Efficiency estimation using probabilistic regression trees with an application to Chilean manufacturing industries

January 28, 2022

Abstract

We propose smooth monotone concave probabilistic regression trees for the estimation of efficiency and productivity. In particular we modify these techniques to allow for the use of panel data which are often encountered in practice. Probabilistic regression trees provide smooth approximations and at the same time they exploit the versatility of standard regression trees in generating efficiently partitions of the space of the regressors to approximate the unknown frontier. We showcase the new techniques in a large sample of Chilean manufacturing firms.

Key Words: Efficiency; Productivity; Regression Trees; Probabilistic Regression Trees.

Acknowledgments: The author is indebted to two anonymous reviewers for their comments on an earlier version.

1 Introduction

The approximation of production functions by monotone concave yet flexible functional forms is a problem that has received considerable attention in the literature. Despite the advances that have been made, we would prefer to use Regression Trees (RTs) that have been proven to be reliable and fast approximations to arbitrary functional forms. Unfortunately, RTs are neither differentiable nor non-decreasing by construction so they do not satisfy concavity properties either. The problem can be solved if we (i) rely on probabilistic trees as we do in this paper, and (ii) adopt flexible functional forms that are monotone concave instead of the step functions used in RTs or random forests and other ensembles. To the best of our knowledge, using monotone concave instead of step functions in RTs is novel. Additionally, there are other problems that might be encountered in practice. *First*, RTs are not straightforward to apply with panel data. *Second*, the incorporation of technical inefficiency in RTs does not appear to be straightforward.We show that in probabilistic RTs this can be overcome. Previous attempts include Esteve et al. (2020) whose method shares some similarities with the Free Disposal Hull technique but, in contrast it overcomes the problem of overfitting by using cross-validation. Another approach, without inefficiency however, is Blanquero et al. (2020).

Emrouznejad and Anouze (2010) use regression trees to understand the factors determining efficiency by applying RTs in a second stage analysis, see also Rebai et al (2019). Esteve et al. (2020) construct a RT approach that is closest to the spirit of constructing an approximation of the production set. This RT is non-smooth, a fact that is not necessarily a constraint but it may, sometimes be more instructive to have smooth approximations. As Esteva et al. (2020) correctly argue, their approach " could be interpreted as a 'pruned' FDH [Free Disposal Hull] or a FDH-type out-of-sample predictor, overcoming its problem of data overfitting if the aim is to estimate the true theoretical frontier" (Esteva et al., 2020, p. 16). In this paper, we are interested in the good approximation properties of RTs but we would like to impose monotonicity and concavity which is critical in many applications (e.g. Kuosmanen, 2008; Kuosmanen and Johnson, 2010; Lee et al., 2013, 2010, 2019). Incorporating productivity into the model is non-trivial in RTs; in fact, incorporating inefficiency is not straightforward either given the state of the art. Our approach to the design of monotone concave approximations using RTs is to use probabilistic node splitting (which guarantees smoothness of the expected value of the response variable) along with Cobb-Douglas production functions at the nodes instead of fitting constants as in traditional RT analysis. Fitting a constant may be a poor approximation and a Cobb-Douglas is much better as, locally, it can approximate to first order any function of interest. Globally, by fitting different Cobb-Douglas production functions at different nodes, does not compromise concavity and monotonicity and yields a flexible approximation (see, for example, Geweke and Petrella, 2014; Geweke and Kean, 2007, and Norets, 2010). Such approximations are of wider interest in logistics management and engineering (Chen et al., 2021; Loyer et al., 2016), education (Masci et al., 2018), etc. Existing methods closely related to this paper can be described as follows. Yagi, Chen, Johnson, and Kuosmanen (2020) use the StoNED approach but modify it using a kernel function in the objective function, in the spirit of the local linear kernel estimator previously

examined in the frontier context by Kumbhakar, Park, Simar, and Tsionas (2007). Their approach builds on earlier work (e.g., Kuosmanen and Kortelainen, 2012) and it is called Shape Constrained Kernel-weighted Least Squares (SCKLS). Valero-Carreras, Aparicio, and Guerrero (2021) propose support vector frontiers based on the concept of support vector regression. Interestingly, they also show that standard FDH and DEA could be reinterpreted as support vector regression techniques. Probabilistic RTs, introduced in this paper, are a different non-parametric technique which approaches the problem of frontier estimation from a different angle. Relative to Yagi et al. (2020) for example, we do not need to introduce the monotonicity and concavity restrictions explicitly. Relative to Valero-Carreras, Aparicio, and Guerrero (2021) the incorporation of restrictions is also easier as Valero-Carreras, Aparicio, and Guerrero (2021) introduced a specific transformation function of the input space to allow determining monotonic non-decreasing step functions as estimator of the production functions; in a second stage, by convexification, they were able to yield concave predictors, which are directly linked to convex production possibility sets and Data Envelopment Analysis (DEA). This resulted in the introduction of the so-called Convexificated Support Vector Frontiers (CSVF). The primary virtue of nonparametric methods is that they provide a principled way to estimate marginal effects (partial derivatives or elasticities); see, for example, Coglianese et al. (2017), Fisher et al. (2017), Schulte (2015), and Chernozhukov et al. (2018), who propose a systematic way to present heterogeneous effects (as, more often than not, only means are reported). Closely related techniques to RTs from a Bayesian approach include the CART and BART techniques, which are implementations of the RT idea (Chipman et al., 1998, 2010; Denison et al., 2008). In particular, Bayesian variants of these methods like Bayesian CART and BART are quite prominent in machine learning applications. "BART is able to detect interactions and nonlinearities in the response surface, which (among other advantages) allows it to more readily identify heterogeneous treatment effects" (Hill, 2011, p. 218). Additionally, "BART also has advantages compared to alternative nonparametric or semiparametric methods that might be used to flexibly model the assignment mechanism and the response surface. BART can handle a large number of both continuous and discrete predictors. Moreover BART overcomes a standard barrier to widespread implementation of new methodology because it requires far less researcher involvement, technical sophistication, and investment of time. The method is accessible to applied researchers who may not have a strong mathematical background and will not require days or weeks of programming to implement (particularly important given that it is difficult to know when a more sophisticated method will actually make a difference in practice)" (Hill, 2011, p. 237). A related technique is Multiple Adaptive Regression Splines (MARS; Friedman, 1991, Friedman et al., 2000) which also uses recursive partitioning and produces a continuous regression function estimate although its results are hard to interpret. Usually, simple RTs, that may provide poor fits are combined through a procedure known as boosting (Freund & Schapire, 1997; Friedman, 2001, 2002; and for a survey, see Bühlmann & Hothorn, 2007). An additional advantage is that RTs allow the modeling of complex nonlinearities, they are insensitive to the inclusion of irrelevant variables, and they are robust to outliers.

RTs (as well as CART and BART) *provide step function approximations* and, therefore, they are not smooth, a

restriction that may be important in practice (e.g., Yagi et al., 2020; Valero-Carreras et al., 2021). In this paper, we propose Probabilistic RTs (abbreviated as PRTs) that avoid the problem without compromising the flexibility properties of RTs per se, and imposing global monotonicity and curvature. Unlike other approximations, PRTs are more like mixtures of regressions (which are semi-parametric approximations) but differ in the selection of adaptive partitioning of the space of regressors. Relative to Yagi et al. (2020) we notice that inefficiency is absent from their model, although other approaches like StoNED can be used to obtain efficiency. Moreover, both Yagi et al. (2020) and Valero-Carreras et al. (2021) do not focus on the problem of panel data, a problem that is often ignored even in the RTs literature. Additionally, relative to standard RTs, our formulation, (i) allows for panel data, (ii) it introduces technical inefficiency, and (iii) allows for smoothness as we use a probabilistic RT. Although it is possible to modify CART and BART for panel data (e.g., Fu and Simonoff, 2015; Segal, 1992; Zhang, 1998; De'Ath, 2002; Hajjem et al., 2011; Sela and Simonoff, 2012), it is not obvious how these approaches can yield a smooth approximation and how they can allow for the presence of technical (in)efficiency. Therefore, our main contributions are smooth RTs for panel data (so, we do not follows preciously suggested techniques) and RTs with technical (in)efficiency. Both features essential in Production Economics (Gunasekaran and Kobu, 2007).

2 Regression trees

Regression Trees (RTs) are part of non-parametric regression methods (Murthy, 1998; Hastie et al., 2001) that work in a fast and computationally efficient way to approximate functional forms like

$$y_i = f(x_i) + e_i,\ i = 1, \ldots, n, \tag{1}$$

where y_i is the dependent variable, $x_i \in \mathcal{X} \subseteq \mathbb{R}^K$ is a vector predictor, e_i is an error term, and $f(\cdot)$ represents an unknown functional form to be approximated; see Breiman et al. (1984). The approximation has the general form

$$f(x_i) = \sum_{g=1}^{G} \gamma_g \mathbb{I}(x_i; \theta_g), \tag{2}$$

where

$$\mathbb{I}(x_i; \theta_g) = \begin{cases} 1, \text{ if } x_t \in \mathcal{R}_g(\theta_g), \\ 0, \text{ otherwise,} \end{cases} \qquad g = 1, \ldots, G, \tag{3}$$

where $\mathcal{R}_g(\theta_g)$ denotes a subregion of $\mathcal{X}$ (viz. a hyperplane that is orthogonal to the axis of the predictor variables), $\gamma_g$and θ_g are parameters, and G is the total number of subregions. Essentially, the relationship between y_t and x_t in (1) is approximated by a linear regression on a set of G dummy variables. Despite the simplicity of the formulation, RTs do not require much tuning and they are computationally fast even in large data sets. A simple tree structure

with $G = 2$ leaves has the following form:

$$y_i = \beta_1 \mathbb{I}(x_i; s_0, c_0) + \beta_2 \left[1 - \mathbb{I}(x_i; s_0, c_0)\right] + e_i, \tag{4}$$

where

$$\mathbb{I}(x_i; s_0, c_0) = \begin{cases} 1, \text{ if } x_{si} \leq c_0, \\ \\ 0, \text{ otherwise,} \end{cases} \tag{5}$$

where $s \in \mathbb{S} = \{1, \ldots, K\}$ is an index for a variable and, of course, $x_{si} \in x_i$. One may write (2) alternatively as

$$f(x_i) = \sum_{g=1}^{G} \gamma_g \mathbb{I}(x_i \in \mathcal{R}_g), \tag{6}$$

where $\mathcal{R}_g$ is a subregion of the predictor space, $\mathcal{X}$. Therefore, a RT fits a constant to each particular subregion. Additionally, we can write a RT as follows.

$$f(x_i) = \sum_{g=1}^{G} \mathfrak{g}(x_i; \mathscr{T}_g, \mathscr{M}_g), \tag{7}$$

where $\mathfrak{g}(x_i; \mathscr{T}_g, \mathscr{M}_g)$ is a function that assigns a predicted value based on x_i, $\mathscr{T}_g$ is a set of splitting rules that defines the gth tree, and $\mathscr{M}_g$ contains the predicted values for all nodes in tree g. The splitting rules that determine the terminal nodes for the tree g, are partitions $\mathscr{P}_{gl}$ with $g(x_i; \mathscr{T}_g, \mathscr{M}_g) = \mu_{gl}$, a certain constant.

3 Panel data and regression trees

Techniques that deal with RTs in the context of longitudinal (panel) data include Fu and Simonoff (2015), Segal (1992), Zhang (1998), De'Ath (2002), Hajjem et al. (2011), Sela and Simonoff (2012). As shown in Figure 1, ignoring the panel data structure can have detrimental consequences for RTs. Therefore, it is clear that we need a systematic procedure to deal with the complexities arising from panel data. From the point of view of applied studies, a widely used panel data model is

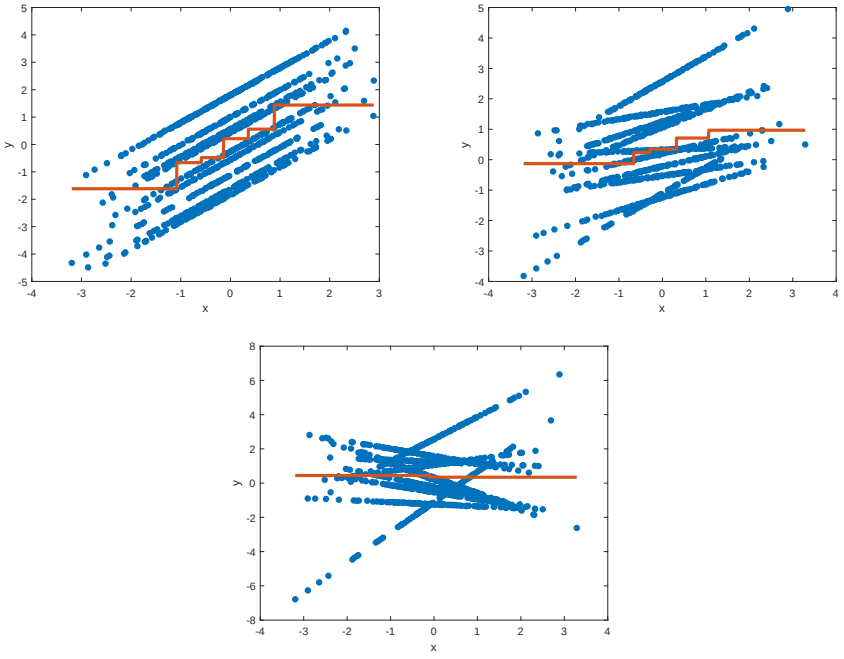
$$y_{it} = \alpha_i + x_{it}'\beta + e_{it},\ i = 1, \ldots, n,\ t = 1, \ldots, T, \tag{8}$$

where α_is represent individual effects. The model assumes common slope coefficients, although this can be taken into account using the more general model

$$y_{it} = \eta_i + x_{it}'\boldsymbol{\beta}_i + e_{it},\ i = 1, \ldots, n,\ t = 1, \ldots, T, \tag{9}$$

Figure 1: Panel data

With panel data, RTs may not perform well. In this experiment we have $y_{it} = \alpha_i + \beta_i x_{it} + e_{it}$, $e_{it} \sim \mathcal{N}(0, 0.01^2)$,$i = 1, \ldots, n = 100$ and $t = 1, \ldots, T$ (where $T = 10$). The slopes and intercepts are generated from standard normal distributions and the same is the case for each regressor, x_{it}. In the middle panel we have panel data with common slopes and different intercepts. Both intercepts and slopes are heterogeneous. Slopes are generated from a standard uniform distribution. In the right panel we have different intercepts but slopes are generated from a standard normal distribution.



where $\boldsymbol{\beta}_i$s are vector of firm-specific slope coefficients. The e_{it}s denote, generically, error terms. When our panel consists of a relatively few units with a large number of temporal observations, instead of (8) or (9) one can fit separate RTs to each unit. Such cases are, however, relatively uncommon, although they may occur some times, e.g. when clusters can be determined or defined in advance. So, when n is large and T is moderate, the alternative procedure is to rely on the residuals, $\hat{e}_{it}$, from panel data models as in (8) or (9) or, perhaps, more complicated structures like Dynamic Panel Data (DPD) models.

To impose monotonicity and concavity in stochastic frontiers, we should allow for functional forms that are provably flexible (Geweke and Keane, 2007; Norets 2010) but *frontiers should also be smooth*. Yagi et al. (2020) for example, build on StoNED to provide monotonically concave frontiers in the spirit of local estimation at a pre-selected grid of points. Valero-Carreras et al. (2021) use a two-step procedure to obtain such frontiers using the concept of support vector regression.

Define, for a vector $z \in \mathbb{R}^K$

$$\Psi(z; \mathcal{R}_g, \sigma) = \frac{1}{\prod_{k=1}^{K} \sigma_k} \cdot \int_{\mathcal{R}_g} \prod_{k=1}^{K} \phi\left(\frac{\zeta_k - z_k}{\sigma_k}\right) \mathrm{d}\boldsymbol{\zeta}, \tag{10}$$

where $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_K]'$ is a vector of scale parameters, $\boldsymbol{\zeta} \equiv [\zeta_1, \ldots, \zeta_K]'$, $\phi(\cdot)$ is any univariate density function (for example, the standard normal, viz. $\phi(u) = (2\pi)^{-1/2} e^{-u^2/2}$, $u \in \mathbb{R}$) and (abusing notation slightly) $\mathcal{R}_g$ denotes a partition of the space $\mathcal{Z}$ of z_{it}s. The z_{it}s are variables that are used in the partitioning instead of the x_{it}s, and they could be particular points in the space of the regressors as in Yagi et al. (2020) or their lagged values ($x_{i,t-1}$). In this paper, we generate randomly (using a uniform distribution) points in the support of the x_{it}s; their number is set to 25% of the original sample size.[1]

Suppose our data is $D = \{y_{it}, x_{it}\}_{1 \leq i \leq n, 1 \leq t \leq T}$. If we were to fit a RT like (2), the problem becomes

$$\min_{\Theta} \sum_{i=1}^{n} \sum_{t=1}^{T} \left(y_{it} - \eta_i - \sum_{g=1}^{G} \gamma_g P_{it,g} \right)^2, \tag{11}$$

where η_i is a firm effect, $\Theta = \left[\{\mathcal{R}_g\}_{1 \leq g \leq G}, \boldsymbol{\beta}, \boldsymbol{\sigma} \right]$, and $P_{it,g}$ encodes the relation between z_{it} and $\mathcal{R}_g$ so that $P_{it,g} = \Psi\left(z_{it}; \mathcal{R}_g, \boldsymbol{\sigma}\right) \geq 0$ and $\sum_{g=1}^{G} P_{it,g} = 1$ (for all $i = 1, \ldots, n$). If we fit a monotone concave function in each region, then we have[2]

$$\min_{\Theta} \sum_{i=1}^{n} \sum_{t=1}^{T} \left(y_{it} - \sum_{g=1}^{G} \left\{ \eta_{i,(g)} + \varphi_g(x_{it}; \boldsymbol{\beta}_{(g)}) \right\} P_{it,g} \right)^2, \tag{12}$$

[1] We could use the x_{it}s themselves in imposing monotonicity and curvature globally. To the extent that x_{it} and $x_{i,t-1}$ are, usually, highly correlated, this is not a restrictive assumption and facilitates the imposition of global properties of the functional form.

[2] The functions $\varphi_g(x_{it}; \boldsymbol{\beta}_{(g)})$ can be Cobb-Douglas for simplicity as monotonicity and concavity can be easily imposed. We follow this practice below, and we assume that all inputs and output are in log form so that $\varphi_g(x_{it}; \boldsymbol{\beta}_{(g)})$ becomes, effectively, a linear function.

where[3]

$$\Theta = \left[\{\mathcal{R}_g\}_{1\leq g\leq G}, \boldsymbol{\beta}, \boldsymbol{\sigma}\right], \tag{13}$$

$\boldsymbol{\beta} = \left[\boldsymbol{\beta}'_{(g)}\right]_{1\leq g\leq G}$, and we allow individual effects $\eta_{i,(g)}$ to depend on the particular subregion. Despite the superficial similarity of (12) with kernel-based methods in the spirit of, say, local estimation, the kernel function is part of the regression problem and, it is not used to weight squared residuals from the regression. *The distinction is fundamental primarily because it allows smooth regression trees.*

One important feature of (11) which is not present in local estimation procedures, SCKLS or CSVF, is the following. If we define $\tilde{y} = [y_{it} - \eta_i; i = 1, ..., n, t = 1, ..., T]$, we set $\varphi_g(x_{it}; \boldsymbol{\beta}_{(g)}) = m_g$, i.e., a local constant, $m = [m_g, g = 1, ..., G]$, and $\mathbb{P} = [P_{it,g}]$, we have the important property that the local constants can be updated by least squares as follows:

$$\hat{m} = (\mathbb{P}'\mathbb{P})^{-1}\mathbb{P}'\tilde{y}, \tag{14}$$

provided matrix $\mathbb{P}$ has full rank. [4]

Conditional on G, $\boldsymbol{\sigma}$ and $\{\mathcal{R}_g\}_{1\leq g\leq G}$ (and, therefore, given $\left\{P_{it,(g)}\right\}$), it is not difficult to update the parameters $\boldsymbol{\eta} = \left\{\eta_{(i,g)}\right\}_{1\leq i\leq n; 1\leq g\leq G}$ and $\boldsymbol{\beta} = \left[\beta_{(g)}\right]$ through MCMC (see Technical Appendix A). From this representation, we see that PRTs resemble mixtures of regressions rather than traditional non-parametric procedures (Norets, 2010). However, although mixtures themselves are flexible in the semi-parametric sense, they are not fully non-parametric. Therefore, we generate the partitions of the space of regressors in a way that is more faithful to the RT approach. Based on the new representation we see that *we can introduce technical inefficiency* ($u_{it,(g)} \geq 0$) as follows.

$$\min_{\Theta} \sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{it} - \sum_{g=1}^{G}\left[\eta_{i,(g)} + \varphi_g(x_{it}; \boldsymbol{\beta}_{(g)}) - u_{it,(g)}\right]P_{it,g}\right)^2, \tag{15}$$

Relative to standard RTs, this formulation, (i) allows for panel data, (ii) it introduces technical inefficiency, and (iii) allows for smoothness as we use a probabilistic RT. In a sampling-theory context, the $u_{it,(g)}$s can be treated as parameters and estimated using a penalty function like, for example, the LASSO or the elastic net. A similar approach can be followed in a Bayesian context as follows. Define the "odds ratio" $\psi_{it,(g)} \equiv \frac{r_{it,(g)}}{1-r_{it,(g)}}$. In turn, we assume that the "odds ratios" are assumed to have a half-Laplace distribution with parameter λ which is equivalent to the LASSO:

$$p(\psi) \propto e^{-\lambda|\psi|}, \tag{16}$$

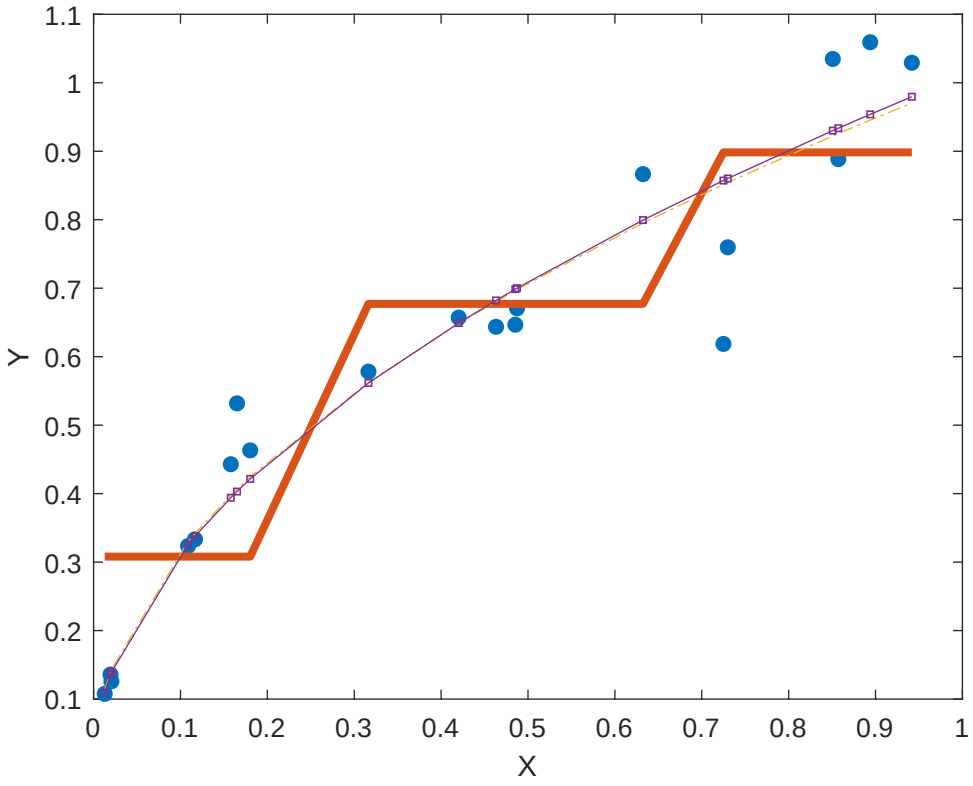
where λ is a parameter related to the usual penalty in LASSO optimization. Although efficiency is defined only in (0,1], the "odds ratios" are defined across the real line. Of course, the result in (14) generalized provided we define $\tilde{y}_{it,(g)} = y_{it} - \eta_i + u_{it,(g)}$ and then we stack all $\tilde{y}_{it,(g)}$s together (see Technical Appendix A). *An efficient way to*

[3] A more restrictive formulation for the individual effects would have been: $\min_{\Theta} \sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{it} - \eta_i - \sum_{g=1}^{G}\varphi_g(x_{it}; \boldsymbol{\beta}_{(g)})P_{it,g}\right)^2$.

[4] If not, we can use the estimator $\hat{m} = (\mathbb{P}'\mathbb{P} + \kappa I)^{-1}\mathbb{P}'\tilde{y}$,where κ is related to the prior precision of the m coefficients being close to zero (Zellner, 1971, pp. 75–76).

Figure 2: Different approximations to a Cobb-Douglas production function

Notes: The blue dots correspond to the actual points. The step-wise function corresponds to a standard RT approximation. The dotted line corresponds to the actual Cobb-Douglas production function. The lines with squares correspond to the Probabilistic Regression Tree approximation.



generate partitions of the space of the regressors is to use the classical RT approach and update the probabilities $P_{it,g}$. Our posterior analysis for this step, is summarized in Appendix A.

To make the difference between standard RTs and PRTs clear, we present in Figure 2, observations from the Cobb-Douglas data generating process $Y_i = aX_i^b e^{v_i}$, where $a = 1$, $b = \frac{1}{2}$, the X_is are generated from a standard uniform distribution, the sample size is n=20, and $v_i \sim i.i.d\, N(0.1^2)$.

4 Data and empirical results

4.1 Data

We use the data from Instituto Nacional de Estadistica which covers all Chilean manufacturing plants with more than ten employees during 1979 - 1996. These data have been used in Levinsohn and Petrin (2003), Gandhi et al. (2020), and Ackerberg et al. (2015). For each of the 10,927 plants in the sample, the data include gross output, material inputs, capital stock and investments, fuels and electricity, and labor (measured in person-years, skilled as well as unskilled) converted where necessary into real values using industry-specific price deflators. A more detailed

description of the data is available in Levinsohn and Petrin (2003, pp. 323–325) as well as in Lee et al. (2019). We use the four largest industries (excluding petroleum and refining). The three-digit level industries and their ISIC codes are Metals (381), Textiles (321), Food Products (311) and Wood Products (331). The data are observed annually and they include gross revenue (the output index), indices of labor and capital inputs, and a measure of the intermediate inputs electricity, materials, and fuels. Quasi-fixed inputs in our setting, are capital stock, and skilled and unskilled labor. Materials, fuels and electricity are assumed to be variable inputs.

4.2 Main empirical results

To compare with previous approaches, like Levinsohn and Petrin (2003), Gandhi et al. (2020), and Ackerberg et al. (2015) and to create a useful benchmark, we estimate a model without PRTs using a translog production function. In turn, we estimate the model using particle-filtering Markov Chain Monte Carlo (MCMC; see Appendix A). Our prior for θ is

$$\theta \sim \mathcal{N}_d(\mathbf{0}, h^2\mathbf{I}_d), \tag{17}$$

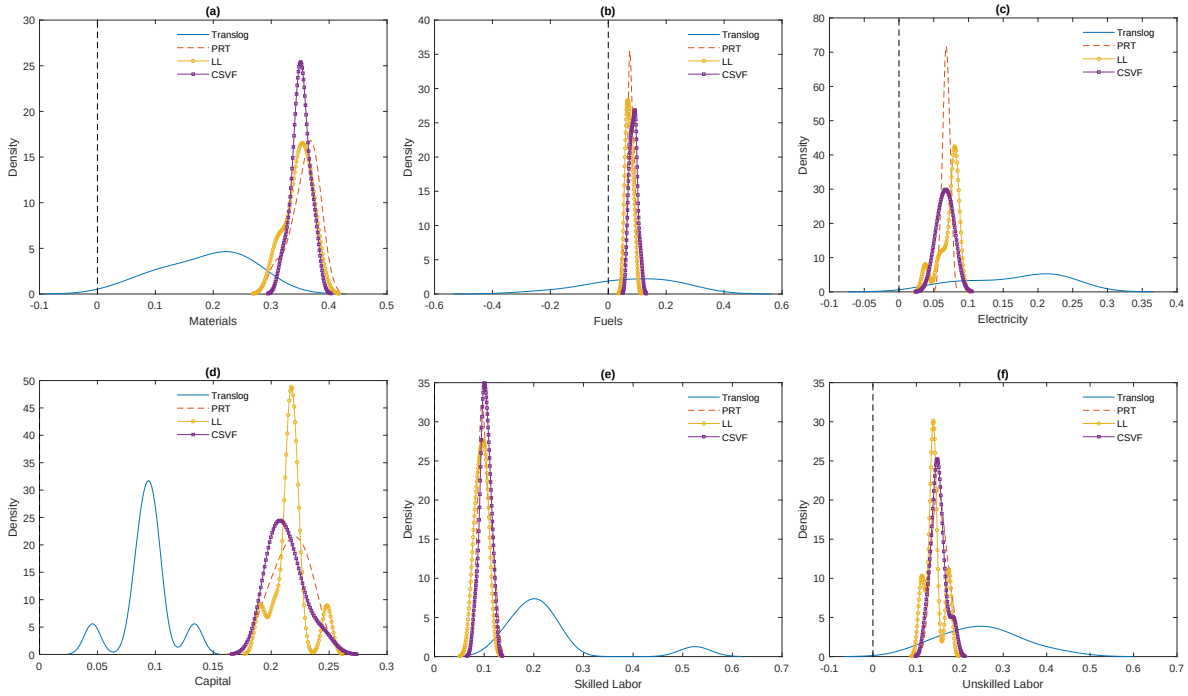
where $\mathbf{I}_d$ is the identity matrix (d is the dimensionality of the parameter space) and h is a scale parameter that we set to 10 so that the prior is proper bit diffuse. We do not impose other prior information, except for the fact that the translog parameters are restricted to provide monotonicity and concavity at the means of the data and 50 other randomly selected points in the support of $\mathcal{X}$. We use the translog instead of the more traditional Cobb-Douglas because of its better approximation properties. First, we compare the results from this approach (which we call **translog**) and the PRT approach, in terms of elasticities and other functions of interest. In Figure 3, we present sample distributions of posterior mean estimates of input elasticities. The translog violates monotonicity in several cases (most prominently for materials, fuels, and skilled labor) but these conditions are satisfied by the RTs as they anchor on Cobb-Douglas production functions at the terminal nodes. In this case, it is easy to impose monotonicity and concavity. Before proceeding, it is useful to define technical change (TC_{it}) as the partial derivative of the production function with respect to time, efficiency change as $EC_{it} = \frac{r_{it}-r_{i,t-1}}{r_{i,t-1}}$ where $r_{it} = e^{-u_{it}}$ and, finally, productivity growth as $PG_{it} \equiv \omega_{it} = TC_{it} + EC_{it}$. We compare our model with the translog model, as well as a local mlinear (LL) model; see SCKLS of Yagi et al. (2020), and CSVF of Valero-Carreras et al. (2021) in Figure 3. Inefficiency cannot be compared with Yagi et al. (2020) as in their paper they do not allow for technical inefficiency.[5]

From 3, the first conclusion is that the translog often violates monotonicity as some input elasticities can be negative. Other than that, PRT, SCKLS and CVSF seem to provide similar sample distributions of input elasticities, at least for the most part.

Estimates of returns to scale, reported in panel (a) of Figure 4 differ widely between the translog and RTs. For the translog, they range between 0.4 and 1.6 while for PRTs they range from 0.8 to 1.1. Productivity growth

[5]See Kuosmanen (2008) and Kuosmanen and Kortelainen (2012) on how inefficiency can be estimated in a *second stage* using distributional assumptions. Here, we opt for a single stage which is, necessarily, quite different.

Figure 3: Input elasticities



(reported in panel (b)) is also overstated by the translog (averages nearly 3% and ranges between zero and 7%). For PRTs, productivity growth averages close to zero and ranges from about –1% to 2%. Efficiency change ans technical change (reported in panels (c) and (d)) do not seem to be markedly different. Efficiency change averages close to zero and ranges roughly from –4% to 4%. Technical change (see panel (d)) averages close to 2% for the translog (close to zero for PRT, LL and CVSF) and ranges between –2% and 7% which is, of course, substantial, it indicates also substantial heterogeneity among different plants but it hard to believe in virew of the different evidence provided by PRT, LL, and CVSF. Finally, in panel (e) we report sample distributions of posterior mean estimates of inefficiency. PRTs provide an average close to 25% (ranging from about 10% to 40% with confidence 0.95) while the translog provides an average close to 15% and ranges roughly between 5% and 25%.

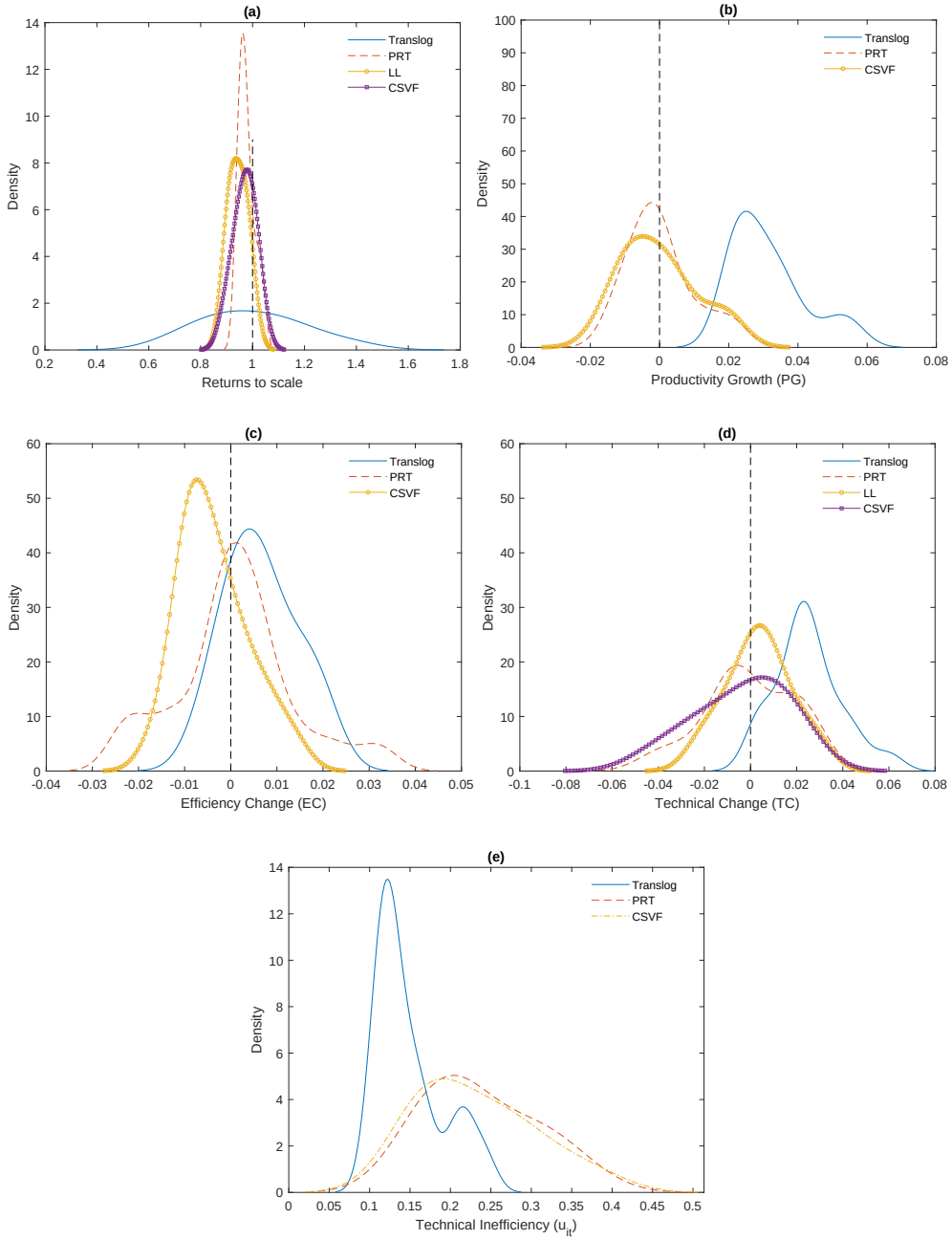
To examine the relationship between efficiency and productivity, we estimate a second-stage panel vector autoregression (PVAR) of the form

$$\begin{aligned} \log u_{it} &= \rho_{u0,i} + \rho_{uu} \log u_{i,t-1} + \rho_{u\omega} \log \omega_{i,t-1} + v_{it,u}, \\ \log \omega_{it} &= \rho_{\omega 0,i} + \rho_{\omega u} \log u_{i,t-1} + \rho_{\omega\omega} \log \omega_{i,t-1} + v_{it,\omega}, \end{aligned} \tag{18}$$

where $v_{it,u}$ and $v_{it,\omega}$ are error terms, and $\omega_{it} = PG_{it}$. Parameters $\rho_{u0,i}$ and $\rho_{\omega 0,i}$ represent fixed effects and the model in (18) is estimated using the system generalized method of moments (GMM) technique.

Posterior moments related to selected parameters of interest, including persistence of inefficiency and productivity are reported in Table 1. In general, both productivity and inefficiency are highly persistent (viz. they exhibit

Figure 4: Distributions of other functions of interest



Notes: Efficiency Change is defined as $EC_{it} = r_{it} - r_{i,t-1}$ where $r_{it} = e^{-u_{it}}$ where $u_{it} \geq 0$ represents technical inefficiency. Technical change (TC_{it}) is given by the derivative of the production function with respect to time trend. Productivity growth is defined as $PG_{it} \equiv \omega_{it} = TC_{it} + EC_{it}$.

Table 1: Posterior moments of selected parameters

	PRT	translog	CVSF
$\rho_{\omega\omega}$	0.772	0.885	0.770
	(0.044)	(0.055)	(0.032)
ρ_{uu}	0.817	0.913	0.744
	(0.023)	(0.044)	(0.013)
$\rho_{\omega u}$	0.051	0.030	0.144
	(0.012)	(0.022)	(0.032)
$\rho_{u\omega}$	0.085	0.132	0.023
	(0.022)	(0.144)	(0.030)

Table 2: Rank correlation coefficients of technical change estimates

	translog	PRT	SCKLS	CVSF
translog	1.000	0.032	0.035	0.029
PRT		1.000	0.618	0.545
LL			1.000	0.747
CVSF				1.000

path dependence, Tsekouras et al., 2016, 2017) and there is significant cross-dependence between them according to RTs. The translog does not allow for significant cross-dependence as parameters $\rho_{\omega u}$ and $\rho_{u\omega}$ seem to be "statistically insignificant". Moreover, and perhaps because of this shortcoming of the translog, persistence parameters are overstated. From these results, it turns out that although PRTs and CSVF deliver similar estimates of input elasticities, inefficiency, technical change etc., there are also some important differences, otherwise the results in 1 would be approximately the same for PRT and CVSF.

From the rank correlation coefficients reported in Table 2, we see that although the methods provide positively correlated estimates of technical change, the correlations do not seem to be large enough.

In Table 3 we present marginal effects in the form of elasticities derived from the production function, along with corresponding results obtained from the translog functional form.

In Table 3, the mean values of EC and PG under PRT have opposite values comparing to that under CSVF.[6] The reason is that PRT and CSVF have different approximation properties at least in finite samples. As we mentioned earlier, Valero-Carreras, Aparicio, and Guerrero (2021) introduced a *second stage* to impose convexification. The two-stage approach depends on the behavior of the first stage, particularly in finite samples.

As the results from the PRT and the translog are quite different, it is reasonable to inquire as to which model

[6]The author is grateful to an anonymous referee for raisin this point.

Table 3: Marginal effects

Notes: Reported are sample means with standard deviations in parentheses. PRT stands for probability regression trees, LL stands for local linear estimation as in Yagi et al. (2020) and CVSF is the method of Vallelo-Carreras et al. (2021). TC is technical change, EC is efficiency change and PG=EC+TC is productivity change. Technical change is denoted by u_{it}. Regular fonts represent elasticities from the regression tree model. Italics represent results from the translog specification. Finally, ω_{it} represents productivity growth (PG_{it}).

inputs				
	translog	PRT	LL	CSVF
Materials	0.187	0.358	0.348	0.351
	(0.072)	(0.024)	(0.022)	(0.015)
Fuels	0.077	0.077	0.073	0.087
	(0.147)	(0.012)	(0.011)	(0.013)
Electricity	0.165	0.068	0.072	0.066
	(0.068)	(0.005)	(0.015)	(0.009)
Capital	0.093	0.217	0.216	0.213
	(0.021)	(0.014)	(0.012)	(0.015)
Skilled labor	0.229	0.102	0.094	0.102
	(0.109)	(0.012)	(0.011)	(0.010)
Unskilled labor	0.249	0.155	0.141	0.150
	(0.087)	(0.015)	(0.021)	(0.017)
TC	0.025	-0.0007	0.0034	-0.0043
	(0.014)	(0.019)	(0.014)	(0.019)
EC	0.066	0.010	—	-0.0037
	(0.013)	(0.008)		(0.0075)
PG	0.031	0.0003	—	-0.0002
	(0.010)	(0.009)		(0.0108)
u_{it}	0.149	0.238	—	0.232
	(0.042)	(0.065)		(0.069)

Table 4: Cross-validation RMSEs

method	RMSE
translog	0.336
PRT	0.025
LL	0.027
CVSF	0.029

is "best". We use the criterion of marginal likelihood and Bayes factor (also known as posterior odds ratio when the prior odds for the two models are 1:1) to answer this question. The marginal likelihood is standard output of sequential Monte Carlo (Andrieu et al., 2010) so it is not difficult to compute Bayes factors (Kass and Raftery, 1995; DiCiccio et al., 1997; O'Hagan, 1995).[7]

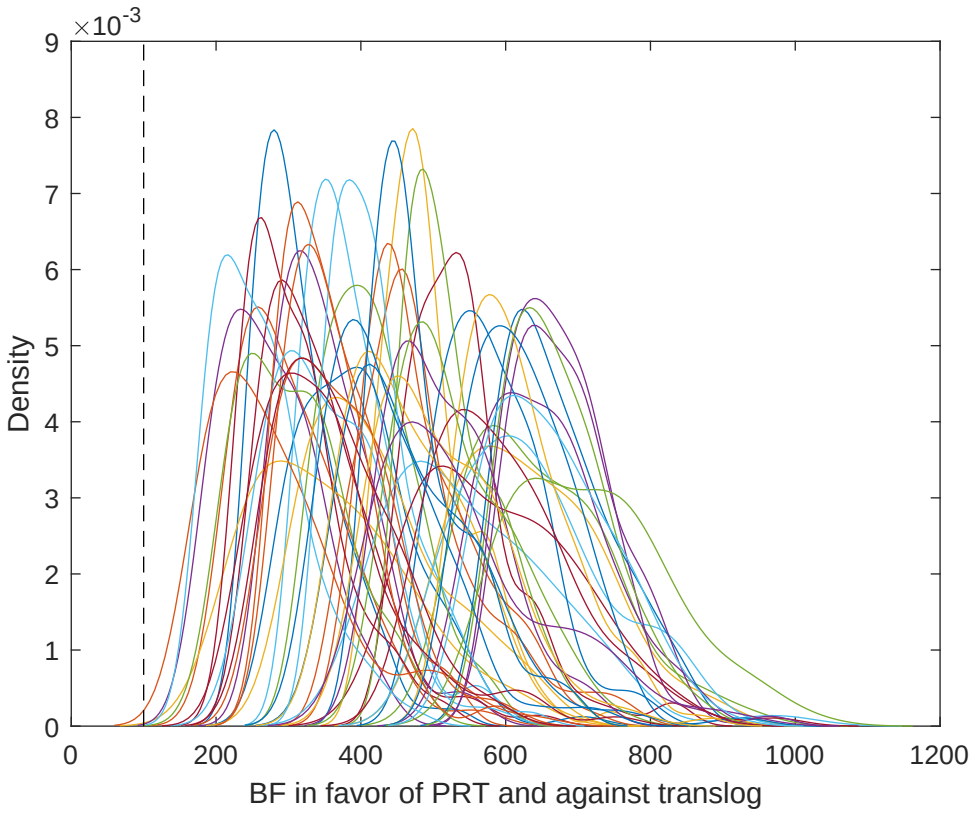
We use as estimation sample the 50% of available observations in each sector, and we re-estimate the model 1,000 times corresponding to different estimation samples. The distribution of Bayes factors in favor of RT and against the translog is reported in Figure 5. In addition, we consider different priors as described in Appendix B. The results are reported in Figure 5. For the most part, Bayes factors exceed 100 which is a conventional benchmark for "decisive evidence" in favor of a given model. As Bayes factors range up to roughly 1200, there is overwhelming evidence in favor of RTs and against the translog. Of course, this cannot be attributed to the fact that the translog violates the monotonicity restrictions in several instances, as imposing restrictions cannot improve the fit of the translog (which is also heavily underparametrized relative to RTs).

As there are no Bayesian versions of SCKLS or LL and CVSF it is not possible to use Bayes factors for model comparison or model selection. We can use, however, (tenfold, say) cross-validation root mean-squared errors (RMSEs) to compare the different models. To allow comparison we shut down inefficiency in translog, PRT, and CVSF and we focus on comparing RMSEs of predicted dependent variables. The tenfold cross-validation RMSEs are reported in Table 4 and they are based on the posterior mean estimates for both the translog and the PRT. The smallest RMSE is attained by PRTs, followed by SCKLS and CVSF. So, the evidence in Tables 1, 2 and 4 suggests that sampling distributions of functions of interest are somewhat similar for PRT, LL and VSCF, there are, nevertheless, important differences as well. The differences can be attributed to the different sub-division of the regressor space implied by PRTs and the way convexification is obtained in CVSF relative to both PRTs but primarily LL. A further comparison of the methods is left for future research along with the fact that SCKLS should be modified appropriately to allow for technical inefficiency.

To examine how PRTs achieve smoothness and satisfy the properties of monotonicity and concavity, we provide some examples in Figure 6.

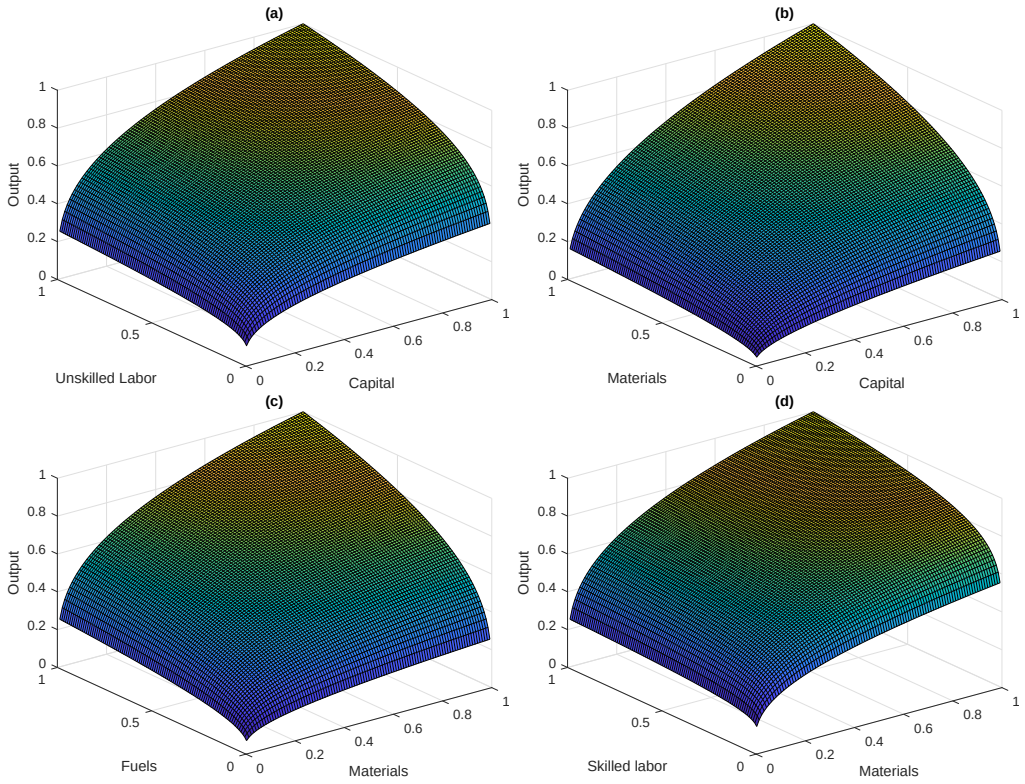
[7]For a model with parameters $\boldsymbol{\theta}$,data D, likelihood $L(\boldsymbol{\theta}; D)$ and prior $p(\boldsymbol{\theta})$, the marginal likelihood is $\mathcal{M}(D) = \int L(\boldsymbol{\theta}; D)p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$, i.e., the integrating constant of the posterior. For two models, say "1" and "2" (estimated with the same data with possibly different parameters and, therefore, likelihood functions and priors) the Bayes factor in favor of model "1" and against model "2" is $\mathcal{B}_{1:2} = \frac{\mathcal{M}_1(D)}{\mathcal{M}_2(D)}$.

Figure 5: Bayes factors



Notes: The broken vertical line corresponds to 100. Shown are 50 densities corresponding to 50 representative different priors as described in Appendix B.

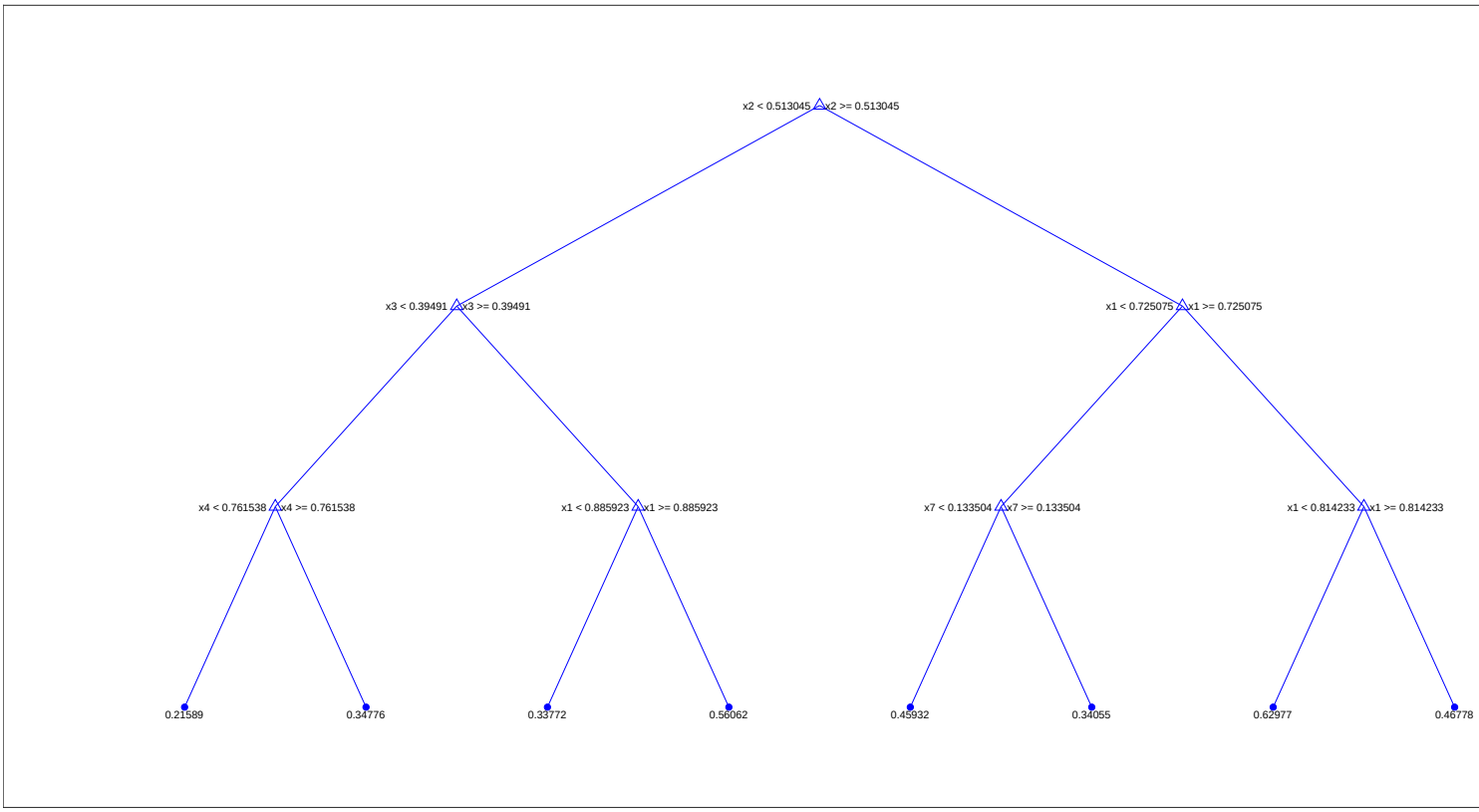
Figure 6: Aspects of production function



Notes: For visual clarity all variables are normalized in the unit interval. In different panels, we present the production function in terms of two variables when all other variables all fixed at their average values. The third imaginary axis represents output so, the plots are surfaces that show that inputs must be used in tandem to produce more output.

Figure 7: Regression Tree for the production function

Notes: x_1 through x_6 correspond to the inputs (capital, unskilled labor, skilled labor, materials, fuels, electricity) scaled in the unit interval, and x_7 is inefficiency (u) in original units.



The PRT is shown in Figure 7.[8] This visualization shows that unskilled labor and capital are critical in determining the partition of the input space, and the final nodes are determined by inefficiency and productivity. Specifically, the final partition depends mostly on relatively inefficient and less productive plants versus their better counterparts. Inefficiency values around 10% or 25% seem to be critical along with productivity levels around 0.65%.

Concluding remarks

We have proposed a probabilistic regression tree (RT) approach to nonparametric smooth and monotone concave approximation to production functions. We combine the RT approach with Markov Chain Monte Carlo methods which allow computationally efficient sub-divisions of the regressor space. If we abstract from productivity and inefficiency, the model is a reasonable parametrization for the application of RTs in panel data which is a persistent problem in the literature.

[8]More detailed results are presented in Online Appendix C.

In terms of future research, there are at least four directions in which the current approach can be generalized. *First*, one can allow easily for endogeneity (e.g. Soytas et al., 2019) in the variable inputs using a systems approach consisting of the first order conditions for cost minimization, which is an alternative to the control function approach as it can handle more one variable input. This will involve, most likely, a separation between variable and quasi-fixed inputs. At the same time, the Bayesian approach allows for measurement error in the variable inputs, unlike the control function approach. *Second*, introduction of productivity as a dynamic latent variable is not difficult and brings the model closer to the spirit of previous studies (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015; Gandhi et al., 2020). *Third*, inefficiency can be parametrized in terms of environmental variables and one can even allow for dynamic inefficiency, in the spirit of first-order Markov processes used in the literature for productivity. *Fourth*, further model comparison can be performed using marginal likelihoods and Bayes factors provided we can give a Bayesian interpretation to SCKLS and CSVF.

Geweke, J., and Petrella, L. (2014). Likelihood-based inference for regular functions with fractional polynomial approximations. Journal of Econometrics 183 (1), 22–30.

Appendix A. Posterior analysis

Let us define Θ as in (13), θ denotes all other structural parameters of the model. We treat the $u_{it,(g)}$s as parameters and we place a Laplace prior on the absolute "odds ratios" $\psi_{it,(g)} = \frac{r_{it,(g)}}{1-r_{it,(g)}}$ where $r_{it,(g)} = e^{-u_{it,(g)}}$.The prior on the tree structure is specified along the lines suggested in Chipman et al. (1998) and Chipman et al. (2010). The posterior distribution of the model is as follows.

$$p(\Theta, u, \eta, u|Y) \propto \sigma_\epsilon^{-nT} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[y_{it} - \eta_o - \sum_{g=1}^{G}\left(\eta_{i,(g)} + \{\varphi_g(x_{it};\beta_{(g)} - u_{it,(g)}\}\right)P_{it,(g)}\right]^2\right\} \cdot p(\Theta, \sigma, \eta, u), \tag{A.1}$$

where $p(\Theta, \sigma, \eta, u) \propto p(\Theta)p(\sigma)p(\eta)p(u)$ (viz., we assume prior independence) where the prior $p(\sigma_\epsilon) \propto \sigma_\epsilon^{-1}$ and the prior $p(\Theta)$ is defined as proceed. Moreover, $u = \left[u_{it,(g)}\right]$ and $\eta = \left[\eta_{i,(g)}\right]$. Notice that we have introduced an overall intercept η_o so that the individual effects $\eta_{i,(g)}$ can be reasonably assumed to have zero prior means. The prior on the u parameters is given by a Laplace distribution on the "odds ratios":

$$p(\psi_{it,(g)}) \propto e^{-\lambda|\psi_{it,(g)}|}, \tag{A.2}$$

where $\psi_{it,(g)} = \frac{e^{-u_{it,(g)}}}{1-e^{-u_{it,(g)}}}$ as in (16). For the individual effects we follow a similar approach as we want to exploit properties of the LASSO:

$$p(\eta_{i,(g)}) \propto e^{-\lambda_\eta|\eta_{i,(g)}|}, \tag{A.3}$$

where λ and λ_η are unknown parameters whose priors are standard exponential. Updating $\beta_{(g)}, u_{it,(g)}, \eta_{i,(g)}$ is performed using the fast MCMC technique of Durmus et al. (2017). Updating σ_ϵ is straightforward as

$$\frac{Q}{\sigma_\epsilon^2}\Big|\Theta, \eta, u, Y \sim \chi^2_{nT}, \tag{A.4}$$

where

$$Q = \sum_{i=1}^{n}\sum_{t=1}^{T}\left[y_{it} - \sum_{g=1}^{G}\left(\eta_{i,(g)} + \{\varphi_g(x_{it}; \beta_{(g)} - u_{it,(g)}\}\right) P_{it,(g)}\right]^2. \tag{A.5}$$

An important part of our MCMC is that the $\boldsymbol{\sigma}$ parameters in (10) are updated as well. Each element, say σ_g of $\boldsymbol{\sigma} = [\sigma_g,\ 1 \leq g \leq G]$ is given the following improper prior:

$$p(\sigma_g) \propto \sigma_g^{-1} e^{-\bar{q}/(2\sigma_g^2)},\ 1 \leq g \leq G, \tag{A.6}$$

where $\bar{q} \geq 0$ is a prior parameter that we set to $\bar{q} = 10^{-4}$ (for the prior, see Zellner, 1971, p. 371, equation A.37b). After reparametrizing to $\sigma_g = e^{\xi_g}$ ($\xi_g \in \mathbb{R}$,$1 \leq g \leq G$), the ξ_gs can be treated as unconstrained parameters in MCMC; otherwise in a non-Bayesian context one would have to select $\boldsymbol{\sigma}$ by computationally expensive cross-validation techniques.

Implementation of the regressor sub-division, is the most critical, uses a standard RT approach as implemented in Akhoury et al. (2020). Of course, there are other probabilistic alternatives to build RTs (e.g. Linero and Yang, 2018). For RTs, the steps and calculations are performed on values of a single response variable. [9]

MCMC is implemented using 150,000 iterations the first 50,000 of which are discarded in the burn-in phase to mitigate possible start up effects. Convergence and numerical performance of MCMC is monitored using Geweke's (1992) diagnostics.

Next, we describe our approach to tree construction and sub-division of the regressor space in more detail. Let $\mathfrak{g}(X, \mathscr{T}_j, m_j)$ denote a single tree model with $\mathscr{T}_j$ standing for the tree structure associated with the jth binary tree and $m_j = [\mu_{j1}, \ldots, \mu_{jb_j}]'$ is the vector of terminal node parameters associated with $\mathscr{T}_j$ and b_j are the leaves of the jth tree. In standard BART, the m_js consist of fixed parameters so that the RT is a step function approximation. We approximate a function using

$$f(X) = \sum_{j=1}^{N} \mathfrak{g}(X, \mathscr{T}_j, m_j). \tag{A.7}$$

Binary trees have the form $\{X \in \mathcal{A}_{jg}\}$ or $\{X \notin \mathcal{A}_{jg}\}$. The sets $\mathcal{A}_{jg}$ are defined by selecting particular columns of X, say $X_{.j}$ ($1 \leq j \leq K$) so these rules have the form $\{X_{.j} \geq c\}$ or $\{X_{.j} \geq c\}$ for some threshold value c. The

[9] See the `bartMachine` package in R, the `BayesTree` package (Linear and Yang, 2020) as well as `https://github.com/theodds/SoftBART`.

function $\mathfrak{g}$ has the following form:

$$\mathfrak{g}(X, \mathscr{T}_j, m_j) = \varphi(X; \beta_{jg}), \text{ if } X \in \mathcal{A}_{jg}, \tag{A.8}$$

where $\varphi(X; \beta_{jg})$ is a Cobb-Douglas functional form with parameters β_{jg}. In standard BART (Chipman et al., 2010), we have instead

$$\mathfrak{g}(X, \mathscr{T}_j, m_j) = \mu_{jg}, \text{ if } X \in \mathcal{A}_{jg}, \tag{A.9}$$

where μ_{jg} is a constant.

In the interest of generality we consider a model with multiple dependent variables. We write the jth equation as

$$f_j(X) = \sum_{g=1}^{G} \mathfrak{g}_{jg}(X, \mathscr{T}_{jg}, m_{jg}), \tag{A.10}$$

where $\mathfrak{g}_{jg}(\cdot)$ denotes an equation-specific Cobb-Douglas function with arguments $\mathscr{T}_{jg}$ and $m_{jg} (1 \leq g \leq G)$. We write the entire system as

$$Y = f(X) + \xi, \tag{A.11}$$

where Y denotes all observations on the dependent variables, X is the matrix of observations on the predictors and ξ is the error term. As the errors are correlated, we use the reparametrization (Carriero et al., 2019; Huber et al., 2020):

$$\xi = \epsilon A_o', \tag{A.12}$$

where A_o is a lower triangular matrix with ones on the main diagonal, so that $\mathbf{\Sigma} = A_o H A_o'$ where H is a diagonal matrix and ϵ are normal errors with zero means and the same diagonal covariance matrices H. This construction permits equation-by-equation estimation as the errors are now independent. The jth equation ($j > 1$) can now be written as

$$y_{.j} = \sum_{g=1}^{G} \mathfrak{g}_{jg}(X, \mathscr{T}_{jg}, m_{jg}) + \sum_{l=1}^{j-1} a_{jl} \xi_{.l} + \epsilon_{.j}, \tag{A.13}$$

where $y_{.j}$, $\xi_{.l}$, $\epsilon_{.j}$ refer to the jth or lth column of Y, ξ or ϵ, and a_{jl} is the (j,l)th elements of A_o. The formulation in (A.13) is a standard BART except for the fact that we have the additional term $\sum_{l=1}^{j-1} a_{jl} \xi_{.l}$ which is just a parametric regression part.

For each equation j, the joint prior is

$$\begin{aligned} &p\left((\mathscr{T}_{j1}, m_{j1}), \ldots, (\mathscr{T}_{jN}, m_{jN}), c_j, \sigma_j^2, (\beta_{jk}), A_o\right) \propto \\ &p\left((\mathscr{T}_{j1}, m_{j1}), \ldots, (\mathscr{T}_{jN}, m_{jN})\right) \cdot p\left(c_j, \sigma_j^2, (\beta_{jk}), A_o\right). \end{aligned} \tag{A.14}$$

Following Chipman et al. (2010) we assume

$$\begin{aligned} p\left((\mathscr{T}_{j1}, m_{j1}), \ldots, (\mathscr{T}_{jN}, m_{jN})\right) &= \textstyle\prod_{k=1}^{N} p(m_{jk}|\mathscr{T}_{jk}) p(\mathscr{T}_{jk}), \\ p(m_{jg}|\mathscr{T}_{jg}) &= \textstyle\prod_{q=1}^{b_q} p(\beta_{jk,q}|\mathscr{T}_{jk}), \end{aligned} \tag{A.15}$$

where we introduce the prior $p(\beta_{jg,q}|\mathscr{T}_{jk})$ instead of $p(\mu_{jg,q}|\mathscr{T}_{jk})$. This prior allows for explicit integration of m_{jg} out of the posterior of the trees. Our prior on the tree structure is specified as in Chipman et al. (1998, 2010). Specifically, we use a stochastic process that consists of three steps to grow trees. Let $s = 0$ be the first iteration of this tree generating process. We start with a tree that contains a single terminal node, which is denoted by η_{jg} (for each equation j and $g = 1, \ldots, G$). The process works as follows.

1. We split the terminal node η_{jg} with probability

$$p_{split}(\eta_{jg}, \mathscr{T}_{jg}^{(s)}) = \alpha(1+d)^{-\delta}, \quad \alpha \in (0,1), \quad \delta \geq 0,\ d \in \{0, 1, 2, \ldots\}, \tag{A.16}$$

where d is the depth of the tree, and α, δ are prior parameters (we set $\alpha = 0.95$ and $\delta = 2$. This implies that that the probability that a given node is non-terminal decreases quadratically if the trees become more complicated (i.e. for increasing values of d).

2. If the current node is split, we choose a splitting variable $X_{\cdot j}$ with probability $\frac{1}{K}$. As we need to determine a threshold for this variable, we assume that the threshold is uniformly distributed in the range of $X_{\cdot j}$.[10]

3. Once we obtain all terminal nodes (i.e. there are no nodes to split further), we denote the new tree by $\mathscr{T}_{jg}^{(s+1)}$ and return to step 2.

On the terminal node parameters $\beta_{jg,q}$ we place a flat prior subject to non-negativity restrictions. For the free parameters in matrix A_o we use a standard normal prior. We sample all quantities related to the trees (i.e. $\mathscr{T}_{jg}$ and m_{jg} for all j, k) using the algorithm in Chipman et al. (2010). Specifically, we consider candidate tree, say $\mathscr{T}_{jg}^{c}$ from a proposal distribution $q(\mathscr{T}_{jg}, \mathscr{T}_{jg}^{c})$ and we accept the proposal using the Metropolis-Hastings acceptance probability

$$\min\left\{1, \quad \frac{p(R_{jg}|X, \mathscr{T}_{jg}^{c}, m_j) p(\mathscr{T}_{jg}^{c}) q(\mathscr{T}_{jg}^{c}, \mathscr{T}_{jg})}{p(R_{jg}|X, \mathscr{T}_{jg}, m_j) p(\mathscr{T}_{jg}) q(\mathscr{T}_{jg}, \mathscr{T}_{jg}^{c})}\right\}. \tag{A.17}$$

The proposal distribution $q(\mathscr{T}_{jg}, \mathscr{T}_{jg}^{c})$ is constructed as in Chipman et al. (1998). In the first step we grow the tree by splitting a node. This step is chosen with probability 0.25. The second step combines two non-terminal nodes into one terminal node. This step is chosen with probability 0.25. The third step interchanges splitting rules between two terminal nodes with probability 0.4. The fourth step changes the splitting rule of a single non-terminal node with probability 0.1.

[10] For this reason, all observed variables are scaled in the unit interval.

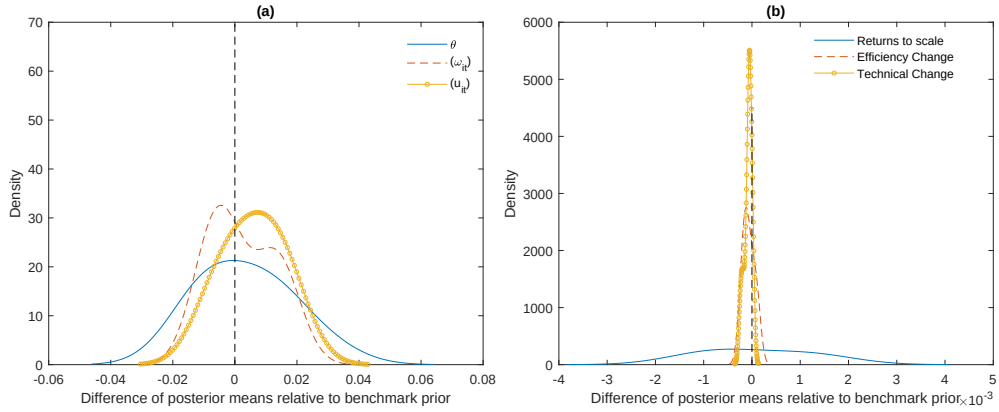
Appendix B. Prior sensitivity analysis

Our main interest is in examining sensitivity to the prior in (17) which we reformulate here as follows.

$$\theta \sim \mathcal{N}_d(\mathbf{a},\, h^2\mathbf{I}_d), \tag{B.1}$$

where $\mathbf{a} \in \mathbb{R}^d$ denotes the prior mean which has been set to zero in (17). We vary h between 1 and 100 using a uniform distribution, and we choose the different elements of $\mathbf{a}$ using uniform distributions in the interval $[-10, 10]$. For the LASSO parameters λ and λ_η we assume they vary in the interval [0.1, 10] and for $\bar{q}$ in (A.6) we assume that it ranges between 10^{-7} and 10. We perform this exercise 1,000 times to obtain 1,000 different priors and we re-estimate the model, re-computing posterior moments of parameters and functions of interest. In Figure B.1 we present the distributions of (median) change in posterior means for θ, $\{\omega_{it}\}$ and $\{u_{it}\}$ in panel (a), and returns to scale, technical change and efficiency change in panel (b). In both cases, differences in posterior means due to different priors seem small enough implying that the model is reasonably robust with respect to the priors.

Figure B.1: Prior sensitivity



References

[1] Ackerberg, D. A., K. Caves, & G. Frazer (2015). Identification Properties of Recent Production Function Estimators. Econometrica 83 (6), 2411–2451.

[2] Akhoury, S., E. Devijver, M. Clausel, M. Tami, E. Gaussier, & G. Oppenheim (2020). Smooth And Consistent Probabilistic Regression Trees. 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

[3] Andrieu C., Doucet A., Holenstein R. (2010). Particle Markov chain Monte Carlo methods. Journal of the Royal Statistical Society Series B Statistical Methodology 72, 269–342.

[4] Blanquero, R., E. Carrizosa, C. Molero-Río, D. R. Morales (2020). Sparsity in optimal randomized classification trees. European Journal of Operational Research 284 (1), 255–272.

[5] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Belmont Wadsworth Int. Group, New York.

[6] Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. Statistical Science, 22(4), 477–505.

[7] Carriero, A., Clark, T. E., and Marcellino, M. (2019). Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors. Journal of Econometrics, 212(1):137–154.

[8] Chen, Y-T, E. W.Sun, M-F Chang, Y-B Lin (2021). Pragmatic real-time logistics management with traffic IoT infrastructure: Big data predictive analytics of freight travel time for Logistics 4.0. International Journal of Production Economics 238, 108157.

[9] Chernozhukov, V., Fernandez-Val, I. & Luo, Y. (2018). The sorted effects method: discovering heterogeneous effects beyond their averages, Econometrica 86 (6), 1883–1909.

[10] Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. Journal of the American Statistical Association, 93(443):935–948.

[11] Chipman, H. A, E. I. George, and R. McCulloch (2010). BART: Bayesian additive regression trees. Annals of Applied Statistics, 4:266–298.

[12] Coglianese, J., Davis, L. W., Kilian, L. & Stock, J. H. (2017), Anticipation, tax avoidance, and the price elasticity of gasoline demand, Journal of Applied Econometrics 32(1), 1–15.

[13] De'Ath, G., 2002. Multivariate regression trees: a new technique for modeling species–environment relationships. Ecology 83, 1105–1117.

[14] Denison, D., B. Mallick, and A. Smith. (1998). A Bayesian CART algorithm. Biometrika 85, 363–377.

[15] DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. Journal of the American Statistical Association 92 903–915.

[16] Durmus, A., G. O. Roberts, G. Vilmart, and K. C. Zygalakis (2017). Fast Langevin based algorithm for MCMC in high dimensions. The Annals of Applied Probability 27 (4), 2195–2237.

[17] Emrouznejad, A., & Anouze, A. L. (2010). Data envelopment analysis with classification and regression tree–a case of banking efficiency. Expert Systems 27(4), 231–246.

[18] Esteve. M., J. Aparicio, A. Rabasa, & J. J. Rodriguez-Sala (2020). Efficiency analysis trees: A new methodology for estimating production frontiers through decision trees. Expert Systems with Applications Volume 162, 30 December 2020, 113783.

[19] Fisher, M., Gallino, S. & Li, J. (2017). Competition-based dynamic pricing in online retailing: A methodology validated with field experiments, Management Science 64(6), 2496– 2514.

[20] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119–139.

[21] Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). Annals of Statistics 19, 1–41.

[22] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232.

[23] Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics and Data Analysis, 38(4), 367–378.

[24] Fu, W. and J. S. Simonoff (2015). Unbiased regression trees for longitudinal and clustered data. Computational Statistics and Data Analysis 88, 53–74.

[25] Friedman, J. H., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. The Annals of Statistics, 28(2), 337–407.

[26] Gandhi, A., S. Navarro, and D. A. Rivers (2020). On the Identification of Gross Output Production Functions. Journal of Political Economy 128 (8), 2973–3016.

[27] Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4* (J. M. Bernado, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Clarendon Press, Oxford, UK, 169–193.

[28] Geweke, J. and Keane, M. (2007). Smoothly mixing regressions. Journal of Econometrics 138, 252–290.

[29] Geweke, J., and Petrella, L. (2014). Likelihood-based inference for regular functions with fractional polynomial approximations. Journal of Econometrics 183 (1), 22–30.

[30] Gunasekaran, A., Kobu, B. (2007). Performance measures and metrics in logistics and supply chain management: a review of recent literature (1995-2004) for research and applications. International Journal of Production Research 45 (12), 2819–2840.

[31] Hajjem, A., Bellavance, F., Larocque, D. (2011). Mixed effects regression trees for clustered data. Statist. Probab. Lett. 81, 451–459.

[32] Hastie, T., Tibshirani, R., Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, NewYork.

[33] Huber, F., Koop, G., and Onorante, L. (2020). Inducing sparsity and shrinkage in time-varying parameter models. Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2020.1713796.

[34] Kass, R. and Raftery, A. (1995). Bayes Factors. Journal of the American Statistical Association 90, 773–795.

[35] Kumbhakar, S. C. & Park, B. U. & Simar, L. & Tsionas, M. G. (2007). Nonparametric stochastic frontiers: A local maximum likelihood approach, Journal of Econometrics, Elsevier, vol. 137 (1), 1–27.

[36] Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. Econometrics Journal 11, 308–325.

[37] Kuosmanen, T., Johnson, A.L. (2010). Data envelopment analysis as nonparametric least-squares regression. Operations Research 58 (1), 149–160.

[38] Kuosmanen, T., M. Kortelainen (2012) Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. J Prod Anal 38,11–28.

[39] Lee, C.-Y., A.L. Johnson, E. Moreno-Centeno, T. Kuosmanen (2013). A more efficient algorithm for Convex Nonparametric Least Squares, European Journal of Operational Research 227, 391–400.

[40] Lee, Y., A. Stoyanov, N. Zubanov (2019). Olley and Pakes-style Production Function Estimators with Firm Fixed Effects. Oxford Bulletin of Economics and Statistics 81 (1), 79–97.

[41] Lee, B. K., Lessler, J., and Stuart, E. A. (2010), Improving Propensity Score Weighting Using Machine Learning, Statistics in Medicine, 29, 337–346.

[42] Levinsohn, J., and A. Petrin (2003). Estimating Production Functions Using Inputs to Control for Unobservables. Review of Economic Studies 70 (2), 317–341.

[43] Linero, A. R., & Y. Yang (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. J. R. Statist. Soc. B 80, Part 5, 1087–1110.

[44] Loyer, J-L., E. Henriques, M. lFontul, & S. Wiseall (2016). Comparison of Machine Learning methods applied to the estimation of manufacturing cost of jet engine components. International Journal of Production Economics 178, 109–119.

[45] Masci, C., G. Johnes, & T. Agasisti (2018). Student and school performance across countries: A machine learning approach. European Journal of Operational Research 269 (3), 1072–1085.

[46] Murthy, S.K., 1998. Automatic construction of decision trees from data: a multi-disciplinary survey. Data Mining and Knowledge Dicovery 2, 345–389.

[47] Norets, A. (2010). Approximation of conditional densities by smooth mixtures of regressions. The Annals of Statistics 38 (3), 1733–1766.

[48] Schulte, P. M. (2015). The effects of temperature on aerobic metabolism: towards a mechanistic understanding of the responses of ectotherms to a changing environment, Journal of Experimental Biology 218 (12), 1856–1866.

[49] O'Hagan, A. (1995). Fractional Bayes Factor for Model Comparison, Journal of the Royal Statistical Society, Series B, 57, 99–138.

[50] Olley, G., and A. Pakes (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. Econometrica 64 (6), 1263–1297.

[51] Rebai, S., Yahia, F. B., & Essid, H. (2019). A graphically based machine learning approach to predict secondary schools' performance in Tunisia. Socio-Economic Planning Sciences, 100724.

[52] Segal, M.R., 1992. Tree-structured methods for longitudinal data. J. Amer. Statist. Assoc. 87, 407–418.

[53] Sela, R.J., Simonoff, J.S., 2012. RE-EM trees: a data mining approach for longitudinal and clustered data. Machine Learning 86, 169–207.

[54] Soytas, M. A., Denizel, M., and Usar, D. D. (2019). Addressing endogeneity in the causal relationship between sustainability and financial performance. International Journal of Production Economics 210, 56–71.

[55] Tsekouras, K., N. Chatzistamoulou, K. Kounetas, & D. C. Broadstock (2016). Spillovers, path dependence and the productive performance of European transportation sectors in the presence of technology heterogeneity. Technological Forecasting and Social Change, 102, 261–274.

[56] Tsekouras, K., N. Chatzistamoulou, & K. Kounetas (2017). Productive performance, technology heterogeneity and hierarchies: Who to compare with whom. International Journal of Production Economics 193, 465–478.

[57] Valero-Carreras D, Aparicio J, Guerrero N M. (2021). Support vector frontiers A new approach for estimating production functions through support vector machines. Omega, 2021, 104: 102490.

[58] Yagi, D., Y. Chen, A.L. Johnson, T. Kuosmanen (2020). Shape-Constrained Kernel-Weighted Least Squares: Estimating Production Functions for Chilean Manufacturing Industries. Journal of Business & Economic Statistics 38, 43–54.

[59] Zhang, H. (1998). Classification trees for multiple binary responses. Journal of the American Statistical Association 93, 180–193.