BAYESIAN ANALYSIS OF HETEROGENEITY IN STOCHASTIC FRONTIER MODELS

Jorge Eduardo Galán Camacho

Department of Statistics

University Carlos III de Madrid

A thesis submitted for the degree of

PhilosophiæDoctor (PhD) in Business Administration and Quantitative Methods

2014, July

Dedicated to my family. Specially, to my son, David Andrés, who was born during the doctorate program and has been my great inspiration and motivation to finish my doctorate program. To my loving wife, Faride, who showed me that no matter where I decide to go, she will be there for me. A special feeling of gratitude to my mother, Angela, who has always been an example to follow and has supported me throughout the entire doctorate program. To my father, Jorge Luis, who always believed in me and always had words of encouragement. To all my relatives and friends in Colombia who I know, from the distance, have always supported and believed in me.

Preface

Frontier models and efficiency measurement are closely related fields, which are currently a wide research area with plenty of applications. The first thoughts on efficiency were documented through some management theories at the beginning of the 20th century and from economics by authors such as John Hicks in the 1930s. In general, the main concern was the failure to achieve the theoretical production and profits maxima. However, it was not until the 50s when a formal definition of technical efficiency was introduced in Koopmans (1951), followed by some applications of this concepts developed in Debreu (1951) and Farrell (1957). These initial attempts provided the theoretical basis for the introduction of robust methods of efficiency measurement through frontier models in the late 70s. Since then, two different methodological approaches have been widely developed and applied: nonparametric and parametric methods (see Coelli et al., 2005; Fried et al., 2008; Kumbhakar and Lovell, 2000, for some excellent books on these techniques).

Nonparametric approaches are flexible but they have the disadvantage of providing, in general, deterministic inefficiency measures. The most common nonparametric method is Data Envelopment Analysis (DEA) introduced in Charnes et al. (1978). On the other hand, parametric approaches have the advantage of introducing an error term, which may account for measurement errors, omitted variables and functional form errors. The most important parametric method is Stochastic Frontier Analysis (SFA) initially introduced in Aigner et al. (1977) and Meeusen and van den Broeck (1977) (see O'Donnell, 2014, for an interesting explanation of the assumptions underpinning DEA and SFA models).

Recently, SFA has been studied from a Bayesian point of view due to some of the attractive features of this approach such as exact inference and individual distributions of inefficiencies, formal specification of uncertainty, easy incorporation of prior ideas and restrictions, and computation of predictive distributions of inefficiency. Since the introduction of the Bayesian approach to SFA in van den Broeck et al. (1994) there have been an increasing number of theoretical studies and applications of SFA from this perspective and it is currently a very influential approach.

Regarding methodological issues, one of the most interesting topics in SFA literature is the treatment of heterogeneity and its effects on efficiency estimations. The omission of heterogeneity in stochastic frontier models has been well documented to produce biased efficiency estimations (see Greene, 2008; Kumbhakar and Lovell, 2000, for complete reviews). Empirical studies have also shown relevant implications of heterogeneity in the estimation of both efficiency levels and rankings (see Bos et al., 2009; Greene, 2004, for examples in the health and banking sectors, respectively). Also, the location of firm heterogeneity variables has been found to drive the estimations (see Coelli et al., 1999). In fact, the question of *"where to put the z's* (covariates)*?"*, as remarked by Greene (2008), is still an open issue. This issue has been widely studied before. However, unobserved inefficiency heterogeneity has been little explored.

Another topic that has received little attention within the frontier literature is the persistence of inefficiency over time. Traditionally, time dependency of inefficiency has been studied through deterministic specifications of time. In the Bayesian context, to the best of our knowledge, only the study by Tsionas (2006) considers dynamic effects in a frontier model, which also includes some exogenous variables representing heterogeneity. However, the implications of including observed covariates in these models and accounting for unobserved sources of heterogeneity have been less explored.

In this thesis, we put forward the modeling of heterogeneity in a Bayesian context by capturing both observed and unobserved heterogeneity in the inefficiency distribution under static and dynamic formulations. We propose several novel specifications which permit the identification of heterogeneity in these contexts. The first of our proposed methods captures unobserved heterogeneity in the inefficiency by modeling a random parameter in the inefficiency distribution. Results suggest that this method is successful in identifying unobserved heterogeneity and that it also can be used as a way to test the relevancy of observed covariates. Also, the location of heterogeneity is found to have important effects on efficiency estimations which are more evident when unobserved heterogeneity is accounted for. The second proposal captures unobserved heterogeneity sources related to firm-specific effects of observed covariates in the inefficiency. This is performed by modeling random coefficients in the inefficiency. It is found that allowing random coefficients for the inefficiency covariates captures firm-specific effects which remain unidentified under the regular fixed coefficients models. This specification distinguishes properly firms in term of the effects of inefficiency drivers and separates unobserved heterogeneity related to these effects from efficiency. Our third proposal relies on the framework of dynamic SFA and specifies a model that is able to capture unobserved heterogeneity in the inefficiency persistence and unobserved technological heterogeneity. Both unobserved effects are found to be very relevant in explaining inefficiency and its evolution over time. Finally, the implications of including observed covariates in dynamic models were studied by mean of an inefficiency specification that allows separating observed inefficiency heterogeneity from the dynamic process. The model allows identifying those firm characteristics that may have persistence effect in the inefficiency from those that can be rapidly adjusted. In general, location of observed covariates is found to have important implications in the identification of inefficiency drivers and posterior efficiency estimations.

The proposed models are implemented in very different applications such as health performance, airlines, banking and electricity distribution and our results have important implications for companies, regulators and policy makers in these sectors.

The inference of all the models is carried out using Bayesian methods and the WinBUGS software package is used for the implementation throughout. We provide the codes used in each chapter of the thesis at the end of the corresponding chapters.

This thesis has the following structure. Chapter 1 presents an introduction to the most important concepts on frontier efficiency, the measuring methods, SFA and its Bayesian approach, and a literature review on the treatment of observed and unobserved heterogeneity in SFA models. Chapter 2 presents the problem of observed heterogeneity in SFA by analyzing the effects of including observed covariates in the frontier, and in different parameters and distributions of the inefficiency. Chapter 3 presents the models proposed to identify unobserved heterogeneity in the inefficiency. Firstly, by modeling a random parameter in the inefficiency; and secondly, by allowing coefficients of inefficiency drivers to vary randomly across firms. Chapter 4 extends the analysis of heterogeneity in the dynamic framework by proposing two specifications: one that identifies unobserved heterogeneity in the inefficiency persistence and in the technology and another one that is able to separate observed heterogeneity from the dynamic behaviour of inefficiency. Finally, Chapter 5 discusses the main conclusions, contributions and further lines of research.

Acknowledgements

I gratefully acknowledge Helena and Mike, who were more than my supervisors during the doctorate program. They were always there for me, giving me useful advices, helping me when I needed and supporting me in every step of this process. I specially acknowledge Michael Pollitt, who was my host advisor during my visit to the University of Cambridge for all his kindness help in my research project. I want to thank Luis Orea, from University of Oviedo, who I had the fortune to know at the beginning of my program, and since then has been very special with me and have provided me special help during these years. I specially acknowledge Christopher O'Donnell from the University of Queensland, who gave me valuable and very kind advices. I thank Camilla Mastromarco, from Univeristy of Salento, for her important and useful comments. I also thank Mark Steel, from University of Warwick, for his interesting suggestions during the first stage of my thesis. I also acknowledge all the people from whom I have received valuable comments and suggestions to improve my work in the different seminars, conferences and workshops that I have attended during these years. I thank Esther Ruiz, director of Department of Statistics, for her support to assist to different conferences, courses and seminars during my doctorate program. Finally, I acknowledge all professors, colleagues and friends from University Carlos III, from whom I have learned many things and that in one or other way have helped and supported me during the process.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this Introduction we define the basic concepts and techniques involved in efficiency modeling, describe the methodology of SFA, the Bayesian approach to this method, and present a literature review on the treatment of heterogeneity in SFA.

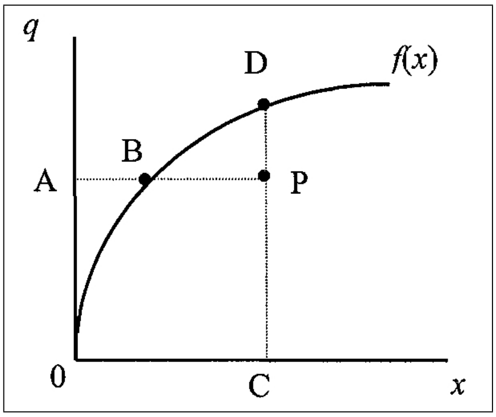
1.1 The concept of efficiency

Efficiency is a relative concept related to the way resources are used in order to obtain a final result. Therefore, it is closely linked with the performance of private or public organizations in the sense that improving efficiency can lead to higher profits, more competitive performance and better service provision. Consequently, efficiency measurement is very important for making managerial and policy decisions and for the identification of areas that can be improved.

Efficiency should be measured with respect to some objective such as maximizing produced output, revenues or earned profits, or minimizing inputs or costs. In general, three main types of efficiency can be measured. These are technical, allocative, and economic efficiency. Technical efficiency is related to the quantities of inputs and outputs employed in a production process. In this case, we can obtain either an input-oriented or an output-oriented measure. The former assesses the quantities of inputs used to produce a given output; while the latter measures the output levels produced with some given inputs. These measures are then linked to the economic concept of a production function.

To illustrate the concept of technical efficiency, Figure 1.1 plots a single-input single-output case, where the output (y) is represented in the vertical axis and the input (x) in the horizontal axis. The curve is the optimal technology representing a production frontier where every output level is produced with the minimum possible amount of input and vice versa. Therefore, point P represents a technically inefficient firm, which could either produce the same output level A using less input quantity, or produce more output with the amount of input. In the first case, the distance $\overline{BP}$ is a measure of input-oriented technical inefficiency for this firm. In the second case, output-oriented technical inefficiency is represente by the distance $\overline{DP}$.

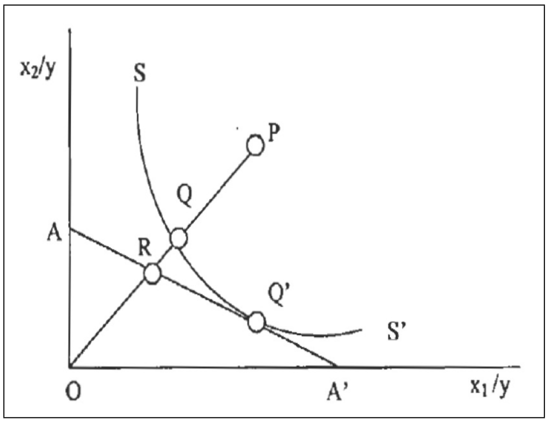
Figure 1.1: Technical efficiency



Source: Coelli et al. (2005)

Allocative efficiency measures the way a firm chooses optimum input or output levels given their prices in the market. This is associated with the ability to fulfill the marginal conditions for a cost minimization function or a revenue or profit maximization. Finally, economic efficiency is a concept involving both types of efficiency. To illustrate these three concepts, Figure 1.2 shows the possible combinations of two inputs (x_1 and x_2) for producing one output (y). Line $\overline{AA'}$ is the isocost representing all the input combinations which generate the minimum cost for the firm given the market prices of those inputs. The curve SS' is the isoquant representing the minimum feasible combination of inputs generating a given output amount. Thus, a firm located at point P is both technical and allocative inefficient, while a firm at point Q is technical efficient because it is on the isoquant. Then, the distance $\overline{PQ}$ is a measure of the input reduction the firm should perform to be technical efficient. However, at Q the firm is still allocative inefficient because given the input prices the firm may use a different combination of inputs which allows it to produce the same output amount but incurring in lower costs.Then, the distance $\overline{QR}$ is a measure of the allocative inefficiency. Therefore, only Q' is the overall economic efficient point, where the isoquant is tangent to the isocost.

Figure 1.2: Technical, allocative and economic efficiency



Source: Coelli et al. (2005)

It is important to distinguish the concepts of efficiency and productivity. Although highly related, the concept of productivity is more general and, as well as technical and allocative efficiency, it takes into account technical change, scale economies and output mix.[1] Here, we focus on technical and economic efficiency measurement, although measures of technical change and scale efficiency are derived in the applications.

1.2 Efficiency measurement

In general, measurement techniques for efficiency are separated into two main approaches: nonparametric and parametric.

The nonparametric approach has flexibility as its main advantage, but the main drawback is that, in general, it provides deterministic measures of inefficiency. Under

[1] Technical change is related to changes in the technology that may shift the frontier. Regarding scale economies, a firm operates at constant returns to scale if an increase of input amounts leads to a proportional increase of the output; at increasing returns if the output increases more than proportional; and at decreasing returns if the output increases less than proportional. As to output mix, a firm is defined as scope efficient if it produces jointly multiple products at proportional less cost than producing them separately.

this approach the main methodology used is DEA. It was introduced by Charnes et al. (1978) and the purpose is to determine which observations or decision making units (DMU's) are efficient in the performance of some activity. These DMU's have common inputs and outputs. It is important to notice that this method allows easily evaluating multiple outputs, introducing inputs different to production factors, and using ratios or other indicators.

In particular, in this method, the efficient DMUs compose a frontier enveloping all inefficient observations, and technical inefficiency is measured by their distance to the frontier. That distance is a measure of the proportion of inputs the firm might reduce or of the proportion of output the firm might increase in order to be efficient. It is important to remark that efficiency under DEA is considered as relative to the other units evaluated and then it is highly sensitive to outliers. There exist two types of envelopment surfaces, one for constant returns to scale and other for variable returns to scale. This choice is determined a priori based on the knowledge of the sector analyzed. The application of this method implies solving a linear programming problem and is underpinned by the assumption that the frontier is locally linear. Its main drawback is that if outputs, inputs and/or environmental variables are measured with error or are unobserved, as it is usually the case, then the estimators are inconsistent (see O'Donnell, 2014, for a discussion). Moreover, modeling environmental or heterogeneity variables affecting the inefficiency is troublesome under this approach.[2]

The parametric approach uses econometric methods for the estimations. Although, within this approach there are some deterministic methods, the most interesting characteristic is that it easily allows the inclusion of an stochastic error term. The most common parametric and stochastic method is SFA, introduced by Aigner et al. (1977) and Meeusen and van den Broeck (1977). As it will be described below, its main advantage is introducing an error term, which may account for measurement errors, omitted variables and functional form errors. It also allows to model, in an easy and consistent way, heterogeneity variables in a single stage (see Wang, 2002, for a discussion and experiments).

[2]Traditionally, two-stage procedures have been used, where the obtained inefficiency measures are regressed over a set of heterogeneity variables. However, this usually leads to inconsistent estimators (see Simar and Wilson, 2007, for a discussion on this issue and a possible solution using bootstrapping).

Recently, some methods that try to incorporate the advantages of both approaches have been proposed. Robust and stochastic DEA-type models as introduced in Cazals et al. (2002); Daraio and Simar (2005), and Kuosmanen and Kortelainen (2012), respectively, are able to deal with outliers, noise and include environmental variables. However, the robust versions do not allow for a rigorous analysis of an stochastic error term and, as the stochastic versions, the modeling of heterogeneity factors that affect directly the inefficiency rather than the production activity itself, is still troublesome. Other methods mixing both worlds are the Bayesian and frequentist approaches to semiparametric SFA models (see Griffin and Steel, 2004; Park et al., 2007, respectively). These methods allow more flexible structures for either the inefficiency or the frontier. However, at least in the first case, modeling heterogeneity variables in a nonparametric specification for the inefficiency component is not easy and is still an open issue.

In this thesis, we focus only on SFA given the advantages it provides for modeling an idiosyncratic error and heterogeneity variables affecting the inefficiency. The inference of all the models is carried out via Bayesian methods, which provides interesting advantages for the incorporation of uncertainty and the analysis of the results that are discussed further.

1.3 Stochastic Frontier Analysis

The SFA method is motivated by the idea that deviations from the frontier may not be entirely under the control of the firm. This approach supposes an error term which can be decomposed as the sum of two components. One component is the idiosyncratic random error assumed to be normally distributed and the second component is nonnegative and considered as inefficiency. The way the error component is divided depends on the distribution assumption of the error component capturing the inefficiency.

As mentioned before, parametric models are based on a functional form; in particular, the stochastic frontier models are usually derived from a production or a cost function. As described in Section 1.1, in the first case, technical efficiency is measured while in the second case, economic efficiency is evaluated. The former is simpler; thus, it is used to explain below the basics of SFA (see Kumbhakar and Lovell, 2000). Both, production and cost function models can be evaluated for cross sectional or panel data. The latter provide more efficient estimations, and allows introducing time varying inefficiencies.

Consider a cross section production frontier model such that:

$$y_i = f(x_{1i}, x_{2i}, \cdots, x_{ki}; \boldsymbol{\beta}) \cdot TE_i, \tag{1.1}$$

where y_i is the total output produced by firm $x_{1i}, x_{2i}, \cdots, x_{ki}$ are the k inputs used by firm i, $\boldsymbol{\beta}$ is the vector of technology parameters to be estimated, and TE_i is the technical efficiency of firm i. Thus, rearranging the previous equation we have that:

$$TE_i = \frac{y_i}{f(x_{1i}, x_{2i}, \cdots, x_{ki}; \boldsymbol{\beta})}. \tag{1.2}$$

Then, technical efficiency is defined as the ratio of the observed output to the maximum feasible output. This means that if the firm is producing at the maximum, technical efficiency is also maximum and equal to 1, otherwise it is less than 1 and some degree of inefficiency is presented. However, in this case the frontier is deterministic because it is not considering possible random shocks out of the control of the firm. To introduce this, the specification of the stochastic frontier and the technical efficiency is as follows:

$$y_i = f(x_{1i}, x_{2i}, \cdots, x_{ki}; \boldsymbol{\beta}) \cdot \exp(v_i) \cdot TE_i \tag{1.3}$$

$$TE_i = \frac{y_i}{f(x_{1i}, x_{2i}, \cdots, x_{ki}; \boldsymbol{\beta}) \cdot \exp(v_i)}, \tag{1.4}$$

where, $\exp(v_i)$ captures the effect of measurement errors or random shocks not controlled by the firm.

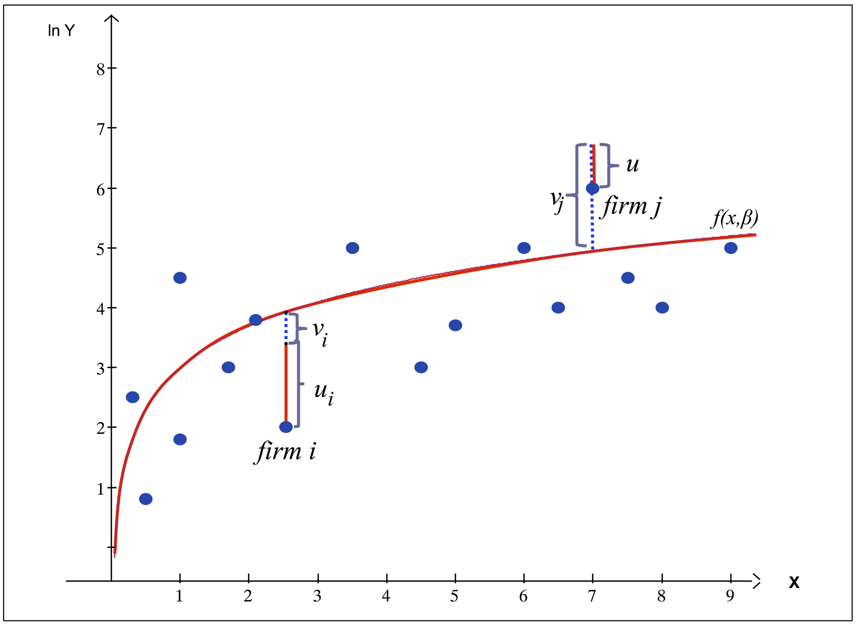
The theoretical specification of an economic production function is $y = f(L, K)$, where y is the output produced, L is the amount of labor input used to produce y, and K is the amount of capital input used to produce y. It is possible to derive this function from a Cobb-Douglas or a Translog function. Taking logarithms to any of these functions, the econometric model for the stochastic frontier would be:

$$\ln y_i = \mathbf{x_i}\boldsymbol{\beta} + \epsilon_i; \quad \epsilon_i = v_i - u_i, \tag{1.5}$$

where the matrix $\mathbf{x_i}$ contains the logarithm terms of the inputs associated to labor and capital and ϵ_i is the error term which is split into two components: v_i is an unrestricted random variable and u_i is the inefficiency component. Given that the technical efficiency cant be greater than 1, u_i is defined to be always nonnegative. Also, recalling the technical efficiency definition we have that $TE_i = \exp(-u_i)$. It is worth observing that in the case of deriving the frontier from a cost function, the sign of the inefficiency component u_i is reversed given that in that case we have a minimum frontier such as that illustrated in Figure 1.2.

Figure 1.3 shows the SFA representation of a production function where it can be observed that for firms i and j not all the distance from their observed outputs to the frontier is attributed to inefficiency but only part of it. In the first case, the idiosyncratic error term is negative while in the second case it is positive. This random part v_i is assumed to be normally distributed $N(0, \sigma_v^2)$ and fulfill the classical assumptions as the usual error under OLS. However, for the inefficiency part u_i some assumption is required given that it must be non negative. The most common alternatives proposed in the literature for the inefficiency distribution are: half normal (Aigner et al., 1977), exponential (Meeusen and van den Broeck, 1977), truncated normal (Stevenson, 1980), and gamma (Greene, 1990) distributions. In all the cases, v_i and u_i are assumed to be independently distributed from each other and from the regressors.

Figure 1.3: SFA approach for the derivation of technical efficiency



1.4 The Bayesian approach to SFA

The Bayesian approach to SFA was introduced in van den Broeck et al. (1994). The main advantages in this context are the exact inference on inefficiencies especially with small samples, the straightforward incorporation of prior ideas and restrictions, the formal specification of uncertainty on the parameters and the model, getting a distribution of the inefficiency for every firm, and the direct computation of the average inefficiencies through the predictive posterior distribution of u_i. Dealing with stochastic frontier models in this context requires using numerical integration methods such as Markov Chain Monte Carlo (MCMC). In particular, the Gibbs sampling algorithm with data augmentation as introduced in Koop et al. (1995) is used very often on the literature of Bayesian stochastic frontiers. The implementation of these models in this framework can be easily carried out through free software packages such as WinBUGS (Lunn et al., 2000). The implementation of the most common SFA models in the Bayesian context using this package is introduced in Griffin and Steel (2007).

The Bayesian formulations for the gamma, exponential and truncated normal distributions are presented by van den Broeck et al. (1994). It is important to remark that the exponential is just a particular case of the gamma distribution when the shape parameter is equal to 1, and that the half normal is a particular case of the truncated normal when the truncation is set at 0. The Bayesian approach assuming the exponential distribution for inefficiencies is presented below.

Let us represent the stochastic frontier model presented in (1.5) as:

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + v_i - u_i, \tag{1.6}$$

where y_i is the log of the output variable for the firm i, $\mathbf{x_i}$ contains the explanatory variables and v_i, u_i are the error terms as considered in (1.5).

Thus the joint distribution of y_i and u_i conditional on $\mathbf{x_i}$ and the parameters is:

$$p(y_i, u_i|\mathbf{x}_i, \boldsymbol{\theta}) = f_N(y_i|f(\mathbf{x}_i, \boldsymbol{\beta}) - u_i, \sigma^2) f_G(u_i|\lambda^{-1}), \tag{1.7}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \lambda)$ are the parameters to be estimated, σ^2 is the variance of the composed error ϵ_i and λ is the unknown scale parameter of the exponential distribution.

From here,we deduce the following conditional density for u_i:

$$p(u_i|y_i, \mathbf{x}_i, \boldsymbol{\theta}) = \Phi^{-1}\left(\frac{m_i}{\sigma}\right) f_N(u_i|m_i, \sigma^2) I(u_i \geqslant 0), \tag{1.8}$$

where $m_i = f(\mathbf{x}_i, \boldsymbol{\beta}) - y_i - \sigma^2/\lambda$. From the two previous conditional distributions the sampling density of y_i is obtained:

$$p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{\lambda{-}1}{\Gamma(1)} \exp\left(-\frac{m_i}{\lambda} - \frac{\sigma^2}{2\lambda^2}\right) \Phi\left(\frac{m_i}{\sigma}\right). \tag{1.9}$$

Now, the likelihood function can be represented as the product of the previous densities:

$$L(\boldsymbol{\theta}|y, \mathbf{x}) = \prod_{1}^{N} p(y_i|\mathbf{x}_i, \boldsymbol{\theta}). \tag{1.10}$$

Given this likelihood function and some prior density $p(\boldsymbol{\theta})$, the posterior distribution is:

$$p(\boldsymbol{\theta}|y, \mathbf{x}) \propto L(\boldsymbol{\theta}|y, \mathbf{x}) p(\boldsymbol{\theta}). \tag{1.11}$$

This distribution includes all the information about the parameters contained in the prior and the data. However, under a frontier analysis the most important are not the parameters but the individual efficiencies, which are measured by $TE_i = \exp(-u_i)$. Therefore, the conditional posterior distributions of u_i and r_i are the following:

$$p(u_i|y, \mathbf{x}) = \int p(u_i|y, \mathbf{x}_i, \boldsymbol{\theta}) p(\boldsymbol{\theta}|y, \mathbf{x}) d\boldsymbol{\theta} \tag{1.12}$$

$$p(r_i|y, \mathbf{x}) = TE_i^{-1} \int p(u_i|y, \mathbf{x}). \tag{1.13}$$

The main difference from the frequentist method is that instead of conditioning over an estimate of $\boldsymbol{\theta}$, the Bayesian approach averages out the uncertainty about $\boldsymbol{\theta}$ in a natural way by marginalizing with respect to the posterior density of the parameters. Finally, in frontier analysis the average inefficiency is usually examined. An advantage of the Bayesian context is that it can be obtained directly through the predictive posterior distribution of u_i, which has the following formulation:

$$p(u_f|y, \mathbf{x}) = \int f_G(u_f|\lambda^{-1})p(\lambda|y, \mathbf{x})d\lambda. \tag{1.14}$$

1.5 Heterogeneity in Stochastic Frontier Models

In stochastic frontier models, the estimated inefficiency component often includes some firm characteristics other than outputs, inputs, or prices defined from the production or cost function, which should not be attributed to inefficiency. These are exogenous variables (e.g. type of ownership, GDP level in the country of operation) that have an effect on the technology used by firms or directly on their inefficiency. If these variables are not taken into account in the model specification, this may affect the estimation of the inefficiencies or of the frontier significantly.

Firm characteristics can be modeled in the frontier if they imply heterogenous technologies or in the one-sided error component if they affect the inefficiency. In the former case, covariates are directly included in the functional form and the main interest is to model unobserved heterogeneity (see Greene, 2005). In the case of heterogeneity in the inefficiency, covariates are usually included in the parameters of the one-sided error distribution (see Huang and Liu, 1994).

Heterogeneity in stochastic frontier models has also been studied in the Bayesian context. The Bayesian approach to stochastic frontiers introduced by van den Broeck et al. (1994) presents advantages in terms of formally deriving posterior densities for individual efficiencies, incorporating economic restrictions, and in the easy modeling of random parameters through hierarchical structures. Hierarchical models have been used to capture heterogeneous technologies (see Tsionas, 2002) and heterogeneity in the inefficiency has been considered through covariates in the distribution of the non-negative error component (see Koop et al., 1997). Modeling observed heterogeneity using non parametric and flexible mixtures of inefficiency distributions are other interesting recent contributions (see Griffin and Steel, 2004, 2008).

On the other hand, unobserved heterogeneity in the non-negative error component has been very little explored in the literature from a frequentist or a Bayesian approach. However, ignoring its existence means that heterogeneity which is not captured by observed covariates is wrongly attributed to inefficiency and consequently leads to bad efficiency estimates.

1.5.1 Literature on Observed Heterogeneity

The original stochastic frontier model introduced by Aigner et al. (1977) and Meeusen and van den Broeck (1977) has the following form:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + v_{it} - u_{it} \tag{1.15}$$

where y_{it} represents the output of firm i at time t, $\mathbf{x}_{it}$ is a vector that contains the input quantities used in the production process, v_{it} is an idiosyncratic error that is typically assumed to follow a normal distribution and u_{it} is the one-sided component representing the inefficiency and follows some non-negative distribution.

Firm specific heterogeneity not specified in (1.15) can be mistaken for inefficiency if it is not identified. Heterogeneity can either shift the efficiency frontier or change the location and scale of the inefficiency estimations (see Greene, 2008; Kumbhakar and Lovell, 2000, for complete reviews). In general, when external factors are supposed to capture technological differences and these are out of the firms' control, heterogeneity should be specified in the frontier. In this case, the main interest is capturing unobserved effects. In the classical context, this has been modeled through fixed and random effects or models with random parameters (see Greene, 2005). Bayesian approaches have been based on frontier models with hierarchical structures (see Huang, 2004; Tsionas, 2002).

When heterogeneity is more related to efficiency and thus more likely to be under firms' control, then this should affect directly the one-sided error term. In the parametric context, inefficiency heterogeneity is often included in the location or scale parameters of the inefficiency distribution. For example, covariates shift the underlying mean of inefficiency in Kumbhakar et al. (1991), Huang and Liu (1994) and Battese and Coelli (1995). A reduced form of these models assumes that the location parameter of the distribution of u_{it} depends on vectors of covariates $\mathbf{z}_{it}$ and parameters $\boldsymbol{\delta}$ as follows:

$$\begin{aligned} u_{it} &\sim N^{+}(\mu_{it}, \sigma_u^2) \\ \mu_{it} &= \mathbf{z}_{it}\boldsymbol{\delta}. \end{aligned} \tag{1.16}$$

The scale parameter of the one-sided error component has also been modeled as a function of firm characteristics. Reifschnieder and Stevenson (1991) provided one of the first linear specifications where this parameter varies across firms. A similar model was proposed by Caudill et al. (1995) with the aim of treating heteroscedasticity in frontier models. These authors found biased inefficiency estimations when heteroscedasticity was not accounted for.[3] The proposed model specifies the variance of a half-normal distributed inefficiency as an exponential function of time invariant covariates:

$$\begin{aligned} &u_i \sim N^+(0, \sigma_{u_i}^2) \\ &\sigma_{u_i} = \sigma_u \cdot \exp(\mathbf{z}_i \gamma). \end{aligned} \tag{1.17}$$

Although the original proposal in (1.17) was presented in a cross sectional framework, it can be easily extended to include time-varying covariates and inefficiencies (see Hadri et al., 2003a,b, for an extension to panel data). It is also possible to define $u_{it} = u_i \cdot g(t)$ where $g(t)$ is a function of time (e.g. the parametric function introduced by Battese and Coelli, 1992). The specification in (1.17) has the characteristic of changing the scale of the inefficiency distribution while preserving its shape and is referred in the literature as the scaling property (see Alvarez et al., 2006; Wang and Schmidt, 2002). In general, this property allows us to think about inefficiency as being composed of two parts: $u_{it} = u_{it}^* \cdot f(\mathbf{z}_{it}, \boldsymbol{\delta})$. The first component is a base inefficiency, which is not affected by firm characteristics and captures random managerial skills, while the second component is a function of heterogeneity variables determining how well management is performed under these conditions. Another important feature of this property is that the interpretation of the effects of covariates on the inefficiency is direct and independent of the inefficiency distribution. The scaling property also holds when the inefficiency is exponentially distributed (see Simar et al., 1994), or in a particular case of truncated normal inefficiency where both parameters are an exponential function of firm characteristics as follows (see Alvarez et al., 2006; Wang and Schmidt, 2002):

$$\begin{aligned} &u_{it} \sim N^+(\mu_{it}, \sigma_{u_{it}}^2) \\ &\mu_{it} = \mu \cdot exp(\mathbf{z}_{it} \boldsymbol{\delta}) \\ &\sigma_{u_{it}} = \sigma_u \cdot \exp(\mathbf{z}_{it} \boldsymbol{\delta}). \end{aligned} \tag{1.18}$$

Specification (1.18) for the inefficiency is a variation of a previous proposal by Wang (2002) where both the mean and the variance of truncated normal inefficiencies are simultaneously affected by the same covariates but with different coefficients. Other authors have also proposed heterogeneity specifications that include firm characteristics

[3] In a previous study, Caudill and Ford (1993) also found biased estimates of the frontier parameters.

in the variance of the idiosyncratic error with the aim of treating heteroscedasticity in frontier models (see Hadri, 1999).

In the Bayesian context, Koop et al. (1997) presented different structures for the mean of the inefficiency component as Bayesian counterparts to the classical fixed and random effects models. One of these specifications is the varying efficiency distribution model, which includes firm specific covariates in the parameter of an exponential distribution. These covariates link the firm effects and only the inefficiencies of firms sharing common characteristics are drawn from the same distribution. The distribution below presents a time invariant inefficiency that depends on vectors of binary covariates $\mathbf{z_i}$ and parameters $\boldsymbol{\gamma}$:

$$\begin{aligned} u_i &\sim Ex(\lambda_i^{-1}) \\ \lambda_i &= \exp(\mathbf{z}_i\boldsymbol{\gamma}). \end{aligned} \tag{1.19}$$

Since this model is intended to be a counterpart of a frequentist random effects model, it is specified to obtain time invariant inefficiencies. However, as in the case of (1.17), it is possible to define $u_{it} = u_i \cdot g(t)$ or to include time-varying covariates. Also, it would be possible to draw inefficiencies for every firm and period of time from the distribution with a firm specific parameter.

1.5.2 Literature on Unobserved Heterogeneity

Unobserved heterogeneity in SFA has been mainly modeled in the frontier, recognizing the existence of unobserved sources related to heterogeneous technologies among firms. From the frequentist approach, the most common way to include these effects is through panel data models with fixed and random effects. In particular, Greene (2005) proposes two models able to capture these effects and distinguish them from inefficiency: the True Fixed Effects (TFE) and the True Random Effects (TRE) models. This author also proposes a Random Parameters (RP) model where all the coefficients are allowed to be firm specific. The specifications for these models are the following:

$$y_{it} = \alpha_i + \boldsymbol{\beta}'\mathbf{x}_{it} + v_{it} \pm u_{it} \tag{1.20}$$

$$y_{it} = (\alpha + w_i) + \boldsymbol{\beta}'\mathbf{x}_{it} + v_{it} \pm u_{it} \tag{1.21}$$

$$y_{it} = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_{it} + v_{it} \pm u_{it}. \tag{1.22}$$

Equation (1.20) is the TFE specification, which can be solved by creating the dummy variables for every firm. As Greene (2005) states the problem of the possible presence of many parameters to be estimated could be solved through the use of Newton's method. On the other hand, (1.21) is the TRE specification, which is a regular random effects model where the composed error $\epsilon_{it} = v_{it} \pm u_{it}$ follows an asymmetric distribution. Finally, (1.22) presents the RP specification where $(\alpha_i, \boldsymbol{\beta}_i) = (\overline{\alpha}, \overline{\boldsymbol{\beta}}) + \Delta_{\alpha,\boldsymbol{\beta}}\mathbf{q}_i + \Gamma_{\alpha,\boldsymbol{\beta}}\mathbf{w}_{\alpha,\boldsymbol{\beta}_\mathbf{i}}$. Here, Δ_j is a matrix of parameters to be estimated, $\mathbf{q}_i$ include related variables, Γ_j is the covariance matrix and $\mathbf{w}_{j,i}$ is a random vector parameterizing random variation.

In the Bayesian context, random effects models have also been proposed to model unobserved technological heterogeneity. Tsionas (2002) proposed a model with the same form in (1.22), where hierarchical structures can be easily defined in the frontier parameters. Therefore, in this case we would define the following structure of prior and hyperprior distributions:

$$\begin{aligned} p(\boldsymbol{\beta}) = &\ f_N(\bar{\boldsymbol{\beta}}, \boldsymbol{\sigma}^2_{\boldsymbol{\beta}}) \\ p(\bar{\boldsymbol{\beta}}) = &\ f_N(0, \boldsymbol{\Sigma}_{\bar{\boldsymbol{\beta}}}) \\ p(\boldsymbol{\sigma}^{-2}_{\boldsymbol{\beta}}) = &\ f_G(a_{\boldsymbol{\sigma}^{-2}_{\boldsymbol{\beta}}}, b_{\boldsymbol{\sigma}^{-2}_{\boldsymbol{\beta}}}), \end{aligned} \tag{1.23}$$

where, $\bar{\boldsymbol{\beta}}$ is the common mean from where the firm specific variables vary randomly with variance equal to $\boldsymbol{\sigma}^2_{\boldsymbol{\beta}}$. Similarly it is defined for α.

As previously mentioned, the case of unobserved heterogeneity in the inefficiency has been less explored. The RP model introduced by Greene (2005) allows us to account for random parameters in the inefficiency component. This model defines a truncated normal distribution for the one-sided error such that:

$$\begin{aligned} u_{it} \sim &\ N^+(\mu_i, \sigma^2_{u_i}) \\ \mu_i = &\ \boldsymbol{\mu}_i'\mathbf{z}_i; \sigma_{u_i} = \sigma_u \exp(\theta_i'\mathbf{h}_i) \\ \boldsymbol{\mu}_i = &\ \overline{\boldsymbol{\mu}} + \Delta_{\boldsymbol{\mu}}\mathbf{q}_i + \Gamma_{\boldsymbol{\mu}}\mathbf{w}_{\boldsymbol{\mu}_\mathbf{i}} \\ \boldsymbol{\theta}_i = &\ \overline{\boldsymbol{\theta}} + \Delta_{\boldsymbol{\theta}}\mathbf{q}_i + \Gamma_{\boldsymbol{\theta}}\mathbf{w}_{\boldsymbol{\theta}_\mathbf{i}} \end{aligned} \tag{1.24}$$

In the Bayesian framework, although not presented as a model to capture unobserved heterogeneity, Koop et al. (1997) propose a marginal independent effects (MIED) model that may capture in some extent unobserved sources in the inefficiency. In this case, the inefficiency is assumed to be exponentially distributed with firm specific mean and independent priors as follows:

$$\begin{aligned} u_{it} \sim & \quad Exp(\lambda_i) \\ p(\lambda_i) = & \quad f_G(1, -\ln(r^*)) \end{aligned} \tag{1.25}$$

Chapter 2

Observed Inefficiency Heterogeneity in Stochastic Frontier Models*

This chapter focuses on the treatment of observed heterogeneity related to inefficiency. Several models have been proposed to deal with this problem using different distributions for the inefficiency component. In general, accounting for observed heterogeneity in the inefficiency has been found to have important consequences in efficiency estimations. Most of these studies have described how these estimations may present biases when relevant information about firm characteristics is omitted. However, the consequences that using a particular heterogeneity SFA model has on the efficiency estimations and rankings provided is still an interesting topic of study. Its understanding is key in this area given that the final aim in most empirical applications is precisely obtaining efficiency scores and comparing firms through efficiency rankings. This chapter analyzes these effects under a Bayesian framework when observed covariates are included in different parameters of the inefficiency distribution and also when different distributions are used for the one-sided error term.

This chapter is divided into five sections. Section 2.1 presents a brief literature review on the treatment of observed heterogeneity in stochastic frontier models from both the frequentist and the Bayesian approach. Section 2.2 presents a general SFA model which

*Much of the work in this chapter has been published in the *Journal of Productivity Analysis* (see Galán et al., 2014b).

allows to include observed covariates in different parameters of the inefficiency distribution. Bayesian inference for this model and three different model selection criteria are presented. Section 2.3 presents a detailed analysis of the effects of including observed covariates in different parameters of the inefficiency distribution on efficiency estimations and rankings. This is illustrated using two real samples previously studied under the frequentist approach. In Section 2.4, the effects on the estimations when different distributions are used for modeling observed inefficiency heterogeneity are presented. Finally, Section 2.5 concludes the chapter.

2.1 Analysis of observed inefficiency heterogeneity

In this section, a general stochastic frontier model for panel data that allows the inclusion of observed inefficiency heterogeneity is presented. The aim is to study the effects on the estimations of including observed covariates in the different parameters of the inefficiency distribution.

For the evaluation of the effects on different parameters of the inefficiency distribution, we focus on a truncated normal distribution, which is one of the most used distributions in studies involving observed heterogeneity in the inefficiency. In particular, covariates are often included in the location parameter of this distribution following the Battese and Coelli (1995) model. However, it is not clear in which parameter of the inefficiency distribution heterogeneity should be included. Wang (2002) proposed modeling the covariates simultaneously in the location and scale parameters of the truncated distribution. Alvarez et al. (2006) analyze a particular specification of truncated normal distributed inefficiencies that has the property of preserving the shape while changing the scale of the inefficiency, and also estimate a model where heterogeneity is captured only by the scale parameter of this distribution. We think that at an individual level, the moments of the distributions affected have different effects on the posterior efficiency distributions of each firm. Since this is possible to be studied from a Bayesian context, our aim here is to analyze the effects on the posterior efficiency distributions of including observed heterogeneity in the location, scale or both parameters of the truncated normal distribution. For the latter case, we extend to the Bayesian framework the scaling property model proposed by Alvarez et al. (2006). This allows us to think of the inefficiency as being composed of two parts, one component capturing natural managerial skills and other component which depends on observed firm characteristics.

For the one-sided error we use an exponential specification of a truncated normal distribution where the location, scale, or both parameters can model firm heterogeneity. The general model in the case of a production function is:

$$\begin{aligned} y_{it} = &\quad \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{z^*}_{it}\boldsymbol{\delta} + v_{it} - u_{it};\ v_{it} \sim N(0, \sigma_v^2) \\ u_{it} \sim &\quad N^+(\mu \cdot \exp(\mathbf{z}_{it}\gamma I_1), \sigma_u^2 \cdot (\exp(\mathbf{z}_{it}\gamma I_2))^2), \end{aligned} \tag{2.1}$$

where y_{it} is the output of firm i at time t, $\mathbf{x}_{it}$ is the row vector of input quantities, $\mathbf{z^*}_{it}$ is a row vector of the observed heterogeneity variables that affect the technology; $\mathbf{z}_{it}$ is a row vector of observed covariates with effects in the inefficiency; and, $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, and $\boldsymbol{\gamma}$ are the corresponding parameter column vectors. I_1 and I_2 are indicator variables taking the value of 1 when either observed covariates are accounted for in the location or scale parameters, respectively, and 0 otherwise.

This model nests other specifications in the literature that capture only observed heterogeneity. When I_2 is equal to 0, the model reduces to an exponential specification of the Battese and Coelli (1995) model in (1.16). If I_1 is equal to 0, the model allows only the scale parameter to include heterogeneity. This specification has only been studied before by Alvarez et al. (2006) in the framework of testing the scaling property. If additionally the location parameter μ is set to zero, our model becomes an extension of the half-normal model proposed by Caudill et al. (1995) in (1.17). Finally, if both parameters are allowed to include simultaneously the same type of heterogeneity ($I_1, I_2 = 1$) our proposal becomes an extension of the scaled Stevenson model in (1.18). In case heterogeneity is considered time invariant, the vector of observed covariates $\mathbf{z_{it}}$ can be set to vary only across firms.

It is easy to extend this specification to a hierarchical model which also allows for additional, unobserved, firm effects in the technology. However, in practical applications, mean posterior efficiencies are found to be very close to 1 for almost all firms (see Huang, 2004; Tsionas, 2002, for similar results). From our point of view, these results are inconclusive as they do not allow us to get reliable efficiency rankings.

2.1.1 Bayesian inference

All the models derived from the general specification in (3.1) are fitted by Bayesian methods. In order to do this, we first need to introduce prior distributions for the model parameters. We assume proper but relatively disperse prior distributions throughout. In particular, the distributions assumed for the parameters in the frontier function are as follows: $\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, $\boldsymbol{\delta} \sim N(\mathbf{0}, \Sigma_{\boldsymbol{\delta}})$ with diffuse, inverse gamma priors for the variances. Finally, the variance of the idiosyncratic error term is inverse gamma, that is equivalent to $\sigma_v^{-2} \sim G(a_{\sigma_v^{-2}}, b_{\sigma_v^{-2}})$ with low values for the shape and scale parameters.

Regarding observed inefficiency heterogeneity, the distribution of the one-sided error component for the truncated normal model is: $u_{it}|\boldsymbol{\gamma}, \mathbf{z}_{it} \sim N^{+}(\mu \cdot \exp(\mathbf{z}_{it}\boldsymbol{\gamma}), \sigma_u^2 \cdot (\exp(\mathbf{z}_{it}\boldsymbol{\gamma}))^2)$, where $\mu = \psi/\sqrt{\sigma_u^{-2}}$ with $\psi \sim N(0,1)$ and $\sigma_u^{-2} \sim G(5, 5 \cdot \ln^2 r^*)$ which centres efficiency at r^* (see Griffin and Steel, 2007, for this implementation). When models include heterogeneity in the inefficiency $\boldsymbol{\gamma}$ is $N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$ distributed with a diffuse prior for the covariance matrix.[1]

The complexity of these models makes it necessary to use numerical integration methods such as Markov Chain Monte Carlo (MCMC), and in particular the Gibbs sampling algorithm with data augmentation as introduced by Koop et al. (1995). For our models, implementation was carried out using the WinBUGS package following the general procedure outlined in Griffin and Steel (2007). The MCMC algorithm involved 50,000 iterations where the first 10,000 were discarded in a burn-in phase. We present in the Appendix of this chapter an example of the code used for the implementation of this model. Sensitivity analysis of our results to changes in prior parameters and values for r^* was also carried out. Results showed that the posterior inference was relatively insensitive to small changes in these parameters.

2.1.2 Model selection

The different models are evaluated in terms of three criteria, the DIC_3, which is a variant of the Deviance Information Criterion (DIC), the Log Predictive Score (LPS) and the Mean Square Error (MSE) of predictions.

The standard choice for comparing competing models in Bayesian statistics is to use the Bayes factor, that is the ratio of the posterior odds to the prior odds in favour of

[1] Griffin and Steel (2007) use priors for $\boldsymbol{\gamma}$ similar to that used for μ here. In our exercises we found that using either alternative lead to roughly the same posterior results.

the first model. However, the accurate calculation of the Bayes factor is very difficult in complex models which need MCMC techniques for parameter estimation such as those we examine here. Therefore, we prefer to use an alternative Bayesian model choice criterion based on the DIC_3. This is a variant of the DIC which is a within sample measure of fit introduced by Spiegelhalter et al. (2002) commonly used in Bayesian analysis.

Defining the deviance of a model with parameters θ as $D(\theta) = -2\log f(\mathbf{y}|\theta)$, where $\mathbf{y}$ are the data, then $DIC = 2\overline{D(\theta)} - D(\bar{\theta})$ where $\bar{\theta}$ represent some mean posterior parameter estimates. However, the DIC is well known to possess a number of stability problems in certain cases such as random effects models and mixture models (see Celeux et al., 2006). In particular, we can note here that the representation we use for the parameters of the inefficiency term is a type of random effects model in the cases where we include an unobserved heterogeneity term. Furthermore and more recently, Li et al. (2012) also remark on the lack of robustness of the original DIC in models with data augmentation such as those we examine here. For such cases, Celeux et al. (2006) recommend the use of the DIC_3 criterion as one of the best choices among various alternatives to the DIC. The formulation for this criterion is:

$$DIC_3 = -4E_\theta[\log f(\mathbf{y}|\theta)|\mathbf{y}] + 2\ \log \widehat{f}(\mathbf{y}).$$

This criterion is based on the expected deviance and an estimate of the predictive density $\hat{f}(\cdot)$ which are both easy and stable to calculate from the MCMC output provided by WinBugs.

We also compare the models in terms of their predictive performance. In order to do this, we calculate the LPS and the MSE of predictions. The LPS is a proper scoring rule developed in Good (1952) that assesses the post-sample behaviour of the models associated with the Kullback-Leibler divergence between the actual sampling density and the predictive density (see Ferreira and Steel, 2007; Griffin and Steel, 2004, for previous applications of LPS in stochastic frontier models).[2] In general, LPS examines how well a model performs when its implied predictive distribution is compared with observations not used in the inference sample. The procedure consists of partitioning the sample into two sets. The first, is a training data set used to fit the model and the second is a prediction set used to evaluate the predictive performance of the first

[2]More details on this criterion and an approximate lower bound for the LPS are described in Fernandez et al. (2001).

set. In our implementation for the panel data models, the training data set contains the observations up to the penultimate time period at which data are observed for each firm. Then, if t_i represents the index of the last time point when data are observed for firm i, the predictive set contains the set of observations y_{1,t_1} to y_{k,t_k} for the k firms in the sample. The average of the log predictive density functions evaluated at observed out-of-sample values are calculated and the formulation is the following:

$$LPS = \frac{-1}{k} \sum_{i=1}^{k} \log f(y_{i,t_i} | \text{previous data}).$$

Finally, the calculation of the predictive MSE involves again the partition of the sample into two parts as earlier. The models are fitted using the training sample and their estimated parameters are used to predict the data for the last observation of every firm. The MSE is calculated as follows:

$$MSE = \frac{1}{k} \sum_{i=1}^{k} \left(y_{i,t_i} - E\left[(\beta' x_{i,t_i} - \overline{u}_{i,t_i}) | \text{previous data } \right] \right)^2,$$

where k is the number of firms as earlier and $\overline{u}_{i,t_i}$ is the mean of the inefficiency component, which is different depending on the distribution and varies with the firm for models with heterogeneity in the inefficiency.

2.2 Empirical applications

In this section, we analyze two data sets, estimate the model presented in (3.1) and interpret the results.

2.2.1 Application to WHO data set

Evans et al. (2000) estimated the technical efficiency of 191 countries in the provision of health by using a classical fixed effects stochastic frontier model for an unbalanced panel. The original data set covers 5 years from 1993 to 1997 and the production function model proposed was the following:

$$\ln(DALE_{it}) = \alpha_i + \beta_1 \ln(HExp_{it}) + \beta_2 \ln(Educ_{it}) + \beta_3 \frac{1}{2} \ln^2(Educ_{it}) + v_{it},$$

where $DALE$ is the disability adjusted life expectancy, a measure that considers mortality and illness and represents health output. Input amounts are measured by $HExp$ and $Educ$, which are health expenditure and the average years of education, respectively.

Their results were reported by the WHO and suffered from several criticisms since the authors did not consider the effects of heterogeneity in their study, even though the sample included countries with very different characteristics such as Switzerland, China, or Zimbabwe. This led to unexpected country health system performance rankings.

Greene (2004) proposed to capture differences among countries in this sample by including eight exogenous variables: $Tropics$, $PopDen$, $GEff$, $Voice$, $Gini$, GDP, $PubFin$, and $OECD$. $Tropics$ is a binary variable that takes the value 1 if the country is located in the tropic and 0 otherwise. This is out of the control of the countries and distinguishes them by the type of diseases found in this region. $PopDen$ is the country population density, which may capture effects of dispersion but also congestion in the provision of health. These two variables are characteristics of the health provision in each country and then they are included as covariates in the production function following Greene (2004). Regarding the other variables, $GEff$ is an indicator of government efficiency; $Voice$ is a measure of political democratization and freedom; $Gini$ is the income inequality coefficient; GDP is the per capita country gross domestic product; $PubFin$ is the proportion of health care financed with public resources, and $OECD$ is a binary variable that takes the value 1 if the country belongs to the organization and 0 otherwise. These variables are policy related and more likely to be drivers of the efficiency in the sense that income, inequality and government characteristics may affect the way health services are managed. However, in this field there is no theory on where these variables should be placed at (see Greene, 2004).[3]

For this application the general model is:

$$\begin{aligned} \ln(DALE_{it}) = \quad & \alpha + \beta_1 \ln(HExp_{it}) + \beta_2 \ln(Educ_{it}) + \beta_3 \tfrac{1}{2} \ln^2(Educ_{it}) + \beta_4 Tropics_i \\ & + \beta_5 \ln(PopDen_i) + \mathbf{z}_i \boldsymbol{\delta} + v_{it} - u_{it};\ v_{it} \sim N(0, \sigma_v^2) \\ & u_{it} \sim N^+(\mu \cdot exp(\mathbf{z}_i \boldsymbol{\gamma} I_1), \sigma_u^2 \cdot (exp(\mathbf{z}_i \boldsymbol{\gamma} I_2))^2). \end{aligned} \tag{2.2}$$

We estimate five different models. Model I is the heterogeneity free base model where I_1, I_2, and $\boldsymbol{\delta}$ are also equal to zero. Model II includes the covariates in the frontier

[3] After performing some tests Greene (2004) chose a model that includes Gini and GDP in the inefficiency and the rest of covariates in the production function.

as technology heterogeneity variables but not in the inefficiency ($I_1, I_2 = 0$). Models III to V consider observed heterogeneity in the inefficiency distribution and not in the production function ($\boldsymbol{\delta} = \mathbf{0}$). In particular, Model III does it only through the location parameter, that is $I_1 = 1$ and $I_2 = 0$. Model IV includes the observed covariates through the scale parameter ($I_1 = 0$,$I_2 = 1$). Finally, Model V preserves the scaling property since both parameters of the inefficiency distribution includes the same covariates and coefficients ($I_1, I_2 = 1$).

Table 2.1 reports the estimation results. They show that models considering observed heterogeneity improve from the base model in terms of fit and predictive performance. In particular, models including heterogeneity in the inefficiency distribution exhibit the lowest values for the three model comparison criteria. This suggests that covariates in $\mathbf{z}_i$ are inefficiency related. Regarding the estimated frontier coefficients, we observe decreasing returns to scale in health provision for all models and countries. This implies that efforts of countries in terms of increasing health expenditure or education are reflected in less than proportional life expectancy improvements. Results for the inefficiency covariates suggest that higher equality, income, government efficiency or pertaining to the OECD increase the efficiency of health provision. However, higher levels of democracy and public finance of health services lead to lower efficiency.

Focusing on models III to V which are those including inefficiency heterogeneity, we observe that the best fit and predictive performance is obtained by the scaling property model (Model V). Results for the predictive efficiency distribution suggest that including covariates in the location parameter of the inefficiency increases its mean, while including them in the scale parameter decreases its dispersion. In particular, the scaling property model which includes covariates in both parameters of the one-sided error distribution presents the highest mean and the lowest dispersion of the predictive efficiency distribution among all models.

The most clarifying insights come from the efficiency rankings since they allow country comparisons. Figure 2.1 shows efficiency rankings' scatter plots comparing the base model against the other four models. For Model II, which includes the covariates in the frontier, most countries preserve a similar position except for small changes in the middle rankings. Spearman's rank correlation with the base model is 0.92. In contrast, models III to V differ widely from the base model in the top and middle positions and

Table 2.1: Posterior mean and S.D. of the parameter distributions

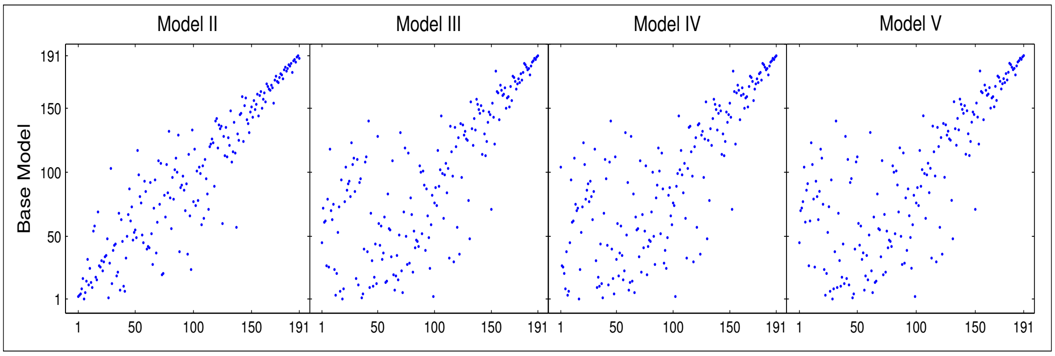
Parameter	**Model I**		**Model II**		**Model III**		**Model IV**		**Model V**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Frontier										
α	3.574	0.585	3.479	0.512	3.844	0.627	3.711	0.608	3.769	0.596
β_1	0.061	0.021	0.026	0.013	0.024	0.011	0.064	0.029	0.041	0.018
β_2	0.226	0.085	0.236	0.090	0.249	0.795	0.248	0.081	0.160	0.077
β_3	-0.039	0.012	-0.049	0.017	-0.061	0.019	-0.046	0.014	-0.033	0.011
β_4	-0.017	0.008	-0.014	0.007	-0.005	0.002	-0.043	0.002	-0.009	0.004
β_5	0.001	0.001	-0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001
$\delta_1(Gini)$			-0.147	0.007						
$\delta_2(\ln GDP)$			0.062	0.029						
$\delta_3(GEff)$			-0.014	0.010						
$\delta_4(Voice)$			0.018	0.011						
$\delta_5(OECD)$			-0.026	0.018						
$\delta_6(\ln PubFin)$			-0.036	0.021						
Inefficiency										
$\gamma_1(Gini)$					3.779	0.527	8.212	1.025	5.054	0.960
$\gamma_2(\ln GDP)$					-0.266	0.074	-0.279	0.061	-0.662	0.119
$\gamma_3(GEff)$					-0.043	0.014	-0.132	0.046	-0.054	0.017
$\gamma_4(Voice)$					0.077	0.032	0.159	0.070	0.030	0.017
$\gamma_5(OECD)$					-0.092	0.046	-3.389	1.603	-1.049	0.518
$\gamma_6(\ln PubFin)$					0.062	0.030	0.376	0.192	0.076	0.037
μ	-1.584	0.348	-1.411	0.253	-0.620	0.214	-1.423	0.288	-0.372	0.146
σ_u^2	0.238	0.054	0.214	0.059	0.406	0.098	0.054	0.017	0.058	0.019
Pred. eff.	0.878	0.104	0.877	0.103	0.908	0.138	0.785	0.081	0.914	0.072
DIC_3	-2517.282		-2809.581		-3015.727		-2989.307		-3094.403	
LPS	-122.890		-130.452		-180.507		-169.215		-185.983	
MSE	0.139		0.105		0.103		0.093		0.087	

the Spearman's rank correlations with the base model are 0.76, 0.77 and 0.75, respectively.[4] However, badly performing countries are always roughly the same regardless of the model used. This latter group is composed mainly of central African countries (e.g. Zambia, Botswana, Zimbabwe), which share some characteristics related to low income, tropical diseases, etc.

In order to observe in detail the changes that occur in the top ranked countries under the different models, Table 2.2 shows the top 20 most efficient countries under all five models. Although there are differences, the ranking is quite stable when we consider the first two models. They include countries such as Oman, Yemen and Cape Verde and other developing countries from Middle East, Asia, North of Africa and Latin America in the top positions. However, this changes completely when observed heterogeneity affects the inefficiency. In models III to V, developed countries rank in the first positions, as might

[4] Among models with inefficiency heterogeneity, rank correlation is very high (0.99).

Figure 2.1: Efficiency rankings - Base model vs. heterogeneity models



be intuitively expected, and for the scaling property model all top 20 countries are from this group. Differences are important compared to the base model. For example, Japan, Norway and Sweden which are the top 3 countries under Model V, rank in positions 45, 70 and 72, respectively, under the base model.

Using a scaling property model with heterogeneity in both parameters of the inefficiency distribution has an important effect over the ranking. Figure 2.2 shows that while most of the African countries continue to exhibit low efficiency; there is a significant change in the positions of the top and middle ranked observations. The best performing countries (developed countries) are very sensitive to the inclusion of relevant covariates such as income and inequality that distinguish them from developing countries.

The main evidence is that models including inefficiency heterogeneity lead to important moves and shrinkages of the individual posterior efficiency distributions changing the estimated mean efficiency scores and rankings. Figure 2.3 shows 90% probability intervals of efficiencies for some selected countries. It can be seen that when covariates affect the location parameter (Model III), the gap between the worst and the best performing countries increases, which leads to a separating effect on the posterior distributions. On the other hand, the intervals are narrower when the observed heterogeneity affects the scale parameter of the inefficiency (Model IV), which implies that estimation uncertainty diminishes. For the scaling property model (Model V) both effects are observed. This leads to less dispersion and overlapping of posterior efficiency distributions, which allow for more reliable conclusions about efficiency scores and rankings.[5]

[5]Similar results were obtained from other scaling-type models following half-normal and exponential distributions but they performed a bit worse in terms of fit and predictive performance.

Table 2.2: Top 20 most efficient countries

Model I	Model II	Model III	Model IV	Model V
1. Oman	1. Yemen	1. Japan	1. Luxembourg	1. Japan
2. Solomon Islands	2. Jamaica	2. Sweden	2. Spain	2. Norway
3. Yemen	3. Morocco	3. Italy	3. Greece	3. Sweden
4. Jamaica	4. Armenia	4. France	4. Malta	4. Austria
5. Morocco	5. Turkey	5. Spain	5. Armenia	5. Luxembourg
6. Cape Verde	6. Oman	6. Iceland	6. Cyprus	6. Italy
7. Georgia	7. Cape Verde	7. Greece	7. Jamaica	7. Belgium
8. Indonesia	8. Honduras	8. Germany	8. Georgia	8. Finland
9. Armenia	9. Cuba	9. Norway	9. Japan	9. Spain
10.Sri Lanka	10.China	10.United Kingdom	10.Slovakia	10.France
11.Venezuela	11.Nicaragua	11.Ireland	11.Italy	11.Denmark
12.China	12.El Salvador	12.Singapore	12.France	12.Switzerland
13.Saudi Arabia	13.Sri Lanka	13.Jamaica	13.New Zealand	13.Iceland
14.El Salvador	14.Moldova	14.Malta	14.Ireland	14.Greece
15.Honduras	15.Mexico	15.Portugal	15.Norway	15.Canada
16.Azerbaijan	16.Costa Rica	16.Czech Republic	16.Sweden	16.Netherlands
17.Turkey	17.Azerbaijan	17.Georgia	17.Oman	17.United Kingdom
18.Costa Rica	18.Colombia	18.Slovakia	18.Singapore	18.Australia
19.Dominican Rep.	19.Spain	19.Oman	19.Portugal	19.Germany
20.Egypt	20.Greece	20.Armenia	20.Czech Republic	20.New Zealand

As mentioned previously, one of the advantages of preserving the scaling property is the decomposition of the one-sided error term into a base and a heterogeneity component. In particular, for Model V, $u_{it} = u^*_{it} \cdot \exp(\mathbf{z}_i\boldsymbol{\gamma})$ where $u^*_{it} \sim N^+(\mu, \sigma^2_u)$. Table 2.3 presents this decomposition in terms of efficiency for countries in Figure 2.3. We observe that countries such as Yemen and Brazil present higher base efficiency but lower total efficiency than developed countries. This may indicate that these countries present good managerial skills in health provision but under their specific characteristics, they exploit their management abilities to a lesser extent than the developed countries. One of the countries taking great advantage of environmental characteristics is the USA, where efficiency in health provision is highly dependent in their particular attributes. These results are in line with those obtained by contrasting the base model and Model V. Other group of countries, mainly from Africa exhibit low base and low total efficiency. This may indicate both, poor natural managerial abilities, and inability to perform well under their relative bad conditions. Consequently, these countries present very bad performance under all models whether heterogeneity is considered or not.

Overall, we observe that observed heterogeneity variables are inefficiency related and their inclusion in the parameters of the one sided error component distribution has a large

Figure 2.2: Heat map of efficiency rankings - Base model vs. Model V

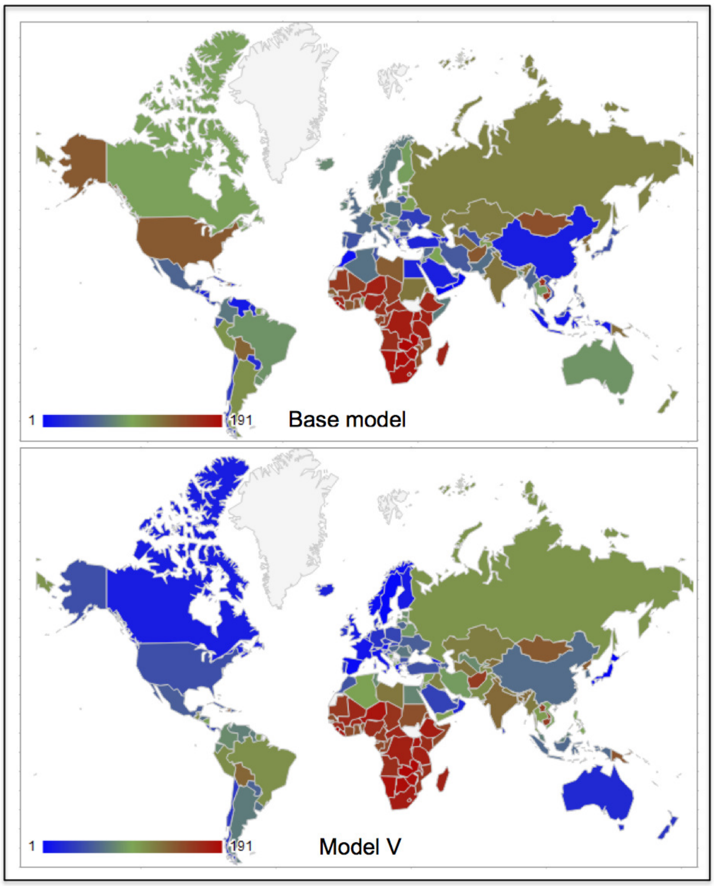
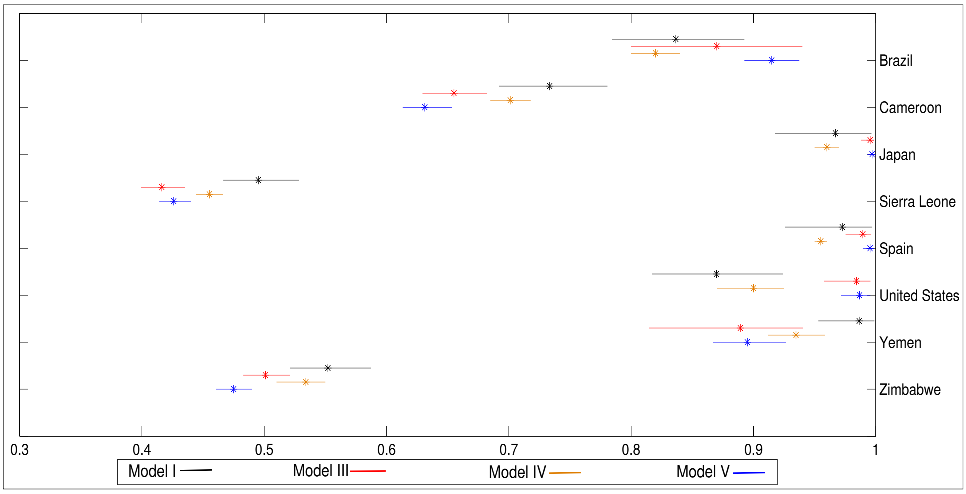


Table 2.3: Posterior mean of base and total efficiency for selected countries

Country	Base efficiency	Total efficiency
Brazil	0.6716	0.9149
Cameroon	0.2543	0.6313
Japan	0.6371	0.9970
Sierra Leone	0.2808	0.4260
Spain	0.6579	0.9953
United States	0.3702	0.9867
Yemen	0.7312	0.8950
Zimbabwe	0.2491	0.4750

impact on the countries' efficiency ranking. Moreover, allowing observed heterogeneity to affect simultaneously both the location and scale parameters of the one-sided error distribution in a way such that the scaling property is preserved has relevant effects on

Figure 2.3: 90% probability intervals of the posterior efficiency distributions for selected countries



shrinking and separating the distributions of posterior individual efficiencies.

2.2.2 Application to Airlines

The airline industry is an interesting sector where performance and efficiency have been studied in the literature using parametric and non-parametric methods. Usually, production functions are employed to evaluate technical efficiency and environmental covariates are often included in the frontier as exogenous variables (see Coelli et al., 1999).

In this application we use a Cobb-Douglas cost function with an output quadratic term to evaluate economic efficiency of the airline industry. The model in (3.1) can be easily extended to a cost function and as in the previous application we consider individual characteristics to capture firms heterogeneity. We use a data set of 24 US domestic airlines over 15 years, from 1970 to 1984, with a total of 246 observations. This is a revised sample obtained from a data set used by Greene (2008).[6]

[6] The original data set includes 256 observations, ten years of observations for an extra airline company. We excluded this firm since we do not have data for the exogenous variables of this airline.

The general model for this application is the following:

$$\begin{aligned} \ln C_{it} = \quad & \alpha + \beta_1 \ln Pm_{it} + \beta_2 \ln Pf_{it} + \beta_3 \ln Pl_{it} + \beta_4 \ln Pe_{it} + \\ & \beta_5 \ln(y_{it}) + \beta_6 \tfrac{1}{2} \ln^2(y_{it}) + \beta_7 t + \beta_8 t^2 + \mathbf{z}_{it}\boldsymbol{\delta} + v_{it} + u_{it} \\ & v_{it} \sim N(0, \sigma_v^2) \\ & u_{it} \sim N^+(\mu \cdot exp(\mathbf{z}_{it}\boldsymbol{\gamma} I_1), \sigma_u^2 \cdot (exp(\mathbf{z}_{it}\boldsymbol{\gamma} I_2))^2), \end{aligned} \tag{2.3}$$

where C_{it} is the total cost supported by airline i at time t in the output production, and Pm_{it}, Pf_{it}, Pl_{it}, Pe_{it} are the input prices of material, fuel, labor and equipment, respectively. Cost and prices are normalized by the property price. y_{it} is the output of airline i at time t and it is an index that aggregates regular passenger, mail, charter, and other freight services. In order to capture possible technological changes over the 15 years covered by the sample we include a trend and its square into the model.

Regarding heterogeneity, $\mathbf{z}_{it}$ is a vector containing information of three observed covariates (load factor, average stage length and points served). Load factor is the effective performed tonne-passenger per kilometer by the airline as a proportion of the total available tonne-passenger per kilometer. Stage length is the ratio of total performed kilometers to the total number of departures. And, points served is the number of destinations.

Variables in $\mathbf{z}_{it}$, as well as other variables of size, are commonly used in productivity and efficiency analysis of the airlines sector but their behavior as drivers of either the frontier or the inefficiency is an open issue. Coelli et al. (1999) present a review on studies using environmental variables in both cases and note that variables in $\mathbf{z}_{it}$ may be argued to have effects on costs and inefficiency.[7] In particular, airlines face high fix but low variable costs, thus we would expect airlines with high load factor to incur in lower costs to transport the same outputs than airlines with a low value for this variable. Its effect on inefficiency would also be negative since a higher load factor implies a higher capital utilization ratio. Airlines operating with high stage length would incur in lower takeoff, landing, parking and other airport costs. Also, they are expected to be more efficient since their aircrafts are being productive for longer time periods. Finally, points served are expected to have a positive effect on total costs since a larger network requires more resources but also more managerial skills which may result on higher or lower inefficiency depending on the routes optimization carried out.

[7] Coelli et al. (1999) evaluate both alternatives for a technical efficiency analysis and conclude statistically in favor of a model including them in the inefficiency term.

Similarly to the WHO application, we estimate five different models. The base model (Model I) does not consider any type of heterogeneity; therefore, $\boldsymbol{\delta} = \mathbf{0}$ and $I_1, I_3 = 0$. Model II considers only frontier heterogeneity by including the observed covariates in the cost function. Models III to V consider covariates in $\mathbf{z}_{it}$ as determinants of the inefficiency and include them in the location, scale or both parameters of the one-sided error distribution, respectively.[8]

Table 2.4: Posterior mean and S.D. of the parameter distributions

Parameter	**Model I**		**Model II**		**Model III**		**Model IV**		**Model V**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Frontier										
α	1.777	0.328	2.463	0.495	1.341	0.289	0.757	0.214	1.623	0.202
$\beta_1(\ln Pm)$	0.359	0.176	0.148	0.070	0.270	0.135	0.112	0.048	0.289	0.162
$\beta_2(\ln Pf)$	0.176	0.055	0.195	0.048	0.197	0.062	0.195	0.064	0.224	0.070
$\beta_3(\ln Pl)$	0.236	0.043	0.484	0.051	0.299	0.048	0.449	0.085	0.217	0.053
$\beta_4(\ln Pe)$	0.052	0.027	0.189	0.064	0.116	0.049	0.136	0.052	0.137	0.058
$\beta_5(\ln y)$	0.942	0.240	0.959	0.218	0.894	0.293	0.861	0.299	0.965	0.305
$\beta_6(\frac{1}{2}\ln^2 y)$	0.088	0.036	0.039	0.012	0.042	0.018	0.038	0.012	0.044	0.017
$\beta_7(t)$	-0.029	0.012	-0.038	0.011	-0.020	0.010	-0.023	0.010	-0.037	0.016
$\beta_8(t^2)$	0.001	0.001	0.001	0.001	-0.001	0.001	-0.001	0.001	0.000	0.000
$\delta_1(Load)$			-0.914	0.253						
$\delta_2(\ln Stage)$			-0.217	0.056						
$\delta_3(\ln Points)$			0.149	0.050						
Inefficiency										
$\gamma_1(Load)$					-0.625	0.296	-0.872	0.408	-0.805	0.395
$\gamma_2(\ln Stage)$					-0.206	0.152	-0.366	0.210	-0.492	0.283
$\gamma_3(\ln Points)$					0.252	0.199	0.250	0.184	0.306	0.204
μ	0.021	0.005	0.209	0.004	0.351	0.008	0.284	0.007	0.351	0.009
σ_u^2	0.184	0.004	0.125	0.003	0.152	0.004	0.123	0.003	0.127	0.004
$Pred.eff.$	0.869	0.101	0.786	0.128	0.681	0.177	0.754	0.099	0.710	0.087
DIC_3	-605.287		-815.394		-697.857		-674.067		-704.846	
LPS	-13.734		-33.652		-19.792		-18.647		-21.669	
MSE	0.026		0.009		0.013		0.019		0.018	

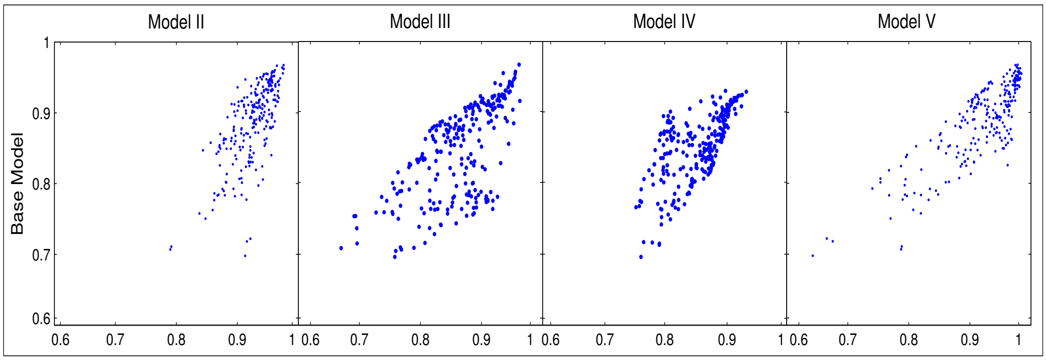
Table 2.4 reports the estimation results. We observe that Model II which includes the observed heterogeneity variables in the cost function present the best fit and predictive performance, suggesting variables in $\mathbf{z}_{it}$ to be drivers of the frontier.[9] Nevertheless, models with inefficiency covariates also improve results from the base model. Among

[8]For all models, monotonicity conditions were found to be not satisfied because of negative signs obtained for prices coefficients. This result was also obtained by Greene (2008). Therefore, we impose regularity conditions by requiring the cost function to have positive elasticities on prices ($\partial c_{it}/\partial p_{it} > 0$). We follow the procedure described in Griffin and Steel (2007) by restricting coefficients β_1 to β_4 to be positive through truncated normal prior distributions for these parameters.

[9]In fact, most of the efficiency studies applied to airlines have treated size and network environment variables as frontier drivers (see Coelli et al., 1999).

these models, the one that includes covariates in both parameters of the inefficiency distribution and preserves the scaling property (Model V) presents the best values in terms of DIC_3 and LPS. However, differences are narrower than in the previous application, in particular compared to Model III, which exhibits the lowest value of MSE. As in the WHO application, models including observed heterogeneity in the scale parameter of the inefficiency exhibit lower dispersion of the predictive efficiency distribution. Regarding the estimated coefficients, we identify increasing returns to scale and expected effects of covariates on costs and inefficiency as discussed above. From the estimation results obtained for Model II we conclude that load factor and stage length affect negatively costs, while the network size has the opposite effect. Overall, considering heterogeneity has effects on the estimations of posterior mean efficiencies with respect to the base model, as we observe in Figure 2.4.

Figure 2.4: Posterior mean efficiencies - Base model vs. heterogeneity models



2.3 Effects of different inefficiency distributions

Besides including these covariates in the different parameters of the same distribution, we also evaluate the effects of assuming different distributions for the one-sided error component. Inefficiencies are assumed to follow: a) a half normal distribution, following the specification for the scale parameter in (1.17), b) a truncated normal distribution, using the scaled Stevenson model in (1.18), and c) an exponential distribution following the model in (1.19).

The general model for panel data in this case is the following:

$$\begin{aligned} y_{it} = \;& \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{z^*}_{it}\boldsymbol{\delta} + v_{it} - u_{it} \\ & v_{it} \sim N(0, \sigma^2) \\ & a)\;\; u_{it} \sim N^+(0, \sigma_u^2 \cdot (\exp(\mathbf{z}_{it}\boldsymbol{\gamma}))^2) \\ & b)\;\; u_{it} \sim N^+(\mu \cdot \exp(\mathbf{z}_{it}\boldsymbol{\gamma}), \sigma_u^2 \cdot (\exp(\mathbf{z}_{it}\boldsymbol{\gamma}))^2) \\ & c)\;\; u_{it} \sim Exp(\lambda \cdot \exp(\mathbf{z}_{it}\boldsymbol{\gamma})), \end{aligned} \tag{2.4}$$

$\mathbf{z^*}_{it}$ is a row vector of the observed heterogeneity variables that affect the technology; $\mathbf{z}_{it}$ is a row vector of observed covariates with effects in the inefficiency; and, $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, and $\boldsymbol{\gamma}$ are vectors of the estimated parameters.

In this case, the distributions of the one-sided error component are: $u_{it}|\boldsymbol{\gamma}, \mathbf{z}_{it} \sim N^+(0, \sigma_u^2 \cdot (\exp(\mathbf{z}_{it}\boldsymbol{\gamma}))^2)$ for the half-normal model; $u_{it}|\boldsymbol{\gamma}, \mathbf{z}_{it} \sim N^+(\mu \cdot \exp(\mathbf{z}_{it}\boldsymbol{\gamma}), \sigma_u^2 \cdot (\exp(\mathbf{z}_{it}\boldsymbol{\gamma}))^2)$ for the truncated normal model; and, $u_{it}|\boldsymbol{\gamma}, \mathbf{z}_{it} \sim Exp(\lambda \cdot \exp(\mathbf{z}_{it}\boldsymbol{\gamma}))$ for the exponential case. μ, σ_u^2 and λ are defined as in Griffin and Steel (2007). The MCMC is performed as in the previous case, running 50,000 iterations and discarding the first 10,000 in a burn-in phase.

2.3.1 Application to WHO data set

The model in (3.2) is adjusted in order to allow the inefficiency component to follow: a) half-normal, b) truncated normal, or c) exponential distributions:

$$\begin{aligned} \ln(DALE_{it}) = \;& \alpha + \beta_1 \ln(HExp_{it}) + \beta_2 \ln(Educ_{it}) + \beta_3 \tfrac{1}{2} \ln^2(Educ_{it}) + \beta_4 Tropics_i \\ & + \beta_5 \ln(PopDen_i) + \mathbf{z}_i\boldsymbol{\delta} + v_{it} - u_{it} \\ & v_{it} \;\sim N(0, \sigma_v^2) \\ & a)\;\; u_{it} \sim N^+(0, \sigma_u^2 \cdot (exp(\mathbf{z}_i\boldsymbol{\gamma}))^2) \\ & b)\;\; u_{it} \sim N^+(\mu \cdot exp(\mathbf{z}_i\boldsymbol{\gamma}), \sigma_u^2 \cdot (exp(\mathbf{z}_i\boldsymbol{\gamma}))^2) \\ & c)\;\; u_{it} \sim \exp(\lambda \cdot exp(\mathbf{z}_i\boldsymbol{\gamma})). \end{aligned} \tag{2.5}$$

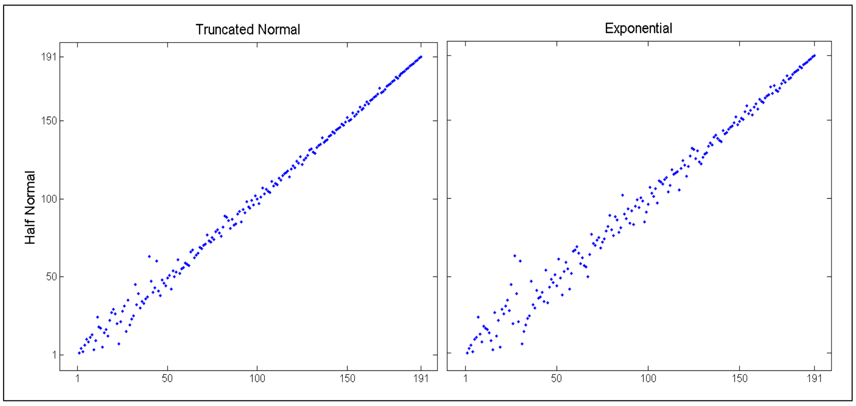
Model comparison criteria for the four models and the three distributions are presented in Table 2.5. In general, similar conclusions are obtained from the three criteria. Results show that models including either observed or unobserved heterogeneity improve from the base model. In particular, the model that exhibits the best fit and predictive performance includes observed heterogeneity in the inefficiency, which suggests that covariates in $\mathbf{z}_i$ are inefficiency related. Regarding the inefficiency distributions, the half-normal and truncated normal models present better indicators and seem to be

Table 2.5: Model comparison criteria assuming different inefficiency distributions

Distribution		**Model I**	**Model II**	**Model III**	**Model IV**
Half normal	DIC	-2251.7150	-2598.3080	-2423.3160	-2914.7370
	LPS	-97.1690	-132.7610	-154.8950	-196.4420
	MSE	0.1382	0.0864	0.0906	0.0736
Truncated normal	DIC	-2292.7710	-2593.1280	-2495.1400	-2884.9030
	LPS	-122.8900	-130.4520	-146.7710	-185.9830
	MSE	0.1387	0.1051	0.1084	0.0869
Exponential	DIC	-2223.7420	-2568.4380	-2231.4950	-2580.1720
	LPS	-95.9810	-121.5150	-123.3560	-132.2700
	MSE	0.1392	0.1153	0.1281	0.1085

better alternatives, specially for those models considering observed heterogeneity in u_{it}. However, efficiency rankings are almost perfectly correlated across distributions as we can observe for Model IV in Figure 2.5.

Figure 2.5: Efficiency rankings in Model IV across distributions



2.4 Conclusions

In stochastic frontier analysis the inefficiency component may be erroneously estimated when firm characteristics are not taken into account. These firm characteristics induce heterogeneity that might result in different firm frontiers, or may have an impact directly on the inefficiencies.

In this chapter, the effects of including observed heterogeneity in different parameters of a truncated normal distributed inefficiency were studied. The models were fitted to two data sets previously studied only in the frequentist context and the results were compared to those obtained with models that ignore heterogeneity or include it in the frontier.

Differences in efficiency rankings and mean scores were observed when inefficiency heterogeneity was included in different parameters of the one-sided error distribution. This was found to be related to effects in the posterior efficiency distributions. In particular, considering firms' heterogeneity in the location parameter of the inefficiency has an effect on separating the firm specific posterior efficiency distributions from each other, which leads to more reliable rankings. On the other hand, when heterogeneity affects only the scale parameter of the inefficiency, an important shrinking effect is observed on the individual posterior efficiency distributions. This results in less uncertainty around mean individual efficiency scores. Finally, including the heterogeneity in both parameters of the inefficiency distribution in models that preserve the scaling property leads to both separating and shrinking effects. This allows less overlapping of the posterior efficiency distributions and provide both more reliable efficiency scores and rankings. Models with this property were extended to the Bayesian context and preserving the scaling property was found to lead to better fit and predictive performance indicators.

2.5 Appendix

A. WinBUGS code for a truncated normal model with observed heterogeneity - Airlines application

```
model{
for (i in 1:N) {
 m[i] <- mu*exp(gamma[1]*z1[i]+gamma[2]*z2[i]+gamma[3]*z3[i])
 sigmau[i]<- (1/sqrt(lambda))*exp(gamma[1]*z1[i]+gamma[2]*z2[i]+gamma[3]*z3[i])
 nu[i]<-1/(sigmau[i]*sigmau[i])
 u[i] ~ djl.dnorm.trunc(m[i],nu[i],0,1000)
 eff[i] <- exp(- u[i]) }

for ( i in 1:N ) {
 mc[i] <- alpha + u[i]+ beta[1]*x1[i]+beta[2]*x2[i]+beta[3]*x3[i]+beta[4]*x4[i]
          +beta[5]*y[i]+0.5*beta[6]*y[i]*y[i]+beta[7]*(t[i])+beta[8]*(t[i]*t[i])
          +beta[9]*z1[i]+beta[10]*z2[i]+beta[11]*z3[i]
 lnc[i] ~ dnorm(mc[i], prec) }

mu <- psi/sqrt(lambda)
psi ~ dnorm(0.0,1)
lambda~dgamma(5,lambda0)
lambda0 <- 5*log(rstar)*log(rstar)
for (i in 1:3) {
 gamma[i] ~ dnorm(0.0, 0.1)}
#Alternative prior for gamma:
#for (i in 1:3) {
 #gammastar[i] ~ dnorm(0.0, 0.1)
 #gamma[i] <- gammastar[i] / sqrt(lambda) }

alpha ~ dnorm(0.0, 1.0E-06)
for (i in 1:5) {
 beta[i] ~ djl.dnorm.trunc(0.0, 1.0E-06,0,1000) }
for (i in 6:11) {
 beta[i] ~ dnorm(0.0, 1.0E-06) }
prec ~ dgamma(0.001, 0.001)
sigmasq <- 1 / prec  }
```

Chapter 3

Modeling Unobserved Inefficiency Heterogeneity

Unobserved heterogeneity in the non-negative error component has been very little explored in the literature from a frequentist or a Bayesian approach. However, ignoring its existence means that heterogeneity which is not captured by observed covariates is wrongly attributed to inefficiency and consequently leads to bad efficiency estimates.

The literature on modeling unobserved firm characteristics in the inefficiency is still scarce. In the frequentist context, Greene (2005) proposed a model where the coefficients of the observed covariates are allowed to be firm specific and vary randomly. In the Bayesian framework, Koop et al. (1997) propose a model that may capture unobserved inefficiency heterogeneity. In this case, the inefficiency is assumed to be exponentially distributed with firm specific mean and independent priors.

In this chapter, we introduce two different ways to account for unobserved inefficiency heterogeneity. The first is the inclusion of a random parameter in the distribution of the inefficiency, which can be included alone or along with observed covariates. The second proposal is modeling random coefficients of the observed inefficiency covariates. This is performed through hierarchical structures in the parameters associated to the observed covariates. Both models are then analyzed using Bayesian inference techniques.

This chapter is divided into three sections. In Section 3.1, we present the model including a random parameter in the inefficiency intended to capture latent heterogeneity. This specification is studied in the two samples presented in Chapter 2 and the effects of its inclusion in the location or scale parameter of a truncated-normal distributed inefficiency are also analyzed. Section 3.2 presents the random coefficients model and its application to the relationship between risk-taking and efficiency in the Colombian banking sector. Finally, Section 3.3 presents some conclusions.

3.1 A stochastic frontier model with a random parameter in the inefficiency*

In this section, a proposal to model unobserved heterogeneity in the inefficiency through the inclusion of a random parameter is presented. This parameter has three main characteristics. It can be allowed to be time-varying, it can be included simultaneously with observed covariates in the inefficiency distribution in order to distinguish observed from unobserved heterogeneity and it can indicate whether or not observed covariates do a good job in capturing the existing heterogeneity.

Following the formulation in Chapter 2, we present a general stochastic frontier model for panel data that allows the modeling of both observed and unobserved inefficiency heterogeneity. For the one-sided error we use an exponential specification of a truncated normal distribution where the location, scale, or both parameters can model firm heterogeneity. The general model in the case of a production function is:

$$\begin{aligned} y_{it} = &\ \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{z^*}_{it}\boldsymbol{\delta} + v_{it} - u_{it};\ v_{it} \sim N(0, \sigma_v^2) \\ u_{it} \sim &\ N^+(\mu \cdot \exp(\mathbf{z}_{it}\boldsymbol{\gamma} I_1 + \tau_{it} I_2), \sigma_u^2 \cdot (\exp(\mathbf{z}_{it}\boldsymbol{\gamma} I_3 + \tau_{it} I_4))^2), \end{aligned} \tag{3.1}$$

where y_{it} is the output of firm i at time t, $\mathbf{x}_{it}$ is the row vector of input quantities, $\mathbf{z^*}_{it}$ is a row vector of the observed heterogeneity variables that affect the technology; $\mathbf{z}_{it}$ is a row vector of observed covariates with effects in the inefficiency; τ_{it} is a random parameter that captures time-varying unobserved firm effects in the inefficiency; and, $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, and $\boldsymbol{\gamma}$ are the corresponding parameter column vectors. I_1 to I_4 are indicator variables taking the value of 1 when either observed covariates or unobserved heterogeneity are accounted for in the location or scale parameters, respectively, and 0 otherwise.

*Much of the work in this section has been published in the *Journal of Productivity Analysis* (see Galán et al., 2014b).

3.1.1 Bayesian inference

All the models derived from the general specification in (3.1) are fitted by Bayesian methods. Proper but relatively disperse prior distributions throughout. The distributions assumed for the parameters in the frontier ($\boldsymbol{\beta}$, $\boldsymbol{\delta}$), the variance of the idiosyncratic error term (σ_v^{-2}), and the inefficiency covariates ($\boldsymbol{\gamma}$) are the same described in Chapter 2.

In the case of unobserved heterogeneity in the inefficiency, the unknown parameter is specified to have a hierarchical structure: $\tau_{it} \sim N(\overline{\tau}, \sigma_\tau^2)$, where $\overline{\tau} \sim N(0, 10)$ and $\sigma_\tau^{-2} \sim G(0.5, 0.5)$. The random parameter τ_{it} can be defined to be either time-varying or not.

As in Chapter 2, the MCMC for these models is carried out using the WinBUGS package following the general procedure outlined in Griffin and Steel (2007). However, the hyperparameters $\overline{\tau}$ and σ_τ^{-2} present slow convergence and high autocorrelation. In particular, if initial values are set far from the posterior mean, convergence is observed only after 50,000 iterations and autocorrelations of order around 20 are identified. Therefore, for these models 550,000 iterations were used for the MCMC, thinning every 25 iterations and discarding the first 50,000. We computed the MCMC convergence diagnostic in Geweke (1992) for the hyperparameteres and the obtained numerical standard errors show very low values that suggest reasonably precise estimates.[1] In Appendix A we show the MCMC iterations for different initial values of $\overline{\tau}$ in both empirical applications. Finally, sensitivity analysis of our results to changes in other prior parameters was also carried out. Results showed that the posterior inference was relatively insensitive to small changes in these parameters.

Finally, the model selection criteria used for these models are DIC_3, LPS and MSE as described in Chapter 2.

3.1.2 Empirical applications

For illustration, we use the same, WHO and airlines, data sets described in Chapter 2. In particular, in the WHO application, since the observed covariates are inefficiency related and time invariant, we include them in different parameters of the inefficiency distribution together with a time invariant random parameter. On the other hand, in

[1] We computed the numerical standard errors employing a 4% autocovariance tapered estimate.

the second application observed heterogeneity variables are time-varying and frontier drivers, so the unobserved heterogeneity component is allowed to change over time and its effects in the posterior efficiency distributions are evaluated when it is included in the location, scale or both parameters of the one-sided error distribution.

3.1.2.1 The WHO data set

The results obtained in Chapter 2 allow us to test our proposal to capture latent heterogeneity through a random parameter. Since previous results favor the scaling property model, we analyze unobserved heterogeneity in models that satisfy this property.

For this application the general model is:

$$\begin{aligned} \ln(DALE_{it}) = \quad & \alpha + \beta_1 \ln(HExp_{it}) + \beta_2 \ln(Educ_{it}) + \beta_3 \tfrac{1}{2} \ln^2(Educ_{it}) + \beta_4 Tropics_i \\ & + \beta_5 \ln(PopDen_i) + \mathbf{z}_i \boldsymbol{\delta} + v_{it} - u_{it};\; v_{it} \sim N(0, \sigma_v^2) \\ & u_{it} \sim N^+(\mu \cdot exp(\mathbf{z}_i \boldsymbol{\gamma} I_1 + \tau_i I_2), \sigma_u^2 \cdot (exp(\mathbf{z}_i \boldsymbol{\gamma} I_3 + \tau_i I_4))^2). \end{aligned} \tag{3.2}$$

First, we estimate Model A where we assume no information about observed heterogeneity variables in $\mathbf{z}_i$. That is, we impose $I_2, I_4 = 1$ and $I_1, I_3 = 0$ in Equation (3.2). Notice that these covariates are time invariant, so for this application the random parameter capturing unobserved effects is defined to be firm specific and constant over time, as well.

We propose to estimate two additional models, where observed covariates are also considered to affect inefficiency. In these cases all indicator variables in Equation (3.2) are equal to 1. This allows us to analyze the efficacy of the parameter τ_i to capture information from omitted covariates and to identify those which are relevant. Model B considers the variables $Gini$ and GDP in addition to the random parameter. These two variables capture the most relevant aspects of inequality and income distinguishing countries and were also found to be the most inefficiency related by Greene (2004) after performing a frequentist based test. Finally, we estimate Model C where τ_i is estimated along with all the covariates in $\mathbf{z}_i$.

Results are presented in Table 3.1. In general, we observe that all model comparison criteria improve compared to models I and II when the unobserved component is included in the inefficiency distribution. This implies that the random component captures part of the heterogeneity identified by covariates in $\mathbf{z}_i$ and therefore, it is a good alternative when no observed heterogeneity variables are available.

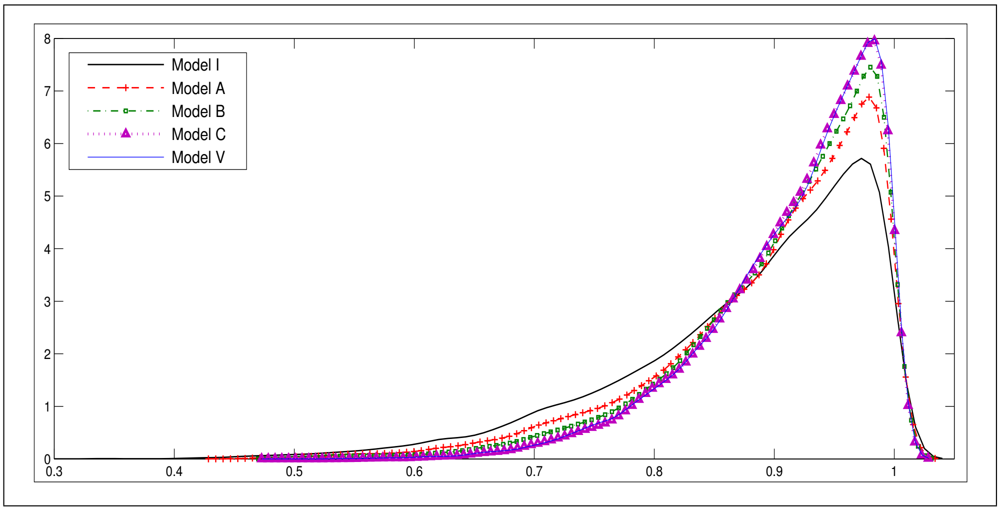
Table 3.1: Posterior mean and S.D. of the parameter distributions for unobserved heterogeneity models

Parameter	**Model A**		**Model B**		**Model C**	
	Mean	SD	Mean	SD	Mean	SD
Frontier						
α	3.846	0.517	3.753	0.504	3.732	0.498
β_1	0.026	0.012	0.024	0.010	0.025	0.010
β_2	0.212	0.085	0.372	0.137	0.413	0.132
β_3	-0.036	0.013	-0.085	0.020	-0.099	0.026
β_4	-0.005	0.003	-0.003	0.002	-0.008	0.003
β_5	-0.002	0.001	-0.005	0.002	-0.006	0.003
Inefficiency						
$\gamma_1(Gini)$			1.950	0.542	1.261	0.469
$\gamma_2(\ln GDP)$			-0.542	0.106	-0.363	0.093
$\gamma_3(GEff)$					-0.070	0.021
$\gamma_4(Voice)$					0.024	0.011
$\gamma_5(OECD)$					-0.746	0.321
$\gamma_6(\ln Pubfin)$					0.083	0.022
$\overline{\tau}$	-4.617	0.953	-0.803	0.197	-0.745	0.175
σ_τ^{-2}	1.039	0.428	2.192	0.806	1.976	0.760
μ	-1.655	0.294	-1.486	0.351	-0.392	0.147
σ_u^2	0.079	0.020	0.099	0.021	0.057	0.013
$Pred.eff.$	0.833	0.090	0.877	0.099	0.915	0.071
DIC_3	-2957.820		-3017.610		-3085.190	
LPS	-146.717		-152.950		-180.427	
MSE	0.104		0.101		0.088	

A second finding is that when τ_i is included simultaneously with observed variables in the inefficiency distribution, this parameter can be used as an indicator of the suitability of the observed covariates to capture inefficiency heterogeneity. In fact, it is observed that Model B, which includes only two covariates in $\mathbf{z}_i$ besides the random parameter, improves in terms of fit and predictive performance in comparison to Model A but it is not as good as Model V that include six covariates. This would mean that $Gini$ and GDP are relevant heterogeneity variables but they are not able to capture all the inefficiency heterogeneity. On the other hand, Model C that includes all observed covariates plus the parameter τ_i performs a little worse than Model V (see model comparison criteria in tables I and IV). This would imply that the six covariates in $\mathbf{z}_i$ capture all the relevant inefficiency heterogeneity.

These conclusions are the same when we compare the posterior predictive efficiencies of models including the unobserved component to those of models I and V (see Figure 3.1). It can be seen that the predictive efficiency distribution becomes less disperse

Figure 3.1: Kernel densities of posterior efficiency distributions



to the extent inefficiency heterogeneity is better identified by the random parameter, observed covariates or a combination of both. Also, it is observed that the predictive efficiency distribution of Model C is very close to that of Model V, which suggests that the parameter τ_i is irrelevant when the observed covariates are able to capture most of the inefficiency heterogeneity.

3.1.2.2 The airlines data set

Since in Chapter 2, the observed covariates were found to be related to frontier heterogeneity, our benchmark here is Model II. However, we assume that it may still exist inefficiency heterogeneity in the sector related to other factors not considered by variables in $\mathbf{z}_{it}$. Therefore, we evaluate the inclusion of a time-varying random parameter in the distribution of the inefficiency when it is specified in the location, scale, or both parameters of the one-sided error distribution. The general model for this application considering unobserved inefficiency heterogeneity is:

$$\begin{aligned} \ln C_{it} = \; & \alpha + \beta_1 \ln Pm_{it} + \beta_2 \ln Pf_{it} + \beta_3 \ln Pl_{it} + \beta_4 \ln Pe_{it} + \\ & \beta_5 \ln(y_{it}) + \beta_6 \tfrac{1}{2} \ln^2(y_{it}) + \beta_7 t + \beta_8 t^2 + \mathbf{z}_{it}\boldsymbol{\delta} + v_{it} + u_{it} \\ & v_{it} \sim N(0, \sigma_v^2) \\ & u_{it} \sim N^+(\mu \cdot exp(\tau_{it} I_1), \sigma_u^2 \cdot (exp(\tau_{it} I_2))^2). \end{aligned} \tag{3.3}$$

In contrast to the WHO application, here the random parameter τ_{it} is allowed to vary over time and is modeled without any observed covariates in the inefficiency. We estimate there different models derived from (3.3). Model A includes τ_{it} only in the location parameter of the inefficiency distribution and then the indicator $I_2 = 0$; Model B includes it only in the scale parameter ($I_1 = 0$); and Model C includes it in both the location and the scale parameters of the inefficiency distribution ($I_1 = I_2 = 1$).

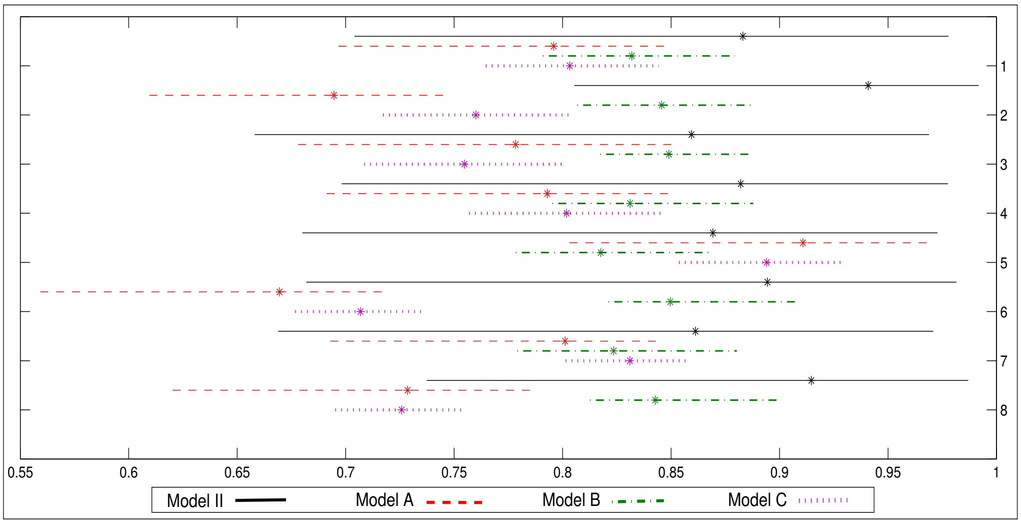
Results for the three estimated models are presented in Table 3.2. It can be observed that all three models improve their fit and predictive performance in comparison to Model II. In particular, models A and C exhibit the best values for the three criteria. However, when the random parameter is included in the scale parameter of the inefficiency distribution (models B and C), a decrease in the dispersion of the predictive efficiency distribution is observed.

Table 3.2: Posterior mean and S.D. of the parameter distributions for unobserved heterogeneity models

Parameter	**Model A**		**Model B**		**Model C**	
	Mean	SD	Mean	SD	Mean	SD
Frontier						
α	1.666	0.201	0.470	0.125	2.914	0.345
$\beta_1(\ln Pm)$	0.436	0.227	0.303	0.162	0.153	0.080
$\beta_2(\ln Pf)$	0.194	0.063	0.197	0.061	0.237	0.069
$\beta_3(\ln Pl)$	0.155	0.048	0.236	0.076	0.330	0.081
$\beta_4(\ln Pe)$	0.147	0.054	0.154	0.059	0.204	0.062
$\beta_5(\ln y)$	0.871	0.301	0.878	0.283	0.976	0.265
$\beta_6(\frac{1}{2}\ln^2 y)$	0.045	0.019	0.026	0.013	0.043	0.019
$\beta_7(t)$	-0.032	0.010	-0.013	0.004	-0.027	0.012
$\beta_8(t^2)$	0.001	0.001	-0.001	0.001	-0.001	0.000
$\delta_1(Load)$	-1.096	0.262	-1.142	0.254	-0.856	0.197
$\delta_2(\ln Stage)$	-0.247	0.054	-0.235	0.049	-0.205	0.049
$\delta_3(\ln Points)$	0.106	0.048	0.071	0.030	0.135	0.057
Inefficiency						
$\overline{\tau}$	-3.491	0.924	-4.221	0.931	-3.514	0.806
σ_τ^{-2}	1.729	0.562	0.895	0.391	1.255	0.389
μ	0.611	0.167	0.343	0.095	0.321	0.082
σ_u^2	0.105	0.031	0.052	0.016	0.076	0.017
$Pred.eff.$	0.774	0.089	0.835	0.019	0.797	0.047
DIC_3	-971.711		-938.855		-984.369	
LPS	-40.528		-36.780		-39.647	
MSE	0.009		0.009		0.009	

The effects on the individual posterior efficiencies using the random parameter are similar to those found in the previous application using observed covariates. That is, when τ_{it} is considered in the location parameter of the one-sided error distribution, the posterior efficiencies of different airlines are more separated from each other, and when it is included in the scale parameter, we observe a shrinking effect and consequently a decrease in the dispersion of the posterior efficiency distributions. Figure 3.2 shows these effects for some selected airlines. We can observe that Model C, which includes the random parameter in both parameters of the inefficiency distribution and satisfies the scaling property, separates and shrinks the individual posterior efficiency distributions providing both more reliable efficiency scores and rankings.

Figure 3.2: 90% credible intervals of the posterior efficiency distributions for selected airlines



Preserving the scaling property makes it possible to decompose inefficiency for Model C. In this case, $u_{it} = u^*_{it} \cdot \exp(\tau_{it})$ where $u^*_{it} \sim N^+(\mu, \sigma^2_u)$. Table 3.3 exhibits the decomposition in terms of efficiency for the airlines plotted above. The difference between the base and total efficiency allows us to distinguish the way unobserved firm effects are handled by airlines managers. For instance, airline 12 presents lower base efficiency but higher total efficiency than airline 17, suggesting that the former handles their specific characteristics better.

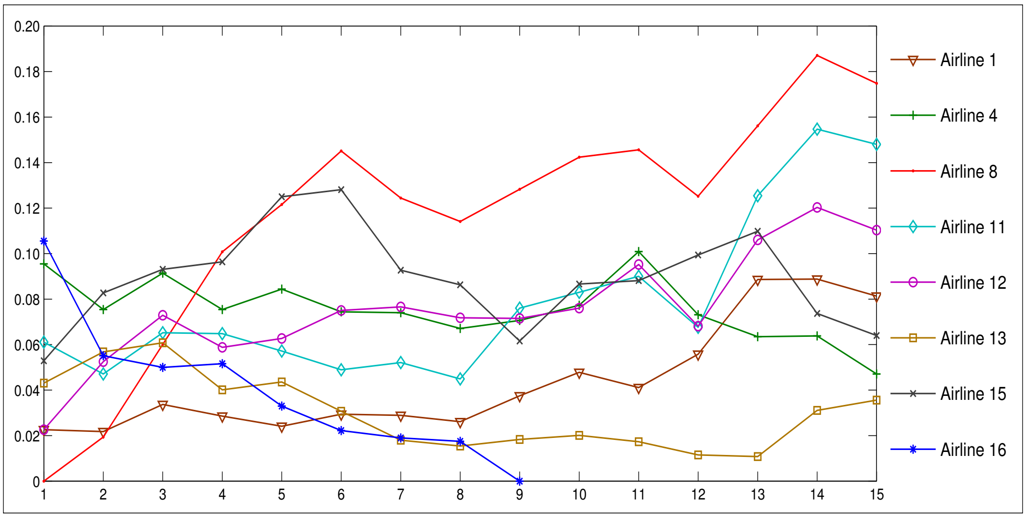
Finally, using the results of Model C, in Figure 3.3 we plot the probabilities of being the most efficient airline in the sample period for some selected firms. This can be easy calculated in the Bayesian context from the posterior individual distributions of efficiencies and might be very useful in empirical studies. We observe that for the last

Table 3.3: Posterior mean of base and total efficiency for selected airlines

Airline ID	Base efficiency	Total efficiency
1	0.4837	0.8245
2	0.3052	0.7669
5	0.4017	0.7614
8	0.6238	0.8092
12	0.3571	0.8970
17	0.5466	0.7194
18	0.5824	0.8352
19	0.3920	0.7317

10 years of the sample period, airline 8 is the most likely to be the benchmark firm. Also, it is possible to see improvements and declines in the airlines' performance along time. For instance, airline 11 presents a high relative improvement of its performance especially in the last 3 years, while airline 16 starts being the most likely benchmark firm and decreases very fast its probability up to being zero in year 9.

Figure 3.3: Probability of being the most efficient firm in the sample period



Summing up, the performance indicators suggest that firm characteristics such as the distance between destinations, the capacity offered, and the size of the network differentiate the airlines in terms of the cost frontier they face. However, there is still latent inefficiency heterogeneity related to unobserved factors. This is captured through a time varying random parameter that improves fit and predictive performance. The way this parameter is included in the inefficiency has different effects in terms of separating and shrinking the individual posterior efficiency distributions. The most desirable effects are obtained when the unobserved heterogeneity component is included both in the location and scale parameters of the inefficiency distribution in models that satisfy the scaling property.

3.2 A stochastic frontier model with random inefficiency coefficients*

Unobserved heterogeneity sources in the inefficiency may also be associated to differences in the way observed covariates affect the inefficiency. That is, covariates modeled in the inefficiency distribution may produce different effects on the inefficiency depending on some unobserved firm-specific characteristics.

This has been previously studied for unobserved firm-specific heterogeneity in the technology. Tsionas (2002) proposed, in the Bayesian context, a model with random coefficients in the frontier, which captures different effects of technology factors for every firm. That model is the one in (1.23) described in Chapter 1. In this Section, this specification is extended by including random coefficients in the covariates of the inefficiency distribution rather than in the frontier.

The proposed stochastic frontier model assuming an exponential distribution for the one-sided error term is the following:

$$\begin{aligned} y_{it} = \quad & \mathbf{x}_{it}\boldsymbol{\beta} - u_{it} + v_{it} \\ & v_{it} \sim N(0, \sigma_v^2) \\ & u_{it} \sim Exp(\lambda_{it}) \\ & \lambda_{it} = \exp\left(\begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\gamma}_i^* \end{pmatrix}' \begin{pmatrix} \mathbf{z}_{it} & \mathbf{0} \\ \mathbf{0} & \mathbf{z^*}_{it} \end{pmatrix} \right), \end{aligned} \tag{3.4}$$

where y_{it} represents the output for firm i at time t, $\mathbf{x}_{it}$ is a row vector that contains the input quantities, $\boldsymbol{\beta}$ is a vector of parameters, v_{it} is an idiosyncratic error assumed to follow a normal distribution, and u_{it} is the inefficiency component. The inefficiency is assumed to follow an exponential distribution with a firm specific and time-varying parameter λ_{it}, which depends on a vector including two sets of parameters and a matrix

*Much of the work in this section is joint with Miguel Sarmiento from the Colombian central bank (see Sarmiento and Galán, 2014) that includes two types of heterogeneity variables. $\boldsymbol{\gamma}$ is a vector of parameters which are common to all firms, including the constant; and, $\boldsymbol{\gamma}_i^*$ is a vector of firm-specific parameters intended to capture differences in the effects of covariates across firms on the inefficiency. Therefore, $\mathbf{z}_{it}$ is a vector of heterogeneity variables whose effects are assumed to be constant across firms, and $\mathbf{z}_{it}^*$ contains a set of heterogeneity variables with firm-specific effects. In the case of assessing cost efficiency, y_{it} would represent costs and the sign of the inefficiency component is reversed.

The novel specification with random coefficients in the parameter of the inefficiency distribution is flexible in the sense that some covariates can be associated to firm specific coefficients while other heterogeneity variables may be modeled with fixed coefficients. In particular, the random specification for the inefficiency coefficients is intended to capture differences in the way some specific characteristics affect efficiency of different types of firms. Therefore, the model is able to identify, not only the effects of observed covariates in the inefficiency, but also the type of firms that are more affected by each of these characteristics.

3.2.1 Bayesian inference

We assume proper but relatively disperse prior distributions throughout. In particular, the distributions assumed for the parameters in the frontier are: $\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}$ is a precision diagonal matrix with priors set to 0.001 for all coefficients. Finally, the variance of the idiosyncratic error term is inverse gamma, that is equivalent to $\sigma_v^{-2} \sim G(a_{\sigma_v^{-2}}, b_{\sigma_v^{-2}})$ with priors set to 0.01 for the shape and rate parameters, respectively.

Regarding the inefficiency component, its distribution is assumed to be exponential: $u_{it}|\boldsymbol{\gamma}, \boldsymbol{\gamma}^*, \mathbf{z}_{it}, \mathbf{z}^*{}_{it} \sim Exp(\exp(\mathbf{z}_{it}\boldsymbol{\gamma} + \mathbf{z}^*{}_{it}\boldsymbol{\gamma}_i^*))$. The prior distribution of the vector of common parameters $\boldsymbol{\gamma}$ is chosen to be centered in a given prior mean efficiency value r^* following the procedure in Griffin and Steel (2007) i.e. $\exp(\boldsymbol{\gamma}) \sim Exp(-\ln r^*)$. For the firm-specific inefficiency heterogeneity coefficients, a hierarchical structure is defined, where: $\exp(\boldsymbol{\gamma}_i^*) \sim Exp(\boldsymbol{\gamma}^*)$, and $\boldsymbol{\gamma}^* \sim Exp(-\ln r^*)$. Therefore, the firm-specific parameters are centered a priori in a common parameter, which at its turn, is centered in a given prior mean efficiency value. In this particular application, r^* is set at 0.65, following other Bayesian SFA studies in banking (see Marzec and Osiewalski, 2001; Tabak and Tecles, 2010). Sensitivity analysis is performed to the use of a normal prior distribution for the inefficiency parameters such that: $\boldsymbol{\gamma}$ is $N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$ with priors for the diagonal precision matrix $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}$ equal to 0.1 for all the coefficients. In this case, the hierarchical structure used for the firm-specific parameters is $\boldsymbol{\gamma}_i^* \sim N(\boldsymbol{\gamma}^*, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}^*})$ where $\boldsymbol{\gamma}^*$ is defined in the same way that $\boldsymbol{\gamma}$.

The implementation is carried out using the WinBUGS package. The MCMC algorithm involves 50,000 iterations where the first 10,000 are discarded and a thinning equal to 4 is used to remove autocorrelations. Therefore, 10,000 iterations are used for the posterior inference. We assess the fit and predictive performance of the different models using a version of the DIC_3 and the LPS as earlier.

3.2.2 Application to bank risk-taking in the Colombian banking sector

After the global financial crisis, understanding bank risk taking has gained more attention among researchers and practitioners because of the regulatory framework proposed in Basel III, which limit and monitor bank risk taking by imposing higher capital requirements and more liquid assets holdings in their portfolios (BIS, 2010).[2] These higher requirements would reduce bank risk exposure but also their profitability in the short run. In the case of emerging economies, there is a growing interest on the effects of the increasing risk appetite exhibited recently by financial institutions, which is mainly associated to higher capital inflows from advanced economies, where financial fragility coexists with prolonged lower interest rates, especially in the Euro area (Ahmed and Zlate, 2013; Bruno and Song, 2013). As a result, the analysis of bank efficiency incorporating bank risk exposure constitute a key element to identify bank performance under risk taking environment which may contribute to the proper design of macroprudential policies to enhance financial stability.

Modern banking theory highlights that risk taking is an inherent element of banking production which should be properly modeled into the efficiency measurement (Hughes et al., 2001). Recent studies have shown that failing to account for risk taking leads to biased estimations of bank efficiency as well as mislead estimates of scale economies and

[2]The Basel III framework promotes higher and better-quality capital, risk coverage and leverage ratios to increase resilience in periods of stress. Likewise, states the introduction of a Liquidity Coverage Ratio for short term (30 days) and a Net Stable Funding Ratio (NSFR) for long-term (one year), which will be implemented gradually during the 2015-2018 period according with the evolution of the economic activity in each member country. The recent initiatives by the European Systemic Risk Board (ESRB), the International Organization of Securities Commissions (IOSCO) and the Financial Stability Board (FSB) are also aligned to limit the risk-taking behavior and contagion in financial market through micro and macro-prudential policies.

cost elasticities (Hughes and Mester, 2013; Koetter, 2008; Malikov et al., 2013). However, studies that incorporate bank risk taking in efficiency measurements have traditionally included only proxies of credit risk exposure (i.e. non-performing loans or loan losses provisions), omitting other important risks faced by banks (e.g. liquidity, market and capital exposures).

Credit risk proxies are usually included into costs and profit functions as a measure of output quality that directly affects the technology (Hughes and Mester, 1998; Mester, 1996) or as an undesirable output where reductions are desirable (see applications in Assaf et al., 2013; Park and Weber, 2006; Zago and Dongili, 2011). Under that approach, risk taking is assumed as an exogenous component of the banking production process. This contrasts with recent empirical literature that illustrates how most of the bank inefficiency corresponds to poor management (or riskier strategies), which is reflected in a higher ex-post credit risk, i.e. elevated share of NPLs (Lepetit et al., 2008).

When risk taking is modeled as endogenous, we can find two approaches in the literature: Firstly, structural models of banking production that account for managerial-risk preferences and endogenous risk taking. In these models, bank performance is measured in the risk-return space that incorporates the trade-off between expected profit and risk.[3] Secondly, SFA models in which firm-specific characteristics are modeled as elements that affect the inefficiency distribution instead of the production technology. This framework avoids additional assumptions on firms behavior and their impact on the production technology. Recently, Radić et al. (2012) applied this approach to assess cost and profit efficiency of a sample of G-7 investment banks. This paper included a set of measures of risk exposure and other firm-specific and macro-related factors as environmental variables and found that those variables affect the inefficiency distribution rather than the production technology. Moreover, it was observed that omitting risk taking from the efficiency estimation leads to underestimate profit efficiency, and that liquidity and capital are the most relevant risk exposures explaining efficiency. Overall, these findings suggest that bank risk taking can be modeled as an endogenous bank characteristic without imposing additional assumptions on their behavior or technology.

Recent literature recognizes that risk exposure may also affect banks with different characteristics in different ways. Foreign banks may deal better with risk exposure given

[3]These models were developed by Hughes et al. (1996, 2001) and have been recently applied by Koetter (2008).

cheaper access to funding sources or more diversification (see Chen and Liao, 2011). Similar effects could be faced by large institutions or those operating in different markets, mainly associated to scale economies (Bos and Kool, 2006; Wheelock and Wilson, 2012). In addition, there is evidence supporting the fact that high-leveraged institutions (or lower capitalized) tend to take on more risk when they can adjust their capital structures and also in presence of market power (Borio and Zhu, 2012; Dell'Ariccia et al., 2011).

Banks with a higher risk propensity may choose to produce less fixed interest bearing loans and engage more in securities or derivatives trading, increasing their market risk exposure. Likewise, lower capitalized banks may increase their risk of insolvency due to credit losses or sudden security price deterioration (Mester, 2008). Risky banks tend to attract more deposits because bank creditors demand higher interest rate as a way to exert market discipline (Demirgüç-Kunt and Huizinga, 2004). Therefore, it is relevant not only to account for risk exposure measures as possible inefficiency drivers, but also to recognize differences in the way risk exposure may affect different banks.

In this context, the proposed SFA model with random inefficiency coefficients allows us to identify the role of bank risk-taking on driving inefficiency and different effects of risk taking on banks involved in merges and acquisitions (M&A) and banks with different size and affiliation. We account for an integral group of risk exposure covariates (i.e. risks of credit, liquidity, capital, and market), and bank characteristics related to size and affiliation (i.e. domestic or foreign-owned bank). By comparing models with fixed and random coefficients we identify the impact that risk taking has on cost and profit efficiency of each bank depending on their specific characteristics. To the best of our knowledge, this proposal constitutes the first SFA model that incorporates endogenous risk taking within a Bayesian framework and that accounts for individual effects of covariates in the inefficiency.

3.2.2.1 Evidence from the Colombian banking sector

The efficiency of the Colombian banking sector has been extensively studied because of: i) the increase on M&A in the banking industry as a result of the growing affluence of capital flows from advanced and other emerging economies; and ii) for regulatory objectives trying to identify micro and macro prudential measures to reduce bank default episodes and to mitigate contagion among financial institutions.[4]

Recent studies have applied the standard stochastic frontier approach using alternatively Cobb-Douglas or translog functions to characterize the technology and find evidence of low cost efficiency during the 90s (Clavijo et al., 2006; Estrada and Osorio, 2004). The most recent studies for the Colombian banking sector find a general improvement of both technical and cost efficiency along with a greater heterogeneity among banks. Cepeda et al. (2013) use a non-parametric frontier model to evaluate the efficiency of Colombian banks for the period 2000-2009. They found that technical efficiency gradually improved during the decade up to the global financial crisis of 2008-09, when all estimated measures of efficiency decreased and a negative productivity change was found. Also, a high heterogeneity in efficiency scores was observed among banks irrespective of their size and affiliation, and M&A were found to have a significant and positive impact on bank efficiency for merged or acquired banks.

For the same period, Galán et al. (2014c) estimated input-oriented technical efficiency using a dynamic Bayesian SFA model. It was found that foreign ownership has positive and persistent effects on efficiency, while the effects of size are positive but rapidly adjusted. High inefficiency persistence in Colombian banks with important differences between institutions were also identified. In particular, merged banks were found to exhibit low costs of adjustment that allow them to recover rapidly the efficiency losses derived from merging processes.

Moreno and Estrada (2013) studied the role of market power in explaining efficiency gains in Colombian banks during the 2004-2012 period. By using alternative SFA and nonparametric models, it was found that there is a positive relationship between market power and efficiency, which is explained by the product differentiation that allows banks to gain in efficiency while they do not charge excessive credit prices.

However, none of these studies have yet incorporated the role of bank-risk-taking on the banking production to estimate efficiency. As mentioned in the previous section risk taking plays a major role in explaining the inefficiency of banks and it should be properly accounted for in efficiency estimations.

[4] In particular, several bank regulatory measures were adopted by the end of nineties as a result of the Mexican and Asian crisis that also affected the Colombian banking sector (see Clavijo et al., 2006, for a detailed review on the evolution of M&A, regulation and performance of the Colombian banking industry during the regional and local financial crises).

3.2.3 Data and empirical model

We employ annual data from 31 commercial banks for the period 2002-2012. This is an unbalanced panel data set from the local central bank (*Banco de la República*) and the financial supervisory agency (*Superintendencia Financiera de Colombia*). We follow the financial intermediation approach in which banks employ deposits, labor and physical capital to produce loans, securities investments and other financial services (Berger, 2007).[5] We consider as input prices: the price of deposits (p_1), which is the ratio of interest expenses divided by total deposits; the price of labor (p_2), which is personnel expenses divided by the total number of employees, and the price of physical capital (p_3), which is calculated as the ratio of operating expenses (i.e. non-interest reduced by personnel) to total fixed assets. As outputs we consider: loans (y_1) including consumer, commercial, mortgage, and microcredit; securities (y_2), which includes public and private bonds holdings, and other securities investments; and off-balance-sheet (OBS) activities (y_3) measured as the ratio of non-interest income over total income. Non-interest income includes securitization, brokerage services, and management of financial assets for clients which represent an important source of income for Colombian banks.[6] Total costs are considered as the sum of interest and non-interest costs and total profit as the earned net profit.

We define a set of bank-specific characteristics including: size (z_1), measured as the level of total assets; and foreign ownership (z_2), which is a binary variable taking the value of 1 if more than 50% of bank shares are foreign owned, and 0 otherwise. As aforementioned, these effects have been found to be relevant inefficiency drivers in previous studies.

Additionally, we include several specific measures of credit, liquidity, capital and market risk according to recent literature, the Colombian financial regulation and the Basel III standards. Credit risk (z_1^*) is measured as risky loans over total loans. We use risky loans instead of NPLs because it is a measure of ex-ante credit risk assumed by banks when they assign loans, which is based on internal loan ratings associated to their

[5]Hughes and Mester (1993) provide evidence that confirm that deposits should be treated as inputs (see Sealey and Lindley, 1977, for a discussion on the intermediation approach).

[6]In a recent study, Tabak and Tecles (2010) find that omitting OBS as an output over (under) estimate cost (profit) efficiency results.

probability of default.[7] In addition, the regulation establishes that loan losses provisions are required for each loan according with its rating. Thus, higher credit risk exposure is associated to more provisions for potential loan losses. As we mentioned before, the use of NPLs in the estimation of bank efficiency may lead to biased estimates of bank technology (Malikov et al, 2013). Liquidity (z_2^*) is measured as liquid assets over total assets.[8] Higher liquid assets prevent banks from losses due to rapid price deterioration and also for the maturity mismatch. Capital risk exposure (z_3^*) is measured as capital equity over total assets. Capital risk is considered as a proxy for regulatory conditions that may affect bank inefficiency. Lower capitalization is usually associated to higher inefficiency.[9] Finally, market risk exposure (z_4^*) is measured as securities investments over total assets. Operating costs associated to securities investments are generally lower that those involved in monitoring and assessing of loans which may induce less efficient banks to engage on more securities investments. However, higher holdings of securities by banks also entail higher market risk exposure.

Table 3.4 exhibits the summary statistics of the main variables described above, where all monetary values are expressed in thousands of U.S. dollars at constant prices of year 2012.

Table 3.4: Summary statistics

Variable	Mean	SD	Min	Max
Total loans	3,342,012	4,206,436	11,553	28,267,020
Securities	1,265,349	1,339,794	563	6,461,458
OBS	0.0354	0.0299	0.0266	0.0587
Price of deposits	0.0248	0.0121	0.0009	0.0923
Price of labour	36.44	22.30	3.13	142.03
Price of capital	1.92	2.66	0.29	17.30
Total assets	5,503,680	6,425,746	39,699	41,786,469
Credit risk exposure	0.0988	0.0667	0.0019	0.3839
Liquidity risk exposure	0.2296	0.0667	0.0019	0.3839
Capital risk exposure	0.1211	0.0757	0.0448	0.7854
Market risk exposure	0.2381	0.1368	0.0013	0.7478
Total cost	1,132,776	1,402,621	15,673	7,722,227
Total profit	76,927	377,974	- 784,642	2,809,771

Source: Colombian central bank and financial supervisory agency.

[7] This measure of ex-ante credit risk has been used in the literature to identify bank risk-taking in the credit market (Ioannidou and Penas, 2010).

[8] Liquid assets include cash holdings, negotiable and available to sell public and private debt instruments and pledged collateral in repurchase agreement operations.

[9] We use both Tier I and Tier II capital requirements as measure of capital equity.

We evaluate cost and profit efficiency for the Colombian banking sector. Thus, we use cost and profit functions for the frontier specification in (3.4), and we choose to represent them with translog multi-product functions. The estimated model is:

$$\begin{aligned}
\ln c_{it} = \quad & \beta_0 + \textstyle\sum_{m=1}^{3} \beta_m \ln y_{m_{it}} + \sum_{r=1}^{2} \delta_r \ln p_{r_{it}} + \frac{1}{2} \sum_{m=1}^{3} \sum_{n=1}^{3} \beta_{mn} \ln y_{m_{it}} \ln y_{n_{it}} \\
& + \tfrac{1}{2} \textstyle\sum_{r=1}^{2} \sum_{s=1}^{2} \delta_{rs} \ln p_{r_{it}} \ln p_{s_{it}} + \sum_{m=1}^{3} \sum_{r=1}^{2} \eta_{mr} \ln y_{m_{it}} \ln p_{r_{it}} + \kappa_1 t \\
& + \tfrac{1}{2} \kappa_2 t^2 + \textstyle\sum_{m=1}^{3} \phi_m t \ln y_{m_{it}} + \sum_{r=1}^{2} \varphi_r t \ln p_{r_{it}} + \sum_{j=1}^{4} \omega_j z^*_{j_{it}} + v_{it} + uit \\
& v_{it} \sim N(0, \sigma_v^2) \\
& u_{it} \sim \exp(\lambda_{it}) \\
& \lambda_{it} = exp(\gamma_0 + \textstyle\sum_{h=1}^{2} \gamma_h z_{h_{it}} + \sum_{j=1}^{4} \gamma^*_{j_i} z^*_{j_{it}}),
\end{aligned} \tag{3.5}$$

where c_{it} is the total cost or the total profit, y are outputs, p are input prices and t is a time trend in order to account for technological change. Linear homogeneity of the cost function is achieved by normalizing total costs and input prices by the price of capital (p_3). We include two types of heterogeneity variables: i) those related to size (z_1) and foreign ownership (z_2), which are modeled in the inefficiency distribution and have common effects to all banks; and, ii) those capturing banks risk-exposure ($z_1^*, z_2^*, z_3^*, z_4^*$), which may be included either in the frontier or in the inefficiency. In the latter case, they are able to be modeled either with common or firm-specific effects on banks inefficiency. In order to overcome the problem of calculations of logarithms of negative profits, we correct profit values by a factor corresponding to the absolute value of the lowest profit plus one (see Tecles and Tabak, 2010). Symmetry of the cross-effects is accomplished by imposing $\beta_{mn} = \beta_{nm}, \delta_{rs} = \delta_{sr}$.

3.2.4 Results

From the general model in (3.5) we estimate four models intended to evaluate cost efficiency (C1 to C4) and four models assessing profit efficiency (P1 to P4). Models C1 and P1 do not include any risk-exposure variable, so $\omega_1, \omega_2, \omega_3, \omega_4 = 0$ and $\gamma^*_{1i}, \gamma^*_{2i}, \gamma^*_{3i}, \gamma^*_{4i} = 0$. Models C2 and P2 include the risk-exposure variables only in the frontier and then $\gamma^*_{1i}, \gamma^*_{2i}, \gamma^*_{3i}, \gamma^*_{4i} = 0$. Models C3 and P3 include the risk covariates only in the inefficiency but restrict them to have a common effect on the inefficiency of all banks; thus, $\omega_1, \omega_2, \omega_3, \omega_4 = 0$ and $\gamma^*_{1i}, \gamma^*_{2i}, \gamma^*_{3i}, \gamma^*_{4i} = \gamma^*_1, \gamma^*_2, \gamma^*_3, \gamma^*_4$. Finally, our proposed specification to model random inefficiency coefficients is estimated in models C4 and P4 ($\omega_1, \omega_2, \omega_3, \omega_4 = 0$). This allows the effects of risk exposure to be different among banks.

Table 3.5: Posterior mean and standard deviation of parameter distributions in cost models

Parameter	**Model C1**		**Model C2**		**Model C3**		**Model C4**	
	No risk		Risk in frontier		Risk in inefficiency		Random coefficients	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
β_0	5.6560	0.6469	6.1830	0.5926	5.8790	0.5956	5.3440	0.7086
β_1	0.0533	0.0398	0.1317	0.0391	0.0296	0.0185	0.0873	0.0593
β_2	0.0927	0.0652	0.0995	0.0719	0.0792	0.0677	0.0401	0.0316
β_3	0.0475	0.0349	0.0315	0.0214	0.0516	0.0329	0.0571	0.0425
β_{11}	0.0712	0.0275	0.1238	0.0257	0.0780	0.0304	0.0873	0.0415
β_{12}	0.0186	0.0275	-0.0752	0.0296	0.0119	0.0300	0.0022	0.0383
β_{13}	-0.0048	0.0030	-0.0014	0.0027	-0.0044	0.0031	-0.0029	0.0033
β_{22}	0.0116	0.0241	0.1217	0.0322	0.0157	0.0255	0.0033	0.0369
β_{23}	0.0018	0.0020	-0.0014	0.0019	0.0015	0.0020	0.0012	0.0021
β_{33}	0.0011	0.0010	0.0005	0.0010	0.0010	0.0010	0.0014	0.0009
δ_1	0.1544	0.1243	0.2440	0.1524	0.1484	0.1025	0.0959	0.0812
δ_2	0.1802	0.1620	0.1125	0.0704	0.1726	0.1272	0.1515	0.1315
δ_{11}	0.2195	0.0596	0.1356	0.0689	0.1866	0.0707	0.0397	0.0977
δ_{12}	-0.2212	0.0430	-0.1651	0.0434	-0.2062	0.0453	-0.1467	0.0580
δ_{22}	0.2010	0.0518	0.1204	0.0519	0.1857	0.0528	0.1831	0.0523
η_{11}	0.1508	0.0257	0.1401	0.0253	0.1492	0.0270	0.1623	0.0379
η_{12}	-0.0298	0.0234	-0.0216	0.0205	-0.0289	0.0239	-0.0354	0.0310
η_{21}	-0.0175	0.0252	-0.0473	0.0251	-0.0283	0.0311	-0.0979	0.0365
η_{22}	-0.0836	0.0235	-0.0521	0.0220	-0.0752	0.0256	-0.0439	0.0292
η_{31}	0.0012	0.0039	0.0007	0.0032	0.0017	0.0037	0.0033	0.0038
η_{32}	0.0035	0.0036	0.0048	0.0031	0.0023	0.0038	-0.0016	0.0041
κ_1	-0.3458	0.1273	-0.3620	0.1074	-0.3589	0.1279	-0.3092	0.1086
κ_2	0.0022	0.0046	-0.0007	0.0045	0.0022	0.0046	0.0051	0.0038
ϕ_1	0.0364	0.0114	0.0355	0.0099	0.0373	0.0115	0.0364	0.0105
ϕ_2	-0.0344	0.0085	-0.0300	0.0077	-0.0349	0.0085	-0.0359	0.0075
ϕ_3	0.0002	0.0008	0.0000	0.0007	0.0001	0.0008	-0.0006	0.0007
φ_1	-0.0401	0.0141	-0.0424	0.0131	-0.0417	0.0142	-0.0388	0.0133
φ_2	0.0167	0.0122	0.0108	0.0106	0.0166	0.0120	0.0099	0.0106
ω_1 (credit)			0.1344	0.3519				
ω_2 (liquidity)			-0.1887	0.2062				
ω_3 (capital)			1.7340	1.0197				
ω_4 (market)			-2.0900	1.1383				
γ_0	1.0350	0.4927	0.5318	0.2616	0.8946	0.7756	1.0090	0.6995
$\gamma_1 (\ln assets)$	-0.1981	0.0531	-0.1731	0.0574	-0.2318	0.0693	-0.1595	0.0873
γ_2 (foreign)	-0.8144	0.3824	-1.1700	0.5051	-0.6873	0.2202	-0.1918	0.0856
γ_1^* (credit)					0.2002	1.0850	0.3162	0.7598
γ_2^* (liquidity)					0.3692	0.6539	-1.9667	1.5432
γ_3^* (capital)					1.5380	0.6128	2.9476	0.9247
γ_4^* (market)					0.0432	1.0600	-0.0168	1.1441
Posterior eff.	0.8934	0.0653	0.9092	0.0570	0.7923	0.1466	0.7102	0.2251
DIC_3	2982.76		2916.44		2497.59		2007.75	
LPS	-9.62		-29.34		-65.19		-90.67	

Table 3.6: Posterior mean and s.d. of parameter distributions in profit models

Parameter	**Model P1**		**Model P2**		**Model P3**		**Model P4**	
	No risk		Risk in frontier		Risk in inefficiency		Random coefficients	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
β_0	9.7890	7.1260	9.8730	7.3150	3.6980	9.0530	7.2050	6.0130
β_1	3.0250	1.1470	3.4950	1.0970	4.0310	1.3970	2.9140	0.9397
β_2	-3.3910	0.9323	-4.1220	0.8941	-4.5860	1.0290	-3.5380	0.6589
β_3	-0.1987	0.1027	-0.2175	0.1001	-0.2612	0.1112	-0.2122	0.0797
β_{11}	-0.4570	0.1279	-0.5658	0.1180	-0.5109	0.1588	-0.3992	0.1253
β_{12}	0.2672	0.0993	0.3844	0.0952	0.3070	0.1290	0.2313	0.1010
β_{13}	0.0173	0.0076	0.0176	0.0078	0.0152	0.0073	0.0128	0.0062
β_{22}	0.0104	0.0686	-0.0778	0.0718	0.0288	0.0889	0.0477	0.0806
β_{23}	-0.0034	0.0072	-0.0050	0.0049	-0.0011	0.0067	0.0011	0.0039
β_{33}	0.0006	0.0020	0.0017	0.0020	0.0011	0.0017	-0.0014	0.0013
δ_1	-1.6780	1.8830	-2.9660	1.6890	-4.3810	2.0620	-3.3070	1.2900
δ_2	0.8204	1.2470	1.6330	1.1820	2.2720	1.2190	1.6450	0.7651
δ_{11}	0.1789	0.1828	0.1024	0.1445	-0.0292	0.1694	-0.0434	0.1290
δ_{12}	-0.0006	0.1784	0.0439	0.1391	0.0949	0.1584	0.0831	0.0942
δ_{22}	-0.0767	0.1830	-0.1739	0.1760	-0.1500	0.1826	-0.2043	0.1071
η_{11}	0.1704	0.1354	0.2448	0.1256	0.3185	0.1649	0.2352	0.0979
η_{12}	0.1229	0.0991	0.0634	0.0926	0.0364	0.1116	0.1109	0.0725
η_{21}	0.0051	0.0684	-0.0042	0.0538	-0.0443	0.0694	-0.0298	0.0461
η_{22}	-0.1236	0.0698	-0.0946	0.0602	-0.0905	0.0717	-0.1137	0.0465
η_{31}	-0.0087	0.0113	-0.0118	0.0113	-0.0157	0.0113	-0.0091	0.0071
η_{32}	-0.0055	0.0090	0.0006	0.0099	0.0040	0.0106	-0.0033	0.0066
κ_1	-0.3355	0.3613	-0.3630	0.2986	-0.5431	0.3371	-0.5106	0.2075
κ_2	0.0066	0.0109	0.0031	0.0090	-0.0052	0.0091	-0.0010	0.0065
ϕ_1	0.0658	0.0282	0.0587	0.0267	0.0825	0.0289	0.0776	0.0205
ϕ_2	-0.0225	0.0185	-0.0163	0.0167	-0.0224	0.0177	-0.0304	0.0129
ϕ_3	-0.0031	0.0018	-0.0027	0.0017	-0.0025	0.0018	-0.0018	0.0010
φ_1	0.0486	0.0381	0.0335	0.0323	0.0436	0.0327	0.0241	0.0201
φ_2	-0.0538	0.0422	-0.0466	0.0310	-0.0523	0.0339	-0.0436	0.0185
ω_1 (credit)			-0.7570	0.6517				
ω_2 (liquidity)			-1.1490	0.6117				
ω_3 (capital)			-0.4682	0.7637				
ω_4 (market)			0.1237	0.9031				
γ_0	-0.9668	0.4019	-1.4530	0.6724	-1.2240	1.1640	-1.4140	1.2740
$\gamma_1 (\ln assets)$	0.0556	0.0429	0.0826	0.0527	0.0721	0.0593	0.1407	0.0848
γ_2 (foreign)	1.0120	0.1713	1.1330	0.1966	1.0410	0.2180	1.0180	0.4408
γ_1^* (credit)					-3.2610	1.2480	-1.9774	0.6750
γ_2^* (liquidity)					0.2454	0.5587	-1.0640	1.2801
γ_3^* (capital)					1.6180	0.5611	2.1365	0.7649
γ_4^* (market)					-0.9284	0.3005	-1.9271	0.5147
Posterior eff.	0.5146	0.2638	0.5313	0.2719	0.6067	0.2818	0.6409	0.3281
DIC_3	3168.01		3085.10		2466.83		2368.70	
LPS	-180.01		-282.42		-305.94		-401.56	

Results of the models using both the cost and profit functions derived from (3.5) are presented in tables 3.5 and 3.6, where posterior means and probability intervals are showed for all the parameters. We observe that loans, investments and OBS affect cost positively in all models as well as input prices. In the case of profits, the relationship is also positive for loans and investments but negative, although not significant, for OBS. This result for OBS was also found by Tabak and Tecles (2010) in an application to the Indian banking sector. However, in that work loans and investments are also found to be not significant when OBS is included in the model in both cost and profit models. Regarding input prices the coefficients are not relevant in any of the profit models.

We also found decreasing returns to scale in all the models, which may suggest low margin for more M&A processes in the sector. When scale economies are analyzed by groups of banks with similar characteristics of size, ownership, and involvement in M&A we find that while big, domestic and merged institutions operate at decreasing returns to scale; small, foreign and non-merged banks present increasing returns to scale.[10] These results coincide with those reported in Galán et al. (2014c) using an input distance function, and suggest that M&A processes carried out mainly by domestic and large institutions may led them to be oversized; while small and foreign banks may still present some potential scale gains.

We observe that size and foreign ownership are important inefficiency drivers in all the models. Their effects are negative on cost inefficiency and positive on profit inefficiency. This suggests that large and foreign banks are more cost efficient but less profit efficient than their counterparts. Previous studies have also found similar effects. Chen and Liao (2011) find that foreign banks perform better than local banks because they may deal better with risk exposure given cheaper access to funding sources or more diversification. Fries and Taci (2005) find similar results for banks with majority of foreign ownership in emerging economies. Regarding size, previous studies have found that large institutions tend to exhibit greater efficiency associated to higher scale economies (see Bos and Kool, 2006; Glass and McKillop, 2006; Hughes and Mester, 2013; Wheelock and Wilson, 2012).In previous applications to Colombian banks, both foreign and large banks have been also found to be more cost efficient than local and small banks (see Galán et al., 2014c; Moreno and Estrada, 2013). However, the fact that larger banks are found to operate on decreasing returns to scale while they exhibit higher cost efficiency may

[10]Small and large banks are those below and above the median of the total assets, respectively.

suggest that those banks benefit from lower funding costs. This relatively advantage over smaller banks has been recently reported in the literature as evidence of the *too-important-to-fail* dilemma where larger banks take advantage of their size for funding at lower cots and for taking on more risk (IMF, 2014; Santos, 2014). Bertay et al. (2013) analyzed a large sample of banks for 90 countries during the period 1992-2011 and found that banks interest costs tend to decline with systemic size. This result was found for all banks except for those with very low capitalization level. Interestingly, as it will be mentioned further, we also find that lower capitalization (i.e. higher capital risk exposure) is associated with lower cost and profit efficiency.

Regarding the effects of risk exposure, posterior results for risk coefficients and model comparison indicators lead to similar conclusions in both the cost and profit models.[11] That is, models including risk exposure improve from a model omitting these variables; and, from these models, the one including these covariates in the inefficiency distribution exhibit better fit and predictive performance. Moreover, no important effects of any risk measure are observed when they are modeled in the frontier, while some risk exposures are found to be very relevant when they are included in the inefficiency distribution. This suggests that risk-taking is an important driver of banks inefficiency. Also, cost and profit efficiency are found to be over and under estimated, respectively, when risk exposure measures are not modeled in the inefficiency distribution.

We identify that greater capital risk exposures lead to lower cost and profit efficiency. There is evidence showing that highly capitalized banks tend to be more efficient than thinly capitalized banks (Kwan and Eisenbeis, 1997). This may be associated to the fact that lower capitalized banks may increase their risk of insolvency due to credit losses or sudden security price deterioration.[12]

Credit and market risks are also found to be key drivers of profit efficiency, providing evidence in favor of risk-taking in both the credit and securities markets. This may be related to the *skimping* hypothesis in Berger and DeYoung (1997), showing that when banks relax credit standards (or expend lower resources in the analysis of loan applications) they try to increase their quantity of loans and perform better (more profitable). This ex-ante credit risk is reflected in a growing proportion of risky loans

[11]Lower values for DIC_3 and LPS indicate better fit and predictive performance.

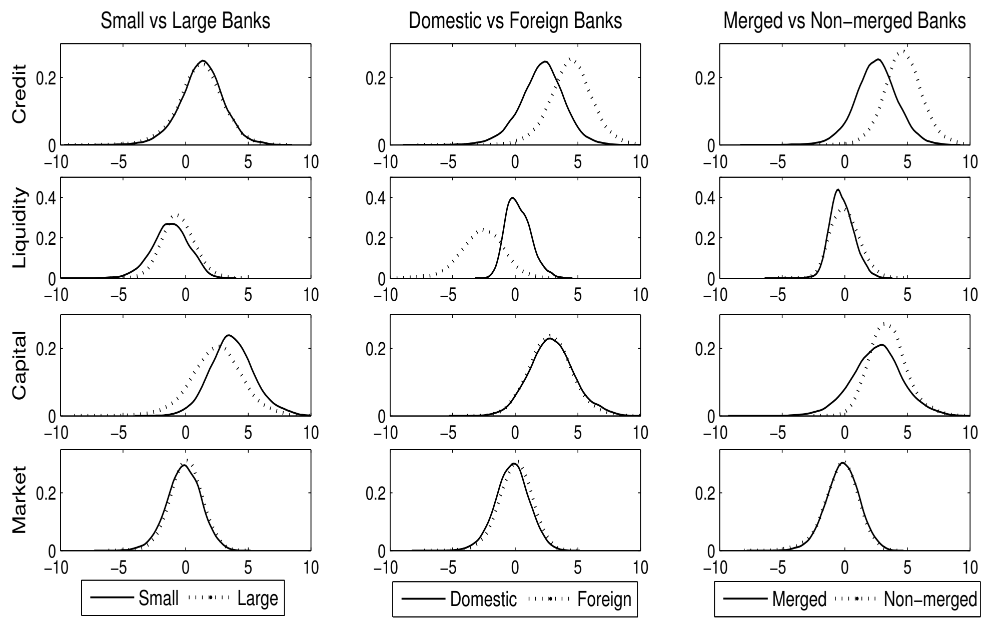
[12]See evidence for the U.S banks in Hughes and Mester (2001) and for German banking industry in Koetter (2008).

which also tend to increase during periods of lower interest rates. A similar strategy is adopted by banks in the securities market. Banks with a higher risk appetite may choose to produce less fixed interest bearing loans and engage more in securities, increasing their market risk exposure (Mester, 2008).

3.2.5 Analysis of risk random coefficients

Results of DIC_3 and LPS favour our proposed inefficiency specification with random coefficients for risk covariates in both cost and profit models. These results suggest not only that risk exposure measures are important inefficiency drivers but also that risk has different effects on cost and profit inefficiency of banks with different characteristics.

Figure 3.4: Average posterior distributions of risk random coefficients by groups of banks under cost model C4

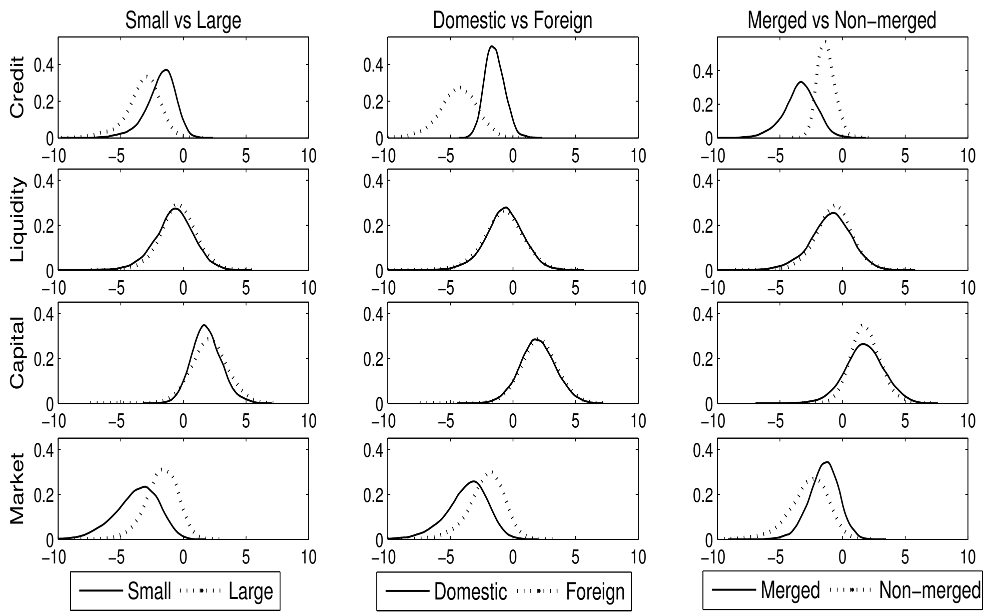


Figures 3.4 and 3.5 exhibit the posterior distributions of the four risk exposure bank-specific coefficients averaged by groups of banks, in models C4 and P4, respectively. The analysis is performed by groups of banks with different characteristics of size (small vs large banks), ownership (domestic vs foreign banks), and involvement in M&A processes (merged vs non-merged banks).[13] We observe two main results when heterogeneity in

[13]Small and large banks are those below and above the median of the total assets, respectively.

the effects of risk on efficiency is accounted for: firstly, some groups of banks are more affected than others at the same risk exposure levels; and secondly, some types of risk become relevant as inefficiency drivers for some groups of banks.

Figure 3.5: Average posterior distributions of risk random coefficients by groups of banks under profit model P4



In particular, credit risk is identified as a key cost inefficiency driver for foreign and non-merged banks, in the sense that higher exposures to credit risk reduce cost efficiency of these types of banks. Likewise, liquidity exposure is only identified as having relevant negative effects on cost inefficiency of foreign banks. Moreover, the effect on domestic banks is almost nonexistent. This result indicates that holding less liquid assets is more costly for foreign banks which could be associated to their incentives to engage on more risk taking. Regarding capital risk exposures, the positive effects on cost inefficiency are similar between types of banks. However, increasing capital in the same proportion is more likely to affect non-merged institutions. Finally, market risk is not relevant for any of the analyzed types of banks following the conclusion obtained from the fixed coefficients model.

Regarding bank-specific effects of risk on profit efficiency, it is observed in Figure 3.7 that credit risk affects more large, foreign and merged banks. Thus, these types of banks benefit more by assuming the same credit risk exposures than their counterparts. On the other hand, small, domestic and non-merged banks find more benefits when they increase market risk in the same proportions. As to liquidity and capital exposures, no differences are identified between these groups of banks. As in the fixed coefficients model, liquidity continues to be non-relevant explaining profit efficiency; while increasing capital has similar positive effects on profit inefficiency for all banks.

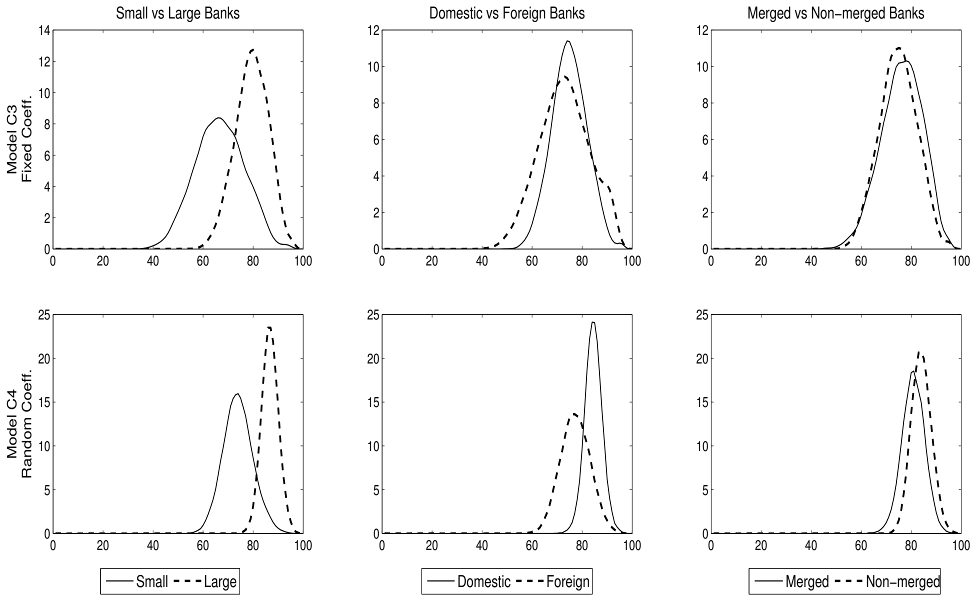
3.2.6 Analysis of efficiency

The most important changes in the posterior efficiency estimations are observed when the model is allowed to estimate bank-specific risk coefficients in the inefficiency distribution. In these cases, the average posterior cost efficiency decreases and profit efficiency increases with respect to the other models. This suggests that considering heterogeneity in the way risk affects inefficiency has important effects on estimations. It is also observed that in both cost and profit models, the dispersion of the posterior efficiency presents important increases in the random coefficient models. This suggests that these models are recognizing differences between banks in terms of their risk exposure and that these differences have effects on their efficiency estimations.

Figures 3.6 and 3.7 exhibit the average posterior distributions of cost and profit efficiency, respectively. Posterior efficiency is analyzed by groups of banks and results from models with fixed and random coefficients of risk covariates are presented.

In general, bank specific characteristics are found to be relevant factors that differentiate banks performance. Large banks exhibit higher costs efficiency levels than small banks in both fixed and random coefficients models. However, the random coefficient models show a higher difference among large and small institutions. A possible explanation for the differences between banks with different size may be associated to the fact that large banks are considered by creditors as *too-important-to-fail* and then, they are willing to offer funds at lower costs. In the case of small banks the result can be seen as opposite in the sense that creditors and depositors may ask for higher return as a way to exert market discipline (see evidence in Hughes and Mester, 2013; Wheelock and Wilson, 2012). Regarding affiliation, domestic institutions present higher costs efficiency than foreign banks but this difference is only important under the random coefficients model, suggesting that domestic banks benefit from the differences in the way credit and liquidity risk affect these banks. On the other hand, no important differences are

Figure 3.6: Average posterior distributions of cost efficiency by groups of banks in Models C3 and C4

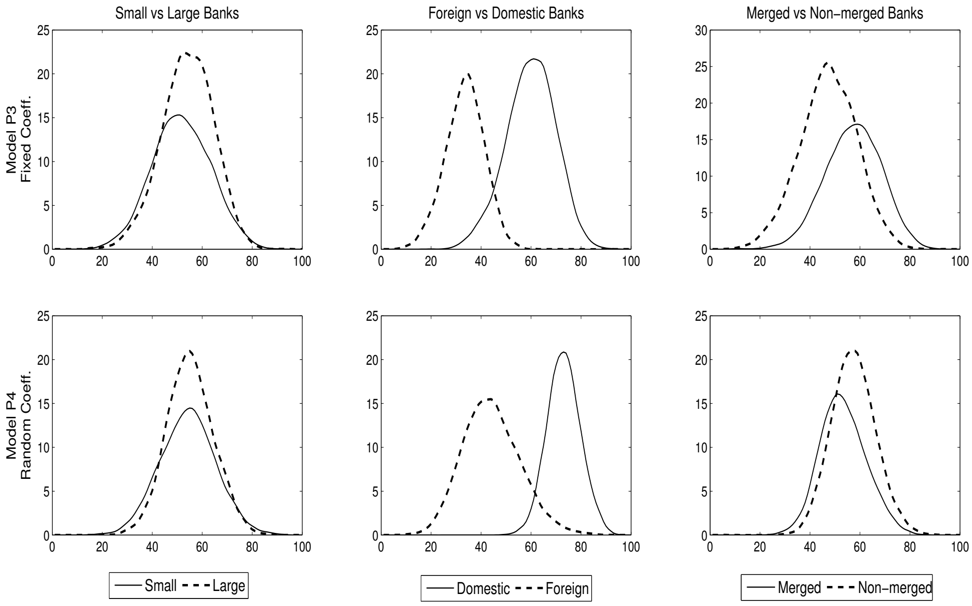


observed in cost efficiency between merged and non-merged banks. As presented further, differences are more evident in the evolution of their efficiency.

Regarding profit efficiency, the random coefficients model identifies some differences in the location and dispersion of the posterior efficiency distributions. The main difference in profit efficiency is observed between domestic and foreign banks. Domestic banks almost double profit efficiency of foreign banks in both models.

Finally, focusing on the results of our preferred models with random coefficients, the evolution of cost and profit efficiency over time is presented in Figure 3.8 by groups of banks. Small banks have been more volatile in both cost and profit efficiency over time, while large banks have been more stable and present higher cost efficiency during all the period. This may suggest that large banks are less sensitive to environmental conditions. Foreign banks present lower profit efficiency with the lowest value presented in 2008 coinciding with the global financial crisis. This suggest that foreign institutions could be affected by their operations and investments in international markets. Nevertheless, in the last years, foreign banks exhibit an increasing trend in profit efficiency and their scores are very close to those of local banks.

Figure 3.7: Average posterior distributions of profit efficiency by groups of banks in Models P3 and P4

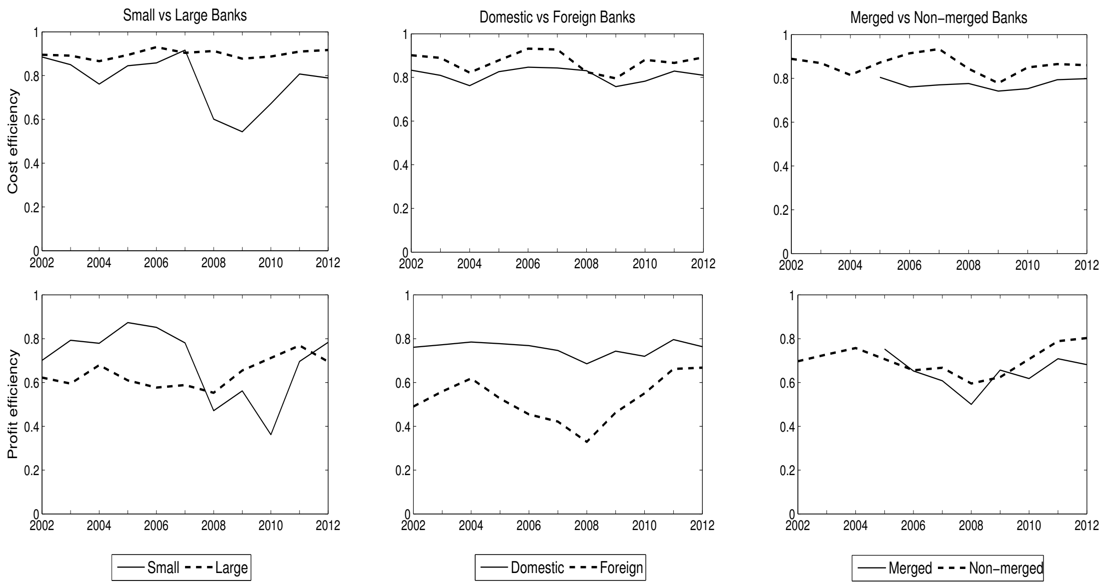


Regarding merged banks, we observe that they present decreases in cost and profit efficiency in the first years after these processes have been carried out (2005 - 2006). However, in the last two years they seem to recover part of these losses. This pattern was also found in Galán et al. (2014c) by using an input-oriented technical efficiency approach under a dynamic SFA model. They find that merged banks are able to recover very fast their efficiency levels and present higher efficiency than non-merged institutions due to lower adjustment costs. Cuesta and Orea (2002) had also found a similar pattern in merged Spanish banks after evaluating output-oriented technical efficiency. Here, we find that these effects are even more evident when we assess integrally costs and revenues in a profit efficiency analysis.

3.2.6.1 Empirical implications

Our findings remark the importance of considering different types of bank risk exposure as cost and profit inefficiency drivers. In particular, large and foreign banks exhibit higher costs efficiency, which can be associated to scale economies but also to "too-big-too-fail" considerations that benefit large banks from lower deposit and funding costs.

Figure 3.8: Evolution of mean posterior cost and profit efficiency by groups of banks in random coefficient models



Regarding risk effects on inefficiency, we find that greater exposures to credit and market risks are found to be key drivers of profit efficiency. These findings suggest that banks may have incentives for risk-taking in both the credit and securities markets. We also find evidence to support that lower capital risk conduct to higher efficiency in both costs and profits. Finally, our proposal to include random coefficients in the inefficiency capture differences in the way risk affects cost and profit efficiency of banks involved in M&A processes, and banks with different size and type of ownership. We identify large, foreign and merged banks to benefit more by assuming the same credit risk exposures than their counterparts; while small, domestic and non-merged banks institutions to take advantage of assuming higher market risk.

Regulators should take into account not only the impact of requirements in capital and liquid assets on cost and profit efficiency of banks, but also that these policies have different, and sometimes opposite, effects on banks with different characteristics. Moreover, the fact that large and merged banks are found to face lower costs and to have incentives to take on more risk in credit and securities markets constitute a signal for regulators to be alert on these institutions. Regulators should also consider alternative measures to limit risk taking incentives associated to the fact that large banks exploit the benefits from being considered as *too-important-to-fail.* This is even more important given the recent local expansion of financial conglomerates, which makes more difficult for regulators to monitor their behavior and may enhance regulatory arbitrage and expose other market participants (non-banking institutions) to higher risk exposure, boosting systemic risk. Work is currently in progress on this area as well as on the relationship between risk taking and the *too-important-to-fail* dilemma in the interbank market.

3.3 Conclusions

In stochastic frontier analysis, unobserved inefficiency heterogeneity has been little explored. In this chapter we have put forward the modeling of heterogeneity in a Bayesian context by proposing two alternative methods to capture unobserved sources of heterogeneity in the inefficiency distribution: i) through a random parameter which can be allowed to be time-varying depending on the application, and ii) through firm-specific random coefficients of observed covariates.

In the first case, the effects of its inclusion in different parameters of a truncated normal distributed inefficiency were studied. Our findings suggest that unobserved inefficiency heterogeneity can be properly captured by a random parameter. Models including this parameter whether alone or simultaneously with observed covariates improve in terms of fit and predictive performance as long as latent heterogeneity remains unidentified. In this sense, it can be used to distinguish unobserved heterogeneity from inefficiency and to validate the suitability of observed covariates to capture it. As observed in Chapter 2, differences in efficiency rankings and mean scores were found when inefficiency heterogeneity was included in different parameters of a truncated normal inefficiency distribution. Also for unobserved heterogeneity, we find that its inclusion in the location parameter of the inefficiency has an effect on separating the firm specific posterior efficiency distributions from each other; while a shrinking effect on the individual posterior efficiency distributions is identified when it affects the scale parameter. In the case, of the model preserving the scaling property by affecting both the location and scale parameters, both effects are observed. Therefore, these results are consistent whether we use observed covariates or our proposal to model unobserved heterogeneity.

In general, models preserving the scaling property can be used with our proposal to capture unobserved inefficiency heterogeneity. This allows to decompose inefficiency into a base component measuring natural managerial skills and other measuring the effect of latent factors causing unobserved heterogeneity.

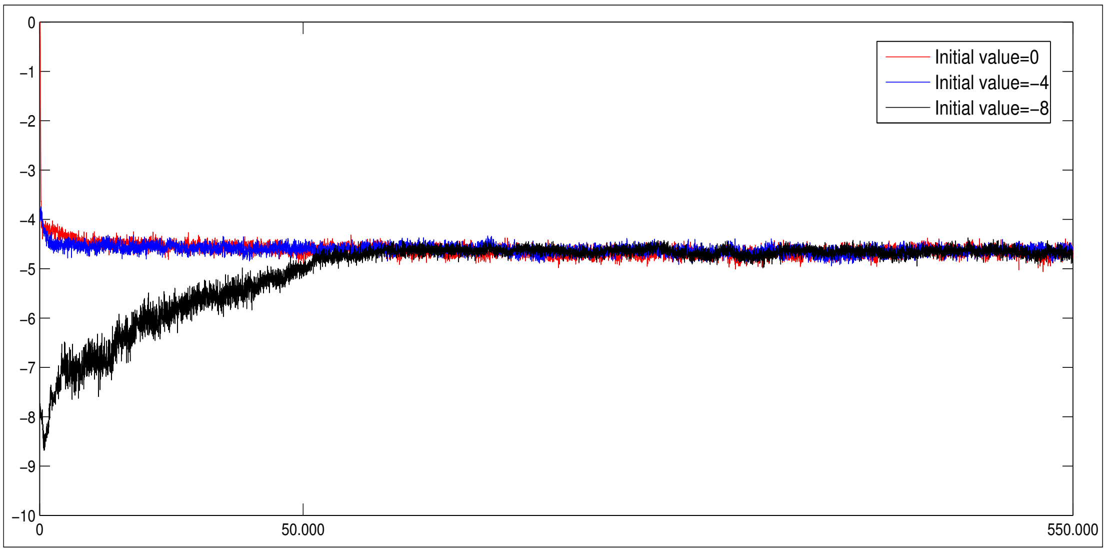
Regarding our second proposal, we find that modeling random coefficients for the inefficiency covariates captures firm-specific effects which remain unidentified under the regular fixed coefficients models. That is, the random coefficients model identifies inefficiency drivers that result as not relevant in models with fixed coefficients. Moreover, some of these heterogeneity variables are found to be very important explaining inefficiency for some firms. Also, the magnitude of the effects of these covariates may change among firms. These effects remain unobserved if they are not modeled and affect the posterior efficiency estimations. Overall, this specification distinguishes firms in terms of the effects of inefficiency drivers and separates unobserved heterogeneity related to these effects from efficiency.

3.4 Appendix

A. MCMC results for the unobserved random parameter τ

As described in Section 3.1.1, we identified very slow convergence in the MCMC simulations for the unobserved random parameter τ. We present in Figures 3.9 and 3.10 plots of the MCMC chains for the hyper-parameter $\overline{\tau}$ at three different starting points for the models preserving the scaling property (Model A in Section 3.1.2.1 for the WHO application and Model C in Section 3.1.2.2 for the airlines application). In the empirical applications we used a starting value equal to zero.

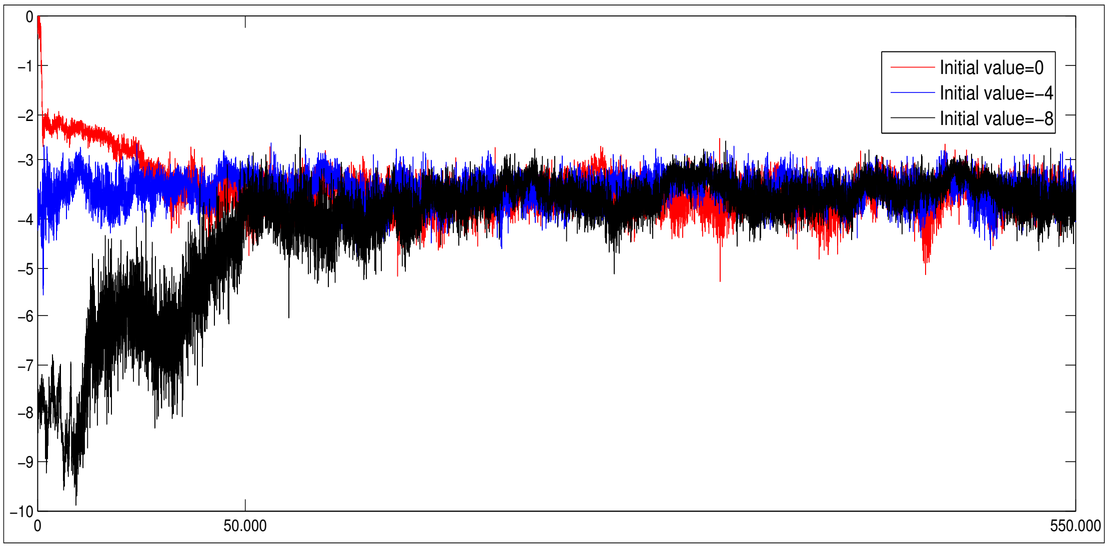
Figure 3.9: MCMC iterations for different initial values of parameter $\overline{\tau}$ - WHO application - Model A



The conclusion we got in both cases is that the parameters are identifiable but the convergence is very slow. In fact, the parameters started to converge around 40,000 iterations and we also found a high autocorrelation of order around 20. Therefore, we decided to consider a thinning equal to 25, to discard the first 50,000 iterations and run a total of 550,000 iterations. This implies that we retain data from 20,000 iterations as in models without the random parameter.[14] When the time-varying specification is used

[14] In those cases we remain running 50,000 with a thinning equal to 2 and discarding the first 10,000.

Figure 3.10: MCMC iterations for different initial values of parameter $\overline{\tau}$ - Airlines application - Model C



the main difference is that the dispersion of the MCMC iterations is higher. This was observed in all the models containing this parameter.

B. WinBUGS code a truncated normal model with a random parameter capturing unobserved heterogeneity - WHO application

```
model {
for (i in 1:N) {
 m[i]<- mu*exp(tao[firm[i]])
 sigmau[i] <- (1/sqrt(lambda))*exp(tao[firm[i]])
 #In case of using also observed covariates:
 #z[i]<- gamma[1]*z1[i]+gamma[2]*z2[i]+gamma[3]*z3[i]
 #m[i]<- mu*exp(z[i]+tao[firm[i]])
 #sigmau[i] <- (1/sqrt(lambda))*exp(z[i]+tao[firm[i]])
 nu[i]<-1/(sigmau[i]*sigmau[i])
 u[i] ~ djl.dnorm.trunc(m[i],nu[i],0,1000)
 eff[i] <- exp(- u[i]) }
for ( i in 1:N ) {
 mp[i] <- alpha - u[i]+beta[1]*x1[i]+beta[2]*x2[i]+beta[3]*x3[i]+beta[4]*zp1[i]
         +beta[5]*zp2[i]+beta[6]*ze1[i]+beta[7]*ze2[i]+beta[8]*ze3[i]
         +beta[9]*ze4[i]+beta[10]*ze5[i]+beta[11]*ze6[i]
 y[i] ~ dnorm(mp[i], prec) }
mu <- psi/sqrt(lambda)
psi ~ dnorm(0.0,1)
lambda~dgamma(5,lambda0)
lambda0 <- 5*log(rstar)*log(rstar)
#In case of using also observed covariates:
#for (i in 1:6) {
 #gamma[i] ~ dnorm(0.0, 0.1) }
#Alternative prior for gamma:
#for (i in 1:6) {
 #gammastar[i] ~ dnorm(0.0, 0.1)
 #gamma[i] <- gammastar[i] / sqrt(lambda) }
alpha ~ dnorm(0.0, 1.0E-06)
for (i in 1:5) {
 beta[i] ~ dnorm(0.0, 1.0E-06) }
#For firm specific random parameter:
for (i in 1:K){
 tao[i] ~ dnorm(mutao,prectao) }
mutao ~ dnorm(0.0, 0.1)
prectao~ dgamma(0.5, 0.5)
prec ~ dgamma(0.001, 0.001)
sigmasq <- 1 / prec  }
```

C. WinBUGS code for a model with random inefficiency coefficients - Banks risk-taking application

```
model{
for (i in 1:N) {
 z[i]=gamma[1]*z1[i]+gamma[2]*z2[i]
 zstar[i]=gammai[1,firm[i]]*z3[i]+gammai[2,firm[i]]*z4[i]+gammai[3,firm[i]]*z5[i]
         +gammai[4,firm[i]]*z6[i]
 lambda[i]<-exp(gamma0+z[i]+zstar[i])
 u[i] ~ dexp(lambda[i])
 eff[i]<- exp(-u[i]) }

for (i in 1:N) {
#+u for costs and -u for profit
 mu[i] <- alpha - u[i]+ beta[1]*lny1[i]+beta[2]*lny2[i]+beta[3]*lny3[i]
         +delta[1]*lnp1[i]+delta[2]*lnp2[i]+betamn[i]+deltars[i]
         +etamr[i]+kappa[1]*t[i]+0.5*kappa[2]*t[i]*t[i]+tcross[i]
 betamn[i] <- 0.5*betam[1]*lny1[i]*lny1[i]+betam[2]*lny1[i]*lny2[i]
         +betam[3]*lny1[i]*lny3[i]+0.5*betam[4]*lny2[i]*lny2[i]
         +betam[5]*lny2[i]*lny3[i]+0.5*betam[6]*lny3[i]*lny3[i]
 deltars[i] <- 0.5*deltar[1]*lnp1[i]*lnp1[i]+deltar[2]*lnp1[i]*lnp2[i]
         +0.5*deltar[3]*lnp2[i]*lnp2[i]
 etamr[i] <- eta[1]*lny1[i]*lnp1[i]+eta[2]*lny1[i]*lnp2[i]+eta[3]*lny2[i]*lnp1[i]
         +eta[4]*lny2[i]*lnp2[i]+eta[5]*lny3[i]*lnp1[i]+eta[6]*lny3[i]*lnp2[i]
 tcross[i] <- phi[1]*t[i]*lny1[i]+phi[2]*t[i]*lny2[i]+phi[3]*t[i]*lny3[i]
         +varphi[1]*t[i]*lnp1[i]+varphi[2]*t[i]*lnp2[i]
 y[i] ~ dnorm(mu[i], prec) }

lambda0 <- -log(rstar)
gamma0<-log(expgamma0)
expgamma0~dexp(lambda0)
for (j in 1:2) {
 gamma[j]<- log(expgamma[j])
 expgamma[j] ~ dexp(lambda0) }
for (j in 1:4) {
 gammastar[j]~dexp(lambda0)
 for (i in 1:K) {
  expgammai[j,i]~dexp(gammastar[j])
  gammai[j,i]<- log(expgammai[j,i]) } }

alpha ~ dnorm(0.0, 0.001)
```

```
for (i in 1:3) {
 beta[i] ~ dnorm(0.0, 0.001) }
for (i in 1:2) {
 delta[i] ~ dnorm(0.0, 0.001) }
for (i in 1:6) {
 betam[i] ~ dnorm(0.0, 0.001) }
for (i in 1:3) {
 deltar[i] ~ dnorm(0.0, 0.001) }
for (i in 1:6) {
 eta[i] ~ dnorm(0.0, 0.001) }
for (i in 1:2) {
 kappa[i] ~ dnorm(0.0, 0.001) }
for (i in 1:3) {
 phi[i] ~ dnorm(0.0, 0.001) }
for (i in 1:2) {
 varphi[i] ~ dnorm(0.0, 0.001) }

prec ~ dgamma(0.01, 0.01)
sigmasq <- 1 /prec  }
```

Chapter 4

Inefficiency Heterogeneity in Dynamic Models*

The decision making process followed by producers is dynamic in nature. Technology and environment change continuously and variations with respect to their current production conditions have to be considered by firms. However, firms face restrictions and costs in the adjustment process. Regulation, quasi-fixed or indivisible inputs, and transaction, information and other adjustment costs are important factors preventing firms from making free and instant adjustments towards optimal conditions. In this context, firms may not only be inefficient at some point, but this inefficiency may persist from one period to the next, and firms may find it optimal to remain partly inefficient in the short-run.

This issue has been little studied in the efficiency measurement literature but has recently become an important concern. In stochastic frontier models, we can find two alternative approaches to deal with time dependent inefficiencies. The first of these defines deterministic time specifications for the evolution of efficiency. As examples we find the proposals by Kumbhakar (1990) and Battese and Coelli (1992) where a time invariant inefficiency measure is multiplied by a parametric function of time, the model by Cornwell et al. (1990) that defines producer specific parameters, and the proposal by Lee and Schmidt (1993) where time dummies are used. These models have the problem of imposing arbitrary restrictions on the short-run efficiency and are not able to model

*Part of the work in this chapter has been accepted for publication in the *European Journal of Operational Research* (see Galán et al., 2014c).

firm-level dynamic behaviour. A more recent approach involves the dynamic behaviour of inefficiency by considering models that estimate long-run efficiency. These models recognize a persistence effect of firms' inefficiency over time and specify its evolution as an autoregressive process. In this context, Ahn et al. (2000) defined an error structure intended to capture the relationship between the short and long-run dynamics. This pioneer proposal has been criticized for its economic foundations and for modeling autoregressive processes on nonnegative variables. An alternative proposal that avoids this problem and argues that improvements on efficiency depend on the costs of adjustment was introduced by Tsionas (2006). This model was applied to a sample of US banks and very high inefficiency persistence was found out, suggesting the presence of high adjustment costs in the banking sector.

In this context, the model proposed by Tsionas (2006) becomes very relevant in accounting for inefficiency persistence. This model presents two main characteristics: It assumes a constant persistence parameter for all firms in the sector, and it allows the inclusion of observed heterogeneity in the inefficiency. However, unobserved sources of heterogeneity may also affect efficiency estimations under a dynamic framework. In this chapter we present an extension of the dynamic SFA specification in Tsionas (2006) in order to account for unobserved sources of heterogeneity. In particular, firm specific inefficiency persistence and unobserved technological heterogeneity. Finally, we also study the effects of including observed covariates in or out the inefficiency dynamics and propose a general specification able to model these differences.

This chapter is composed of three sections. Section 4.1 presents a dynamic SFA model able to capture unobserved sources of heterogeneity. We apply the proposed model to the electricity sector by using a sample of Colombian distribution utilities. Section 4.2 presents a inefficiency specification that allows modeling observed covariates in and out the inefficiency dynamics. The effects on the efficiency posterior distributions are assessed using a sample of Colombian banks during the last decade. Empirical results in both applications are of great interest not only for understanding the effects of the treatment of heterogeneity in dynamic models, but also for regulators and firms in these sectors. Finally Section 4.3 concludes the chapter.

4.1 A Dynamic Model with Unobserved Heterogeneity[1]

We propose a dynamic stochastic frontier model that accounts for both observed and unobserved heterogeneity sources. This is mainly an extension of the model introduced by Tsionas (2006) that combines it with other recent proposals in the literature of dynamic SFA models. In particular, the proposed specification accounts for observed firm characteristics in the inefficiency dynamics, as in Tsionas (2006), but also captures two additional sources of unobserved heterogeneity: the first one is related to differences in the adjustment costs among firms. It is possible that firms with different characteristics face different costs of adjustment. This would introduce a source of unobserved heterogeneity among firms which may have relevant effects on the efficiency estimations (see Galán et al., 2014c).

The second unobserved heterogeneity source is related to technological heterogeneity and we model it in a similar way to the dynamic model in Emvalomatis (2012). This author presented a dynamic model with no observed covariates in the inefficiency where unobserved technological heterogeneity is introduced. His findings reveal important biases in the efficiency estimations when this unobserved effects is not considered. The general model is given by the following equations:

$$y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + v_{it} - u_{it}, \qquad v_{it} \sim N(0, \sigma_v^2) \tag{4.1}$$

$$\log u_{it} = \omega + \mathbf{z}_{it}\boldsymbol{\gamma} + \rho_i \log u_{i,t-1} + \xi_{it}, \qquad \xi_{it} \sim N(0, \sigma_\xi^2), \qquad t = 2...T \tag{4.2}$$

$$\log u_{i1} = \frac{\omega + \mathbf{z}_{it}\boldsymbol{\gamma}}{1-\rho_i} + \xi_{i1}, \qquad \xi_{i1} \sim N\left(0, \frac{\sigma_\xi^2}{1-\rho_i^2}\right), \; t = 1. \tag{4.3}$$

Equation (4.1) represents the stochastic frontier, where in the case of a production function y_{it} is the output for firm i at time t, α_i is the firm specific parameter intended to capture unobserved technological heterogeneity, $\mathbf{x}_{it}$ is a row vector of the input quantities, $\boldsymbol{\beta}$ is a vector of parameters, v_{it} is the idiosyncratic error assumed to follow a normal distribution, and u_{it} is the inefficiency component. The dynamic specification for the inefficiency is represented by (4.2), where ω is a constant term, $\mathbf{z}_{it}$ is a row vector of firm specific heterogeneity variables, $\boldsymbol{\gamma}$ is a vector of parameters, ρ_i is the heterogeneous

[1]Much of the work in this section is joint work with Professor Michael Pollitt from University of Cambridge (see Galán and Pollitt, 2014a).

persistence parameter capturing, for every firm, the proportion of inefficiency that is transmitted from one period to the next, and ξ_{it} is a white noise process with constant variance σ^2_ξ, which may capture unobserved random shocks in the dynamic component. Finally, equation (4.3) represents the specification of the inefficiency in the first period and is intended to initialize a stationary dynamic process.

Stationarity is imposed by requiring the persistence parameters to satisfy $|\rho_i| < 1$. This is important in order to avoid possible divergence of $\log u_{it}$ to positive or negative infinity, which would lead to efficiencies equal to zero or to one. These results are not desirable since in the first case they would mean that completely inefficient firms remain in the market, and in the second case that firms may be fully efficient, contradicting the adjustment cost theory behind the formulation. In general, if a firm has a value of ρ_i close to 1 it would suggest that this firm presents high adjustment costs, which translates into a high proportion of inefficiency being transmitted from one period to the next. On the other hand, if this value is close to 0, a low proportion of inefficiency is persistent in time, implying that the firm may move quicker towards more optimal conditions.

The general model in (4.2) and (4.3) allows to evaluate different specifications by imposing restrictions over some parameters. If $\alpha_i = \alpha$ is assumed, then unobserved technological heterogeneity is not accounted for. If $\rho_i = \rho$ is imposed, homogeneous persistence is assumed for all companies in the sector. If $\rho = 0$ the model reduces to a static model where the inefficiency follows a log-normal distribution with firm specific mean. Finally, if no inefficiency covariates are observed, then $\boldsymbol{\gamma} = \mathbf{0}$ would be assumed.

4.1.1 Bayesian inference

Inference of the model in (4.1) till (4.3) is carried out using the Bayesian approach as earlier. In general, we assume non-informative but proper prior distributions for all the parameters. For the parameter capturing unobserved heterogeneity in the frontier we define a hierarchical structure where $\alpha_i \sim N(\alpha, \lambda^{-1}_{\alpha_i})$ and the hyperparameter $\alpha \sim N(0, \lambda^{-1}_\alpha)$. Priors for the precision parameters λ are set to 0.1 and 0.001 for the firm specific parameters and the hyperparameter, respectively. For parameters in $\boldsymbol{\beta}$ we assume a normal prior distribution $\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\Lambda}^{-1}_{\boldsymbol{\beta}})$ where $\boldsymbol{\Lambda}_{\boldsymbol{\beta}}$ is a precision diagonal matrix with priors set to 0.001 for all parameters. The variance of the idiosyncratic error component is assumed to follow an inverse gamma distribution $\sigma^2_v \sim IG(a, b)$ with priors set to 0.01 and 100 for the shape and scale parameters.

The inefficiency component as defined in (4.2) follows a log-normal distribution where $u_{it}|u_{i,t-1}, \omega, \mathbf{z}_{it}, \boldsymbol{\gamma}, \rho_i, \sigma_\xi^2 \sim LN(\omega + \mathbf{z}_{it}\boldsymbol{\gamma} + \rho_i \log u_{i,t-1}, \sigma_\xi^2)$ for $t = 2...T$. For $t = 1$, the inefficiency is distributed $u_{i1}|\omega, \mathbf{z}_{i1}, \boldsymbol{\gamma}, \rho_i, \sigma_\xi^2 \sim LN\left(\frac{\omega+\mathbf{z}_{i1}\boldsymbol{\gamma}}{1-\rho_i}, \frac{\sigma_\xi^2}{1-\rho_i^2}\right)$.

Regarding the parameters in the inefficiency, the distribution for the common constant term is $\omega \sim N(\mu_\omega, \lambda_\omega^{-1})$ with priors set to -1.5 and 1 for the mean and precision parameters, respectively. The distribution for the parameters of observed heterogeneity is: $\boldsymbol{\gamma} \sim N(\mathbf{0}, \boldsymbol{\Lambda}_\gamma^{-1})$ where $\boldsymbol{\Lambda}_\gamma^{-1}$ is a diagonal matrix of precisions with priors set to 0.1 for every precision parameter. For the persistence parameters, we impose $|\rho_i| < 1$ to assure stationarity and we define a hierarchical structure with $\rho_i = 2k_i - 1$, where $k_i \sim \beta(k, 1-k)$. The hyperparameter is distributed $k \sim \beta(r, s)$ with priors set to 0.5 for shape parameters. The variance of the inefficiency component is assumed to follow an inverse gamma distribution where $\sigma_\xi^2 \sim IG(n, d)$ with priors set to 10 and 100 for the shape and scale parameters, respectively.[2]

Sensitivity analysis is performed on priors in the inefficiency component. Different values are used for prior parameters in the distributions of ω, k and σ_ξ^2 and posterior results are found to converge to approximately the same values.[3] We also found posterior results to be robust to the use of a truncated normal distribution for parameters ρ_i and ρ.

The specification proposed accounts for firm specific effects in the frontier and the inefficiency persistence. However, firms in the sector share a common long-run dynamic component ω, common elasticities for the covariates given by $\boldsymbol{\gamma}$, and are linked through common parameters ρ and α that are present in the hierarchical structures defined.

As earlier, the models are run using the WinBUGS package. For all the estimated models we use 5,000 iterations for posterior inference. The MCMC algorithm involves 50,000 iterations with 10,000 discarded in a burn-in phase and a thinning equal to 8 is used to remove autocorrelations.

Model choice is carried out using DIC_3 and LPS as in the previous chapters.

4.1.2 Application to Colombian electricity distribution utilities

The electricity market reform introduced in Colombia in 1994 established a new structure of the sector and new conditions for private participation and competition. The

[2] This is the same prior used by Tsionas (2006) and Galán et al. (2014c).
[3] The priors used centre the efficiency prior distributions at 0.8.

reform was mainly motivated by an energy crisis suffered in 1992-1993 that caused major blackouts as a consequence of extreme droughts. This situation revealed the inefficiency and inability of the state-owned industry to satisfy an increasing demand and to deal with weather events. The regulatory reform adapted a version of the UK model with the creation of a pool where prices are settled in a bidding process. The Electric Law of 1994 created the regulatory commission *Comisión Reguladora de Energía y Gas* (CREG) and split the traditional vertically integrated and monopolistic system into the activities of generation, transmission, distribution and retailing. As a consequence, the seven major public holdings in charge of multiple activities from generation to distribution previous to the reform were divested into eleven companies performing only one of these activities and two companies involved in both generation and distribution. Although generation and distribution were allowed to be performed by the same company, limits to the amount of electricity that the distributor could buy from its own generation firm were set and separate managerial and accounting procedures were required.

However, privatization and competition have been slow processes in Colombia. After the reform only two of the new companies were fully privatized and, although in the following years several companies were open to private capital, in most of the cases private investors are minority shareholders and firms remain under the control of municipalities and regional governments. Certainly, privatization and competition have been identified as pending issues in Colombia in previous studies analyzing the effects of the first years of the reform (see Larsen et al., 2004; Pombo and Taborda, 2006).

Nevertheless, these processes have accelerated in recent years. From 2010 to 2012, the number of generating and retailing firms has increased by 23% and 32%, respectively, and most of the companies involved in these activities are classified as private-owned. In distribution, companies with a majority of public capital account for 62% of total firms and serve 51% of the total users. Currently there are 54 generation, 33 distribution and 85 retailing companies. Of the generation firms, 12 are also involved in distribution and 15 combine generation exclusively with retailing activities.[4]

In general, the effects of the reform have been positive in terms of the ability of the electricity sector to overcome extreme weather conditions and to satisfy the increasing

[4]Information provided by the national supervisory agency of public services *Superintendencia de Servicios Públicos Domiciliarios (SSPD)* in 2013.

demand. Since the reform, Colombia has not experienced blackouts in spite of some severe droughts that have affected the region during the 1997-1998 and 2009-2010 periods, and that have seriously affected neighbor countries. Moreover, Colombia has become an electricity exporter to Ecuador and Venezuela and it is currently planning to export electricity to other Central American and Caribbean countries.[5]

On the other hand, the effects of the reform in terms of energy losses and service quality have not been successful until recent years. During the first ten years of the reform, energy losses and electricity interruptions did not present reductions and were even higher than previous to the reform. Colombia also exhibited very bad performance in these aspects when compared to other countries in the region (see Dyner et al., 2006; Larsen et al., 2004). Only from 2008, can important reductions in energy losses be observed. In terms of the length of interruptions, although it is possible to identify some improvements since 2005, it is only until 2011 that significative reductions are evident. In both cases, these improvements are consequence of changes in the regulation, as is discussed further below.

Meeting the quality requirements and satisfying the increases in electricity consumption and users has required distribution companies to make important investments. In fact, capital and operational expenses have increased by more than 30% during the period 1998 - 2012. This suggests the need to study the effects of the reform and the latest regulations established by CREG on efficiency. Concerning this issue, some few previous studies have quoted the effects on efficiency of the reform in Colombia and no major gains have been identified. Pombo and Taborda (2006) use Data Envelopment Analysis (DEA) to perform an analysis of technical efficiency of Colombian distribution firms during the period from 1985 to 2001. The authors find no major changes during the period and highlight that the most efficient firms previous to the reform continue to be in the best-practice frontier but firms which were inefficient have not been able to change this condition and present even lower efficiency scores. A similar result was found by Melo and Espinosa (2005), who measure the technical efficiency of Colombian distributors from 1999 to 2003 using Stochastic Frontier Analysis (SFA). The authors find out that public companies perform better than those privately owned but that there have not been major changes in technical efficiency in the immediate years after the reform. This Colombian evidence contrasts with the effects of the electricity reforms on

[5]In 2011, Colombia exported 1.740 GWh. Information from the Ministry of Mines and Energy.

performance in other South American countries (see Mota, 2003; Pérez-Reyes and Tovar, 2009; Pollitt, 2004, 2008, for the cases of Brazil, Chile, Argentina and Peru, respectively.)

Findings from these studies may suggest the presence of high adjustment costs in the Colombian distribution sector that imply inefficiency to be highly persistent in time. In this context, it is costly for firms to move towards optimal conditions and they may find it optimal to remain inefficient in the short-run. These studies have also evidenced the existence of important differences among firms with different characteristics in terms of their performance.

Therefore, this application has two main aims: first, to identify the presence of adjustment costs in the distribution sector after the reform and distinguish heterogeneity in the technology and the inefficiency among Colombian distributors; second, to estimate measures of efficiency that consider costs and quality of service in the Colombian electricity sector and their evolution from the first years after the reform into the following fifteen years. In particular, we focus on the last five years, when most of the changes in terms of quality, demand and costs have occurred.

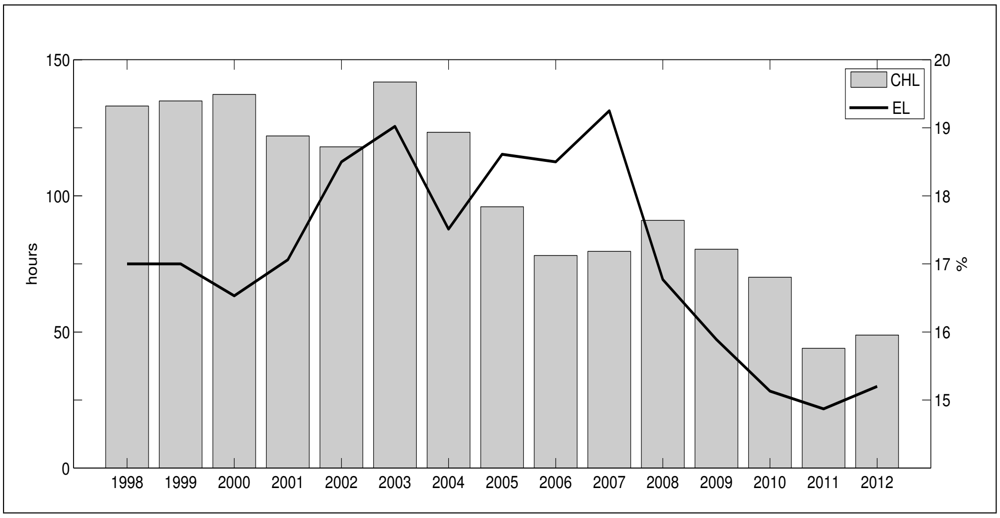
4.1.2.1 The Colombian electricity distribution sector

The activity of electricity distribution in Colombia is defined by CREG as the transportation of electricity from the national transmission system, which operates at voltages above 220 Kv, to the final user. There are four different levels of tension operated by the distributor. That is, from level 1, which involves tension levels below 1 Kv, to level 4 with tension levels between 57.5 Kv and 115 Kv. CREG follows a cost of service type of regulation and establishes the pricing formula for distributors for each of the tension levels considering demand, investments, and administration, operation and maintenance costs. The length of the price review is five years and the first pricing period was 1998-2002.[6]

Besides prices, service quality and energy losses have also been under regulation. In 1998 CREG established maximum values for both duration and number of interruptions by tension level, as well as compensations to users when companies exceeded these maximums.[7] However, small and slow improvements motivated CREG to modify this scheme in 2008. The new regulation introduced quality incentives in the pricing formula

[6]CREG resolution 031 of 1997.
[7]CREG resolution 070 of 1998.

Figure 4.1: Average CHL and EL ratio per firm

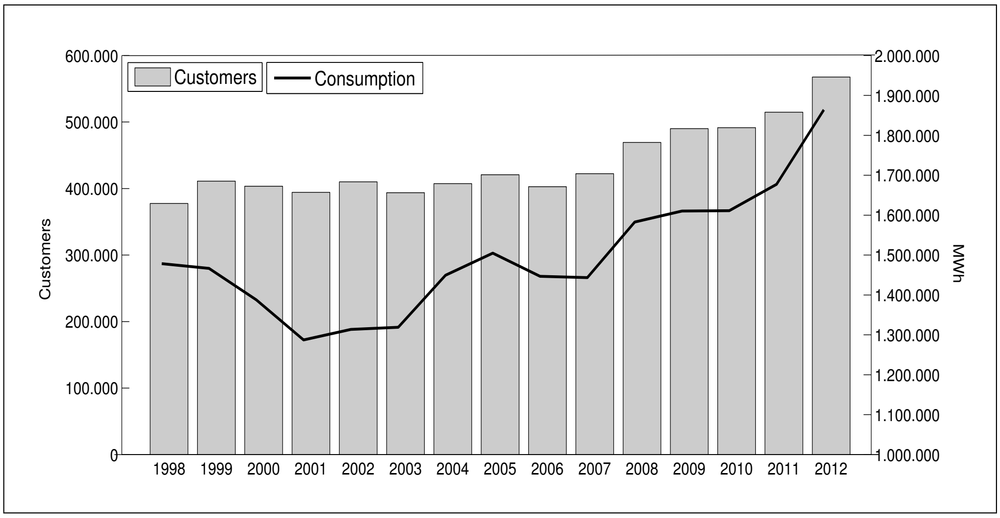


and compensations for the most affected users.[8] Under this model, an index of service discontinuity is calculated quarterly and three ranges of values for this index are set: if distribution companies exceed an acceptable range their pricing formula is revised down; if they perform better than the acceptable values their formula is revised up; and if their discontinuity index is within the acceptable range their formula does not change. The implementation of this mechanism has been postponed and only from 2011 have all companies had to report this index. The effects of this last regulatory scheme are still uncertain. In the literature, some studies have found this direct mechanism of incentive regulation to have negative effects on quality of service (see Ter-Martirosyan and Kwoka, 2010). However, the most important reductions in the length of interruptions have occurred since then. This can be observed in Figure 4.1, where the evolution in customer hours lost (CHL) and energy losses (EL) from 1998 to 2012 is presented for the sample of distribution companies described in Section 4.1.5.

Regarding energy losses, new regulations were also set by CREG in 2008 by establishing a program for reducing losses and setting upper limits for the percentage of losses recognized by users via tariff.[9] The effects of this regulation also seem to be positive (see Figure 4.1).

[8] CREG resolution 097 of 2008.
[9] CREG resolutions 199 and 121 of 2007.

Figure 4.2: Average number of customers and electricity consumption per firm



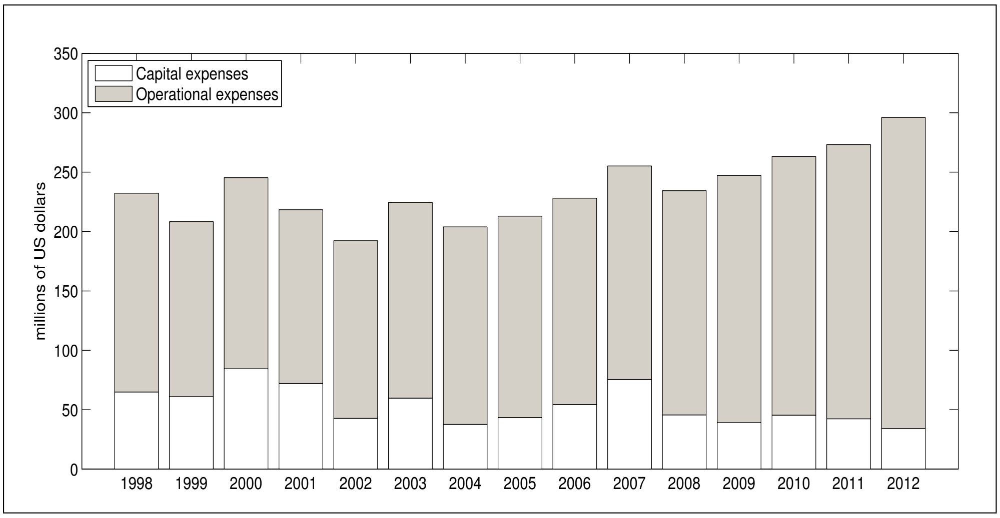
During the period 1998-2012, the electricity consumption and the number of connected users have also presented important increases (27% and 51%, respectively). Figure 4.2 presents this evolution for the same firms above. We can observe that, after a period characterized by economic recession and low growth rates (1999-2003), consumption and customers exhibit an upward trend with high growth in the most recent years.

Satisfying the demand and meeting the quality requirements have had effects on the costs of distribution firms. Figure 4.3 presents the evolution of capital and operational expenses in real US dollars of 2012 for the same companies in the figures above. We observe important increases, mainly in operational expenses, from 2007, when relatively higher capital expenses were made. The overall increase in real total expenses from 1998 to 2012 was 31%.

Higher distribution costs have had an impact on the tariff for the final user. Figure 4.4 plots the evolution of the tariff per kWh by decomposing it into each of their components. Although almost all the components of the tariff have increased in real terms, the proportion of the distribution component has raised from 33% to 40% during the period, with a particular increase in 2011 and 2012.

Regarding tariffs, it is important to remark that CREG establishes their value only for regulated users. After the reform, customers were separated into regulated and non-regulated users, which are differentiated in terms of their power demand and con-

Figure 4.3: Average operational and capital expenses per firm



sumption. Since 2000, CREG has defined regulated users as those with power demands under 0.1 MW and monthly consumption below 55 MWh.[10] Non-regulated users are allowed to negotiate sale prices with retailing companies.

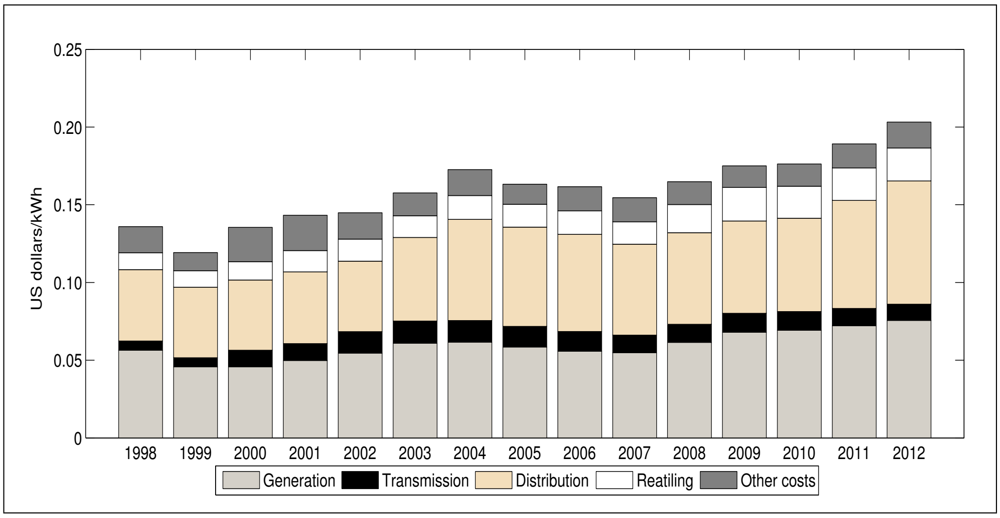
4.1.3 Heterogeneity in the electricity sector

Accounting for both observed and unobserved heterogeneity in stochastic frontier models is still a concern since efficiency estimations are sensitive to the modeling of sources of heterogeneity. In the case of observed heterogeneity, previous applications to the electricity distribution sector have studied the effects of including different types of covariates in the frontier, in the inefficiency or both. Hattori (2002) found out that heterogeneity sources related to the load factor, customer density and consumption density affect both, the shape of the frontier and the level of technical efficiency. Goto and Tsutsui (2008) found only customer density to have impacts on the technical efficiency of U.S. electricity distribution firms in a model that also includes consumption density, time and a deregulation index in the inefficiency distribution. In a recent study, Growitsch et al. (2012) considered weather factors and found them to be influential on costs but having limited effects in the efficiency estimations.

However, Growitsch et al. (2012) achieved more sensitivity in the efficiency estimations when unobserved heterogeneity is included by using a True Random Effects (TRE)

[10]CREG Resolution 131 of 1998.

Figure 4.4: Evolution of tariff per kWh in Colombia in real terms of 2012



model as proposed by Greene (2005). Other recent studies in electricity distribution have also been found to be relevant to considering this latent source of heterogeneity in SFA models. Kopsakangas-Savolainen and Svento (2011) perform a good analysis of the effect of observed and unobserved heterogeneity and warn of the high changes produced in rankings of cost efficiency under different models.

In the context of dynamic inefficiency models, Emvalomatis et al. (2011) studied the effect of including technological unobserved heterogeneity in an application to power generation plants in the US. Their findings reveal high persistence of inefficiency over time but also biases in the efficiency estimations when unobserved factors are not considered. However, it is also possible to think of heterogeneity regarding the persistence parameters. This would be related to possible differences in the adjustment costs among firms. The only studies considering this issue have been applications to the banking sector, where this type of heterogeneity has been found to be relevant (see Galán et al., 2014c; Huang and Chen, 2009).[11]

[11] Huang and Chen (2009) include firm specific persistence parameters in the context of models with forward-looking rational expectations while Galán et al. (2014c) include them in relation to the theory of adjustment costs.

4.1.4 Stochastic input distance function

Given that electricity distributors do not have control over electricity consumption and the number of users, which are their natural outputs, it is only possible to use input-oriented models for measuring technical efficiency. In this context, we assume that distribution firms use an $N \times 1$ vector of inputs $\mathbf{x} = (x_1, x_2, \ldots, x_N)'$ to provide an $M \times 1$ vector of outputs $\mathbf{q} = (q_1, q_2, \ldots, q_M)'$. Thus, we define an input set as follows:

$$L_g(q) = \mathbf{x} : \mathbf{x} \, \text{and technology} \, g \, \text{can produce} \, \mathbf{q}, \tag{4.4}$$

where the technology g satisfies the axioms of closeness, boundedness, strong disposability and convexity as described by Färe and Primont (1995). This technology can be represented by an input distance function, which is defined as:

$$D_I(\mathbf{x}, \mathbf{q}, g) = \sup_{\lambda} \{\lambda : \mathbf{x}/\lambda \in L_g(\mathbf{q}) \geq 1\}, \tag{4.5}$$

where λ denotes the maximum amount by which an input vector can be radially contracted while the output vector remains constant. We assume that every distribution firm employs the best available technology in each period. Thus, the Debreu-Farrell input-oriented measure of technical efficiency (TE) for firm i in period t is:

$$TE(x_{it}, q_{it}, t) \equiv 1/D_I(x_{it}, q_{it}, t). \tag{4.6}$$

The input distance function has the following features: it is homogeneous of degree one, a non-decreasing concave function of inputs, and a non-increasing quasi-concave function of outputs (see Färe and Primont, 1995). Linear homogeneity implies that it is possible to normalize all the inputs in the distance function by an arbitrarily chosen input $x_{N_{it}}$:

$$1/x_{N_{it}} = D_I(x_{it}/x_{N_{it}}, q_{it}, t) \exp(-u_{it}), \tag{4.7}$$

where $u_{it} \equiv \ln D_I(x_{it}/x_{N_{it}}, q_{it}, t) \geq 0$. Then, a firm is technically efficient if and only if $u_{it} = 0$ or similarly, $TE(x_{it}, q_{it}, t) = 1$.

Regarding the technology representation, we use a translog functional form to parameterize the distance function. So, we define $v_{it} \equiv \ln D_I(x_{it}/x_{N_{it}}, q_{it}, t) - TL(x_{it}/x_{N_{it}}, q_{it}, t)$, where $TL(.)$ is the translog function. In this case, (4.7) becomes:

$$y_{it} = TL(x_{it}/x_{N_{it}}, q_{it}, t) + v_{it} - u_{it}, \tag{4.8}$$

where $y_{it} \equiv -\ln x_{N_{it}}$. If any outputs or normalized inputs are stochastic then v_{it} is stochastic and (4.8) becomes a standard translog stochastic frontier model. For estimation purposes, the random noise term v_{it} is assumed to follow a normal distribution and the inefficiency component u_{it} is assumed to follow a nonnegative distribution. Using the results for individual inefficiencies, TE in each period is calculated as:

$$TE_{it} = \exp(-u_{it}). \tag{4.9}$$

Changes in productivity may also be computed from a stochastic distance function (see Balk, 2001; Orea, 2002, for a parametric approach to the computation of the Malmquist productivity index). In this context an input-oriented Malmquist productivity index can be computed and decomposed into technical efficiency change (TEC), technical change (TC), scale efficiency change (SEC) and an input mixed effect (IME) as follows:

$$MPI_I^t = TEC_I \cdot TC_I^{t,t+1} \cdot SEC_I^t \cdot ME_I^t, \tag{4.10}$$

where I denotes the input orientation, and the four components can be defined using a parametric translog specification as:[12]

$$TEC_I = \frac{D_I(x^t, q^t, t)}{D_I(x^{t+1}, q^{t+1}, t+1)} \tag{4.11}$$

$$TC_I^{t,t+1} = \frac{D_I(x^{t+1}, q^{t+1}, t+1)}{D_I(x^{t+1}, q^{t+1}, t)} \tag{4.12}$$

$$SEC_I^t = \frac{\breve{D}_I(x^t, q^t, t)}{\breve{D}_I(x^t, q^{t+1}, t)} \cdot \frac{D_I(x^t, q^{t+1}, t)}{D_I(x^{t+1}, q^t, t)} \tag{4.13}$$

$$ME_I^t = \frac{\breve{D}_I(x^t, q^{t+1}, t)}{\breve{D}_I(x^{t+1}, q^{t+1}, t)} \cdot \frac{D_I(x^{t+1}, q^{t+1}, t)}{D_I(x^t, q^{t+1}, t)}, \tag{4.14}$$

[12] Lovell (2003) defines the combined effect of SEC_I^t and ME_I^t as the volume effect.

where $\breve{D}_I$ stands for an input distance function associated with constant returns to scale (see Pantzios et al., 2011, for the derivation using the translog input oriented model).

4.1.5 Data and empirical model

Information on expenses, consumption, users, network length and quality indicators was collected for a sample of 21 electricity distribution firms during the period 1998 - 2012. The main data sources are CREG, SSPD and annual reports of the companies. Firms in the sample distributed 81% of the total consumed KwH in Colombia during the period and share 98% of total customers in the country. The data set is an unbalanced panel with a total of 246 observations. Table 4.4 presents a summary of statistics of the main variables. Monetary values are expressed in thousands of US dollars in real terms of 2012 after deflating by the consumer price index.

Table 4.1: Summary statistics

Variable	Mean	SD	Minimum	Maximum
Residential consumption (MWh)	785,665	1,118,006	13,499	4,687,938
Non-residential consumption (MWh)	729,120	1,138,132	9,069	5,637,621
Residential customers (#)	405,457	491,828	34,365	2,247,024
Non-residential customers (#)	40,672	57,430	2,935	294,734
Network length (Km)	16,587	15,673	232	70,795
Customer hours lost ($hours$)	89.12	101.94	6.20	580.89
Energy losses (%)	16.25	7.45	4.02	38.57
Consumption density ($kWh/user$)	2,836	1,120	436	6,642
Customer density ($users/Km$)	43.41	45.42	9.85	194.42
Total Expenses (thousands USD)	239,034	363,063	1,395	1,768,163

From these variables two outputs and three inputs are selected for the specification of the input distance function. Consumption and number of customers are the standard outputs in electricity distribution; however, they are usually highly correlated (0.95 in our sample) and one of them should be chosen to avoid collinearity problems. In our case, we select the number of users divided into residential (y_1) and non residential users (y_2). Inputs are total expenses (x_1), energy losses (x_2) and customer hours lost (x_3). Total expenses is the sum of operational and capital expenses. The former include administrative, operative and maintenance expenditures and the latter corresponds to the value of new investments in network cables, lines, ducts, tunnels and other machinery, plant and equipment. Considering overall total expenses is desirable for benchmarking electricity utilities (see Giannakis et al., 2005). Moreover, since we also account for quality measures, including total expenses recognizes that distribution firms adopt different strategies mixing capital and operating investment inputs in order to improve quality of service (see Jamasb et al., 2012). We also include energy losses and the length of interruptions as inputs where reductions are desirable. This approach has been used before in applications to the electricity sector using SFA models with distance functions (see Growitsch et al., 2009; Tovar et al., 2011; von Hirschhausen et al., 2006). Giannakis et al. (2005) and Yu et al. (2009) have also found these variables to be relevant in performing electric utilities benchmarking analysis explicitly including quality of service. Energy losses is the percentage of energy lost due to technical reasons and customer hours lost is the duration of service interruptions measured in hours per customer. We also include the network length measured in kilometers (km) as a characteristic of the output which is not directly under the control of firms.

Finally, we consider two inefficiency heterogeneity variables. These are consumption density (z_1) and customer density (z_2). Consumption density is measured as the number of KwH consumed per customer and customer density is measured as the number of users per kilometer. Both variables are expected to affect the inefficiency negatively in the sense that firms serving areas with low customer and consumption density may face a higher input-output relationship and more managerial difficulties in providing optimal service quality and resources allocation. Previous studies have also modeled these variables in the inefficiency distribution. Hattori (2002) and Goto and Tsutsui (2008) found these density characteristics to be relevant technical inefficiency drivers in the US and to produce changes in the results when they are omitted from the inefficiency distribution. Growitsch et al. (2009) found similar effects for eight European countries when including customer density in the mean of a truncated normal distributed inefficiency. In the case of Colombia, Melo and Espinosa (2005) have tested the inclusion of both density variables in the frontier and the inefficiency and have concluded about relevant effects of these variables as inefficiency drivers.

We use a translog representation of the technology for the input distance function derived in (4.8). The estimated model with the dynamic specification presented in (4.1) till (4.3) is the following:

$$\begin{aligned}
-\ln x_{1_{it}} = \; & \alpha_i + \textstyle\sum_{m=1}^{2} \beta_m \ln y_{m_{it}} + \beta_{m+1} \ln km_{it} + \sum_{r=1}^{2} \delta_r \ln \left(\frac{x_{r_{it}}}{x_{1_{it}}}\right) \\
& + \tfrac{1}{2} \textstyle\sum_{m=1}^{2} \sum_{n=1}^{2} \beta_{mn} \ln y_{m_{it}} \ln y_{n_{it}} + \tfrac{1}{2} \sum_{r=1}^{2} \sum_{s=1}^{2} \delta_{rs} \ln \left(\frac{x_{r_{it}}}{x_{1_{it}}}\right) \ln \left(\frac{x_{s_{it}}}{x_{1_{it}}}\right) \\
& + \textstyle\sum_{m=1}^{2} \sum_{r=1}^{2} \eta_{mr} \ln y_{m_{it}} \ln \left(\frac{x_{r_{it}}}{x_{1_{it}}}\right) + \kappa_1 t + \tfrac{1}{2} \kappa_2 t^2 + \sum_{m=1}^{2} \phi_m t \ln y_{m_{it}} \\
& + \textstyle\sum_{r=1}^{2} \varphi_r t \ln \left(\frac{x_{r_{it}}}{x_{1_{it}}}\right) - u_{it} + v_{it} \\
& \log u_{it} = \omega + \textstyle\sum_{p=1}^{2} \gamma_p z_{p_{it}} + \rho_i \log u_{i,t-1} + \xi_{it};\ \xi_{it} \sim N(0, \sigma_\xi^2);\ t = 2...T \\
& \log u_{i1} = \frac{\omega + \sum_{p=1}^{2} \gamma_p z_{p_{i1}}}{1-\rho_i} + \xi_{i1};\ \xi_{i1} \sim N\left(0, \frac{\sigma_\xi^2}{1-\rho_i^2}\right);\ t = 1.
\end{aligned} \tag{4.15}$$

Total expenses are used as a numeraire to accomplish linear homogeneity in inputs and cross-effects symmetry is imposed by requiring $\beta_{mn} = \beta_{nm}$ and $\delta_{rs} = \delta_{sr}$.

4.1.6 Estimation Results

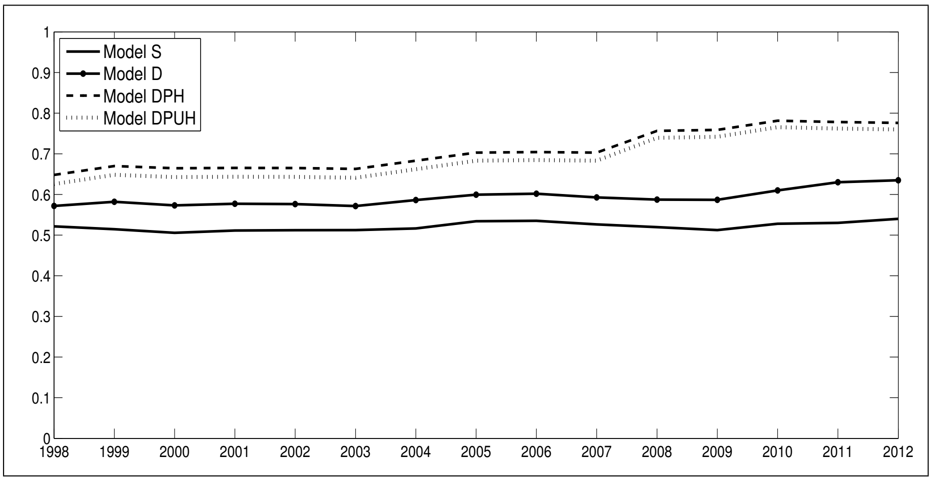
We estimate four different models derived from (4.15). The first three models do not account for unobserved technological heterogeneity, that is, $\alpha_i = 0$. In addition, model (S) restricts $\rho_i = 0$, so the model becomes static and the inefficiency term follows a log-normal distribution with observed heterogeneity in its location parameter. The second model (D) restricts $\rho_i = \rho$, which implies a dynamic model with fixed persistence parameter. The third model (DPH) allows heterogeneous persistence through ρ_i. Finally, the fourth model (DPUH) is the complete model in (4.15), which is dynamic and allows for heterogeneous persistence and unobserved heterogeneity. Results of the estimations are presented in Table 4.5.

Table 4.2: Posterior mean and standard deviation of parameter distributions

Parameters	**Model S**		**Model D**		**Model DPH**		**Model DPUH**	
	$\alpha_i = \alpha, \rho_i = 0$		$\alpha_i = \alpha, \rho_i = \rho$		$\alpha_i = \alpha, \rho_i \neq \rho$		$\alpha_i \neq \alpha, \rho_i \neq \rho$	
Parameter	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
ID function								
α	-13.4149	1.2091	-12.6924	0.7935	-11.4653	0.6624	-11.4045	0.5543
$\beta_1(\ln y_1)$	-0.1902	0.1215	-0.0379	0.0257	-0.0346	0.0219	-0.1082	0.0266
$\beta_2(\ln y_2)$	-0.0968	0.0991	-0.1200	0.0806	-0.0712	0.0530	-0.0463	0.0248
$\beta_3(\ln x_2)$	0.0115	0.0087	0.0244	0.0135	0.0060	0.0050	0.0149	0.0134
$\beta_4(\ln x_3)$	0.0116	0.0088	0.0485	0.0168	0.0232	0.0197	0.0075	0.0056
$\beta_5(\ln km)$	-0.3494	0.0739	-0.3265	0.1074	-0.1265	0.0491	-0.1413	0.0625
$\beta_6(t)$	-0.1724	0.1217	-0.0932	0.1336	-0.0616	0.0808	-0.0730	0.0684
$\beta_7(t^2)$	0.0032	0.0010	0.0046	0.0012	0.0049	0.0006	0.0050	0.0005
$\phi_1(1/2 \ln y_1^2)$	-1.0098	0.3705	-1.3391	0.5202	1.6021	0.7925	1.5440	0.6968
$\phi_2(\ln y_1 \ln y_2)$	0.4733	0.3262	0.8353	0.5289	-1.4377	0.6969	-1.3677	0.6227
$\phi_3(1/2 \ln y_2^2)$	0.1132	0.3291	-0.2584	0.5504	1.2588	0.6821	1.2503	0.6303
$\phi_4(1/2 \ln x_2^2)$	0.0868	0.0463	0.0470	0.0450	0.0105	0.0362	0.0005	0.0346
$\phi_5(\ln x_2 \ln x_3)$	-0.0951	0.0224	-0.0652	0.0321	-0.0160	0.0147	-0.0037	0.0147
$\phi_6(1/2 \ln x_3^2)$	0.0302	0.0174	0.0209	0.0194	0.0164	0.0112	0.0138	0.0124
$\delta_1(\ln y_1 \ln x_2)$	-0.2636	0.1341	-0.2488	0.1303	0.2395	0.1451	0.1911	0.1275
$\delta_2(\ln y_2 \ln x_2)$	0.4149	0.0977	0.3551	0.1001	-0.2212	0.1136	-0.1622	0.0967
$\delta_3(\ln y_1 \ln x_3)$	0.0175	0.0822	-0.0168	0.0767	-0.0375	0.0563	0.0140	0.0554
$\delta_4(\ln y_2 \ln x_3)$	-0.2235	0.0728	-0.1163	0.0623	0.0371	0.0542	0.0035	0.0525
$\kappa_1(t \ln y_1)$	0.0252	0.0211	0.0353	0.0238	0.0192	0.0157	0.0175	0.0141
$\kappa_2(t \ln y_2)$	-0.0238	0.0196	-0.0233	0.0211	-0.0142	0.0138	-0.0150	0.0126
$\kappa_3(t \ln x_2)$	-0.0063	0.0075	0.0032	0.0074	0.0020	0.0047	0.0004	0.0041
$\kappa_4(t \ln x_3)$	0.0064	0.0040	0.0045	0.0040	0.0022	0.0025	0.0025	0.0022
Inefficiency								
ω	-1.4049	0.8467	0.0205	0.0050	0.0017	0.0002	0.0028	0.0002
ρ			0.8366	0.0846	0.6532	0.0850	0.6507	0.0868
$\gamma_1(\ln z_1)$	-0.3443	0.1008	-0.0424	0.0081	-0.0317	0.0024	-0.0314	0.0168
$\gamma_2(\ln z_2)$	-0.4407	0.0838	-0.1277	0.0394	-0.1258	0.0553	-0.1009	0.0452
σ_v	0.1653	0.0315	0.1314	0.0194	0.0943	0.0017	0.0977	0.0018
σ_ϵ	0.1610	0.0517	0.0613	0.0023	0.0406	0.0038	0.0347	0.0029
Mean eff.	0.5173		0.5841		0.6478		0.6373	
SD eff.	0.1205		0.1551		0.2600		0.2420	
DIC_3	-119.12		-253.28		-339.49		-349.86	
LPS	35.79		21.06		9.74		6.53	

We observe that the more flexible is the model in terms of accounting for dynamic effects and heterogeneity, the better the values obtained for DIC_3 and LPS. Lower values for these criteria suggest better fit and predictive performance. Moreover, high inefficiency persistence is estimated by the dynamic models suggesting the presence of important adjustment costs in the Colombian distribution sector. Model D estimates around 84% of the inefficiency being transmitted from one period to the next, which is very similar to the average firm specific persistence estimated under models DPH and DPUH.[13] It can be also seen that not only is the average technical efficiency in the whole sector higher in the more flexible models, but also its dispersion. This may suggest that introducing dynamic effects and unobserved heterogeneity sources distinguishes the presence of adjustment costs and heterogeneity from technical inefficiency and also differentiates firms depending on their specific characteristics. These effects can also be observed in Figure 4.5, where the evolution of efficiency over time under the four models is plotted. We can also observe that the dynamic models accounting for persistence heterogeneity (DPH and DPUH) identify larger improvements in TE during the period.

Figure 4.5: Evolution of posterior mean TE under different models

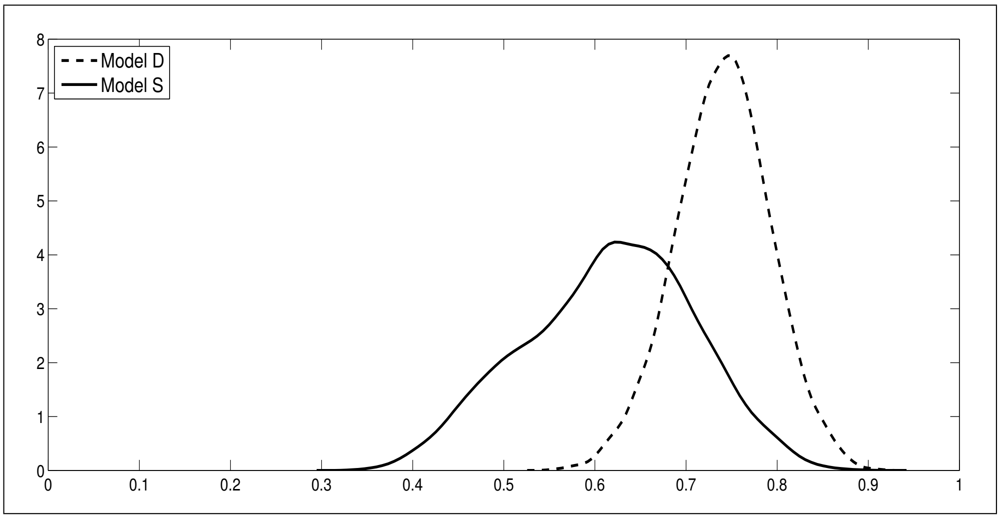


In order to understand better the effects of the different specifications on the efficiency estimates, we analyze the results at firm level and their evolution over time by comparing the models derived from (4.15) from the most to the least restrictive. In Figure 4.6, we compare the posterior efficiency distribution for a firm with median values for

[13]Recently, Poudineh et al. (2014) found also very high inefficiency persistence in an application of a dynamic model to Norwegian electricity utilities.

customer and consumption density in 2012 under static and dynamic formulations. We observe that introducing dynamic effects alter not only the location of the distribution, by estimating higher values for technical efficiency, but also that the dispersion is lower, which allows more certainty on the individual efficiency estimations.

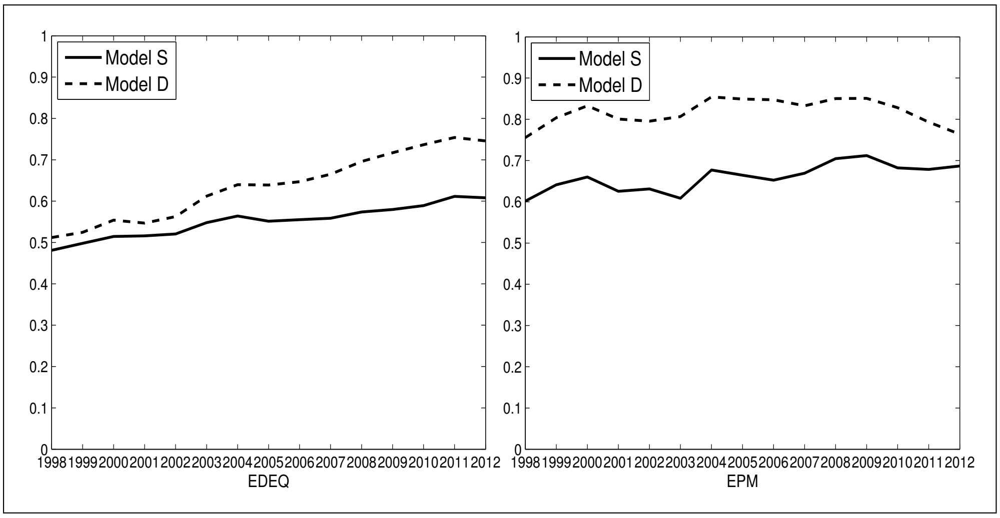
Figure 4.6: Posterior efficiency distribution for a representative firm in 2012



These differences in the posterior distributions also affect the estimation of the evolution of technical efficiency over time. Figure 4.7 presents the posterior mean efficiency estimations during the period for two firms, *Electrificadora del Quindio* (EDEQ) and *Empresas Públicas de Medellín* (EPM). We observe that for EDEQ, the dynamic specification estimates gains in technical efficiency that are not identified under the static model. This may suggest that the improvements made by this firm during the period are more important in relative terms given the presence of high adjustment costs in the sector. In the case of EPM, results imply that, given the adjustment costs faced by all firms in the sector, this firm did not improve enough to identify efficiency gains.These findings are important from the point of view of the regulator because they suggest that firms could not explain poor performance on the basis of modelled adjustment costs.

The dynamic model analyzed assumes that all distribution firms face the same adjustment costs in terms of being able to adjust the same proportion of inefficiency from one period to the next. However, firms with different characteristics may present different adjustment costs, so Model DPH allows for firm specific persistence parameters. Figure 4.16 in the appendix exhibits the 95% probability intervals for the persistence

Figure 4.7: Evolution of posterior mean efficiencies for EDEQ and EPM

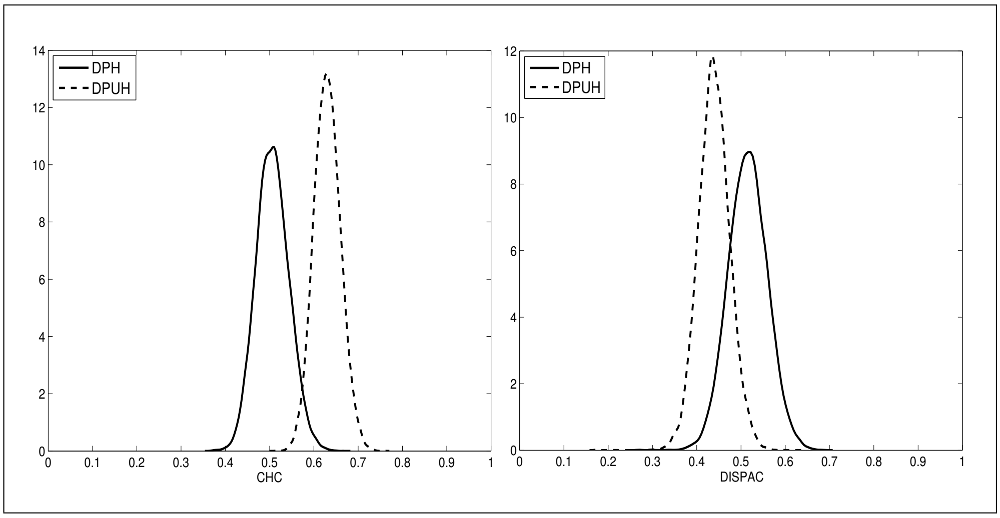


estimations of every firm. Important differences in the individual posterior estimations of persistence are found, ranging from 0.31 to 0.99. This suggests large heterogeneity in the adjustment costs of electricity distributors that could be related to certain characteristics of these firms and the incentive regulation that they have faced, as is discussed further below. These findings illustrate the importance of accounting for firm specific persistence parameters, which have implications for the efficiency estimations and their evolution over time as is observed in Figure 4.5.

Finally, the full model in (4.15) is estimated accounting not only for heterogeneous persistence but also for unobserved technological heterogeneity. Although the evolution of efficiency is similar to that estimated under Model DPH (see Figure 4.5), Model DPUH identifies unobserved firm effects that distinguish them in terms of the estimated efficiency. Figure 4.8 compares the posterior efficiency distributions for a low and a high efficient firm under models DPH and DPUH.[14] We observe that their posterior distributions move and shrink, implying a reduction in the uncertainty of the individual estimations. It is also important to notice that estimations of firm specific persistence parameters do not present important changes compared to those obtained in Model DPH.

[14]The selected firms are *Central Hidroeléctrica de Caldas* (CHC) and *Empresa Distribuidora del Pacífico* (DISPAC).

Figure 4.8: Posterior efficiency distributions for CHC and DISPAC



Focusing on our preferred model (DPUH), we can identify some links between differences in adjustment costs and firm characteristics. We plot in Figure 4.9 the average posterior distributions of the persistence parameter by groups of firms. In general, we observe that firms with a higher proportion of rural and small customers present lower adjustment costs than those which are mainly urban and serve larger customers. In contrast, by type of ownership and number of customers, no major differences can be observed between firms in terms of inefficiency persistence. This would imply that rural firms and those with small customers have been able to adapt more easily towards optimal performance.

Differences between groups of Colombian utilities are also observed in terms of efficiency. Figure 4.10 exhibits the average posterior technical efficiency during the period by groups of firms. We observe that mainly urban distributors and firms serving high consumption customers have been more efficient during the period than their counterparts. Although differences are smaller, this is also the case of private and large distributors.

Figure 4.9: Average posterior distribution of ρ_i by groups of firms

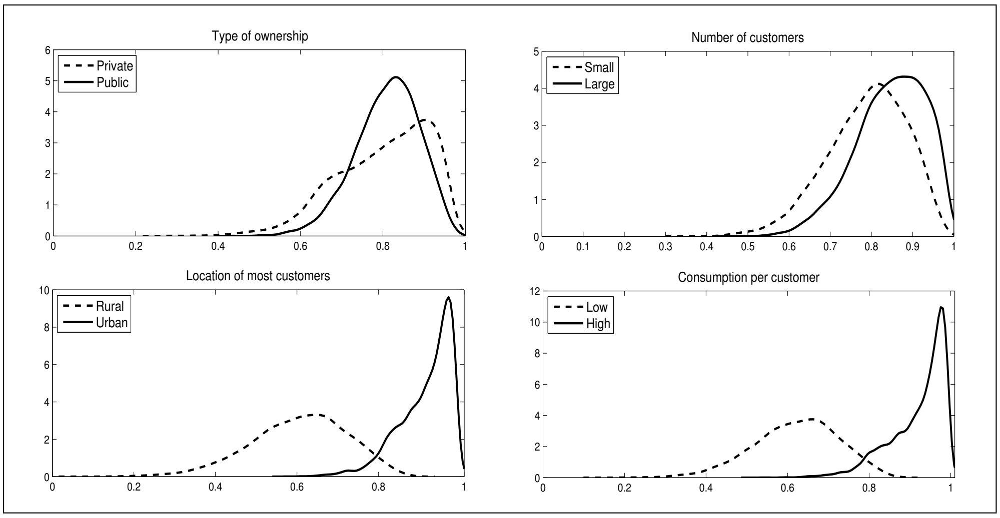
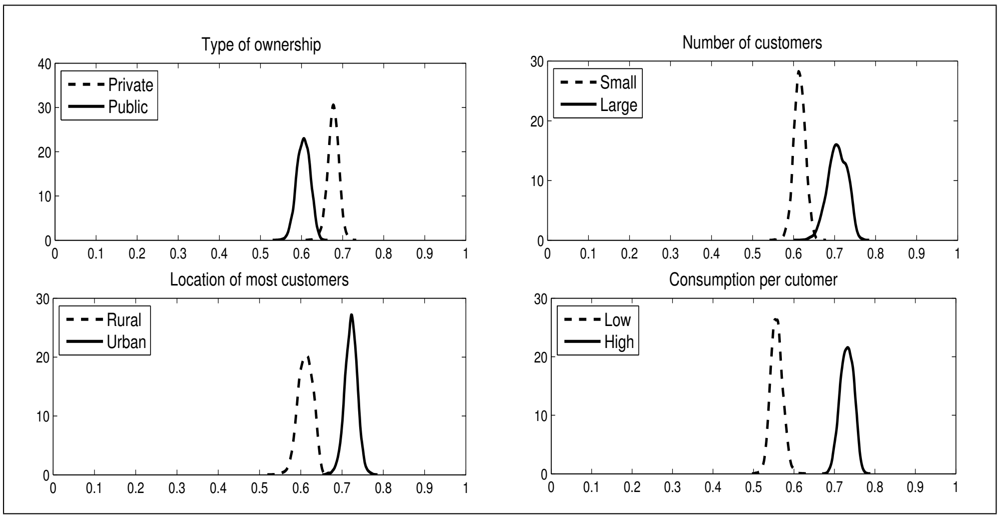


Figure 4.10: Average posterior distribution of efficiency by groups of firms



However, what it is more interesting in our dynamic analysis is the change that these firms have exhibited from 1998 to 2012 in terms of efficiency and productivity. We compute the MPI and its decomposition as described in (4.10) till (4.14) and present the results by group of firms in Table 4.3. We observe that all types of firms except those urban and serving high consumption customers have increased their productivity during the period. This improvement has been driven in all the cases by increases in technical efficiency, which has compensated the technological regress suggested by the results for technical change for all firms. Regarding scale efficiency change, only private, rural and firms with low consumption customers exhibit some improvement in terms of operating at the efficient scale. Finally, the input-mix effect exhibit values very close to 1 for all groups which suggest that changes in the input mix during the period have kept scale efficiency almost unaltered.

Table 4.3: Decomposition of the Malmquist Productivity Index by groups of firms

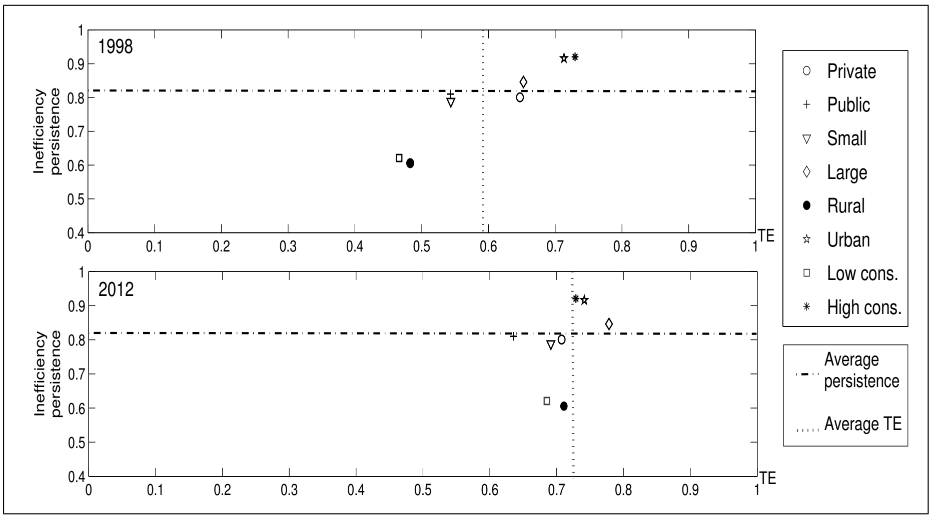
Firms	**TEC**	**TC**	**SEC**	**ME**	**MPI**
Private	1.0930	0.9399	1.0167	0.9956	1.0445
Public	1.1694	0.9417	0.9736	1.0023	1.0722
Small	1.2714	0.9419	0.9753	1.0024	1.1680
Large	1.1932	0.9386	0.9572	0.9996	1.0720
Rural	1.4723	0.9384	1.0343	1.0007	1.4290
Urban	1.0396	0.9407	0.9694	1.0058	0.9480
Low cons	1.4701	0.9440	1.0102	0.9987	1.4019
High cons	0.9986	0.9387	0.9666	1.0073	0.9061
Total	1.2161	0.9388	1.0332	0.9972	1.1795

In general, rural firms and those serving low consumption customers exhibit very important increases in productivity during the post-reform period explained by improvements in their scale efficiency but mainly due to large increases in technical efficiency. These firms are also those exhibiting lower inefficiency persistence and therefore those with higher scope for improvement.

This relationship between inefficiency persistence and changes in technical efficiency is presented in Figure 4.11, where the average posterior inefficiency persistence is plotted against their average posterior TE in 1998 and 2012 for every group of firms. We observe that firms with high inefficiency persistence has barely presented changes in TE. This is in particular noticeable for urban distributors and firms with high consumption customers. On the other hand, rural companies and firms with more small customers seem to catch up with their counterparts in terms of efficiency during the period.[15] This suggests that incentives introduced by the regulator during the period, mainly in terms of service quality, have helped distributors with low consumption customers and in rural areas to improve their efficiency but that they have not been effective for their counterparts, which in absence of new incentives may become stuck in terms of technical efficiency.

[15] In Table 4.7 of the Appendix we present these results for each firm along with other firm characteristics.

Figure 4.11: Inefficiency persistence and technical efficiency by groups of firms (1998 and 2012)



4.1.7 Empirical implications

The electricity reform in Colombia introduced the separation of activities in the electricity sector and set the conditions for privatization and competition. In general, the reform has had positive effects on the ability of the sector to overcome extreme weather conditions and meet demand requirements. However, for distribution companies, competition and privatization have been slow processes and the users did not benefit from improvements in service quality for the first ten years after the reform. In fact, previous studies measuring consequences of the reform on efficiency have not found evidence of improvements, although large differences in efficiency have been found among firms.

This may indicate the presence of high adjustment costs in the sector in Colombia and important heterogeneity factors among distributors. We include these conditions in a stochastic frontier model that accounts for dynamic effects and unobserved heterogeneity. Our findings suggest high inefficiency persistence in the sector that could be related to adjustment costs and inadequate incentive regulation. However, important differences are found among firms. In particular, firms operating mainly in rural markets and serving small customers present lower adjustment costs than firms with the opposite characteristics. This condition has allowed these firms to catch up urban firms and firms serving large users, which should draw the attention of the regulator because they seem to be stuck in terms of technical efficiency. In fact, customer density and consumption density are found to be important inefficiency drivers in the sector and unobserved heterogeneity sources to be relevant in distinguishing heterogeneity from inefficiency and identifying differences among firms.

Our findings may be helpful for the Colombian regulator and the Ministry of Mines and Energy, which have been recently working on the composition of groups of distribution firms that would share the same prices.[16] These groups have been formed following geographical criteria. However, our results suggest that the design of these groups should mainly consider the inefficiency persistence level of each firm and their characteristics in terms of customer density and consumption density.

Overall, efficiency in the Colombian distribution sector has been found to exhibit improvements. However, efficiency gains can only be clearly identified in the last five years. This period coincides with the main reductions in the length of interruptions and energy losses, and the highest rates of increase in the number of customers. Although very preliminary, these results may favour the recent incentive regulation policies for improving quality of service and reducing energy losses. Nevertheless, the last five years have also been characterized by important increases in the electricity tariff for regulated users. Not only the tariff per kWh has presented important increases during the period, but also the proportion derived from distribution costs has increased relative to the other tariff components. This implies that Colombian users are now receiving a better service but that they are paying the costs of these improvements via higher tariffs.[17] These results suggest that incorporating willingness to pay into the efficiency analysis of the Colombian distribution sector would be of interest for future research.

4.2 Separating Heterogeneity from Inefficiency Dynamics

The inclusion of variables that capture firm characteristics in the inefficiency component is important to distinguish properly heterogeneity from inefficiency but they do not necessarily capture inefficiency adjustment processes or have persistent effects. This can be particularly important in a dynamic framework, since including covariates as inefficiency drivers in an autoregressive specification implies that they have persistent effects over

[16]CREG resolution 058 of 2008 and Ministry of Mines and Energy resolutions: 182306 of 2009, 181347 of 2010, 180686 of 2011 and 180574 of 2012.

[17]In fact, for the case of UK, Yu et al. (2009) have found the social cost of outages to be considerably higher than the utilities' incentives.

time. In this context, it would mean that the effects that these firm characteristics and environmental conditions produce on inefficiency can not be easily adjusted by firms. This would be the case of a public bank that is less efficient because of attending rural customers in remote places. This characteristic may not be easily altered and may consequently induce high adjustment costs. However, changing other conditions such as some managerial practices or the risk exposure of short- run investment portfolios may be easier to adjust. In these cases, heterogeneity sources should be allowed to be inefficiency drivers but modeled out of the dynamic specification.

We propose a specification for the inefficiency in the context of dynamic stochastic frontier models that is flexible in terms of the treatment of heterogeneity. The first characteristic is the separation of two components within a log-linear specification for the inefficiency. One component, which is unobserved, will follow an autoregressive process that captures the portion of inefficiency that is transmitted from one period to the next. The second component is a vector of observed covariates driving the inefficiency level. In this sense this could be seen as an extension of the model by Galán et al. (2014b) where an unobserved random parameter is modeled in the inefficiency distribution along with observed covariates, but where the unobserved part follows a first order autoregressive process. This component can also include observed variables that will capture persistent effects of heterogeneity in the inefficiency as in the specification followed by Tsionas (2006) for the whole inefficiency. The second characteristic is that we account for heterogeneity in the adjustment costs by allowing the persistence parameter to be firm-specific as in the model in the previous section (see Galán and Pollitt, 2014a).

The model is given by:

$$
\begin{aligned}
y_{it} &= \mathbf{x}_{it}\boldsymbol{\beta} + v_{it} - u_{it}, & v_{it} &\sim N(0,\sigma_v^2) && (4.16)\\
\log u_{it} &= \theta_{it} + \mathbf{z}_{it}\boldsymbol{\gamma} + \xi_{it}, & \xi_{it} &\sim N(0,\sigma_\xi^2) && (4.17)\\
\theta_{it} &= \omega + \mathbf{h}_{it}\boldsymbol{\psi} + \rho_i\,\theta_{i,t-1} + \eta_{it}, & \eta_{it} &\sim N(0,\sigma_\eta^2), \quad t = 2...T && (4.18)\\
\theta_{i1} &= \frac{\omega + \mathbf{h}_{i1}\boldsymbol{\psi}}{1-\rho_i} + \eta_{i1}, & \eta_{i1} &\sim N\left(0, \frac{\sigma_\eta^2}{1-\rho_i^2}\right), \ t = 1. && (4.19)
\end{aligned}
$$

The stochastic frontier is represented by (4.16) where y_{it} represents the output for firm i at time t, $\mathbf{x}_{it}$ is a row vector that contains the input quantities, $\boldsymbol{\beta}$ is a vector of parameters, v_{it} is an idiosyncratic error assumed to follow a normal distribution, and u_{it} is the inefficiency component. Equation (4.17) is the log-linear specification for the inefficiency where θ_{it} represents the dynamic component, $\mathbf{z}_{it}$ is a row vector of firm specific heterogeneity variables, $\boldsymbol{\gamma}$ is a vector of parameters and ξ_{it} is a white noise process with constant variance σ^2_ξ. The unobserved dynamic component θ_{it} follows an autoregressive process represented by (4.18) where ω is a constant, $\mathbf{h}_{it}$ is a row vector of observed covariates, $\boldsymbol{\psi}$ is a vector of parameters, ρ_i is the firm specific persistence parameter measuring the proportion of the dynamic part of the inefficiency that is transmitted from one period to the next for every firm, and η_{it} represents unobserved random shocks in the dynamic component and follows a normal distribution with variance σ^2_η.[18] The dynamic process is assumed to be stationary and (4.19) initializes it.

Stationarity ensures that the dynamics of the log-inefficiency do not diverge to negative or positive infinity. If this condition is not imposed, efficiency scores could be equal to one or zero in the long-run. The first case would contradict the adjustment costs theory that motivates the dynamic formulation and the second case would imply that totally inefficient firms do not exit the market. Therefore, the persistence parameters are required to satisfy $|\rho_i| < 1$. A value close to 1 for this parameter means high persistence of the inefficiency dynamic component and slow adjustment of firms towards optimal conditions.

Modeling firm specific persistence parameters imply that even when no covariates are included, firms may present differences in their adjustment of common factors and therefore different long-run inefficiencies. However, as we present in Section 4.2.1 we model these parameters using a hierarchical structure common in Bayesian statistics that links them to a common parameter for the whole sector. Moreover, firms in the sector share also a common long-run dynamic component ω and common elasticities for the covariates given by the vectors $\boldsymbol{\psi}$ and $\boldsymbol{\gamma}$.

The proposed dynamic specification given by equations (4.17) till (4.19) encompasses other models in the literature and permits us to compare some assumptions by including restrictions. For instance, homogeneous costs of adjustments for all banks can be studied by imposing $\rho_i = \rho$. If $\rho = 0$ the model is reduced to a static formulation with no

[18] It can be noticed that if unobserved random shocks are only allowed to affect the inefficiency via the scale parameter of its distribution, then $\eta_{it} = 0$ and the inefficiency can be modeled following other nonnegative distributions with firm specific and time-varying mean. However, other distributions can change the interpretation of the inefficiency parameters.

adjustment costs but where an unobserved component θ_{it} captures latent inefficiency heterogeneity as in Galán et al. (2014b). Additionally, if η_{it} is also equal to 0, the model takes the form of the Battese and Coelli (1995) static formulation. Finally, if $\rho_i = \rho$, $\xi_{it} = 0$ and $\boldsymbol{\gamma} = \mathbf{0}$ the model reduces to the dynamic model in Tsionas (2006).

4.2.1 Bayesian inference

We assume proper but relatively disperse prior distributions throughout. In particular, the distributions assumed for the parameters in the distance function are: $\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}$ is a diagonal matrix with precision priors set to 0.001 for all coefficients. The variance of the idiosyncratic error term is inverse Gamma distributed, that is equivalent to $\sigma_v^{-2} \sim G(a, b)$ where the priors for shape and rate parameters are set to 0.01.

The specification in (4.17) implies that the inefficiency follows a log-normal distribution. Then $u_{it}|\theta_{it}, \mathbf{z}_{it}, \boldsymbol{\gamma}, \sigma_\xi^2 \sim LN(\theta_{it} + \mathbf{z}_{it}\boldsymbol{\gamma}, \sigma_\xi^2)$, where the location component is composed of the unobserved dynamic parameter and the observed heterogeneity component. The distribution for the unobserved parameter modeling the dynamics is $\theta_{it}|\theta_{i,t-1}, \omega, \mathbf{h}_{it}, \boldsymbol{\psi}, \rho_i, \sigma_\eta^2 \sim N(\omega + \mathbf{h}_{it}\boldsymbol{\psi} + \rho_i\, \theta_{i,t-1}, \sigma_\eta^2)$ for $t = 2...T$. Given stationarity we have $\theta_{i1}|\omega, \mathbf{h}_{i1}, \boldsymbol{\psi}, \rho_i, \sigma_\eta^2 \sim N\left(\frac{\omega+\mathbf{h}_{i1}\boldsymbol{\psi}}{1-\rho_i}, \frac{\sigma_\eta^2}{1-\rho_i^2}\right)$. The distribution for the common constant ω is normal with priors set to -1.5 and 1 for the mean and precision.[19] The distribution of the parameter vector of the observed covariates in the dynamic component is: $\boldsymbol{\psi} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\psi}})$ where $\boldsymbol{\Sigma}_{\boldsymbol{\psi}}^{-1}$ is a diagonal matrix with precision priors set to 0.1. Finally, the distribution for the firm characteristic parameters in the inefficiency are: $\boldsymbol{\gamma} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$ where $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}$ is a diagonal with priors set to 0.1 for every precision coefficient.

Regarding the persistence parameters, we assume $|\rho_i| < 1$ to assure stationarity. Since the persistence parameters are allowed to vary across firms, we define a hierarchical structure where $\rho_i = 2k_i - 1$ and $k_i \sim \beta(k, 1-k)$ with $k \sim \beta(r, s)$ and priors for these parameters set to 0.5. In the case that the homogeneous persistence restriction $\rho_i = \rho$ is imposed, we assume $\rho = 2k - 1$ with k defined as previously.

The variances are assumed to follow inverse gamma distributions where $\sigma_\eta^{-2}, \sigma_\xi^{-2} \sim G(n, d)$ with priors set to $n = 10, d = 0.01$ and $n = 0.5, d = 0.005$, respectively.[20]

[19] These values center the efficiency prior distributions at 0.8 similar to other Bayesian empirical applications in banking.

[20] The first is the same prior used by Tsionas (2006) for the random shocks variance in the inefficiency equation and the second is that suggested by West and Harrison (1997) for the state equation of Bayesian

Sensitivity analysis was performed by allowing changes in the priors of the parameters in the inefficiency component. In particular, different priors for ω imply different priors on the efficiency but in our experiments, no important differences were obtained in the posterior distributions. For the persistence parameter ρ we studied the sensitivity to the use of a truncated normal distribution and posterior results were also found to be robust to the use of this alternative. Small changes in the values of n and d in the priors of σ_η^{-2} and σ_ξ^{-2} were also examined with no evidence of posterior sensitivity. Finally, we also checked the posterior distribution of the idiosyncratic errors $\mathbf{v}$ to check the normality and non autocorrelation assumptions. We found no evidence of non-normality or of autocorrelation in this case. Note however that in cases where the idiosyncratic errors do not appear to be normal, one possibility is to model using a heavy tailed distribution such as student-t (see Griffin and Steel, 2007, for the implementation of this assumption under a Bayesian framework). In the case of autocorrelation, it is also possible to think of an autoregressive structure for this component.

The implementation of the models were carried out using WinBUGS with an MCMC sample of 50.000 iterations discarding the first 20.000 and a thinning of 6 to remove autocorrelations. As model comparison criteria we use DIC_3 and LPS as earlier.

4.2.2 Application to the Colombian banking sector

We apply the new specification to a sample of Colombian banks during the last decade. Inefficiency persistence in the banking sector has been precisely identified by Tsionas (2006) in his application to US banks. His findings reveal very high inefficiency persistence, suggesting the presence of high adjustment costs in the banking sector. Previous studies have also found evidence of inefficiency persistence in financial institutions. Tortosa-Ausina (2002), in an analysis of transition probabilities of efficiency, found that most of Spanish banks remain in the same state of relative inefficiency in consecutive periods.

In particular, the Colombian banking sector is of interest since it has been characterized by the arrival of foreign institutions and several mergers and acquisition (M&A) processes that have increased the differences in terms of size among banks during this period.

dynamic linear models.

Certainly, the effects of foreign ownership, size and M&A on banks efficiency have been studied previously under static formulations. Regarding foreign ownership and size, divergent results have been obtained previously. On the one hand, using Bayesian stochastic frontiers, Tecles and Tabak (2010) and Assaf et al. (2013) found foreign and large banks to be more cost and profit efficient in Brazil and Turkey, respectively. Using nonparametric methods, Ray and Das (2010) and Sathye (2003) also found positive effects of foreign ownership on efficiency in Indian banks. On the other hand, negative effects of foreign ownership have been found by Lensik et al. (2008) in a study including a sample of 105 countries; while, a negative impact of size was found by Hartman and Storbeck (1996) for banks in Sweden, and no size effects were concluded for the case of Brazil by Staub et al. (2010). In the dynamic context, Tsionas (2006) identified size to have persistent effects on cost efficiency of US banks. Concerning M&A, previous studies have found none or very little improvement on input-oriented technical efficiency or cost efficiency (see Amel et al., 2004, for a review). However, merged banks have been found to present different time patterns than non-merged institutions and their efficiency to be highly dependent on time (see Cuesta and Orea, 2002). In this context, introducing time dependency into a dynamic structure and allowing merged banks to follow their own dynamics may lead to different conclusions.

In particular, we have two main aims with this application: firstly, to evaluate the impact of adding more flexibility to the persistence parameter and separating heterogeneity from the dynamics on efficiency estimations; secondly, to identify the effects of size, foreign ownership and M&A on input-oriented technical efficiency of Colombian banks.

4.2.2.1 The Colombian banking sector

The Colombian banking sector has experienced major changes in the last thirty years. It passed from a high regulated and low competitive system in the eighties to a more flexible and foreign capital open system in the nineties. From 1998 to 2002 the country suffered a deep financial crisis that lead to a rearrangement of the banking sector. This implied a reduction of the number of banks and a concentration of commercial and mortgage activities under the same institution. This reorganization process occurred during the period 2002 till 2009, which was characterized by an environment of economic recovery, high foreign capital flows and an increase of the services provided by banks. During these years, several M&A processes took place leading to a reduction of the number of financial institutions, which passed from more than forty mortgage and commercial banks in the mid 90's to less than twenty in 2009. Foreign capital banks had also played an important role in Colombia during the period of study and they accounted for almost 40% of the banking entities in 2009.

As mentioned in Chapter 3, previous efficiency studies of Colombian banking system have mainly evaluated costs and profit efficiency and have focused on the crisis and immediate post-crisis periods (see Janna, 2003, for a review on applications to the Colombian banking sector). All of these studies have shown similar results in terms of an increase in the efficiency of the sector during the mid-nineties, decreases in efficiency during the crisis period and a recovery on these indicators in the following years. The effect of bank mergers in Colombia has been studied by Estrada (2005) for the $1994-2004$ period who found gains in cost efficiency specially for relatively inefficient pre-merger banks. Clavijo et al. (2006) also studied M&A from 1990 to 2005 finding decreases in efficiency in the subsequent periods to the processes. However, most of these occurred during the crisis period.

4.2.2.2 Data and model specification

As the sample presented in Chapter 3, the data set used in this section is also from the local central bank and the supervisory agency and contains most of the banks in the application presented earlier. However, some of them are different as well as the period of study. For this application the data set contains information from thirty one commercial banks, which represents 87% of the total assets in the Colombian banking sector. This is an unbalanced panel data set of quarterly data from 2000 to 2009. During the period, nineteen M&A processes were carried out involving banks in the sample and here, we shall consider post-merged institutions as different banks. Regarding ownership, nine banks in the sample are foreign-owned and only one is public-owned.

As in the application in Chapter 3, we use the intermediation approach in Sealey and Lindley (1977). However, here we represent the technology in (4.16) with an input distance function. This allows us to consider input quantities while accounting for multiple products and avoiding using firm specific prices. The derivation of the stochastic input distance function is the same presented in Section 4.1.4.

We select three inputs and two outputs. Inputs are quantities of deposits (x_1), labor (x_2) and physical capital (x_3), including premises and other fixed assets. As outputs we consider the total loans (y_1), including consumer, industrial, commercial and real state loans; and the total investments and other securities (y_2). All monetary variables are expressed in thousand of millions of pesos and are in real terms of 2009 by deflating by the consumer price index. Regarding the inefficiency heterogeneity variables, they are included either inside the dynamic inefficiency component or out of it but the same variable is not simultaneously included in both parts. These variables are the log of total banks assets (z_1/h_1), its square (z_2/h_2) and foreign ownership (z_3/h_3). Public ownership is not considered since the sample contains only one bank with public capital. Table 4.4 reports summary statistics of these variables.

Table 4.4: Summary statistics

Variable	Mean	Std. Deviation	Minimum	Maximum
Deposits	3 886 117.8	4 661 207.2	146 005.1	29 600 000
Labor	2 984.5	2 676.6	79	20 780
Physical Capital	93 036.7	101 070.5	5 359.075	710 837.1
Loans	3 305 469.4	4 195 981.4	132 508.6	27 900 000
Investments	1 357 952.7	1 472 720.2	32 466.74	8 277 268
Assets	5 643 177.9	6 576 929.2	319 757.3	41 700 000

Input x_3 is used as a numeraire to accomplish linear homogeneity in inputs and a translog input distance function derived from (4.8) is used. The estimated model including the dynamic specification in (4.17) to (4.19) for the inefficiency distribution is the following:

$$\begin{aligned}
-\ln x_{3_{it}} = \;& \beta_0 + \textstyle\sum_{m=1}^{2} \beta_m \ln y_{m_{it}} + \sum_{r=1}^{2} \delta_r \ln\left(\frac{x_{r_{it}}}{x_{3_{it}}}\right) + \frac{1}{2} \sum_{m=1}^{2} \sum_{n=1}^{2} \beta_{mn} \ln y_{m_{it}} \ln y_{n_{it}} \\
& + \tfrac{1}{2} \textstyle\sum_{r=1}^{2} \sum_{s=1}^{2} \delta_{rs} \ln\left(\frac{x_{r_{it}}}{x_{3_{it}}}\right) \ln\left(\frac{x_{s_{it}}}{x_{3_{it}}}\right) + \sum_{m=1}^{2} \sum_{r=1}^{2} \eta_{mr} \ln y_{m_{it}} \ln\left(\frac{x_{r_{it}}}{x_{3_{it}}}\right) \\
& + \kappa_1 t + \tfrac{1}{2} \kappa_2 t^2 + \textstyle\sum_{m=1}^{2} \phi_m t \ln y_{m_{it}} + \sum_{r=1}^{2} \varphi_r t \ln\left(\frac{x_{r_{it}}}{x_{3_{it}}}\right) - u_{it} + vit \\
& \log u_{it} = \theta_{it} + \textstyle\sum_{p=1}^{3} \gamma_p z_{p_{it}} + \xi_{it};\; \xi_{it} \sim N(0, \sigma_\xi^2) \\
& \theta_{it} = \omega + \rho_i \theta_{i,t-1} + \textstyle\sum_{p=1}^{3} \psi_p h_{p_{it}} + \eta_{it};\; \eta_{it} \sim N(0, \sigma_\eta^2);\; t = 2...T \\
& \theta_{i1} = \frac{\omega + \sum_{p=1}^{3} \psi_p h_{p_{i1}}}{1-\rho_i} + \eta_{i1};\; \eta_{i1} \sim N\left(0, \frac{\sigma_\eta^2}{1-\rho_i^2}\right);\; t = 1.
\end{aligned} \tag{4.20}$$

In addition to linear homogeneity in inputs, we impose cross-effects symmetry by requiring $\beta_{mn} = \beta_{nm}$ and $\delta_{rs} = \delta_{sr}$.

4.2.2.3 Estimation results

Using the specification in (4.20) we estimate four different models by adding some restrictions. Model I follows the same inefficiency specification in Tsionas (2006) by including all three heterogeneity variables in the inefficiency dynamics. Therefore, $\rho_i = \rho$ and γ_1, γ_2, γ_3 and ξ_{it} are all equal to 0. Model II consider all heterogeneity variables out of the dynamic component but still restricts persistence to be common for all banks. Thus, $\rho_i = \rho$ and ψ_1, ψ_2 and ψ_3 are equal to zero. Model III combines heterogeneity variables in and out the dynamic component. In particular, following results in models I and II, we set γ_3, ψ_1, and ψ_2 equal to zero and we keep $\rho_i = \rho$. Finally, Model IV uses the same combination of heterogeneity variables in Model III but allows for bank specific persistence parameters ($\rho_i \neq \rho$).

Table 4.5 presents the estimation results for all models. If we compare Model I to Model II we observe two main relevant results. First, Model II exhibits lower values for DIC_3 and LPS suggesting better fit and prediction performance when heterogeneity is modeled out of the inefficiency dynamics. Second, variables regarding size become relevant as technical inefficiency drivers and seem to present negative but decreasing effects. This would suggest that size affects the inefficiency level at every period but that its effects can be rapidly adjusted. On the other hand, foreign ownership presents relevant negative effects in technical inefficiency under both models. Consequences of these differences in the technical efficiency estimations are explored by selecting banks with different characteristics in terms of size and ownership.

Figure 4.12 compares the posterior technical efficiency distributions for two banks with different sizes obtained from both models. One bank from the first quartile (Bank A) and one bank from the fourth quartile (Bank B) of the sample are selected in terms of assets level. We observe that while in Model I the posterior distributions of the technical efficiencies of both banks are almost undistinguishable, in Model II Bank B seems to have a high probability of being more efficient than Bank A. This shows that size becomes important for differentiating banks in terms of their technical efficiency only when it is modeled out of the dynamic component. This warns us of the possibility of biases in efficiency estimations and wrong conclusions about the effects of heterogeneity variables in dynamic inefficiency models when their effect is only considered as part of the dynamics.

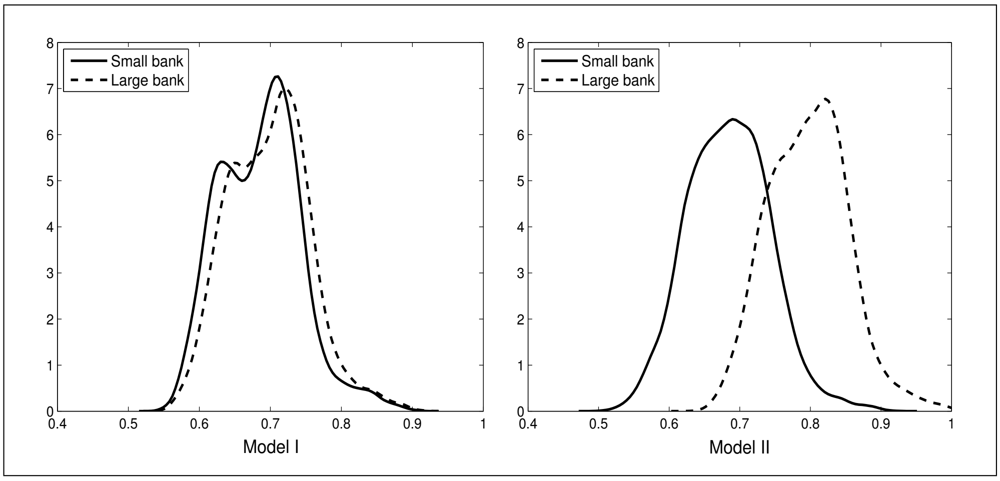
Table 4.5: Posterior mean and standard deviation of parameter distributions

Parameter	Model I $\psi_1, \psi_2, \psi_3 = 0$ $\rho_i = \rho$		Model II $\gamma_1, \gamma_2, \gamma_3 = 0$ $\rho_i = \rho$		Model III $\psi_1, \psi_2, \gamma_3 = 0$ $\rho_i = \rho$		Model IV $\psi_1, \psi_2, \gamma_3 = 0$ $\rho_i \neq \rho$	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Distance function								
β_0	-5.3801	0.3807	-5.3883	0.3813	-5.4324	0.2145	-5.2458	0.3438
$\beta_1(\ln y_i)$	-0.0216	0.0010	-0.0443	0.0021	-0.0344	0.0019	-0.0167	0.0013
$\beta_2(\ln y_c)$	-0.0065	0.0027	-0.0029	0.0012	-0.0061	0.0005	-0.0080	0.0007
$\beta_3(\ln dep)$	0.1290	0.0169	0.2284	0.0299	0.2240	0.0560	0.1773	0.0260
$\beta_4(\ln l)$	0.1029	0.0243	0.1172	0.0277	0.1198	0.0341	0.0947	0.0272
$\beta_5(t)$	0.0206	0.0097	0.0283	0.0133	0.0153	0.0070	0.0193	0.0031
$\beta_6(t^2)$	-0.0004	0.0002	-0.0003	0.0002	-0.0002	0.0001	-0.0003	0.0001
$\phi_1(1/2 \ln y_i^2)$	-0.0808	0.0232	-0.0844	0.0243	-0.0857	0.0200	-0.0873	0.0234
$\phi_2(\ln y_i \ln y_c)$	0.0734	0.0225	0.0713	0.0218	0.0692	0.0179	0.0644	0.0153
$\phi_3(1/2 \ln y_c^2)$	-0.1429	0.0223	-0.1458	0.0227	-0.1346	0.0219	-0.1343	0.0153
$\phi_4(1/2 \ln dep^2)$	-0.2878	0.0988	-0.3098	0.1064	-0.3230	0.0491	-0.4601	0.0433
$\phi_5(\ln dep \ln l)$	0.1136	0.0491	0.1334	0.0577	0.1293	0.0379	0.1572	0.0569
$\phi_6(1/2 \ln l^2)$	-0.0699	0.0437	-0.1004	0.0628	-0.0753	0.0299	-0.0380	0.0158
$\delta_1(\ln y_i \ln dep)$	-0.0332	0.0367	-0.0296	0.0327	-0.0122	0.0164	-0.0163	0.0175
$\delta_2(\ln y_i \ln l)$	-0.0088	0.0123	-0.0187	0.0260	-0.0251	0.0157	-0.0360	0.0127
$\delta_3(\ln y_c \ln dep)$	0.1530	0.0306	0.1707	0.0342	0.1420	0.0165	0.1486	0.0233
$\delta_4(\ln y_c \ln l)$	-0.0171	0.0244	-0.0170	0.0244	-0.0118	0.0123	-0.0460	0.0130
$\kappa_1(\ln y_i t)$	0.0007	0.0012	0.0011	0.0018	0.0004	0.0011	0.0028	0.0047
$\kappa_2(\ln y_c t)$	-0.0050	0.0022	-0.0058	0.0025	-0.0043	0.0010	-0.0054	0.0018
$\kappa_3(\ln dep t)$	0.0096	0.0054	0.0092	0.0051	0.0097	0.0028	0.0117	0.0074
$\kappa_4(\ln lt)$	-0.0068	0.0037	-0.0062	0.0034	-0.0068	0.0027	-0.0016	0.0114
Inefficiency								
ω	0.0022	0.0003	0.0024	0.0002	0.0015	0.0002	0.0012	0.0002
$\psi_1(\ln assets)$	-0.0001	0.0003						
$\psi_2(\ln assets^2)$	0.0000	0.0001						
$\psi_3(foreign)$	-0.0033	0.0008			-0.0041	0.0003	-0.0032	0.0003
ρ	0.9954	0.0172	0.9749	0.0168	0.9670	0.0183	0.8867	0.0452
$\gamma_1(\ln assets)$			-0.0245	0.0011	-0.0189	0.0010	-0.0458	0.0017
$\gamma_2(\ln assets^2)$			0.0002	0.0001	0.0009	0.0002	0.0023	0.0007
$\gamma_3(foreign)$			-0.2061	0.0162				
σ_v^2	0.0012	0.0008	0.0011	0.0007	0.0013	0.0009	0.0017	0.0009
σ_η^2	0.0006	0.0002	0.0003	0.0001	0.0007	0.0002	0.0010	0.0002
σ_ξ^2			0.0005	0.0002	0.0002	0.0001	0.0002	0.0001
DIC_3	-2789.60		-2949.20		-2961.52		-3094.10	
LPS	-213.72		-253.48		-261.79		-295.63	

Note: In Model IV the values reported for ρ correspond to the posterior means of the individual ρ_i's

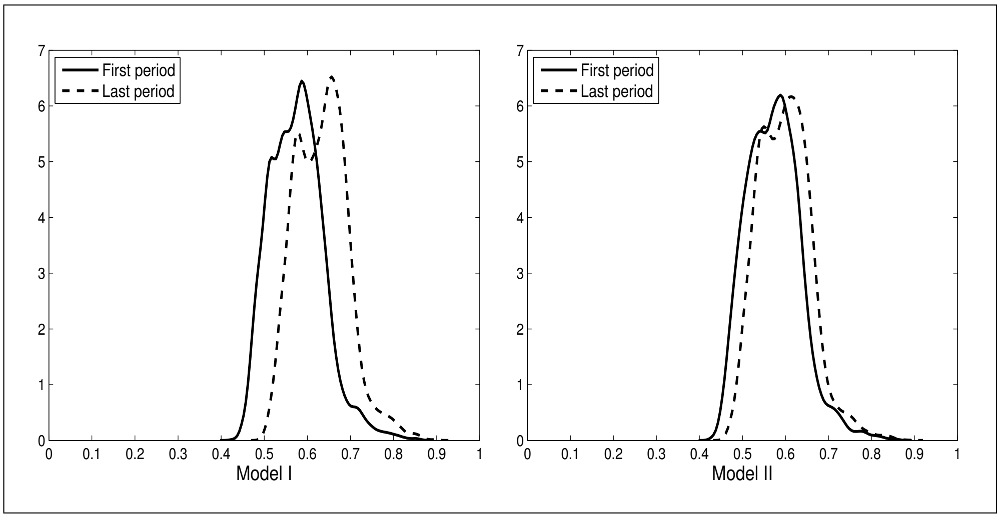
Foreign ownership is found to be a relevant inefficiency driver when it is included both in and out of the dynamic component. Thus, we explore the consequences for the efficiency estimations of a foreign bank under both models and in two different periods. Figure 4.13 shows that posterior efficiency distributions are quite similar in both models in the first period but that they become different at the end of the sample period. Since the value of the variable for this bank is the same during all the sample, this may suggest

Figure 4.12: Posterior technical efficiency distribution for a small and large bank



that the effect of including a covariate in the inefficiency dynamics could be cumulative after many periods despite of the lower estimate for the coefficient. A possible reason is that given that persistence is very high, most of the effect of the covariate is transmitted to the next period, where once more it affects the inefficiency.

Figure 4.13: Posterior technical efficiency distribution of a foreign bank



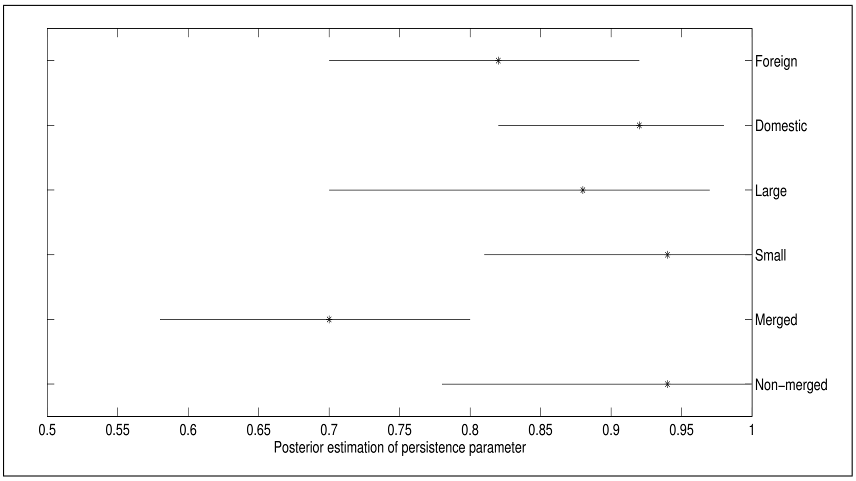
Since the location of foreign ownership may lead to different efficiency estimates, we estimate a third model. Model III includes foreign ownership in the inefficiency dynamics while keeps the assets variables out of this component. Results in terms of fitting performance and prediction improve compared to those of the previous models and the coefficient for the variable remains relevant as inefficiency driver. This suggests that foreign-owned banks present lower technical inefficiency and that the effects derived from this type of ownership are persistent over time.

In general, inefficiency persistence is found to be very high in all models. This result is very similar to that obtained by Tsionas (2006) for US banks. However, we explore differences among banks by including a firm-specific persistence parameter in Model IV. We observe that this model exhibits lower DIC_3 and LPS values compared to Model III, which suggests that recognizing heterogeneous costs of adjustment may improve the fit and predictive performance of the model and have effects on the evolution of the efficiency if the estimated persistence parameters are very different among banks. In order to identify these differences, in Figure 4.14 we plot the posterior mean and 95% probability intervals for ρ_i. We classify banks in three main categories: foreign and domestic banks, large and small banks, and merged and non-merged institutions.[21] We observe differences in the average posterior mean between all complementary groups. In particular, foreign, large and merged banks are more likely to present lower adjustment costs than their counterparts. However, the persistence parameters between merged and non-merged institutions are those with the highest probability of being different as suggested from the very small overlapping between both probability intervals. The average posterior mean for ρ_i among merged banks (0.71) is the lowest compared to those of the other groups and it is not only far from that estimated for non-merged banks (0.94) but also from that estimated for ρ in Model III (0.97) when the parameter is assumed to be common to all banks.

These differences may have important effects in the dynamic behavior of inefficiency over time between both groups of banks. To illustrate these effects we plot in Figure 4.15 the evolution of the mean posterior technical efficiency estimated from models III and IV for merged and non-merged banks. It is observed that efficiency of merged banks decreases immediately in both models after these processes are carried out. However, Model IV identifies a rapid recovery of the efficiency of merged banks that starts around three years after the merging process and reaches the non-merged efficiency levels after five years. This pattern is totally different from that identified in Model III, where technical efficiency of merged institutions seems to remain lower than that of non-merged banks.

[21] We define small and large banks as those below and above the median of assets level, respectively.

Figure 4.14: Posterior median and 95% probability intervals for firm specific persistence parameters by type of bank

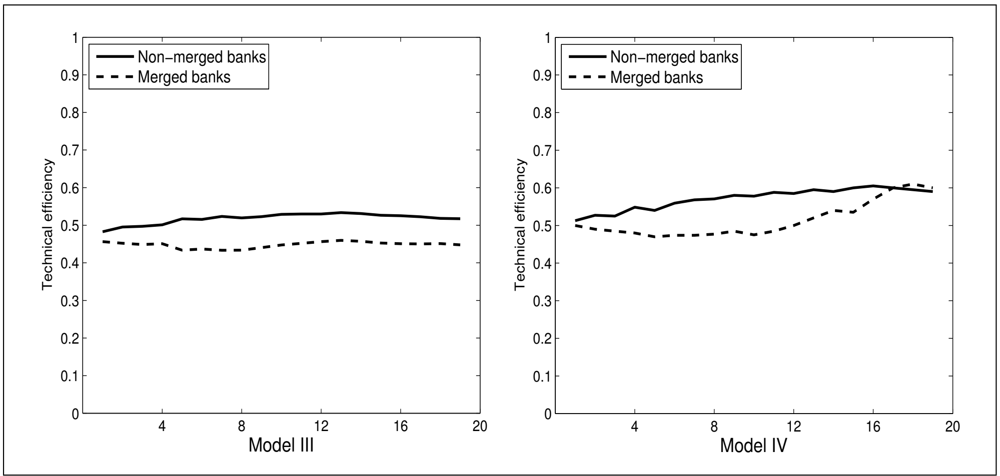


These results contrast with those of previous studies that measure the effect of banks M&A whether on cost or input-oriented technical efficiency (see Amel et al., 2004, for a review on studies in developed countries). However, the pattern on the evolution of input-oriented technical efficiency that we see for Colombian merged banks in Model IV is similar to that identified by Cuesta and Orea (2002) in a study of output-oriented technical efficiency of Spanish merged banks. In that study, technical efficiency was found to exhibit a concave pattern with negative but decreasing effects during the first six years after mergers, and positive increasing effects after that point. Although the model estimated by Cuesta and Orea (2002) is not dynamic in nature, it allows merged banks to follow a different temporal pattern to that of non-merged institutions. This may suggest that mergers lead to different evolution processes of the inefficiency and that models recognizing these differences are more appropriate.

With respect to the inefficiency drivers in the Colombian banking sector, foreign ownership and size are found to have positive effects on technical efficiency. However, the impact is decreasing for size. Moreover, we identify that the effects of size on inefficiency can be rapidly adjusted by Colombian banks, while the advantages presented by foreign banks are difficult and costly to reach or adjust.

Finally, we compare TE, TC and RTS by groups of banks following the results from Model IV. Table 4.6 summarizes these findings. We observe that foreign banks in Colombia present higher technical efficiency than domestic institutions, as well as

Figure 4.15: Evolution of mean posterior efficiencies for merged and non-merged banks



higher technical change during the last decade. These findings coincide with those reported in other recent studies for developing countries using non-dynamic models (see Claessens and Horen, 2012). Beyond managerial practices, the reasons could be related to more diversification, parents expertise, or access to cheaper and multiple sources of financial resources (see Chen and Liao, 2011). In contrast to domestic institutions, foreign banks also present increasing returns to scale, suggesting that these institutions have more room to raise their production scale and possibly to take M&A decisions. Foreign banks in Colombia are characterized for being specialized in corporate clients, offer complex products and have few branches with low operations. In a recent study, Das and Kumbhakar (2011) also found similar scale economies for foreign banks in India.

Table 4.6: TE, TC and RTS by type of bank

Bank type	TE	TC	RTS
Foreign	0.6011	0.0307	1.0986
Domestic	0.5476	0.0251	0.9180
Large	0.5853	0.0285	0.9202
Small	0.5304	0.0278	1.0316
Merged	0.5076	0.0326	0.9021
Non-Merged	0.5512	0.0267	1.0633

In terms of size, we find that large banks present higher TE and TC than small institutions during the period. However, large institutions are found to operate at decreasing returns to scale in contrast to small banks. Higher efficiency of large banks and potential scale gains for small banks were also recently found by Tabak and Tecles

(2010) and Tecles and Tabak (2010) in India and Brazil, respectively. In particular, in Colombia most large domestic banks are those involved in merger processes. Since merged institutions also present decreasing returns to scale, this may suggest that these processes led banks to be oversized. Also, on average, merged banks exhibit lower technical efficiency than non-merged institutions. However, they are found to present lower adjustment costs that allow them to adjust quicker towards optimal conditions. Thus, they would be able to present higher efficiency after some periods. Finally, technical change is also found to be higher for merged than for non-merged banks and it can be a consequence of the reorganization processes implied by mergers.

4.2.2.4 Empirical implications

Our findings suggest that modeling covariates in and out of the inefficiency dynamics have implications on the identification of inefficiency determinants and on the efficiency estimations. In particular, we find that foreign ownership has negative and persistent effects in technical inefficiency of Colombian banks. This may suggest that some characteristics of foreign banks such as country diversification and access to cheaper funding sources can be difficult and costly to obtain or change. On the other hand, the effects of bank size on technical inefficiency are found to be rapidly adjusted.

Colombian banks are also found to present very high inefficiency persistence, coinciding with previous findings in the US and Spanish banking sectors (see Tortosa-Ausina, 2002; Tsionas, 2006, respectively). However, important differences are observed among banks with different characteristics when firm specific persistence parameters are modeled. Foreign, large and merged institutions are found to present lower adjustment costs than their counterparts. This suggests that these institutions may benefit from diversification or economies of scale when carrying out adjustments in their short-run operations as these are costly for domestic, small and non merged banks. This finding is particularly important for merged banks since this characteristic allows them to recover rapidly from efficiency losses observed after merger processes.

These results are of interest not only for financial institutions, but also for regulators given the importance that M&A have had in the sector in recent years and the role of foreign banks in developing countries. In particular, although, our findings reveal important decreases in efficiency of merged institutions during the initial years after these processes are carried out, the lower inefficiency persistence of banks involved in

M&A and the non-persistent effects of size on inefficiency may validate these processes in the mid-term. However, the Colombian regulator should be aware of the results on economies of scale, which leave little margin for non-merged institutions to increase their size and reveal decreasing returns for merged and large banks. Exploring market power considerations would be also of interest for future policy decisions in the sector. In general, bank efficiency may be a useful indicator for financial stability considerations given that banks with low efficiency have been found to be more prone to future defaults (see Berger and DeYoung, 1997). In this regard, those banks with high inefficiency persistence should be drawn to the attention of the regulator.

4.3 Conclusions

In the presence of adjustment costs, firms do not find it optimal to adapt their processes towards efficiency. This behaviour can be captured through a dynamic specification for the inefficiency term. One of the most relevant contributions in this context is that by Tsionas (2006) where the inefficiency is allowed to have persistent effects over time and to be driven by inefficiency covariates. In this work we have extended this idea in order to recognize heterogeneity in the adjustment costs among firms and non-persistent effects of observed heterogeneity.

Our findings suggest that accounting for unobserved sources of heterogeneity is also very relevant under a dynamic framework. In particular, allowing the inefficiency persistence parameter to be firm-specific recognizes differences in the adjustment costs among firms, which drive the posterior efficiency estimations. In both applications, firm-specific persistence parameters are found to be very different among firms with different characteristics. This allow us to identify differences in the way firm characteristics affect the evolution of individual efficiencies and important implications for regulators and firms in the electricity and banking sectors.

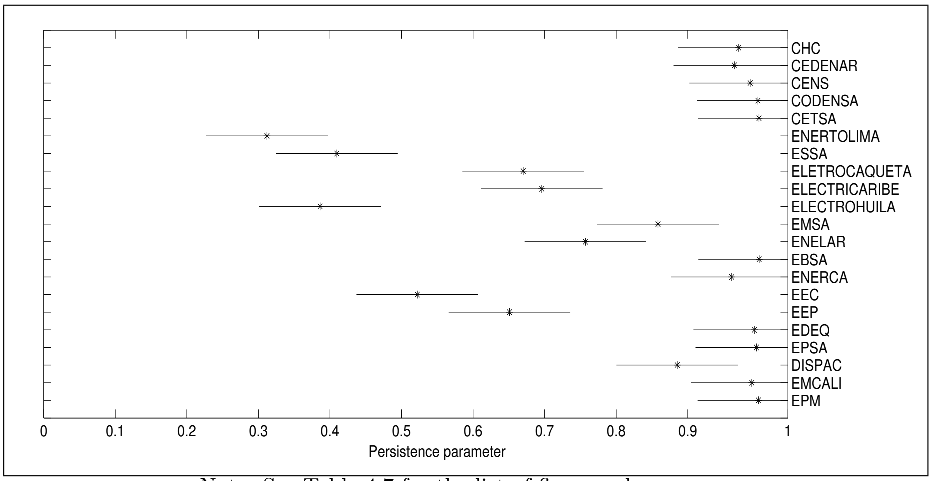
Modeling covariates in and out of the inefficiency dynamics was also found to have relevant effects on the identification of inefficiency determinants and on the efficiency estimations. This points out the implications of including observed firm characteristics in dynamic specifications.

Overall, the proposed specifications encompass other models in the literature and adds more flexibility in terms of considering inefficiency heterogeneity in a dynamic context. This improves the fit and predictive performance of the models and allows us to capture effects that have not been previously identified. Extensions of dynamic inefficiency models such as using alternative distributions for the inefficiency are aspects of interest for future research.

4.4 Appendix

A. Complimentary results from the electricity distribution application

Figure 4.16: 95% probability intervals for firm specific persistence parameters under Model DPH



Note: See Table 4.7 for the list of firms and acronyms

Table 4.7: Average posterior mean estimations of TE and inefficiency persistence (IP) under model DPUH, customer density (users/km) and consumption density (kWh/user) by firm for the period 1998-2012

Firm	TE	IP	Cust. Dens.	Cons. Dens.
Central Hidroeléctrica de Caldas S.A. E.S.P. (CHC)	0.5520	0.9713	31.199	341
Centrales Eléctricas de Nariño S.A. E.S.P. (CEDENAR)	0.3045	0.9651	23.072	474
Centrales Eléctricas del Norte de Santander S.A. E.S.P. (CENSA)	0.6118	0.9872	13.996	94
CODENSA S.A. E.S.P. (CODENSA)	0.9894	0.9981	47207	830
Compañía de Electricidad de Tuluá S.A. E.S.P. (CETSA)	0.9892	0.9996	47.355	2116
Compañía Energética del Tolima S.A E.S.P (ENERTOLIMA)	0.4667	0.3120	13.205	77
Electrificadora de Santander S.A. E.S.P. (ESSA)	0.4624	0.4096	33639	152
Electrificadora del Caquetá S.A. E.S.P. (ELECTROCAQUETA)	0.4977	0.6700	20.120	209
Electrificadora del Caribe S.A. E.S.P. (ELECTRICARIBE)	0.4506	0.6960	40.553	336
Electrificadora del Huila S.A. E.S.P. (ELECTROHUILA)	0.4720	0.3862	16.663	94
Electrificadora del Meta S.A. E.S.P. (EMSA)	0.5033	0.8584	39.699	261
Empresa de Energía de Arauca E.S.P (ENELAR)	0.4260	0.7571	21.334	981
Empresa de Energía de Boyacá S.A. E.S.P. (EBSA)	0.9960	0.9999	21.356	237
Empresa de Energía de Casanare S.A. E.S.P. (ENERCA)	0.3677	0.9615	13.352	110
Empresa de Energía de Cundinamarca S.A. ESP (EEC)	0.4760	0.5221	42.579	153
Empresa de Energía de Pereira S.A. E.S.P. (EEP)	0.4913	0.6509	21.193	299
Empresa de Energía del Quindío S.A.E.S.P. (EDEQ)	0.6487	0.9930	33.337	452
Empresa de Energía del Pacífico S.A. E.S.P. (EPSA)	0.7303	0.9959	50.925	269
Empresa Distribuidora del Pacífico S.A. E.S.P (DISPAC)	0.4233	0.8853	22.464	475
Empresas Municipales de Cali E.I.C.E E.S.P (EMCALI)	0.7328	0.9895	61.707	2331
Empresas Públicas de Medellín E.S.P. (EPM)	0.9015	0.9988	82.735	389

B. WinBUGS code for a dynamic model with unobserved inefficiency and technological heterogeneity - Colombian electricity distribution application

```
model {
#K=number of firms
#N=number of observations
#M=N-K
#obs=N*1 vector with numbers of observations from 1 to N
#obsK=K*1 vector with the number of the first observation of every firm
#obsM=M*1 vector with the number of the rest of observations of every firm

#Inefficiency specification for the first observation of every firm:
for (i in 1:K) {
 log(u[obsK[i]])<-(c/(1-rhofirm[obs[obsK[i]]]))+z[obsK[i]]+eta[obsK[i]]
 z[obsK[i]] <- gamma[1]*lnz1[obsK[i]]+gamma[2]*lnz2[obsK[i]]
 tao1[i] <- tao*(1-rhofirm[obs[obsK[i]]]*rhofirm[obs[obsK[i]]])
 eta[obsK[i]] ~ dnorm(0,tao1[i]) #(c),(d) }
#Inefficiency specification for the rest of observations of every firm:
for (i in 1:M) {
 log(u[obsM[i]])<-c+rhofirm[obs[obsM[i]]]*log(u[obsM[i]-1])+z[obsM[i]]
                +eta[obsM[i]]
 z[obsM[i]]<-gamma[1]*lnz1[obsM[i]]+gamma[2]*lnz2[obsM[i]]
 eta[obsM[i]] ~ dnorm(0,tao) }

for (i in 1:N) {
 mu[i] <- alphafirm[obs[i]]- u[i]+ beta[1]*lny1[i]+beta[2]*lny2[i]+beta[3]*lnx2[i]
         +beta[4]*lnx3[i]+beta[5]*lnkm[i]+beta[6]*t[i]+beta[7]*(t[i]*t[i])+tl[i]
         +tlcross[i]+tlt[i]
 tl[i] <- 0.5*phi[1]*(lny1[i]*lny1[i])+phi[2]*(lny1[i]*lny2[i])
         +0.5*phi[3]*(lny2[i]*lny2[i])+0.5*phi[4]*(lnx2[i]*lnx2[i])
         +phi[5]*(lnx2[i]*lnx3[i])+0.5*phi[6]*(lnx3[i]*lnx3[i])
 tlcross[i] <- delta[1]*(lny1[i]*lnx2[i])+delta[2]*(lny2[i]*lnx2[i])
         +delta[3]*(lny1[i]*lnx3[i])+delta[4]*(lny2[i]*lnx3[i])
 tlt[i] <- kappa[1]*(t[i]*lny1[i])+kappa[2]*(t[i]*lny2[i])+kappa[3]*(t[i]*lnx2[i])
         +kappa[4]*(t[i]*lnx3[i])
 lnx1[i] ~ dnorm(mu[i], prec)
 eff[i]<- exp(-u[i]) }

c ~ dnorm(-1.5, 1)
for (i in 1:2) {
```

```
 gamma[i] ~ dnorm(0.0, 0.1) }

for (k in 1:K) {
 hfirm[k]~ dbeta(h,h1)
 rhofirm[k] <- 2*hfirm[k]-1
 alphafirm[k]~ dnorm(alpha, 0.1) }

h ~ dbeta(0.5,0.5)
h1<-1-h
rho <- 2*h-1
tao ~ dgamma(10,0.01)

alpha ~ dnorm(0.0, 0.001)
for (i in 1:7) {
 beta[i] ~ dnorm(0.0, 0.001) }
for (i in 1:6) {
 phi[i] ~ dnorm(0.0, 0.001) }
for (i in 1:4) {
 delta[i] ~ dnorm(0.0, 0.001) }
for (i in 1:4) {
 kappa[i] ~ dnorm(0.0, 0.001) }
prec ~ dgamma(0.01, 0.01)
sigmasq <- 1 /prec  }
```

C. WinBUGS code for a dynamic model with heterogeneous persistence and separating heterogeneity from the inefficiency dynamics- Colombian banking sector application

```
model{
#K=number of banks
#N=number of observations
#M=N-K
#obs=N*1 vector with numbers of observations from 1 to N
#obsK=K*1 vector with the number of the first observation of every bank
#obsM=M*1 vector with the number of the rest of observations of every bank

#Inefficiency specification for the first observation of every bank:
for (i in 1:K) {
 log(u[obsK[i]])<-theta[obsK[i]]+z[obsK[i]]+xi[obsK[i]]
 theta[obsK[i]]<-((c+s[obsK[i]])/(1-rhof[obs[obsK[i]]]))+eta[obsK[i]]
 z[obsK[i]]<-gamma[1]*z1[obsK[i]]+gamma[2]*z2[obsK[i]]
 s[obsK[i]]<-omega*z3obsK[i]]
 tao1[i]<-tao*(1-(rhof[obs[obsK[i]]]*rhof[obs[obsK[i]]]))
 eta[obsK[i]]~dnorm(0,tao1[i])
 xi[obsK[i]]~dnorm(0,lambda) }
#Inefficiency specification for the rest of observations of every bank:
for (i in 1:M) {
 log(u[obsM[i]])<-theta[obsM[i]]+z[obsM[i]]+xi[obsM[i]]
 theta[obsM[i]]<-c+s[obsM[i]]+rhof[obs[obsM[i]]]*theta[obsM[i]-1]+eta[obsM[i]]
 z[obsM[i]]<-gamma[1]*z1[obsK[i]]+gamma[2]*z2[obsK[i]]
 s[obsM[i]]<-omega*z3[obsM[i]]
 eta[obsM[i]]~dnorm(0,tao)
 xi[obsM[i]]~dnorm(0,lambda) }

for (i in 1:N) {
 mu[i]<-alpha - u[i]+ beta[1]*lnyi[i]+beta[2]*lnyc[i]+beta[3]*lndep[i]
        +beta[4]*lnl[i]+beta[5]*t[i]+beta[6]*(t[i]*t[i])+tl[i]+tlcross[i]+tlt[i]
 tl[i]<-0.5*phi[1]*(lnyi[i]*lnyi[i])+phi[2]*(lnyi[i]*lnyc[i])
        +0.5*phi[3]*(lnyc[i]*lnyc[i])+0.5*phi[4]*(lndep[i]*lndep[i])
        +phi[5]*(lndep[i]*lnl[i])+0.5*phi[6]*(lnl[i]*lnl[i])
 tlcross[i]<-delta[1]*(lnyi[i]*lndep[i])+delta[2]*(lnyi[i]*lnl[i])
        +delta[3]*(lnyc[i]*lndep[i])+delta[4]*(lnyc[i]*lnl[i])
 tlt[i]<-kappa[1]*(t[i]*lnyi[i])+kappa[2]*(t[i]*lnyc[i])+kappa[3]*(t[i]*lndep[i])
        +kappa[4]*(t[i]*lnl[i])
 lnx1[i]~dnorm(mu[i], prec)
```

```
 eff[i]<-exp(-u[i]) }

c~dnorm(-1.5,1)
for (i in 1:2) {
 gamma[i]~dnorm(0.0,1) }
omega~dnorm(0.0,1)
lambda~dgamma(0.5,0.005)
h~dbeta(0.5,0.5)
rho<-2*h-1
tao~dgamma(10,0.01)
h1<-1-h
for (k in 1:K) {
 hf[k]~dbeta(h,h1)
 rhof[k]<-2*hf[k]-1 }
alpha~dnorm(0.0,0.001)
for (i in 1:2) {
 beta[i]~dnorm(0.0,0.001) }
for (i in 3:4) {
 beta[i]~dnorm(0.0,0.001) }
for (i in 5:6) {
 beta[i]~dnorm(0.0,0.001) }
for (i in 1:6) {
 phi[i]~dnorm(0.0,0.001) }
for (i in 1:4) {
 delta[i]~dnorm(0.0,0.001) }
for (i in 1:4) {
 kappa[i]~dnorm(0.0,0.001) }
prec~dgamma(0.01,0.01)
sigmasq<-1/prec  }
```

Chapter 5

Discussion and further research

In stochastic frontier analysis the inefficiency component may be erroneously estimated when firm characteristics are not taken into account. These firm characteristics induce heterogeneity that might result in different firm frontiers, or may have an impact directly on the inefficiencies. This issue has been widely studied before. However, unobserved inefficiency heterogeneity has been little explored. In this thesis, we have put forward the modeling of heterogeneity in a Bayesian context by capturing both observed and unobserved heterogeneity in the inefficiency distribution under static and dynamic formulations.

The first of our proposed methods captures unobserved heterogeneity in the inefficiency by modeling a random parameter in the inefficiency distribution, which can be allowed to be whether time invariant or time-varying. Our findings suggest that unobserved inefficiency heterogeneity can be properly captured by this random parameter. Models including this parameter whether alone or simultaneously with observed covariates improve in terms of fit and predictive performance as long as latent heterogeneity remains unidentified. In this sense, it can be used to distinguish unobserved heterogeneity from inefficiency and to validate the suitability of observed covariates to capture it.

Also, the effects of including both types of heterogeneity in different parameters of the inefficiency distribution were studied. Differences in efficiency rankings and mean scores were observed when inefficiency heterogeneity was included in different parameters of the one-sided error distribution. This was found to be related to effects in the posterior efficiency distributions. In particular, considering firms' heterogeneity in the location parameter of the inefficiency has an effect on separating the firm specific posterior efficiency distributions from each other, which leads to more reliable rankings. On the other hand, when heterogeneity affects only the scale parameter of the inefficiency, an important shrinking effect is observed on the individual posterior efficiency distributions. This results in less uncertainty around mean individual efficiency scores. Finally, including the heterogeneity in both parameters of the inefficiency distribution in models that preserve the scaling property leads to both separating and shrinking effects. This allows less overlapping of the posterior efficiency distributions and provide both more reliable efficiency scores and rankings. These results are consistent whether we use observed covariates or our proposal to model unobserved heterogeneity.

Our second proposal introduces random firm specific coefficients for covariates in the inefficiency. It was found that allowing random coefficients for the inefficiency covariates captures firm-specific effects which remain unidentified under the regular fixed coefficients models. This specification distinguishes firms in terms of the effects of inefficiency drivers and separates unobserved heterogeneity related to these effects from efficiency. This was found to have relevant implications for regulation an policy making.

We extended the study of observed and unobserved heterogeneity to dynamic SFA models. This is a topic with relative little attention within the frontier literature. Here previous proposals were extended in order to capture two possible sources of unobserved heterogeneity. One related to heterogeneity in the inefficiency persistence and the second one related to unobserved technological heterogeneity. Both unobserved sources were found to be very relevant in an empirical application to electricity distribution utilities. In particular, heterogeneity in the inefficiency persistence was found to be very important in explaining inefficiency and its evolution over time. Finally, the implications of including observed covariates in dynamic models were studied by mean of an inefficiency specification that allows separating observed inefficiency heterogeneity from the dynamic process. The model allows identifying those firm characteristics that may have persistent effects in the inefficiency from those that can be rapidly adjusted. In general, location of observed covariates was found to have important implications in the identification of inefficiency drivers and posterior efficiency estimations.

Overall, new specifications to model both observed and unobserved heterogeneity in stochastic frontier models have been proposed in a Bayesian context. These models identify effects that are not captured with other models in the literature, which are shown to be very useful for firms, regulators, and policy maker in different sectors, from health to electricity and banking.

We identify interesting future research lines from both empirical and theoretical standpoints. Applications of the proposed models to any regulated sector are of relevance due to important concerns on accounting for unobserved heterogeneity sources when comparing firms in these sectors. Also applications to less explored areas such as investment funds performance may have important implications. Models proposed here can be easily implemented in the identification of persistence effects in the efficiency of the risk-return relationship of funds and the differences among types of firms.

Theoretically, we think that extending our models to include common time specific factors as in Bai (2009) would be interesting in order to account for spillover spatial effects. This would allow to model heterogeneity related to geographical aspects. Also, an extension of the dynamic models to the use of Bayesian vector autoregressive and state space representations would allow to model endogeneity issues in a more proper way (see Mastromarco and Woitek, 2012a,b, for the use of these techniques in efficiency measurement). Finally, modeling inefficiency heterogeneity in nonparametric SFA models is an area of great interest. These models have the advantage of adding flexibility to the inefficiency component while preserving an stochastic error term. In this context, there are some recent contributions from both the frequentist and Bayesian approaches, but modeling inefficiency heterogeneity and studying its implication in these models is still an open topic.

References

Ahmed, S. and A. Zlate (2013). Capital flows to emerging market economies: A brave new world? Working Paper No. 1081, International Finance Discussion Papers, Board of Governors of the Federal Reserve System. 48

Ahn, S., D. Good, and R. Sickles (2000). Estimation of long-run inefficiency levels: A dynamic frontier approach. *Econometrics Review 19*, 461–492. 74

Aigner, D., C. Lovell, and P. Schmidt (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics 6*, 21–37. iii, 4, 7, 11

Alvarez, A., C. Amsler, L. Orea, and P. Schmidt (2006). Interpreting and testing the scaling property in models where inefficiency depends on firm characteristics. *Journal of Productivity Analysis 25*, 201–212. 12, 18, 19

Amel, D., C. Barnes, F. Panetta, and C. Salleo (2004). Consolidations and efficiency in the financial sector: A review of the international evidence. *Journal of Banking and Finance 28*, 2493–2519. 103, 110

Assaf, A., R. Matousek, and E. Tsionas (2013). Turkish bank efficiency: Bayesian estimation with undesirable outputs. *Journal of Banking and Finance 37*, 506–517. 49, 103

Bai, J., 2009. Panel Data Models With Interactive Fixed Effects. Econometrica 77, 1229–1279. 123

Balk, BM., 2001. Scale Efficiency and Productivity Change. Journal of Productivity Analysis 15, 159–183. 86

Battese, G. and T. Coelli (1992). Frontier production function, technical efficiency and panel data: With application to paddy farmers in India. *Journal of Productivity Analysis 3*, 153–169. 12, 73

Battese, G. and T. Coelli (1995). A model for technical inefficiency effects in a stochastic frontier production model for panel data. *Empirical Economics 20*, 325–332. 11, 18, 19, 101

Berger, A. and R. DeYoung (1997). Problem loans and cost efficiency in commercial banks. *Journal of Banking and Finance 21(6)*, 849–870. 58, 113

Bertay, A., A. Demirgüç-Kunt, and H. Huizinga (2013). Do we need big banks? Evidence on performance, strategy and market discipline. *Journal of Financial Intermediation 22*, 532–558. 58

BIS (2010). Basel III: International framework for liquidity risk measurement, standards and monitoring. Basel Committee on Banking Supervision December, Bank for International Settlements. 48

Borio, C. and H. Zhu (2012). Capital regulation, risk-taking and monetary policy: A missing link in the transmission mechanism? *Journal of Financial Stability 8 (4)*, 236–251. 50

Bos, J., M. Koetter, J. Kolari, and C. Kool (2009). Effects of heterogeneity on bank efficiency scores. *European Journal of Operational Research 195*, 251–261. iv

Bos, J. and C. Kool (2006). Bank efficiency: The role of bank strategy and local market conditions. *Journal of Banking and Finance 30*, 1953–1974. 50, 57

Bruno, V. and H. Song (2013). Capital flows and the risk-taking channel of monetary policy. NBER Working Paper No. 18942. 48

Caudill, S. and J. Ford (1993). Biases in frontier estimation due to heteroscedasticity. *Economics Letters 41*, 17–20. 12

Caudill, S., J. Ford, and D. Gropper (1995). Frontier estimation and firm-specific inefficiency measures in the presence of heteroskedasticity. *Journal of Business and Economics Statistics 13*, 105–111. 12, 19

Cazals, C., J. Florens, and L. Simar (2002). Nonparametric Frontier Estimation: A Robust Approach. *Journal of Econometrics 106*, 1–25. 5

Celeux, G., F. Forbes, C. Robert, and D. Titterington (2006). Deviance information criteria for missing data models. *Bayesian Analysis 4*, 651–674. 21

Cepeda, A., H. Mutis, J. Perez, and M. Sarmiento (2013). Nueva evidencia sobre la eficiencia de la banca colombiana: Una medición con modelos de frontera no-paramétricos. Archivos de Economía 392, Departamento Nacional de Planeación. 51

Charnes, A., W. Cooper, and E. Rhodes (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research 2*, 429–444. iii, 4

Chen, S. and C. Liao (2011). Are foreign banks more profitable than domestic banks? Home- and host-country effects of banking market structure, governance, and supervision. *Journal of Banking and Finance 35*, 819–839. 50, 57, 111

Claessens, S. and N. Horen (2012). Foreign banks: Trends, impact and financial stability. Working Paper, International Monetary Fund. 111

Clavijo, S., C. Rojas, C. Salamanca, G. Montoya, and C. Rizo (2006). Mergers and acquisitions in the colombian financial sector: Impact on efficiency (1990-2005). *Icfai Journal of Mergers and Acquisitions 3*, 38–61. 51, 104

Coelli, T., S. Perelman, and E. Romano (1999). Accounting for environmental influences in stochastic frontier models: With application to international airlines. *Journal of Productivity Analysis 11*, 251–273. iv, 29, 30, 31

Coelli, T., D. Rao, C. O'Donnell, and G. Battese (2005). *An Introduction to Efficiency and Productivity Analysis*. New York: Springer. iii, 2, 3

Cornwell, C., P. Schmidt, and R. Sickles (1990). Production frontiers with cross-sectional and time-series variation in efficiency levels. *Journal of Econometrics 46*, 185–200. 73

Cuesta, R. and L. Orea (2002). Mergers and technical efficiency in spanish savings banks: A stochastic distance function approach. *Journal of Banking and Finance 26*, 2231–2247. 63, 103, 110

Daraio, C. and L. Simar (2005). Introducing environmental variables in nonparametric frontier models: A probabilistic approach *Journal of Productivity Analysis 24*, 93–121. 5

Das, A. and S. Kumbhakar (2011). Productivity and efficiency dynamics in indian banking: An input distance function approach incorporating quality of inputs and outputs. *Journal of Applied Econometrics 93*, 205–234. 111

Debreu, G. (1951). The coefficient of resource utilization. *Econometrica 19*, 273–292. iii

Dell'Ariccia, G., L. Laeven, and R. Marquez (2011). Monetary policy, leverage, and bank risk taking. CEPR, Discussion Paper Series, No.8199. 50

Demirgüç-Kunt, A. and H. Huizinga (2004). Market discipline and deposit insurance. *Journal of Monetary Economics 51 (2)*, 375–399. 50

Dyner, I., S. Arango, and E. Larsen (2006). *Understanding The Argentinean and Colombian Electricity Markets. Electricity Market Reform: An International Perspective*, Chapter 17, pp. 595–616. Elsevier Ltd. 79

Emvalomatis, G. (2012). Adjustment and unobserved heterogeneity in dynamic stochastic frontier models. *Journal of Productivity Analysis 37*, 7–16. 75

Emvalomatis, G., S. Stefanou, and A. Lansink (2011). A reduced-form model for dynamic efficiency measurement: Application to dairy farms in Germany and The Netherlands. *American Journal of Agricultural Economics 93*, 131–174. 84

Estrada, D. (2005). An analysis of banking mergers in Colombia. Financial Stability Report July, 69-79, Banco de la República's Monetary and Reserves Division. 104

Estrada, D. and P. Osorio (2004). Efectos del capital financiero en la eficiencia del sistema bancario colombiano. *Ensayos Sobre Política Económica 47*, 162–201. 51

Evans, D., A. Tandon, C. Murray, and J. Lauer (2000). The comparative efficiency of national health systems in producing health: An analysis of 191 countries. Discussion paper no. 29, World Health Organization, EIP/GPE/EQC. 22

Färe, R. and D. Primont (1995). *Multi-output production and duality: Theory and applications*. Netherlands: Kluwer Academic Publishers. 85

Farrell, M. (1957). The measurement of productive efficiency. *Journal of Royal Statistical Society 3*, 253–290. iii

Fernandez, C., E. Ley, and M. Steel (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics 100*, 381–427. 21

Ferreira, J. and M. Steel (2007). Model comparison of coordinate-free multivariate skewed distributions with an application to stochastic frontiers. *Journal of Econometrics 137*, 641–673. 21

Fried, H., C. Lovell, and S. Schmidt (2008). *The Measurement of Productive Efficiency and Productivity Growth*. New York: Oxford University Press. iii

Fries, S. and A. Taci (2005). Cost efficiency of banks in transition: Evidence from 289 banks in 15 post-communist countries. *Journal of Banking and Finance 29*, 55–81. 57

Galán, J. and M. Pollitt (2014a). Inefficiency persistence and heterogeneity in Colombian electricity distribution utilities. Cambridge Working Paper in Economics, EPRG Working Paper 1403, University of Cambridge. 75, 99

Galán, J., H. Veiga, and M. Wiper (2014b). Bayesian estimation of inefficiency heterogeneity in stochastic frontier models. *Journal of Productivity Analysis 42*, 85–101. 17, 38, 99, 101

Galán, J., H. Veiga, and M. Wiper (2014c). Dynamic effects in inefficiency: Evidence from the Colombian banking sector. *European Journal of Operational Research DOI: 10.1016/j.ejor.2014.07.005*, in press. 51, 57, 63, 73, 75, 77, 84

Sarmiento, M. and J. Galán (2014). Heterogeneous Effects of Risk-Taking on Bank Efficiency: A Stochastic Frontier Model with Random Coefficients. Statistics and Econometrics Series (13) WP 14-20, Universidad Carlos III de Madrid. 46

Geweke, J. (1992). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Bernando, J., Berger, J., Dawid, A., and Smith, A., editors, Bayesian Statistics 4*, pp. 641–649. Clarendon Press, Oxford, UK. 39

Giannakis, D., T. Jamasb, and M. Pollitt (2005). Benchmarking and incentive regulation of quality of service: an application to the UK electricity distribution networks. *Energy Policy 33*, 2256–2271. 88

Glass, J. and D. McKillop (2006). The impact of differing operating environments on us credit union performance. *Applied Financial Economics 16 (17)*, 1285–1300. 57

Good, I. (1952). Rational decisions. *Journal of the Royal Statistical Society B 14*, 107–114. 21

Goto, M. and M. Tsutsui (2008). Technical efficiency and impacts of deregulation: An analysis of three functions in U.S. electric power utilities during the period from 1992 through 2000. *Energy Economics 30*, 15–38. 83, 88

Greene, W. (1990). A gamma-distributed stochastic frontier model. *Journal of Econometrics 46*, 141–164. 7

Greene, W. (2004). Distinguishing between heterogeneity and inefficiency: Stochastic frontier analysis of the World Health Organization's panel data on national health care systems. *Health Economics 13*, 959–980. iv, 23, 40

Greene, W. (2005). Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics 126*, 269–303. 10, 11, 13, 14, 37, 84

Greene, W. (2008). *The Econometric Approach to Efficiency Analysis. The Measurement of Productive Efficiency and Productivity Growth*, Chapter 2, pp. 959–980. Oxford University Press, Inc. iv, 11, 29, 31

Griffin, J. and M. Steel (2004). Semiparametric Bayesian inference for stochastic frontier models. *Journal of Econometrics 123*, 121–152. 5, 10, 21

Griffin, J. and M. Steel (2007). Bayesian stochastic frontier analysis using WinBUGS. *Journal of Productivity Analysis 27*, 163–176. 8, 20, 31, 33, 39, 47, 102

Griffin, J. and M. Steel (2008). Flexible mixture modelling of stochastic frontiers. *Journal of Productivity Analysis 29*, 33–50. 10

Growitsch, T., T. Jamasb, and M. Pollitt (2009). Quality of service, efficiency and scale in network industries: An analysis of European electricity distribution. *Applied Economics 41*, 2555–2570. 88

Growitsch, T., T. Jamasb, and H. Wetzel (2012). Efficiency effects of observed and unobserved heterogeneity: Evidence from Norwegian electricity distribution networks. *Energy Economics 34*, 542–548. 83

Hadri, K. (1999). Estimation of a doubly heteroscedastic stochastic frontier cost function. *Journal of Productivity Analysis 17*, 359–363. 13

Hadri, K., C. Guermat, and J. Whittaker (2003a). Estimating farm efficiency in the presence of double heteroscedasticity using panel data. *Journal of Applied Economics 2*, 255–268. 12

Hadri, K., C. Guermat, and J. Whittaker (2003b). Estimation of technical inefficiency effects using panel data and doubly heteroscedastic stochastic production frontiers. *Empirical Economics 28*, 203–222. 12

Hartman, T. and J. Storbeck (1996). Input congestion in loan operations. *International Journal of Production Economics 46-47*, 413–421. 103

Hattori, T. (2002). Relative performance of U.S. and Japanese electricity distribution: An application of stochastic frontier analysis. *Journal of Productivity Analysis 18*, 269–284. 83, 88

Huang, H. (2004). Estimation of technical inefficiencies with heterogeneous technologies. *Journal of Productivity Analysis 21*, 277–296. 11, 19

Huang, H. and J. Liu (1994). Estimation of a non-neutral stochastic frontier production function. *Journal of Productivity Analysis 5*, 171–180. 10, 11

Huang, T. and Y. Chen (2009). A study on long-run inefficiency levels of a panel dynamic cost frontier under the framework of forward-looking rational expectations. *Journal of Banking and Finance 33*, 842–849. 84

Hughes, J., W. Lang, L. Mester, and C. Moon (1996). Efficient banking under interstate branching. *Journal of Money, Credit and Banking 28 (4)*, 1045–1071. 49

Hughes, J. and L. Mester (1993). A quality and risk-adjusted cost function for banks: Evidence on the "too-big-to-fail" doctrine. *Journal of Productivity Analysis 4 (3)*, 293–315.

Hughes, J. and L. Mester (1998). Bank capitalization and cost: Evidence of scale economies in risk management and signaling. *Review of Economics and Statistics 80 (2)*, 314–325. 49

Hughes, J. and L. Mester (2013). Who said large banks don't experience scale economies? Evidence from a risk-return-driven cost function. *Journal of Financial Intermediation 22 (4)*, 559–585. 49, 57, 61

Hughes, J., L. Mester, and C. Moon (2001). Are scale economies in banking elusive or illusive?: Evidence obtained by incorporating capital structure and risk-taking into models of bank production. *Journal of Banking and Finance 25 (12)*, 2169–2208. 48, 49

IMF (2014). Global financial stability report: Moving from liquidity -to growth- driving markets. Technical Report April, International Monetary Fund. 58

Ioannidou, M. and M. Penas (2010). Deposit insurance and bank risk-taking: Evidence from internal loan ratings. *Journal of Financial Intermediation 19*, 95–115. 53

Jamasb, T., L. Orea, and M. Pollitt (2012). Estimating the marginal cost of quality improvements: The case of the UK electricity distribution companies. *Energy Economics 34*, 1498–1506. 88

Janna, M. (2003). Banking efficiency in Colombia: A review of the literature. Financial Stability Report December, 78-85, Banco de la República's Monetary and Reserves Division. 104

Koetter, M. (2008). The stability of efficiency rankings when risk-preferences and objectives are different. *The European Journal of Finance 14 (2)*, 115–135. 49, 58

Koop, G., J. Osiewalski, and M. Steel (1997). Bayesian efficiency analysis through individual effects: Hospital cost frontiers. *Journal of Econometrics 76*, 77–106. 10, 13, 14, 37

Koop, G., M. Steel, and J. Osiewalski (1995). Posterior analysis of stochastic frontier models using Gibbs sampling. *Computational Statistics 10*, 353–373. 8, 20

Koopmans, T. (1951). *Activity analysis of production and allocation*, Chapter An analysis of production as efficient combination of activities. Cowles Commission for Research in Economics, Monograph no. 13. New York. iii

Kopsakangas-Savolainen, M. and R. Svento (2011). Observed and unobserved heterogeneity in stochastic frontier models: An application to the electricity distribution industry. *Energy Economics 33*, 304–310. 84

Kumbhakar, S. (1990). Production frontiers, panel data, and time-varying technical inefficiency. *Journal of Econometrics 46*, 201–212. 73

Kumbhakar, S., S. Ghosh, and J. McGuckin (1991). A generalized production frontier approach for estimating determinants of inefficiency in U.S. dairy farms. *Journal of Business and Economic Statistics 9*, 279–286. 11

Kumbhakar, S. and C. Lovell (2000). *Stochastic Frontier Analysis.* New York: Cambridge University Press. iii, iv, 5, 11

Kuosmanen, T. and M. Kortelainen (2012). Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. *Journal of Productivity Analysis 38*, 11–28. 5

Kwan, S. and R. Eisenbeis (1997). Bank risk, capitalization and operating efficiency. *Journal of Financial Services Research 12 (2)*, 117–131. 58

Larsen, E., I. Dyner, L. Bedoya, and C. Franco (2004). Lessons from deregulation in Colombia: successes, failures and the way ahead. *Energy Policy 32*, 1767–1780. 78, 79

Lee, Y. and P. Schmidt (1993). *The Measurement of Productive Efficiency: Techniques and Applications*, Chapter Production Frontier Model with Flexible Temporal Variation in Technical Inefficiency. Oxford University Press, Inc. 73

Lensik, R., A. Meesters, and I. Naaborg (2008). Bank efficiency and foreign ownership: Do good institutions matter? *Journal of Banking and Finance 32*, 834–844. 103

Lepetit, L., E. Nys, P. Rous, and A. Tarazi (2008). Bank income structure and risk: An empirical analysis of European banks. *Journal of Banking and Finance 32 (8)*, 1452–1467. 49

Li, Y., T. Zeng, and J. Yu (2012). Robust deviance information criterion for latent variable models. Singapore Management University, Research Collection School of Economics (Open Access). Paper 1403. 21

Lovell, CAK., 2003. The Decomposition of Malmquist Productivity Indexes. Journal of Productivity Analysis 20, 437–458. 86

Lunn, D.J., A. Thomas, N. Best, and D. Spiegelhalter (2000). WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing 10*, 325–337. 8

Malikov, E., D. Restrepo-Tobón, and S. Kumbhakar (2013). Estimation of banking technology under credit uncertainty. Working Paper, No. 13-19, Center for Research in Economics & Finance (CIEF), Universidad EAFIT. 49

Marzec, J. and J. Osiewalski (2001). Bayesian inference on technology and cost efficiency of bank branches. Working paper of Department of Econometrics, Kracow University of Economics, Poland. 47

Mastromarco, C. and U. Woitek (2012a). A Stochastic Frontier Model with Time-Varying Vectorautoregressive Inefficiency. Department of Economics, University of Salento and University of Zurich, November. 123

Mastromarco, C. and U. Woitek (2012b). A State Space Approach to Productivity and Efficiency Measurement: The Italian Economy, 1950-2003. Department of Economics, University of Salento and University of Zurich, November. 123

Meeusen, W. and J. van den Broeck (1977). Efficiency estimation from Cobb-Douglas production functions with composed errors. *International Economic Review 8*, 435–444. iii, 4, 7, 11

Melo, L. and N. Espinosa (2005). Ineficiencia en la distribución de energía eléctrica: una aplicación de las funciones de distancia estocástica. Ensayos sobre Política Económica, Banco de la República. No. 49. 79, 88

Mester, L. (1996). A study of bank efficiency taking into account risk-preferences. *Journal of Banking and Finance 20 (6)*, 1025–1045. 49

Mester, L. (2008). *Handbook of Financial Intermediation and Banking*, Chapter Optimal Industrial Structure in Banking, pp. 133–162. Elsevier. 50, 59

Moreno, D. and D. Estrada (2013). Colombian bank efficiency and the role of market structure. Temas de Estabilidad Financiera, No. 76, Banco de la República. 51, 57

Mota, R. (2003). The restructuring and privatisation of electricity distribution and supply business in Brazil: A social cost-benefit analysis. CMI working paper EP16, University of Cambridge. 80

O'Donnell, C. (2014). Technologies, markets and behaviour: Some implications for estimating efciency and productivity change. Paper presented at the Annual Conference of Australian Agricultural and Resource Economics Society, Port Macquarie, 5-7 February. iii, 4

Orea, L., 2002. Parametric Decomposition of a Generalized Malmquist Productivity Index. Journal of Productivity Analysis 18, 5–22. 86

Pantzios, CJ., Karagiannis, G., Tzouvelekasin, V., 2011. Parametric decomposition of the input-oriented Malmquist productivity index: with an application to Greek aquaculture. Journal of Productivity Analysis 36, 21–31. 87

Park, B., R. Sickles, and L. Simar (2007). Semiparametric efficient estimation of dynamic panel data models. *Journal of Econometrics 136*, 281–301. 5

Park, K. and W. Weber (2006). A note on efficiency and productivity growth in the Korean banking industry, 1992-2002. *Journal of Banking and Finance 30 (8)*, 2371–2386. 49

Pérez-Reyes, R. and B. Tovar (2009). Measuring efficiency and productivity change (PTF) in the Peruvian electricity distribution companies after reforms. *Energy Policy 37*, 2249–2261. 80

Pollitt, M. (2004). Electricity reform in Chile: Lessons for developing countries. *Journal of Network Industries 5*, 221–262. 80

Pollitt, M. (2008). Electricity reform in Argentina: Lessons for developing countries. *Energy Economics 30*, 1536–1567. 80

Pombo, C. and R. Taborda (2006). Performance and efficiency in Colombia's power distribution system: Effects of the 1994 reform. *Energy Economics 28*, 339–369. 78, 79

Poudineh, R., G. Emvalomatis, and T. Jamasb (2014). Dynamic efficiency and incentive regulation: An application to electricity distribution networks. EPRG working paper 1402, Cambridge Working Paper in Economics, University of Cambridge. 91

Radić, N., F. Fiordelisi, and C. Girardone (2012). Efficiency and risk-taking in pre-crisis investment banks. *Journal of Finance Service Research 41(1)*, 81–101. 49

Ray, S. and A. Das (2010). Distribution of cost and profit efficiency: Eevidence from Indian banking. *European Journal of Operational Research 201*, 297–307. 103

Reifschnieder, D. and R. Stevenson (1991). Systematic departures from the frontier: A framework for the analysis of firm inefficiency. *International Economic Review 32*, 715–723. 11

Santos, J. (2014). Evidence from the bond market on banks' "too big to fail" subsidy. *Economic Policy Review 20(2)*, forthcoming. 58

Sathye, M. (2003). Efficiency of banks in a developing economy: The case of India. *European Journal of Operational Research 148*, 662–671. 103

Sealey, C. and J. Lindley (1977). Inputs, outputs and a theory of production and cost at depository financial institutions. *Journal of Finance 32*, 1252–1266. 52, 104

Simar, L., C. Lovell, and P. van den Eeckaut (1994). Stochastic frontiers incorporating exogenous influences on efficiency. Discussion paper no. 9403, Institut de Statistique, Universit Catholique de Louvain. 12

Simar, L. and P. Wilson (2007). Estimation and inference in two-stage, semi parametric models of production processes. *Journal of Econometrics 136*, 31–64. 4

Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B 64*(4), 583–639. 21

Staub, R., G. da Silva e Souza, and B. Tabak (2010). Evolution of bank efficiency in Brazil: A DEA approach. *European Journal of Operational Research 202*, 204–213. 103

Stevenson, R. (1980). Likelihood functions for generalized stochastic frontier estimation. *Journal of Econometrics 13*, 57–66. 7

Tabak, B. and P. Tecles (2010). Estimating a Bayesian stochastic frontier for the Indian banking system. *International Journal of Production Economics 125*, 96–110. 47, 52, 57, 111

Tecles, P. and B. Tabak (2010). Determinants of bank efficiency: The case of Brazil. *European Journal of Operational Research 207*, 1587–1598. 54, 103, 112

Ter-Martirosyan, A. and J. Kwoka (2010). Incentive regulation, service quality, and standards in U.S. electricity distribution. *Journal of Regulatory Economics 38*, 258–273. 81

Tortosa-Ausina, E. (2002). Exploring efficiency differences over time in the Spanish banking industry. *European Journal of Operational Research 139*, 643–664. 102, 112

Tovar, B., F. Ramos-Real, and E. Fagundes de Almeida (2011). Firm size and productivity. Evidence from the electricity industry in Brazil. *Energy Policy 39*, 826–833. 88

Tsionas, E. (2002). Stochastic frontier models with random coefficients. *Journal of Applied Econometrics 17*, 127–147. 10, 11, 14, 19, 46

Tsionas, E. (2006). Inference in dynamic stochastic frontier models. *Journal of Applied Econometrics 21*, 669–676. iv, 74, 75, 77, 99, 101, 102, 103, 106, 109, 112, 113

van den Broeck, J., G. Koop, J. Osiewalski, and M. Steel (1994). Stochastic frontier models: A Bayesian perspective. *Journal of Econometrics 61*, 273–303. iii, 8, 10

Van Roy, P. (2008). Capital requirements and bank behaviour in the early 1990s: Cross-country evidence. *International Journal of Central Banking 4(3)*, 29–60.

von Hirschhausen, C., A. Cullmann, and A. Kappeler (2006). Efficiency analysis of German electricity distribution utilities - non-parametric and parametric tests. *Applied Economics 38*, 2553–2566. 88

Wang, H. (2002). Heteroscedasticity and non-monotonic efficiency effects of a stochastic frontier model. *Journal of Productivity Analysis 18*, 241–253. 4, 12, 18

Wang, H. and P. Schmidt (2002). One step and two step estimation of the effects of exogenous variables on technical efficiency levels. *Journal of Productivity Analysis 18*, 129–144. 12

West, M. and J. Harrison (1997). *Bayesian Forecasting and Dynamic Models*. Springer Series in Statistics. New York: Springer-Verlag. 101

Wheelock, D. C. and P. W. Wilson (2012). Do large banks have lower costs? New estimates of returns to scale for U.S. banks. *Journal of Money, Credit and Banking 44(1)*, 171–199. 50, 57, 61

Yu, W., T. Jamasb, and M. Pollitt (2009). Willingness to pay for quality service: an application to efficiency analysis of the UK electricity distribution utilities. *Energy Journal 30*, 1–48. 88, 98

Zago, A. and P. Dongili (2011). Credit quality and technical efficiency in banking. *Empirical Economics 40(2)*, 537–558. 49