BRQ Business Research Quarterly (2018) 21, 251–261

BUSINESS RESEARCH QUARTERLY
www.elsevier.es/brq

REGULAR ARTICLES

Measuring the performance of local administrative public services

Jos L.T. Blank [a,b,c,*]

[a] *Delft University of Technology, Delft, Netherlands*
[b] *Erasmus University Rotterdam, Rotterdam, Netherlands*
[c] *Institute of Public Sector Efficiency Studies, Delft, Netherlands*

Received 12 March 2018; accepted 13 September 2018
Available online 28 October 2018

JEL CLASSIFICATION
C33;
D24;
I12;
O39

KEYWORDS
Weighted least squares;
Frontier analysis;
Efficiency;
Local public services

Abstract The academic literature provides excellent methodologies to identify best practices and to calculate inefficiencies by stochastic frontier analysis. However, these methodologies are regarded as a black box by policy makers and managers and therefore results are hard to accept. This paper proposes an alternative class of stochastic frontier estimators, based on the notion that some observations contain more information than others about the true frontier. If an observation is likely to contain much information, it is assigned a large weight in the regression analysis. In order to establish the weights, we propose an iterative procedure. The advantages of this more intuitive approach are its transparency and its easy application. The method is applied to Dutch local administrative services (LAS) in municipalities. The method converges quickly and produces reliable estimates. About 25% of the LAS are designated as efficient. The average efficiency score is 93%. For the average sized LAS no economies of scale exist.

Introduction

The recent financial and economic crises are forcing many administrations to cut budgets in various areas of public services. Specifically, the European countries that are, or were, under direct budgetary supervision by the Euro Group and/or the IMF, such as Greece and Spain, are experiencing a tremendous impact on service levels in education, healthcare, and infrastructure industries. The pressure on these services is great, as they are of great importance to the structural improvement of their economies – or in a broader sense, in the maintenance of their social welfare. Good physical infrastructures, well-functioning law enforcement, healthy and well-trained personnel are among the many aspects that are important assets for

* Correspondence to: Delft University of Technology, PO Box 5015, 2600 GA Delft, Netherlands.
E-mail addresses: j.l.t.blank@tudelft.nl, j.blank@ipsestudies.nl

https://doi.org/10.1016/j.brq.2018.09.001

economic development and social well-being. The only way to balance shrinking budgets and the need for structural improvement is to enhance performance in these sectors. This implies that more effort must be put into finding ways to improve performance in the public sector. This involves good, supportive policies at both the government level and the management level of executive public institutions. Academics can play an important role in identifying best practices in order to increase knowledge about which types of internal and external governance, incentive structures, market regulations and capacity planning might improve performance.

However, one might surmise (not based on solid empirical evidence) that, in many cases, governments and those managing public institutions are operating in the dark. Academics fail, not only in bridging the gap between practice and theory, but also in providing policymakers and management with evidence-based policy and management measures to strive for optimal strategies and business conduct (see e.g. Curristine et al., 2007). The academic literature provides excellent methodologies to identify best practices (see e.g. Fried et al., 2008; Parmeter and Kumbhakar, 2014) in stochastic frontier analysis (SFA) and data envelopment analysis (DEA). There are numerous examples of applications in public service industries, such as in the health industry (Blank and Valdmanis, 2008; Jacobs et al., 2006) and water and power utilities (Bottasso et al., 2011; Murillo-Zamorano and Vega-Cervera, 2001). Other very interesting public sector applications can be found in Blank (2000), Levitt and Joyce (1987) and Ganley and Cubbin (1992). The technique is also being applied to the comparison of the performance of countries or industries in different countries (Chen and Lin, 2009; Shao and Lin, 2016).

For local public services, there are interesting opportunities on hand. Local public services, depending on the country, provide a substantial part of a country's public services. Aside from their financial relevance, local public services generally provide good ground for conducting best practice research due to the large number of observations and the (mostly) obligatory uniform registration of financial and production data. Further, many data are available on all kinds of contextual variables (including population, social conditions, geographical and climate data). For these reasons, research on local government productivity and efficiency has some popularity amongst researchers (a few examples: Bel and Mur, 2009; Bikker and van der Linde, 2016; Niaounakis and Blank, 2017; Pérez-López et al., 2016; Veenstra et al., 2016; Zafra-Gómez et al., 2013). In this paper we focus on the productivity and efficiency of local administrative services in the Netherlands.

Unfortunately, most of the researchers in this field have ''lost'' themselves in their methods, as opposed to paying attention to practical and policy-relevant issues: and no connection is made with management research. Almost two decades ago, Meier and Gill (2000) complained that frontier or best practice techniques were not being applied to public administration research. They state that ''it has fallen notably behind research in related fields in terms of methodological sophistication. This hinders the development of empirical investigations into substantive questions of interest to practitioners and academics.''

One may wonder why frontier techniques have not become common practice in management or public administration research. A possible explanation is that these techniques are based on sophisticated mathematical economics, econometrics, and statistics. Besides the technical problems researchers might face in applying these techniques, the fact that policymakers and managers do not have faith in the results derived from these complex and rather non-transparent methodologies plays a significant role. It is not the mathematics that are involved that cause acceptance problems but rather the conceptual issues behind these techniques. Apart from the seminal work by Meier and Gill (2000) in their *What Works: A New Approach to Program and Policy Analysis*, few serious attempts have been made to introduce more accessible and transparent methodologies that produce the same results as existing state-of-the-art frontier techniques. Therefore, in this paper, we present a more attractive technique that is based on the original ideas of Gill and Meier, and that provides results that are similar to SFA while presenting fewer computational problems.

In this paper we focus on Dutch local public administrative services. The main service of the local administrative services (LAS) is the provision of passports, driving licenses, and national identity cards, as well as birth, death, and marriage certificates that are retrieved from the local registry upon the request of citizens. From a research perspective, this is an interesting part of local public services, since municipalities are strongly regulated. Every citizen requesting one of these services must be served, and security considerations, for instance with respect to identity theft, ensure that each municipality follows the same procedures. Furthermore, the production of services is unambiguous and good data are available. So, the question is whether municipalities are capable of further improving efficiency by copying best practice behaviour of other (efficient) municipalities. In addition, it is to be expected that in this sector, which is dominated by administrative processes, productivity gains can be achieved with the use of improved information and communication technology.

We define three specific outputs: the (unweighted) sum of passports, identity cards, and driving licenses; the (unweighted) number of extracts from municipal databases (such as birth and death certificates); and the number of marriages (which is included because arranging civil marriage ceremonies is an important activity of this part of local government).

This paper is organised as follows. In the next section, we present a brief literature overview of methodologies of productivity and efficiency measurement, and various types of frontier analysis techniques. Readers who are solely interested in the application of local public services can skip this section, as it is not essential in order to understand the empirical analysis. It merely provides a conceptual justification for the proposed technique. In the consecutive section, we discuss the conceptual and global technical issues concerning the proposed alternative method. Then we apply the model to Dutch local administrative public services by discussing the empirical model, the estimation procedure, the data and the results. We conclude the paper in the final

section. Appendix A discusses the proposed methodology in more detail.

A brief literature review of methodologies

Best practices in public sector service delivery can be identified by various techniques. One of the most popular is the stochastic frontier analysis (SFA) methodology suggested by Aigner et al. (1977) and Meeusen and Van den Broeck (1977). This technique has become a standard in the parametric estimation of production and cost (or any other value) function. It is based on the idea that production (or cost) can be empirically described as a function of a number of inputs (or outputs and input prices); a stochastic term reflecting errors; and a stochastic term reflecting efficiency. Maximum likelihood or least squares techniques can be used to estimate the parameters of the function and the parameters of the distribution of the stochastic components. To put it simply, this technique is essentially a multivariate regression technique, but instead of drawing a graph through the ''middle of all data points'', the graph envelopes them. By doing so the graph does not represent production or cost of the average firm but that of the best performing firms (with highest production or lowest cost, conditional on all other variables). For extensive discussions on this technique (see e.g. Kumbhakar and Lovell, 2000; Fried et al., 2008; Blank and Valdmanis, 2017; Parmeter and Kumbhakar, 2014).

SFA has become very popular, and it has been applied in a great deal of empirical work. Nevertheless, the approach has been widely criticised. The criticisms focus on two major points, namely the a priori specification of the production (or cost) function (why should economic reality behave like a smooth mathematical function?), and the assumptions concerning the distribution of the stochastic term representing efficiency (can efficiency be described as a stochastic distribution function? see e.g. Ondrich and Ruggiero, 2001). A third area of criticism, which is not expressed as often, is of a conceptual nature: the methodology suggests the observation of an unobservable (the efficiency), which can be derived from another unobservable (the measurement and specification error), within a relatively complex econometric framework. Those who try to explain this approach to the non-initiated, such as managers and policymakers, are met with scepticism and disbelief. A technique such as data envelopment analysis (DEA), which actually seeks (existing) observations that form the envelope, is far more attractive and more transparent. This is why DEA has become a very popular tool in applied work on real-life problems. However, DEA has some serious drawbacks, such as measurement errors that substantially affect outcomes, or the lack of ways to correct for contextual variables. Of course, researchers have found some (even more) complex solutions to these problems. However, there may be another way to tackle the problem using another conceptual framing of SFA that makes the technique more accessible to non-experts.

If all firms operate at full efficiency estimating a production, cost, or profit frontier (hereinafter ''frontier'') would not be a big deal, just apply OLS. Although one could use OLS to estimate the parameters of the model, in reality some firms are inefficient, which makes the estimation of the frontier a challenging task. This problem could be solved by neglecting the inefficient firms and only taking efficient firms into account. However, this method implies a priori knowledge of whether a firm is efficient, and knowledge about the efficiency of firms is generally not available prior to the estimation of a production frontier. Therefore, other methods for addressing this problem have been proposed.

An alternative to the original SFA approach is the thick frontier analysis (TFA) developed by Berger and Humphrey (1991). This approach is based on the idea of selecting efficient firms from a first stage of regression analysis. The technique uses a selection of firms in the top 10% (or any other percentage) and the bottom 10%. In a second stage, the production (or cost) function for both subsamples is estimated separately. Cost efficiencies are subsequently derived by taking the ratio of the average cost of the worst practice firms and the best practice firms. TFA does not require any rigid assumptions about the distributions of the efficiency component. It is a conceptually very transparent and attractive approach, although it does have some serious drawbacks. It does not provide firm-specific cost efficiencies, but only more general cost efficiency scores. Further there is a loss of information, due to the discarding of a large subset of observations, and it is questionable whether the researcher can permit him/herself the luxury of losing so much information.

Another approach to estimating a frontier – one that can be regarded as a successor to TFA – is provided by Wagenvoort and Schure (2006), who show how efficient firms can be identified if panel data are available. They use a recursive thick frontier approach (RTFA), dropping the most inefficient firm at each iteration. In each step, the firm-specific efficiency is calculated by averaging the residuals of each individual firm over the whole time period. Their final step consists of using the fully efficient firms to estimate the frontier. Although it is intuitively appealing, RTFA also has some serious drawbacks. It can only be applied to panel data. Furthermore, it is assumed that inefficiency is time-invariant. This implies that a firm cannot change its efficiency over time – which is a fairly rigid assumption, particularly when dealing with a long time span. Another drawback is that it still depends on the assumption of a 0–1 probability of being efficient.

Another complex alternative is quantile regression (see e.g. Koenker and Hallock, 2001). The key issue here is that quantile regression provides an estimate of the conditional median or any other quantile instead of the conditional mean (as in standard regression analysis). To put it simply, the graph does not go through the middle of the cloud of data points but through the upper (or lower) 10 or 25% of the data points. The interesting aspect of this method is that it actually assigns more weight to observations that are close (conditionally on the explanatory variables) to the desired quantile. Thus, in contrast to TFA, it does not drop or ignore a number of observations. Although promising results have been achieved with this method, it lacks transparency, perhaps even more so than SFA. The concept is very hard to understand, calculations are based on linear programming techniques, and no straightforward statistical inferences can be made.

Our proposed method also has a strong resemblance to earlier work by Meier and Gill (2000), who focused on investigating subgroups in a given sample by applying a method called substantively weighted least squares (SWLS).

In an iterative procedure, SWLS selects the outliers from standard least squares (e.g., observations with residuals above 3 times the standard deviation of the residuals), and re-estimates the model by assigning weights equal to 1 to observations in the selection, and weights smaller than 1 to observations outside the selection. In an iterative procedure, the weights corresponding to the observations outside the selection are successively decreased. Although this method is quite attractive, it has no direct link to standard productivity and efficiency literature, and weights are handled in the iterations in a somewhat ad hoc way.

Our approach combines the best of many worlds. We argue that whether a firm is fully efficient or not does not concern a 0–1 casus, but is probabilistic. We therefore introduce weights to the observations and show the way in which a weighting scheme can be implemented in order to determine which firms are likely to be efficient and which are likely to be inefficient. At the same time, we are able to preserve the transparency of the RTFA and the SWLS method by applying standard least squares techniques and without losing any degrees of freedom, which occurs in RTFA (by creating a subsample of selected observations). With respect to the SWLS method, our approach does not assign common and rather arbitrary weights to the observations outside the selection. Instead, we use weights that reflect the probability of being efficient or nearly efficient, which implies a minimum loss of information, and therefore leads to more efficient estimates of the model parameters.

Our concept also translates to a cross-section setting so as to avoid the need for panel data. This also implies that we do not need to assume that inefficiency is time-invariant, which can be regarded as a somewhat restrictive assumption in many efficiency models that are based on panel data.

Thus, our approach is related to the concept of stochastic frontier analysis, but is far more conceptually appealing. Our alternative incorporates information derived from all the available data. It is based on an iterative weighted least squares (IWLS) method and can easily be programmed in standard statistical software.

Alternative methodology

Economic framework

We start with the cost function, although the method may be applied to any other model (production model, profit model). The cost function is a mathematical description between cost on one hand and services delivered and input prices on the other hand. In the context of local administrative services, a cost function approach is probably most appropriate, since outputs and input prices are exogenous. Every citizen requesting an administrative service must be served, by any means necessary. So municipalities cannot influence outputs, but only inputs. It is even impossible to affect outputs by creating waiting lists since municipalities are required to deliver within a limited number of days.

We assume that total cost can be represented by a cost function $c(y, w)$, where y and w are a vector of various output and input prices, respectively, that meets all the requirements it entails. For convenience, we rewrite the cost equations in terms of logarithms and add an error term (representing measurement errors and possible inefficiencies).

$$\ln(C) = c(\ln(y), \ln(w)) + \varepsilon \tag{1}$$

with C = total costs; y = vector of outputs; w = vector of input prices; ε = error term.

The parameters of Eq. (1) can be estimated by a least squares method. However, if certain firms are inefficient – that is, they have a cost that is higher than that which can be accounted for – the cost function will cause biases in the estimated parameters of Eq. (1). In the estimation procedure we take this into account by attributing less weight to the observations that are expected to be inefficient.

Applying iteratively weighted least squares

So we can reduce these biases by estimating Eq. (1) with weighted least squares, and assigning the relatively inefficient observations a small weight and the relatively efficient observations a large weight. Weighted least squares (WLS), which is also referred to as generalised least squares (GLS), is a widely used econometric technique to deal with this heterogeneity in data; however, since the weights are generally not observable, they have to be estimated (see e.g. Verbeek, 2017). Our proposed weighting scheme is based on the residuals $\hat{\varepsilon}$ obtained after equation (1) has been estimated in the first stage with least squares (LS),[1] as we know that the firms that are highly inefficient, and thus likely to bias the results, will have a large residual $\hat{\varepsilon}$. The transformation of residuals into weights can be reflected by a weighting function $\omega(\hat{\varepsilon})$. A possible candidate for this weighting function is:

$$w = \frac{1}{\left(1 + \frac{\hat{\varepsilon}}{\sigma_{\hat{\varepsilon}}}\right)} \quad \text{if} \quad \hat{\varepsilon} > 0, \qquad \text{else} \quad w = 1 \tag{2}$$

where $\hat{\varepsilon}$ = residuals, from the former estimation; $\sigma_{\hat{\varepsilon}}$ = the standard deviation of the least squares residuals.

The residuals are divided by the standard error in order to standardise them. Eq. (2) states that observations with actual costs lower than expected costs $\left(\hat{\varepsilon} \leq 0\right)$ are assumed to be efficient ($w = 1$) and observations with actual costs higher than expected costs ($\hat{\varepsilon} > 0$) are inefficient, and the corresponding weights decline with larger residuals.

Although not strictly necessary for estimation, we should also like to impose a direct correspondence between the weights and the probability of firms being efficient. After each WLS estimation, new $\hat{\varepsilon}$s are calculated, which are then used to generate new weights, which in turn are used in a next stage WLS estimation, until the convergence criterion is met. The convergence criterion we use requires that the parameter estimates do not differ by more than 1% from the previous stage. Note that if the parameter estimates are stable or almost stable, the residuals and the corresponding weights are also stable, implying that there is no more information available in the data to identify a firm that is probably more efficient than another.

[1] If Eq. (1) is estimated with fixed effects, the weights can also be based on the fixed effects, which would make our estimator into a generalised version of the estimator, as suggested by Wagenvoort and Schure (2006).

Implementing the weights in the estimation procedure is straightforward. Instead of minimising the sum of the squared residuals, the sum of the squared weighted residuals is minimised. Observations that show large deviations from the frontier will therefore contribute less to establishing the parameters of the cost function.

A detailed technical explanation of the methodology can be found in Blank (2018).

Deriving cost efficiency

We also want to gain insight in the levels of inefficiency, rather than simply the parameters of the cost function. We therefore implement the following procedure. We assume that observations with actual costs smaller than estimated costs are efficient (observations with negative residuals): they receive an efficiency score of 1. Within this subset we can derive the variance of the residuals and regard them as an estimate of the measurement errors for the full sample. In the subsample with actual costs higher than estimated costs (residuals greater than zero) the efficiency scores are less than 1 and directly related to the value of the residual. An observation with a large residual implies low efficiency. The factor to transform the residuals into efficiency scores depends on the ratio between the variance of the residuals in the efficient subset and the variance in total sample. It makes sense that when the variance of the residuals in the efficient subset is low (i.e. the variance in the error component in the inefficient subsample is low) only a small part of the residuals can be counted for measurement errors. A large part of the residuals can then be accounted for inefficiency. Please refer to the appendix for the exact formulas and a complete theoretical derivation of the efficiency scores.

Deriving economies of scale

Economies of scale refer to the relation between resources and scale (range) of output. They indicate by which factor the costs change when there is a proportional change in all outputs. In other words, when the costs change by the same factor as the outputs, we speak of constant economies of scale. When the change is less than proportional, we speak of economies of scale. Diseconomies of scale indicate that the costs grow faster than the increased employment of resources. Economies of scale in smaller firms can be explained by increasing opportunities to redistribute labour and by making more efficient use of buildings and equipment. Diseconomies of scale in larger firms may be due to increased bureaucracy or to distractions among many more employees. Between these two extremes, we often speak of an optimal scale corresponding with a maximum benefit from the distribution of labour without the negative influences of bureaucracy.

There are different ways to evaluate economies of scale from the cost function. Here we follow the most intuitive way to get an insight in economies of scale by using the concept of average costs. As longs as economies of scale prevail then average costs will drop and as long as diseconomies of scale prevail average costs will increase. So if we are able to derive average costs then we will also have a clear picture of economies of scale. As we have multiple outputs, we cannot simply divide costs by the amount of output. Instead we define a bundle that consist of the average amount of each separate output. We put a value of one to this particular bundle. When all outputs in the bundle are doubled then the bundle will be assigned a value of two. Costs of bundles with different values can be calculated from the cost function and average cost can consecutively be computed by dividing the estimated costs by the value of the bundle. By assigning a range of different values to the bundle we will also be able to calculate a range of corresponding average costs and show the pattern related to size.

A formal way is to derive the so-called cost flexibilities or cost elasticities. For further explanation and an example see e.g. Blank and Valdmanis (2017).

Application to Dutch local administrative services

Model specification

We apply the well-known translog cost function model (Christensen et al., 1973; Christensen and Greene, 1976). In general, the model includes first- and second-order terms, as well as cross-terms between outputs and input prices on the one hand, and a time trend on the other hand. These cross-terms with a time trend represent the possible different natures of technical change. Cross-terms with outputs refer to output-biased technical change, while cross-terms with input prices refer to input-biased technical change.

In the application we cannot distinct between different input prices. We therefore discard terms with input prices. Instead the annual price changes are accounted for by deflating the costs by a general price index.

In many applications the cost function also includes terms representing so-called environmental variables controlling for differences in environmental conditions. The most illustrative example is road maintenance where maintenance costs are heavily depending on the intensity of road use and the condition of the soil (clay or sand). In this case environmental influences are very limited. Possible environmental variables are education level and age composition of the population. Lower educated or older people may face more problems in filling in request forms and therefore appeal for more assistance from local service employees. Since this only corresponds to a very small proportion of resource usage in the production process, we ignore these influences. The consequences of these assumptions are reflected in the specification in Appendix A (see Eq. (A.3)).

Data

The data for this study cover the period 2005–10. They were obtained from the municipal annual accounts at Statistics Netherlands (CBS). Annual financial and services data were collected by means of surveys covering all the local administrative services (LASs) in the Netherlands. For the purpose of this study, the data were checked for missing or unreliable data. Various consistency checks were performed on the data, in order to ensure that changes in average

Table 1 Descriptives.

	Mean	Std dev.	Minimum	Maximum
Documents (Doc.)	11,044.3	16,657.3	280	223,050
Excerpts (Exc.)	4180.0	9219.0	72	116,995
Marriages (Mar.)	150.7	228.6	10	3397
Total cost (×1000 euro)	1322.7	3540.7	15.2	67,206.5

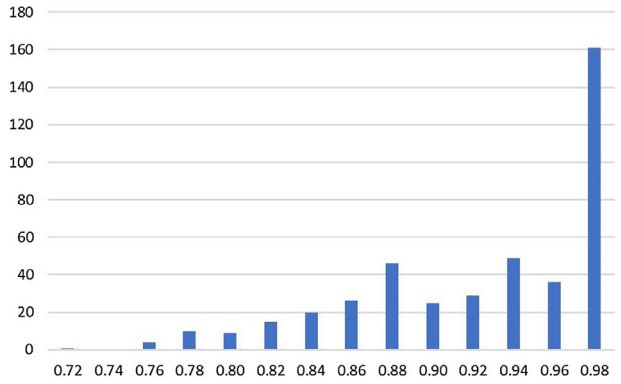

Figure 1 Distribution of cost efficiency scores, 2010.

values and in the distribution of values across time were not excessive. After eliminating observations whose dataset contained inaccurate or missing values, we had an unbalanced panel dataset of 2683 observations over the 6 years of the study. There are approximately 400 observations for each year.

As mentioned in the introduction, the main service of the LASs is the provision of passports, driving licenses, and national identity cards, as well as birth, death, and marriage certificates that are retrieved from the local registry upon the request of citizens. We define three specific outputs: the (unweighted) sum of passports, identity cards, and driving licenses; the (unweighted) number of excerpts from municipal databases (such as birth and death certificates); and the number of marriages (which is included because arranging civil marriage ceremonies is an important activity of this part of local government).

Resources include all types of staff, material supplies, and capital input. Unfortunately, the data do not allow a distinction to be made between these different resources; therefore, the total input of resources is expressed by total costs only. Since we are dealing with data from a number of years, costs are deflated by the GDP price index (for more details see van Hulst and de Groot, 2011). We do not distinguish any environmental factors in our analysis. Table 1 provides the statistical descriptives of the data.

Our pooled dataset for 2005–10 contains 2683 cases.

Estimation results and diagnostics

The model will be estimated by weighted least squares. Since we are dealing with a relatively large number of cross-sectional units (>400) and a limited number of periods (6 years), we ignore the fact that we are dealing with panel data (with respect to intra-firm correlations): the between variance is far more important than the within variance. So some of the standard errors of the estimated parameters may be slightly underestimated. We estimate the cost frontier for 2005–10, with year fixed effects to allow for an annual shift of the frontier due to technological progress or other relevant changes to the production structure.

As explained in the theoretical section, the weighting scheme is such that the weights are directly related to the efficiency scores. Efficient firms have weights equal to 1, while inefficient firms have efficiency scores equalling the weights multiplied by a constant (equal to the ratio of variances).

However, it is a simple matter to implement other weighting schemes and to see whether the results differ. As it turns out, our results were quite robust when another weighting scheme was used, based on rank numbers. In the case of IWLS estimation, we assume convergence if the maximum change in the parameters is less than 1% and the procedure stops. For convergence we needed 12 iterations in our application. So far, we have not found any problems with convergence whatsoever, which is a persistent problem in numerous SFA applications.

In order to get some insight between possible differences between SFA and IWLS we also estimated the cost function model with SFA, assuming that the efficiency component follows a half normal distribution. Both frontier methods

Table 2 Estimates of frontier cost function by SFA and IWLS.

		SFA		IWLS	
		Est.	St. err.	Est.	St. err.
2006	a_2	0.034	0.021	0.037	0.016
2007	a_3	−0.097	0.025	−0.119	0.019
2008	a_4	−0.021	0.022	−0.056	0.017
2009	a_5	0.022	0.024	−0.014	0.019
2010	a_6	0.098	0.023	0.060	0.018
Constant	a_0	−0.412	0.028	−0.362	0.015
Documents (Doc.)	b_1	0.598	0.103	0.638	0.086
Excerpts (Exc.)	b_2	0.238	0.091	0.227	0.071
Marriages (Mar.)	b_3	0.122	0.035	0.128	0.024
Doc. × Doc.	b_{11}	0.311	0.317	0.161	0.262
Doc. × Exc.	b_{12}	−0.096	0.268	−0.095	0.211
Doc. × Mar.	b_{13}	−0.120	0.085	−0.063	0.058
Exc. × Exc.	b_{22}	0.102	0.242	0.240	0.180
Exc. × Mar.	b_{23}	0.002	0.080	−0.130	0.052
Mar. × Mar.	b_{33}	0.192	0.056	0.347	0.033
Sigma	σ_ε	0.368	0.014	0.292	
σ_u/σ_v	λ	1.211	0.156	0.624	

are estimated using standard maximum likelihood and least squares methods with TSP software. Table 2 shows the estimates according to both estimation procedures.

A comparison of the outcomes of the SFA estimates and the IWLS shows that a number of the estimated parameters are very similar, in particular the parameters corresponding to the production terms in the equation (b_1, b_2 and b_3). Consequently, the calculated cost flexibilities for the average firm are almost identical ($\sum b_m = 0.96$ versus 0.99). The parameters corresponding to the cross terms may show some differences, but none of them are significantly different (b_{11}, b_{12}, b_{22}, b_{23} and b_{33}). The same holds for the trend parameters (a_2–a_6), representing the frontier shift from year to year. As expected, all the parameter estimates according to the IWLS estimation are more efficient.

In order to underline the plausibility of the estimates, we derived a few other economically relevant outcomes. The first concerns the cost efficiency scores. Fig. 1 shows the distribution of the efficiency scores in 2010.

Fig. 1 shows that in 2010, approximately one quarter of the LASs were efficient or almost efficient. Furthermore, the inefficient LASs show a plausible pattern of inefficiencies. The average efficiency is 94%, with a standard deviation of 6%. The minimum efficiency score is 69%. The efficiency scores between the years are very robust (not presented in the figure): the average efficiency scores over the years vary between 0.94 and 0.95. Comparing the IWLS efficiency scores to the SFA scores, it appears that the IWLS scores are higher. The average difference is 7 percentage points. However, this difference refers only to the absolute level of the efficiency scores. The correlation between both types of efficiency scores equals almost 100% and the rank correlation equals 98%. Further, it shows that all the SFA identified efficient firms are also IWLS efficient, and that 81% of the IWLS efficient firms are also SFA efficient.

In the theoretical section we mentioned that one of the major drawbacks of TFA is that it requires sampling from a stratified sample. Since in this procedure we do not stratify the sample at all, it is questionable whether, regardless of certain characteristics, each LAS has an equal probability of being identified as an efficient LAS. It might appear that this approach suffers from the same drawback as TFA. Characteristics that may affect the probability of being (in)efficient are the size and the year. We therefore inspected the distribution of the efficiency scores in relation to year and size. Fig. 2 shows the number of efficient LASs in each year of the sample.

Fig. 2 shows that the final selection of efficient LASs is fairly uniformly distributed over the years, varying between 116 and 124, indicating that there is an equal probability of a municipality in a certain year to belong to the frontier. This shows that the procedure does not tend to favour a particular year.

Fig. 3 shows the frequency distribution with respect to the size of the LASs (divided into four quartiles with respect to total cost).

Fig. 3 also shows that all the size categories are well represented by a substantial number of efficient LASs.

One of the restrictive assumptions in RTFA concerns the firm-specific efficiency through time. Since in our approach we allow for time varying efficiency, we are able to check this assumption. Based on the calculated total variance

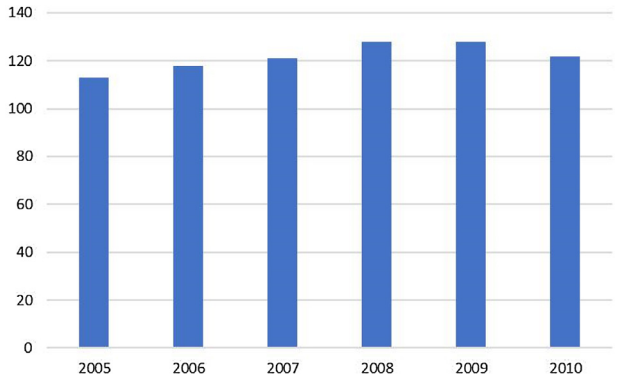

Figure 2 Number of efficient local administrative services by year.

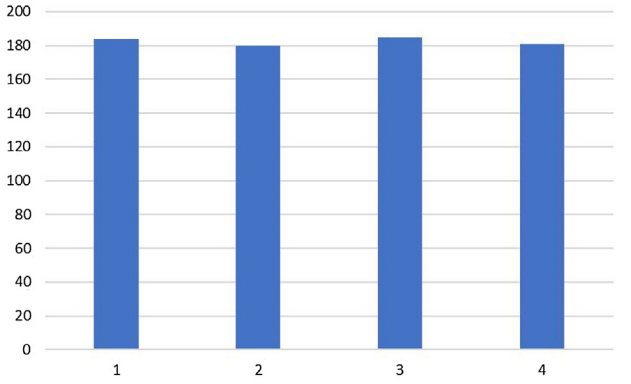

Figure 3 Number of efficient local administrative services by size, 2010.

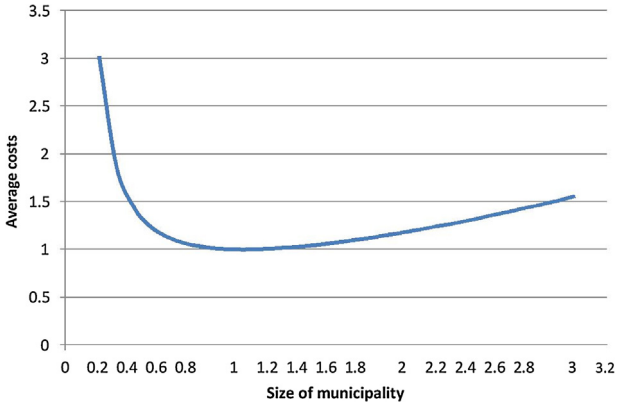

Figure 4 Relationship between municipality size and average costs.

(0.0028), between variance (0.0021), and within variance (0.0007) of the residuals, it shows that one quarter of total variance can be attributed to the within variance and three quarters to the between variance. From this we can conclude that there is some consistency in the municipality efficiency through time, but that the assumption of constant firm-specific efficiency does not hold.

Another interesting result that can be derived from these outcomes is the relationship between (municipality) scale and average costs. Fig. 4 represents the average cost and scale, expressed in an index number. Average size is

represented by 1 and the average cost is represented by 1. So, an index of 1.10 with respect to scale describes a municipality that is producing 10% more than the average municipality, whereas 1.20 with respect to average cost implies 20% higher average cost than the mean of average costs.

From Fig. 4 we see that average costs are substantial in case of a small scale. A municipality only producing 20% of the average municipality (size = 0.2) has average costs that are three times as high. The average cost graph has a typical U-shape. As scale increases average costs decline, up to a certain level. Beyond this level further scale increase would lead to an increase of average costs. So large municipalities also have high average costs.

Policy outcomes and recommendations

From the outcomes we conclude that, on average, efficiency scores are rather high, indicating that there is not much room for improvement. In our introduction we already hypothesised that this would be the case, since these services are strongly regulated due to security risks regarding identity theft and privacy concerns. The production process of the documents itself is completely centralised. The only practice variation that occurs comes from the front office, where citizens have to submit their request and can pick up their documents. Nevertheless, there are a number of municipalities operating far from best practice. They could accomplish some major efficiency gains just by comparing their business conduct with municipalities that are identified as being efficient.

The optimal scale of a municipality with respect to administrative services is about the average scale. From this perspective we may recommend that it would be wise to merge small municipalities and to split large municipalities. However, from other research, we know that the optimal scale of other local services may substantially differ from the average scale. The optimal size of the municipality for levying local taxes is about five times average (Niaounakis and Blank, 2017). So there is no such thing as one size fits all. However, it might be worthwhile investigating whether some form of collaboration between small municipalities might also lead to cost savings thanks to scale economies (depending on whether legislation allows for such a collaboration). See Niaounakis and Blank (2017) for an interesting example of successful collaboration between municipalities exploiting scale economies without merging.

A striking result concerns the productivity change through years, represented by the estimated parameters a_2–a_6. They strongly fluctuate over the years and have large standard errors, implicating that there is no general shift of best practice and no consistent trend over the years. The only reasonable explanation for this is the strong fluctuation in production levels in the course of years, not only on a macro, but also on a micro level. They are sometimes the result of the completion of new residential areas that may lead to extra registration of inhabitants and extra issuing of birth certificates. Even on a macro level, particular waves in the issuing of drivers' licenses are visible. If this explanation holds, then the measured productivity change is probably a reflection of changes in occupation rates rather than of technical change. If we add up all the productivity changes over the years ($= a_2 + \cdots + a_6$) we must conclude that overall productivity change in this period is negligible (the test that the above sum equals zero could not be rejected). In the introduction we hypothesised that technical change would be positive due the many improvements in information and communication technology. This has not been the case, which might be due to lack of incentives in this entirely monopolistic service.

Conclusions

In this paper we focus on the productivity and efficiency of Dutch local public administrative services. The main service of the LASs is the provision of passports, driving licenses, and national identity cards, as well as birth, death, and marriage certificates that are retrieved from the local registry upon the request of citizens. From a research perspective this is an interesting part of local public services, as municipalities are strongly regulated. Every citizen requesting one of these services must be served, and due to security reasons, for instance identity theft, each municipality must follow the same procedures. Furthermore, the production of services is unambiguous and good data are available. So the question is whether municipalities are still capable of improving efficiency by copying the best practice behaviour of other (efficient) municipalities. Additionally, it is to be expected that in this sector, which is dominated by administrative processes, productivity gains over time can be achieved by the use of improved information and communication technology.

This paper proposes an alternative way to derive productivity and efficiency of public services. It is stated that broadly accepted academic methodologies, such as stochastic frontier analysis, are not very attractive to policy makers and public sector managers. The methodologies are regarded as a black box, not just because of the statistics and mathematics involved but mostly because of the lack of conceptual transparency. This paper describes a method that is based on standard (weighted) regression analysis. The key notion is that some observations (the efficient ones) contain more information than others about the ''true'' frontier. If an observation is likely to contain a lot of information, it is assigned a large weight in the regression analysis. In order to establish the weights, we propose an iterative procedure. We simply repeat the regression analysis with adjusted weights in each step until a particular convergence criterion is met. If you would visualise this procedure by presenting the graph of the frontier cost function at each step, you would see that the cost function is shifting downwards to the lower region of the observations. At a certain point the graph stops moving, representing the frontier. Observations with costs lower than the frontier costs reflect measurement and specification errors. When the frontier is established, efficiency scores can be derived from the residuals.

The advantages of this approach include its high transparency. It allows the direct ascertainment of which observations largely determine the frontier. Its flexibility pertains to the use of several alternative weighting functions and the ease of testing for the sensitivity of the outcomes.

The model was applied to a set of Dutch local administrative services data that comprised 2683 observations. The outcomes are promising. The model converges quickly and

presents reliable estimates of the parameters, the cost efficiencies, and the error components. We also conducted a Stochastic Frontier Analysis on the same data set. It shows that the IWLS methodology produces comparable results to SFA.

About 25% of local administrative services are designated as efficient. The average efficiency score is approximately 93%. For the average sized LAS, no economies of scale exist.

Acknowledgements

I would like to thank Aljar Meesters for his substantial input in preliminary versions of this article. Further I would like to thank Bart van Hulst for putting the data set at my disposal. I also acknowledges Vivian Valdmanis and the referees for their valuable comments and suggestions.

Appendix A. Technical explanation and details of the methodology

As mentioned in the text we apply a cost function (Eq. (1)). Here we present some additional explanation on the estimation of (1). For an even more detailed discussion we refer to Blank (2018).

Eq. (1) can be estimated by a certain minimum distance estimator or, if one wants to check for heterogeneity, with fixed or random effects, which will result in consistent estimates of the parameters if $E[\varepsilon|y, w] = 0$. However, if some firms are inefficient – that is, they have a cost that is higher than can be explained – the cost function or random noise with $E[\varepsilon] > 0$, will cause biases in the estimated parameters of Eq. (1). In the estimation procedure we take this into account by putting less weight on the observations that are expected to be inefficient. So we can reduce these biases by estimating Eq. (1) with weighted least squares, and assigning the relatively inefficient observations a small weight and the relatively efficient observations a large weight. Since the weights are generally not observable, they have to be estimated (see e.g. Verbeek, 2017). Our proposed weighting scheme is based on the residuals obtained after Eq. (1) has been estimated in the first stage with least squares (LS),[2] as we know that firms that are highly inefficient, and thus likely to bias the results, will have a large residual $\hat{\varepsilon}$, where $\hat{\varepsilon}$ is the estimate of ε. The transformation of residuals into weights can be reflected by a weighting function $\omega(\hat{\varepsilon})$, which satisfies the requirements that it is monotonously non-decreasing in $\hat{\varepsilon}$ and always non-negative. We also impose a direct correspondence between the weights and the probability of firms being efficient. If actual cost is below estimated cost (i.e. $\hat{\varepsilon} < 0$), the firm is assumed to be efficient and the corresponding weight is set at 1. Formally, $\omega(\hat{\varepsilon}) = 1$ if $\hat{\varepsilon} < 0$. In our analysis, we use the weighting scheme according to Eq. (2).

Since the weighting scheme depends on $\hat{\varepsilon}$, which is not an independent observable variable, an iterative reweighted least squares procedure should be implemented. This procedure is used for some robust regression estimators, such as the Huber W estimator (Guitton, 2000). This similarity is not a coincidence, since our proposed estimator can also be considered a robust type of regression. This implies that, after each WLS estimation, new $\hat{\varepsilon}$s are calculated, which are then used to generate new weights, which in turn are used in a next stage WLS estimation, until the convergence criterion is met. The convergence criterion we use requires that the parameter estimates do not differ by more than 1% from the previous stage. Note that if the parameter estimates are stable or almost stable, the residuals and the corresponding weights are also stable, implying that there is no more information available in the data to identify a firm that is probably more efficient than another.

A.1. Deriving efficiency

Ondrich and Ruggiero (2001) showed that if a normal distribution is assumed to be noise, the ranking of $\hat{\varepsilon}$ is equal to the ranking of the efficiency measure μ. We use this insight in deriving efficiency scores, just by assuming the efficiency scores (u) have a relationship with the residuals ($\hat{\varepsilon}$). We apply the following procedure.

Since we have identified the cost frontier, we are able to select a subsample of efficient observations that satisfy $u = 0$, that is, all observations with an observed cost lower than or equal to frontier cost ($v \leq 0$) and thus a weight of one. This sample can be seen as the fully efficient sample, which is in accordance with Kumbhakar et al. (2013), who developed a model that allows for fully efficient firms. Note that we are not able to identify observations that satisfy $u = 0$ and $v \geq 0$, namely efficient firms with an observed cost greater than the frontier cost. We therefore assume that $|v|$ in the subsample is distributed as $N^{+}\left(0, \sigma_v^2\right)$. The variance σ_v^2 can now be estimated by the sum of squared residuals divided by the number of observations in the subsample (denoted as $\hat{\sigma}_v^2$). Furthermore, in the full sample, we assume that the subsample is representative of the variance of the random errors, and that random errors are distributed as $N\left(0, \hat{\sigma}_v^2\right)$. Since we now have an estimate of the variance of the random errors, we are also able to conditionally derive the expected efficiency from the residuals by applying, for instance, Materov's formula (Kumbhakar and Lovell, 2000, p. 78):

$$M\left(\hat{u}_i|\hat{\varepsilon}_i\right) = \hat{\varepsilon}_i\left(\frac{\hat{\sigma}_u^2}{\hat{\sigma}_\varepsilon^2}\right) \quad \text{if} \quad \hat{\varepsilon}_i \geq 0; = 0 \quad \text{otherwise} \tag{A.1}$$

with

$$\hat{\sigma}_u^2 = \hat{\sigma}_\varepsilon^2 - \hat{\sigma}_v^2$$

The efficiency score then equals:

$$Eff_i = \exp(-M\left(\hat{u}_i|\hat{\varepsilon}_i\right) \tag{A.2}$$

There are, of course, other alternatives (see e.g. Kumbhakar and Lovell, 2000). Note that in our model we have swapped the roles of the random error and efficiency components with respect to the original paper by Jondrow et al. (1982). It is important to stress that we do not apply

[2] If Eq. (1) is estimated with fixed effects, the weights can also be based on the fixed effects, which would render our estimator into a generalised version of the estimator, as suggested by Wagenvoort and Schure (2006).

the distributional assumptions a priori to the errors and efficiency components in the estimation procedure as Jondrow et al. (1982) do, but do so only in the derivation of the efficiency scores. We can also apply less complicated techniques such as corrected ordinary least squares. Further technical explanations are provided in Blank (2018).

Note that the proposed approach here shows its great advantages in the estimation procedure, and less in the derivation of efficiency scores. For the efficiency scores we still need the distributional assumptions.

A.2. Model specification

We apply the well-known translog cost function model (Christensen et al., 1973; Christensen and Greene, 1976) with some modifications due to the fact that there is only one general price index (used for deflating costs) and no environmental variables included. This leads to the following simplified form:

$$\ln(C/W) = a_0 + \sum_{m=1}^{M} b_m \ln(Y_m) + \frac{1}{2}\sum_{m=1}^{M}\sum_{m'=1}^{M} b_{mm'} \ln(Y_m)\ln(Y_{m'}) + \sum_{t=2}^{6} a_t (YR = 2004 + t) \quad (A.3)$$

where C = total costs; Y_m = output m ($m = 1, \ldots, M$); YR = year of observation; W = general price index; a_0, b_m, b_{mn}, a_t parameters to be estimated.

Symmetry is imposed by applying constraints to some of the parameters to be estimated. In formula:

$$b_{mm'} = b_{m'm}$$

References

Aigner, D., Lovell, C.A.K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production function models. J. Econometr. 6 (1), 21–37, http://dx.doi.org/10.1016/0304-4076(77)90052-5.

Bel, G., Mur, M., 2009. Intermunicipal cooperation, privatization and waste management costs: evidence from rural municipalities. Waste Manage. 29 (10), 2772–2778, http://dx.doi.org/10.1016/j.wasman.2009.06.002.

Berger, A.N., Humphrey, D.B., 1991. The dominance of inefficiencies over scale and product mix economies in banking. J. Monet. Econ. 28 (1), 117–148, Retrieved from http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VBW-45D0KVW-24&_user=499885&_coverDate=08%2F31%2F1991&_rdoc=1&_fmt=high&_orig=gateway&_origin=gateway&_sort=d&_docanchor=&view=c&_searchStrId=1719629222&_rerunOrigin=scholar.google&_acct=C00002450.

Bikker, J., van der Linde, D., 2016. Scale economies in local public administration. Local Gov. Stud. 42 (3), 441–463, http://dx.doi.org/10.1080/03003930.2016.1146139.

Blank, J.L.T., 2000. Public Provision and Performance: Contributions from Efficiency and Productivity Measurement. Elsevier, Amsterdam.

Blank, J.L.T., 2018. Iteratively Weighted Least Squares as an Alternative Frontier Methodology: Applied to the Local Administrative Public Services Industry. IPSE Studies Working Papers, Delft, Retrieved from http://www.ipsestudies.nl/research/publications/research-reports/.

Blank, J.L.T., Valdmanis, V.G., 2008. Evaluating Hospital Policy and Performance: Contributions from Hospital Policy and Productivity Research. Elsevier JAI, Oxford http://doi.org/BO0701.

Blank, J.L.T., Valdmanis, V.G., 2017. Principles of Productivity Measurement: An Elementary Introduction to Quantative Research on the Productivity, Efficiency, Effectiveness and Quality of the Public Sector, second rev. IPSE Studies, Delft.

Bottasso, A., Conti, M., Piacenz, M., Vannoni, D., 2011. The appropriateness of the poolability assumption for multiproduct technologies: evidence from the English water and sewerage utilities. Int. J. Prod. Econ. 130 (1), 112–117, http://dx.doi.org/10.1016/j.ijpe.2010.12.002.

Chen, Y.H., Lin, W.T., 2009. Analyzing the relationships between information technology, inputs substitution and national characteristics based on CES stochastic frontier production models. Int. J. Prod. Econ. 120 (2), 552–569, http://dx.doi.org/10.1016/j.ijpe.2008.07.034.

Christensen, L., Greene, W.H., 1976. Economies of scale in U.S. electric power generation. J. Polit. Econ. 84 (4), 655–676, Retrieved from http://www.jstor.org/stable/1831326.

Christensen, L.R., Jorgenson, D.W., Lau, L.J., 1973. Transcendental logarithmic production frontiers. Rev. Econ. Stat. 55 (1), 28–45, Retrieved from http://www.jstor.org/stable/1927992.

Curristine, T., Lonti, Z., Joumard, I., 2007. Improving public sector efficiency: challenges and opportunities. OECD J. Budg. 7 (1), 1–42, http://dx.doi.org/10.1787/budget-v7-art6-en.

Fried, H.O., Lovell, C.A.K., Schmidt, S.S., 2008. The Measurement of Productive Efficiency and Productivity Growth. Oxford University Press, New York.

Ganley, J.A., Cubbin, J., 1992. Public Sector Efficiency Measurement: Applications of Data Envelopment Analysis. Elsevier Science Publishers, Amsterdam.

Guitton, A., 2000. Stanford Lecture Notes on the IRLS Algorithm, Retrieved from http://sepwww.stanford.edu/public/docs/sep103/antoine2/paper_html/index.html.

Jacobs, R., Smith, P.C., Street, A., 2006. Measuring Efficiency in Health Care. Analytic Techniques and Health Policy. Cambridge University Press, Cambridge/New York, Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=econ&AN=0873679.

Jondrow, J., Lovell, C.A.K., Materov, I.S., Schmidt, P., 1982. On the estimation of technical inefficiency in the stochastic frontier production function model. J. Econom. 19 (2–3), 233–238, http://dx.doi.org/10.1016/0304-4076(82)90004-5.

Koenker, R., Hallock, K.F., 2001. Quantile regression. J. Econ. Perspect. 15 (4), 143–156, Retrieved from http://www.jstor.org/stable/2696522.

Kumbhakar, S.C., Lovell, C., 2000. Stochastic Frontier Analysis. Cambridge University Press, New York.

Kumbhakar, S.C., Parmeter, C.F., Tsionas, E.G., 2013. A zero inefficiency stochastic frontier model. J. Econom. 172 (1), 66–76, http://dx.doi.org/10.1016/j.jeconom.2012.08.021.

Levitt, M.S., Joyce, M.A.S., 1987. The Growth and Efficiency of Public Spending. Cambridge University Press, Cambridge.

Meeusen, W., Van den Broeck, J., 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. Int. Econ. Rev. 8, 435–444.

Meier, K.J., Gill, J., 2000. What Works: A Now Approach to Program and Policy Analysis. Westview Press, Boulder.

Murillo-Zamorano, L.R., Vega-Cervera, J.A., 2001. The use of parametric and non-parametric frontier methods to measure the productive efficiency in the industrial sector: a comparative study. Int. J. Prod. Econ. 69 (3), 265–275, http://dx.doi.org/10.1016/S0925-5273(00)00027-X.

Niaounakis, T.K., Blank, J.L.T., 2017. Inter-municipal cooperation, economies of scale and cost efficiency: an application of stochastic frontier analysis to Dutch municipal tax departments. Local Gov. Stud., http://dx.doi.org/10.1080/03003930.2017.1322958.

Ondrich, J., Ruggiero, J., 2001. Efficiency measurement in the stochastic frontier model. Eur. J. Oper. Res. 129 (2), 434–442, http://dx.doi.org/10.1016/s0377-2217(99)00429-4.

Parmeter, C., Kumbhakar, S., 2014. Efficiency Analysis: A Primer on Recent Advances. Miami, New York.

Pérez-López, G., Prior, D., Zafra-Gómez, J.L., Plata-Díaz, A.M., 2016. Cost efficiency in municipal solid waste service delivery: alternative management forms in relation to local population size. Eur. J. Oper. Res. 255, 583–592, http://dx.doi.org/10.1016/j.ejor.2016.05.034.

Shao, B.B.M., Lin, W.T., 2016. Assessing output performance of information technology service industries: productivity, innovation and catch-up. Int. J. Prod. Econ. 172, 43–53, http://dx.doi.org/10.1016/j.ijpe.2015.10.026.

van Hulst, B.L., de Groot, H., (IPSE Studies Research Reeks No. 2011-7) 2011. Benchmark burgerzaken. Een empirisch onderzoek naar de kostendoelmatigheid van burgerzaken. IPSE Studies, Delft.

Veenstra, J., Koolma, H.M., Allers, M.A., 2016. Scale, mergers and efficiency: the case of Dutch housing corporations. J. Hous. Built Environ., 1–25, http://dx.doi.org/10.1007/s10901-016-9515-4.

Verbeek, M., 2017. A Guide to Modern Econometrics, 5th ed. John Wiley and Sons, Hoboken, NJ.

Wagenvoort, R.J.L.M., Schure, P.H., 2006. A recursive thick frontier approach to estimating production efficiency*. Oxf. Bull. Econ. Stat. 68 (2), 183–201, http://dx.doi.org/10.1111/j.1468-0084.2006.00158.x.

Zafra-Gómez, J.L., Prior, D., Díaz, A.M.P., López-Hernández, A.M., 2013. Reducing costs in times of crisis: delivery forms in small and medium sized local governments' waste management services. Public Adm. 91 (1), 51–68, http://dx.doi.org/10.1111/j.1467-9299.2011.02012.x.