochastic Frontier with regression trees

Using the NUTS sampler, available through the PyMC framework, the posterior distribution of the parameters were estimated. The number of trees of the BART model was set to 100. The priors for the σ_u^2 and σ_v^2 were defined as $\beta_{\sigma_u^2} = \beta_{\sigma_v^2} = 1$ for both and $\alpha_{\sigma_u^2} = 3.276$ and $\alpha_{\sigma_v^2} = 3.831$. Those definitons of α suggest prior distributions to σ_u^2 and σ_v^2 where the prior means are equal to the estimated values of those coefficients in the standard SFA-CD approach (Table 4.1.1).

Convergence analysis is implemented to check if the MCMC procedure generates trustful results, as well as posterior predictive checks. The results are shown in the Appendix

A.1.

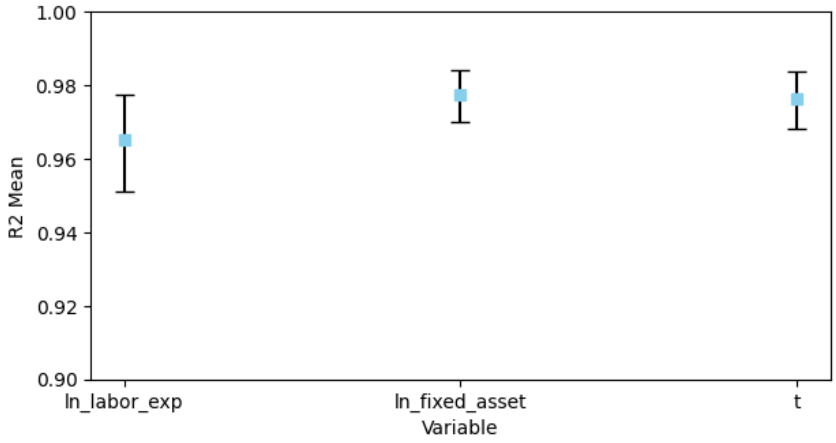
The model shows a better fit to the data, with RMSE equal to 0.4101 and MAPE equal to 2.3053% - taking the posterior distribution's mean of the target variable as point estimates of y, which implies in a quadratic loss function. Much of this better performance can be explained due to the better adjustment of the frontier, once tree-based models are capable of identify and explore complex relationships between the variables, while the previous approaches relies on the assumption of linear parametric frontiers (Codd-Douglas and translog). In this sense, misspecification issues may arise and lead to results that, although are easy to interpret, are not precise. The more flexible modelling of the error components u and v also bring benefits, allowing for distinct behaviors for each observation - as illustrated by Figure A.1.1.

The posterior distribution of the parameter u, which measures the deviations caused by ineffiency, has mean 0.3025, smaller than its counterparts in the SFA-CD and SFA-T models (0.3945 and 0.4954, respectively). The posterior mean of the proportion of error variance caused by inefficiency - the parameter γ - is equal to 0.4755, smaller than both previous results and suggesting that most of the deviations from the frontier are caused by the stochastic error v, and not the stochastic inefficiency u. Those results are not in consonance with the previous estimations and could also be partially explained by the better adjustment of the frontier, reducing the whole error component $\varepsilon_{it} = v_{it} - u_{it}$ and it's dynamics, illustrating how nonlinearities and a more flexible structure can lead to more robust results.

One of the main drawbacks of nonparametric models is the lack of interpretability on the parameters. In a linear regression model, like the standard SFA-CD, the weight of each covariate is easy to visualize and validate its statistical significance. In more complex formulations - as the SFA-T model - the partial derivatives help to address this problem. However, the same could not be done for estimation methods such as BART. However, this framework allows an intuitive way to measure feature importance: considering that, at each iteration, the covariates compete among each other for being chosen to the split rules, the number of times each covariate appears in the final model is a good measure of its importance to predict the values of the dependent variable.

Figure 4.3.1 ranks the three covariates based on their inclusion proportion in the sum-of-trees model, ploting the estimated R^2 evolution as the features are included in the BART estimation. It is clear the relevance of the labor expenditures as the main predictor of revenues, as well as the insignificance of the time-related variable t - in consonance to the previous results.

Figure 4.3.1: R^2 BART estimates with 94% Credibility Intervals

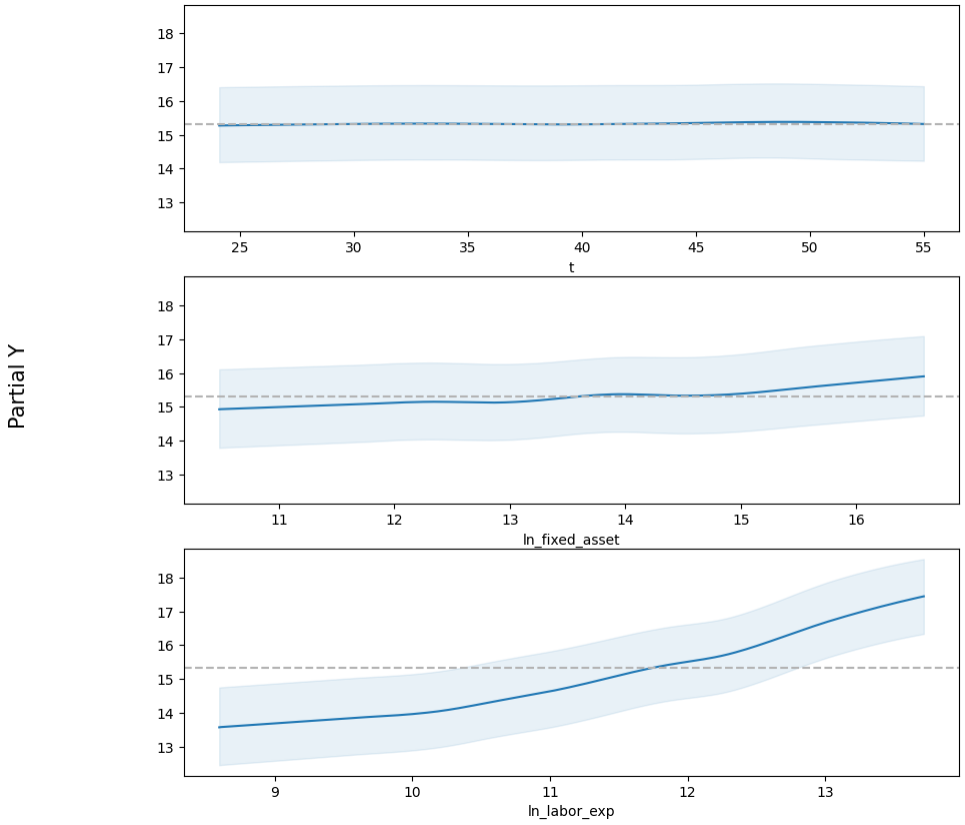


Partial dependence plots in Figure 4.3.2 are used to identify the relationship between the estimated revenues and the inputs. The results are strictly positive, which is also aligned with the basic assumptions in economic theory and the previous results. The relevance of labor expenditure is once again highlighted, representing the largest impact on expected revenues. Moreover, the plot indicates a nonlinear effect of the log of labor expenditures over logged revenues - that also got captured by the SFA-T model, but could not be estimated via SFA-CD. As the standard SFA results, the bayesian semiparametric model does not indicate time-related changes in efficiency. However, partial dependences rely on the assumption of uncorrelated covariates, a possible drawback for several applications - specially on high dimmensional models.

Also focused on the relationship beteween inputs and output, the Forward Marginal Effects (FME) and Nonlinearity Measures (NLM) methods proposed by Scholbeck et al. [2024] were also implemented. These approaches provide a set of novel agnostic strategies to identify and measure the expected impacts on the target given changes in the exogeneous variables - specially for nonlinear models such as machine learning or "black-box" deep learning algorithms. The idea is to estimate new outputs given changes in one or more inputs and compare the results locally or globally (FME), also investigating how those effects dist from linear predictions (NLM). The methods does not assume independence between the covariates and in this implementation we simulate 0.20 increases on both inputs, individually and then combined. Since the variables are logged - as well as in both SFA-CD and SFA-T - the intepretation is similar to elasticities: the average effects presented on Table 4.3.1 represent percentual changes expected in the output given a 20% shift on inputs [2]. For calculating the NLMs, we consider a simpler

[2]The authors suggest a few approaches on selecting the steps for calculating FMEs. In this case, we implement a 0.2 shift because it represents a significant technological/strategical change for companies while keeping the

Figure 4.3.2: Partial Dependence Plots



approach than Scholbeck et al. [2024], comparing the results from the 0.20 shifts estimated through the SFA-BART model versus the linear prediction made considering the SFA-CD estimation.

Table 4.3.1: Average FMEs and NLMs by Market Segment

Segment (B3)	Labor		Capital		Both inputs	
	FME	NLM	FME	NLM	FME	NLM
Food	0.1233	0.9120	0.0180	0.9094	0.1442	0.9111
Home appliances	0.2070	0.9451	0.0295	0.9451	0.2411	0.9446
Transport materials	0.2754	0.9378	0.0415	0.9278	0.3218	0.9384
Medicines and other products	0.1633	0.9336	0.0276	0.9335	0.1939	0.9373
Miscellaneous	0.2034	0.9438	0.0164	0.9366	0.2233	0.9428
Fabrics, clothing and other products	0.1579	0.9353	0.0182	0.9310	0.1798	0.9344
Total	0.1606	0.9371	0.0316	0.9274	0.1861	0.9254

The FME results also indicate the greater relevance of labor factor over expected revenues, but in a smaller degree than the SFA-T model. It also brings evidence that investments on fixed assets, individually, lead to almost irrelevant growth on revenues, while a combined

interpretation of logged deltas as percentual changes (as this approximation fails for large variations)

increase in both inputs lead to more significant shifts. Once again, the benefits of a nonlinear frontier are highlighted, as the FMEs on both inputs are significantly different from the sum of FMEs for each input, capturing how those variables interact in the sample in a more complex way than the previous models could achieve.

In our segmented analysis, the Transport Material firms repeat the best returns to scale performance, with the biggest gains on fixed assets increases. The SFA-BART model did not capture the negative marginal effects of capital investment on revenues that are present in the SFA-T results. As Scholbeck et al. [2024] mention while advocate for the FME approach, marginal effects in nonlinear models (as the translog SFA) may provide unreal results when local effects are assumed as global. Then, the FME method, when fine tuned for relevant steps, can lead to more trustful analysis.

The Food segment, highlighted in the SFA-T results due to the large elasticity on labor factor and increasing returns to scale, do not indicate the same in SFA-BART. The lowest NLM - an indicator that measures how much nonlinearities are captured in the model, with $NLM = 1$ representing perfect linearity - could explain this divergence, as the nonparametric tree-based frontier allows for much more flexible results than the previous methods. Again, the elasticities from SFA-T could indicate a local result, that vanished when nonlinear effects are considered.

Finally, the companies were ranked according to their average efficiency measures, sampled from the posterior distributions of each observation. Although some of the results previously presented diverge from the SFA-CD and SFA-T models, the efficiency measures presented in Table 4.3.2 are alligned with the first evidences, ranking the firms in a similar way and corroborating the existence of a significant segmentation effect over the efficiency of brazillian retail companies.

The main difference on efficiency measures lies on the scale of u: the size of inefficiency impacts over revenues is significantly smaller, given the much more accurate production frontier. This could be an evidence of an overfitted frontier, that fails to isolate the relationship between inputs and output and also captures noise. To verify this hypothesis, pseudo-out-of-sample (OOS) predictions were made, training the same model on half the dataset and measuring out-of-sample predictions on the other half, randomly splitted: the results do not indicate overfitting, generating an OOS-RMSE equal to 0.7577 and OOS-MAPE equal to 4.2577%, naturally above the in-sample error metrics, but still surpassing both SFA-CD and SFA-T models.

Table 4.3.2: Technical Effiency - Semiparametric Bayesian SFA

Company	Avg. Efficiency	Std. Efficiency	Segment (B3)
Allied	0.4155	0.0675	Home appliances
Viveo	0.3931	0.0627	Medicines and other products
Grupo Mateus	0.3528	0.0647	Food
Pague Menos	0.352	0.0584	Medicines and other products
Dimed	0.3415	0.0612	Medicines and other products
Magaz Luiza	0.338	0.05	Home appliances
Wlm Ind Com	0.3268	0.072	Transport materials
Carrefour BR	0.3173	0.0511	Food
Assai	0.305	0.0497	Food
RaiaDrogasil	0.3045	0.0562	Medicines and other products
Minasmaquina	0.3015	0.0579	Transport materials
Casas Bahia	0.2997	0.0506	Home appliances
P.Acucar-Cbd	0.2941	0.0477	Food
Azzas 2154	0.2927	0.0414	Fabrics, clothing and other products
Cea Modas	0.2922	0.059	Fabrics, clothing and other products
Blau	0.2887	0.0575	Medicines and other products
Le Biscuit	0.2855	0.0524	Miscellaneous
Petz	0.2785	0.0423	Miscellaneous
Lojas Marisa	0.2759	0.0498	Fabrics, clothing and other products
Guararapes	0.273	0.0442	Fabrics, clothing and other products
Quero-Quero	0.273	0.0457	Miscellaneous
Grupo Sbf	0.2714	0.0518	Miscellaneous
Lojas Renner	0.2703	0.0515	Fabrics, clothing and other products
Grazziotin	0.2622	0.0523	Fabrics, clothing and other products
Embpar S/A	0.262	0.0762	Transport materials
Hypera	0.2473	0.0384	Medicines and other products
Veste	0.2436	0.0443	Fabrics, clothing and other products
Espacolaser	0.2361	0.035	Miscellaneous

5 Conclusion

With the main objective of analysing the technical efficiency of Brazilian retail listed companies, this research implemented two methods: the first, a well stablished approach, familiar to most of econometricians and that figures as one of the most - if not *the* most - popular econometric model designed for such task; the latter, an extension of the first, bringing flexibility to some of the assumptions that the traditional stochastic frontier models make.

The gains on dealing with such a less restrictive structure got evident through the fit performance, getting a much better adjustment with the Bayesian semiparametric method - specially due to the estimation of a nonparametric frontier. Beyond that, the capability of measuring in a more straighforward way the uncertainty of the results is also a benefit. On the

other hand, the computational cost of choosing such approach must be highlighted, beeing one of the biggest issues that Bayesian inference methods face ever since it's first developments.

The questions that this research aimed to answer got coherent answers from both models - also alligned to expectations one might had from Economic theory and previous studies: a significant segmentation effect is noted among the evaluated firms, as well as decreasing returns to scale, stable efficiency levels throughout the time, and highly labor-intense technologies. The nonparametric Bayesian method also provided evidence about nonlinearities in the relationship between inputs and output, advocating for production frontiers able to capture nonlinear effects (such as the translog frontier), as well as nonparametric frontiers such as the BART estimation we proppose. On the other hand, the better fit obtained with the nonparametric frontier led to smaller inefficiency measures, indicating that the less flexible methods - the SFA-CD and SFA-T models - tend to overestimate the role efficiency plays on determining the performance of the analysed firms. For practioners, we recommend the adoption of strategies to avoid overfitted frontiers, as it could capture noisy elements that do not reflect the true interaction between inputs and outputs.

In this sense, we advocate that although similar results could be obtained with both methods, dealing with less restrictions may lead to more trustful results, gaining empirical adherence and allowing to capture more complex relationships between the variables. As the SFA literature keeps receiving new developments, it is clear that practioners interested in analysing efficiency among firms, countries, or any other decision-making units must be aware of the assumptions made by the wide range of models available and how they could impact the results. Beyond that, the usage of accounting and financial data, nowadays easily obtained for most of publicly traded companies, gets consolidated as the main source of information for operational aspects of such firms.

We also highlight how the interpretability of nonparametric methods is less problematic nowadays then it was previously, as novel tools such as partial dependences and forward marginal effects allow for a straightforward understanding on the estimations. More than good predictors, machine learning algorithms gain space in the literature as tools for causal inference and other econometrics applications. In addition, the uncertainty quantification that Bayesian methods allow is a relevant piece on the framework we applied, providing robust results and estimulating the search for more precise model specifications.

Further analysis on the subject of this research could bring some light on the notori-

ous segment-effect over efficiency, an aspect that other researches do not bring special attention to and could be an idiossincracy of the Brazilian market. Furthermore, the results presented in this paper may suggest size-effects over efficiency, which in other works tend to be negative (bigger firms usually show smaller efficiency measures). Another topic that demands attention is the stability of efficiency measures throughout the time - a result we verify in the three SFA models estimated. As previously mentioned, this could be a result from the absence of technical improvements in the sector, which definitely represents a preoccupant hypothesis that should be explored in further investigations, specifically designed to address the determinants of efficiency among those firms. Finally, to extend the comparison of both models to a wider range of firms, not restricting it to sectors or markets, could lead to new insights on the benefits of less restrictive methods against the advantage of well known parametric approaches, as specific applications might recquire some additional structure (as the monotonic constrained BART models implemented by Wei et al. [2024]).

References

D. Aigner, C. Lovell, and P. Schmidt. Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 6(1):21–37, 1977. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(77)90052-5. URL https://www.sciencedirect.com/science/article/pii/0304407677900525.

A. A. Alshammari, S. M. bin Syed Jaafar Alhabshi, and B. Saiti. The impact of competition on cost efficiency of insurance and takaful sectors: Evidence from gcc markets based on the stochastic frontier analysis. *Research in International Business and Finance*, 47:410–427, 2019. ISSN 0275-5319. doi: https://doi.org/10.1016/j.ribaf.2018.09.003. URL https://www.sciencedirect.com/science/article/pii/S0275531918304100.

L. Alvarez, E. HUAMANÍ, and Y. CORONADO. Heterogeneous effects of e-commerce over technical efficiency in developing countries: Stochastic frontier analysis approach. In *2020 The 4th International Conference on Business and Information Management*, ICBIM 2020, page 101–105, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450387972. doi: 10.1145/3418653.3418678. URL https://doi.org/10.1145/3418653.3418678.

B3. Critério de classificação, 2023. URL https://www.b3.com.br/pt_br/produtos-e-servicos/negociacao/renda-variavel/acoes/consultas/criterio-de-classificacao/.

G. Battese and T. Coelli. Frontier production functions, technical efficiency and panel data: With application to paddy farmers in india. *Journal of Productivity Analysis*, 3(1/2):153–169, 1992. ISSN 0895562X, 15730441. URL http://www.jstor.org/stable/41770578.

L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. Taylor & Francis, 1984. ISBN 9780412048418. URL https://books.google.com.br/books?id=JwQx-WOmSyQC.

J. Breivik, N. M. Larsen, S. B. Thyholdt, and O. Myrland. Measuring inventory turnover efficiency using stochastic frontier analysis: building materials and hardware retail chains in norway. *International Journal of Systems Science: Operations & Logistics*, 10(1):1964635,

2023. doi: 10.1080/23302674.2021.1964635. URL https://doi.org/10.1080/23302674.2021.1964635.

P. Carvalho and R. C. Marques. Estimating size and scope economies in the portuguese water sector using the bayesian stochastic frontier analysis. *Science of The Total Environment*, 544:574–586, 2016. ISSN 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2015.11.169. URL https://www.sciencedirect.com/science/article/pii/S0048969715311505.

H. A. Chipman, E. I. George, and R. E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266 – 298, 2010. doi: 10.1214/09-AOAS285. URL https://doi.org/10.1214/09-AOAS285.

A. d. S. Costa, G. Sâmia da Silva Fôro, and J. de Lima Vieira. Covid-19 e as cadeias de suprimentos:: uma revisão bibliográfica dos principais impactos no brasil. *Revista Vianna Sapiens*, 11(2):28, ago. 2020. doi: 10.31994/rvs.v11i2.687. URL https://viannasapiens.emnuvens.com.br/revista/article/view/687.

W. L. d. M. Cruz. Crescimento do e-commerce no brasil: desenvolvimento, serviços logísticos e o impulso da pandemia de covid-19. *GeoTextos*, 17(1), jul. 2021. doi: 10.9771/geo.v17i1.44572. URL https://periodicos.ufba.br/index.php/geotextos/article/view/44572.

M. D. de Barcellos, K. Basso, and L. B. Espartel. *Food Retailing in Brazil: A General View*, pages 119–139. Springer Fachmedien Wiesbaden, Wiesbaden, 2015. ISBN 978-3-658-09603-8. doi: 10.1007/978-3-658-09603-8_6. URL https://doi.org/10.1007/978-3-658-09603-8_6.

F. L. N. B. de Melo, R. M. B. Sampaio, and L. M. B. Sampaio. Efficiency, productivity gains, and the size of brazilian supermarkets. *International Journal of Production Economics*, 197:99–111, 2018. ISSN 0925-5273. doi: https://doi.org/10.1016/j.ijpe.2017.12.016. URL https://www.sciencedirect.com/science/article/pii/S092552731730422X.

O. Delardas, K. S. Kechagias, P. N. Pontikos, and P. Giannos. Socio-economic impacts and challenges of the coronavirus pandemic (covid-19): an updated review. *Sustainability*, 14 (15):9699, 2022.

M. A. Diaz and R. Sanchez. Firm size and productivity in spain: A stochastic frontier analysis. *Small Business Economics*, 30(3):315–323, Jul 2007. doi: 10.1007/s11187-007-9058-x.

P. M. Dunne, R. F. Lusch, and J. R. Carver. *Retailing*. South-Western Cengage Learning, 7 edition, 2011.

E. M. Farina, R. Nunes, and G. F. d. A. Monteiro. Supermarkets and their impacts on the agrifood system of brazil: The competition among retailers. *Agribusiness*, 21(2):133–147, 2005. doi: https://doi.org/10.1002/agr.20039. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/agr.20039.

P. Fenn, D. Vencappa, S. Diacon, P. Klumpes, and C. O'Brien. Market structure and the efficiency of european insurance companies: A stochastic frontier analysis. *Journal of Banking Finance*, 32(1):86–100, 2008. ISSN 0378-4266. doi: https://doi.org/10.1016/j.jbankfin.2007.09.005. URL https://www.sciencedirect.com/science/article/pii/S0378426607002610. Dynamics of Insurance Markets: Structure, Conduct, and Performance in the 21st Century.

A. S. Gupta and J. Mukherjee. Long-term changes in consumers' shopping behavior post-pandemic: an exploratory study. *International Journal of Retail & Distribution Management*, 50(12):1518–1534, 2022.

C. Hevia, P. A. Neumeyer, et al. A perfect storm: Covid-19 in emerging economies. *COVID-19 in developing economies*, 1(1):25–37, 2020.

IBGE. volume 33. 2021.

IPEA. Boletim mercado de trabalho : conjuntura e análise : n. 78, 2024.

A. Kapelner and J. Bleich. bartmachine: Machine learning with bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40, 2016. doi: 10.18637/jss.v070.i04. URL https://www.jstatsoft.org/index.php/jss/article/view/v070i04.

G. Koop and M. F. Steel. *Bayesian Analysis of Stochastic Frontier Models*, chapter 24, pages 520–537. John Wiley Sons, Ltd, 2003. ISBN 9780470996249. doi: https://doi.org/10.1002/9780470996249.ch25. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470996249.ch25.

G. Koop, J. Osiewalski, and M. Steel. The components of output growth: A stochastic frontier analysis. *Oxford Bulletin of Economics and Statistics*, 61:455–487, 1999. doi: 10.1111/1468-0084.00139.

S. C. Kumbhakar and C. A. K. Lovell. *Stochastic Frontier Analysis*. Cambridge University Press, 2000.

L.-F. Lee and W. G. Tyler. The stochastic frontier production function and average efficiency. *Journal of Econometrics*, 7(3):385–389, Apr 1978. doi: 10.1016/0304-4076(78)90061-1.

D. Leite, J. Pessanha, P. Simões, R. Calili, and R. Souza. A stochastic frontier model for definition of non-technical loss targets. *Energies*, 13(12), 2020. ISSN 1996-1073. doi: 10.3390/en13123227. URL https://www.mdpi.com/1996-1073/13/12/3227.

C. F. Parmeter and S. C. Kumbhakar. *Recent Advances in the Construction of Nonparametric Stochastic Frontier Models*, pages 165–181. Springer International Publishing, Cham, 2023. ISBN 978-3-031-29583-6. doi: 10.1007/978-3-031-29583-6_10. URL https://doi.org/10.1007/978-3-031-29583-6_10.

F. Pavelescu. Some aspects of the translog production function estimation. *Romanian Journal of Economics*, 32(1(41)):131–150, 2011. URL https://EconPapers.repec.org/RePEc:ine:journl:v:1:y:2011:i:41:p:131-150.

M. R. Pinto, P. K. Salume, M. W. Barbosa, and P. R. de Sousa. The path to digital maturity: A cluster analysis of the retail industry in an emerging economy. *Technology in Society*, 72:102191, 2023. ISSN 0160-791X. doi: https://doi.org/10.1016/j.techsoc.2022.102191. URL https://www.sciencedirect.com/science/article/pii/S0160791X22003323.

M. Robaina-Alves, V. Moutinho, and P. Macedo. A new frontier approach to model the eco-efficiency in european countries. *Journal of Cleaner Production*, 103:562–573, 2015. ISSN 0959-6526. doi: https://doi.org/10.1016/j.jclepro.2015.01.038. URL https://www.sciencedirect.com/science/article/pii/S0959652615000426. Carbon Emissions Reduction: Policies, Technologies, Monitoring, Assessment and Modeling.

A. Santos and C. Costa. Características gerais do varejo no brasil. *BNDES Setorial*, 5:55–69, Mar 1997. URL https://web.bndes.gov.br/bib/jspui/bitstream/1408/

7125/2/BS%2005%20Caracteristicas%20gerais%20do%20varejo%20no%20Brasil_P.pdf.

A. M. Schmidt, A. R. Moreira, S. M. Helfand, and T. C. Fonseca. Spatial stochastic frontier models: Accounting for unobserved local determinants of inefficiency. *Journal of Productivity Analysis*, 31(2):101–112, Dec 2008. doi: 10.1007/s11123-008-0122-6.

I. N. Schneider, D. K. Baggio, J. S. T. D. Silveira, and M. M. B. Brizolla. Assessing market timing performance of brazilian multi-asset pension funds using the battese and coelli's stochastic frontier model (1995). *Economics Bulletin*, 40:50–60, 2020. URL https://api.semanticscholar.org/CorpusID:215413071.

C. A. Scholbeck, G. Casalicchio, C. Molnar, B. Bischl, and C. Heumann. Marginal effects for non-linear prediction functions. *Data Mining and Knowledge Discovery*, 38(5):2997–3042, Sep 2024. ISSN 1573-756X. doi: 10.1007/s10618-023-00993-x. URL https://doi.org/10.1007/s10618-023-00993-x.

L. Simar and P. W. Wilson. Nonparametric, stochastic frontier models with multiple inputs and outputs. *Journal of Business amp; Economic Statistics*, 41(4):1391–1403, Oct 2022. doi: 10.1080/07350015.2022.2110882.

T. G. Taylor and J. S. Shonkwiler. Alternative stochastic specifications of the frontier production function in the analysis of agricultural credit programs and technical efficiency. *Journal of Development Economics*, 21(1):149–160, Apr 1986. doi: 10.1016/0304-3878(86)90044-1.

B. Tovar, F. Javier Ramos-Real, and E. F. de Almeida. Firm size and productivity. evidence from the electricity distribution industry in brazil. *Energy Policy*, 39(2):826–833, Feb 2011. doi: 10.1016/j.enpol.2010.11.001.

E. Tsionas. Combining dea and stochastic frontier models: An empirical bayes approach. *European Journal of Operational Research*, 147:499–510, 06 2003. doi: 10.1016/S0377-2217(02)00248-5.

E. Tsionas. An introduction to efficiency measurement using bayesian stochastic frontier models. *Global Business and Economics Review*, 3:287–311, 02 2005. doi: 10.1504/GBER.2001.006177.

E. G. Tsionas, K. C. Tran, and P. G. Michaelides. Bayesian inference in threshold stochastic frontier models. *Empirical Economics*, 56(2):399–422, Feb 2019. ISSN 1435-8921. doi: 10.1007/s00181-017-1364-9. URL https://doi.org/10.1007/s00181-017-1364-9.

M. Tsionas. Efficiency estimation using probabilistic regression trees with an application to chilean manufacturing industries. *International Journal of Production Economics*, 249: 108492, Jul 2022. doi: 10.1016/j.ijpe.2022.108492.

G. J. D. Vries. Small retailers in brazil: Are formal firms really more productive? *The Journal of Development Studies*, 46(8):1345–1366, 2010. doi: 10.1080/00220380903147668. URL https://doi.org/10.1080/00220380903147668.

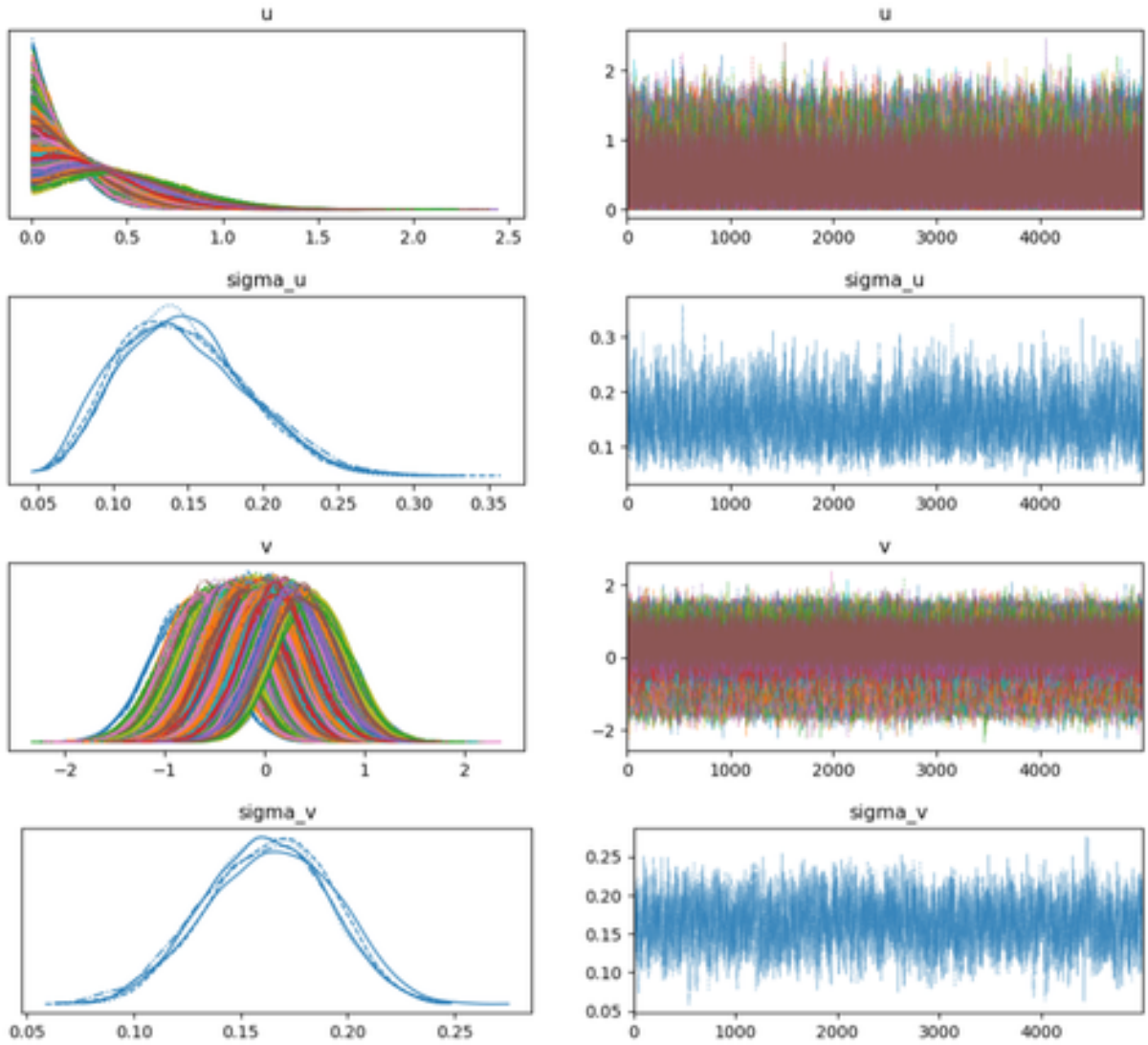
Z. Wei, H. Sang, and N. Coulibaly. Nonparametric machine learning for stochastic frontier analysis: A bayesian additive regression tree approach. *Econometrics and Statistics*, 2024. ISSN 2452-3062. doi: https://doi.org/10.1016/j.ecosta.2024.06.002. URL https://www.sciencedirect.com/science/article/pii/S2452306224000388.

A Appendix

A.1 Convergence analysis

The model was estimated with 5.000 draws, after 2.000 burn-in iterations that are not considered in the posterior analysis, in 5 parallel chains. The Geweke's diagnostic, calculated for each chain on the four parameters that are not part of the BART model (u, σ_u, v, σ_v) generates scores that indicate convergence. The trace plots in Figure A.1.1 illustrate the diagnosis.

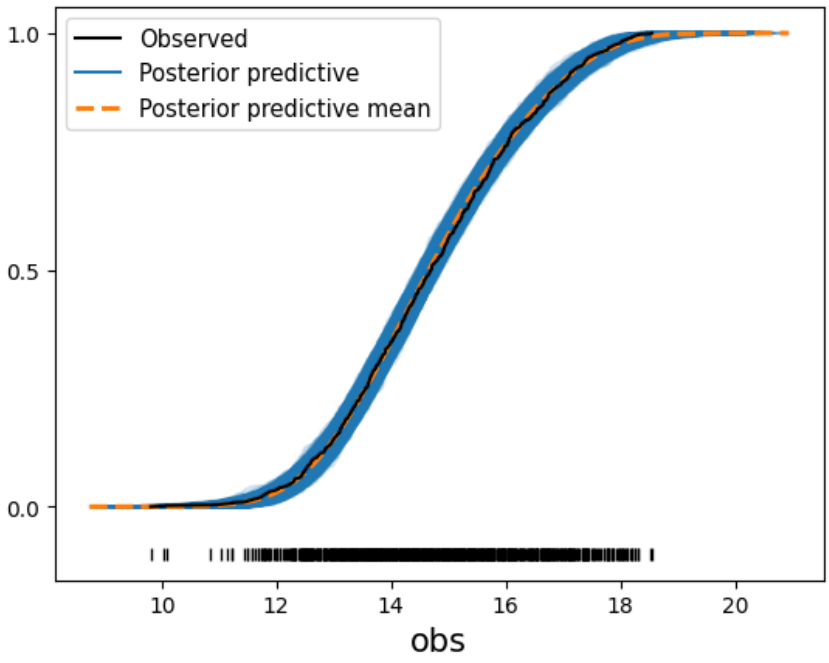
Figure A.1.1: Convergence Analysis



Each colored line in u and v represents the MCMC samples for the 796 observations. The visual inspection makes evident that the distributions of those inefficiency and error terms vary across observations, achiveing convergence for all of them.

Posterior predictive checks were also implemented to visualize how well the model describes the observed distribution of the target variable - the natural logarithm of net revenues. The results are shown in Figure A.1.2, evaluating the cumulative distribution observed in the dataset (black), the posterior samples (blue) and the posterior mean (dashed yellow). The plot highlights the good fit performance.

Figure A.1.2: Posterior Predictive Check



A.2 Translog Frontier Results

Table A.2.1 presents the average elasticities and returns to scale by firm. Since there were no evidence of time related changes on efficiency, the averages should properly describe the firms technology throughout the whole sample.

Table A.2.2 ranks the firms based on the efficiency measures estimated through the SFA-T model. The results don't indicate any significant change from the SFA-CD model, in Table 4.1.2.

Table A.2.1: Average Elasticities and Returns to Scale by Firm

Company	e(Labor)	e(Capital)	Returns to Scale (e(Labor) + e(Capital))
Allied	0.7968	0.1281	0.9249
Azzas 2154	0.8441	0.0638	0.9079
Casas Bahia	1.1896	-0.0775	1.1121
Cea Modas	1.0188	-0.0662	0.9526
Embpar S/A	0.4984	0.2151	0.7135
Espacolaser	0.7903	0.0529	0.8432
Grazziotin	0.5861	0.0303	0.6164
Grupo Sbf	1.0053	-0.0446	0.9607
Guararapes	1.0858	-0.0725	1.0133
Le Biscuit	0.7067	-0.0062	0.7005
Lojas Marisa	0.8793	0.0957	0.975
Lojas Renner	1.1129	-0.0964	1.0165
Magaz Luiza	1.1217	-0.0624	1.0593
Minasmaquina	0.4497	0.1789	0.6286
Petz	0.9159	-0.0575	0.8584
Quero-Quero	0.8766	0.0353	0.9119
Veste	0.8355	0.0464	0.8819
Wlm Ind Com	0.641	0.0193	0.6603
Assai	1.2292	-0.1758	1.0534
Blau	0.7491	0.0419	0.791
Carrefour BR	1.3022	-0.1911	1.1111
Dimed	0.8406	0.0093	0.8499
Grupo Mateus	1.0855	-0.0819	1.0036
Hypera	1.0666	-0.0423	1.0243
P.Acucar-Cbd	1.2593	-0.1716	1.0877
Pague Menos	1.0179	0.0003	1.0182
RaiaDrogasil	1.1496	-0.0875	1.0621
Viveo	0.9057	0.0561	0.9618
Total	0.9189	-0.0055	0.9134

Table A.2.2: Technical Effiency Ranking - SFA-T

Company	**Avg. Technical Effic.**	**Segment (B3)**	**Market Share.**[1]
Allied	Home appliances	0.9416	1.17%
Carrefour BR	Food	0.899	23.86%
Viveo	Medicines and other products	0.8787	2.45%
Assai	Food	0.8562	15.18%
Grupo Mateus	Food	0.7956	6.61%
Magaz Luiza	Home appliances	0.7772	7.71%
P.Acucar-Cbd	Food	0.7133	3.84%
Pague Menos	Medicines and other products	0.6422	2.63%
Minasmaquina	Transport materials	0.6315	0.28%
Dimed	Medicines and other products	0.6153	1.02%
RaiaDrogasil	Medicines and other products	0.6147	8.14%
Wlm Ind Com	Transport materials	0.5861	0.67%
Casas Bahia	Home appliances	0.5718	5.44%
Azzas 2154	Fabrics. clothing and other products	0.3984	1.41%
Petz	Miscellaneous	0.3877	0.69%
Le Biscuit	Miscellaneous	0.3843	0.44%
Cea Modas	Fabrics. clothing and other products	0.3836	1.44%
Lojas Renner	Fabrics. clothing and other products	0.3636	2.78%
Blau	Medicines and other products	0.359	0.37%
Lojas Marisa	Fabrics. clothing and other products	0.3442	0.26%
Grazziotin	Fabrics. clothing and other products	0.3292	0.14%
Embpar S/A	Transport materials	0.325	0.01%
Guararapes	Fabrics. clothing and other products	0.3242	1.87%
Quero-Quero	Miscellaneous	0.3076	0.55%
Grupo Sbf	Miscellaneous	0.3016	1.41%
Hypera	Medicines and other products	0.2355	1.68%
Veste	Fabrics. clothing and other products	0.2038	0.23%
Espacolaser	Miscellaneous	0.1928	0.22%

[1] Note: Measured as the company's share in the sum of revenues of the last available quarter (2024q3).