Empirical Economics (2021) 60:2665–2700
https://doi.org/10.1007/s00181-021-02026-2

Benchmarking the performance of US Municipalities

Caitlin T. O'Loughlin[1] · **Paul W. Wilson**[2]

Received: 25 February 2020 / Accepted: 22 January 2021 / Published online: 10 February 2021

Abstract

This paper examines the performance of US municipal governments over 1997–2012, and hence prior to, during and following the financial crisis of 2007–2008. Fully nonparametric methods are employed to estimate technical efficiencies of cities utilizing recently developed statistical results. The results strongly suggest non-convexity of the local governments' production set, calling into question the results of previous studies examining municipal efficiency that do not allow for non-convexity. We find strong evidence that production sets for municipal governments are different across time and across regions of the USA. Overall, we find that municipalities in the Midwest and South on average out-performed those in the Northeast and West in terms of both efficiency and productivity, and both before and after the financial crisis.

Keywords Cities · Efficiency · Productivity · Nonparametric

1 Introduction

Municipal governments provide varying bundles of local public goods in terms of services and amenities for residents. In addition to police and fire protection, municipal governments may provide roads and streets, traffic management, trash collection, street cleaning, water services, libraries and other services. However, while different municipal governments provide different levels of services, some are more adept at providing amenities while others are relatively inept. These differences are often reflected in lists of "best places to live" that appear in the popular press.

The views expressed in this paper do not necessarily reflect those of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. We thank the Cyber Infrastructure Technology Integration group at Clemson University for operating the Palmetto cluster used for our computations. We are grateful for comments from Kathy Hayes, Shawna Grosskopf and participants at the XVI European Workshop on Efficiency and Productivity Analysis held in London 10–13 June 2019. Of course, any remaining errors are our fault.

✉ Paul W. Wilson
pww@clemson.edu

Extended author information available on the last page of the article

The idea that municipal governments should—but sometimes do not—provide services efficiently is an old one. Bruère et al. (1912) argue that "the efficiency movement in cities grew out of recognition of the dependence of community welfare upon government activity" beginning in 1906, and that the efficiency movement "aims to remove city government from its isolation, and to make it the customary and accepted common agency for 'getting things done'." Tiebout (1956) suggests that households sort into different jurisdictions based on their preferences for bundles of public goods and services offered by different local governments, and argues that competition among municipalities to attract residents increases overall efficiency, and these efficiency gains are often capitalized in local property values.

Empirical support for the link between competition among municipalities and efficiency in local public good provision is provided by Hayes et al. (1998), Grossman et al. (1999) and others who find evidence that competition among local governments tends to enhance efficiency. Davis and Hayes (1993) argue that citizens are more likely to closely monitor local governments where taxes are high, and that housing owners are more likely to watch closely than renters because they have a larger stake in outcomes. Davis and Hayes (1993), Grossman and West (1994), Hayes and Wood (1995) and Hayes et al. (1998) find evidence that increased monitoring (proxied by tax rates or the degree of centralization by local governments) is associated with smaller or more efficient local governments. Grosskopf et al. (2001) note that competition among municipalities may create incentives to provide services efficiently by influencing citizens' willingness to pay for public services or their inclination to remain in the jurisdiction, and suggest that monitoring by voters may encourage efficiency among municipalities. Government officials may increase their probability of remaining in office by running local governments efficiently, particularly where residents are personally affected by local policy.

On the other hand, a number of factors may work against efficient provision of local public goods. Friction caused by real estate transaction fees, costs of commuting and job search and other factors may reduce competition among municipalities, perhaps contributing to inefficient provision of services. Fiscal pressure due to pension obligations and other factors in some cities were exacerbated by the financial crisis of 2007–2008, and many municipalities struggled to meet expenses in the face of falling tax revenues during the crisis. In the aftermath of the financial crisis, several US cities filed for Chapter 9 bankruptcy, including Vallejo, CA in 2008; Harrisburg, PA in 2011; Central Falls, RI in 2011; Stockton, CA in 2012; San Bernardino, CA in 2012; Detroit, MI in 2013; and Hillview, KY in 2015.[1] In addition, municipal governments provide a classic example of the principal–agent problem. Residents pay taxes and consume services, but must delegate management of municipal governments to politicians, bureaucrats and functionaries. The potential for rent-seeking behavior is ever-present since these functionaries may have their own agendas and interests. Moreover, claimants on local government tax revenues are typically ill-specified, creating further potential for waste (i.e., inefficiency) due to rent-seeking behavior which may

[1] The Harrisburg case was subsequently dismissed.

in turn affect cities' ability to provide local public goods efficiently.[2] To the extent that local public goods are capitalized into housing prices, cities that are worse (or, conversely, better) in providing services and amenities have a negative (or positive) impact on the wealth of their residents.

While there are a number of factors that might lead to inefficiency in provision of local public goods, the question of how large this inefficiency might be, how it might vary across regions and how it might have been affected by the financial crisis of 2007–2008, remains. At present, local governments are facing similar pressures due to disruptions of business activity (and tax revenue) caused by the on-going COVID-19 pandemic. We use new, recently developed statistical methods to examine the efficiency of local-public good provision by US municipal governments in 1997, 2002, 2007 and 2012, i.e., before, during and after the 2007–2008 financial crisis. As such, our results may provide some insight into how well municipalities may be likely to weather the ongoing pandemic and its effects.

Necessarily, not all questions can be answered in a single paper, and this paper is not different from others in this regard. But, our paper contributes to the existing literature by employing a rigorous statistical framework that distinguishes our paper from many that examine municipal efficiency. Specifically, we use nonparametric estimators and inference to examine not only the level of inefficiency in each year of our data, but how this varies over time and over regions. In addition, we investigate the *productivity* of municipal governments, and how their productivity has changed over time. Our use of rigorous statistical methods combined with nonparametric estimation stands in contrast to virtually all of the previous literature on efficiency of local-public good provision. While many papers (e.g., Charnes et al. 1989; Chalos and Cherian 1995; De Sousa and Stošić 2005; Fang et al. 2013) use nonparametric estimators in such studies, the choice of the *particular* nonparametric estimator is often ad hoc, and few if any provide statistical inference, presenting instead only point estimates or perhaps sample means of point estimates. The results of some of our hypothesis tests provide useful guidelines for future research on municipal government efficiency.

The existing literature on efficiency of local public good provision is extensive. In addition to the papers cited above, see Tang (1997), De Borger and Kerstens (2000), Afonso (2008), Da Cruz and Marques (2014), de Oliveira Junqueira (2015) and Narbón-Perpiñá and De Witte (2018) for reviews of this literature. Previous empirical analyses of municipal efficiency can be broadly divided into two categories, i.e., those that employ fully parametric methods along the lines of Aigner et al. (1977) and Meeusen and van den Broeck (1977) versus those that use fully nonparametric envelopment estimators such as the free disposal hull (FDH) estimators proposed by Deprins et al. (1984) or data envelopment analysis (DEA) estimators proposed by Farrell (1957) and made popular by Charnes et al. (1978) and Banker et al. (1984). Among studies of municipal efficiency, nonparametric methods are used more frequently than parametric methods. Among those that use nonparametric methods, DEA estimators (which require convexity of the production set) are used far more frequently than FDH estimators (which do not require convexity). For example, the review by Narbón-Perpiñá and

[2] See Lee and Wilson (1990, 1991) for theoretical analysis of the economic loss created by this type of rent-seeking behavior.

De Witte (2018, Table A2) lists 97 empirical studies of local government efficiency. Among these, 66 use nonparametric estimators, while only 31 use parametric methods. Among the 66 papers employing nonparametric estimators, 50 use DEA estimators, 14 use FDH estimators, and 2 use both. As discussed below, the choice between FDH and DEA estimators is not innocuous—DEA estimators are not statistically consistent if the production set is not convex, while FDH estimators are consistent regardless of whether the production set is convex.[3] As discussed below, our results, specifically our results from several hypothesis tests, provide useful suggestions for how empirical researchers should use nonparametric methods in the future.

The remainder of our paper unfolds as follows. Section 2 presents the nonparametric methods we use for estimation and inference. We discuss the underlying statistical model and the various nonparametric efficiency estimators and their properties in Sect. 2.1, and in Sect. 2.2 we discuss methods for hypothesis testing in the context of our nonparametric model. The data used for estimation and inference are discussed in Sect. 3, and empirical results are presented in Sect. 4. Summary and conclusions are given in Sect. 5.

2 Methods for estimation and inference

2.1 Nonparametric estimators and their properties

To establish notation, let $X \in \mathbb{R}_+^p$ and $Y \in \mathbb{R}_+^q$ denote random vectors of input and output quantities, and similarly let $x \in \mathbb{R}_+^p$ and $y \in \mathbb{R}_+^q$ denote corresponding fixed, nonstochastic vectors of input and output quantities. The (unconditional) production set is the set of feasible combinations of input and output quantities, i.e.,

$$\Psi := \{(x, y) \mid x \text{ can produce } y\}, \tag{1}$$

which gives the set of possible inputs and outputs. We adopt standard assumptions from microeconomic theory of the firm (e.g., see Shephard 1970 or Färe 1988). We assume (i) Ψ is closed; (ii) any nonzero production requires the use of some inputs (i.e., $(x, y) \notin \Psi$ if $x = 0,\ y \geq 0,\ \ y \neq 0$); and (iii) both inputs and outputs are freely disposable so that (i) $\widetilde{x} \geq x \Rightarrow (\widetilde{x}, y) \in \Psi$ and (ii) $\widetilde{y} \leq y \Rightarrow (x, \widetilde{y}) \in \Psi$. $\forall\ (x, y) \in \Psi$.

[3] The studies that employ parametric methods typically specify translog response functions. Among the 31 papers listed by Narbón-Perpiñá and De Witte (2018, Table A2) that employ parametric methods to assess local government performance, 25 use a translog specification, and one (Nikolov and Hrovatin 2013) uses a Cobb–Douglas specification (which of course is nested by the translog specification). However, municipal governments vary widely in terms of size, and several studies have noted that the parameters of a translog function are unlikely to be stable when the function is fit globally across units of widely varying size. See, for example, Guilkey et al. (1983) and Chalfant and Gallant (1985) for Monte Carlo evidence, and Cooper and McLaren (1996) and Banks et al. (1997) for empirical evidence involving consumer demand, Wilson and Carey (2004) for empirical evidence involving hospitals, and McAllister and McManus (1993), Mitchell and Onvural (1996), and Wheelock and Wilson (2001, 2012, 2018) for empirical evidence involving banks.

The assumption that Ψ is closed ensures that the *efficient frontier* or *technology*

$$\Psi^{\partial} := \left\{(x, y) \mid (x, y) \in \Psi,\ (\gamma^{-1}x, \gamma y) \notin \Psi \text{ for any } \gamma \in (1, \infty)\right\} \tag{2}$$

consisting of the set of extreme points of Ψ is contained in Ψ.[4] The requirement that some strictly positive amount of input must be used to produce any output greater than zero rules out the existence of free lunches. Free disposability of inputs and outputs implies weak monotonicity of the frontier.

The frontier Ψ^{∂} provides a benchmark against which production units' performance can be measured. Units operating on the frontier are said to be *technically efficiency*, while those operating under the frontier, in the interior of Ψ, are said to be *technically inefficient*. Several measures of technical efficiency are employed in the literature. The Farrell (1957) input efficiency measure

$$\theta(x, y \mid \Psi) := \inf\{\theta \mid (\theta x, y) \in \Psi\} \tag{3}$$

gives the proportion by which input levels can be feasibly reduced without reducing output levels. Alternatively, the Farrell (1957) output efficiency measure

$$\lambda(x, y \mid \Psi) := \sup\{\lambda \mid (x, \lambda y) \in \Psi\} \tag{4}$$

gives the feasible proportion by which output levels can be increased without increasing input quantities. The hyperbolic measure

$$\gamma(x, y \mid \Psi) := \inf\left\{\gamma > 0 \mid (\gamma x, \gamma^{-1}y) \in \Psi\right\} \tag{5}$$

proposed by Färe et al. (1985) gives the proportion by which output quantities can be increased while simultaneously reducing input quantities by the same proportion. Clearly, $\theta(x, y \mid \Psi) \leq 1$, $\lambda(x, y \mid \Psi) \geq 1$, and $\gamma(x, y \mid \Psi) \leq 1$ for all $(x, y) \in \Psi$, with strict equality holding for any $(x, y) \in \Psi^{\partial}$.[5]

It is important to note that Ψ, and hence Ψ^{∂} as well as the measures $\theta(x, y \mid \Psi)$, $\lambda(x, y \mid \Psi)$ and $\gamma(x, y \mid \Psi)$ are unobserved. Consequently, they must be estimated from a sample $\mathcal{S}_n = \{(X_i, Y_i)\}_{i=1}^{n}$ of observed input–output pairs. There are several possibilities. First, Deprins et al. (1984) estimate Ψ by the free disposal hull of the sample observations in $\mathcal{S}_n$, i.e.,

$$\widehat{\Psi}_{\text{FDH},n} := \bigcup_{(X_i, Y_i) \in \mathcal{S}_n} \left\{(x, y) \in \mathbb{R}_{+}^{p+q} \mid x \geq X_i,\ y \leq Y_i\right\}. \tag{6}$$

[4] Inequalities involving vectors are defined on an element-by-element basis.

[5] In principle, for some $(x, y) \in \Psi^{\partial}$ one might have $\theta(x, y \mid \Psi) < 1$ or $\lambda(x, y \mid \Psi) > 1$ if the frontier is parallel to either all of the input axes or all of the output axes in some regions. However, this is ruled out by additional assumptions required to define a statistical model in which the efficiency measure defined in (3)–(5) are statistically efficiency; see Kneip et al. (1998) and Kneip et al. (2008) for details.

This estimator does not impose convexity on Ψ. Substituting $\widehat{\Psi}_{\text{FDH},n}$ for Ψ in (3)–(5) provides FDH estimators of the input, output and hyperbolic efficiency measures.

The resulting FDH efficiency estimators can be computed easily by letting let $\mathcal{D}_{x,y}$ denote the set indices of points in $\mathcal{S}_n$ dominating (x, y), i.e., $\mathcal{D}_{x,y} = \{i \mid (X_i, Y_i) \in \mathcal{S}_n,\ X_i \leq x,\ Y_i \geq y\}$. Then, the input-oriented FDH efficiency estimator is computed by solving

$$\widehat{\theta}_{\text{FDH}}(x, y \mid \mathcal{S}_n) = \min_{i \in \mathcal{D}_{x,y}} \max_{j=1,\ldots,p} \left(\frac{X_i^j}{x^j} \right), \tag{7}$$

where for a vector a, a^j denotes its j-th component, and the output-oriented FDH estimator can be computed by solving

$$\lambda_{\text{FDH}}(x, y \mid \mathcal{S}_n) = \max_{i \in \mathcal{D}(x,y)} \min_{j=1,\ldots,q} \left(\frac{Y_i^j}{y^j} \right). \tag{8}$$

The hyperbolic FDH estimator can be computed by solving

$$\widehat{\gamma}_{\text{FDH}}(x, y \mid \mathcal{S}_n) = \min_{i=1,\ \ldots,\ n} \left(\max_{\substack{j=1,\ \ldots,\ p \\ k=1,\ \ldots,\ q}} \left(\frac{X_i^j}{x^j}, \frac{y^k}{Y_i^k} \right) \right) \tag{9}$$

as shown by Wilson (2011).

Second, the variable returns to scale (VRS) version of the DEA estimator $\widehat{\Psi}_{\text{VRS},n}$ of Ψ is given by the convex hull of $\widehat{\Psi}_{\text{FDH},n}$. Substituting this for Ψ in (3)–(4) gives the familiar VRS-DEA estimators $\widehat{\theta}(x, y \mid \mathcal{S}_n)$ and $\widehat{\lambda}(x, y \mid \mathcal{S}_n)$ of $\theta(x, y \mid \mathcal{S}_n)$ and $\lambda(x, y \mid \mathcal{S}_n)$, respectively, and these estimators can be computed by solving the resulting linear programs. Substituting the convex hull of $\widehat{\Psi}_{\text{FDH},n}$ for Ψ in (5) gives the hyperbolic VRS-DEA estimator $\widehat{\gamma}(x, y \mid \mathcal{S}_n)$ of $\gamma(x, y \mid \mathcal{S}_n)$, which can be computed using numerical methods described by Wilson (2011). Finally, constant returns to scale (CRS) versions of DEA estimators of $\theta(x, y \mid \mathcal{S}_n)$, $\lambda(x, y \mid \mathcal{S}_n)$ and $\gamma(x, y \mid \mathcal{S}_n)$ are obtained by substituting the conical hull of $\widehat{\Psi}_{\text{VRS},n}$, denoted by $\widehat{\Psi}_{\text{CRS},n}$, for Ψ in (3)–(5), and these CRS-DEA estimators can be computed by solving the resulting linear programs. See Simar and Wilson (2013, 2015) for discussion and details.

Thus, there are three categories of nonparametric efficiency estimators: FDH, VRS-DEA and CRS-DEA. For each of these categories, there are estimators for efficiency measures in input, output and hyperbolic directions. Of course, estimation reveals nothing; learning from data can occur only when statistical inference is made, and inference requires a statistical model as made clear by the results of Bahadur and Savage (1956). The appropriate statistical model consists of the assumptions on Ψ described above as well as a number of assumptions on the statistical process (i.e., the data-generating process) that yields observed data. Among these assumptions, we require that observed data on inputs and outputs consist of identically, independently distributed (iid) realizations of random variables $(X, Y) \in \mathbb{R}_+^{p+q}$ with bounded support on Ψ. We require also that the joint density of (X, Y) must be strictly positive along the frontier Ψ^{∂} and continuously differentiable in a neighborhood of Ψ^{∂}. We also require that the frontier Ψ^{∂} be sufficiently smooth, depending on the class of estimators—FDH, VRS-DEA or CRS-DEA—being used. Technical details are given by Kneip et al. (2008, 2015).

The statistical properties of the nonparametric efficiency estimators are developed in a number of papers. In all cases, the estimators converge at rate n^{κ} and have limiting distributions. For the case of FDH estimators, Park et al. (2000) and Daouia et al. (2017) establish that $\kappa = 1/(p+q)$, and the limiting distributions are Weibull, but with parameters that depend on features of the model being estimated (e.g., curvature of Ψ^{∂}) and which are difficult to estimate. These results hold regardless of whether Ψ is convex. For the case of VRS-DEA estimators, under VRS and convexity of Ψ, Kneip et al. (1998) prove that $\kappa = 2/(p+q+1)$ and Kneip et al. (2008) establish existence of and characterize limiting distributions. However, the limiting distributions do not have explicit, analytical expressions. Park et al. (2010) establish corresponding results for CRS-DEA estimators under CRS and convexity of Ψ, where $\kappa = 2/(p+q)$. Kneip et al. (2016) prove that under CRS and convexity of Ψ, the VRS-DEA estimators converge at the faster rate achieved by CRS-DEA estimators under the same conditions. The existence of limiting distributions is needed to ensure validity of bootstrap methods developed by Kneip et al. (2008, 2011) and Simar and Wilson (2011a) for estimating confidence intervals for the efficiency of individual producers.

2.2 Testing hypotheses about model structure

By construction, the FDH, VRS-DEA and CRS-DEA efficiency estimators are biased. This is due to the fact that $\widehat{\Psi}_{\text{FDH},n} \subseteq \Psi$ and $\widehat{\Psi}_{\text{FDH},n} \subseteq \widehat{\Psi}_{\text{VRS},n} \subseteq \Psi$ when Ψ is convex. If Ψ is convex and CRS prevail, then $\widehat{\Psi}_{\text{FDH},n} \subseteq \widehat{\Psi}_{\text{VRS},n} \subseteq \widehat{\Psi}_{\text{CRS},n} \subseteq \Psi$. Kneip et al. (2015) provide results on moments of the input-oriented efficiency estimators, and the results extend trivially to the output-oriented estimators. Kneip et al. (2020) extend these results to the hyperbolic VRS efficiency estimator, and Wilson (2021) extends the results to the hyperbolic FDH estimator. In all cases, the bias of the estimators disappears at the same rate at which the estimators converge. Consequently, for a convergence rate of n^{κ}—where $\kappa = 1/(p+q)$ for the FDH estimators, $\kappa = 2/(p+q+1)$ for the VRS estimators, or $\kappa = 2/(p+q)$ for the CRS estimators—standard central limit theorem (CLT) results (e.g., the Lindeberg–Feller CLT) do not hold for mean efficiency unless $\kappa > 1/2$. In the CRS case, this means that the usual CLTs hold only if $(p+q) < 4$. In the VRS case, the usual CLTs hold only if $(p+q) < 3$. In the FDH case, standard CLTs hold only if $p+q < 2$.[6] Consequently, standard CLT results can be used with sample means of estimated efficiencies to make inference about expected efficiency only if the number of inputs is smaller than 3 or 4 when VRS-DEA or CRS-DEA estimators are used, and can *never* be used in any realistic setting when FDH estimators are used. When $\kappa = 1/2$, sample means of estimated efficiencies have constant, nonzero bias, and when $\kappa < 1/2$, this bias is no longer constant, but explodes toward infinity as sample size increases. Kneip et al.

[6] In other words, standard CLT results hold in the FDH case if and only if $p = 1$ and output is fixed and constant, or $q = 1$ and input is fixed and constant.

(2015) develop a set of new CLTs that incorporate generalized jackknife estimates of bias, and which involve subsampling when κ is too small. See Kneip et al. (2015, 2020) and Wilson (2021) for additional discussion and technical details.

Kneip et al. (2016) use the results of Kneip et al. (2015) to develop (i) tests of differences in mean efficiency across groups of producers, where the two groups face either the same or different frontiers, (ii) a test of convexity versus non-convexity of Ψ and (iii) a test of CRS versus VRS (provided Ψ is convex). Tests (ii) and (iii) involve comparing differences in sample means of different classes of efficiency estimators. For example, to test convexity of Ψ, a test statistic based on the difference between a sample mean of FDH estimates and a sample mean of VRS-DEA estimates is used. The test of CRS versus VRS involves a similar difference of sample means of VRS-DEA and CRS-DEA estimates. In both cases, the underlying theory requires independence of the two sample means used to build the test statistics, and this in turn requires randomly splitting the original sample into two independent subsamples. Details are given in Kneip et al. (2016).

The test statistics developed by Kneip et al. (2016) are asymptotically normally distributed. The tests are valid for *any* random split of the original sample into two subsamples, but repeated splitting of a given sample can yield different results, i.e., some splits may result in a value of the test statistic that leads to rejection of the null hypothesis, while another split may yield a value that fails to reject the null. This is not surprising when one considers that in any statistical test, the p value is a random variable with a non-degenerate distribution in any interesting case. Unfortunately, one cannot simply average test statistics across several splits of the same sample, as the resulting statistics cannot be independent which precludes any inference since the dependence is of a complicated form.

Simar and Wilson (2020a) solve this problem by developing a bootstrap method to remove (most of) the randomness due to the sample splitting. The idea is to split the sample m times, and then compute the sample average of a given test statistic over the m splits. The bootstrap proposed by Simar and Wilson (2020a) can be used to obtain an appropriate critical value, or to estimate a p value, for this average of the m individual test statistics. One can also save the p values corresponding obtained for the particular test statistic on each sample split, resulting in m p values, and then compute the corresponding Kolmogorov–Smirnov (KS) one-sample test statistic to compare the distribution of the p values against the uniform distribution on $[0, 1]$ (which is necessarily the distribution of the p values under the null hypothesis due to the probability integral transform as discussed by Simar and Wilson 2020a). Since the m p values used to construct the KS statistic are not independent, the resulting KS statistic does not have its usual distribution, but its distribution can be estimated by the bootstrap developed by Simar and Wilson (2020a). Either approach can be used to eliminate much of the randomness introduced by random sample splits required by the tests developed by Kneip et al. (2016), and extensive simulation results in Simar and Wilson (2020a) indicate that the method yields good size and power properties. Additional simulation experiments indicate that $m = 100$ splits typically eliminates most of the uncertainty surrounding the random splitting. Below, in all of the tests where we rely on random sample splits, we use $m = 1,000$ splits, and none of our qualitative results are changed from when we use only $m = 100$ splits.

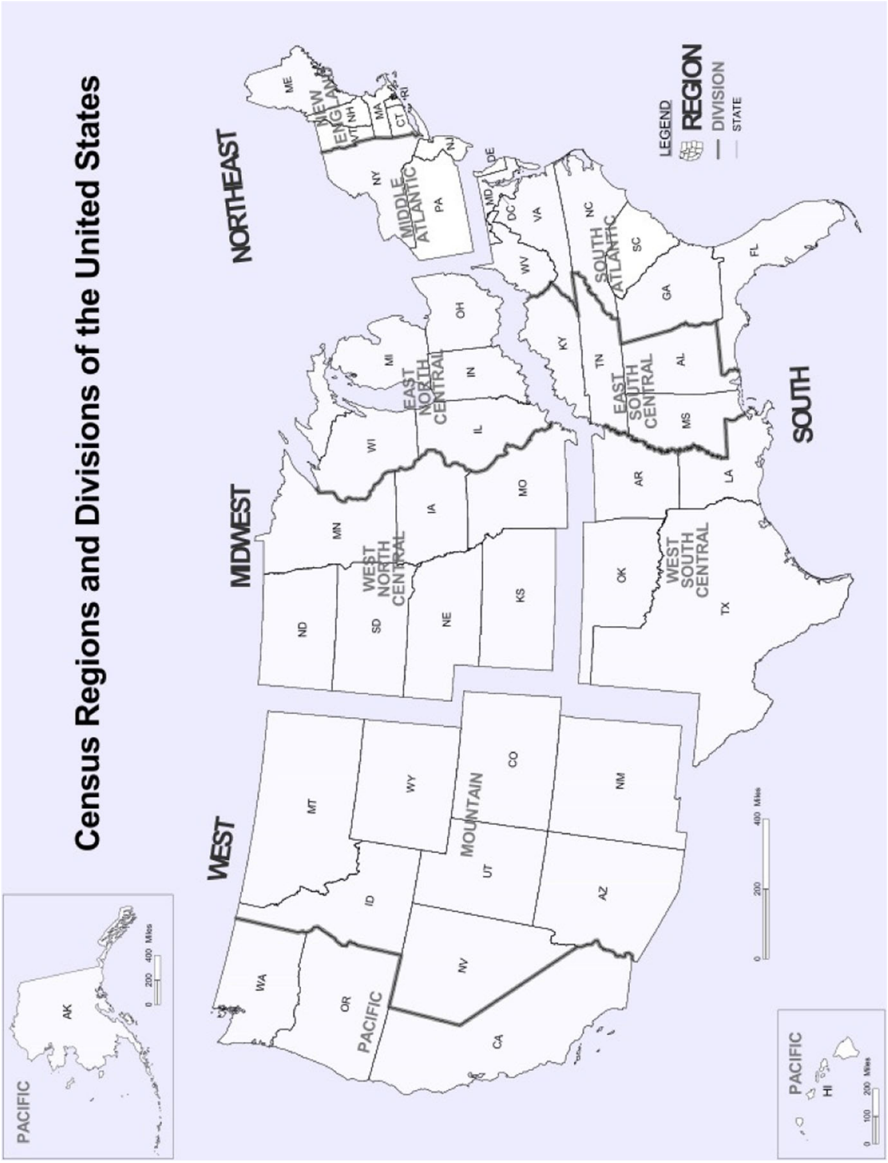

Fig. 1 Census regions of the USA

In our data on municipalities, we have in addition to input and output variables two additional variables to consider, i.e., time (or year) and US census region. These are both discrete, categorical variables, and in both cases, there are four categories (i.e., we have data for 4 years, and there are four census regions).[7] Denote these

[7] We include municipalities in the 48 states excluding Alaska and Hawaii. The census regions are illustrated in Fig. 1. We denote the Northeast, Midwest, South and West regions as regions 1, 2, 3 and 4, respectively.

"environmental" variables by Z, a bivariate, discrete random variable with support on the Cartesian product of $\{1,\ 2,\ 3,\ 4\}$ with itself, which we denote by $\mathcal{Z}$. Hence, we have $4^2 = 16$ categories. Let

$$\Psi^z = \{(x, y) \mid x \text{ can produce } y \text{ when } Z = z\} \tag{10}$$

denote the *conditional* (on Z) production set. This is related to the *unconditional* production set defined in (1) in the sense that

$$\bigcup_{z \in \mathcal{Z}} \Psi^z = \Psi. \tag{11}$$

Two mutually exclusive, collectively exhaustive possibilities exist. Necessarily, either

$$\Psi^z = \Psi \text{ for all } z \in \mathcal{Z}, \tag{12}$$

or

$$\Psi^z \subset \Psi \text{ for some } z \in \mathcal{Z}, \tag{13}$$

but obviously one and only one of these two conditions can hold. Possibility (12) is the well-known "separability" condition described by Simar and Wilson (2007, 2011b). In this case, the environmental variables in Z can be ignored when estimating efficiency, and the efficiency estimators described above can be used. But, (12) is a strong condition, and as noted by Simar and Wilson (2007), one should test this rather than merely assuming that the separability condition holds. If (12) does not hold, then (13) must hold. In this case, *conditional* (on Z) efficiency estimators must be used to estimate efficiency, as the unconditional efficiency estimators described above estimate distance to a frontier that is unattainable in some cases whenever (13) holds.

Conditional efficiency estimators were first described by Cazals et al. (2002) and extended by Daraio and Simar (2007a, b). The idea involves localizing the FDH, VRS-DEA or CRS-DEA estimator in terms of Z. When Z is continuous, this involves bandwidth parameters, but when Z is discrete, no bandwidths are needed; one simply divides the sample into categories defined by observed values of Z, and estimates efficiency independently for observations in each category. For example, if there are only two categories, one would estimate efficiency for each observation in the first category using only the data for that category, and then similarly estimate efficiency for each observation in the second category, using only the data in the second category.

Daraio et al. (2018) develop CLTs for conditional efficiency estimators, and use the new CLTs to develop tests of the null hypothesis of separability given by (12) versus the alternative hypothesis of non-separability given by (13). Similar to the tests of convexity of Ψ versus non-convexity and CRS versus VRS developed by Kneip et al. (2016), the tests of separability developed by Daraio et al. (2018) involve comparing sample means of unconditional and conditional efficiency estimators, and depend on independence between the two sample means. This in turn requires randomly splitting the original sample into two independent subsamples, but the bootstrap method of Simar and Wilson (2020a) can be used to remove the ambiguity or uncertainty resulting from a single split of the original sample. In our tests below, we again use $m = 1,000$ sample splits. See Daraio et al. (2018) and the accompanying appendices for specific details.

3 Data and variable specification

Our data are taken from the Annual Survey of State and Local Government Finances, the Annual U.S. Building and Permit Survey, the U.S. Census of Governments, the U.S. Bureau of Labor Statistics (BLS) and the Federal Bureau of Investigation (FBI) to define input and output variables. Our variable specifications are broadly similar to those use in other studies of local governments' efficiency levels. All dollar amounts are measured in terms of thousands of constant, 2010 US dollars. We use the US consumer price index (CPI) for all urban consumers and all items to adjust for price differences across municipalities.[8]

Empirical assessments of municipal efficiency in the literature typically take one of two approaches, either (i) examining overall efficiency or (ii) focusing on the efficiency with which municipalities provide a specific good or service. We take the first route in order to learn about the overall efficiency of municipal governments in providing an array of services. The majority of local government efficiency studies employ a single input variable to account for resources used to produce goods and services. Total current operating expenditures is the most widely specified input; examples include Štastná and Gregor (2015) and Radulović and Dragutinović (2015). Alternatively, a few studies have specified input as total expenditures (e.g., Hayes and Chang 1990) or financial expenditures (e.g., De Borger and Kerstens 1996). We adopt the former, more typical approach and specify a single input (denoted by X, and hence $p = 1$) consisting of total current operating expenditures reported in the Annual Survey of State and Local Government Finances. This survey, conducted by the U.S. Census Bureau, is the only comprehensive source of data on local government finances collected on a nationwide scale using uniform definitions, concepts and procedures. (U.S. Census Bureau 2020). Data are collected annually from a sample of state and local governments, but every 5 years, in years ending in "2" or "7," a census of all state and local governments is conducted. Alternatively, one might use administrative data which might provide more detail on expenditures for various items, but such data are difficult to collect, and in the end would reflect different definitions and accounting systems used by municipal governments; in addition, this would likely result in a smaller sample size.

In recent years, local governments have seen rising costs, exceeding the rate of increasing costs in the private sector, which is likely reflected in our input measure. Berry and Lowery (1984) speculate that Baumol's "cost disease" may be driving this difference, since many local public goods are labor-intensive or hand-produced. Moreover, there has been an increasing level of concentration of government provision of goods and services at the federal level. Baicker et al. (2012) suggest this may be due to the growing importance of certain budget components including education, health

[8] Problems with the CPI are well known, e.g., the CPI covers a small number of items compared to the gross domestic product deflator, but this is not available at local levels. Where a city-specific CPI is not available, we use the CPI for the region in which a given city is located.

and welfare programs. By contrast, total and financial expenditures (as opposed to operating expenditures, which we use) include expenditures on total capital outlays, which is susceptible to volatility due to the nature of government spending. In this sense, our variable captures short-run operating costs.

We specify $q = 6$ output variables, including total population (Y_1), total charges for sewerage and waste management (Y_2), the reciprocal of the total crime rate (Y_3), total land area (in square miles) (Y_4), total building permits (Y_5) and the employment rate (Y_6). Our specification of outputs reflects the wide variety of goods and services provided by municipal governments. Total population is one of the most frequently specified outputs in the literature on local government efficiency (e.g., see De Borger et al. 1994; Kalseth and Rattsø 1995; Athanassopoulos and Triantis 1998; Grossman et al. 1999; Geys and Moesen 2009; lo Storto 2013) and serves as a proxy for the scope of demand for publicly provided goods and services. We use total charges for sewerage and waste removal or treatment to account for communal service administration. Worthington (2000) utilizes a similar output measure related to the municipal sewerage system and waste collection, while Balaguer-Coll et al. (2007) use the number of street lights.[9] Data for Y_1 and Y_2 are obtained from the Annual Survey of State and Local Government Finance.

We use the reciprocal of the total municipal crime rate, Y_3, to capture the degree of public safety provided through law enforcement services. These data are obtained from the FBI's Uniform Crime Reporting Statistics which are voluntarily reported by police departments at the local level. The FBI defines total violent crimes to include murders and non-negligent manslaughter, legacy rape, revised rape, robbery and aggravated assault. Total property (i.e., "non-violent") crimes include offenses such as burglary, larceny and motor-vehicle theft. The total crime rate includes both total violent as well as total property crimes. We employ the reciprocal of the total crime rate so that our measure of safety, Y_3, increases as crime decreases.[10]

We use total land area (in square miles), Y_4, obtained from the U.S. Census of Governments. Land area is often included in studies of local public good provision; examples include Grossman et al. (1999), Ibrahim and Karim (2004), Moore et al. (2005), Sung (2007), Nakazawa (2013) and Da Cruz and Marques (2014). Larger land areas require more infrastructure and public good servicing, such as highway repair and sewerage connections. In addition, while we include population (Y_1), providing services to a highly concentrated population is likely very different from providing services to a dispersed population, and including land area controls for this. Moreover, our model is fully nonparametric, and so no restrictions are imposed on how these two variables (i.e., Y_1 and Y_4) might interact. The total number of unit-level building permits issued in a given year, Y_5, provides a measure of the amount of adminis-

[9] Alternative measures of communal service administrative include number of highway miles and square footage of green space. Unfortunately, these data are not readily available for US municipalities.

[10] One might also consider two separate measures of crime, one for violent crime and another for property crimes such as theft. We expect that there would be a high degree of collinearity between these measures, and as discussed below, there is a high degree of collinearity among our six output measures. Consequently, splitting Y_3 into two measures may not add much information.

trative services provided.[11] Data on building permits are obtained from the Annual U.S. Building and Permit Survey and are summed over months to obtain annual figures. Finally, data on the annual employment rate, Y_6, are obtained from the BLS Local Area Unemployment Statistics. All of our data are at the level of municipal governments in a given year.[12] After eliminating observations with missing values for one or more of our variables, we have 648, 730, 746 and 800 observations for 1997, 2002, 2007 and 2012, respectively, for a total of 2924 observations.[13]

Summary statistics for our variables are presented in Table 1. For each variable, Table 1 shows the minimum value, first quartile ($Q1$), median, mean, third quartile ($Q3$) and the maximum value. The wide range of city sizes and our use of data spanning 15 years results in substantial variation in each of the variables as reflected in the table. Comparing differences between the median and $Q1$ and between $Q3$ and the median for the input and output variables reveals that the marginal distributions are skewed to the right, reflecting the distribution of city sizes in the USA. In addition, Table 2 gives the number of observations in our sample for each region, year and region–year. Table 2 reveals a good bit of variation in the number of observations across regions and across time. This is addressed in Sect. 4.2.

4 Empirical results

4.1 Preliminary results

4.1.1 Making our estimates more accurate

We begin with some preliminary investigation using our data in order to determine how to proceed toward our main results discussed in Sect. 4.2. Although our nonparametric model is highly flexible, there is a price to pay in terms of the well-known "curse of dimensionality." Wilson (2018) discusses dimension reduction in the context of nonparametric efficiency estimation, and presents diagnostics to indicate whether reducing dimensionality might be advantageous. As discussed in Sect. 2, the FDH, VRS and CRS estimators converge at rate n^{κ}, where $\kappa = 1/(p+q)$ for FDH estimators, $\kappa = 2/(p+q+1)$ for VRS estimators and $\kappa = 2/(p+q)$ for CRS estimators.

[11] We use the number of permits issued for individual units, rather than number of buildings for which permits have been issued. In dense urban environments, one might observe multiple units (e.g., condominium units) in a single building for which building permits have been issued.

[12] We use the U.S. Census Bureau's definition of local government, corresponding to type code 2 in the U.S. Census of Governments' 14-digit government ID code.

[13] Starting with data from the Annual Survey of State and Local Government Finance, we merge data from other sources listed above. Necessarily, there are mismatches across the various sources, and in particular some of the municipal governments observed in the Annual Survey of State and Local Government Finance in a given year are not observed in one or more of the other sources. Due to this, we must discard observations for 165–363 cities in each year. As one might expect, these are predominantly smaller cities, but due to the skewed distribution of city-sizes in the USA, the discarded observations are a small proportion of the number of smaller cities for which observations remain in our sample. In the end, our sample size is larger than many that have been used in the literature, e.g., Moore et al. (2005) use a sample of 46 US cities observed over 6 years, giving a total sample size of 276 observations, whereas we have 648–800 observations in each of the 4 years covered in our sample.

Table 1 Summary statistics

Variable	Min	$Q1$	Median	Mean	$Q3$	Max
X	2.7265×10^{2}	5.0395×10^{4}	1.0966×10^{5}	2.7265×10^{2}	5.0395×10^{4}	1.0966×10^{5}
Y_1	7.1800×10^{2}	4.6496×10^{4}	7.9362×10^{4}	7.1800×10^{2}	4.6496×10^{4}	7.9362×10^{4}
Y_2	4.0000×10^{0}	6.8280×10^{3}	1.3872×10^{4}	4.0000×10^{0}	6.8280×10^{3}	1.3872×10^{4}
Y_3	1.1500×10^{-6}	2.3444×10^{-4}	3.4525×10^{-4}	1.1500×10^{-6}	2.3444×10^{-4}	3.4525×10^{-4}
Y_4	5.4000×10^{-1}	2.1820×10^{1}	3.9510×10^{1}	5.4000×10^{-1}	2.1820×10^{1}	3.9510×10^{1}
Y_5	1.0000×10^{0}	1.9000×10^{2}	4.7150×10^{2}	1.0000×10^{0}	1.9000×10^{2}	4.7150×10^{2}
Y_6	6.4800×10^{1}	9.4700×10^{1}	9.6100×10^{1}	6.4800×10^{1}	9.4700×10^{1}	9.6100×10^{1}
Y_*	7.2119×10^{2}	4.6995×10^{4}	7.9903×10^{4}	7.2119×10^{2}	4.6995×10^{4}	7.9903×10^{4}

Table 2 Sample sizes by year and US census region

Year	n	Region 1	Region 2	Region 3	Region 4
1997	648	90	122	200	236
2002	730	75	158	289	208
2007	746	74	160	288	224
2012	800	72	200	299	229
Total	2924	311	640	1076	897

With the $(p + q) = 7$-dimensional specification described above, the convergence rates are $n^{1/7}$, $n^{1/4}$ and $n^{2/7}$ for FDH, VRS and CRS estimators, respectively.

Moreover, as noted above, the number of observations in each period range from 649 to 800. The *effective parametric sample size* defined by Wilson (2018) is then, in the worst case, $649^{1/7} \approx 6$ for FDH estimators, $649^{1/4} \approx 25$ for VRS estimators and $649^{2/7} \approx 40$ for VRS estimators. In other words, with a sample size of 649, FDH estimators should be expected to result in estimation error of the same order one would achieve with a typical parametric estimator (converging at the root-n rate) and only 6 observations. With VRS (or CRS) estimators, one should expect estimation error of the same order that 25 (or 40) observations would provide in a parametric model. Of course, consistency of the VRS estimators requires convexity of Ψ, and consistency of the CRS estimators requires in addition CRS. It remains to be seen whether Ψ satisfies such restrictions. Of course, the notion of *effective parametric sample size* defined by Wilson (2018) presupposes that one has a correctly specified parametric model. As Robinson (1988) notes, the root-n parametric convergence rate means that estimators converge quickly to the wrong thing in a mis-specified model; Robinson refers to this as root-n *in*consistency.

Wilson (2018) also suggests examining the ratio R_y of the largest eigenvalue of the moment matrix $\boldsymbol{Y}'\boldsymbol{Y}$ to the sum of eigenvalues for $\boldsymbol{Y}'\boldsymbol{Y}$, where $\boldsymbol{Y}$ is the $(n \times 6)$ matrix of output observations. Our data yield values 0.9961, 0.9965, 0.9934 and 0.9916 for R_y in 1997, 2002, 2007 and 2012, and when we pool data over all years we obtain 0.9935 for R_y. To understand the meaning of these results, consider the set of rays from the origin passing through each observation in the six-dimensional output space $\mathbb{R}^6_+$ for each year. The values of R_y indicate that these rays lie in a very tight bundle and are very similar in terms of their angles with respect to each axis; there is little difference among the rays, even when the data are pooled over all 4 years. Consequently, the results indicate a high degree of collinearity among our six output variables, and as such the data contain almost no information about marginal rates of transformation between outputs. The smallest of the values for R_y, 0.9916, is well above the level needed for dimension reduction to likely reduce mean square error of either DEA or FDH estimates as indicated by the simulation results reported by Wilson (2018). Hence, rather than using all six of output variables, we use only the first principle component corresponding to the largest eigenvalue of the moment matrix $\boldsymbol{Y}'\boldsymbol{Y}$ (for the pooled data, over all years), denoted by Y_*, as a measure of municipalities' outputs.[14]

[14] Specifically, $Y_* = YE_y$ where E_y is the $(q \times 1)$ eigenvector corresponding to the largest eigenvalues of the moment matrix $Y'Y$.

Table 3 Numbers of observations with estimated hyperbolic technical efficiency equal to 1 in each year

Year	n	Without dimension reduction			With dimension reduction		
		FDH	VRS	CRS	FDH	VRS	CRS
1997	648	339	38	9	25	7	1
2002	730	364	35	6	24	5	1
2007	746	359	39	9	26	6	1
2012	800	381	35	7	28	6	1

The value 0.9935 for R_y obtained with the pooled data ensures that we retain 99.35% of the independent information in our six output variables, but we gain substantially in terms of the speed of convergence (and hence, reduction in estimation error) of our efficiency estimators. Using only X and Y_* for our estimation, we have $p = q = 1$ and convergence rates $n^{1/2}$, $n^{2/3}$ and n^1 for the FDH, VRS-DEA and CRS-DEA estimators. Hence, after dimension reduction, the FDH estimator has the parametric root-n rate, while the DEA estimators converge even faster! The simulation results of Wilson (2018) provide clear evidence that in our study, relying only on X and Y_* for estimation likely results in less estimation error than would be the case with seven dimensions.[15]

In order to see some of the effect of our use of dimension reduction, we compute estimates of hyperbolic efficiency defined in (5) for municipalities in each year using the full-dimensional data with our single input variable (X) and the six output variables $Y_1, \ldots, Y_6$. We then repeat this exercise using only X and the reduced-dimensional output variable Y_*. In both cases, we obtain estimates from the FDH, VRS-DEA and CRS-DEA estimators described in Sect. 2.1. Table 3 shows the number of observations in each year as well as counts of the number of estimates equal to one in each of the six scenarios. As discussed by Wilson (2018), large proportions of efficiency estimates equal to one, especially among FDH estimates, may indicate the need for dimension reduction, and this is exactly what we see in Table 3.

Each of the FDH, VRS-DEA and CRS-DEA estimators are biased toward one (thereby overstating efficiency), but the bias is largest for the FDH estimator, and smallest for the CRS estimator. The counts in Table 3 where the full data are used reveal that about half of the FDH estimates are equal to one in each year. By contrast, a much smaller proportion of the VRS-DEA estimates equal one, and less than 1.5% of the CRS-DEA estimates are equal to one in each year. Taken together, the results in columns 3–5 of Table 3 indicate that most of the inefficiency that one would find using either the VRS-DEA or CRS-DEA estimators is merely an artifact of the convexity imposed by both estimators, and in addition the assumption of CRS imposed by the CRS-DEA estimator. As such, the estimates in Table 3 obtained with the full-dimensional data make clear the need for dimension reduction, i.e., the evidence is

[15] Färe and Lovell (1988) and Olesen and Petersen (2016) suggest that dimension reduction is only appropriate when the technology is homothetic. However, the technologies simulated by Wilson (2018) are not homothetic, and yet the simulations in Wilson (2018) show that substantial reductions in estimation error can be achieved with dimension reduction as implemented here.

clear that there are too many dimensions for the sample size (see Wilson 2018 for discussion). Furthermore, the differences in proportions of estimates equal to one obtained with the FDH estimator and those obtained with the VRS-DEA estimator suggest that Ψ may not be convex. This is investigated further in Sect. 4.1.3.

In the last three columns of Table 3, where X and Y_* are used for estimation, the counts for the FDH estimator are larger than the counts for the VRS-DEA estimator, which in turn are larger than the counts for the CRS-DEA estimator. This is to be expected since the bias decreases as one moves from the FDH to the VRS-DEA and then the CRS-DEA estimators. However, the counts of FDH estimates equal to one with the reduced-dimensional data amount to only about 10% of the counts of FDH estimates equal to one with the full data. This provides clear indication of the reduction in bias obtained with dimension reduction. Overall, the results in Table 3 provide evidence (in addition to the values of R_y and the effective parametric sample sizes discussed earlier in Sect. 3) that dimension reduction likely reduces estimation error relative to what would be obtained working in the full, seven-dimensional space. Consequently, we employ dimension reduction and work in the two-dimensional space of the variables X and Y_* for the remainder of the paper. Given that we have chosen our output variables to reflect variable specifications typically used in the literature on municipal efficiency, it seems likely that other studies might similarly benefit from use of dimension-reduction methods along the lines used here.

4.1.2 One frontier or many?

Before deciding on whether to use FDH, VRS-DEA or CRS-DEA estimators, it is important to consider which might be appropriate. FDH estimators require neither convexity nor CRS, and hence are more flexible than the other two estimators. VRS-DEA estimators require convexity of the production set, while CRS-DEA estimators require both convexity as well as CRS. But before testing whether the production set is convex or whether the technology is CRS, we must first consider whether the production set is the same across time and across regions. If it is not, then the separability condition in (12) does not hold, and we must estimate conditionally as described in Sect. 2.2. With our reduced-dimensional data, the FDH estimator attains the parametric rate of convergence, and is the most flexible among the three classes of estimators, so we use FDH estimators for testing separability.

We have 4 years and four regions. Rather than dividing our data into 16 categories, which would result in small sample sizes for some categories, we first test separability with respect to time for each of the four regions using the test of Daraio et al. (2018) with multiple sample splits and the bootstrap of Simar and Wilson (2020a) as described in Sect. 2.2. Results are given in Table 4, where "Test #1" indicates the test based on averaging the Daraio et al. (2018) test statistic over the sample splits, and "Test #2" refers to tests based on the distribution of p values obtained on each sample split. We test in each of three directions—input, output and hyperbolic—since tests on one direction may not find evidence of non-separability, while another direction does; this can result when data are not uniformly distributed over the production set.

The results in Table 4 provide clear evidence that separability does not hold across time for regions 2–4. For region 1, the evidence is mixed, but separability is clearly

Table 4 Tests of separability with respect to time by US census regions (1000 splits, 1000 bootstrap replications)

Region	Input		Output		Hyperbolic	
	Statistic	*p* value	Statistic	*p* value	Statistic	*p* value
Test #1						
1	1.323	0.239	4.281	0.000	2.424	0.222
2	5.190	0.000	5.512	0.000	5.958	0.000
3	4.671	0.000	5.528	0.000	4.036	0.000
4	4.162	0.000	5.474	0.000	4.970	0.000
Test #2						
1	0.414	0.262	0.722	0.026	0.589	0.375
2	0.817	0.000	0.904	0.000	0.939	0.000
3	0.811	0.001	0.786	0.000	0.789	0.002
4	0.677	0.000	0.810	0.000	0.745	0.000

Table 5 Tests of separability with respect to US census regions, by time (1000 splits, 1000 bootstrap replications)

Year	Input		Output		Hyperbolic	
	Statistic	*p* value	Statistic	*p* value	Statistic	*p* value
Test #1						
1997	8.387	0.000	8.325	0.000	9.056	0.000
2002	9.694	0.000	7.707	0.000	9.158	0.000
2007	9.495	0.000	9.136	0.000	11.024	0.000
2012	10.604	0.000	9.048	0.000	11.713	0.000
Test #2						
1997	0.850	0.000	0.929	0.000	0.900	0.000
2002	0.913	0.000	0.939	0.000	0.923	0.000
2007	0.939	0.000	0.934	0.000	0.907	0.000
2012	0.974	0.000	0.952	0.000	0.963	0.000

rejected when testing in the output direction, and failure to reject in the input or hyperbolic directions does not imply the null hypothesis of separability is true (all statistical tests can either reject or fail to reject the null, but cannot "accept" the null). Table 5 presents analogous results for tests of separability with respect to census region for each year. Here, the results are conclusive, i.e., separability is rejected in all cases with *p* values smaller than 0.000. Overall, the results in Tables 4 and 5 indicate that separability does not hold, and that the frontiers are different for each region–year. It is perhaps not surprising that the technology shifts over time. When one considers how different the regions of the USA shown in Fig. 1 are in terms of the age of their cities, their cultures, their demographics and other features, it is perhaps also not surprising that municipalities' frontiers vary across regions.

These results are important for our subsequent investigation. Since separability does not hold, neither with respect to time nor region, we must estimate conditionally on both time and region. This means that we must treat each region–year independently, as each region–year involves a different production set and a different frontier. This raises some interesting questions, e.g., why are the frontiers different? And among the factors listed above, which cause differences among the frontiers for different regions? We leave these for future research as they are beyond the scope of this paper, but they are important questions.

4.1.3 Which estimator should be used?

As noted above, we must decide which class of estimators (i.e., FDH, VRS-DEA or CRS-DEA) to use for estimating efficiency. In many applied studies in the literature, the choice between FDH, VRS-DEA and CRS-DEA estimators often appears to be made arbitrarily, or worse, perhaps to avoid excessive numbers of estimates equal to one. The choice, however, is not innocent, as both classes of DEA estimators require convexity of the production set for statistical consistency, while FDH estimators do not. Therefore, we test the convexity of the production set, for each region–year since our data reject separability, using the test of Kneip et al. (2016) described in Sect. 2.2. Results are reported in Table 6, where again "Test #1" indicates the test based on averaging the test statistic from Kneip et al. (2016) over the sample splits, and "Test #2" refers to tests based on the distribution of p values obtained on each sample split. We again test in each of three directions for completeness; a test in one direction may reject the null while a test in another direction does not depend on the distribution of the data and other factors.

The evidence shown in Table 6 is mixed. Among the 96 tests, convexity is rejected with a p value less than .1 in 30 cases, or in just over 30% of the tests. Of course, as noted above, failure to reject the null does not provide evidence that the null is true, and if the null were true here (and if the tests were independent, which they are not), one would expect to reject in only 10% of the tests for tests of size .1. An additional consideration is provided by Wilson (2018), where simulation results indicate than in many cases after dimension reduction, FDH estimators yield smaller mean square error than VRS-DEA estimators, even if the production set is convex. Therefore we use FDH estimators for the remainder of our analysis.

The evidence against convexity in Table 6 may be surprising to some, but it is not unheard of. In microeconomic theory, convexity is often assumed for mathematical convenience as opposed to any other reason. In addition, evidence of convexity of production sets has been found in a number of studies of banks (e.g., Wheelock and Wilson 2012, 2018). But why might municipal governments face non-convex frontiers, and what is the nature of the non-convexity? We speculate that the finding of non-convexity may be related to the fact that municipal boundaries often expand incrementally rather than continuously. This occurs when a municipality annexes a new development and then must provide roads, sewers and other services, not just to one or two new residents but to many. This may result in "lumpy" output in the sense discussed by Shephard (1970). Further research is needed to resolve these questions.

Table 6 Tests of convexity by year and census region (1000 splits, 1000 bootstrap replications)

Year	Region	Input		Output		Hyperbolic	
		Statistic	*p* value	Statistic	*p* value	Statistic	*p* value
Test #1							
1997	1	0.083	0.713	0.978	0.880	0.580	0.940
	2	1.249	0.853	1.047	0.199	1.454	0.909
	3	0.178	0.742	0.616	0.960	0.403	0.794
	4	2.731	0.000	1.639	0.996	1.358	0.015
2002	1	0.322	0.683	0.902	0.970	0.029	0.374
	2	0.351	0.490	0.824	0.189	1.354	0.953
	3	0.340	0.279	0.168	0.807	0.093	0.345
	4	1.572	0.099	0.270	0.475	1.389	0.039
2007	1	0.457	0.580	1.501	0.999	0.363	0.507
	2	1.164	0.907	0.805	0.164	1.666	0.978
	3	0.111	0.673	0.030	0.418	0.701	0.958
	4	1.138	0.125	0.948	0.807	2.305	0.002
2012	1	1.295	0.994	0.229	0.516	1.805	0.999
	2	0.950	0.039	0.544	0.148	0.266	0.646
	3	0.446	0.108	0.295	0.395	0.455	0.621
	4	1.306	0.038	1.416	0.991	1.421	0.031
Test #2							
1997	1	0.218	0.592	0.603	0.005	0.280	0.464
	2	0.403	0.166	0.372	0.254	0.460	0.115
	3	0.228	0.642	0.411	0.010	0.245	0.500
	4	0.522	0.003	0.458	0.001	0.457	0.003
2002	1	0.222	0.274	0.444	0.049	0.291	0.110
	2	0.270	0.336	0.322	0.247	0.397	0.076
	3	0.216	0.326	0.193	0.327	0.183	0.425
	4	0.512	0.048	0.330	0.208	0.455	0.019
2007	1	0.219	0.470	0.503	0.026	0.279	0.248
	2	0.370	0.118	0.334	0.195	0.464	0.037
	3	0.259	0.427	0.238	0.221	0.303	0.112
	4	0.511	0.014	0.437	0.028	0.605	0.001
2012	1	0.394	0.003	0.354	0.206	0.475	0.003
	2	0.503	0.002	0.356	0.069	0.310	0.161
	3	0.390	0.024	0.243	0.268	0.237	0.452
	4	0.553	0.000	0.527	0.001	0.620	0.000

For present purposes, the finding of evidence against convexity points to use of FDH estimators instead of VRS-DEA or possibly CRS-DEA estimators.

It is important to note that we assume that all municipalities in the same region–year operate in the same (conditional) production set Ψ^z defined by (10), and consequently face the same frontier. Municipalities in a given region–year may have very different scales and budget plans, and hence may operate in different regions of the production set or under different parts of the frontier. The view here contrasts with studies that assume different frontiers for different producers. Those that do so invariably rely on fully parametric estimation methods, and allowing different frontiers buys some flexibility. But, the model described in Sect. 2.1 is fully nonparametric, and hence quite flexible. The assumptions underlying our model and described in Sect. 2.1 impose only minimal restrictions involving free disposability, continuity and some smoothness of the frontier. Moreover, as noted above, our use of FDH estimators for the remainder of our analysis means that we avoid assuming convexity of the production sets.

4.2 Main results

Our findings in Sect. 4.1 have implications for how previous studies of municipal efficiency should be regarded, as well as implications how future research should proceed. Having determined (i) that dimension reduction can reasonably be employed with expectation of reduced estimation error, (ii) that our data reject separability with respect to time and regions and hence require independent estimation within each region–year, and (iii) that FDH estimators are appropriate since the data provide evidence against convexity of production sets in individual region–years, we now turn to our main results. We first estimate efficiency for each municipality in a given region–year, using only the observations in that region–year. We use input, output and hyperbolic FDH estimators, and report summary statistics for these estimates in Tables 7 and 8. In Table 8 where we report summary statistics for the output-oriented estimates, we first take reciprocals of the estimates in order to facilitate comparison with the results in Tables 7 and 9.[16]

Since our input measure X measures cost, the input-oriented efficiency estimates summarized in Table 7 can be interpreted as estimates of cost efficiency or input overall efficiency as discussed by Simar and Wilson (2020b). As such, our input-oriented estimates give an idea of how much municipalities might feasibly reduce costs while holding output constant. Nonetheless, for a municipality operating in the interior of the set of feasible costs and outputs, we can also consider other directions to the boundary of this set, e.g., output and hyperbolic oriented measures. In the output direction, we estimate how far municipalities might expand outputs while holding cost fixed, and in the hyperbolic direction, we estimate by how much municipalities might reduce cost while increasing output by the same proportion.

Readers who wish to focus on the cost-efficiency interpretation of the input-oriented efficiency estimates may do so, but we believe the output and hyperbolic efficiency estimates give some additional insight. In particular, note that the minimum values

[16] Recall from the discussion in Sect. 2.1 that by construction, $\theta(x, y \mid \Psi) \leq 1$ and $\gamma(x, y \mid \Psi) \leq 1$, while $\lambda(x, y \mid \Psi) \geq 1$.

Table 7 Summary statistics for input-oriented FDH technical efficiency estimates by time and US census region

Year	Min	Q1	Median	Mean	Q3	Max	Var
Region 1: Northeast							
1997	0.116	0.279	0.365	0.468	0.580	1.000	0.071
2002	0.190	0.309	0.448	0.534	0.758	1.000	0.074
2007	0.197	0.348	0.474	0.568	0.759	1.000	0.073
2012	0.195	0.360	0.555	0.590	0.807	1.000	0.068
Region 2: Midwest							
1997	0.172	0.500	0.655	0.688	0.972	1.000	0.063
2002	0.166	0.453	0.608	0.637	0.816	1.000	0.051
2007	0.119	0.492	0.669	0.658	0.855	1.000	0.053
2012	0.104	0.423	0.549	0.581	0.725	1.000	0.051
Region 3: South							
1997	0.068	0.330	0.530	0.544	0.716	1.000	0.068
2002	0.046	0.288	0.471	0.505	0.699	1.000	0.072
2007	0.056	0.295	0.480	0.515	0.732	1.000	0.073
2012	0.049	0.283	0.494	0.511	0.703	1.000	0.073
Region 4: West							
1997	0.059	0.203	0.330	0.419	0.587	1.000	0.074
2002	0.028	0.173	0.297	0.415	0.621	1.000	0.085
2007	0.039	0.218	0.341	0.436	0.594	1.000	0.077
2012	0.039	0.223	0.332	0.451	0.688	1.000	0.087

for the input-oriented estimates in Table 7 are quite small, in some cases less than 0.1. This is also true for the output-oriented estimates shown in Table 8. But looking at the hyperbolic estimates in Table 9, the minimum values for each region–year (as well as the quartiles and means) are larger than for either the input or output-oriented estimates. As discussed by Wilson (2011), the hyperbolic estimates are less sensitive to the curvature and slope of the frontier in different locations, and hence yield less-extreme values.[17] Our (reciprocal) output-oriented estimates in Table 8 are slightly larger than the input-oriented estimates in Table 7, but either set of estimates signal more inefficiency than do the hyperbolic estimates in Table 9. One might reasonably doubt that any municipalities could have efficiency on the order of 0.1 as indicated for some region–years in Tables 7 and 8, but comparison of these estimates with the corresponding hyperbolic estimates in Table 9 suggests that the very small values obtained with the input and output-oriented estimators are likely consequences of the phenomenon discussed above and in footnote 17.

[17] To illustrate, consider a municipality operating near the frontier where both X and Y_* are small. Here, the frontier is very steep, and the municipality may be close to the frontier in the input direction, but far from the frontier in the output direction. For a municipality operating near the frontier but where both X and Y_* are large, the reverse is true since the frontier is relatively flat in this region. But the hyperbolic measure will be similar in both cases. See Wilson (2011) for additional discussion, and in particular see Wilson (2011, Fig. 6.1).

Table 8 Summary Statistics for Output-Oriented FDH Technical Efficiency Estimates by Time and US Census Region

Year	Min	Q1	Median	Mean	Q3	Max	Var
Region 1: Northeast							
1997	0.105	0.380	0.493	0.531	0.650	1.000	0.058
2002	0.204	0.356	0.488	0.565	0.802	1.000	0.068
2007	0.210	0.361	0.519	0.587	0.810	1.000	0.069
2012	0.136	0.321	0.524	0.585	0.865	1.000	0.083
Region 2: Midwest							
1997	0.151	0.577	0.762	0.742	0.977	1.000	0.047
2002	0.254	0.553	0.711	0.705	0.881	1.000	0.043
2007	0.189	0.551	0.704	0.705	0.867	1.000	0.045
2012	0.153	0.538	0.662	0.665	0.784	1.000	0.038
Region 3: South							
1997	0.089	0.426	0.622	0.612	0.792	1.000	0.060
2002	0.106	0.370	0.571	0.579	0.765	1.000	0.060
2007	0.097	0.376	0.585	0.588	0.771	1.000	0.060
2012	0.121	0.402	0.589	0.597	0.794	1.000	0.058
Region 4: West							
1997	0.180	0.377	0.508	0.550	0.669	1.000	0.047
2002	0.093	0.414	0.531	0.569	0.714	1.000	0.052
2007	0.093	0.381	0.496	0.556	0.718	1.000	0.053
2012	0.099	0.373	0.531	0.577	0.763	1.000	0.057

Reported values reflect inverse efficiency estimates in order to facilitate comparison with hyperbolic and input-oriented efficiency estimates

Overall, our results so far suggest that there is some inefficiency among municipalities, as well as some differences over time and regions, but aside from the remarks made above, we caution readers not to make too much of the results in Tables 7, 8 and 9. The results in these tables are of the same type reported in previous analyses of municipal efficiency in the sense that they reflect only point estimates, with no inference. Moreover, recall from Table 2 that there is substantial variation in the number of observations available for each region–year, and this in turn causes variation in the bias of our efficiency estimates across region–years. Direct comparison of the results in Tables 7, 8 and 9 across time or regions is thus problematic without inference and without accounting for bias.

Fortunately, the test of differences in mean efficiency described by Kneip et al. (2016, Sect. 3.1.1) includes explicit, generalized jackknife estimates of bias as discussed in Sect. 2.2, and can be used to compare mean efficiencies across pairs of region–years with unequal numbers of observations. Moreover, the test does not restrict the frontier to be the same across a pair of region–years. Table 10 reports results from tests of differences in mean efficiency across pairs of census regions for each year, while Table 11 reports results from tests of differences in mean efficiency across pairs of years for

Table 9 Summary statistics for hyperbolic FDH technical efficiency estimates by time and US census region

Year	Min	$Q1$	Median	Mean	$Q3$	Max	Var
Region 1: Northeast							
1997	0.368	0.582	0.676	0.695	0.798	1.000	0.028
2002	0.453	0.558	0.658	0.711	0.893	1.000	0.034
2007	0.452	0.597	0.692	0.739	0.896	1.000	0.030
2012	0.370	0.642	0.726	0.752	0.899	1.000	0.030
Region 2: Midwest							
1997	0.434	0.723	0.835	0.834	0.986	1.000	0.021
2002	0.488	0.725	0.827	0.819	0.943	1.000	0.019
2007	0.426	0.734	0.828	0.821	0.910	1.000	0.017
2012	0.350	0.712	0.783	0.781	0.860	1.000	0.018
Region 3: South							
1997	0.333	0.611	0.752	0.745	0.883	1.000	0.029
2002	0.267	0.584	0.724	0.713	0.846	1.000	0.033
2007	0.225	0.591	0.742	0.725	0.876	1.000	0.034
2012	0.275	0.603	0.732	0.728	0.873	1.000	0.031
Region 4: West							
1997	0.318	0.550	0.673	0.688	0.802	1.000	0.029
2002	0.278	0.543	0.640	0.673	0.825	1.000	0.033
2007	0.291	0.496	0.661	0.661	0.807	1.000	0.039
2012	0.269	0.514	0.659	0.682	0.845	1.000	0.041

each census region. In both cases, we consider tests based on efficiency estimated in the input, output and hyperbolic directions.[18]

Values of test statistics and corresponding p values are shown in Table 10 for 72 cases, and the p values are less than .1 in 62 (more than 87%) of these cases. Each line in Table 10 involves a test for different mean efficiencies in region j versus region k. In the input and hyperbolic directions, positive (negative) test statistics indicate than region j is more (less) efficient than region k on average. For the output direction, positive (negative) test statistics indicate region j is less (more) efficient than region k on average. In each of the 4 years covered by our data, the tests comparing region 2 versus region 3, region 2 versus region 4 and region 3 versus region 4 provide

[18] Note that the test for differences in mean efficiency developed by Kneip et al. (2016) requires independence between the two means to avoid complications arising from covariance. Our tests of different mean efficiencies across regions involve different municipalities in different regions, so there is little reason to suspect independence does not hold. However, our tests of differences in mean efficiency across time in a particular region involve many of the same municipalities observed at two points in time, and consequently covariance is an issue. On the other hand, inertia likely plays a role here, i.e., a municipality that performs poorly (or well) in one period is likely to also perform poorly (or well) in the next period. Consequently, any covariance is likely to be positive, and since we are testing for *differences*, ignoring positive covariance makes our tests conservative by biasing toward failure to reject the null hypothesis of no difference in mean efficiencies.

Table 10 Tests of differences in mean efficiency across US census region by year

Year		Input		Output		Hyperbolic	
		Statistic	*p* value	Statistic	*p* value	Statistic	*p* value
1997	1 v. 2	− 10.309	6.425×10^{-25}	11.040	2.149×10^{-28}	− 11.052	2.149×10^{-28}
	1 v. 3	− 4.724	2.307×10^{-6}	4.101	1.880×10^{-7}	− 5.211	1.880×10^{-7}
	1 v. 4	1.921	5.477×10^{-2}	4.724	3.037×10^{-1}	− 1.029	3.037×10^{-1}
	2 v. 3	7.482	7.322×10^{-14}	− 8.243	2.670×10^{-14}	7.613	2.670×10^{-14}
	2 v. 4	14.500	1.215×10^{-47}	− 12.120	1.557×10^{-40}	13.330	1.557×10^{-40}
	3 v. 4	8.339	7.484×10^{-17}	− 0.354	9.604×10^{-9}	5.738	9.604×10^{-9}
2002	1 v. 2	− 4.850	1.233×10^{-6}	10.705	2.127×10^{-19}	− 9.007	2.127×10^{-19}
	1 v. 3	− 1.368	1.714×10^{-1}	4.125	2.551×10^{-4}	− 3.657	2.551×10^{-4}
	1 v. 4	3.364	7.694×10^{-4}	3.498	9.528×10^{-1}	− 0.059	9.528×10^{-1}
	2 v. 3	4.894	9.899×10^{-7}	− 8.909	5.967×10^{-14}	7.509	5.967×10^{-14}
	2 v. 4	11.102	1.227×10^{-28}	− 9.042	5.649×10^{-38}	12.882	5.649×10^{-38}
	3 v. 4	7.010	2.386×10^{-12}	− 0.635	2.217×10^{-9}	5.981	2.217×10^{-9}
2007	1 v. 2	− 4.429	9.475×10^{-6}	9.232	8.017×10^{-14}	− 7.470	8.017×10^{-14}
	1 v. 3	− 0.652	5.145×10^{-1}	3.603	1.284×10^{-3}	− 3.220	1.284×10^{-3}
	1 v. 4	3.466	5.287×10^{-4}	1.931	1.523×10^{-2}	2.427	1.523×10^{-2}
	2 v. 3	5.914	3.338×10^{-9}	− 7.392	2.570×10^{-10}	6.323	2.570×10^{-10}
	2 v. 4	10.798	3.514×10^{-27}	− 9.491	5.859×10^{-43}	13.740	5.859×10^{-43}
	3 v. 4	5.713	1.110×10^{-8}	− 1.922	2.493×10^{-16}	8.196	2.493×10^{-16}

Table 10 continued

Year		Input		Output		Hyperbolic	
		Statistic	*p* value	Statistic	*p* value	Statistic	*p* value
2012	1 v. 2	1.434	1.515×10^{-1}	6.301	4.089×10^{-2}	− 2.045	4.089×10^{-2}
	1 v. 3	2.360	1.828×10^{-2}	4.164	5.325×10^{-1}	− 0.624	5.325×10^{-1}
	1 v. 4	4.460	8.209×10^{-6}	2.671	1.208×10^{-3}	3.237	1.208×10^{-3}
	2 v. 3	0.948	3.433×10^{-1}	− 4.807	1.589×10^{-2}	2.412	1.589×10^{-2}
	2 v. 4	4.254	2.102×10^{-5}	− 6.900	2.200×10^{-14}	7.638	2.200×10^{-14}
	3 v. 4	3.666	2.467×10^{-4}	− 2.107	1.444×10^{-7}	5.260	1.444×10^{-7}

clear evidence that municipalities region 2 are *more* efficient on average than those in regions 3 or 4, and those in region 3 are *more* efficient on average than those in region 4; the results agree regardless of whether the input, output or hyperbolic direction is used.[19] Tests for region 1 versus regions 2 or 3 are also in agreement in each of the three directions and in each year where tests are statistically significant, and indicate that municipalities in region 1 are *less* efficient than those in regions 2 or 3. For test involving region 1 versus region 4, only the input direction yields significant results in 1997 and 2002, and the positive values obtained in the input direction indicate that region 1's municipalities are *more* efficient on average than those in region 4. In 2007 and 2012, tests involving region 1 versus region 4 are statistically significant in all three directions, but produce conflicting results, with the input and hyperbolic directions indicating that average efficiency is larger in region 1 than region 4, and the output direction indicating the opposite.

To summarize, we find overall that in each year, municipalities in region 2 are more efficient on average than municipalities in region 3, which are more efficient on average than those in region 4. The evidence for 1997 and 2002 suggests that municipalities in region 1 are more efficient on average than those in region 4, but less efficient on average than those in either region 2 or 3. The evidence for 2007 similarly suggests that municipalities in region 1 are less efficient on average than those in region 2 or 3, but the comparison between regions 1 and 4 is not clear. The evidence for 2012 suggests clearly that municipalities in region 1 are less efficient on average than those in region 2, but the comparison between region 1 and regions 3 and 4 is less clear.

There are perhaps numerous reasons why municipal efficiencies vary across census regions, and further work is needed to examine the sources of this variation. Nonetheless, cities in region 1 (Northwest) tend to be older, with higher population density than those in the Midwest (region 2) or West (region 4). In the South (region 3), cities are typically older and have higher densities in many cases near the Atlantic coast than further west (e.g., in Texas). These differences reflect the history of urbanization in the USA, westward migration and coinciding decreases in transportation costs, and likely contribute to lower efficiency on average in the Northeast (region 1) than in the Midwest (region 2). Cities in the South (region 3) are more heterogeneous than those in the Northeast or Midwest in terms of their ages, but many of those in the South are younger and less dense than those in the Northeast, perhaps contributing to the larger average efficiency we find in the South than in the Northeast. The largest cities in the West (region 4) are concentrated along the Pacific coast. Moreover, in recent decades, residents of states along the Pacific coast and in the Northeast have tended to vote Democratic, while those in the Midwest and South have tended to vote Republican. Conceivably, the resulting differences in local government policy might have some effect on our finding that municipalities in regions 2 and 3 are on average more efficient than those in regions 1 and 4. More research is needed, as it is beyond the scope of this paper to answer all of these questions here.

Turning to results from the tests for differences in mean efficiency over time (by regions) in Table 11, signs on the test statistics should be interpreted similarly to those in Table 10. In other words, for a test of different mean efficiencies between time t_1

[19] Among these tests, only one—for region 2 versus region 3 in 2012—is not statistically significant.

Table 11 Tests of differences in mean efficiency across time by US census region

Period	Input		Output		Hyperbolic	
	Statistic	p value	Statistic	p value	Statistic	p value
Region 1: Northeast						
1997–2002	-2.317	2.053×10^{-2}	1.634	6.996×10^{-1}	-0.386	6.996×10^{-1}
2002–2007	-0.904	3.662×10^{-1}	0.719	1.082×10^{-1}	-1.607	1.082×10^{-1}
2007–2012	-1.917	5.522×10^{-2}	-0.085	2.617×10^{-2}	-2.224	2.617×10^{-2}
1997–2012	-5.274	1.331×10^{-7}	1.943	2.307×10^{-5}	-4.233	2.307×10^{-5}
Region 2: Midwest						
1997–2002	3.281	1.036×10^{-3}	-1.302	1.053×10^{-1}	1.620	1.053×10^{-1}
2002–2007	-1.291	1.967×10^{-1}	-0.871	5.727×10^{-1}	0.564	5.727×10^{-1}
2007–2012	5.745	9.211×10^{-9}	-1.875	1.231×10^{-5}	4.372	1.231×10^{-5}
1997–2012	7.299	2.901×10^{-13}	-3.332	1.072×10^{-8}	5.719	1.072×10^{-8}
Region 3: South						
1997–2002	0.051	9.597×10^{-1}	0.974	5.650×10^{-1}	0.575	5.650×10^{-1}
2002–2007	-0.228	8.200×10^{-1}	0.240	2.683×10^{-1}	-1.107	2.683×10^{-1}
2007–2012	0.786	4.318×10^{-1}	1.496	9.713×10^{-1}	-0.036	9.713×10^{-1}
1997–2012	0.327	7.438×10^{-1}	2.417	4.999×10^{-1}	-0.675	4.999×10^{-1}
Region 4: West						
1997–2002	-1.169	2.423×10^{-1}	0.935	4.276×10^{-1}	0.793	4.276×10^{-1}
2002–2007	-1.666	9.568×10^{-2}	-0.930	2.646×10^{-1}	1.116	2.646×10^{-1}
2007–2012	-0.976	3.290×10^{-1}	1.430	1.205×10^{-2}	-2.511	1.205×10^{-2}
1997–2012	-4.142	3.443×10^{-5}	1.413	3.270×10^{-1}	-0.980	3.270×10^{-1}

and t_2, positive (negative) test statistics indicate greater (smaller) mean efficiency in t_1 than in t_2 for the input and hyperbolic directions, while the reverse is true for the output direction. Among the tests for region 3, none are significant. Although the technology changed over time for municipalities in region 3 as indicated by the separability tests in Table 4, we find no evidence of changes in mean efficiency over time within region 3. Apparently, to the extent that the technology moved up or down, municipalities in region 3 moved in similar directions, following the frontier as it shifted up or down. For region 1, the results in Table 11 suggest that mean efficiency worsened from 1997 to 2002 as well as 1997 to 2012. The results for 2002–2007 are insignificant, and the evidence for 2007–2012 is unclear (the results are significant in all three directions, but indicate worsening efficiency in the input and hyperbolic directions, and improving efficiency on average in the output direction). For the Midwest (region 2), there is some evidence of improving average efficiency from 1997 to 2002, and strong evidence for improvement from 2007 to 2012 as well as over the entire period 1997 to 2012. The results for 2002 to 2007 are insignificant in all three directions. For region 4, to the extent that results are significant, we find evidence of worsening efficiency on average for 2002–2007, 2007–1012 and 1997–2012, with no significant changes from 1997 to 2002. The finding that average efficiency declined over the years following the financial crisis of 2007–2008 in the Northeast and West is in accordance with the discussion in Sect. 1 regarding bankruptcy filings by municipal governments; among those listed in Sect. 1, all except Detroit, MI and Hillview, KY are in either region 1 or 4. Moreover, as noted above, in the regions where average efficiency declined (Northeast and West) voters tend toward the Democratic Party, while in the regions where average efficiency either improvise or did not decline, voters tend toward the Republican Party. More research is needed to examine this further, but the results are suggestive.

Since we use reduced-dimensional data with $p = q = 1$, we can measure productivity simply by dividing Y_* by X for each observation. Table 12 gives summary statistics on productivity by year and by census region. In terms of mean productivity, there are some differences across years and across regions, but inference is needed to know whether these differences are significant. Since we measure productivity by a simple ratio, without using the nonparametric efficiency estimators, we can use the standard Lindeberg–Feller CLT to make inference about differences in mean efficiency across regions and years. In Table 13, we examine differences between regions, by year. For a test of region j versus region k, a positive (negative) test statistic indicates larger (smaller) productivity in region k than in region j. Among the 24 tests in Table 13, all but four yield p values smaller than .1. The results in Table 13 indicate that for 1997, regions 2, 3 and 4 are more productive on average than region 1, but we find no significant evidence to rank regions 2–4. For 2002, 2007 and 2012, however, the results are clear and crisp: in each of these years, we find evidence that municipalities in the Midwest (region 2) are more productive on average than those in the South (region 3), which in turn are more productive on average than those in the West (region 4), which dominate those in the Northeast (region 1) in terms of average productivity. The patterns here are similar to our findings regarding differences in mean efficiency, and may be related to the same factors discussed above.

Table 12 Summary statistics for productivity estimates by year and US census region

Year	Min	*Q*1	Median	Mean	*Q*3	Max
Region 1: Northeast						
1997	0.408	0.729	0.970	1.316	1.573	5.221
2002	0.508	0.738	0.908	1.326	1.839	3.875
2007	0.421	0.693	0.925	1.317	1.580	11.286
2012	0.401	0.661	0.867	1.136	1.483	3.827
Region 2: Midwest						
1997	0.291	0.711	0.993	1.056	1.289	2.075
2002	0.224	0.728	0.868	0.950	1.184	2.460
2007	0.259	0.897	1.126	1.143	1.332	2.621
2012	0.267	0.895	1.103	1.127	1.308	3.376
Region 3: South						
1997	0.201	0.755	1.190	1.229	1.565	4.135
2002	0.163	0.711	1.151	1.197	1.500	3.803
2007	0.147	0.704	1.219	1.270	1.663	4.126
2012	0.141	0.669	1.153	1.173	1.537	6.992
Region 4: West						
1997	0.236	0.833	1.063	1.232	1.421	6.714
2002	0.106	0.707	0.999	1.146	1.338	7.239
2007	0.162	0.722	0.966	1.130	1.357	6.659
2012	0.132	0.768	1.024	1.160	1.361	6.405

We also examine differences in mean productivity across time, by region, in Table 14, again using standard CLT results.[20] For a test of difference in mean productivity between years t_1 and t_2 (where $t_2 > t_1$), a statistic with a positive (negative) value indicates improving (worsening) productivity. For the Northeast, we find no change from 1997 to 2002, a decrease in productivity from 2002 to 2007, and an offsetting improvement from 2007 to 2012, with no net change from 1997 to 2012. It is perhaps surprising that productivity declined *before* the 2007–2008 financial crisis, then improved afterward. Perhaps tighter budgets forced administrators to waste less. In the West, we also see offsetting changes, but earlier than in the Northeast with a decrease in productivity from 1997 to 2002 followed by an increase from 2002 to 2007, and no apparent net change from 1997 to 2012. By contrast, we find significant increases in average productivity in regions 2 and 3, with improvement during 1997–2002 and 2007–2012 in the Midwest, and in 1997–2002 in the South. Both the Midwest and South show evidence in net improvements in average productivity over the entire period 1997–2012. The patterns seen in Table 14 are similar to those seen in Tables 10, 11 and 12, and are likely due to some of the same factors driving patterns in those tables.

[20] Similar reasoning regarding covariance across time discussed in footnote 18 applies here. To the extent that there is positive covariance between municipalities' productivity over time, our tests are conservative.

Table 13 Tests of differences in mean productivity across US census region by years

Year	Regions	Productivity	
		Statistic	*p* value
1997	1 v. 2	9.353	8.560×10^{-21}
	1 v. 3	9.060	1.301×10^{-19}
	1 v. 4	8.034	9.436×10^{-16}
	2 v. 3	– 0.292	7.701×10^{-1}
	2 v. 4	– 1.319	1.873×10^{-1}
	3 v. 4	– 1.026	3.048×10^{-1}
2002	1 v. 2	13.577	5.486×10^{-42}
	1 v. 3	12.230	2.161×10^{-34}
	1 v. 4	4.908	9.221×10^{-7}
	2 v. 3	– 1.347	1.779×10^{-1}
	2 v. 4	– 8.669	4.345×10^{-18}
	3 v. 4	– 7.322	2.444×10^{-13}
2007	1 v. 2	18.195	5.611×10^{-74}
	1 v. 3	16.403	1.829×10^{-60}
	1 v. 4	10.864	1.704×10^{-27}
	2 v. 3	– 1.793	7.302×10^{-2}
	2 v. 4	– 7.331	2.284×10^{-13}
	3 v. 4	– 5.538	3.054×10^{-8}
2012	1 v. 2	16.789	2.922×10^{-63}
	1 v. 3	11.104	1.197×10^{-28}
	1 v. 4	7.140	9.362×10^{-13}
	2 v. 3	– 5.685	1.307×10^{-8}
	2 v. 4	– 9.650	4.928×10^{-22}
	3 v. 4	– 3.965	7.352×10^{-5}

5 Summary and conclusions

Taken together, our findings in Sect. 4.2 suggest that on average, municipalities in the Midwest and the South out-performed those in the Northeast and West in terms of both efficiency and productivity. There are many differences between these regions, including cultural, demographic and political differences. It is well known that in recent presidential and congressional elections, Democratic votes have been somewhat concentrated in the Northeast and in states along the Pacific coast, while Republican votes have prevailed in much of the South and Midwest. Local elections have followed similar patterns, and the resulting differences in municipal management may explain some of the differences we have found. But, there are other differences, and these should be explored in future research.

Table 14 Tests of differences in mean productivity across time by US census region

Period	Productivity	
	Statistic	*p* value
Region 1: Northeast		
1997–2002	0.294	7.690×10^{-1}
2002–2007	− 4.145	3.393×10^{-5}
2007–2012	4.254	2.098×10^{-5}
1997–2012	0.403	6.873×10^{-1}
Region 2: Midwest		
1997–2002	4.518	6.238×10^{-6}
2002–2007	0.473	6.361×10^{-1}
2007–2012	2.848	4.398×10^{-3}
1997–2012	7.839	4.529×10^{-15}
Region 3: South		
1997–2002	3.463	5.342×10^{-4}
2002–2007	0.028	9.778×10^{-1}
2007–2012	− 1.044	2.963×10^{-1}
1997–2012	2.446	1.443×10^{-2}
	Region 4: West	
1997–2002	− 2.833	4.616×10^{-3}
2002–2007	1.811	7.007×10^{-2}
2007–2012	0.529	5.965×10^{-1}
1997–2012	− 0.492	6.228×10^{-1}

We find some differences in efficiency and productivity before and after the financial crisis of 2007–2008, but the evidence is mixed and unclear. The financial crisis affected the entire country, but we find no significant evidence of decreased productivity by municipalities (in any of the four census regions) in the years following the crisis. While we find some evidence of changes in efficiency after the financial crisis, we do not see clear patterns across all regions. To the extent that the financial crisis had an effect on municipalities, it may be been outweighed by other, more local considerations suggested above.

A large number of previous papers examining municipal efficiency have used nonparametric efficiency estimators, but the majority have used either VRS-DEA or CRS-DEA estimators, and the choice seems to be ad hoc in most cases. By contrast, we carefully test model features such as separability and convexity of production sets, and the results of these tests inform us regarding the choice among the classes of FDH, VRS-DEA and CRS-DEA estimators. In addition, we exploit collinearity in the data to reduce estimation error. Future research should examine the efficacy of dimension reduction using the diagnostics provided by Wilson (2018); we expect that in many cases, dimension reduction will be advantageous due to the common, inherent collinearity in economic data. Finally, we have used a well-defined nonparametric model to allow rigorous statistical testing of various hypotheses using the results of Kneip et al. (2015, 2016), Daraio et al. (2018) and Simar and Wilson (2020a). Statistical tools for hypothesis testing analogous to those long-used by parametric modelers are now available for the nonparametric efficiency and productivity framework, and should be used.

As we have noted at several points in Sect. 4, we have left some questions for future research. This is unavoidable in a journal-length paper, but we believe there is some value in suggesting directions for future research. In particular, the differences we have found across census regions strike us as interesting. Our prior was that our most important findings would revolve around the 2007–2008 financial crisis, but as noted above, to the extent the crisis had an effect on municipal governments, the effects were mixed. We have suggested several differences among the four census regions that might be related to differences we see across the regions, but more detailed data seem likely to reveal more insight. More research is needed.

Funding This study was funded by NSF Award No. 1633608.

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

Afonso FS (2008) Assessing and explaining the relative efficiency of local government. J Socio-Econ 37:1946–1979

Aigner D, Lovell CAK, Schmidt P (1977) Formulation and estimation of stochastic frontier production function models. J Econom 6:21–37

Athanassopoulos AD, Triantis KP (1998) Assessing aggregate cost efficiency and the related policy implications for Greek local municipalities. INFOR Inf Syst Oper Res 36:66–83

Bahadur RR, Savage LJ (1956) The nonexistence of certain statistical procedures in nonparametric problems. Ann Math Stat 27:1115–1122

Baicker K, Clemens J, Singhal M (2012) The rise of the states: US fiscal decentralization in the postwar period. J Public Econ 96:1079–1091

Balaguer-Coll MT, Prior D, Tortosa-Ausina E (2007) On the determinants of local government performance: a two-stage nonparametric approach. Eur Econ Rev 51:425–451

Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manage Sci 30:1078–1092

Banks J, Blundell R, Lewbel A (1997) Quadratic Engel curves and consumer demand. Rev Econ Stat 79:527–539

Berry WD, Lowery D (1984) The growing cost of government: a test of two explanations. Soc Sci Q 65:735–749

Bruère H, Allen WH, Cleveland FA, Baker SJ (1912) Efficiency in city government. Ann Am Acad Polit Soc Sci 41:3–22

Cazals C, Florens JP, Simar L (2002) Nonparametric frontier estimation: a robust approach. J Econom 106:1–25

Chalfant JA, Gallant AR (1985) Estimating substitution elasticities with the Fourier cost function. J Econom 28:205–222

Chalos P, Cherian J (1995) An application of data envelopment analysis to public sector performance measurement and accountability. J Account Public Policy 14:143–160

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2:429–444

Charnes A, Cooper WW, Li S (1989) Using DEA to evaluate relative efficiencies in the economic performance of Chinese cities. Socio-Econ Plan Sci 23:325–344

Cooper RJ, McLaren KR (1996) A system of demand equations satisfying effectively global regularity conditions. Rev Econ Stat 78:359–364

Da Cruz NF, Marques RC (2014) Revisiting the determinants of local government performance. Omega 44:91–103

Daouia A, Simar L, Wilson PW (2017) Measuring firm performance using nonparametric quantile-type distances. Econom Rev 36:156–181

Daraio C, Simar L (2005) Introducing environmental variables in nonparametric frontier models: a probabilistic approach. J Prod Anal 24:93–121

Daraio C, Simar L (2007) Advanced robust and nonparametric methods in efficiency analysis. Springer, New York

Daraio C, Simar L (2007) Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach. J Prod Anal 28:13–32

Daraio C, Simar L, Wilson PW (2018) Central limit theorems for conditional efficiency measures and tests of the 'separability condition' in non-parametric, two-stage models of production. Econom J 21:170–191

Davis ML, Hayes KJ (1993) The demand for good government. Rev Econ Stat 75:148–152

De Borger B, Kerstens K (1996) Cost efficiency of Belgian local governments: a comparative analysis of FDH, DEA, and econometric approaches. Reg Sci Urban Econ 26:145–170

De Borger B, Kerstens K (2000) What is known about municipal efficiency? the Belgian case and beyond. In: Blank J (ed) Public provision and performance: contributions from efficiency and productivity measurement. Elsevier, Amsterdam, pp 299–330

De Borger B, Kerstens K, Moesen W, Vanneste J (1994) Explaining differences in productive efficiency: an application to Belgian municipalities. Public Choice 80:339–358

de Oliveira Junqueira M (2015) Efficiency measurement of local public sector in international perspective: a literature survey. In: Unpublished working paper, University of São Paulo, Brazil. Presented at International Conference on Public Policy, Milan, 2015

De Sousa MD, Stošić B (2005) Technical efficiency of the Brazilian municipalities: correcting nonparametric frontier measurements for outliers. J Prod Anal 24:157–181

Deprins D, Simar L, Tulkens H (1984) Measuring labor inefficiency in post offices. In: Pestieau MMP, Tulkens H (eds) The performance of public enterprises: concepts and measurements. North-Holland, Amsterdam, pp 243–267

Fang C, Guan X, Lu S, Zhou M, Deng Y (2013) Input–output efficiency of urban agglomerations in China: an application of data envelopment analysis (DEA). Urban Stud 50:2766–2790

Färe R (1988) Fundamentals of production theory. Springer, Berlin

Färe R, Grosskopf S, Lovell CAK (1985) The measurement of efficiency of production. Kluwer-Nijhoff Publishing, Boston

Färe R, Lovell CAK (1988) Aggregation and efficiency. In: Eichhorn W (ed) Measurement in economics. Physica, Heidelberg, pp 639–647

Farrell MJ (1957) The measurement of productive efficiency. J R Stat Soc A 120:253–281

Geys B, Moesen W (2009) Exploring sources of local government technical inefficiency: evidence from Flemish municipalities. Public Finance Manag 9:1–29

Grosskopf S, Hayes KJ, Taylor LL, Weber WL (2001) On the determinants of school district efficiency: competition and monitoring. J Urban Econ 49:453–478

Grossman PJ, Mavros P, Wassmer RW (1999) Public sector technical inefficiency in large U.S. cities. J Urban Econ 46:278–299

Grossman PJ, West EG (1994) Federalism and the growth of government revisited. Public Choice 79:19–32

Guilkey DK, Lovell CAK, Sickles RC (1983) A comparison of the performance of three flexible functional forms. Int Econ Rev 24:591–616

Hayes KJ, Chang S (1990) The relative efficiency of city manager and mayor-council forms of government. South Econ J 57:167–177

Hayes KJ, Razzolini L, Ross LB (1998) Bureaucratic choice and nonoptimal provision of public goods: theory and evidence. Public Choice 94:1–20

Hayes KJ, Wood L (1995) Utility maximizing bureaucrats: the Bureaucrat's point of view. Public Choice 82:69–83

Ibrahim FW, Karim M (2004) Efficiency of local governments in Malaysia and its correlates. Int J Manag Stud 11:57–70

Kalseth J, Rattsø J (1995) Spending and overspending in local government administration: a minimum requirement approach applied to Norway. Eur J Polit Econ 11:239–251

Kneip A, Park B, Simar L (1998) A note on the convergence of nonparametric DEA efficiency measures. Econom Theory 14:783–793

Kneip A, Simar L, Wilson PW (2008) Asymptotics and consistent bootstraps for DEA estimators in non-parametric frontier models. Econom Theory 24:1663–1697

Kneip A, Simar L, Wilson PW (2011) A computationally efficient, consistent bootstrap for inference with non-parametric DEA estimators. Comput Econ 38:483–515

Kneip A, Simar L, Wilson PW (2015) When bias kills the variance: central limit theorems for DEA and FDH efficiency scores. Econom Theory 31:394–422

Kneip A, Simar L, Wilson PW (2016) Testing hypotheses in nonparametric models of production. J Bus Econ Stat 34:435–456

Kneip A, Simar L, Wilson PW (2020) Inference in dynamic, nonparametric models of production: central limit theorems for Malmquist indices. Econometric Theory Forthcoming

Lee DR, Wilson PW (1990) Rent-seeking and peak-load pricing of public services. Natl Tax J 43:497–503

Lee DR, Wilson PW (1991) Efficient user charges in a rent seeking model. Econ Inq 29:367–374

lo Storto C (2013) Evaluating technical efficiency of Italian major municipalities: a data envelopment analysis model. Proc Soc Behav Sci 81:346–350

McAllister PH, McManus D (1993) Resolving the scale efficiency puzzle in banking. J Bank Finance 17:389–405

Meeusen W, van den Broeck J (1977) Efficiency estimation from Cobb–Douglas production functions with composed error. Int Econ Rev 18:435–444

Mitchell K, Onvural NM (1996) Economies of scale and scope at large commercial banks: evidence from the Fourier flexible functional form. J Money Credit Bank 28:178–199

Moore A, Nolan J, Segal GF (2005) Putting out the trash: measuring municipal service efficiency in U.S. cities. Urban Aff Rev 41:237–259

Nakazawa K (2013) Cost inefficiency of municipalities after amalgamation. Proc Econ Finance 5:581–588

Narbón-Perpiñá I, De Witte K (2018) Local governments' efficiency: a systematic literature review-part I. Int Trans Oper Res 25:431–468

Nikolov M, Hrovatin N (2013) Cost efficiency of Macedonian municipalities in service delivery: does ethnic fragmentation matter? Lex Localis 11:743–776

Olesen OB, Petersen NC (2016) Stochastic data envelopment analysis: a review. Eur J Oper Res 251:2–21

Park BU, Jeong S-O, Simar L (2010) Asymptotic distribution of conical-hull estimators of directional edges. Ann Stat 38:1320–1340

Park BU, Simar L, Weiner C (2000) FDH efficiency scores from a stochastic point of view. Econom Theory 16:855–877

Radulović B, Dragutinović S (2015) Efficiency of local self-governments in Serbia: an SFA approach. Industrija 43:123–142

Robinson PM (1988) Root-*n*-consistent semiparametric regression. Econometrica 56:931–954

Shephard RW (1970) Theory of cost and production functions. Princeton University Press, Princeton

Simar L, Wilson PW (2007) Estimation and inference in two-stage, semi-parametric models of productive efficiency. J Econom 136:31–64

Simar L, Wilson PW (2011) Inference by the *m* out of *n* bootstrap in nonparametric frontier models. J Prod Anal 36:33–53

Simar L, Wilson PW (2011) Two-stage DEA: caveat emptor. J Prod Anal 36:205–218

Simar L, Wilson PW (2013) Estimation and inference in nonparametric frontier models: recent developments and perspectives. Found Trends Econom 5:183–337

Simar L, Wilson PW (2015) Statistical approaches for non-parametric frontier models: a guided tour. Int Stat Rev 83:77–110

Simar L, Wilson PW (2020) Hypothesis testing in nonparametric models of production using multiple sample splits. J Prod Anal 53:287–303

Simar L, Wilson PW (2020) Technical, allocative and overall efficiency: estimation and inference. Eur J Oper Res 282:1164–1176

Šťastná L, Gregor M (2015) Public sector efficiency in transition and beyond: evidence from Czech local governments. Appl Econ 47:680–699

Sung N (2007) Information technology, efficiency and productivity: evidence from Korean local governments. Appl Econ 39:1691–1703

Tang K-L (1997) Efficiency in the private sector: a critical review of empirical evidence from public services. Int Rev Admin Sci 63:459–474

Tiebout CM (1956) A pure theory of local public expenditure. J Polit Econ 64:416–424

U.S. Census Bureau (2020) Annual survey of state and local government finances. https://www.census.gov/programs-surveys/gov-finances.html. Accessed 19 December 2020

Wheelock DC, Wilson PW (2001) New evidence on returns to scale and product mix among U.S. commercial banks. J Monet Econ 47:653–674

Wheelock DC, Wilson PW (2012) Do large banks have lower costs? new estimates of returns to scale for U.S. banks. J Money Credit Bank 44:171–199

Wheelock DC, Wilson PW (2018) The evolution of scale-economies in U.S. banking. J Appl Econom 33:16–28

Wilson PW (2011) Asymptotic properties of some non-parametric hyperbolic efficiency estimators. In: Van Keilegom I, Wilson PW (eds) Exploring research frontiers in contemporary statistics and econometrics. Springer, Berlin, pp 115–150

Wilson PW (2018) Dimension reduction in nonparametric models of production. Eur J Oper Res 267:349–367

Wilson PW (2021) U.S. banking in the post-crisis era: new results from new methods. In: Parmeter C, Sickles R (eds) Advances in efficiency and productivity analysis. Springer Nature, Cham, pp 233–264

Wilson PW, Carey K (2004) Nonparametric analysis of returns to scale and product mix among US hospitals. J Appl Econom 19:505–524

Worthington AC (2000) Cost efficiency in Australian local government: a comparative analysis of mathematical programming and econometrical approaches. Financ Account Manag 16:201–223

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Caitlin T. O'Loughlin[1] · **Paul W. Wilson**[2]

✉ Paul W. Wilson
pww@clemson.edu

Caitlin T. O'Loughlin
caitlin.oloughlin@phil.frb.org

[1] Department of Supervision, Regulation, and Credit, Federal Reserve Bank of Philadelphia, Philadelphia, PA, USA

[2] Department of Economics and School of Computing, Division of Computer Science, Clemson University, Clemson, SC 29634–1309, USA