# *Quo Vadis*, Raters? A Frontier Approach to Identify Overratings and Underratings in Sovereign Credit Risk

Hüseyin Öztürk
*Central Bank of Turkey*

Emili Tortosa-Ausina
*Universitat Jaume I, IIDL and Ivie*

Meryem Duygun
*University of Nottingham*

Mohamed Shaban
*University of Leicester*

February 24, 2020

**Abstract**

*This study analyses overratings and underratings in sovereign credit risk. The analysis uses partial frontier methods, a technique rarely applied in this literature. By combining a robust variant of the free disposal hull (FDH) estimator, we measure both underratings and overratings for individual countries and groups of countries. Particular attention is paid to comparing pre-crisis and crisis years in order to assess possible changes in the magnitude of the deviations. Our findings indicate a remarkable degree of both overratings and underratings during the analysed period (1999–2010), which partially vanish during the last years of the sample (2008–2010)—corresponding to the financial crisis—when many downgrades took place, especially in Eurozone countries. The results allow us to emphasize the importance of monitoring these deviations for sustainable financial stability. Our results also show the potential benefits of using partial frontier methods for measuring both underratings and overratings.*

**Keywords:** credit rating agency, partial frontier, rating, sovereign credit risk.

**JEL Classification Numbers:** C51, E44, G15, G24.

**Communications to**: Emili Tortosa-Ausina, Departament d'Economia, Universitat Jaume I, Campus del Riu Sec, 12071 Castelló de la Plana, Spain. Tel.: +34 964387168, fax: +34 964728591, e-mail: tortosa@uji.es

## 1. Introduction

Globalization has brought about the internationalization of financial markets in the last decades. This change has dramatically increased and differentiated investment opportunities across the world, and at the same time, has created new challenges. Measuring credit risk is now the core challenge prior to any financial investment decision in contemporary financial markets. The need for accurate pricing of financial investments has revealed a further need to appraise the credit risk metrics primarily provided by credit rating agencies (CRAs).

Sovereign credit ratings (SCRs), a small constituent of the considerable credit rating industry, are the benchmark risk indicators of financial markets. A fair assessment of sovereign credit risk certainly benefits both international lenders and sovereigns, as SCRs have a significant impact on portfolio management and fiscal policy. More seriously, however, are the unanticipated repayment problems that can cause financial crises with cross-country spillovers and contagion effects.

Potential persistent errors in SCRs could have a significant impact on the cost of funding for both sovereign and private entities. While SCRs are one of the major pricing parameters for sovereign debt, they impact the cost of funding for private institutions too. CRAs assign country risk ceilings as a function of SCRs, which in the end determine the maximum credit rating achievable for private issuers (Borensztein et al., 2013). Regardless of how sound a private entity's financial condition is, that ceiling may circumvent cheaper funds for that institution (Williams et al., 2013). This anomaly has more severe repercussions for the private institutions in low-rated countries. In addition to the cost of funding, its availability could be worsened by the SCRs of low-rated countries. The abovementioned negative effects are more visible when downgrades occur from investment grade to junk grade, often termed cliff effects (Eijffinger, 2012).

Another linkage from SCRs to financial markets is the wide array of regulations that have extensive assignments to SCRs. In particular, some regulations impose certain restrictions requiring some institutional investors to invest in investment grade instruments only. National regulatory agencies require financial institutions to measure the quality of their assets in accordance with credit ratings, including SCRs. Also, investors may want to rely on SCRs rather than undertake

costly analysis themselves to benefit from the expertise of CRAs. The prevalent use of CRA ratings is defined as "over-reliance" by international regulatory bodies; recently, the Financial Stability Board (FSB) formed a task force to work on reducing over-reliance on CRA ratings.[1] While discussions are underway and some recommendations have been proposed to increase transparency in the rating process, effective policies to reduce reliance on CRA ratings are yet to be taken. Some recommendations on this matter propose that new organizations should take an active role in the assignment of SCRs, e.g. the European Rating Agency, but such initiatives have not weakened the dominance of CRAs. Although US and EU laws acknowledge many agencies as "recognized", undeniably, Moody's, Standard and Poor's (S&P), and Fitch hold the largest stake. The rating industry is, therefore, still far from competitive.

Past downgrades in the EU have also underscored just how contagious they are. The consecutive downgrades of Greece's ratings instigated many other downgrades in the EU, including Spain, Austria, France and Italy, which afterwards jeopardized the success of austerity measures in these countries. Changes to SCRs in a single country had significant knock-on effects on the macro economy and became a self-fulfilling prophecy. The rise in the cost of funding worsened Greek debt dynamics by calling for further rate cuts. Given the intense inter-linkages between the EU countries, the troubles facing the Greek economy at the beginning of 2010 spread into other EU countries and became an issue for the whole EU. Therefore, it is not hard to argue that SCRs are multi-dimensional phenomena that have both intra- and cross-country impacts.

Due partly to the multi-dimensional roles SCRs have in international finance, there have been numerous criticisms of these agencies. Several authors have noted that credit risk measurement suffers from the weaknesses of measurement tools and analyst bias (Bar-Isaac and Shapiro, 2011; Gärtner et al., 2011; Gaillard, 2014), home bias (Vernazza and Nielsen, 2015; Fuchs and Gehring, 2017), and conflict of interest between the *raters* and the *rated* (Darbellay and Frank, 2012; Opp et al., 2013; Bolton et al., 2012). Major criticism has been made of CRAs, which were accepted as

[1] In general, credit ratings were at the top of regulatory institutions' agendas in the early days of the 2008 financial crisis. They again came under fire after fiscal problems in the Eurozone when specific attention was paid to SCRs. A large set of rules is currently in force to regulate and supervise credit ratings in the EU (see Darbellay and Frank, 2012). From a regulatory perspective, one of the most important reasons for improving prediction of credit ratings is underlined by the recommendations of global financial institutions and joint initiatives. Recently the FSB, established under the auspices of the G-20, published a proposal to reduce over-reliance on CRA ratings. The proposal simply recommends that financial institutions carry out their independent credit assessment. In recommending this, the FSB aims to reduce over-reliance on CRA ratings that were pinpointed as one of the potential reasons for financial instability, especially after the recent financial crisis.

the scapegoats of the ongoing financial crisis (Fuchs and Gehring, 2017; Opp et al., 2013; Mora, 2006). The crux of the criticism lies in credit ratings' poor predictive power against possible defaults. These concerns were mainly forwarded to corporate ratings in the early episodes of the 2008 financial crisis; however, the rise of Eurozone crisis unleashed considerable disapproval of sovereign credit ratings' failure to signal major deterioration in country fundamentals (Gaillard, 2014; Gärtner et al., 2011). Arguably, although market indicators suggested a certain degree of deterioration during the early days of the EU crisis, CRAs remained mute on the subject of market developments for a long time.

Vernazza and Nielsen (2015) show that the arguments on the weaknesses of SCRs, especially during the build-up to several crises, were not totally groundless. The authors decompose SCRs into subjective (committee based) and objective (numerical based) components and explore the success of each component in predicting sovereign defaults. Their main result is that while the objective component of SCRs is able to predict sovereign defaults, the subjective component does not seem to have this capacity. The implication of their finding is that the final rating decision might lead to significant distortions in the information content of the objective component of SCRs. This could ultimately have been more useful in signalling sovereign creditworthiness.

In response to criticisms, CRAs defend their actions on two grounds. First, they contend that they try to measure the probability of default rather than temporary market reactions. By this, they imply that they are not concerned about high frequency market indicators of a rated entity during the rating process, but specifically, they look at entities' long-term potential. Therefore, fluctuations in the markets do not concern CRAs as much as they do market players and policy-makers. Second, they point to committee judgements as the main reason behind the discrepancies between the assigned ratings and the market pricing of sovereign credit risk. CRAs argue that what seems to be a divergence in the two assessments stems from the judgements of experts, who are very familiar with the country in question. The implication of this argument is that experts can factor in those idiosyncratic country-specific dynamics that country variables are unable to capture.

This study aims to investigate the discrepancies between objective and subjective components of SCRs in a large country sample for the 1999–2010 period. The objective component of ratings corresponds to a quantitative assessment of sovereign indicators. The subjective component is

the addition to the quantitative output to reach the final rating. We argue that wider and persistent discrepancies between the two may indicate the main source of weaknesses in SCRs that provokes such criticisms. The rich data in this study will enable us to study different groupings (e.g. EU versus non-EU, developed versus developing countries), over a long time horizon (e.g. pre-crisis versus crisis years). Referring to the positive contribution of the subjective component as overrating and the negative contribution as underrating, the analysis will identify which countries, or groups of countries, showed the highest overratings or underratings.[2] Comparing pre- and crisis years will also enable us to assess the changing behaviour of CRAs when the effect of the 2008 financial crisis is taken into account.

One of the essential novelties of this paper is the frontier efficiency framework adopted to study the proposed theme. In this particular setting, ratings obtained by each country are the outputs, and the various fundamental indicators used by CRAs are the inputs. We therefore assume that countries maximize their credibility (SCRs) with minimum input usage (fundamentals). In this process, (in)efficiencies would imply that some overratings and underratings are generated. We use order-$m$ to estimate whether some *inefficiencies* might exist when CRAs assign SCRs (Cazals et al., 2002; Daouia and Gijbels, 2011). In the proposed context, using partial frontier approaches such as order-$m$ is particularly appropriate because they identify both underrated and overrated countries in each period of time. The order-$m$ frontiers allow estimation of both inefficiency (underrating) and superefficiency[3] (overrating)—and, more importantly, identify which countries are underrated and overrated. In addition, with respect to the non-robust alternatives on which they are based (data envelopment analysis, DEA, and its non-convex alternative, free disposal hull, FDH), partial frontier estimators, such as order-$m$, offer several advantages, including their relative immunity to outlying observations, the fact that they are less affected by the curse of dimensionality, and better properties in general.

We contribute to the literature in at least two ways. First, and foremost, we propose an efficiency-based approach to identify discrepancies between objective and subjective components of SCRs. Studies in this area of research have adopted a number of statistical methods, but to our knowledge, SCRs have not previously been classified as underratings and overratings through

[2]Recent contributions using this terminology are, among others, Kunovac and Ravnik (2017).
[3]See Andersen and Petersen (1993).

efficiency scores. Second, the source of weaknesses or potential biases in SCRs is attracting wider interest, although it has received limited attention in the literature. A number of papers have examined this theme in a small group of countries and within a limited context. We broaden the discussions in the present paper by using a representative dataset. The dataset enables us to investigate SCRs with a large group of countries through pre-crisis and crisis years. Since our analysis is cross-sectional, the representative dataset is immune to the requirement that the pre- and crisis periods need to be well-balanced in panel data.

Our findings suggest that exploring overratings and underratings in an efficiency framework is indeed consistent. Although partial frontiers require a trimming parameter to be specified, it is always possible to detect the potentially underrated countries and, more interestingly, underratings that do not correct over the years. Our results also reveal differences among country groups for both underratings and overratings-specifically, developing countries receive lower ratings than their developed peers with respect to their fundamentals. We also find that, on average, the number of overratings and underratings is higher for non-OECD and non-Eurozone countries, when countries are classified as OECD vs. non-OECD countries or Eurozone vs. non-Eurozone countries. Interestingly, the results show that the 2007–2008 financial crisis corrects both overratings and underratings to some extent, due possibly to certain rating changes resulting from the pressure of intensified criticism of SCRs.

The study is structured as follows. Section 2 provides a literature review on SCRs and the motivation behind this study. Section 3 describes the methodology based on a nonparametric partial frontier approach. Section 4 introduces SCRs and several country indicators, defining the main variables considered in the study. Section 5 presents and discusses the results, and Section 6 outlines some conclusions and policy implications.

## 2. Literature review and motivation

The research on SCRs is increasing through a wide array of discussions. While the early literature mainly focused on the determinants of SCRs, recent work covers a variety of themes, such as the relation between market indicators and SCRs, the procyclicality of SCRs and its implications, and the impact of ratings on fiscal policy. The SCR literatures can be broadly

grouped into two areas. The first area deals mainly with the financial market reactions to SCR changes such as government bond markets, swap markets, and interest rate markets etc. (see e.g. Alsakka and ap Gwilym, 2013; Treepongkaruna and Wu, 2012; Candelon et al., 2011). Most of these studies investigate the pattern of causality between market reactions and sovereign credit downgrades/upgrades, in other words, which one leads the other. Other topics include concerns about the information content of SCRs and their association with bond spreads and intensified default risk after the global financial crisis (see e.g. Binici et al., 2018; Aizenman et al., 2013). The second area began with the study by Cantor (1995), and investigates the determinants of SCRs. Many of these investigations conclude that SCRs can be largely explained by the level of GDP per capita, real GDP growth, external debt, the public debt level, and the government budget balance (see Erdem and Varli, 2014; Gültekin-Karakaş et al., 2011; Sy, 2009; Hill, 2004). Other contributions within this category examine the relationship between rating outlook and rating changes (see e.g. Alsakka and ap Gwilym, 2012, 2010). After the 2008 financial crisis, some other attempts falling into this category explored further evidence underlying the downgrades of many Eurozone countries. For instance, Afonso et al. (2012) investigated the relationship between fiscal imbalances and credit rating downgrades, concluding that fiscal imbalances do actually have a negative impact on SCRs, but in different ways for each country.

The topic of potential weaknesses in SCRs, however, did not attract visible interest in the literature until the 2008 financial crisis. Although a few papers identified some biases in SCRs (Fuchs and Gehring, 2017; Ferri et al., 1999), the common wisdom before the 2008 crisis suggested that there were not systematic errors in SCRs. Papers in this line tend to find that much of the variation in SCRs can be explained by a number of variables. Two notable exceptions are those of Vernazza and Nielsen (2015) and Amstad and Packer (2015), who discussed the potential weaknesses of SCRs. Vernazza and Nielsen (2015) split ratings into two parts and focus on the information content of the component assigned by the committee members. The authors find that the additional input from committee members contributes little to predicting sovereign defaults, whereas the component implied by country fundamentals makes successful default predictions in a horizon of one year or more. Amstad and Packer (2015) also acknowledge some systematic bias in SCRs in favour of advanced countries. This is also underlined by Amstad and Packer (2015), who contend that to make ratings more transparent after 2008, SCRs became more

quantitative-oriented and suppress the influence of committee decisions. In a pool of advanced and developing countries, Amstad and Packer (2015) find a notch difference between estimated ratings and actual ratings that prejudices developing countries. The authors, however, argue that this difference should be accepted as negligible.

While we contribute to the newly flourishing discussions on the potential sources of weaknesses in SCRs, another, maybe more important, contribution of this study concerns its analysis. Several statistical methods have been employed in previous studies to predict SCRs. However, as Wang et al. (2011) argue, multivariate normality assumptions are frequently violated in statistical models, and these models do not guarantee normality assumptions for every independent variable. Therefore, the accuracy of predictions is frequently low. The literature generally uses parametric models to estimate SCR changes, such as linear discriminant analysis (Frank and Cline, 1971; Grinols, 1976), principal component analysis (Mellios and Paget-Blanc, 2006), linear regressions (Cantor, 1995; Ferri et al., 1999; Erdem and Varli, 2014), and ordered response models (Afonso et al., 2012, 2007; Alsakka and ap Gwilym, 2010, among others). A small niche in the literature investigates credit rating with artificial intelligence (AI) models (Wang et al., 2011; Huang et al., 2004; Maher and Sen, 1997). Bennell et al.'s (2006) study was one of the first to introduce AI in the estimation of SCRs. According to their findings, AI models estimate SCRs more accurately than other statistical approaches.

The literature on the determinants of SCRs mainly uses panel techniques in the estimation of ratings. Ordered response models are widely preferred since SCRs are discrete variables and are ordered in terms of probability of default. In this framework, the conditional probabilities for all the rating categories are estimated over an unobserved latent variable, the highest probability of which is then selected as the estimated rating. However, as with other panel techniques, ordered response models assume a functional form in estimating SCRs that is hard to defend in the SCR context. The impact of various factors does not follow a linear pattern. The impact of inflation, for instance, can be non-linear, suggesting higher impact for developing countries. In addition, the impact of deteriorating fiscal performance in tandem with large current account deficits may trigger higher downgrades. Some nonparametric methods in the literature address the functional form constraint. Emerging techniques in this field are decision trees—support vector machines, Bayesian learners, etc.—which create a tree structure where the nodes end at final

rating categories (Ozturk et al., 2016b,a). The tree structure, by its nature, represents non-linear relations. Although the decision trees are highly successful in predicting ratings and revealing the nonlinear paths that rating decisions follow, these computational techniques do not allow us to quantify both overratings and underratings. This drawback is covered by the technique proposed in the present paper.

## 3. Methodology

In this paper, we combine the SCRs literature with the literature on activity analysis. One of the essential novelties in this paper is our assessment of both overratings and underratings by exploiting a non-parametric technique. To identify overratings and underratings, we propose a model in an efficiency setting where countries maximize their ratings according to their macroeconomic, financial and fiscal indicators. When evaluating an SCR, the aspects that investors minimize (such as debt stock, current account deficit, low GDP growth etc.) will be considered as inputs and the rating will be the output. By designating this setup, the yields or efficiency scores will summarize a country's performance. The performance of a country in terms of attaining higher or lower scores will implicitly indicate the contribution of the subjective component of each SCR. Both high performance (high ratings versus high risk) and low performance (low ratings versus low risk) will be considered as the deviation from the objective component of SCRs which primarily relies on quantitative analysis.

We simply propose that if a country in a certain rating period (year) is rewarded by a higher credit rating than its fundamentals would suggest, the country in that year will tend to receive a higher efficiency score; the reverse also holds true. By using variants of free disposal hull (FDH) methods, we identify superefficient SCRs. To this end, we first estimate a frontier that matches country fundamentals with efficient SCRs. Ratings in the upper boundary are the superefficient ratings, whereas those in the lower boundary are inefficient ratings. Then, we explore whether certain groups of countries, e.g. EU countries, emerging countries, advanced countries, tend to be over/underrated.

Our methods are based on the set of activity analysis techniques initially devised by Georgescu-Roegen (1951). His ideas were refined in later stages in order to model the productive efficiency

of decision making units (DMUs), which can vary widely. This type of unit could be restricted to countries, as in our case, but can also refer to a broad range of organizations such as banks and other financial institutions, municipalities, hospitals, etc. As a result, measures of performance via efficiency scores have become widespread for operators in business, government, public transportation, infrastructure, energy production and other sectors.

A wide variety of frontier methods can be used to *measure* efficiency. Murillo-Zamorano (2004) provides an excellent review of these methods for the case of economic efficiency. There are two main groups of methods to estimate efficiency scores, namely, stochastic frontier analysis, SFA (Aigner et al., 1977; Meeusen and van den Broeck, 1977), and data envelopment analysis, DEA (Charnes et al., 1978). There has been a long standing division between SFA and DEA. Both methods have advantages and disadvantages—the 'historically' perceived merit of SFA is that the estimator is stochastic, whereas in the case of DEA the estimator is nonparametric (Badunenko et al., 2012). Most comparative studies such as, for instance, Ferrier and Lovell (1990) or Badunenko et al. (2012) conclude that different methods can be preferable under different circumstances.

Although progress has been made both in the parametric (SFA) and nonparametric (DEA) fields, advances have been unequal—especially in terms of applications. According to Badunenko et al. (2012), recent research has seen a relaxation of functional forms in the parametric field (SFA) and the introduction of asymptotics in the nonparametric field (DEA). In asymptotic terms, some of the newest estimators based on linear programming perform better than DEA and, in addition, they overcome some of its disadvantages, including the 'curse of dimensionality' (low number of DMUs relative to number of input-output variables) or the influential role of outliers. The curse of dimensionality results from the fact that, as a given set of $n$ observations are projected in an increasing number of orthogonal directions, the Euclidean distance between the observations should necessarily increase. As for the role of outliers, envelopment estimators such as DEA are very sensitive to outliers and extreme values, which may disproportionately (and misleadingly) influence the evaluation of the performance of other DMUs.[4]

A series of proposals (Cazals et al., 2002; Daraio and Simar, 2005; Aragon et al., 2005; Daouia and Simar, 2007) have put forward two families of robust estimators—i.e., estimators which

[4] As indicated by Simar and Wilson (2008), this drawback is also present in parametric frontier estimators when deterministic frontier models are considered.

are much less sensitive to extreme observations: (i) order-$m$ frontiers (where $m$ can be viewed as a trimming parameter); and (ii) order-$\alpha$ quantile frontiers (analogous to traditional quantile functions but adapted to the frontier problem). These are 'partial' frontier estimators, as opposed to the traditional idea of a 'full' frontier that envelops all the data, given that the goal is not to estimate the absolute lowest (uppermost) technically achievable level of input (output) for a given level of output (input), but rather to estimate something 'close' to these quantities. In addition, apart from not suffering from the curse of dimensionality and being much more robust than either DEA or its non-convex variant (free disposal hull, FDH), both order-$m$ and order-$\alpha$ estimators have generally better properties, since they also allow us to achieve the $\sqrt{n}$ rate of convergence with asymptotic normality.

Because of these advantages, partial frontier methods are particularly well suited to our specific setting, where the number of dimensions in which a country can be evaluated (i.e., the number of inputs and outputs) could be high. Therefore, whereas the resulting DEA or FDH estimators could be affected by the curse of dimensionality, the order-$m$ or order-$\alpha$ estimators are less likely to be.

Following Daraio and Simar (2007),[5] order-$m$ estimators are based on FDH estimators. Supposing there are $m$ decision making units (i.e. credit rating agencies) using at most input level $x$, we define the set:

$$\Psi(x) = \{(x', y') \in \mathbb{R}_+^{N+M} | x' \leq x, Y_i \leq y'\} \tag{1}$$

where $i = 1, \ldots, m$, $Y_i$ are $m$ iid random variables drawn from the conditional $M$-variate distribution $F_Y(\cdot|x)$, and $N$ is the number of inputs and $M$ the number of outputs.

In this context, the output-oriented efficiency score (i.e., our indicator of underrating) can be defined relative to the $\Psi_m(x)$ set (which is random, since it depends on random variables) as:

$$\tilde{\lambda}(x, y) = \sup\{\lambda | (x, \lambda y) \in \Psi(x)\} = \max_{i=1,\ldots,m} \left\{ \min_{j=1,\ldots,M} \left( \frac{Y_i^j}{y^i} \right) \right\} \tag{2}$$

[5]For applications in the field of finance, see for instance, Matallín-Sáez et al. (2014), Abdelsalam et al. (2014) or, more recently, Matallín-Sáez et al. (2019), among others. For a more general view on the relevance of frontier efficiency methods applied to finance, see Eling and Schuhmacher (2007).

For each combination of inputs and outputs, $(x, y) \in \mathbb{R}_{+}^{N+M}$, we will define the output-oriented order-$m$ efficiency score as an expectation for all $x$ in the interior of the support of $X$ (assuming that the expectation exists) as:

$$\lambda_m(x, y) = E(\tilde{\lambda}_m(x, y)|X \leq x) \tag{3}$$

Therefore, in contrast to either FDH or its convex version (DEA), the idea of the order-$m$ is to compare each observation with *part* of the frontier instead of the full frontier—which is why we refer to order-$m$ as a partial frontier.

Interestingly, since the country under analysis is not (necessarily) included in the order-$m$ sample (and there will not necessarily be any other countries that dominate the country analysed in the output), efficiencies can be either higher or lower than one. Specifically, output-oriented efficiencies based on Shephard distance functions (reciprocal to the Farrell distance functions) are either equal to or lower than unity under FDH (or DEA). However, under order-$m$ some outlying observations (countries) can reach efficiency levels higher than one; the literature usually refers to these as superefficient units (Andersen and Petersen, 1993). We will consider those countries classified as inefficient (i.e., with scores lower than one) to be underrated. In contrast, we will refer to the superefficient units (with values higher than one) as overrated.[6]

## 4. Data and variables

SCRs are assigned through a series of qualitative and quantitative analyses. As an initial assessment, agency analysts collect a set of indicators that demonstrate macroeconomic and financial strength and institutional quality. In order to increase the transparency of the rating process, CRAs publicly announce the details of and motivations for using these indicators. Analysts' assessments of country creditworthiness entail assigning different weights to each indicator in the rating process, although the weights attached to these indicators are not publicly available. The quantitative assessments therefore constitute the objective component of SCRs. The final rating is decided by a committee based on quantitative assessment and the view of each mem-

[6] The interested reader can consult the recent paper by Daraio et al. (2019), in which different software options for evaluating efficiency and productivity based on frontier methods are reviewed.

ber. The change made on quantitative assessments during committee meetings is therefore the subjective component of SCRs.

In our study, we build a SCR database with foreign currency ratings[7] of sovereigns provided by a major agency. The rating of a particular year is the rating that was attributed on the last day of that year. In our database there are two main blocks of data. Both SCRs and macroeconomic and financial indicators used in the analysis are obtained from the agency. World Governance Indicators showing the quality of institutions are from the World Bank. These indicators are monitored by CRAs as a measure showing sovereigns' willingness to repay.

There is an agreed consistency in the classification of sovereign creditworthiness. CRA methodologies define 20 possible credit ratings for a country: Aaa, Aa1, Aa2, Aa3, A1, A2, A3, Baa1, Baa2, Baa3, Ba1, Ba2, Ba3, B1, B2, B3, Caa1, Caa2, Caa3, Ca. In this classification, Aaa is the highest rating, reflecting the highest possibility of repayment whereas Ca is the lowest rating and denotes the lowest creditworthiness. The raw data had information on 106 countries for the 1999–2010 period, although the panel was incomplete, ranging from a minimum of 77 observations in 1999 to a maximum of 91 for 2008, 2009 and 2010. Because of the low number of observations for the Caa2, Caa3, and Ca ratings, countries that have received these ratings at least once over the sample period are not included in the analysis. We will refer to the SCRs as $y_1$ in our analysis, and it will be treated as the output variable.

In order to analyse the tendencies of SCRs in certain country groups, the data is divided into two sub-samples of developed and developing countries. This classification is based on the World Bank definition, according to which the high income OECD and high income non-OECD countries are classified in the "developed country" group, while countries in the low income, lower middle income, and upper middle income categories are classified in the "developing country" group. Grouping the countries according to this definition also allows us to discriminate the efficiency with respect to countries' development levels.[8]

[7]The rating of a country is its foreign currency rating which shows the likelihood of repayment on foreign debt. CRAs also assign domestic currency ratings which are equivalent ratings but measure the likelihood of repayment on domestic debt. The use of domestic currency ratings is not suitable as central banks are able to print money and may support domestic currency debt repayments without strengthening economic fundamentals but simply via monetization.

[8] The classification of countries as "developed" and "developing" is a complicated one, since this status is unobservable and subject to judgements. In a rating study, however, it is highly important to classify countries based on their income level, as this is (the most, as argued by the big three) important factor in the likelihood of repayment. The interpretation is straightforward—the higher the income level, the higher the probability of repayment. Based

Table 1 reports the list of countries in the analysis, and Table 2 presents the percentage distribution of SCRs by the variation in income level. A clear pattern suggests that as the income level increases countries are more likely to obtain higher ratings. When the countries are grouped by their income level into developed and developing countries, there is a clear difference in SCRs in favour of developed countries. The developed countries obtain ratings higher than Ba1, whereas A1 is the highest rating obtained by a developing country. Moreover, Aaa is the most commonly assigned credit rating by 20.08% which means that most of the developed countries are ranked in the most creditworthy category.

Tables 3 and 4 present the variables used as the inputs in our analysis. Although CRAs use a vast dataset, these variables can be taken to represent the performance of a country. Below we briefly summarize the motivation behind the variable selection.

**Ratio of current account balance to GDP (*balancegdp*, $x_1$):** the current account (when in deficit) gives a rough indication of how much net import of capital is needed for a country to close the gap between domestic saving and investment. Large and persistent current-account deficits can lead to a distortion of external debt structure, if the deficits cannot be financed by inflows of direct investment or equity positions in local companies. However, rapidly-growing countries with high investment rates can sustain large deficits for many years if the investments are conducive to a growing export capacity that can create the inflow of foreign earnings needed to service a growing debt. Since the nominal current account will vary with the scale of a country's size and openness to trade, we divide it by GDP to allow for cross-country comparisons (Erdem and Varli, 2014; Mora, 2006; Bennell et al., 2006; Gültekin-Karakaş et al., 2011; Afonso et al., 2007).

**Ratio of general government financial balance to GDP ($financialbalancegdp$, $x_2$):** The fiscal balances and debt stocks of the various levels of government are among the most important indicators examined by sovereign risk analysts. The ability of governments to extract revenues from the population of tax payers and users of services, the elasticity of revenue with

on this reasoning, the World Bank classification is quite promising here because the mandate of the institution is development, poverty, income inequality etc. To be on the safe side and to check whether our analyses are robust to the definition of "development", we also classified countries based on the IMF definition. The analysis based on this new definition did not change the results. To save space we do not report the results here, but they are available upon request.

respect to the growth or decline of national income, and the rigidity of the composition of government expenditures are key factors that determine whether central and local governments will be able to make timely payments of interest and principal on outstanding debt. We proxy fiscal balances with three indicators: general government financial balance to GDP, general government primary balance to GDP, and general government debt to GDP (Mora, 2006; Bennell et al., 2006; Gültekin-Karakaş et al., 2011; Afonso et al., 2007). The ratio of general government financial balance to GDP indicates governments' deficit or surplus in GDP. Higher government deficits can create repayment problems that can be solved by inflationary money creation. Inflation on the other hand can distort the dynamics of growth.

**GDP per capita ($gdppc$, $x_3$):** GDP is the standard international measure of the size of an economy. While frequently criticized for understating output by leaving out or underestimating the accumulation of intangible assets (knowledge, organizational innovation, improved product quality, etc.) or for overstating it by ignoring resource depletion and environmental degradation, GDP remains the only internationally comparable standard. Nevertheless, GDP only gives an aggregate level of the economy. We use GDP per capita to show the relative wealth possessed by the average individual within a given country (Erdem and Varli, 2014; Mora, 2006; Bennell et al., 2006; Gültekin-Karakaş et al., 2011; Afonso et al., 2007).

**Inflation, ($inflation$, $x_4$):** inflation is an important indicator of excess demand pressure or of structural distortions in the labour and product markets. Under extreme conditions of monetary instability (in which, for example, central banks create money in order to finance government deficits) inflation can accelerate to "hyperinflationary" levels that undermine normal productive activity. It is well known that an inflationary environment in a national economy leads to high uncertainty where production decisions cannot easily be taken (Erdem and Varli, 2014; Mora, 2006; Gültekin-Karakaş et al., 2011; Afonso et al., 2007).

**Official foreign exchange reserves ($foreignexcreserve$, $x_5$):** foreign exchange reserves held by a country are the first line of defence against withdrawal of foreign credit. Hence foreign exchange reserves act as a cushion especially for sudden outflows. Since the ratings we are studying are those assigned to foreign exchange debts, ample reserves give the country

further flexibility (Erdem and Varli, 2014; Gültekin-Karakaş et al., 2011; Afonso et al., 2007).

**Government effectiveness (*governmenteffectiveness*, $x_6$):** this indicator is one of six measures of institutional quality compiled by the World Bank. The index of government effectiveness combines responses on the quality of public services and the bureaucracy that provides them, the competence and political independence of civil servants, and the credibility of the government's commitment to its policies. Apart from the sovereigns' capacity to pay, CRAs attach importance to their willingness to pay, which can be broadly proxied by the index of government effectiveness (Erdem and Varli, 2014; Gültekin-Karakaş et al., 2011; Afonso et al., 2007).

**Ratio of general government primary balance to GDP (*primarybalancegdp*, $x_7$):** the primary balance figures exclude interest expenditures. Positive general government primary balance figures show how governments progress in narrowing general government deficits.

**Nominal exports of goods and services, % change (*exportsprcnt*, $x_8$):** the percentage change of the nominal exports of goods and services shows the performance of a country by degree to which the country supplements its domestic saving with foreign export revenues in financing capital investment (Gültekin-Karakaş et al., 2011).

**Nominal GDP percentage change (*gdpprcnt*, $x_9$):** the annual percentage change in nominal GDP (in local currency) is important because a decline in nominal GDP that is a combination of weak or negative growth and falling prices may be a distress signal in the economy that results in a rating downgrade. In such circumstances, consumers and businesses may postpone purchases, expecting goods to be cheaper in the future, and the real burden of household and corporate debt will increase (Mora, 2006; Bennell et al., 2006; Gültekin-Karakaş et al., 2011; Afonso et al., 2007).

**Ratio of gross investment to GDP (*investgdpratio*, $x_{10}$):** investments that add to the country's capital stock are a vital contributor to the process of economic growth. Countries with a sustained high investment rate, especially in productive assets in the business sector and in infrastructure, will tend to grow faster over the long term (Gültekin-Karakaş et al., 2011).

**Ratio of domestic savings to GDP (*savinggdpratio*, $x_{11}$):** the real investment undertaken within a country is necessarily equal to the sum of the domestic saving generated within its borders plus the use of foreign saving. If a country cannot generate a high enough saving flow out of the incomes of the domestic population in order to accelerate growth, it may face balance of payment constraints.

**Ratio of general government debt to GDP (*debtgdp*, $x_{12}$):** general government debt to GDP is a broad indicator of a government's total debt stock. High level of debt stock becomes a severe threat to government financing when government revenues are relatively low. If debt is barely rolled over, the risk of default increases.

Previous studies have been very heterogeneous in the relevant factors to be considered in the analysis—which in our case are the inputs. Our empirical strategy consisted of selecting variables that could be deemed "fundamental", since they are consistently used in previous models, and then sequentially introducing other variables with less generalized usage.

We therefore consider an initial model (Model 1, or the "restricted model") in which the relevant factors (inputs) are the ratio of current account balance to GDP ($x_1$), the ratio of general government financial balance to GDP ($x_2$), GDP per capita ($x_3$), inflation ($x_4$), official foreign exchange reserves ($x_5$) and government effectiveness ($x_6$). The rest of the variables are introduced sequentially to constitute a new model (which we refer to as the "unrestricted" model). We then calculate efficiencies using the methods proposed in Section 3 and test whether or not the efficiencies generated by each model are statistically significant.

We use the Li (1996) test to identify differences between the restricted and unrestricted models. Based on kernel smoothing, it tests the null hypothesis that the densities corresponding to the efficiencies generated by each model are equal ($f(\text{restricted model}) = g(\text{unrestricted model})$). For previous applications of these models see, for instance, Thieme et al. (2013). Results, which are provided in Table 5,[9] indicate that efficiencies (overratings and underratings) only differ statistically when introducing variables $x_7$, $x_8$ and $x_9$ (ratio of general government primary balance to GDP, nominal exports of goods and services % change, nominal GDP % change). For the rest of the variables ($x_{10}$, $x_{11}$, $x_{12}$, i.e., ratio of gross investment to GDP, ratio of domestic savings

[9] For the definition of the $T$-statistic see Li (1996).

to GDP, and ratio of general government debt to GDP) the differences among models were not significant and, therefore, were not included in the model.

## 5. Results

### 5.1. General tendencies

Having computed the efficiency scores for each country during the sample period, we first document the summary statistics for the efficiency scores obtained. We then introduce an additional exercise to show how robust the results are to different trimming parameters. We finally identify overratings and underratings, to discuss potential skewness across country groups and the time horizon. Results are reported in Tables 6–10. The first of these tables (Table 6) reports summary statistics (mean, interquartile range, median and standard deviation) for the efficiency scores yielded by the order-$m$ estimators. The last column reports the total number of overrated and underrated countries according to our methods. Results are split into four panels, three of which report information for the different trimming parameters considered—i.e., the selected value for $m$. The fourth panel reports a summary of efficiency scores, each row representing the summary statistics corresponding to the sum of underrated and overrated countries for each $m$ parameter, where the overratings have been inverted for easier comparison with the underratings.[10]

On average, the amount of underrating ranges from 0.7837 (for $m_{\alpha=.90}$) to 0.7372 (for $m_{\alpha=.99}$).[11] Recall that these values represent efficiencies and, therefore, the lower the values, the higher the rating inefficiencies—i.e., the magnitude of the underrating. This would imply that, for the entire sample, ratings could be improved by more than 20%. Since this is an average, for some particular countries underrating is actually quite high, because the standard deviation is also relatively high, ranging between 0.1565 (for $m_{\alpha=.99}$) to 0.1636 (for $m_{\alpha=.90}$). Although one may think these average values are driven by outliers, it is not the case because the median also

[10] Since we adopt an output-oriented approach and efficiency is measured in terms of Shephard (1970) distance functions, inefficient units are those with values lower than 1.

[11] We chose the three values for the trimming parameter ($m$) based on the proposals by Daouia and Gijbels (2011), who consider that order-$\alpha$ and order-$m$ estimators are closely related when $\alpha = \alpha(m) = (1/2)^{1/m}$. Given the general recommendation by Daraio and Simar (2007) to use trimming parameters for order-$\alpha$ equivalent to those generally used in regression analysis (i.e., the usual significance levels), we selected $\alpha = 0.90$, $\alpha = 0.95$ and $\alpha = 0.99$, and the $m$ values are those obtained by substituting in Daouia and Gijbels's formula.

reveals high inefficiencies, and their values are relatively close to those of the mean (they range from 0.7500 to 0.8318). The number of underrated countries is also relatively high (from 72 to 81) compared with the size of the sample (1,023 country-year pairs). These results certainly suggest that overratings and underratings are persistent, since these ratings belong to certain countries. Had the overratings or underratings in the sample varied over time and countries, we would have argued that this was because of measurement errors but not a potential bias. However, the analysis at this stage signals that there are certain countries which tend to be over or underrated persistently. [12]

One of the main advantages of using partial frontier techniques (such as order-$m$) is their ability to identify not only inefficiency but also *superefficiency*. In our particular context, the superefficient units would be those *overrated* countries, whose efficiencies lie above unity. In this case, the amount of overrating is also high, although this partly depends on the choice of trimming parameter, which is particularly high for lower values of $m$ ($m_{\alpha=.90}$). The average values range from 1.0039 (for $m_{\alpha=.90}$) to 1.0580 (for $m_{\alpha=.99}$) and, similarly to the underrating case, these values are not driven by outliers due to the closeness between the values for the mean and the median.

The effect of the trimming parameter is reflected in the varying number of overrated countries (the higher the $m$ value, the lower the number of superefficient units, or outliers) and, therefore, it could be deemed as a pitfall of this technique. However, in our particular setting we consider this might actually be an advantage, since we are obtaining a full ranking not only of overrated but also of *potentially* overrated countries and, therefore, it would be possible to identify those countries whose ratings would have to be corrected in the event of shock—because they were overrated. This is of particular importance for policymakers and especially for the agencies. The results indicate that CRAs assign higher ratings to superefficient units than what their credentials imply. The results also suggest that superefficient units need special scrutiny, above all during

[12] It is worth noting here that we consider our method as an alternative, or complement, to methods used by CRAs, and that it is still subject to several improvements within the context of nonparametric frontier estimation. This implies that it can be tuned to deal with some challenging scenarios such as that corresponding to the 2007/08 crisis. In addition, a key advantage of our approach is that it provides us with a *ranking*, or scale, for underratings/overratings. Therefore, should the financial scenario change, we would have an additional measure giving information as to which countries should be monitored first in terms of both underratings and, particularly, overratings. This implies that we can construct different scenarios according to how "prudent" we want to be, but the scenarios finally materialized, the rankings would remain essentially unchanged.

periods of turmoil. This would also be desirable since what is expected from CRAs is simply early warning of a possible credit event.

### 5.2. Results for different countries and temporal contexts

To augment our results we base our analysis on different country groups and time splits. The evidence that CRAs tend to overrate home countries and closely related nations, motivated us to split our country group into developed versus developing countries. This is a viable strategy since mainly emerging markets are found to be at a certain disadvantage when receiving higher ratings (Vernazza and Nielsen, 2015; Amstad and Packer, 2015; Gültekin-Karakaş et al., 2011). Based on empirical evidence, during the 2008 global financial crisis and the ongoing Eurozone crisis, many countries have faced frequent downgrades. Interestingly, the developed countries have been downgraded the most. We consider the possibility that this is the end result of possible upward bias towards developed countries that were downgraded when advanced economies were harshly criticized. In our analysis, we test whether the techniques proposed in this study can capture this bias via overratings, specifically for developed countries. We also investigate the pre- and crisis periods to observe whether the 2008 global crisis had a disciplinary effect on CRAs.

Tables 7, 8 and 9 report results for different groups of countries, depending on the country's level of development, OECD membership, or whether it has adopted the euro. Table 10 reports results for pre-crisis (1999–2007) and crisis years (2008–2010).

Table 7 reports results for developing and developed countries based on the classification of countries shown in Table 1. Results for the two groups of countries might differ because, as witnessed in the recent crises, developed countries were overrated.[13] On average, there are some remarkable differences which are robust to the choice of parameter between the two groups of countries.[14] The differences are particularly large for underrated countries, especially for lower values of $m$. In the case of $m_{\alpha=.90}$ the average gap between developed and developing countries

[13] As we shall see at the end of this section, we apply the Li (1996) test to ascertain whether or not the differences between different classifications are statistically significant.

[14] We admit that in order to evaluate more precisely the differences between these groups of countries, a specific analysis would be more appropriate. In addition, we also acknowledge that the "developed vs. developing" classification of countries is crude, but given it is an external classification reported by a prestigious international institution, we considered our analysis could shed some light on the (likely) disparate trends for the two categories.

is 0.1437 (resulting from $0.8336 - 0.6899$), but in all cases it is in the vicinity of 0.1 or above. Although in the case of the median these gaps are lower, they always exist and are favourable to developed countries. While this might suggest that there is a tendency to underrate developing countries, a more accurate interpretation would be that, regardless of the $m$ parameter considered, these countries tend to receive both overratings and underratings. This reasonably indicates that CRAs are less certain when rating developing countries, potentially because they are more volatile economies. Therefore, a relatively higher number of overratings and underratings for these countries might be related to over-cautious CRA criteria for deteriorating/improving their fundamentals. Therefore, although CRAs tend to assume relative stability in a longer period of time for developing countries, rapid changes are punished more harshly.

When comparing results for OECD and non-OECD countries (Table 8), results are very similar to those in Table 7, which might have been expected. However, there are some particularities for the comparison based on the euro area criterion. With the exception of $m_{\alpha=.99}$, the number of overrated and underrated countries is relatively similar (especially for $m_{\alpha=.99}$). In addition, although, on average, underrating is lower in the euro area (regardless of the $m$ parameter considered), the median is actually higher (i.e., lower underrating) for the non-euro countries. In contrast, the amount of overrating in the euro area is much lower compared with non-euro countries (regardless of the summary statistic considered: either the mean or the median), and this result is robust to the choice of $m$. Actually, for $m_{\alpha=.99}$ we found no overrated countries.

We also explore whether the crisis might have played a role when assessing SCRs. While massive downgrades in developed countries might reflect a degree of deterioration in fiscal and economic dynamics, as Amstad and Packer (2015) argue, CRAs also changed their rating methodologies, as they are now based more on quantitative inputs. The evolution to a more quantitative-oriented rating could have changed the intensity of overratings and underratings in our sample. This will be a good exercise to test whether CRAs did not properly recognize the signal from country fundamentals before the 2008 financial crisis (Vernazza and Nielsen, 2015). To this end, in Table 10 we compare results during pre-crisis (1999–2007) and crisis (2008–2010) years. On average, the magnitude of both overratings and underratings was higher during pre-crisis years (i.e. both indicators were further from 1), and this result is robust to the value of the trimming parameter. These results could be partly expected because of the frequency (or

the intensity) of the credit assessments, which increased during the crisis.[15] This is in line with the interpretations that CRAs overruled the signal coming from country fundamentals during pre-crisis years (Vernazza and Nielsen, 2015). The crisis years seem to be when ratings were significantly revised, regardless of whether or not they correspond to developed countries. The rating revisions suggest that the subjective component of SCRs is relatively less influential in the final rating decision, which also supports the forceful result that CRAs become more data-reliant during the crisis years. Overall, results for pre- versus crisis time splits are not driven by outlying observations, since the tendencies for the median coincide. This would confirm the virtues of the methods we use to assess under- or overrating, since we are actually quantifying that *ex post feeling* among practitioners, academics and policy-makers that during the pre-crisis years the ratings were not as accurate as they should have been. Therefore by using the definition of superefficiency, nonparametric methods can be applied to detect the distressed countries that need downgrades. This early action could align and allocate risk *ex ante* by reducing the damage of financial mayhem. Specific attention should also be given to inefficient units. Inefficiency in the results indicates that inefficient units deserve upgrades, but for some reason, the CRAs delay taking action. This may be plausible because of the asymmetric information stemming from CRAs' expertise in each country. However, inefficiencies should also be taken into account because underrated units can experience repayment problems due to low ratings.

Following a process akin to that in Section 3 for choosing our model, we consider the Li (1996) test to ascertain whether or not the differences for groups of countries and groups of years are significant. Therefore, we test the null hypothesis that the densities between two particular groups of countries are statistically different—i.e., we do not test whether results differ statistically for a particular statistic (mean, median) but for the entire distributions of overratings.

Results are reported in Table 11. They show strongly significant differences between comparison of euro area and non-euro area countries, and of pre-crisis and crisis years. However, this is not the case for the comparisons based on level of development. When the OECD membership criterion is taken into account, differences are only significant at the 5% significance level; for the level of development comparison, these densities do not differ statistically. For convenience, we also report results for the Kruskal-Wallis test, in order to ascertain whether the medians differ

[15] We are grateful to an anonymous reviewer for this comment.

statistically, and results corroborate those obtained for the Li (1996) test.[16]

As indicated in Section 4, one of the inputs in our models is government effectiveness (*governmenteffectiveness*), an institutional quality variable compiled by the World Bank (from the Worldwide Governance Indicators database, WGIs). The choice of using government effectiveness depends on CRAs' choice as this indicator extensively measures the quality of governmental institutions, which is closely related to willingness to repay debt. Although they are expected to be interrelated in some way, alternatively we use the remaining WGIs, in order to capture potential future shocks to different dimensions of institutional quality. In doing so, we are also able to check whether our results are robust to the choice of indicator that represents different segments of institutional quality.[17] We include them in the model(s) by replacing government effectiveness ($x_6$) in Model 7 with the five other dimensions of governance, namely, control of corruption, voice and accountability, political stability/no violence, regulatory quality, and rule of law. Results are reported in Table 12, in which we test whether the efficiencies found differ when considering government effectiveness or other alternatives. They show that results differ, but not significantly. We must admit, however, that this way of examining the impact of different quality of government variables expands the possible outcomes of our approach and, therefore, deserves further investigation.

A graphical illustration is provided in Figure 1, which displays densities for both over- and underratings corresponding to all the groups considered. Only in the first of these densities (upper left sub-figure) is it apparent that the lines almost overlap. In the other three cases the lines corresponding to each density being compared (solid and dashed lines) are, in general, different, especially when considering the crisis or the euro effect.

## 6. Conclusions

Credit ratings stimulated much debate during and after the 2008 financial crisis. The scope of the debate was broad and the content was mixed, but what is certain is that potential weaknesses in SCRs can create havoc in financial markets. Despite the fact that several measures were taken to

[16] The results of the test reported in Table 11 are not directly comparable with those in Tables 7–10. Because of this, in Table 11 we should also differentiate between overratings and underratings, but we considered this would have increased the number and/or size of the tables in the results' section to unreasonable limits.

[17] We are grateful to the associate editor for this pertinent comment.

curb the adverse effects of the crisis, credit ratings still need proper regulation. The contributions investigating SCRs deal with a variety of topics but have less frequently explored objective and subjective components of SCRs. This is partly because many market players, policymakers and academics take credit ratings for granted.

Our study examined SCRs by proposing a nonparametric partial frontier approach. The proposal to use nonparametric techniques should be considered as an innovative application in the credit ratings literature as it contributes a new approach to the scarce research in this field. The main advantage of the partial frontier analysis conducted in the study is that it measures the contribution of subjective judgement for each SCR in our sample, either over- or underrating, since the order-$m$ estimators provide results for both inefficiencies (underratings) and superefficiencies (overratings). Apart from the originality of this methodology, the study contributes to the investigation and discussion of over- and underratings, which remain at the heart of regulatory debate, even though several years have gone by since the global financial crisis.

Our findings suggest that the magnitude of both overratings and underratings is indeed remarkable. Although partial frontiers require a trimming parameter to be specified, it is always possible to detect potentially underrated countries and, more interestingly, this can be done *contemporaneously*. Our results also reveal differences among groups of countries for both underratings and overratings: specifically, developing countries receive lower ratings than their developed peers with respect to their fundamentals. These differences were significant when comparing OECD vs non-OECD countries or Eurozone vs non-Eurozone countries, and results generally suggest that, on average, the magnitude of overratings and underratings is higher for non-OECD and non-Eurozone countries. The other interesting result suggests that the 2007–2008 financial crisis reduced the impact of subjective revisions in quantitative assessments to some extent. This can be explained by the new business model of CRAs, which became more data-reliant after the 2008 global financial crisis.

The importance of our findings should also be assessed from the point of view of financial stability, which is among the top priorities of policymakers in the aftermath of the 2008 financial crisis. Many precautionary measures were implemented to avoid further global crises in the system. This study complements endeavours in this respect, and recommends vigilant monitoring of credit ratings. The methods we propose for application in this context could serve as an alter-

native (or supplementary) basis for more effective monitoring, since they allow both overratings and underratings to be measured contemporaneously, thus guiding the correction of potential misalignments in SCRs in order to achieve greater financial stability.

## Acknowledgements

We are grateful to Ana Lozano-Vivas and other participants at the Second Santander Chair International Workshop of Efficiency and Productivity 2018 for their comments and suggestions. We acknowledge particularly those made by the referee and the associate editor, which contributed to the overall improvement of the paper. Emili Tortosa-Ausina acknowledges the financial support of the Ministerio de Economía y Competitividad (ECO2017-85746-P), Generalitat Valenciana (PROMETEO 2018/102) and Universitat Jaume I (UJI-B2017-33). The usual disclaimer applies.

## References


Abdelsalam, O., Duygun, M., Matallín-Sáez, J. C., and Tortosa-Ausina, E. (2014). Do ethics imply persistence? The case of Islamic and socially responsible funds. *Journal of Banking & Finance*, 40:182–194.

Afonso, A., Furceri, D., and Gomes, P. (2012). Sovereign credit ratings and financial markets linkages: Application to European data. *Journal of International Money and Finance*, 31(3):606–638.

Afonso, A., Gomes, P. M., and Rother, P. (2007). What 'hides' behind sovereign debt ratings? Working Paper 711, European Central Bank, Frankfurt am Main.

Aigner, D., Lovell, C. A. K., and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production models. *Journal of Econometrics*, 6:21–37.

Aizenman, J., Binici, M., and Hutchison, M. (2013). Credit ratings and the pricing of sovereign debt during the euro crisis. *Oxford Review of Economic Policy*, 29(3):582–609.

Alsakka, R. and ap Gwilym, O. (2010). Leads and lags in sovereign credit ratings. *Journal of Banking & Finance*, 34(11):2614–2626.

Alsakka, R. and ap Gwilym, O. (2012). Rating agencies' credit signals: An analysis of sovereign watch and outlook. *International Review of Financial Analysis*, 21(C):45–55.

Alsakka, R. and ap Gwilym, O. (2013). Rating agencies' signals during the European sovereign debt crisis: Market impact and spillovers. *Journal of Economic Behavior & Organization*, 85(C):144–162.

Amstad, M. and Packer, F. (2015). Sovereign ratings of advanced and emerging economies after the crisis. *BIS Quarterly Review*.

Andersen, P. and Petersen, N. C. (1993). A procedure for ranking efficient units in Data Envelopment Analysis. *Management Science*, 39(10):1261–1264.

Aragon, Y., Daouia, A., and Thomas-Agnan, C. (2005). Nonparametric frontier estimation: A conditional quantile-based approach. *Econometric Theory*, 21(2):358–389.

Badunenko, O., Henderson, D. J., and Kumbhakar, S. C. (2012). When, where and how to perform efficiency estimation. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 175(4):863–892.

Bar-Isaac, H. and Shapiro, J. (2011). Credit Ratings Accuracy and Analyst Incentives. *American Economic Review*, 101(3):120–24.

Bennell, J. A., Crabbe, D., Thomas, S., and ap Gwilym, O. (2006). Modelling sovereign credit ratings: Neural networks versus ordered probit. *Expert Systems with Applications*, 30(3):415–425.

Binici, M., Hutchison, M. M., and Miao, E. W. (2018). Are credit rating agencies discredited? Measuring market price effects from agency sovereign debt announcements. BIS Working Papers 704, Bank for International Settlements.

Bolton, P., Freixas, X., and Shapiro, J. (2012). The credit ratings game. *The Journal of Finance*, 67(1):85–111.

Borensztein, E., Cowan, K., and Valenzuela, P. (2013). Sovereign ceilings "lite"? The impact of sovereign ratings on corporate ratings. *Journal of Banking & Finance*, 37(11):4014–4024.

Candelon, B., Sy, A. N. R., and Arezki, R. (2011). Sovereign rating news and financial markets spillovers: Evidence from the european debt crisis. IMF Working Papers 11/68, International Monetary Fund.

Cantor, R. Packer, F. (1995). Multiple ratings and credit standards: Differences of opinion in the credit rating industry. Federal Reserve Bank of New York Research Paper 9527, Federal Reserve Bank of New York.

Cazals, C., Florens, J.-P., and Simar, L. (2002). Nonparametric frontier estimation: a robust approach. *Journal of Econometrics*, 106:1–25.

Charnes, A., Cooper, W. W., and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6):429–444.

Daouia, A. and Gijbels, I. (2011). Robustness and inference in nonparametric partial frontier modeling. *Journal of Econometrics*, 161(2):147–165.

Daouia, A. and Simar, L. (2007). Nonparametric efficiency analsyis: A multivariate conditional quantile approach. *Journal of Econometrics*, 140:375–400.

Daraio, C., Kerstens, K. H., Nepomuceno, T. C. C., and Sickles, R. (2019). Productivity and efficiency analysis software: An exploratory bibliographical survey of the options. *Journal of Economic Surveys*, 33(1):85–100.

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

Daraio, C. and Simar, L. (2007). *Advanced Robust and Nonparametric Methods in Efficiency Analysis. Methodology and Applications*. Studies in Productivity and Efficiency. Springer, New York.

Darbellay, A. and Frank, P. (2012). Credit rating agencies and regulatory reform. In Hill, C. A. and Krusemark, J. L., editors, *Research Handbook on the Economics of Corporate Law*. Edward Elgar Publishing.

Eijffinger, S. (2012). Rating agencies: Role and influence of their sovereign credit risk assessment in the Eurozone. *Journal of Common Market Studies*, 50(6):912–921.

Eling, M. and Schuhmacher, F. (2007). Does the choice of performance measure influence the evaluation of hedge funds? *Journal of Banking & Finance*, 31(9):2632–2647.

Erdem, O. and Varli, Y. (2014). Understanding the sovereign credit ratings of emerging markets. *Emerging Markets Review*, 20(0):42 – 57.

Ferri, G., Liu, L.-G., and Stiglitz, J. E. (1999). The procyclical role of rating agencies: Evidence from the east asian crisis. *Economic Notes*, 28(3):335–355.

Ferrier, G. D. and Lovell, C. A. K. (1990). Measuring cost efficiency in banking: econometric and linear programming evidence. *Journal of Econometrics*, 46:229–245.

Frank, C. R. and Cline, W. R. (1971). Measurement of debt servicing capacity: An application of discriminant analysis. *Journal of International Economics*, 1(3):327 – 344.

Fuchs, A. and Gehring, K. (2017). The home bias in sovereign ratings. *Journal of the European Economic Association*, 15(6):1386–1423.

Gaillard, N. (2014). What is the value of sovereign ratings? *German Economic Review*, 15(1):208–224.

Gärtner, M., Griesbach, B., and Jung, F. (2011). Pigs or lambs? the european sovereign debt crisis and the role of rating agencies. *International Advances in Economic Research*, 17(3):288–299.

Georgescu-Roegen, N. (1951). The aggregate linear production function and its applications to von Neumann's economic model. In Koopmans, T., editor, *Activity Analysis of Production and Allocation*, pages 98–115. Wiley, New York.

Grinols, E. (1976). International debt rescheduling and discrimination using financial variables. Mimeo, US Treasury Department.

Gültekin-Karakaş, D., Hisarciklilar, M., and Öztürk, H. (2011). Sovereign risk ratings: Biased toward developed countries? *Emerging Markets Finance and Trade*, 47(0):69–87.

Hill, C. A. (2004). Regulating the rating agencies. *Washington University Law Quarterly*, 82:43.

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., and Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, 37(4):543–558.

Kunovac, D. and Ravnik, R. (2017). Are sovereign credit ratings overrated? *Comparative Economic Studies*, 59(2):210–242.

Li, Q. (1996). Nonparametric testing of closeness between two unknown distribution functions. *Econometric Reviews*, 15:261–274.

Maher, J. J. and Sen, T. K. (1997). Predicting bond ratings using neural networks: A comparison with logistic regression. *Intelligent Systems in Accounting, Finance & Management*, 6(1):59–72.

Matallín-Sáez, J. C., Soler-Domínguez, A., and Tortosa-Ausina, E. (2014). On the informativeness of persistence for evaluating mutual fund performance using partial frontiers. *Omega*, 42:47–64.

Matallín-Sáez, J. C., Soler-Domínguez, A., and Tortosa-Ausina, E. (2019). Does active management add value? new evidence from a quantile regression approach. *Journal of the Operational Research Society*, 70(10):1734–1751.

Meeusen, W. and van den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review*, 18(2):435–555.

Mellios, C. and Paget-Blanc, E. (2006). Which factors determine sovereign credit ratings? *The European Journal of Finance*, 12(4):361–377.

Mora, N. (2006). Sovereign credit ratings: Guilty beyond reasonable doubt? *Journal of Banking & Finance*, 30(7):2041 – 2062. Special Section: Banking and Finance in an Integrating Europe.

Murillo-Zamorano, L. R. (2004). Economic efficiency and frontier techniques. *Journal of Economic Surveys*, 18(1):33–77.

Opp, C. C., Opp, M. M., and Harris, M. (2013). Rating agencies in the face of regulation. *Journal of Financial Economics*, 108(1):46 – 61.

Ozturk, H., Namli, E., and Erdal, H. I. (2016a). Modelling sovereign credit ratings: The accuracy of models in a heterogeneous sample. *Economic Modelling*, 54:469–478.

Ozturk, H., Namli, E., and Erdal, H. I. (2016b). Reducing Overreliance on Sovereign Credit Ratings: Which Model Serves Better? *Computational Economics*, 48(1):59–81.

Shephard, R. W. (1970). *Theory of Cost and Production Functions*. Princeton University Press, Princeton.

Simar, L. and Wilson, P. W. (2008). Statistical inference in nonparametric frontier models: Recent developments and perspectives. In Fried, H., Lovell, C. A. K., and Schmidt, S. S., editors, *The Measurement of Productive Efficiency*, chapter 4, pages 421–521. Oxford University Press, Oxford, 2nd edition.

Sy, A. N. R. (2009). The systemic regulation of credit rating agencies and rated markets. IMF Working Papers 09/129, International Monetary Fund, Washington.

Thieme, C., Prior, D., and Tortosa-Ausina, E. (2013). A mutilevel decomposition of school performance using robust nonparametric frontier techniques. *Economics of Education Review*, 32:104–121.

Treepongkaruna, S. and Wu, E. (2012). Realizing the volatility impacts of sovereign credit ratings information on equity and currency markets: Evidence from the Asian Financial Crisis. *Research in International Business and Finance*, 26(3):335–352.

Vernazza, D. R. and Nielsen, E. F. (2015). The Damaging Bias of Sovereign Ratings. *Economic Notes*, 44(2):361–408.

Wang, G., Hao, J., Ma, J., and Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, 38(1):223–230.

Williams, G., Alsakka, R., and ap Gwilym, O. (2013). The impact of sovereign rating actions on bank ratings in emerging markets. *Journal of Banking & Finance*, 37(2):563–577.

**Table 1:** Countries by country segmentation

| Developing countries | | Developed countries | |
|---|---|---|---|
| Albania | Latvia | Australia | Luxembourg |
| Azerbaijan | Lebanon | Austria | Macao |
| Belarus | Lithuania | Bahrain | Malta |
| Bolivia | Malaysia | Barbados | Netherlands |
| Bosnia and Herzegovina | Mauritius | Belgium | New Zealand |
| Botswana | Mexico | Canada | Norway |
| Brazil | Mongolia | Cyprus | Oman |
| Bulgaria | Montenegro | Czech Republic | Portugal |
| Cambodia | Morocco | Denmark | Qatar |
| Chile | Panama | Estonia | Saudi Arabia |
| China | Papua New Guinea | Finland | Slovakia |
| Colombia | Peru | France | Slovenia |
| Costa Rica | Philippines | Germany | South Africa |
| Croatia | Poland | Greece | Spain |
| Dominican Republic | Romania | Hong Kong | Sweden |
| Egypt | Russia | Hungary | Switzerland |
| El Salvador | Singapore | Iceland | Taiwan |
| Fiji Islands | St. Vincent and the Grenadines | Ireland | Trinidad & Tobago |
| Guatemala | Suriname | Israel | United Arab Emirates |
| Honduras | Thailand | Italy | United Kingdom |
| India | Tunisia | Japan | United States of America |
| Indonesia | Turkey | Korea | |
| Jordan | Uruguay | Kuwait | |
| Kazakhstan | Vietnam | | |

Note: Developing country: Low income, lower middle income, and upper middle income, Developed country: high income OECD and high income non-OECD.

**Table 2:** Ratings by country segmentation

| Rating | Developing countries | Developed countries | Full sample |
|---|---|---|---|
| B3/B- | 4.92 | 0.00 | 2.41 |
| B2/B | 8.66 | 0.00 | 4.25 |
| B1/B+ | 13.78 | 0.00 | 6.76 |
| Ba3/BB- | 6.89 | 0.00 | 3.38 |
| Ba2/BB | 11.61 | 0.00 | 5.69 |
| Ba1/BB+ | 14.96 | 1.33 | 8.01 |
| Baa3/BBB- | 12.80 | 2.84 | 7.72 |
| Baa2/BBB | 8.07 | 4.55 | 6.27 |
| Baa1 | 6.89 | 4.92 | 5.89 |
| Baa1/BBB+ | 3.35 | 6.06 | 4.73 |
| A2/A | 6.89 | 8.71 | 7.82 |
| A1/A+ | 1.18 | 10.61 | 5.98 |
| Aa3/AA- | 0.00 | 6.44 | 3.28 |
| Aa2/AA | 0.00 | 10.04 | 5.12 |
| Aa1/AA+ | 0.00 | 5.11 | 2.61 |
| Aaa/AAA | 0.00 | 39.39 | 20.08 |
| Total | 100.00 | 100.00 | 100.00 |

Note: Developing country: Low income, lower middle income, and upper middle income, Developed country: high income OECD and high income non-OECD.

**Table 3:** Descriptive statistics of country specific variables

| | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|
| Financial and Macroeconomic Indicators (provided by the CRA) | | | | |
| *balancegdp* | 0.614 | 11.383 | –39.600 | 131.700 |
| *expendituregdp* | 34.184 | 11.104 | 11.104 | 58.600 |
| *financialbalancegdp* | –1.178 | 6.026 | –23.100 | 48.400 |
| *primarybalancegdp* | 1.381 | 5.511 | –11.400 | 48.800 |
| *exportsprcnt* | 10.800 | 15.049 | –42.700 | 74.300 |
| *gdppc* | 14,729.060 | 17,617.156 | 275.000 | 118,566.000 |
| *gdpprcnt* | 9.215 | 9.368 | –29.000 | 83.400 |
| *inflation* | 4.497 | 5.590 | –4.000 | 68.800 |
| *investgdpratio* | 22.857 | 5.558 | 5.558 | 43.200 |
| *savinggdpratio* | 24.987 | 12.257 | –12.000 | 71.500 |
| *foreignexcreserve* | 35.856 | 87.430 | 0.000 | 947.990 |
| *debtgdp* | 45.439 | 30.327 | 0.000 | 191.600 |
| Governance Indicators (World Bank) | | | | |
| *governmenteffectiveness* | 0.651 | 0.860 | –1.169 | 2.408 |

**Table 4:** Definition of variables

| Variable name | Description (CRA's Financial and Macroeconomic Indicators) |
|---|---|
| *balancegdp* | Ratio of current account balance to GDP |
| *financialbalancegdp* | Ratio of general government financial balance to GDP |
| *primarybalancegdp* | Ratio of general government primary balance to GDP |
| *exportsprcnt* | Nominal exports of goods and services (percentage change, USD) |
| *gdppc* | GDP per capita |
| *gdpprcnt* | Nominal GDP percentage change (local currency) |
| *inflation* | Inflation (CPI) |
| *investgdpratio* | Ratio of gross investment to GDP |
| *savinggdpratio* | Ratio of domestic saving to GDP |
| *foreignexcreserve* | Official foreign exchange reserves (billion USD) |
| *debtgdp* | Ratio of general government debt to GDP |
| Variable name | Description (Governance Indicators, The World Bank) |
| *governmenteffectiveness* | Government effectiveness |

**Table 5:** Model selection results based on the Li (1996) test, restricted vs. unrestricted models

| Null hypothesis | $T$-statistic | $p$-value |
|---|---|---|
| $H_0 : f(\text{Model 1}) = g(\text{Model 2})$ | 1.7787 | 0.0376 |
| $H_0 : f(\text{Model 2}) = g(\text{Model 3})$ | 15.7205 | 0.0000 |
| $H_0 : f(\text{Model 3}) = g(\text{Model 4})$ | 12.6977 | 0.0000 |
| $H_0 : f(\text{Model 4}) = g(\text{Model 5})$ | 0.7868 | 0.2157 |
| $H_0 : f(\text{Model 4}) = g(\text{Model 6})$ | 0.4137 | 0.3396 |
| $H_0 : f(\text{Model 4}) = g(\text{Model 7})$ | 1.0173 | 0.1545 |
| Model 1: $x_1$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$, $y$ | | |
| Model 2: $x_1$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$, $x_7$, $y$ | | |
| Model 3: $x_1$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$, $x_7$, $x_8$, $y$ | | |
| Model 4: $x_1$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$, $x_7$, $x_8$, $x_9$, $y$ | | |
| Model 5: $x_1$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$, $x_7$, $x_8$, $x_9$, $x_{10}$, $y$ | | |
| Model 6: $x_1$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$, $x_7$, $x_8$, $x_9$, $x_{11}$, $y$ | | |
| Model 7: $x_1$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$, $x_7$, $x_8$, $x_9$, $x_{12}$, $y$ | | |

**Table 6:** Order-$m$ efficiencies

| | | Mean | 1st quartile | Median | 3rd quartile | Std.dev. | #[a] |
|---|---|---|---|---|---|---|---|
| $m_{\alpha=.90}$ | Underrated | 0.7837 | 0.7372 | 0.8318 | 0.8919 | 0.1636 | 72 |
| | Overrated | 1.0580 | 1.0064 | 1.0180 | 1.0567 | 0.0982 | 114 |
| $m_{\alpha=.95}$ | Underrated | 0.7604 | 0.7122 | 0.7857 | 0.8573 | 0.1578 | 80 |
| | Overrated | 1.0281 | 1.0020 | 1.0065 | 1.0304 | 0.0484 | 48 |
| $m_{\alpha=.99}$ | Underrated | 0.7372 | 0.6667 | 0.7500 | 0.8571 | 0.1565 | 81 |
| | Overrated | 1.0039 | 1.0011 | 1.0016 | 1.0045 | 0.0050 | 4 |
| Overrating/underrating | $m_{\alpha=.90}$ | 0.8867 | 0.8363 | 0.9337 | 0.9893 | 0.1417 | 186 |
| | $m_{\alpha=.95}$ | 0.8407 | 0.7500 | 0.8639 | 0.9787 | 0.1642 | 128 |
| | $m_{\alpha=.99}$ | 0.7494 | 0.6671 | 0.7500 | 0.8571 | 0.1624 | 85 |

We followed Daouia and Gijbels (2011) in selecting the three values for the trimming parameter ($m$), as explained in footnote 11.

[a] The total number of observations (country-year) is 1,023 for the 1999–2010 period. Therefore, by subtracting the sum of overrated and underrated countries we obtain the country-years that are rated correctly for each choice of $m$, i.e., $1,023 - 72 - 114 = 837$, $1,023 - 80 - 48 = 895$ and $1,023 - 81 - 4 = 938$ (for $m_{\alpha=.90}$, $m_{\alpha=.95}$ and $m_{\alpha=.99}$, respectively).

**Table 7:** Order-$m$ efficiencies, developed vs. developing countries

| | | | Mean | 1st quartile | Median | 3rd quartile | Std.dev. | #[a] |
|---|---|---|---|---|---|---|---|---|
| $m_{\alpha=.90}$ | Developed | Underrated | 0.8336 | 0.7692 | 0.8553 | 0.9128 | 0.0900 | 47 |
| | | Overrated | 1.0434 | 1.0048 | 1.0164 | 1.0450 | 0.0731 | 59 |
| | Developing | Underrated | 0.6899 | 0.6040 | 0.7562 | 0.8327 | 0.2228 | 25 |
| | | Overrated | 1.0736 | 1.0068 | 1.0267 | 1.0793 | 0.1181 | 55 |
| $m_{\alpha=.95}$ | Developed | Underrated | 0.7999 | 0.7475 | 0.8066 | 0.8582 | 0.0994 | 51 |
| | | Overrated | 1.0187 | 1.0017 | 1.0043 | 1.0208 | 0.0361 | 25 |
| | Developing | Underrated | 0.6908 | 0.6000 | 0.7415 | 0.8200 | 0.2116 | 29 |
| | | Overrated | 1.0383 | 1.0045 | 1.0120 | 1.0470 | 0.0581 | 23 |
| $m_{\alpha=.99}$ | Developed | Underrated | 0.7733 | 0.7143 | 0.7505 | 0.8571 | 0.1062 | 51 |
| | | Overrated | 1.0016 | 1.0014 | 1.0016 | 1.0019 | 0.0007 | 2 |
| | Developing | Underrated | 0.6760 | 0.6000 | 0.6905 | 0.8295 | 0.2049 | 30 |
| | | Overrated | 1.0062 | 1.0036 | 1.0062 | 1.0088 | 0.0074 | 2 |
| Overrated/underrated (developed) | | $m_{\alpha=.90}$ | 0.9052 | 0.8557 | 0.9318 | 0.9894 | 0.0969 | 106 |
| | | $m_{\alpha=.95}$ | 0.8601 | 0.7803 | 0.8581 | 0.9787 | 0.1199 | 76 |
| | | $m_{\alpha=.99}$ | 0.7818 | 0.7143 | 0.7857 | 0.8571 | 0.1128 | 53 |
| Overrated/underrated (developing) | | $m_{\alpha=.90}$ | 0.8621 | 0.8087 | 0.9529 | 0.9884 | 0.1830 | 80 |
| | | $m_{\alpha=.95}$ | 0.8124 | 0.7312 | 0.9043 | 0.9755 | 0.2111 | 52 |
| | | $m_{\alpha=.99}$ | 0.6958 | 0.6000 | 0.7143 | 0.8750 | 0.2130 | 32 |

We followed Daouia and Gijbels (2011) in selecting the three values for the trimming parameter ($m$), as explained in footnote 11.

[a] For the interpretation of the numbers in this column, see Table 6 and its corresponding table notes.

**Table 8:** Order-$m$ efficiencies, OECD vs. non-OECD countries

| | | | Mean | 1st quartile | Median | 3rd quartile | Std.dev. | #[a] |
|---|---|---|---|---|---|---|---|---|
| $m_{\alpha=.90}$ | OECD | Underrated | 0.8276 | 0.7857 | 0.8420 | 0.8666 | 0.0773 | 20 |
| | | Overrated | 1.0434 | 1.0039 | 1.0171 | 1.0400 | 0.0782 | 41 |
| | Non-OECD | Underrated | 0.7668 | 0.7173 | 0.8101 | 0.9030 | 0.1844 | 52 |
| | | Overrated | 1.0662 | 1.0067 | 1.0200 | 1.0720 | 0.1075 | 73 |
| $m_{\alpha=.95}$ | OECD | Underrated | 0.8148 | 0.7781 | 0.8182 | 0.8571 | 0.0851 | 25 |
| | | Overrated | 1.0194 | 1.0010 | 1.0032 | 1.0199 | 0.0431 | 14 |
| | Non-OECD | Underrated | 0.7356 | 0.6741 | 0.7500 | 0.8578 | 0.1767 | 55 |
| | | Overrated | 1.0316 | 1.0031 | 1.0092 | 1.0383 | 0.0506 | 34 |
| $m_{\alpha=.99}$ | OECD | Underrated | 0.7932 | 0.7500 | 0.7857 | 0.8571 | 0.0833 | 25 |
| | | Overrated | – | – | – | – | – | 0 |
| | Non-OECD | Underrated | 0.7123 | 0.6429 | 0.7159 | 0.8571 | 0.1748 | 56 |
| | | Overrated | 1.0039 | 1.0011 | 1.0016 | 1.0045 | 0.0050 | 4 |
| Overrated/underrated (OECD) | | $m_{\alpha=.90}$ | 0.9184 | 0.8664 | 0.9615 | 0.9922 | 0.0905 | 61 |
| | | $m_{\alpha=.95}$ | 0.8750 | 0.7873 | 0.8579 | 0.9826 | 0.1080 | 39 |
| | | $m_{\alpha=.99}$ | 0.7932 | 0.7500 | 0.7857 | 0.8571 | 0.0833 | 25 |
| Overrated/underrated (non-OECD) | | $m_{\alpha=.90}$ | 0.8712 | 0.8101 | 0.9286 | 0.9885 | 0.1590 | 125 |
| | | $m_{\alpha=.95}$ | 0.8257 | 0.7300 | 0.8771 | 0.9717 | 0.1819 | 89 |
| | | $m_{\alpha=.99}$ | 0.7312 | 0.6442 | 0.7232 | 0.8750 | 0.1832 | 60 |

We followed Daouia and Gijbels (2011) in selecting the three values for the trimming parameter ($m$), as explained in footnote 11.

[a] For the interpretation of the numbers in this column, see Table 6 and its corresponding table notes.

**Table 9:** Order-$m$ efficiencies, euro vs. non-euro countries

| | | | Mean | 1st quartile | Median | 3rd quartile | Std.dev. | #[a] |
|---|---|---|---|---|---|---|---|---|
| $m_{\alpha=.90}$ | Euro | Underrated | 0.8308 | 0.7729 | 0.8218 | 0.8503 | 0.0685 | 5 |
| | | Overrated | 1.0280 | 1.0042 | 1.0092 | 1.0287 | 0.0422 | 8 |
| | Non-euro | Underrated | 0.7802 | 0.7319 | 0.8327 | 0.8943 | 0.1683 | 67 |
| | | Overrated | 1.0602 | 1.0067 | 1.0187 | 1.0624 | 0.1009 | 106 |
| $m_{\alpha=.95}$ | Euro | Underrated | 0.7965 | 0.7500 | 0.7512 | 0.8182 | 0.086 | 5 |
| | | Overrated | 1.0041 | 1.0007 | 1.0037 | 1.0071 | 0.0040 | 4 |
| | Non-euro | Underrated | 0.7580 | 0.6915 | 0.7857 | 0.8575 | 0.1615 | 75 |
| | | Overrated | 1.0303 | 1.0021 | 1.0081 | 1.0319 | 0.0500 | 44 |
| $m_{\alpha=.99}$ | Euro | Underrated | 0.7886 | 0.7500 | 0.7500 | 0.8182 | 0.0952 | 5 |
| | | Overrated | – | – | – | – | – | 0 |
| | Non-euro | Underrated | 0.7339 | 0.6636 | 0.7500 | 0.8571 | 0.1596 | 76 |
| | | Overrated | 1.0039 | 1.0011 | 1.0016 | 1.0045 | 0.0050 | 4 |
| Overrated/underrated (euro) | | $m_{\alpha=.90}$ | 0.9190 | 0.8503 | 0.9486 | 0.9922 | 0.0875 | 13 |
| | | $m_{\alpha=.95}$ | 0.8852 | 0.7512 | 0.9375 | 0.9934 | 0.1215 | 9 |
| | | $m_{\alpha=.99}$ | 0.7886 | 0.7500 | 0.7500 | 0.8182 | 0.0952 | 5 |
| Overrated/underrated (non-euro) | | $m_{\alpha=.90}$ | 0.8842 | 0.8356 | 0.9328 | 0.9885 | 0.1449 | 173 |
| | | $m_{\alpha=.95}$ | 0.8373 | 0.7500 | 0.8585 | 0.9747 | 0.1669 | 119 |
| | | $m_{\alpha=.99}$ | 0.7470 | 0.6667 | 0.7500 | 0.8571 | 0.1658 | 80 |

We followed Daouia and Gijbels (2011) in selecting the three values for the trimming parameter ($m$), as explained in footnote 11.

[a] For the interpretation of the numbers in this column, see Table 6 and its corresponding table notes.

**Table 10:** Order-$m$ efficiencies, pre-crisis (1999–2007) vs. crisis years (2008–2010)

| | | | Mean | 1st quartile | Median | 3rd quartile | Std.dev. | #[a] |
|---|---|---|---|---|---|---|---|---|
| $m_{\alpha=.90}$ | Pre-crisis | Underrated | 0.7683 | 0.7309 | 0.8011 | 0.8721 | 0.1705 | 58 |
| | | Overrated | 1.0682 | 1.0060 | 1.0180 | 1.0870 | 0.1103 | 80 |
| | Crisis | Underrated | 0.8478 | 0.8286 | 0.8668 | 0.9277 | 0.1152 | 14 |
| | | Overrated | 1.0339 | 1.0069 | 1.0185 | 1.0350 | 0.0550 | 34 |
| $m_{\alpha=.95}$ | Pre-crisis | Underrated | 0.7316 | 0.6866 | 0.7500 | 0.8200 | 0.1610 | 61 |
| | | Overrated | 1.0309 | 1.0026 | 1.0092 | 1.0311 | 0.0522 | 38 |
| | Crisis | Underrated | 0.8529 | 0.8447 | 0.8788 | 0.9288 | 0.1050 | 19 |
| | | Overrated | 1.0175 | 1.0007 | 1.0034 | 1.0112 | 0.0301 | 10 |
| $m_{\alpha=.99}$ | Pre-crisis | Underrated | 0.7076 | 0.6523 | 0.7176 | 0.8101 | 0.1591 | 62 |
| | | Overrated | 1.0068 | 1.0045 | 1.0068 | 1.0091 | 0.0066 | 2 |
| | Crisis | Underrated | 0.8340 | 0.8258 | 0.8571 | 0.8920 | 0.1010 | 19 |
| | | Overrated | 1.0011 | 1.0010 | 1.0011 | 1.0011 | 0.0001 | 2 |
| Overrated/underrated (pre-crisis) | | $m_{\alpha=.90}$ | 0.8702 | 0.8096 | 0.9177 | 0.9884 | 0.1526 | 138 |
| | | $m_{\alpha=.95}$ | 0.8239 | 0.7432 | 0.8462 | 0.9780 | 0.1744 | 99 |
| | | $m_{\alpha=.99}$ | 0.7165 | 0.6538 | 0.7232 | 0.8252 | 0.1643 | 64 |
| Overrated/underrated (crisis) | | $m_{\alpha=.90}$ | 0.9340 | 0.9229 | 0.9706 | 0.9921 | 0.0903 | 48 |
| | | $m_{\alpha=.95}$ | 0.8979 | 0.8571 | 0.9286 | 0.9868 | 0.1065 | 29 |
| | | $m_{\alpha=.99}$ | 0.8497 | 0.8333 | 0.8593 | 0.9231 | 0.1079 | 21 |

We followed Daouia and Gijbels (2011) in selecting the three values for the trimming parameter ($m$), as explained in footnote 11.

[a] For the interpretation of the numbers in this column, see Table 6 and its corresponding table notes.

**Table 11:** Differences among country classifications for the median (Kruskal-Wallis test) and for the distribution (Li test, 1996), $m_{\alpha=.90}$, $m_{\alpha=.95}$ and $m_{\alpha=.99}$

| | Li (1996) test | | Kruskal-Wallis test | |
|---|---|---|---|---|
| Null hypothesis | $T$-statistic | $p$-value | $\chi^2$-statistic | $p$-value |
| $H_0 : f(\text{developed}) = g(\text{developing})$ | –0.9785 | 0.8361 | 0.0012 | 0.9718 |
| $H_0 : f(\text{OECD}) = g(\text{non-OECD})$ | 2.0193* | 0.0217 | 9.2047** | 0.0024 |
| $H_0 : f(\text{euro}) = g(\text{non-euro})$ | 25.8112** | 0.0000 | 56.5069** | 0.0000 |
| $H_0 : f(\text{pre-crisis}) = g(\text{crisis})$ | 4.2560** | 0.0000 | 15.3599** | 0.0000 |

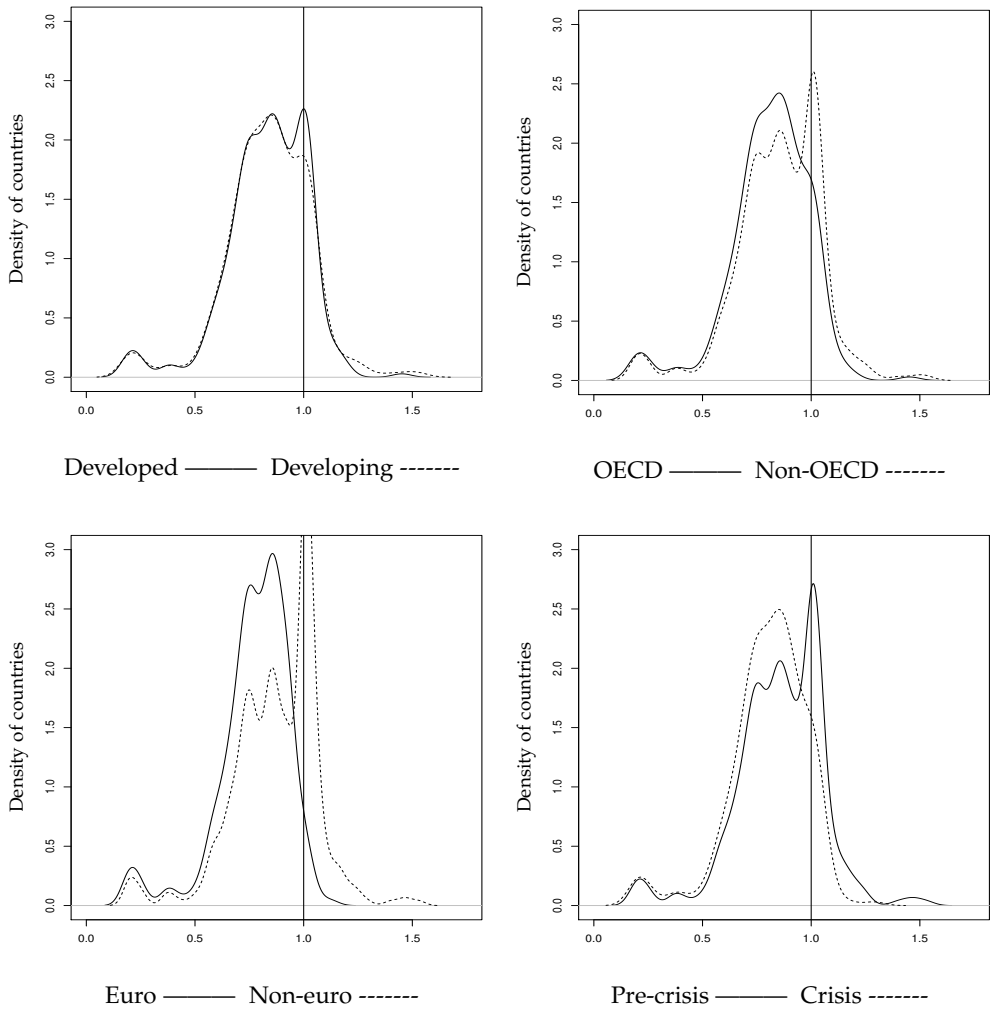
(*), (**): significant differences among medians (Kruskal-Wallis test) and distributions (Li test, 1996) at 5% and 1%, respectively.

**Table 12:** Differences when considering alternative Worldwide Governance Indicators (WGIs) based on the 1996 test, order-$m$ efficiencies ($m_{\alpha=.99}$)

| Model | Model 7 | Model 7(i) | Model 7(ii) | Model 7(iii) | Model 7(iv) | Model 7(v) |
|---|---|---|---|---|---|---|
| | | | | $p$-value | | |
| Model 7 | — | 0.3823 | 0.4766 | 0.4814 | 0.4926 | 0.5000 |
| Model 7(i) | | — | 0.419 | 0.3956 | 0.4566 | 0.3823 |
| Model 7(ii) | | | — | 0.4965 | 0.4949 | 0.4756 |
| Model 7(iii) | | | | — | 0.4931 | 0.4814 |
| Model 7(iv) | | | | | — | 0.4926 |
| Model 7(v) | | | | | | — |

In Models 7(i), 7(ii), 7(iv) and 7(v) the variable government effectiveness was replaced by control of corruption, voice and accountability, political stability/no violence, regulatory quality, and rule of law, respectively.

**Figure 1:** Kernel density plots, overrated and underrated countries




Notes: All figures contain densities estimated using kernel density estimation for the overratings/underratings yielded by the order-$m$ estimators. The vertical lines in each plot represent efficiency. The probability mass below 1 represents the underrated countries; that above 1 represents the overrated countries. A Gaussian kernel was chosen, and bandwidths were estimated using plug-in methods.