Is there a prison size dilemma? An empirical analysis of output-specific economies of scale

Veerle Hennebel*

Department of Economics, KU Leuven, Etienne Sabbelaan 51, 8500 Kortrijk - Belgium

Richard Simper

Centre for Risk, Banking and Financial Services, University of Nottingham, Jubilee Campus, Nottingham, UK

Marijn Verschelde

IÉSEG School of Management, LEM (UMR-CNRS 9221), Socle de la Grande Arche - 1 Parvis de La Défense, Paris La Défense cedex, 92044, France and Department of Economics, KU Leuven, Etienne Sabbelaan 51, 8500 Kortrijk - Belgium

Abstract

We advocate a nonparametric multi-output framework to estimate output-specific economies of scale and we apply this model to male prisons in England and Wales over the sample period 2009-2012. To estimate output-specific returns to scale in prisons, we consider not only the cost-per-place, but also qualitative outputs such as purposeful out-of-cell activity and successful reintegration. Furthermore, we introduce environmental heterogeneity using the characteristics of the prison(ers). England and Wales offers a unique example to study economies of scale in prisons as the UK has started to build new super-size prisons in order to replace the most outdated prisons.

Keywords: data envelopment analysis, economies of scale, multi-output production, UK penology

*Corresponding author

Email addresses: veerle.hennebel@kuleuven-kulak.be (Veerle Hennebel), Richard.Simper@nottingham.ac.uk (Richard Simper), m.verschelde@ieseg.fr (Marijn Verschelde)

1. Introduction

Prisoner numbers are on the rise for decades in the US and many European countries.[1] The increasing prisoner population puts the existing United Kingdom (UK) Criminal Justice System (CJS) under stress and forces policy makers to reconsider the limits both from a cost-per-place and from social perspectives (i.e., providing humane incarceration with prospects for reintegration into society when released). Current public policy mainly consists of building new prisons, reshaping existing prisons and putting less convicts behind bars.

This paper focuses on what we call the potential '*prison size dilemma*'. Since public policy makers could consider returns to scale from either a cost-per-place or a social viewpoint, these two viewpoints could potentially lead to conflicting opinions on the optimal scale size of a prison. We empirically test whether the optimal scale size of a prison differs when the focus is either on costs-per-place, quality of life in prison or successful reintegration. In particular, we study economies of scale of a sample of male prisons in England and Wales, by using publicly available data collected by the Ministry of Justice (MoJ).

Our empirical analysis is timely and warranted as the building of the first titan prison in the UK has started. The name titan refers to not just a single large prison but one consisting of hubs (Lockyer, 2013, page 6). The Labour government in 2007 was forced to abandon 3 titan prisons which would provide up to 2,500 places in five units of approximately 500 offenders. However, the rejection of building these titan prisons was reversed in 2011 under the next UK government - a Conservative/Liberal coalition - where the building of the first titan prison based in Wrexham, Wales was agreed to begin.

Renewing and rescaling the prison estate is part of the strategy of the National Offender Management Service (NOMS), which covers both the prison and probation systems in England and Wales, to reduce costs. The modernization of the prison estate includes the closure of old and inefficient prisons,

[1] See Levitt (1996) and Campbell et al. (2015) for a discussion on US mass incarceration and e.g. the National Audit Office (2013, p. 14) for prisoner figures for England and Wales.

which will be replaced by new large prisons and housing blocks.[2]

The cost reduction strategy of the NOMS, initiated in 2010, also involved a reduction of input waste within the system. Furthermore, the NOMS aimed to introduce more competition by privatization and re-tendering of prisons that were already tendered to the private sector. Rogge et al. (2015) document that there is little empirical support for large cost savings contracting-out prison service to private-run organizations. In our study, we analyze the optimal scale size of prisons.

We advocate a framework that is specially tailored to analyze the multidimensional prison production process. In particular, we propose a DEA-based methodology with an axiomatic basis that fully acknowledges that returns to scale can differ between the different (qualitative) dimensions of production.

We build on the work of Cherchye et al. (2013), who introduce a multi-output methodology that recognizes that each output is characterized by its own production technology. Starting from this multi-output methodology, we will be able to estimate output-specific returns to scale.

An attractive feature of the methodology is that it is nonparametric: there is no need to assume a specific functional representation of the production technology. This is warranted for public sector applications as public firms operate in non-competitive markets and can have a complex structure of public production. Consequently, the imposition of a parametric functional relationship can be intricate. Instead, a minimum set of production axioms is used to test for output-specific economies of scale.

From a methodological perspective, we contribute to the literature by showing that, by making full use of the flexibility of recently introduced production models, we can obtain a multi-output methodology that allows for varying returns to scale over outputs. As such, we can avoid a potential misspecification bias that can result from falsely imposing the assumption of non-varying scale economies over the multiple outputs. This is relevant,

[2]On January 10, 2013, The Ministry of Justice announced the closure of four prisons and partial closure of three prisons. In total, 2,614 places were closed. An announcement on September 4, 2013 showed an even more drastic change of the prison landscape as in the period 2010-2014, the prison (planned) closures consist in total of 6,382 places and total gained places in micro-prisons (housing blocks) or new large prisons are up to 5,945. For more information, see URL: https://www.gov.uk/government/news/modernisation-of-the-prison-estate.

as similar size dilemmas are widely debated in both the public and private sector. That is, related public-sector examples include the debate on re-scaling courts (Peyrache and Zago, 2015) and regional police forces (Drake and Simper, 2002; Verschelde and Rogge, 2012). A private-sector example is the debate on re-scaling bank branches (see the review of Fethi and Pasiouras (2010)).

In the context of prisons, we argue that it is crucial to consider output-specific returns to scale. In line with the key performance areas as posed by the Ministry of Justice (see Section 3), we take three output objectives into account. Naturally, we consider the incarceration of convicts as one of the main outputs of a prison. Besides incarcerating convicts, we consider in our study also qualitative outputs including the provision of a humane prison environment and successful reintegration. In the empirical analysis, we select proxies that in our opinion best reflect these output objectives.

A common motivation for large prisons is a reduction of the cost-per-place. Meanwhile, opponents fear little prospects for reintegration and low quality of life in large-scale prisons. For example, Liebling (2004) questions the moral performance of the so called '*Titan*' prisons that could hold over 2,500 prisoners. In fact, the HM Chief Inspector of Prisons (2009) and The National Audit Office (2013) provide support in England and Wales for a better performance in smaller prisons. Surveys show that prisoners tend to be more engaged in smaller establishments. Moreover small prisons do on average better in independent inspections and in the NOMS's performance ratings, which take reintegration and quality of life in prison into account. By contrast, Lockyer (2013) argues that the age and not the size of a prison determines the performance of a prison. In our opinion there is a need for further research on the relation between prison size and the multiple facets of performance. Doing so, we control for differences in the age of the prison by including prison age as an environmental variable in the analysis.

Ruggiero (2000) emphasizes that environmental variables have a considerable impact on the provision of public services and that without controlling for these environmental factors the estimates of returns to scale will be biased. We advocate a methodology that explicitly takes into account environmental heterogeneity (in contrast to the above mentioned studies). In particular, we control for prison(er) and regional characteristics. For example, next to prison age and prison management, we control in our study for the inflow of prisoners in a particular establishment. Specifically, we include the pre-

dicted rate of re-offending in an establishment. The rate of re-offending is estimated at prison level by the Ministry of Justice, based on prisoner-level data on social background, ethnicity, crime type, etc.

Furthermore, the proposed methodology distinguishes between discretionary and non-discretionary output variables. We therefore measure the performance of prisons only with respect to the output variables that the prison management controls and actually wants to maximize. Examples of non-discretionary variables in our application are the size of the average prison population and the yearly number of discharges.

To our knowledge, we posit an original estimation strategy that adequately models the multidimensional prison production process. The advocated methodology is tailored to all specificities of the prison production process and enables us to meaningfully answer the prison size dilemma, by using publicly available data. Moreover, we discuss in detail how public policy makers can further refine the analysis by adding information on the allocation of expenses to particular outputs.

The remainder of this paper is structured as follows. Section 2 explains the nonparametric multi-output methodology. Section 3 discusses the data and the empirical model and section 4 discusses the results. Section 5 concludes.

2. Methodology

To set the stage, we first intuitively introduce the concept of returns to scale. The concept of returns to scale is directly related to the most productive scale size. A Decision Making Unit (DMU) that is situated on the constant returns to scale technology, is considered to operate on its most productive scale size (Banker, 1984). A DMU which is not situated on its most productive scale size, can improve its productivity by resizing the scale of its operations. The type of returns to scale can be interpreted as the direction of change necessary to achieve its most productive scale size.[3] Increasing returns to scale indicate that the most productive scale size of a DMU is situated at a larger size. Similarly, decreasing returns to scale indicate that the DMU should decrease the scale of its operations to achieve the optimal scale size. The type of returns to scale is therefore very useful information for

[3]We estimate global returns to scale. See Podinovski (2004a) and Podinovski (2004b) for a discussion on the distinction between local and global returns to scale.

the operational manager, indicating how rescaling the operation can improve average productivity and reduce the average cost.[4]

To estimate output-specific returns to scale, we build on the work of Cherchye et al. (2013), who introduce a multi-output methodology that recognizes that each output is characterized by its own production technology. The output-specific production technologies remain linked through the use of joint inputs. In Sections 2.2 to 2.4 we will focus on the production process of one particular output, to come back to the multi-output production process in Section 2.5. We extend Cherchye et al. (2013) by including alternative returns to scale assumptions in the methodology. At this point, it might be worth to note that our approach bears some analogy to Cook and Zhu (2011), who also allow returns to scale type behavior to be different for one output subgroup than for another, by using the notion of component technologies. However, we offer an axiomatic approach to the estimation of output-specific returns to scale.

Furthermore, we include output-specific environmental variables in the methodology. Since we are able to work with output-specific production technologies, we can estimate output-specific returns-to-scale, controlling for output-specific environments. For ease of exposition, we will assume in sections 2.2 to 2.5 that all DMUs are situated in the same environment and introduce environmental influences later in section 2.6.

2.1. Notational preliminaries

Suppose we observe data for N DMUs. Each DMU n ($1 \leq n \leq N$) uses input $\mathbf{x}_n = (x_n^1, \ldots, x_n^L)$ to produce output $\mathbf{y}_n = (\mathbf{y}_n^1, \ldots, \mathbf{y}_n^R)$ and is situated in environment $\mathbf{z}_n = (z_n^1, \ldots, z_n^K)$. Note that output $\mathbf{y}_n^r$ can be a set of outputs having a common production technology.

Following Cherchye et al. (2013) and Cherchye et al. (2015), we distinguish between output-specific, joint and subjoint inputs. Output-specific inputs can be allocated to the production of particular outputs. We use α_l^r,

[4]We estimate qualitative characterizations of returns to scale, such as increasing, decreasing or constant returns to scale. There is a different strand of DEA literature which is directed to quantitative directions of returns to scale. For example Podinovski and Forsund (2010) and Atici and Podinovski (2012) analyze a class of mixed partial elasticity measures. These measures indicate the elasticity of response of a subset of outputs with respect to marginal changes of a subset of inputs.

with $\sum_{r=1}^{R} \alpha_l^r$=1, to represent the fraction of input l that is used to produce output r. Next, joint (or public) inputs simultaneously benefit the production of all outputs. Subjoint inputs also figure as joint inputs, but only for a subset of outputs. The use of joint and subjoint inputs therefore makes the output-specific production processes interdependent. Note that the methodology can also be applied when all inputs are joint (as is the case in our empirical analysis).

We summarize the information on how inputs are allocated to outputs by means of a vector $\mathbf{A}^r$ for each output r. Specifically, $(\mathbf{A}^r)_l = \alpha_l^r$ if input l is output-specific and used to produce output r. Next, $(\mathbf{A}^r)_l = 1$if input l is joint or sub-joint and used to produce output r. Finally, $(\mathbf{A}^r)_l = 0$ otherwise. The element-by-element product $\mathbf{X}^r = \mathbf{A}^r \odot \mathbf{x}$ captures the input quantities used in the production process of output r.

Next, some environmental variables can influence only a part of the outputs, not all. The vector $\mathbf{B}^r$ captures the environmental variables that are relevant for output r. In particular, $(\mathbf{B}^r)_k = 1$ if environmental variable k is relevant for output r and $(\mathbf{B}^r)_k = 0$ otherwise. Summarizing, the element-by-element product $\mathbf{Z}^r = \mathbf{B}^r \odot \mathbf{z}$ captures the environmental variables that are controlled for in the specification of the technology of output r.

Taken together, the empirical analysis starts from the following data set:

$$S = \{(\mathbf{y}_n^1, \ldots, \mathbf{y}_n^R, \mathbf{X}_n^1, \ldots, \mathbf{X}_n^R, \mathbf{Z}_n^1, \ldots, \mathbf{Z}_n^R) \mid n = 1, \ldots, N\}. \tag{1}$$

2.2. Output-specific production technology

We focus on the production technology of output r. For output r, we observe for each DMU n the inputs $\mathbf{X}_n^r$ that are used to produce output $\mathbf{y}_n^r$. We adopt an output-oriented approach[5] and characterize the production technology of output r by output sets $P^r(\mathbf{X}^r)$, which contains the amount of output $\mathbf{y}^r$ that can be produced with input $\mathbf{X}^r$.

$$P^r(\mathbf{X}^r) = \{\mathbf{y}^r | \mathbf{X}^r \text{ can produce } \mathbf{y}^r\}. \tag{2}$$

In practice, the true output sets $P^r(\mathbf{X}^r)$ are not observed. A solution is to construct empirical approximations of these output sets on the basis of some standard production axioms.

[5]In this respect we deviate from Cherchye et al. (2013), who follow an input oriented approach and characterize the production technology by input requirement sets.

Axiom 1 (Monotone output sets)**.**

$$\boldsymbol{y}^r \in P^r(\boldsymbol{X}^r) \text{ and } \boldsymbol{y}^{r*} \leq \boldsymbol{y}^r \Rightarrow \boldsymbol{y}^{r*} \in P^r(\boldsymbol{X}^r)$$

Axiom 2 (Nested output sets)**.**

$$\boldsymbol{X}^r \leq \boldsymbol{X}^{r*} \Rightarrow P^r(\boldsymbol{X}^r) \subset P^r(\boldsymbol{X}^{r*})$$

Axiom 3 (Convex output sets)**.**

$$\boldsymbol{y}^r \in P^r(\boldsymbol{X}^r) \text{ and } \boldsymbol{y}^{r*} \in P^r(\boldsymbol{X}^r) \Rightarrow \forall \lambda \in [0,1] : \lambda \boldsymbol{y}^r + (1-\lambda)\boldsymbol{y}^{r*} \in P^r(\boldsymbol{X}^r)$$

Axiom 4 (Observability means feasibility)**.**

$$(\boldsymbol{y}^r, \boldsymbol{X}^r) \in S \Rightarrow \boldsymbol{y}^r \in P^r(\boldsymbol{X}^r).$$

Essentially, the first two axioms say that inputs and outputs are freely disposable. Axiom 3 states that, if input $\mathbf{X}^r$ can produce both output $\mathbf{y}^r$ and $\mathbf{y}^{r*}$, then it can also produce any convex combination of these outputs. We work in a setting with relaxed convexity assumptions: we assume convex output sets, but we do not impose convexity in the input-output space. A growing strand of literature assumes a weaker form of convexity, see Podinovski and Kuosmanen (2011) for an overview. The motivation is that a fully convex production set is not well suited for modeling economies of scale (Petersen (1990) and Bogetoft (1996)). Axiom 4 states that the observed input-output combinations are certainly feasible.

We add one final axiom, which includes returns to scale in the methodology. We assume either variable returns to scale (vrs), non-increasing returns to scale (nirs), non-decreasing returns to scale (ndrs) or constant returns to scale (crs). We include rts^r in the notation of the output set, indicating which returns to scale assumption we make for output r.

Axiom 5 (Output-specific returns to scale)**.**

$$\boldsymbol{y}^r \in P^r(\boldsymbol{X}^r, rts^r) \Rightarrow k\boldsymbol{y}^r \in P^r(k\boldsymbol{X}^r, rts^r) \text{ for } k \in K(rts^r)$$

where rts^r *='vrs', 'nirs', 'ndrs', or 'crs' and*
where $K(vrs) = \{1\}$, $K(nirs) = [0,1]$, $K(ndrs) = [1,\infty)$ *and* $K(crs) = \mathbb{R}_0$.

The returns to scale assumption describes the change in output resulting from a proportional change in inputs. If input $\mathbf{X}^r$ can produce output

$\mathbf{y}^r$, then $k\mathbf{X}^r$ can produce $k\mathbf{y}^r$ for $k \in K(rts^r)$. Depending on which returns to scale assumption that is made, the potential to scale up or down differs. Variable returns to scale is the weakest assumption, under which the input-output combinations can not be scaled. Under the assumption of non-increasing returns to scale, we can scale down the observations. Similarly, the assumption of non-decreasing returns to scale enables us to scale up the observations. Constant returns to scale is the strongest assumption and allows to scale both up and down.

We define the empirical approximation $\hat{P}^r(\mathbf{X}_n^r, rts^r)$ of the output set as the smallest set that is consistent with Axioms 1-5. This is an application of the minimum extrapolation principle which is commonly used in DEA, see Banker et al. (1984).

Illustrative example. Before giving a formal definition of the empirically constructed output set, we illustrate the construction with a single input, single output example, which is depicted in Figure 1.

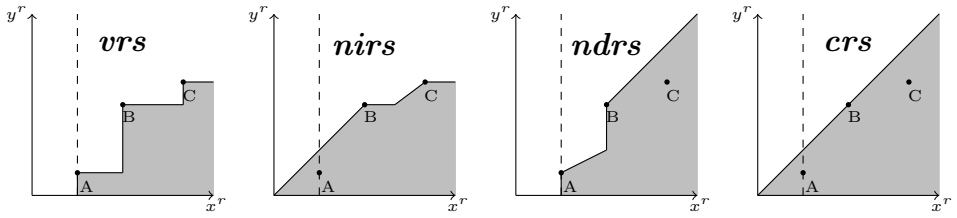


Figure 1: Production technology under variable, non-increasing, non-decreasing and constant returns to scale.

We observe the input-output combinations of DMU A, B and C. The grey area displays the technology set. The relation between the technology set T^r and the output sets $P^r(\mathbf{X}^r, rts^r)$ is the following: $T^r = \{(\mathbf{X}^r, \mathbf{y}^r, rts^r) | \mathbf{y}^r \in P^r(\mathbf{X}^r, rts^r)\}$. Along a vertical line we can therefore read an output set, for a particular input level.

In a first step in the construction, we apply Axiom 4, observability means feasibility. This axiom indicates that the technology set is constructed on the basis of the observed input-output combinations A, B and C. In a second step, Axioms 1 and 2 imply that the input-output combinations to the bottom right of A, B and C are feasible. Since we only have one output in our

example, Axiom 3 adds no additional information here. In a setting with one output, our technology corresponds to a free disposal hull technology. The free disposal hull (FDH) model (see Deprins et al. (1984) and Tulkens (1993)) does not require convexity, in contrast to the popular DEA model. In a final step, we include the returns to scale assumption. Figure 1 shows the technology sets under the four alternative returns to scale assumptions.

Formal construction. Petersen (1990) and Bogetoft (1996) define a scaling function β_s^r, which brings the production of DMU s on a similar scale as the production of DMU n:

$$\beta_s^r(\cdot, rts^r) : \mathbb{R}_0 \to \mathbb{R}_0 \cup \{-\infty\}$$

where

$$\beta_s^r(\mathbf{X}^r, rts^r) = \sup\{\beta \mid \beta \mathbf{X}_s^r \leq \mathbf{X}^r, \beta \in K(rts^r)\},$$

where we let $\sup(\emptyset) = -\infty$. The scaling parameter $\beta_s^r(\mathbf{X}_n^r, rts^r)$ relates the amount of input of DMU n to the input of DMU s and implies that $\beta_s^r(\mathbf{X}_n^r, rts^r)\mathbf{y}_s^r \in P^r(\mathbf{X}_n^r, rts^r)$ for a finite value of β. The scaling parameter therefore determines to what extent we should scale the output produced by DMU s for the scaled output of DMU s to figure as a benchmark for DMU n. The empirical output sets $\hat{P}^r(\mathbf{X}_n^r, rts^r)$ are then constructed on the basis of the scaled observations:

$$\hat{P}^r(\mathbf{X}_n^r, rts^r) = \left\{ \mathbf{y} \;\middle|\; \begin{array}{c} \mathbf{y} \leq \sum_{s \in C_n^r(rts^r)} \lambda_s^r \beta_s^r(\mathbf{X}_n^r, rts^r)\mathbf{y}_s^r \\ \sum_{s \in C_n^r(rts^r)} \lambda_s^r = 1, \lambda_s^r \geq 0 \end{array} \right\},$$

with $C_n^r(rts^r) = \{s | \beta_s^r(\mathbf{X}_n^r, rts^r) > 0\}$ the set of comparison partners for DMU n with respect to output r. Proposition 1 states that the output sets $\hat{P}^r$ satisfy the minimum extrapolation principle, under Axioms 1 to 5.

Proposition 1. *$\hat{P}^r(\boldsymbol{X}_n^r, rts^r)$ satisfies Axioms 1 - 5. Moreover, for any $P^r(\boldsymbol{X}_n^r, rts^r)$ that satisfies Axioms 1 - 5, we have $\hat{P}^r(\boldsymbol{X}_n^r, rts^r) \subseteq P^r(\boldsymbol{X}_n^r, rts^r)$.*

The set $\hat{P}^r(\mathbf{X}_n^r, rts^r)$ thus gives an inner bound approximation of the true output set $P^r(\mathbf{X}_n^r, rts^r)$, under the given technology axioms. We refer to Appendix A for the proof of Proposition 1. Now that we have constructed empirical approximations of the output sets, we can define output-specific efficiency.

2.3. Output-specific technical efficiency

Following Banker and Morey (1986), we allow for both discretionary and non-discretionary outputs. We therefore divide the vector $\mathbf{y}_n^r$ in a discretionary and a non-discretionary part: $\mathbf{y}_n^r = (\mathbf{y}_{Dn}{}^r, \mathbf{y}_{Fn}{}^r)$. We assume that both discretionary and non-discretionary outputs can be scaled. Although the original Banker and Morey model does not allow scaling for the non-discretionary part, a commonly used version of the model does allow scaling. See Syrjänen (2004) for a discussion on non-discretionary factors and scale in data envelopment analysis.

We define the following Farrell (1957) - Debreu (1951) efficiency measure for the production of output r:

$$\hat{\varphi}_n^r(rts^r) = \max\{\varphi | (\varphi \mathbf{y}_{Dn}{}^r, \mathbf{y}_{Fn}{}^r) \in \hat{P}^r(\mathbf{X}_n^r, rts^r)\}, \tag{3}$$

The measure $\hat{\varphi}_n^r(rts^r)$ captures the distance of DMU n to the boundary of the empirically constructed output set. Stated differently, $\hat{\varphi}_n^r(rts^r)$ indicates the equiproportionate expansion of discretionary output that is certainly feasible, under Axioms 1 - 5. In general, $1 \leq \hat{\varphi}_n^r(rts^r)$ with $\hat{\varphi}_n^r(rts^r) = 1$ indicating full output-specific technical efficiency. Since $\hat{P}^r(\mathbf{X}_n^r, rts^r) \subseteq P^r(\mathbf{X}_n^r, rts^r)$, the measure $\hat{\varphi}_n^r(rts^r)$ defines a lower bound for the true, but unobserved measure φ_n^r (with respect to the true, but unobserved output set $P^r(\mathbf{X}_n^r, rts^r)$).

The measure $\hat{\varphi}_n^r(rts^r)$ is straightforward to compute by a two-step procedure. In a first step, compute the values of the functions $\beta_s^r(\mathbf{X}_n^r, rts^r)$ and the DMUs in the set $C_n^r(rts^r)$. In a second step, the measure $\hat{\varphi}_n^r(rts^r)$ can be computed by solving the following linear programming problem:

$$\begin{aligned}
\hat{\varphi}_n^r(rts^r) \quad &= \quad \max_{\varphi_n \geq 0, \lambda_s^r \geq 0} \varphi_n \\
s.t. & \\
\text{(D-1)} \qquad & \sum_{s \in C_n^r(rts^r)} \lambda_s^r \beta_s^r(\mathbf{X}_n^r, rts^r) \mathbf{y}_{Ds}{}^r \geq \varphi_n \mathbf{y}_{Dn}{}^r \\
\text{(D-2)} \qquad & \sum_{s \in C_n^r(rts^r)} \lambda_s^r \beta_s^r(\mathbf{X}_n^r, rts^r) \mathbf{y}_{Fs}{}^r \geq \mathbf{y}_{Fn}{}^r \\
\text{(D-3)} \qquad & \sum_{s \in C_n^r(rts^r)} \lambda_s^r = 1
\end{aligned}$$

2.4. Output-specific returns to scale

To estimate returns to scale, we follow a method discussed by Podinovski (2004a) and Podinovski (2004b), based on Kerstens and Vanden Eeckaut (1999). For both observations on the production frontier and below the frontier, it is possible to determine the returns to scale. In the second case, we actually estimate the returns to scale of the projection of the (inefficient) observation on the production frontier. To determine the most appropriate returns to scale assumption, we asses the efficiency of the observation with respect to several production technologies, each based on an alternative returns to scale assumption. Since each output has its own production technology, we can estimate returns to scale for every individual output.

Traditionally, returns to scale are said to be either constant, increasing or decreasing. In a setting of relaxed convexity assumptions, Podinovski (2004a) and Podinovski (2004b) introduce a fourth option, namely sub-constant returns to scale. Sub-constant returns to scale indicate that the most productive scale size can be achieved by either reducing or increasing its scale. The identification of the returns to scale is based on the definition of the variable, non-increasing and non-decreasing returns to scale technologies:

- Constant returns to scale $\Leftrightarrow \hat{\varphi}_n^r(vrs) = \hat{\varphi}_n^r(nirs) = \hat{\varphi}_n^r(ndrs)$
- Decreasing returns to scale $\Leftrightarrow \hat{\varphi}_n^r(vrs) \leq \hat{\varphi}_n^r(nirs) < \hat{\varphi}_n^r(ndrs)$
- Increasing returns to scale $\Leftrightarrow \hat{\varphi}_n^r(vrs) \leq \hat{\varphi}_n^r(ndrs) < \hat{\varphi}_n^r(nirs)$
- Sub-constant returns to scale $\Leftrightarrow \hat{\varphi}_n^r(vrs) < \hat{\varphi}_n^r(ndrs) = \hat{\varphi}_n^r(nirs)$

Finally, to quantify output-specific scale efficiency, we follow Banker (1984) and we define a measure of scale efficiency as the ratio of the output-specific technical efficiency measure under constant returns to scale and the measure under variable returns to scale:

$$SE_n^r = \hat{\varphi}_n^r(crs)/\hat{\varphi}_n^r(vrs).$$

Comparing these efficiency measures gives an indication of the extent to which a DMU deviates from the point of optimal scale of operation.

Continuing the illustrative example displayed in Figure 1, we estimate that DMU A exhibits increasing returns to scale and DMU C decreasing returns to scale. DMU B exhibits constant returns to scale and therefore has a scale efficiency equal to 1.

2.5. Multi-output technical efficiency

Until now we focused on the production process of the individual outputs. However, the production of the individual outputs is linked through the use of joint (and subjoint) inputs. Following Cherchye et al. (2013) we define in this section multi-output efficiency measures that consider the production of all the outputs. An interesting feature of this methodology is that it allows for returns to scale that are specific to individual outputs. The vector $\mathbf{rts} = (rts^1, \ldots, rts^R)$ captures the returns to scale assumptions rts^r for every output r. We define

$$\hat{\varphi}_n(\mathbf{rts}) = \max\{\varphi | \forall r : (\varphi \mathbf{y}_{D_n}{}^r, \mathbf{y}_{F_n}{}^r) \in \hat{P}^r(\mathbf{X}_n^r, rts^r)\}. \tag{4}$$

In practice, this multi-output technical efficiency measure is computed as follows:

$$\hat{\varphi}_n(\mathbf{rts}) = \min\{\hat{\varphi}_n^1(rts^1), \ldots, \hat{\varphi}_n^R(rts^R)\}. \tag{5}$$

2.6. Robust methodology with environmental variables

Since the estimation of returns to scale is sensitive to outliers, we combine our methodology with the robust order-m method, as introduced by Cazals et al. (2002), discussed in Daraio and Simar (2007a) and elaborated for convex technologies in Daraio and Simar (2007b). The robust measure is computed by repeatedly drawing a sample of potential comparison partners for DMU n. For each random draw, we estimate the efficiency and the returns to scale of DMU n. The robust efficiency measure is then computed as the average over all draws. This procedure allows us to report the statistical significance of the estimations, which is based of the percentage of draws that leads to a particular returns to scale estimate. When discussing the estimation results, we will report the most frequently estimated type of returns to scale and the corresponding significance.

The robust order-m method is also well-suited to include environmental variables in the analysis using kernel weighting (see Daraio and Simar (2005)).[6] With this approach, we repeatedly draw a sample of size m (with replacement), whereby DMUs in a similar environment as the DMU under evaluation will have a larger probability to be drawn as a reference DMU.

[6] An alternative approach to account for environmental variables in a nonparametric efficiency evaluation, is to conduct a two-step procedure. However, this involves a separability assumption. See Simar and Wilson (2007) for an insightful discussion.

This approach is particularly interesting in the absence of information on the direction of influence of environmental factors. Furthermore, the suggested approach can be used to examine the effect of environmental factors. We refer to Appendix C for technical details on the estimation of the robust efficiency measure and how to examine the influence of environmental factors.

3. Empirical prison production model

For the empirical analysis of the prison size dilemma, we collected publicly available data, provided by the Ministry of Justice (MoJ), on 34 prisons in England and Wales. In England and Wales, prisons are divided into categories based on the severity of crime committed by inmates and the risk should the person escape. In order to obtain a sample that is sufficiently comparable, we focus on local male category B and C prisons. We collected data over the book years 2009/10, 2010/11 and 2011/12 and pool the data over the years to obtain 102 observations. By pooling the data, we impose that all observations operate under the same technology, but still allow they can vary with respect to returns to scale, scale efficiency and technical efficiency.

The empirical literature on prison efficiency estimation is scarce. Papers analyzing penitentiary institutions using DEA include Butler and Johnson (1997), who measure prison efficiency in the state of Michigan, Nyhan (2002) who consider juvenile justice facilities in the state of Florida, Hall et al. (2013) who assess Young Offender Institutions (YOIs) in England and Wales and Rogge et al. (2015) who assess various forms of efficiency of category B and C male prisons in England and Wales.

Depending on the setting and research question, these studies use a varying range of inputs and outputs. Though, there is consensus that national or regional crime figures should not be directly included as an output of a prison. A well-established literature, inspired by the seminal work of Becker (1968), shows positive, but highly accelerating diminishing returns from more incarcerations to reduce crime (See e.g. Levitt (1996), Buonanno and Raphael (2013), Di Tella and Schargrodsky (2013), Vollaard (2013), Hansen (2015) and references therein). However, it hard to disentangle the effects of individual prisons on crime rates. Moreover, crime figures relate to many outside prison characteristics such as neighbourhood-level and region-level socio-economic and demographic characteristics, the functioning of police and probation departments, etc.

We base our empirical model on the definition of the key prison performance areas as posed by the Ministry of Justice. The four key prison performance areas are (1) public protection, (2) decency, (3) reducing re-offending and (4) resource Management and operational effectiveness (see Ministry of Justice (2015)). The NOMS Prison Rating System creates a composite index of prison performance, based on imposing a priori chosen weights on quantitative and qualitative information that relates to the key performance areas. By construction, the PRS is not based on a production model that relates input(s) to outputs. Consequently, the PRS does not allow for testing (output-specific) economies of scale.

For the purpose of our scale (dis-)economies analysis, we analyze prison production in a multi-output production framework and we abstain from a priori imposing weights. We presume that the local prisons use resources — that come from different sources — to maximize three outputs: (1) Keeping convicts outside society, which we label as "*incarceration*", (2) To provide a humane prison environment and prepare prisoners for reintegration mainly by organizing purposeful and outside-cell activities, (3) Successfully reintegrating discharged prisoners into society.[7] As such, the key performance areas public protection, decency and reducing re-offending are covered by our three outputs. The fourth key performance area is captured by our multi-output production model. Figure 2 shows the empirical prison model we advocate to approximate the true conduct in local prisons in England and Wales, which is by nature multi-dimensional and complex.

As prisons do not operate in vacuum, we control for both regional and prison(er) characteristics, which we allow to be output-specific. As not all aspects of production are controllable for prison management, we distinguish between '*discretionary*' and '*non-discretionary*' output variables. Non-discretionary output variables are non-discretionary for the prison management, but are discretionary for higher-level decision makers such as the MoJ. This implies the non-discretionary outputs, in contrast to environmental variables, can be re-scaled to improve scale efficiency.

[7]While there is some discussion in the literature on whether harsh prison conditions have a deterrence effect (Katz et al. (2003)), we follow the MoJ and consider efforts to foster reintegration and quality of prison life as 'goods'. In England and Wales it is now fully acknowledged that the high proportion of offenders that re-offend is costly to the tax payers and society (see e.g. Ministry of Justice (2011, 2013, 2015).)

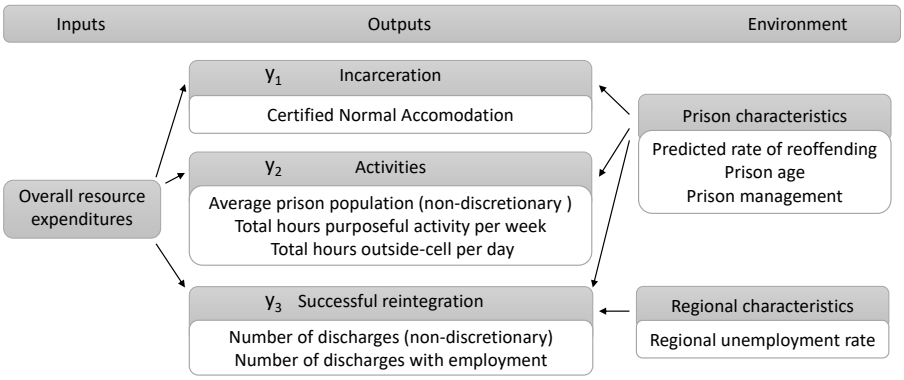


Figure 2: Empirical prison production model

Table 1 shows descriptive statistics for each of the included prison-year level variables. To approximate the inputs, outputs and environmental heterogeneity, we use proxies that are used in policy making and in our opinion best reflect the true production process:

Overall resource expenditures. We include in the model overall expenditures as joint input (tabulated in Table 1 per prisoner), which also includes outside-prison expenditures of collaborating agencies. We thus fully acknowledge that prisons in England and Wales are not stand alone institutions. They closely collaborate with institutions that strife for improving re-integration and reducing re-offending risk (i.e., MAPPA, the Probation Service and Primary Care Trusts). The expenditures are deflated to allow for comparison over time. We used the GDP deflator at market prices for financial years as provided by the HM Treasury and use restated versions of the overall expenditures that improve comparability over time.

A substantial part of the expenditures within prison is payroll. In 2010/11, 49 percent of overall costs relate to direct payroll costs (Freedom of Information Act, Ref. 72845/11), resulting in a correlation between overall resource expenditures and staff of 0.94. As overall resources is a more complete measure of prison input, we do not include prison officers or total staff separately into the analysis.

Prison officers traditionally keep inmates secure and maintain order. However, prison officers also promote anti-bullying and suicide prevention policies, take part in programmes to help prisoners reflect on their offending behaviour and prepare inmates for release through rehabilitation programmes.

We therefore consider the expenditures by prison, of which a large part goes to payroll, as a joint input, which simultaneously contributes to the production of all outputs.

Nevertheless, there is room for refining the analysis and increasing the discriminatory power of the proposed model by allocating resources to outputs. In principle, the methodology allows to include both joint and output-specific expenses in the analysis. For this to be possible, the MoJ could construct a data set – which is consistent over time and over prisons – that breaks down the overall prison costs. For particular costs, allocation to outputs is straightforward. For example, the costs of food provision (which is frequently outsourced) can be allocated to the incarceration output. Costs of lecturers, workshop places, material and leaders can be attributed to the provision of activities. Administrative costs related to the discharges and re-integration into society could be attributed to the particular output on successful re-integration. Further, as many prisons collaborate with outside-prison organizations to provide out-of-cell activities and re-integration programs, the contractual agreements can be used to allocate resources to the outputs concerning purposeful activities and successful re-integration. However, for a substantial part of costs, allocation is less straightforward. To structure the input allocation to the multiple outputs, we advocate the use of '*activity based costing*' (ABC, see Cooper and Kaplan (1988)). The distinguishing feature of ABC is that costs are first attributed to activities and subsequently, these activity costs are allocated to the outputs. In comparison to other costing methodologies, which often are based on the produced output quantities, ABC gives a much clearer and more accurate picture of the production model of the multi-output decision making unit, see Cherchye et al. (2013). As such, ABC offers a framework to allocate expenses to particular outputs.

Incarceration. To approximate the daily operations and administrative work that are needed to keep convicts outside society, we consider the places the prison offers to incarcerate prisons. In particular, we consider the Certified Normal Accommodation (CNA). By the Prison Act 1952, confining prisoners is only allowed in accommodation which is certified by an inspector that considers among others size, lighting, heating, communication-possibilities. Certified Normal Accommodation reflects the number of places the prison should not exceed (Prison Rules, 1999; rule 26). The respective cells are

available for immediate use. Damaged cells, cells affected by building works and cells taken out of use due to staff shortages are excluded.

Activities. We simplify the provision of a humane prison environment and preparation for reintegration to organizing purposeful and outside-cell activities. We include the total hours of purposeful activity per week[8] as an indicator of the effort during imprisonment that is taken to '*break the cycle*' by getting the prisoners to work and train outside their cell (see Ministry of Justice (2011)). In addition, we include time outside-cell[9] to fully acknowledge the beneficial aspects of other outside-cell activities such as sports and recreation. The average prison population is included as non-discretionary output (which is non-discretionary to prison management but can be rescaled by higher-level decision makers) to control for the quantity of inmates for which purposeful and outside-cell activities can be provided.

Note that the average prison population in the sample is generally higher than the number of places according to the certified normal accommodation. We refer to Rogge et al. (2015) for a discussion on capacity (over)-utilization. Overcrowding is a well-known issue in UK prisons and can make rehabilitation more difficult as prisoners have reduced access to purposeful activity. We therefore include the average prison population as a non-discretionary variable for the activities output, to control for the number of prisoners that take part in purposeful activities, but without assuming that policy makers have the intention to maximize the average prison population (which could lead to overcrowding). The study of the discretionary and non-discretionary aspects of overcrowding goes beyond this paper. For the output incarceration, the variable Certified Normal Accommodation is preferred over average prison population as it directly measures the operational capacity of a prison to incarcerate convicts.

Successful reintegration. Successful reintegration is proxied by focusing on employment at discharge. Employment at release is a direct indicator of successful reintegration. Promoting employment at release is challenging.

[8]This is calculated as the average number of hours purposeful activity per prisoner per week times the average prison population. This output follows Nyhan (2002) who also used a like variable "*percentage of all youths who successfully completed the required program or were transferred to aftercare or to a less restrictive level.*" (page. 429).

[9]This is calculated as 24 minus the average time within cell per prisoner, times the average prison population.

At most, 44 percent of prisoners have employment at release date. The number of discharges is included as non-discretionary output to control for the quantity of offenders that are released. Local prisons hold offenders with short sentences resulting in more discharges than the yearly average prison population.

Regional characteristics. Successful reintegration also highly depends on the socio-economic environment in which prisoners are reintegrated. As this study deals with local prisons, we include the regional male unemployment rate as output-specific environmental variable for successful reintegration. This variable directly relates to the employment opportunities of the discharged prisoners. The unemployment rates were retrieved from the Labour Force Survey (LFS), which is the largest household survey in the UK and provides the official measures of unemployment.

Prison characteristics. Input requirements depend on the security level of prisoners (Butler and Johnson, 1997). While we focus on 'similar' local prisons, the three aspects of prison operation can still be conditioned by the heterogeneity in prisoner inflow. To take this heterogeneity into account, we include the '*predicted rate of re-offending*' in an establishment as an environmental variable. The probability of re-offending is estimated at prison level by the Ministry of Justice (2011), based on prisoner-level data on social background, ethnicity, crime type, etc.

Furthermore, we control in our analysis of scale economies for prison age. Renewing and rescaling the prison estate is a key part of the strategy of the NOMS (see the Introduction). Lockyer (2013) argues that prison age directly affects the production process of prisons. Given the large time gaps between construction years in our sample (e.g., no prison was built between 1891 and 1991), we divide the prisons in the sample into 3 age categories: prisons opened before 1837 (category 1), prisons opened between 1837 and 1901 (the Victorian era, category 2) and prisons opened after 1990 (category 3). Our sample consists of 7 prisons in the first category, 18 prisons in the second category and 9 prisons in the third category. We include the age categorization in the analysis as an ordered categorical environmental variable, which may affect all three aspects of prison production.

Last, as the local prisons differ in terms of prison management (see Rogge et al. (2015) for a discussion), we include a dummy that denotes whether the prison is a privately managed or a public prison. For the considered period,

we have five privately managed prisons in our dataset. G4S Justice Services (G4S) manages HMP Altcourse, HMP Birmingham (since October 2011) and HMP Parc. HMP Doncaster is managed by Serco Custodial Services (Serco) and HMP Forest Bank is managed by Sodexo Justice Services (SJS).

Table 1: Descriptive statistics

	Year	Mean	St.Dev.	0%	25%	50%	75%	100%
Prison inputs								
Overall resource expenditures per prisoner	2009/10	36601.76	7321.32	27248.38	32391.31	34962.52	38092.25	63691.60
(overallres)	2010/11	36697.52	8020.31	24938.20	32185.38	34627.37	39343.51	65503.96
	2011/12	34556.44	7150.36	23736.67	30400.02	32253.57	36836.95	59392.22
Incarceration								
Certified normal accommodation	2009/10	666.21	291.94	145.00	449.25	646.67	872.75	1186.00
(cna)	2010/11	688.76	299.74	146.00	464.50	682.00	938.00	1187.00
	2011/12	679.00	299.86	162.00	466.00	642.50	906.00	1187.00
Activities								
Average prison population	2009/10	879.63	351.71	232.92	638.02	843.92	1165.77	1653.58
(avpop)	2010/11	881.38	349.07	228.00	621.00	891.50	1170.25	1621.00
	2011/12	875.53	336.36	223.00	660.75	845.00	1121.75	1544.00
Average hours purposeful activity per week per prisoner	2009/10	20.63	3.72	16.29	18.09	19.91	22.29	34.98
(avpurp)	2010/11	21.32	3.83	16.90	18.68	20.48	23.30	35.20
	2011/12	20.99	3.61	16.81	18.52	20.05	22.11	33.73
Average hours outside-cell per day per prisoner	2009/10	8.16	1.30	5.60	7.30	7.90	9.20	12.10
(outcell)	2010/11	8.40	1.30	5.80	7.80	8.10	9.00	12.90
	2011/12	8.46	1.37	5.50	7.80	8.15	9.10	12.40
Successful reintegration								
Number of Discharges	2009/10	1309.71	558.38	419.00	898.50	1207.00	1593.00	2933.00
(discharges)	2010/11	1417.60	519.15	460.00	1086.38	1376.50	1771.50	2575.50
	2011/12	1525.50	576.55	417.00	1169.25	1421.00	1973.50	2839.00
Percentage of discharges with employment	2009/10	24.89	6.68	12.80	19.93	23.60	28.80	41.70
(emprate)	2010/11	27.68	7.37	14.00	22.00	28.00	31.00	44.00
	2011/12	27.18	7.10	14.00	22.25	27.00	31.00	44.00
Prison characteristics								
Predicted rate of re-offending	2009/10	62.76	2.84	54.16	61.61	63.00	65.05	67.17
(predreof)	2010/11	62.54	5.52	51.00	58.88	62.27	67.33	77.28
	2011/12	62.09	4.76	53.40	58.87	61.20	65.91	74.18
Year of construction (age)		1888.18	68.01	1782	1842	1864.50	1991	2000
Regional characteristics								
Regional male unemployment rate	2009/10	8.61	1.53	6.30	7.10	9.10	9.80	11.20
(regunemp)	2010/11	9.39	1.85	6.70	7.70	10.10	10.50	13.00
	2011/12	8.19	1.42	6.20	6.60	8.30	9.40	10.30

Table 2: Spearman rank correlogram

	overallres	cna	avpop	avpurp	outcell	discharges	emprate	predreof	agecat	regunemp
overallres	1.00	0.87	0.88	0.11	0.11	0.67	0.07	-0.42	0.54	0.49
cna	0.87	1.00	0.94	0.09	0.04	0.78	-0.06	-0.45	0.52	0.48
avpop	0.88	0.94	1.00	0.15	0.09	0.83	-0.13	-0.34	0.48	0.56
avpurp	0.11	0.09	0.15	1.00	0.68	0.27	0.16	-0.06	0.27	-0.06
outcell	0.11	0.04	0.09	0.68	1.00	0.15	0.27	-0.07	0.30	-0.05
discharges	0.67	0.78	0.83	0.27	0.15	1.00	-0.10	-0.33	0.46	0.46
emprate	0.07	-0.06	-0.13	0.16	0.27	-0.10	1.00	-0.42	0.01	-0.31
predreof	-0.42	-0.45	-0.34	-0.06	-0.07	-0.33	-0.42	1.00	-0.25	0.02
agecat	0.54	0.52	0.48	0.27	0.30	0.46	0.01	-0.25	1.00	0.18
regunemp	0.49	0.48	0.56	-0.06	-0.05	0.46	-0.31	0.02	0.18	1.00

Table 2 shows the correlation between outputs, input and environmental variables. Output variables that relate to qualitative aspects of prison production are scaled per prisoner as in Table 1. There is modest correlation between input and outputs (even if not scaled per prisoner), indicating there can be deviations from optimal conduct or effects from the heterogeneity in the operating environment. Overall resources are positively associated with higher numbers of purposeful activity per prisoner (avpurp) and with time outside-cell (outcell). Prison characteristics are related to the size of the prison. The correlation between input and respectively the predicted rate of re-offending and the ordered prison age category, is respectively -0.42 and 0.54. Stated differently, larger prisons are generally more recently built and incarcerate prisoners with characteristics that imply a lower expected rate of re-offending. Conditioning on the operating environment is thus needed to meaningfully analyze returns to scale. The regional unemployment rate is negatively related to employment at discharge. Overall, Table 2 indicates a single output analysis cannot capture the production process of prisons as it would imply an omitted variable bias. We need an empirical analysis that includes multi-output structure and environmental heterogeneity to meaningfully estimate output-specific returns to scale in prison production.

4. Results

The methodology allows for a simultaneous analysis of output-specific scale efficiency and technical efficiency. We first discuss the results on scale efficiency in subsection 4.1 and subsequently turn to technical efficiency in subsection 4.2. As discussed in the methodology section, the order-m subsample bootstrapping routine (with m=50 and 1000 random draws)[10] is ap-

[10]The value of m is chosen on the basis of visually inspecting the relation between the proportion of observations with $\varphi^m(v) < 1$ and the value of m as in Daraio and Simar

plied to lower the sensitivity of the efficiency estimates to potential outliers and extreme noise. In particular, the order-m efficiency measure is computed as the average over all draws. The order-m efficiency score can be interpreted as the expected efficiency score relative to a subsample of $m = 50$ prisons. Table 8 in Appendix C shows the prison-specific and output-specific bandwidth sizes, which are the basis for the weighted bootstrap routine to include influences of environmental variables. Overall, prison age and prison management categories are estimated to have highly dissimilar production environments, as represented by the bandwidth sizes of the Liracine kernels that are close to the lower bound 0, meaning nearly no weight is given to observations from other categories. The direction of influence of the environmental variables is discussed in Appendix C.

4.1. Output-specific returns to scale and scale efficiency

By applying the advocated framework on the empirical prison production model, we can examine whether output-specific returns to scale estimates differ considerably over outputs, implying a prison size dilemma for the public policy maker.

Incarceration. Figure 3(a) shows returns to scale estimates and scale efficiency when the focus is solely on incarceration. The higher the values above 1, the more room for improvement. Scale efficiency estimates for the small and large prisons have values surpassing 1.5, indicating potential efficiency gains of over 50% by rescaling these prisons. These results are not surprising, since the cost per place varies between 31200 and 65500 pounds per year.

Given that the estimates are conditional upon the environment, it is possible for example that a particular prison is characterized by increasing returns to scale and that an even smaller prison is characterized by decreasing returns to scale. The reason is that both prisons are situated in a different environment. Although there is overlap between the returns to scale estimates, Figure 3(a) shows a clear pattern. Smaller prisons are generally characterized by increasing returns to scale and larger prisons are characterized by decreasing returns to scale.

Over the three book years, we find that 52 observations are characterized with decreasing returns to scale and 45 with increasing returns to scale. For respectively 35 and 32 observations the returns to scale estimates are

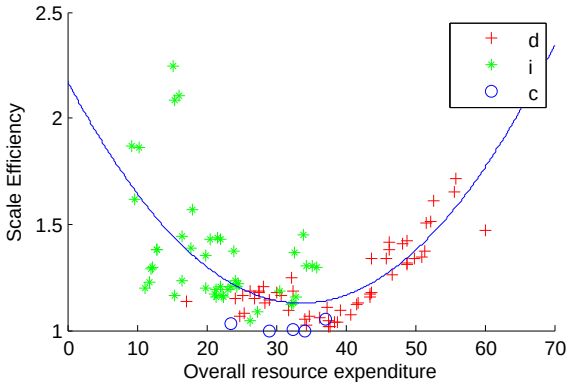
(2007a).

significant at the 95% confidence level (see Table 5, 6 and 7). We find 5 prisons characterized by constant returns to scale, whereof 3 significant at the 95% confidence level. The returns to scale estimates show no year- specific patterns.

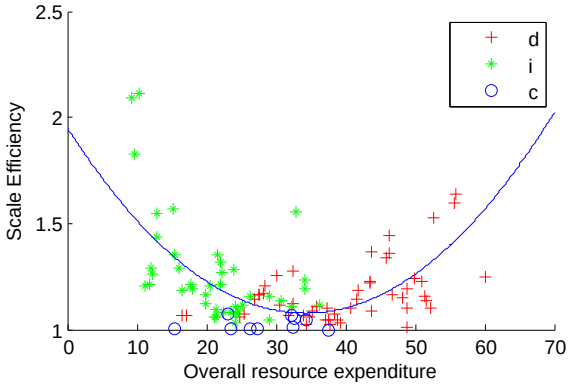
The optimal scale size in terms of resources is situated between 23 and 37 million pounds. In terms of Certified Normal Accommodation, the prisons characterized by constant returns to scale provide between 554 and 1187 places. We therefore conclude that the optimal scale size of a prison with respect to incarceration is medium scale, depending on the environment.

Activities. Figure 3(b) shows the estimates for the models that include the output variables that proxy purposeful and outside-cell activities to promote humane incarceration and to prepare inmates for reintegration into society. With this focus, we characterize 50 observations with *drs*, 42 with *irs* and 10 with *crs*. For respectively 46, 24 and 5 observations this is significant at the 95% confidence level. Overall, we find a similar pattern of returns to scale and scale efficiency as for the model focusing on incarceration. Focusing on purposeful and outside-cell activities, the prisons characterized by constant returns to scale provide between 322 and 1073 places. In terms of resources, the optimal scale size to provide activities is a little more spread than before.

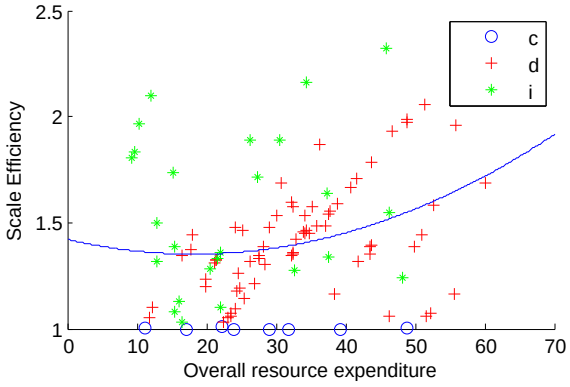
Successful reintegration. For successful reintegration we characterize respectively 69, 25 and 8 observations with respectively *drs*, *irs* and *crs*. For respectively 47, 18 and 7 observations the *rts* estimates are significant. Figure 3(b) shows a pattern of returns to scale and scale efficiency which highly differs in terms of optimal scale size over the considered environmental variables. The most productive scale size is in the broad interval of 11 to 49 million pounds and corresponds to prisons with a number of places between 221 and 1064. Although smaller prisons can be optimal to provide reintegration, we find just as well medium scale prisons with an optimal scale size. Most probably, successful reintegration is highly dependent on unobserved heterogeneity in the effectiveness of reintegration programmes which is given the sample of observations difficult to disentangle from economies of scale.



(a) Incarceration



(b) Activities



(c) Successful reintegration

Figure 3: Returns to scale and scale efficiency, in function of overall resource expenditure, for each of the three output-objectives.

In sum, our results over the three considered outputs reject the idea that public managers are faced with a prison size dilemma, which implies a choice between cost-per-place performance and social performance. We cannot reject medium scale to be optimal. Of course, the optimal scale size depends on the operating environment. For both incarceration and providing purposeful and outside-cell activities, we find supportive evidence that a medium scale size is optimal. For successful reintegration, we find no supportive evidence for drastic productivity gains by moving towards a very small or large prison scale.

Prison-specific estimates. The overall finding that there is no prison size dilemma requires further prison-level consideration. Table 5, 6 and 7 in appendix show that for 14 observations, we do find a prison size dilemma in the sense that we estimate a prison with output-specific technologies to be simultaneously characterized by drs and irs, both significant at the 95% confidence level. For example, for HMP Exeter in book years 2010/11 and 2011/12, we find it is optimal to scale down prison scale to improve successful reintegration and scale up prison scale to improve the provision of purposeful and outside-cell activities and keeping convicts outside society by incarceration. From an operational viewpoint, allowing for output-specific and environment-specific returns to scale can be a valuable tool to provide policy advice on re-scaling prison conduct, taking the complexity of multi-output production into account.

4.2. Technical efficiency

Table 3 shows conditional order-m technical efficiency estimates for the variable returns to scale model ($\varphi^m(v)$) and the constant returns to scale model ($\varphi^m(c)$). Conditional order-m efficiency is reached when $\varphi^m = 1$. Note that the prison under evaluation is not necessarily included in the randomly drawn subsamples. Consequently, the order-m efficiency scores might be smaller than one. If the efficiency score is smaller than one, a prison is called super efficient. Overall, values of φ^m larger than one indicate that, on the basis of a subsample of m prisons, we estimate that there is room to proportionally increase the production, given the input and environment.

The first two columns show the respective vrs and crs results for the three outputs analyzed simultaneously, but allowing for output-specific technologies. The other columns show the results for the models that include only one output next to the input and output-specific environment.

Table 3 shows the technical efficiency of local male prisons in England and Wales is improving over time. Considering the multi-output model, on average, the room for increasing production went from 4 percent in 2009/10 to less than 0 percent (thus indicating super efficiency) in 2011/12. Stated differently, our estimates support the idea that the public policy of the coalition at place since 2010 was, at least partly, successful in reducing inefficiencies. Still, some prisons considerably and persistently underperform (see Table 5, 6 and 7 in appendix). For example, technical efficiency of HMP Belmarsh is estimated to be respectively 1.29, 1.25 and 1.25 in the three consecutive book years considered.

In sum, using the advocated framework to consider multi-output prison production, we are able to pinpoint low performers in terms of both scale efficiency and technical efficiency. The persistent low performers require further attention from public managers. Are there additional prison characteristics that could explain the low figures?

Table 3: Mean (standard deviation) of conditional order-m efficiency scores for m=50 and B=1000

	Multi-Output		Incarceration		Activities		Reintegration	
Year	$\varphi^m(v)$	$\varphi^m(c)$	$\varphi^{m,1}(v)$	$\varphi^{m,1}(c)$	$\varphi^{m,2}(v)$	$\varphi^{m,2}(c)$	$\varphi^{m,3}(v)$	$\varphi^{m,3}(c)$
2009/10	1,04	1,28	1,12	1,43	1,11	1,37	1,67	2,36
	(0,13)	(0,29)	(0,16)	(0,31)	(0,20)	(0,38)	(0,68)	(0,90)
2010/11	1,01	1,19	1,07	1,37	1,09	1,29	1,33	1,93
	(0,06)	(0,24)	(0,11)	(0,30)	(0,16)	(0,33)	(0,48)	(0,89)
2011/12	0,98	1,14	1,05	1,32	1,04	1,23	1,31	1,87
	(0,06)	(0,21)	(0,13)	(0,30)	(0,14)	(0,30)	(0,68)	(1,25)

4.3. Sensitivity analysis

Sensitivity analysis shows that our results are robust when altering the specification of the prison production model and including additional qualitative variables related to the prison production process.

Firstly, we include an additional qualitative variable in the model related to safety and the quality of life within prison, by employing data on the number of prisoner-on-officer assaults. It is worth to note that data on the number of prisoner-on-prisoner assaults is also available, but these numbers are less reliable since incidents between prisoners often remain unreported. The number of prisoner-on-officer assaults per prison per year varies between 3 and 102, with an average over all observations of 35 incidents per year. Since the number of prisoner-on-officer assaults is an undesirable output, we use one divided by the number of prisoner-on-officer assaults as a proxy for

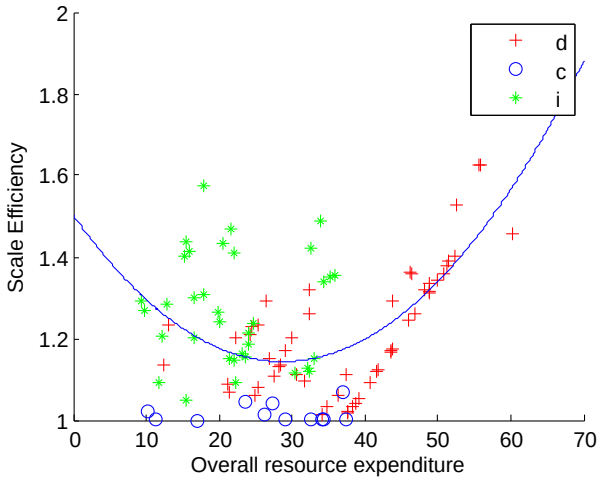


Figure 4: Returns to scale and scale efficiency for incarceration, taking assaults into account.

safety and the quality of life within prison. We include this proxy as an additional variable for the incarceration output and repeat the analysis as a robustness check for our conclusions with respect to the optimal scale size of prisons. We dropped one observation due to a missing value for the number of prisoner-on-officer assaults and consequently performed the robustness check for 101 observations. The average order-m efficiency for incarceration equals 1,08 under variable returns to scale and 1,31 under constant returns to scale. These efficiency scores are slightly lower than the scores in the original model, due to the inclusion of an additional output. Figure 4 plots the returns to scale and scale efficiency for the incarceration output, taking prisoner-on-officer assaults into account. The general pattern remains similar to Figure 3(a). On average, the scale efficiency equals 1,21. When including assaults, 12 observations are characterized by constant returns to scale. Remarkably, 3 of those observations correspond to smaller scale prisons. However, most smaller scale prisons remain scale inefficient.

Similarly, we extend the model to include accommodation at release, which is a necessary condition of successful reintegration. In particular, we include the percentage of prisoners with accommodation at release times the number of discharges in a prison. The percentage of prisoners having accommodation at release varies between 67% and 100%. On average, 88% of prisoners have accommodation at release. The average order-m efficiency

score for the reintegration output is respectively 1,25 and 1,68 under variable and constant returns to scale. Including accommodation at release, 17 observations are characterized by constant returns to scale. Figure 5 shows that the optimal scale sizes remain equally scattered as in the original model, however, medium-scale prisons tend to be slightly more scale efficient when taking accommodation at release into account. We conclude that our main findings are robust to including additional qualitative variables.

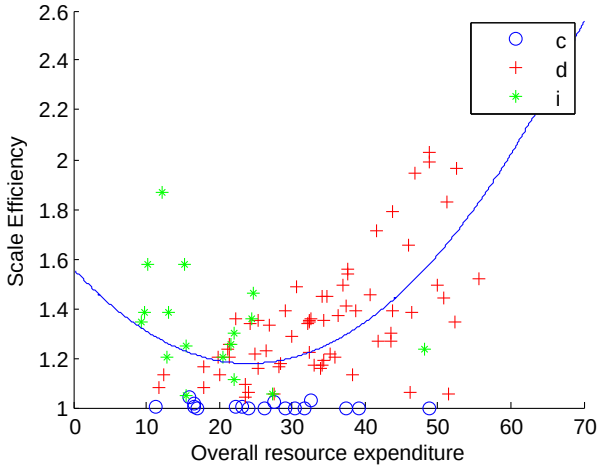


Figure 5: Returns to scale and scale efficiency for successful reintegration, taking accommodation at release into account

Furthermore, we have investigated whether our results are robust to including additional socio-economic environmental variables in our prison production model. In particular, we focused on the regional deflated GDP per capita as a continuous environmental variable and the region as a categorical environmental variable. The results (available upon request) show that our findings are robust for including additional regional characteristics. Overall, we find no indications that our results concerning successful reintegration are driven by regional heterogeneity.

Next, we have tested whether there might be a time lag on the prison production process. To be concrete, we linked the inputs of a particular year to the output of the following year. In the model with time lag, the expenses of the book year 2009/2010 are linked to the output of the year 2010/2011 and the expenses of 2010/2011 are linked to the output of 2011/2012. We obtain a dataset with 68 observations. Since this dataset is smaller than the

original dataset with 102 observations, we select the observations from the book years 2010/2011 and 2011/2012 to obtain a sample also consisting of 68 observations, in which the expenses of the prison are linked to output in the same book year. Table 4 shows the results of the model with and without a time lag. However, both models result into comparable average efficiency scores.

Table 4: Mean (standard deviation) of model with and without time lag (N=68)

.

	Multi-Output		Incarceration			Activities			Reintegration		
	$\varphi(v)$	$\varphi(c)$	$\varphi^1(v)$	$\varphi^1(c)$	SE^1	$\varphi^2(v)$	$\varphi^2(c)$	SE^2	$\varphi^3(v)$	$\varphi^3(c)$	SE^3
Lag	1,01	1,15	1,07	1,33	1,25	1,06	1,26	1,18	1,19	1,47	1,24
	(0,08)	(0,21)	(0,14)	(0,28)	(0,23)	(0,14)	(0,31)	(0,18)	(0,40)	(0,53)	(0,27)
No	1,01	1,14	1,05	1,33	1,27	1,06	1,25	1,17	1,22	1,57	1,26
lag	(0,06)	(0,22)	(0,11)	(0,28)	(0,25)	(0,12)	(0,30)	(0,18)	(0,44)	(0,80)	(0,27)

5. Conclusion

There is little reason to expect public firms to operate on their optimal scale size in the absence of competitive pressure. For prisons in England and Wales, there is a widespread policy debate concerning whether very small housing blocks or very large, so called '*titan*' prisons are the solution to improve the productivity of prisons. The general belief is that small prisons can provide a safe and humane environment wherein prisoners can be well prepared to reintegrate into society and large prisons are especially effective when the focus is on expenditures-per-prisoner. If indeed the case, public managers would face a prison size dilemma. However, it is unclear whether the observed data support the idea of a prison size dilemma, as the policy debate does not go beyond an anecdotal discussion at most supported with partial indicators of reintegration and costs.

We provide a thorough examination of economies of scale using a complete multi-output assessment that allows for interlinkages between the output of incarcerating convicts and qualitative outputs (i.e., purposeful and outside-cell activity, successful reintegration) and allows that economies of scale can differ between the different qualitative dimensions of production.

Although our focus is on economies of scale in prisons, it is worth to note that the advocated methodological framework is more generally applicable to multi-output public sector organizations and multi-output manufacturing plants. For example Duncombe and Yinger (1993) study economies of

scale in different dimensions of public production, with an application to fire protection.

With respect to prisons in England and Wales, we do not find supportive evidence for the idea that public managers are confronted with a prison size dilemma. The main conclusion is that we cannot reject medium scale to be optimal. This conclusion is supported by two observations. Observation 1: depending on the operating environment, we find that medium scale size is optimal for both incarceration and providing purposeful and outside-cell activities. Observation 2: for successful reintegration, the results are mixed, but we do not find indications for drastic productivity gains by moving towards a very small prison scale.

Our results are therefore supportive for a policy oriented towards 'titan' prisons, which are operated as a number of semi-autonomous units sharing a common site and set of services. However, it is worth noting that our results are based on a given set of observations of prison production. The building of very small and very large prisons can coincide with the introduction of new technologies, making extrapolation from the observed set of prison production difficult. Further research is needed to examine the optimal scale size to introduce productivity enhancing technological innovations.

Furthermore, the pillar of the UK 2010 coalition concerning the reduction of technical inefficiencies within-prison is estimated to be, at least partly, successful. The technical efficiency is improving over the considered period 2009/10-2011/12 with the exception of the output successful reintegration.

We demonstrate the value of the multi-output production framework to analyze and test for output-specific scale (dis-)economies, using publicly available local prison data provided by the Ministry of Justice. As such, we provide a framework to test the success of recent policies to lower average costs by changing the scale of prisons.

This paper introduces a framework that fosters further research. First, the Ministry of Justice could further increase the discriminatory power of the methodology by applying the advocated methodology with detailed information on the allocation of expenses to outputs and obtain even more detailed insight into the multi-output production process of prisons. Second, while we consider the prison size dilemma, both public and private sector policy making is confronted with similar rescaling dilemmas. Examples include the debate on rescaling local police departments and the optimal scale size of bank branches. Last but not least, further research is needed on whether key findings from the productivity decomposition literature (e.g. at the firm-

level or aggregated at the sector-level or country-level) are sensitive for the introduction of varying scale efficiency over outputs.

Acknowledgements

We would like to thank three anonymous referees, Laurens Cherchye, Bram De Rock, Victor Podinovski, the participants of the 14th European Workshop on Efficiency and Productivity Analysis in Helsinki and of the 5th workshop on efficiency and productivity analysis in Porto for their insightful comments and suggestions. Veerle Hennebel gratefully acknowledges financial support from the Fund for Scientific Research - Flanders (FWO).

Appendix

A. Proof of proposition 1.

Proof. We first verify that $\hat{P}^r(\mathbf{X}_n^r, rts^r)$ satisfies Axioms 1 - 5. Axiom 4 follows from the definition of β_n^r and $\hat{P}^r(\mathbf{X}_n^r, rts^r)$. Since $\beta_n^r(\mathbf{X}_n^r, rts^r) = 1$ and $n \in C_n^r(rts^r)$, we have that $\mathbf{y}_n^r \in \hat{P}^r(\mathbf{X}_n^r, rts^r)$. Moreover, Axioms 1 and 3 follow directly from the construction of $\hat{P}^r(\mathbf{X}_n^r, rts^r)$ as the convex-monotone hull of the scaled output vectors of the DMUs in the set $C_n^r(rts^r)$. To verify Axiom 2, suppose that $\mathbf{X}^r \leq \mathbf{X}^{r*}$. By definition of β_s^r, we have that $\beta_s^r(\mathbf{X}^r, rts^r) \leq \beta_s^r(\mathbf{X}^{r*}, rts^r)$ and consequently $C_n^r(rts^r) \subset C_n^{r*}(rts^r)$. This implies that $\hat{P}^r(\mathbf{X}^r, rts^r) \subseteq \hat{P}^r(\mathbf{X}^{r*}, rts^r)$. Lastly, Axiom 5, is satisfied since $\beta_s^r(k^r\mathbf{X}^r, rts^r) = k^r\beta_s^r(\mathbf{X}^r, rts^r)$ for all $k^r \in K(rts^r)$. Then $\mathbf{y}^r \in \hat{P}^r(\mathbf{X}^r, rts^r)$ implies that $k^r\mathbf{y}^r \in \hat{P}^r(k^r\mathbf{X}^r, rts^r)$ for $k^r \in K(rts^r)$. We conclude that $\hat{P}^r(\mathbf{X}_n^r, rts^r)$ satisfies Axioms 1 - 5.

It remains to prove that for any $P^r(\mathbf{X}_n^r, rts^r)$ that satisfies Axioms 1 to 5, we have that $\hat{P}^r(\mathbf{X}_n^r, rts^r) \subseteq P^r(\mathbf{X}_n^r, rts^r)$. Take any $\mathbf{y}^{r*} \in \hat{P}^r(\mathbf{X}_n^r, rts^r)$. We need to prove that $\mathbf{y}^{r*} \in P^r(\mathbf{X}_n^r, rts^r)$. By the definition of $\hat{P}^r(\mathbf{X}_n^r, rts^r)$, we have given that

$$\mathbf{y}^{r*} \leq \sum_{s \in C_n^r(rts^r)} \lambda_s^r \beta_s^r(\mathbf{X}_n^r, rts^r)\mathbf{y}_s^r$$

for some $\lambda_s^r \geq 0$ such that $\sum_{s \in C_n^r(rts^r)} \lambda_s^r = 1$. We can now prove that $\mathbf{y}^{r*} \in P^r(\mathbf{X}_n^r, rts^r)$ by using that $P^r(\mathbf{X}_n^r, rts^r)$ satisfies Axioms 1 to 5. First, Axiom 4 implies that $\mathbf{y}_s^r \in P^r(\mathbf{X}_s^r, rts^r) \forall s$. Using Axiom 5 we have that

$\beta_s^r(\mathbf{X}_n^r, rts^r)\mathbf{y}_s^r \in P^r(\beta_s^r(\mathbf{X}_n^r, rts^r)\mathbf{X}_s^r, rts^r)$. By definition of β_s^r we have that $\beta_s^r(\mathbf{X}_n^r, rts^r)\mathbf{X}_s^r \leq \mathbf{X}_n^r$. Together with Axiom 2 this implies that

$$\beta_s^r(\mathbf{X}_n^r, rts^r)\mathbf{y}_s^r \in P^r(\mathbf{X}_n^r, rts^r) \quad \forall s.$$

Combining the definition of $\hat{P}^r(\mathbf{X}_n^r, rts^r)$ with Axiom 3, this results into

$$\mathbf{y}^{r*} \leq \sum_{s \in C_n^r(rts^n)} \lambda_s^r \beta_s^r(\mathbf{X}_n^r, rts^r)\mathbf{y}_s^r \in P^r(\mathbf{X}_n^r, rts^r)$$

for $\lambda_s^r \geq 0$ such that $\sum_{s \in C_n^r(rts^r)} \lambda_s^r = 1$. Finally, Axiom 1 implies that $\mathbf{y}^{r*} \in P^r(\mathbf{X}_n^r, rts^r)$. □

B. Prison-specific estimates

Table 5: Order-m efficiency scores and returns to scale estimates for $m = 50$ and $B = 1000$ (2009/10)

Year 2009/10	Multi Output		Incarceration				Activities				Reintegration			
Prison	$\varphi(v)$	$\varphi(c)$	$\varphi^1(v)$	$\varphi^1(c)$	SE^1	RTS	$\varphi^2(v)$	$\varphi^2(c)$	SE^2	RTS	$\varphi^3(v)$	$\varphi^3(c)$	SE^3	RTS
Altcourse (G4S)	1,00	1,00	1,26	1,79	1,42	d***	1,00	1,01	1,01	d***	1,00	1,00	1,00	c**
Bedford	1,00	1,00	1,00	1,13	1,13	d	1,00	1,07	1,07	d	1,00	1,00	1,00	c***
Belmarsh	1,29	2,14	1,29	2,14	1,65	d***	1,77	2,81	1,59	d***	3,61	4,20	1,16	d***
Birmingham	0,99	1,21	1,07	1,25	1,17	d***	0,99	1,21	1,22	d**	1,17	1,63	1,39	d
Bristol	0,99	1,08	1,17	1,39	1,19	i	0,99	1,08	1,09	i*	1,60	2,12	1,32	d***
Brixton	1,00	1,13	1,00	1,13	1,13	d	1,41	1,70	1,21	d**	2,12	2,75	1,30	d*
Bullingdon	1,00	1,19	1,14	1,19	1,05	d*	1,00	1,20	1,19	i	2,22	3,21	1,45	d
Cardiff	1,02	1,20	1,02	1,23	1,20	d	1,03	1,20	1,17	d**	1,47	2,04	1,39	d*
Chelmsford	1,02	1,10	1,02	1,10	1,08	d	1,21	1,30	1,08	d**	1,30	1,48	1,14	d*
Doncaster (Serco)	1,00	1,14	1,00	1,18	1,19	i***	1,02	1,14	1,12	d**	1,08	2,04	1,89	i**
Dorchester	1,00	1,99	1,07	1,99	1,86	i***	1,00	2,12	2,11	i***	1,14	2,24	1,97	i***
Durham	1,06	1,29	1,09	1,29	1,18	d**	1,06	1,35	1,28	d**	2,48	3,92	1,58	d*
Exeter	0,96	1,15	1,02	1,59	1,57	i***	0,96	1,15	1,19	i*	1,49	2,15	1,44	d
Forest Bank (SJS)	1,00	1,02	1,14	1,16	1,03	d	1,00	1,02	1,02	d***	1,03	2,24	2,16	i***
Gloucester	1,00	1,30	1,00	1,30	1,30	i***	1,05	1,33	1,26	i***	1,39	1,53	1,10	d*
High Down	1,04	1,15	1,04	1,15	1,11	d***	1,45	1,59	1,10	d*	1,29	2,12	1,64	i
Holme House	1,00	1,15	1,00	1,15	1,15	i***	1,09	1,70	1,56	i*	2,22	3,15	1,42	d***
Hull	0,98	1,08	1,00	1,25	1,25	d	0,98	1,08	1,11	i	1,70	2,71	1,60	d**
Leeds	1,00	1,29	1,00	1,30	1,30	i	1,17	1,29	1,11	d*	1,06	1,57	1,48	d
Leicester	0,97	1,28	1,07	1,48	1,38	i***	1,01	1,57	1,55	i***	0,97	1,28	1,32	i***
Lewes	1,00	1,20	1,00	1,20	1,20	i*	1,21	1,31	1,08	i	2,17	2,32	1,07	d
Lincoln	1,06	1,28	1,09	1,56	1,43	i***	1,06	1,28	1,21	i***	1,08	1,47	1,36	i
Liverpool	1,00	1,02	1,00	1,02	1,02	d*	1,01	1,03	1,02	d*	1,76	2,73	1,56	d***
Manchester	0,98	1,21	1,23	1,65	1,34	d***	0,98	1,21	1,24	d***	1,90	2,64	1,39	d
Norwich	1,14	1,32	1,14	1,38	1,21	i	1,23	1,32	1,07	i	1,74	2,06	1,18	d
Nottingham	1,51	1,64	1,63	1,99	1,22	i	1,51	1,64	1,09	i	3,46	4,36	1,26	d
Parc (G4S)	1,00	1,34	1,20	1,61	1,34	d***	1,00	1,34	1,34	d***	1,14	2,66	2,32	i***
Pentonville	1,21	1,49	1,29	1,49	1,16	d***	1,21	1,49	1,23	d***	1,63	2,19	1,35	d
Preston	1,00	1,23	1,37	1,63	1,19	d	1,06	1,23	1,16	i	1,00	1,32	1,32	d*
Swansea	0,85	1,34	0,85	1,92	2,25	i***	0,85	1,34	1,57	i***	1,21	2,09	1,74	i***
Wandsworth	0,98	1,20	1,07	1,44	1,35	d***	0,98	1,20	1,23	d***	2,19	3,15	1,44	d
Winchester	0,97	1,04	1,22	1,41	1,15	i	0,97	1,04	1,08	i	2,29	2,37	1,03	d
Woodhill	1,43	2,11	1,54	2,12	1,38	d***	1,55	2,11	1,36	d*	1,43	2,21	1,55	i
Wormwood Scrubs	1,01	1,13	1,01	1,13	1,12	d***	1,02	1,17	1,15	d***	2,57	4,38	1,70	d**
Mean	*1,04*	*1,28*	*1,12*	*1,43*	*1,29*		*1,11*	*1,37*	*1,23*		*1,67*	*2,36*	*1,44*	
(Standard Deviation)	*(0,13)*	*(0,29)*	*(0,16)*	*(0,31)*	*(0,25)*		*(0,20)*	*(0,38)*	*(0,22)*		*(0,68)*	*(0,90)*	*(0,31)*	

Note: + indicates over 90% of the subsample bootstrap replications show the value, $*$ indicates 95%, $**$ indicates 99% and $***$ indicates 99.9%.

Table 6: Order-m efficiency scores and returns to scale estimates for $m = 50$ and $B = 1000$ (2010/11)

Year 2010/11	Multi Output		Incarceration				Activities				Reintegration			
Prison	$\varphi(v)$	$\varphi(c)$	$\varphi^1(v)$	$\varphi^1(c)$	SE^1	RTS	$\varphi^2(v)$	$\varphi^2(c)$	SE^2	RTS	$\varphi^3(v)$	$\varphi^3(c)$	SE^3	RTS
Altcourse (G4S)	1,00	1,14	1,27	1,91	1,50	d***	1,00	1,14	1,13	d***	1,08	1,14	1,06	d**
Bedford	0,98	1,01	1,00	1,23	1,24	i***	1,00	1,07	1,07	d*	0,98	1,01	1,03	i
Belmarsh	1,25	2,14	1,25	2,14	1,71	d***	1,59	2,61	1,64	d***	2,61	5,12	1,96	d***
Birmingham	0,98	1,07	1,07	1,26	1,18	d***	0,98	1,07	1,09	d*	1,18	2,10	1,79	d**
Bristol	1,00	1,06	1,17	1,36	1,16	i	1,00	1,06	1,06	i	1,04	1,37	1,32	d***
Brixton	1,09	1,28	1,09	1,28	1,18	d*	1,41	1,77	1,26	d**	1,80	2,76	1,53	d*
Bullingdon	1,00	1,01	1,00	1,13	1,13	i***	1,00	1,01	1,01	c	2,07	2,80	1,35	d***
Cardiff	1,01	1,18	1,02	1,20	1,18	d	1,01	1,18	1,16	d**	1,31	1,76	1,35	d
Chelmsford	1,02	1,08	1,02	1,08	1,07	d	1,08	1,14	1,06	d	1,11	1,33	1,19	d
Doncaster (Serco)	1,00	1,00	1,00	1,09	1,09	i	1,00	1,00	1,00	c**	1,01	1,73	1,71	i***
Dorchester	1,00	1,76	1,09	1,76	1,62	i***	1,00	1,82	1,83	i***	1,20	2,20	1,84	i***
Durham	1,00	1,00	1,00	1,10	1,10	d***	1,02	1,09	1,07	d*	1,00	1,00	1,00	c***
Exeter	1,00	1,23	1,00	1,39	1,39	i***	1,01	1,23	1,21	i***	1,37	1,89	1,38	d*
Forest Bank (SJS)	1,00	1,00	1,01	1,10	1,09	d***	1,00	1,03	1,03	d***	1,00	1,00	1,00	c***
Gloucester	1,00	1,22	1,00	1,23	1,23	i***	1,00	1,22	1,22	i***	1,21	1,27	1,05	d*
High Down	1,00	1,07	1,00	1,07	1,07	d*	1,25	1,32	1,06	d*	1,77	2,57	1,45	d**
Holme House	1,00	1,06	1,00	1,06	1,06	d***	1,47	1,64	1,11	i	1,58	2,95	1,87	d***
Hull	0,99	1,12	1,00	1,17	1,17	d	0,99	1,12	1,13	i	1,23	2,08	1,69	d***
Leeds	1,00	1,20	1,00	1,30	1,30	i	1,11	1,20	1,08	d*	1,05	1,66	1,58	d
Leicester	0,98	1,45	1,07	1,48	1,38	i***	1,01	1,45	1,44	i***	0,98	1,47	1,50	i***
Lewes	0,94	1,09	0,94	1,13	1,21	i**	1,15	1,51	1,32	i***	0,99	1,09	1,10	i*
Lincoln	1,00	1,34	1,10	1,58	1,43	i***	1,04	1,41	1,35	i***	1,00	1,34	1,33	i
Liverpool	1,00	1,01	1,00	1,01	1,02	d	1,00	1,04	1,04	d*	1,25	1,92	1,54	d**
Manchester	1,01	1,17	1,23	1,69	1,37	d***	1,01	1,17	1,15	d***	1,28	2,63	2,06	d***
Norwich	0,98	1,01	1,00	1,23	1,23	i**	0,98	1,01	1,03	i	1,14	1,25	1,10	d
Nottingham	1,03	1,10	1,04	1,52	1,45	i	1,03	1,10	1,06	d*	1,64	2,40	1,46	d
Parc (G4S)	1,08	1,51	1,08	1,63	1,51	d***	1,37	1,51	1,10	d***	3,06	3,28	1,07	d***
Pentonville	1,08	1,32	1,27	1,43	1,13	d***	1,11	1,32	1,18	d***	1,08	1,43	1,32	d
Preston	1,00	1,12	1,31	1,52	1,16	d	1,00	1,12	1,12	i	1,00	1,46	1,46	d***
Swansea	0,90	1,26	0,94	1,97	2,09	i***	0,93	1,26	1,35	i***	0,90	1,26	1,39	i***
Wandsworth	1,07	1,28	1,07	1,41	1,32	d***	1,07	1,28	1,19	d***	1,77	3,53	1,99	d***
Winchester	0,98	1,04	1,00	1,16	1,17	i	0,98	1,04	1,06	i	1,24	1,63	1,31	d***
Woodhill	1,03	1,09	1,32	1,86	1,41	d***	1,26	1,82	1,44	d*	1,03	1,09	1,06	d***
Wormwood Scrubs	1,01	1,08	1,01	1,08	1,08	d***	1,12	1,23	1,10	d**	1,24	2,08	1,67	d***
Mean	*1,01*	*1,19*	*1,07*	*1,37*	*1,28*		*1,09*	*1,29*	*1,18*		*1,33*	*1,93*	*1,43*	
(Standard Deviation)	*(0,06)*	*(0,24)*	*(0,11)*	*(0,30)*	*(0,22)*		*(0,16)*	*(0,33)*	*(0,18)*		*(0,48)*	*(0,89)*	*(0,31)*	

Note: + indicates over 90% of the subsample bootstrap replications show the value, $*$ indicates 95%, $**$ indicates 99% and $***$ indicates 99.9%.

Table 7: Order-m efficiency scores and returns to scale estimates for $m = 50$ and $B = 1000$ (2011/12)

Year 2011/12	Multi Output		Incarceration				Activities				Reintegration			
Prison	$\varphi(v)$	$\varphi(c)$	$\varphi^1(v)$	$\varphi^1(c)$	SE^1	RTS	$\varphi^2(v)$	$\varphi^2(c)$	SE^2	RTS	$\varphi^3(v)$	$\varphi^3(c)$	SE^3	RTS
Altcourse (G4S)	1,00	1,15	1,25	1,77	1,41	d***	1,00	1,15	1,15	d***	1,00	1,24	1,24	i***
Bedford	0,98	1,00	0,99	1,15	1,16	i***	1,00	1,00	1,01	c*	0,98	1,06	1,08	i***
Belmarsh	1,25	2,02	1,25	2,02	1,61	d***	1,59	2,43	1,53	d***	2,53	4,00	1,58	d***
Birmingham (G4S)	1,00	1,00	1,00	1,37	1,37	i***	1,00	1,29	1,29	i***	1,00	1,00	1,00	c***
Bristol	0,95	1,08	0,95	1,29	1,35	i***	0,96	1,08	1,12	i	1,00	1,23	1,23	d*
Brixton	1,06	1,21	1,06	1,21	1,15	d	1,39	1,59	1,15	d**	2,03	2,47	1,22	d*
Bullingdon	1,00	1,08	1,00	1,12	1,12	i***	1,01	1,08	1,06	c	1,82	2,44	1,34	d***
Cardiff	1,01	1,19	1,01	1,19	1,18	d	1,15	1,33	1,17	d**	1,17	1,55	1,33	d
Chelmsford	0,99	1,00	0,99	1,02	1,03	c	0,99	1,00	1,01	c*	1,25	1,32	1,06	d*
Doncaster (Serco)	1,00	1,00	1,00	1,04	1,04	i	1,00	1,00	1,00	c**	1,02	1,93	1,89	i***
Dorchester	0,84	1,57	0,84	1,57	1,87	i***	0,91	1,90	2,09	i***	0,98	1,77	1,81	i***
Durham	0,98	1,00	1,00	1,00	1,00	c***	0,98	1,02	1,04	i	1,00	1,00	1,00	c***
Exeter	0,96	1,16	0,96	1,39	1,44	i***	0,98	1,16	1,18	i***	1,00	1,34	1,34	d*
Forest Bank (SJS)	1,00	1,00	1,00	1,04	1,04	d	1,00	1,00	1,00	c***	1,00	1,34	1,34	i***
Gloucester	0,99	1,00	0,99	1,19	1,20	i***	1,00	1,21	1,21	i***	0,99	1,00	1,00	c*
High Down	1,00	1,00	1,00	1,00	1,00	c*	1,15	1,29	1,12	d	1,58	2,15	1,36	d*
Holme House	1,00	1,00	1,00	1,00	1,00	c***	1,00	1,24	1,24	i	1,53	2,35	1,53	d***
Hull	0,94	1,09	0,98	1,12	1,14	d	0,94	1,09	1,15	i	1,28	1,89	1,48	d***
Leeds	0,94	0,98	0,98	1,28	1,31	i	0,94	0,98	1,04	c	1,12	1,65	1,46	d
Leicester	0,96	1,28	1,07	1,38	1,29	i***	0,99	1,28	1,29	i***	0,96	2,01	2,10	i***
Lewes	0,98	1,01	0,98	1,14	1,16	i**	0,99	1,26	1,27	i***	1,00	1,01	1,01	c
Lincoln	0,96	1,16	1,06	1,52	1,43	i***	0,96	1,16	1,21	i***	1,01	1,30	1,28	i
Liverpool	0,95	1,00	0,95	1,00	1,05	c	0,98	1,02	1,05	d	1,16	1,72	1,49	d*
Manchester	1,00	1,10	1,23	1,61	1,31	d***	1,00	1,10	1,10	d***	1,95	3,84	1,97	d***
Norwich	0,93	1,00	1,00	1,19	1,19	i***	0,93	1,00	1,08	c	1,01	1,06	1,05	d
Nottingham	0,93	0,98	1,01	1,38	1,37	i	0,93	0,98	1,05	c	0,98	1,25	1,28	i
Parc (G4S)	1,00	1,45	1,00	1,47	1,47	d***	1,16	1,45	1,25	d***	4,51	7,62	1,69	d***
Pentonville	0,97	1,13	1,29	1,34	1,04	d*	1,10	1,19	1,07	d***	0,97	1,13	1,17	d
Preston	1,00	1,10	1,25	1,43	1,15	d	1,00	1,10	1,11	i	1,12	1,66	1,48	d***
Swansea	0,97	1,19	0,98	2,06	2,10	i***	0,97	1,25	1,29	i***	1,05	1,19	1,13	i*
Wandsworth	1,00	1,16	1,08	1,36	1,26	d***	1,00	1,16	1,16	d***	1,44	2,77	1,93	d***
Winchester	0,93	1,12	0,93	1,12	1,20	i**	0,98	1,14	1,16	i***	1,09	1,31	1,20	d
Woodhill	1,00	1,39	1,51	2,02	1,34	d***	1,25	1,70	1,37	d***	1,00	1,39	1,39	d***
Wormwood Scrubs	1,00	1,05	1,01	1,05	1,04	d*	1,19	1,24	1,05	d***	1,00	1,59	1,59	d***
Mean	*0,98*	*1,14*	*1,05*	*1,32*	*1,26*		*1,04*	*1,23*	*1,18*		*1,31*	*1,87*	*1,38*	
(Standard Deviation)	*(0,06)*	*(0,21)*	*(0,13)*	*(0,30)*	*(0,24)*		*(0,14)*	*(0,30)*	*(0,20)*		*(0,68)*	*(1,25)*	*(0,30)*	

Note: + indicates over 90% of the subsample bootstrap replications show the value, $*$ indicates 95%, $**$ indicates 99% and $***$ indicates 99.9%.

C. Environmental variables

Daraio and Simar (2005) introduce a probabilistic approach to condition on environmental factors. To apply the probabilistic approach, a kernel function and appropriate bandwidth need to be estimated. Following Li and Racine (2007) we use a generalized product kernel to allow for both continuous and discrete environmental variables. For the continuous data, we use Epanechnikov kernel weighting and for the discrete data Liracine kernel weighting. To select the bandwidth sizes, we follow the procedure of Badin et al. (2010), making use of least squares cross-validation. The procedure of Badin et al. (2010) also considers the influence of the environmental variables on the production process. Since our framework employs output-specific production processes, we obtain output-specific bandwidth sizes. Table 8 reports the estimated bandwidths for each environmental variable.

Table 8: Bandwidth sizes

	Mean	St.Dev.	0%	25%	50%	75%	100%
Incarceration							
Predreof	3,07	0,61	1,39	2,68	2,71	4,01	4,01
Agecat	0,02	0,04	0,00	0,00	0,00	0,01	0,15
Management	0,07	0,13	0,00	0,00	0,00	0,10	0,50
Activities							
Predreof	$8{,}23e^5$	$4{,}94e^6$	2,67	2,70	2,79	2,93	$4{,}15e^7$
Agecat	0,02	0,04	0,00	0,00	0,01	0,02	0,22
Management	0,06	0,13	0,00	0,00	0,01	0,03	0,50
Successful reintegration							
Predreof	1,56	0,38	1,15	1,39	1,41	1,64	2,84
Agecat	0,77	0,35	0,08	0,57	0,97	1,00	1,00
Management	0,06	0,13	0,00	0,00	0,00	0,01	0,50
Regunemp	$5{,}01e^6$	$9{,}59e^6$	1,25	1,34	1,48	$7{,}17e^6$	$6{,}25e^7$

An advantage of the kernel weighting procedure is that there is no need to specify a priori a direction of influence of the environmental variables. In fact, we can obtain a posteriori indications on the direction of influence.

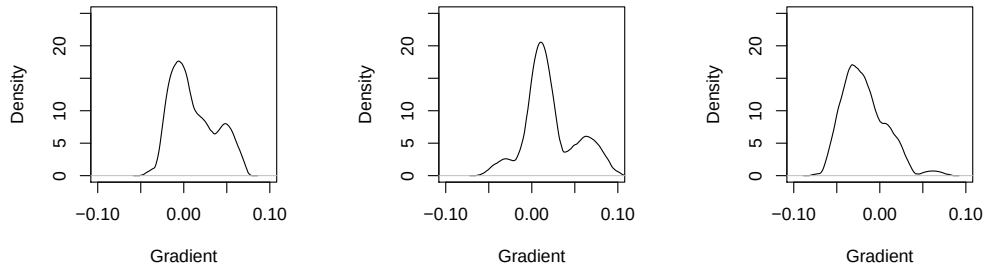
Daraio and Simar (2005, 2007a) explain in detail how to examine the direction of the environmental effect by nonparametrically regressing the environmental variables on the ratio of the conditional to unconditional technical efficiency. We follow this approach and use a generalized kernel local linear

regression (Li and Racine, 2004), again with Epanechnikov kernels for the continuous data and Liracine kernels for the discrete data.[11] A positive (negative) gradient reveals a favorable (unfavorable) effect of the environmental variable. For discrete data, a category with higher (lower) fitted values is indicated as a more (less) favorable environment. Given the data limitations, we interpret the point estimates as indications for a direction of influence rather than as evidence of a causal effect. Figure 6 shows the output-specific density plots of the gradients of the nonparametric regression, for the effect of predicted rate of re-offending. We find that the effect of the predicted rate of re-offending, our proxy for prisoner characteristics, is indicated to be non-monotone. Only for the output successful reintegration, the effect is clearly unfavorable.

Similarly, Figure 7 shows the density plot of the gradients of unemployment rate for the output successful reintegration. Higher regional unemployment rate is for nearly all prisons indicated as a less favorable environment to successfully reintegrate prisoners. This finding corresponds with the intuition that a higher regional unemployment rate leads to more difficulties finding a job at discharge. Further, Figure 8 shows the output-specific box plots of the estimated influence of prison age. The data shows that newer prisons do not necessarily have an operational advantage.

Finally, we consider the environmental influence of prison management (public=0, privately managed=1) in Figure 9. We find that privately managed prisons have an operational disadvantage for incarceration and an operational advantage for the provision of activities. For successful reintegration, we find a less pronounced operational advantage for public prisons. As such, we confirm the results of Rogge et al. (2015) that private prisons do not outperform public prisons when considering all outputs. However, we do not make any causality statements concerning prison management and a detailed study of the causal effect of privatization on the performance of prisons goes beyond the scope of this paper.

[11] We use the '*np*' package of Hayfield and Racine (2008) to implement the bandwidth selection procedure of Badin et al. (2010) and to perform the nonparametric regression, with least squares cross validation.



(a) Incarceration (b) Activities (c) Successful reintegration

Figure 6: The impact of predicted rate of re-offending on the ratio of the conditional to unconditional scores, for each output.

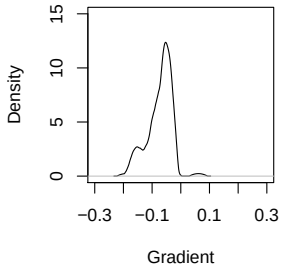
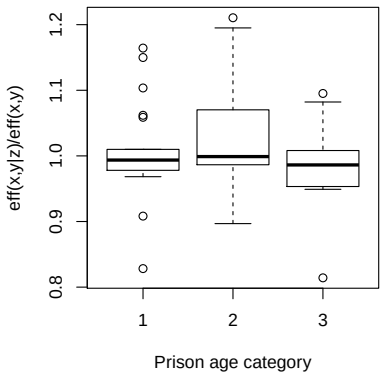
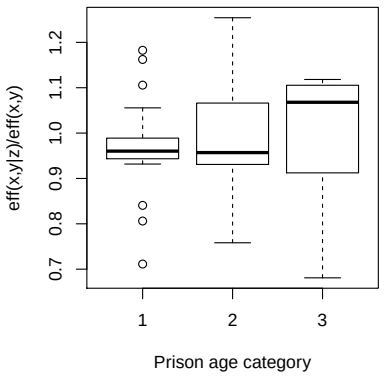


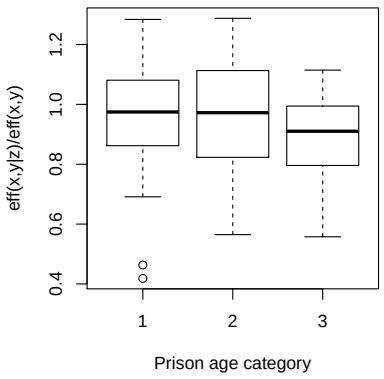
Figure 7: The impact of regional male unemployment rate on the ratio of the conditional to unconditional scores, for successful reintegration



(a) Incarceration

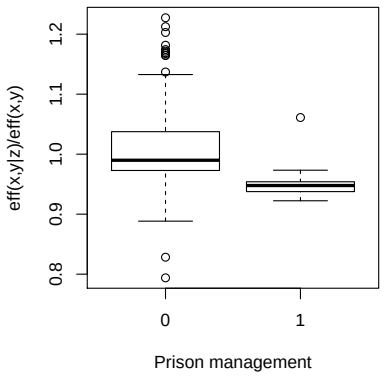


(b) Activities

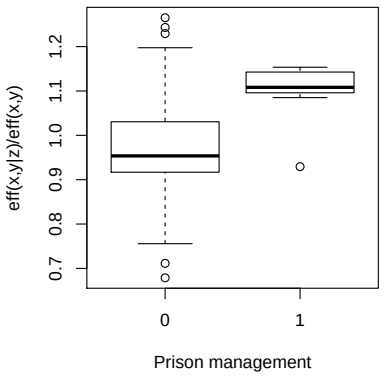


(c) Successful reintegration

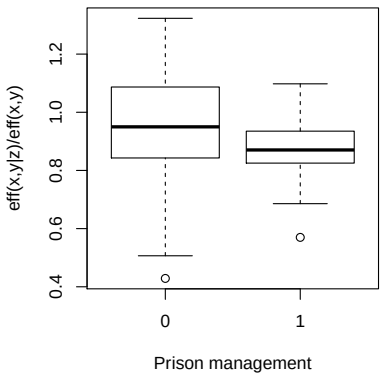
Figure 8: The impact of prison age on the ratio of the conditional to unconditional scores.



(a) Incarceration



(b) Activities



(c) Successful reintegration

Figure 9: The impact of prison management (0=public, 1=privately managed) on the ratio of the conditional to unconditional scores.

References

Atici, K. B., Podinovski, V. V., 2012. Mixed partial elasticities in constant returns-to-scale production technologies. European Journal of Operational Research 220 (1), 262–269.

Badin, L., Daraio, C., Simar, L., 2010. Optimal bandwidth selection for conditional efficiency measures: A data-driven approach. European Journal of Operational Research 201, 633–640.

Banker, R., Morey, R., 1986. Efficiency analysis for exogenously fixed inputs and outputs. Operations Research 34 (4), 513–521.

Banker, R. D., 1984. Estimating most productive scale size using data envelopment analysis. European Journal of Operational Research 17 (1), 35–44.

Banker, R. D., Charnes, A., Cooper, W. W., 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science 30 (9), pp. 1078–1092.

Becker, G. S., 1968. Crime and punishment - economic approach. Journal of Political Economy 76 (2), 169–217.

Bogetoft, P., 1996. DEA on Relaxed Convexity. Management Science 42 (3), 457–465.

Buonanno, P., Raphael, S., 2013. Incarceration and incapacitation: Evidence from the 2006 italian collective pardon. American Economic Review 103 (6), 2437–2465.

Butler, T., Johnson, W., 1997. Efficiency evaluation of michigan prisons using data envelopment analysis. , 22, 1-15. Criminal Justice Review 22 (1), 1–15.

Campbell, M. C., Vogel, M., Williams, J., 2015. Historical contingencies and the evolving importance of race, violent crime, and region in explaining mass incarceration in the united states. Criminology 53 (2), 180–203.

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

Cherchye, L., De Rock, B., Dierynck, B., Roodhooft, F., Sabbe, J., 2013. Opening the "black box" of efficiency measurement: Input allocation in multioutput settings. Operations Research 61 (5), 1148–1165.

Cherchye, L., De Rock, B., Walheer, B., 2015. Multi-output efficiency with good and bad outputs. European Journal of Operational Research 240, 872–888.

Cook, W. D., Zhu, J., 2011. Multiple Variable Proportionality in Data Envelopment Analysis. Operations Research 59 (4), 1024–1032.

Cooper, R., Kaplan, R. S., 1988. Measure costs right: Make the right decision. Harvard Bussiness Review 66 (5), 96–103.

Daraio, C., Simar, L., 2005. Introducing environmental variables in nonparametric frontier models: A probabilistic approach. Journal of Productivity Analysis 24 (1), 93–121.

Daraio, C., Simar, L., 2007a. Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications. Studies in Productivity and Efficiency. Springer.

Daraio, C., Simar, L., 2007b. Conditional nonparametric frontier models for convex and nonconvex technologies : a unifying approach. Journal of Productivity Analysis 28, 13–32.

Debreu, G., 1951. The coefficient of resource utilization. Econometrica 19, 273–292.

Deprins, L., Simar, L., Tulkens, H., 1984. Measuring labor efficiency in post offices. In: Marchand, M., Pestieau, P., Tulkens, H. (Eds.), The Performance of Public Enterprises: Concepts and Measurement. North Holland, Amsterdam, pp. 243–267.

Di Tella, R., Schargrodsky, E., 2013. Criminal recidivism after prison and electronic monitoring. Journal of Political Economy 121 (1), 28–73.

Drake, L., Simper, R., 2002. X-efficiency and scale economies in policing: a comparative study using the distribution free approach and DEA. Applied Economics 34 (15), 1859–1870.

Duncombe, W., Yinger, J., 1993. An analysis of returns to scale in public production, with an application to fire protection. Journal of Public Economics 52 (1), 49–72.

Farrell, L. M. J., 1957. The measurement of productive efficiency. Journal of the Royal Statistical Society Series A-General 120 (3), 253–290.

Fethi, M., Pasiouras, F., 2010. Assessing bank efficiency and performance with operational research and artificial intelligence techniques: A survey. European Journal of Operational Research 204, 189–198.

Hall, M., Liu, W., Simper, R., Zhou, Z., 2013. The economic efficiency of rehabilitative management in young offender institutions in england and wales. Socio-Economic Planning Sciences 47, 2013, 38-49.

Hansen, B., 2015. Punishment and deterrence: Evidence from drunk driving. American Economic Review 105 (4), 1581–1617.

Hayfield, T., Racine, J. S., 2008. Nonparametric econometrics: The np package. Journal of Statistical Software 27 (5).
URL `http://www.jstatsoft.org/v27/i05/`

HM Inspectorate of Prisons, 2009. The prison characteristics that predict prisons being assessed as performing 'well': A thematic review. Tech. rep., HM Chief Inspector of Prisons, London: HM Inspectorate of Prisons.

Katz, L., Levitt, S., Shustorovich, E., 2003. Prison conditions, capital punishment, and deterrence. American Law and Economics Review 5 (2), 318–343.

Kerstens, K., Vanden Eeckaut, P., 1999. Estimating returns to scale using non-parametric deterministic technologies: A new method based on goodness-of-fit. European Journal of Operational Research 113 (1), 206–214.

Levitt, S. D., 1996. The effect of prison population size on crime rates: Evidence from prison overcrowding litigation. Quarterly Journal of Economics 111 (2), 319–351.

Li, Q., Racine, J., 2004. Cross-validated local linear nonparametric regression. Statistica Sinica 14 (2), 485–512.

Li, Q., Racine, J., 2007. Nonparametric Econometrics: Theory and practice. Princeton University Press.

Liebling, A., 2004. Prisons and Their Moral Performance: A Study of Values, Quality, and Prison Life. Oxford: Oxford University Press.

Lockyer, K., 2013. A radical plan to reform the prison estate. Policy exchange, London. UK.

Ministry of Justice, 2011. Breaking the cycle: Government response. London: The Stationery Office.

Ministry of Justice, 2013. Transforming Rehabilitation: A strategy for reform. London: The Stationery Office.

Ministry of Justice, 2015. National offender management service: Prison annual performance ratings 2014/15. Tech. rep., NOMS.

National Audit Office, 2013. Managing the prison estate. Tech. rep., London: The Stationary Office.

Nyhan, R., 2002. Benchmarking tools: an application to juvenile justice facility performance. The Prison Journal 82, 423-439.

Petersen, N. C., 1990. Data Envelopment Analysis on a Relaxed Set of Assumptions. Management Science 36 (3), 305–314.

Peyrache, A., Zago, A., 2015. Large courts, small justice!: The inefficiency and the optimal structure of the italian justice sector. Omega 64, 42–56.

Podinovski, V. V., 2004a. Efficiency and Global Scale Characteristics on the " No Free Lunch " Assumption Only. Journal of Productivity Analysis 22, 227–257.

Podinovski, V. V., 2004b. Local and Global Returns to Scale in Performance Measurement. Journal of the Operational Research Society 55 (2), 170–178.

Podinovski, V. V., Forsund, F. R., 2010. Differential characteristics of efficient frontiers in data envelopment analysis. Operations Research 58 (6), 1743–1754.

Podinovski, V. V., Kuosmanen, T., 2011. Modelling weak disposability in data envelopment analysis under relaxed convexity assumptions. European Journal of Operational Research 211 (3), 577–585.

Rogge, N., Simper, R., Verschelde, M., Hall, M., 2015. An analysis of managerialism and performance in English and Welsh male prisons. European Journal of Operational Research 241 (1), 224–235.

Ruggiero, J., 2000. Nonparametric estimation of returns to scale in the public sector with an application to the provision of educational services. Journal of the Operational Research Society 51 (8), 906–912.

Simar, L., Wilson, P., 2007. Estimation and inference in two-stage, semiparametric models of production processes. Journal of Econometrics 136 (1), 31–64.

Syrjänen, M. J., 2004. Non-discretionary and discretionary factors and scale in data envelopment analysis. European Journal of Operational Research 158 (1), 20–33.

Tulkens, H., 1993. On fdh efficiency analysis: Some methodological issues and applications to retail banking, courts and urban transit. Journal of Productivity Analysis 4, 183–190.

Verschelde, M., Rogge, N., 2012. An environment-adjusted evaluation of citizen satisfaction with local police effectiveness: Evidence from a conditional data envelopment analysis approach. European Journal of Operational Research 223, 214–225.

Vollaard, B., 2013. Preventing crime through selective incapacitation. Economic Journal 123 (567), 262–284.