

Model-based adjustment for conditional benchmarking

DANIEL J. GRAHAM[†] AND RAMANDEEP SINGH

*Transport Strategy Centre, Department of Civil & Environmental Engineering, Imperial College
London, London, SW7 2AZ, UK*

[†]Corresponding author: d.j.graham@imperial.ac.uk

[Received on 14 August 2020; accepted on 28 May 2021]

Quantitative benchmarking is widely used in the industry to compare relative performance across a sample of organizations. A key analytical challenge lies in obtaining accurate measures of intrinsic organizational performance net of contextual or exogenous influences. In this paper, we propose a model-based adjustment approach for comparative benchmarking that allows the analyst to recover targeted metrics for specific aspects of innate performance. We outline the statistical theory underpinning our method, provide simulations to demonstrate its properties and describe practical examples for computation. The managerial relevance of the method is demonstrated via two real-world transport industry applications: adjusting for economies of scale and density in benchmarking average costs of urban metros and for service characteristics in benchmarking metro journey times.

Keywords: benchmarking; model; performance; conditional; unconditional.

1. Introduction

Quantitative benchmarking is widely used in industry to compare relative performance across a sample of organizations. Performance domains of interest range from strategic management functions to highly disaggregate activities (Francis & Holloway, 2007; Kaplan & Norton, 1992; Moriarty & Smallman, 2009). A typical quantitative benchmarking analysis involves computation of a set of metrics, sometimes referred to as key performance indicators (KPIs), using data observed on a sample of organizations. Deviation from the central tendency of a KPI is taken as an indicator of relative performance, and improvement targets are set for units that fall below defined threshold levels (Camp, 1989; Leibfried & McNair, 1992).

KPIs are useful in describing relative performance, for instance in the form of a performance ranking, but are typically of less value in explaining what lies behind the ranking itself. In particular, problems of interpretation can arise when benchmarking metrics do not readily partition the innate performance of organizations from other exogenous determinants. Under these conditions, KPIs can produce a somewhat obscured understanding of performance, or worse, can actually be highly misleading (Bogetoft, 2012). It is this well-known limitation of quantitative benchmarking that we aim to address in this paper.

To motivate our study, we begin by making a distinction between *intrinsic* and *composite* performance. By intrinsic performance, we mean the innate performance of organizations that the benchmarking analyst ultimately wants to measure and rank. Composite performance, on the other hand, is a combination of both intrinsic performance and of other exogenous determinants of performance that are not themselves of interest for benchmarking purposes. Such exogenous determinants could, for instance, include characteristics of the context or local environment in which an organization operates,

but which are largely outside the control of the organization itself. In practice, it is often the case that available data allow us to benchmark composite performance but leave intrinsic performance unobserved. Our aim here is to describe an approach that can be used to recover a measure of intrinsic performance from observed data on composite performance.

To do so, we develop a simple model-based adjustment approach for comparative benchmarking. We achieve this by first removing ‘contamination’ from exogenous determinants via regression analysis and then by recovering adjusted performance metrics with reduced variance that are measured on the same scale as the original metric. Our adjusted benchmarking approach allows comparisons to be made on specially tailored definitions of performance and can generate useful insights through comparison of conditional and unconditional benchmarks.

The paper is structured as follows. Section 2 provides a brief review of relevant literature on efficiency measurement. Section 3 outlines our model-based approach for benchmarking. It first develops the basic statistical theory and then provides some practical examples for computation. Section 4 presents simulation results to demonstrate the properties of our proposed method. In Section 5 we demonstrate the managerial relevance of our approach through two real-world applications for the transport industry: one that adjusts for economies of scale and density in benchmarking average costs of urban metros and one that adjusts for service characteristics in benchmarking metro journey times. Conclusions are drawn in the final section.

2. Literature review

A wide range of quantitative methods may be adopted to evaluate and compare performance. The approaches most relevant to the work we develop in this paper are those concerned with efficiency frontier estimation, including stochastic frontier analysis (SFA) and data envelopment analysis (DEA) (Agrenell & Bogetoft, 2017; Bogetoft, 2012).

Frontier techniques compare the efficiency of firms (or production entities) relative to a technological or production frontier. The frontier represents the maximum output attainable for a given level of inputs, and the position of the firms or units relative to the frontier indicates their relative efficiency. The SFA technique, first proposed by Aigner *et al.* (1977) and Meeusen & Broeck (1977), uses a parametric model to define the production frontier and includes terms that distinguish between efficiency and stochastic noise. The DEA method, introduced by Charnes *et al.* (1978), is non-parametric and consequently invokes less a priori assumptions to establish the production frontier, but typically does not explicitly model stochastic terms. Example applications of SFA and DEA include evaluating the efficiency of private firms or public agencies in finance (Tsolas, 2011), healthcare (Alexander *et al.*, 2003), energy (Kuo, 2011; Tsolas, 2010) and transport (Canavan *et al.*, 2015; Graham, 2008).

Exogenous determinants of productivity can be modelled using SFA and DEA approaches. For example, in SFA frameworks, exogenous determinants can be included as random and fixed effects terms or as latent variables as per Tsionas (2002), Greene (2005) and Schmidt *et al.* (2009). In DEA models, exogenous factors can be incorporated using one and two stage approaches (see Daraio & Simar, 2005, for further references) or through probabilistic conditional models as per Daraio & Simar (2005), Daraio & Simar (2007b) and Daraio & Simar (2007a).

However, while it is possible to distinguish between intrinsic and exogenous determinants of performance in frontier models, it is difficult to apply frontier techniques universally because in many cases performance cannot easily be conceptualized in an input–output-type formulation. In this paper, we propose a general model-based adjustment framework for conditional benchmarking, which to our knowledge is the first time such a formal framework has been set out.

3. Model-based adjustment for benchmarking

Throughout the paper we use the standard convention that upper case denotes random variables and lower case the realized values of random variables. The setting we address is one in which the full data for units under study would comprise realization of the random vector, $Z_i = (Y_i, P_i, X_i)$, $i = (1, \dots, n)$, where outcome variable Y_i is a composite performance metric for unit i , P_i is a vector of covariates that determine the intrinsic performance of unit i and X_i is a vector of covariates that capture measured exogenous influences on composite performance.

For the purposes of benchmarking we would like to reveal as much as possible about variation in intrinsic performance, which we denote $Y_i(P_i)$, but this quantity is not observed, and in fact even data on the vector P_i we also assume to be unavailable. Rather, we observe only Y_i and X_i and we are therefore left with no option but to benchmark performance using Y_i . To obtain a better understanding of intrinsic performance, our objective is to adjust Y_i to reduce the influence of exogenous factors X_i , and thus get closer to a measure of unobserved $Y_i(P_i)$. In addition, we want both our adjusted and unadjusted benchmarking metrics to be on the same scale to allow for direct comparison.

3.1 Statistical theory

We define observed composite performance as a function of the intrinsic performance determinants, exogenous influences and random error (e_i), $y_i = f(p_i, x_i, e_i; \theta)$, where θ is an unknown parameter vector. If $f(p_i, x_i, e_i; \theta)$ is additively separable in the covariates and the error, we can write it in the form of a conditional expectation function as

$$y_i = \mathbb{E}[Y_i|P_i, X_i] + e_i,$$

where for the moment we will assume that $e_i \sim N(0, \sigma_e^2)$. The mean of Y is denoted μ_Y and the variance σ_Y^2 .

The data generating process (DGP) is correctly captured by the following model for the expectation of response given intrinsic and exogenous performance,

$$\mathbb{E}[Y_i|P_i, X_i] = m_A(P_i, X_i; \theta_A), \quad (1)$$

where $m_A()$ is a linear or additive regression function and $\theta_A = (\theta_{A1}, \dots, \theta_{AQ})$ is a Q dimensional parameter vector.

Our objective is to empirically adjust random variable Y so that it captures variance in intrinsic performance but not in exogenous influences, and also retains the scale of the original variable. In other words, we want to reduce the variance of Y by removing that associated with X . To do this, we simply apply an iterated expectation to form the adjusted quantity

$$y_i(p_i) = \mathbb{E}_X [\mathbb{E}[Y_i|P_i = p_i, X_i]] + e_i, \quad (2)$$

with sample counterpart

$$\hat{y}_i(p_i) = \frac{1}{n} \sum_{i=1}^n m_A(p_i, X_i; \hat{\theta}_A) + e_i.$$

The adjusted metric is therefore derived by averaging the predicted value of the regression model for $\mathbb{E}[Y_i|P_i, X_i]$ over covariates X . If we write the mean of $y_i(p_i)$ as μ_Y and the variance as $\sigma_Y^2(p)$, then note that while the means of Y_i and $Y_i(p_i)$ are the same, $\sigma_{Y(p)}^2 \leq \sigma_Y^2$ because $\text{Var}\{\mathbb{E}_X[\mathbb{E}[Y_i|P_i = p_i, X_i]]\} \leq \text{Var}\{\mathbb{E}[Y_i|P_i, X_i]\}$.

If full data on Y_i , P_i and X_i were available, then consistent estimates of θ_A could be obtained as solutions to estimating equations of the form

$$\sum_{i=1}^n \frac{\partial m_A(p_i, x_i; \theta_A)}{\partial \theta_A^T} \frac{1}{\phi} [y_i - m_A(p_i, x_i; \theta_A)] = 0, \quad (3)$$

where ϕ is a scale parameter in $\text{Var}[Y_i|P_i, X_i]$ and $\hat{\theta}_A$ then used to calculate an estimate of (2). Many parametric or nonparametric optimization routines could yield estimating equations consistent with the form shown in equation (3). In the parametric setting, the score (or quasi-score) equations of maximum likelihood estimation (MLE), maximum quasi-likelihood and restricted MLE for linear mixed models all provide appropriate optimization solutions (e.g. McCulloch *et al.*, 2008).

However, the scenario we are interested in in this paper is one in which P_i is unobserved, or partially observed, so direct estimation of (1), and therefore of θ_A and $y_i(p_i)$, is not possible. We do have a measure of composite performance in Y_i , but it is ‘contaminated’ by exogenous influences and therefore unsuitable as a measure of intrinsic performance. Our proposed solution is to use model-based adjustment to recover a more representative intrinsic performance metric. We will develop the theory under the assumption that exogenous influences X_i are independent of P_i , implying that X_i is a non-confounding vector. Later in the section we will discuss the implications that arise when this assumption fails.

We write the correctly specified model for the observed data $z'_i = (y_i, x_i)$ as

$$\mathbb{E}[Y_i|X_i] = m_B(X_i; \theta_B), \quad (4)$$

where $m_B()$ is a linear or additive regression function and $\theta_B = (\theta_{B1}, \dots, \theta_{BK})$ is a K dimensional parameter vector. In addition, we define $\theta_X = \theta_A \cap \theta_B$ as the subset of both θ_A and θ_B that determines the relationship between Y_i and X_i . Under the key identifying condition of independence between intrinsic performance and exogenous influences, i.e. $P_i \perp\!\!\!\perp X_i$, estimation of (4) will yield consistent estimates of parameter vector θ_B , and hence of θ_X , even in the absence of representation of P_i within the model.

Since our regression models are additively separable we can isolate the specific component related to X_i , and note that under $P_i \perp\!\!\!\perp X_i$, $m_A(X_i, \theta_X) = m_B(X_i, \theta_X)$. From this result we can conduct model-based adjustment in the following way:

$$\hat{y}_i(p_i) = \frac{1}{n} \sum_{i=1}^n m_A(p_i, X_i; \hat{\theta}_A) + e_i \quad (5a)$$

$$= \frac{1}{n} \sum_{i=1}^n m_A(p_i; \hat{\theta}_X^c) + \frac{1}{n} \sum_{i=1}^n m_A(X_i; \hat{\theta}_X) + e_i \quad (5b)$$

$$= m_A(p_i; \hat{\theta}_X^c) + \frac{1}{n} \sum_{i=1}^n m_A(X_i; \hat{\theta}_X) + e_i \quad (5c)$$

$$= m_A(p_i, X_i; \hat{\theta}_A) - m_A(X_i; \hat{\theta}_X) + \frac{1}{n} \sum_{i=1}^n m_A(X_i; \hat{\theta}_X) + e_i \quad (5d)$$

$$= m_A(p_i, X_i; \hat{\theta}_A) - m_B(X_i; \hat{\theta}_X) + \frac{1}{n} \sum_{i=1}^n m_B(X_i; \hat{\theta}_X) + e_i \quad (5e)$$

$$= y_i - m_B(X_i; \hat{\theta}_X) + \frac{1}{n} \sum_{i=1}^n m_B(X_i; \hat{\theta}_X) \quad (5f)$$

$$\xrightarrow{P} \mathbb{E}_X [\mathbb{E}[Y_i | P_i = p_i, X_i]] + e_i, \quad (5g)$$

where (5a) assumes a correctly specified model for consistent estimation of θ_A , (5b) requires $P_i \perp\!\!\!\perp X_i$ and additive separability of the regression model (with $\theta_X^c = \{\theta_{Aq} \in \theta_A | \theta_{Aq} \notin \theta_X\}$), (5e) follows from consistent estimation of θ_X via estimation of model (4) and (5g) follows by applying a law of large numbers.

In this way we can remove the effect of exogenous influences from our composite performance measure and return this adjusted metric to the original Y_i scale. The resulting quantity, $\hat{y}_i(p_i)$ has contamination from random noise, but not from exogenous influences and is thus a viable metric for intrinsic performance.

3.2 Practical calculations

To provide concrete examples of these calculations we derive them in the context of estimates obtained via linear and semiparametric regression.

If our composite metric can be adequately represented by the first-order linear approximation

$$y_i = \theta_0 + P_i^\top \theta_P + X_i^\top \theta_X + e_i$$

then given a consistent estimate of θ_X obtained from a model on observed data, either estimated directly or reported in the literature, our model-based calculation simply transforms y_i to $\hat{y}_i(p_i)$ using

$$\begin{aligned} \hat{y}_i(p_i) &= \hat{\theta}_0 + p_i^\top \hat{\theta}_P + \bar{x}^\top \hat{\theta}_X + e_i \\ &= y_i + (\bar{x}_m - x_{mi})^\top \hat{\theta}_X. \end{aligned}$$

If the full unobserved semiparametric model with K intrinsic performance covariates and M exogenous covariates is

$$\mathbb{E}[Y_i | P_i, X_i] = \theta_0 + \sum_{k=1}^K g_k(P_{ki}; \theta_{kP}) + \sum_{m=1}^M h_m(X_{mi}; \theta_{mX}),$$

where $g_k()$ and $h_m()$ are smooth functions to be estimated; while the model for observed data is

$$\mathbb{E}[Y_i | X_i] = \theta_1 + \sum_{m=1}^M h_m(X_{mi}; \theta_{mX}).$$

Then the quantity

$$y_i(p_i) = \theta_0 + \sum_{k=1}^K g_k(p_{ki}; \theta_{kP}) + \sum_{m=1}^M \mathbb{E}[h_m(x_{mi}; \theta_{mX})] + e_i$$

can be estimated using

$$\begin{aligned} \hat{y}_i(p_i) &= \hat{\theta}_0 + \sum_{k=1}^K g_k(p_{ki}; \hat{\theta}_{kP}) + \sum_{m=1}^M h_m(x_{mi}; \hat{\theta}_{mX}) - \left[\hat{\theta}_1 + \sum_{m=1}^M h_m(x_{mi}; \hat{\theta}_{mX}) \right] \\ &\quad + \frac{1}{n} \sum_{i=1}^n \left[\hat{\theta}_1 + \sum_{m=1}^M h_m(x_{mi}; \hat{\theta}_{mX}) \right] + e_i \\ &= y_i - \left[\hat{\theta}_1 + \sum_{m=1}^M h_m(x_{mi}; \hat{\theta}_{mX}) \right] + \frac{1}{n} \sum_{i=1}^n \left[\hat{\theta}_1 + \sum_{m=1}^M h_m(x_{mi}; \hat{\theta}_{mX}) \right]. \end{aligned}$$

3.3 Dependence between P_i and X_i

When it is infeasible to assume that $P_i \perp\!\!\!\perp X_i$, and we do not observe P_i , then we face the problem of confounding and it is no longer possible to obtain consistent estimates of β from the marginal model (4). Such estimates will always be biased as long as $P_i \not\perp\!\!\!\perp X_i$.

Solutions to this problem can be found via application of methods for causal inference in the presence of unobserved confounders. A review of this material is outwith the scope of this article, but comprehensive descriptions of relevant estimators, including instrumental variables (IV), longitudinal random, fixed and dynamic methods, regression discontinuity designs (RDD) and difference-in-differences (DID), can be found in Diggle *et al.* (2002), Tsiatis (2006), Imbens & Wooldridge (2009), Imbens & Rubin (2015) Pearl *et al.* (2016) and Peters *et al.* (2017).

Whatever causal method is applied, the ultimate objective is to obtain consistent estimates of parameter vector β and follow the same procedure outlined above. Note that in situations for which causal inference is not possible, or is unlikely to produce appropriate estimates, recourse to the literature may provide statistical evidence that can be used to perform model-based adjustment and remove the effect of exogenous influences.

4. Simulations

In this section we present simulations to demonstrate the properties of the approach to model-based benchmarking described above. The simulation setup is as follows:

$$P \sim N(\mu_P = 10, \sigma_P^2 = 5)$$

$$X \sim N(\mu_X = 9, \sigma_P^2 = 2)$$

$$Y = 2 + P + 2X + e, \quad e \sim N(0, 1)$$

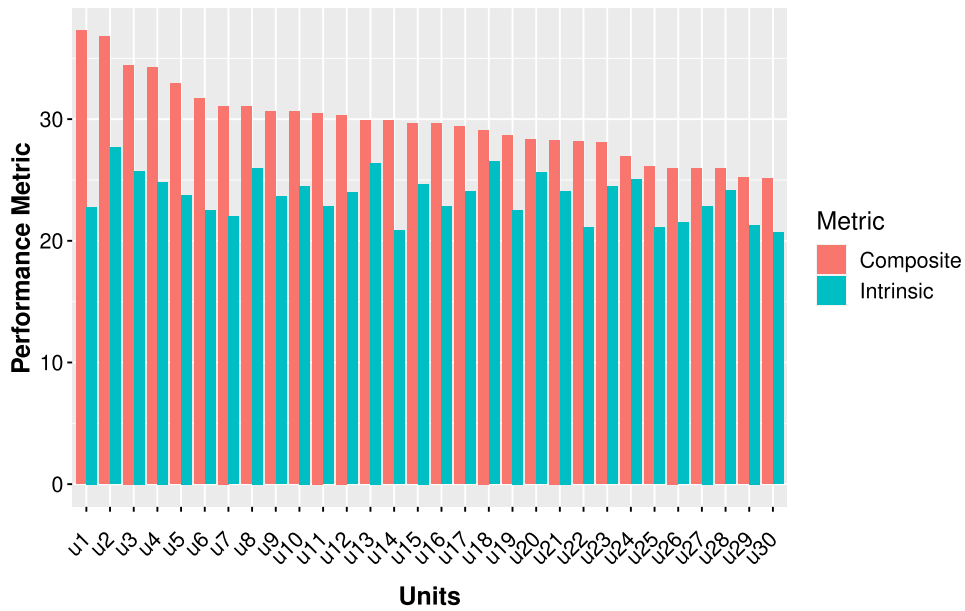


FIG. 1. Comparison of composite and intrinsic performance metrics for simulated data.

$$Y \sim N(\mu_Y = 30, \sigma_Y^2 = 14)$$

$$Y(p) = 2 + P + 2\mathbb{E}[X] + e$$

$$Y(p) \sim N(\mu_{Y(p)} = 30, \sigma_{Y(p)}^2 = 6).$$

Thus, we simulate a scenario in which composite performance is determined by both intrinsic performance and exogenous influences.

Figure 1 shows a benchmarking comparison of ‘true’ composite and intrinsic performance for 30 units sampled from our simulated population. The simulated data of course provide exact values of intrinsic performance. In simulating our model, we want to test how well we can represent these values if they were unobserved, and instead we used our model-based method.

Figure 1 shows that, if the objective is to reveal variance in intrinsic performance, then composite performance is not a particularly good metric to use. Composite performance is of course correlated with intrinsic performance, $\rho_{P,Y} = 0.60$, but because it also reflects exogenous influences it obscures the variance we wish to reveal and provides an incorrect ranking of units on P .

To simulate the properties of model-based adjustment, we use the linear marginal model, $\mathbb{E}[Y, X] = \alpha + \beta X$, to estimate α and β and we recover predicted values $\hat{y}_i$ from this model. We then calculate the

TABLE 1 *Simulation results for model-based adjusted benchmarking*

	Truth	Est.	Var.	MSE
β	2.000	1.960	0.031	0.032
μ_Y	30.000	30.063	0.130	0.134
$\mu_{Y(p)}$	30.000	30.063	0.130	0.134
σ_Y^2	14.000	13.688	3.952	4.046
$\sigma_{Y(p)}^2$	6.000	5.965	0.685	0.686
$\rho_{P,Y}$	0.598	0.589	0.004	0.004
$\rho_{P,Y(p)}$	0.913	0.908	0.000	0.000

adjusted performance metric using

$$\begin{aligned}
 y_i(p) &= y_i - \hat{y}_i + \frac{1}{n} \sum_{i=1}^n \hat{y}_i \\
 &= y_i + \beta(\bar{x} - x_i).
 \end{aligned}$$

The results are derived from 1000 simulated runs on samples of size 100 from a generated population dataset of size 10,000. We report results for estimation of β from the marginal model, the means of Y and $Y(p)$, the variances of Y and $Y(p)$ and Pearson correlation coefficients between the unadjusted and adjusted performance metrics and ‘true’ intrinsic performance. The results are shown in Table 1.

Estimation of β is, as expected, unbiased and consistent because X and P are uncorrelated, and thus the marginal model with P omitted is correctly specified. The adjusted metric has the same simulated mean as the unadjusted metric, because as discussed above, it is on the same scale, but has smaller variance. This is because variance due to exogenous influences X has been removed. The mean simulated correlation coefficient between the unadjusted metric and intrinsic performance is 0.589, while the corresponding value for the adjusted metric is 0.913. This demonstrates that model-based adjustment leads to a substantial improvement in our ability to measure intrinsic performance.

5. Managerial relevance

In this section we apply the method outlined above to demonstrate its managerial relevance through two particular real-world benchmarking comparisons. First, using a unique dataset for 32 international metro operators collected annually by the Transport Strategy Centre (TSC) at Imperial College London, we adjust for the influence of metro size (e.g. scale economies) in benchmarking the average cost of urban metros. Second, using open data on passenger trips from the London Underground metro network, we adjust for service characteristics in benchmarking mean journey time performance across 22 origin-destination routes in Central London.

5.1 *Adjusting for scale economies in benchmarking average costs*

In this application our objective is to compare performance in cost efficiency across units, in this case metros, and to do this we construct a measure of the average cost of producing a metro car kilometre (e.g. the average cost of operating a single car of a metro train for a distance of one kilometre). Cost data for 2015 have been converted to international dollar equivalent using a purchasing power parity

(PPP) currency converter published by the World Bank. The unadjusted benchmark is total operating cost per car kilometre. We find considerable variation in this unadjusted metric across systems, with a mean value of \$5.76, a minimum of \$3.04 and a maximum of \$16.52. Thus, across our sample of urban metros we observe a fivefold difference in average cost with a variance of \$6.21.

We know from the theoretical and empirical literature that the costs of urban metro operations are strongly influenced by the size of metro systems. A recent study by the TSC at Imperial College London finds evidence of considerable scale economies both from system size (measured by total route length in kms) and from the density of use of the system (measured by car kms operated per km of route length) (Anupriya *et al.*, 2019). This study estimate an elasticity of average cost with respect to scale of 0.85 and with respect to density of 0.60. These results are consistent with economic theory and can be explained by the prevalence of fixed and quasi-fixed costs in metro operations.

Since there is typically little metros can do in the short run to make large-scale changes to system scale or density, it is informative to consider average costs that are adjusted for differences in scale and density. Essentially, this amounts to asking the following question: if we remove the influence of scale and density from our measure of composite performance (e.g. nullify the effect of two known strong exogenous influences on performance) how will this change the distribution of our cost performance metric for metros? Other things being equal, since there are positive scale and density effects, we would expect the adjusted metric to reduce the relative cost performance of large and densely used systems since they benefit most from these two exogenous influences.

To adjust for differences in scale and density, we apply the model-based adjustment method outlined above using the parameter values from Anupriya *et al.* (2019). A comparison of adjusted and unadjusted average cost metrics is shown in Fig. 2.

Two aspects of this comparison are worth emphasizing. First, by removing scale effects the variance in average cost is reduced from \$6.21 to \$4.71, thus demonstrating the considerable influence of scale on the unadjusted performance metric. Substantively, this suggest a priori that larger metros and metros with higher densities of traffic will tend to have lower average costs, *ceteris paribus*. Second, the figure shows that performance rankings are radically altered through adjustment.

From a managerial perspective, adjustment for scale and density is relevant for two reasons. First, since metro operators typically do not make large shifts in operational scale in the short run, the influence of scale on performance is largely not actionable, and therefore arguably should not be something we incorporate in a benchmark used to target performance improvements. Second, the revised performance ranking based on the conditional benchmark gives managers a much better undertaking of how well their organization is doing, net of scale effects, and thus a more realistic picture of where they sit within the industry.

5.2 Adjusting for service characteristics in benchmarking metro journey times

In the second application of our model-based adjustment method, the objective is to quantify journey time performance by comparing mean journey times on a sample of the busiest routes on the Central, Jubilee and Victoria lines of the London Underground metro network. Data on passenger journey times are obtained from the Transport for London (TfL) open source data repository, and cover 2 weeks in October in 2015 (Transport for London, 2015). Average journey times are recorded for each unit, in this case an origin–destination route, for each half hour period per day, and a sample of 22 of the busiest origin–destination routes within central London are analysed. Unadjusted benchmarks are calculated at a route level as the mean journey time in minutes across all trips undertaken on the route over the 2-week time period. The unadjusted mean journey times range from a minimum of 5.61 mins to a maximum of

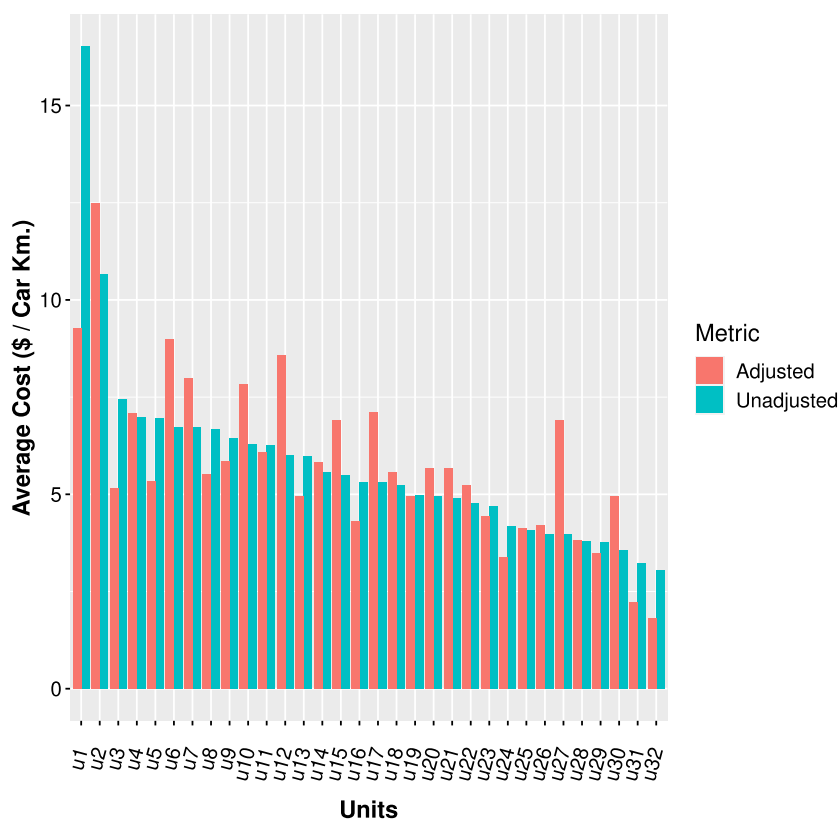


Fig. 2. Unadjusted and adjusted average cost metrics for metro operators.

17.17 mins. Across all routes, the mean value of unadjusted mean journey times is 12.10 mins and the variance is 11.07 mins².

Comparing mean journey times across routes does not in itself provide useful information, since the values primarily represent differences in route length and operating conditions. In their analysis of metro journey times on the London Underground, [Singh et al. \(2020\)](#) demonstrated that along with route length, operational characteristics such as train frequency and congestion from passenger demand have the most influential impact on passenger journey times. In assessing journey time performance, it is of interest for operators to understand whether there are inherent differences across routes that may not be within operator control after conditioning for known operational characteristics. We therefore apply the model-based adjustment method to compute adjusted mean journey times conditioning for route length, train frequency and passenger flows. The data on passenger route flows per half hour per day are obtained from the same 2015 TfL open data set on journey times ([Transport for London, 2015](#)). Open data on average train frequencies are only available for 2018 ([Transport for London, 2018](#)) and these data are considered appropriate to approximate operating frequencies in 2015 for use in the model. Data on route lengths are publicly available from [Transport for London \(2008\)](#). Figure 3 illustrates the resulting adjusted and unadjusted journey time benchmarks.

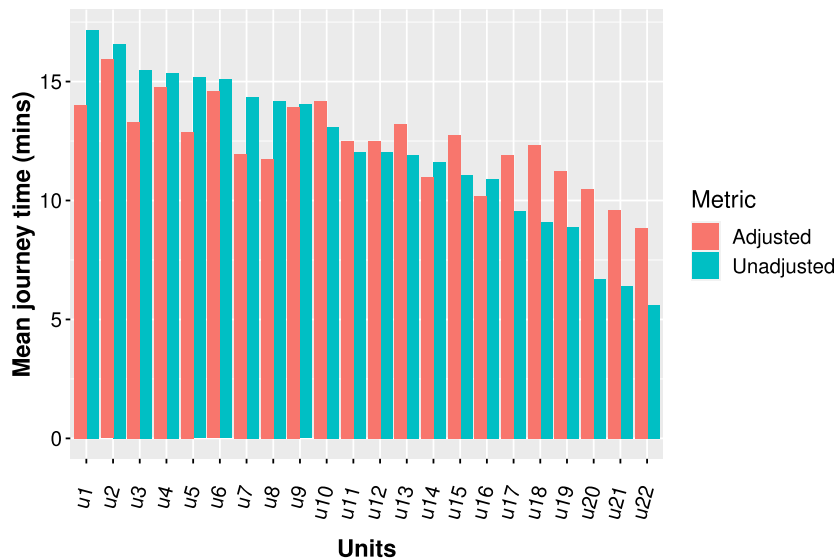


FIG. 3. Unadjusted and adjusted mean journey time metric for sample of routes on London Underground.

In the adjusted model, route-specific effects are significant at a level of 95% ($\alpha=0.05$), indicating that routes do possess inherent characteristics that create prevailing differences in journey time performance over and above sources of variance from known service characteristics. As with the previous example, the adjusted benchmarks yield a different ranking of routes compared to the unadjusted benchmarks. The variance across rankings reduces from 11.07 mins² for the unadjusted benchmarks to 3.14 mins² for the adjusted benchmarks.

Model-based adjusted benchmarking is potentially informative to metro operators in this case, because in showing that the operational factors included in the adjusted model plausibly capture the majority of variance in journey times, it indicates where else operators can best focus their efforts to improve journey time performance.

6. Conclusions

In this paper we have proposed a model-based approach for comparative benchmarking that adjusts for exogenous determinants of performance to produce more targeted benchmarking metrics. We have described the statistical theory underpinning our approach and proposed some example calculations for computation. Simulations have demonstrated the properties of our model-based adjustment approach showing that it performs well in removing ‘contamination’ through variance reduction via marginalization. The managerial relevance of the method is demonstrated via two real-world transport industry applications: adjusting for economies of scale and density in benchmarking average costs of urban metros and for service characteristics in benchmarking metro journey times.

Model-based adjusted benchmarking can help managers recover more focused measures of performance that they can then target for improvement and best practice. It also allows them to compare conditional and unconditional benchmarks, which can be illuminating in identifying areas for further investigation.

REFERENCES

- AGRELL, P. J. & BOGETOFT, P. (2017) Regulatory benchmarking: models, analyses and applications. *Data Envelop. Anal. J.*, **3**, 49–91.
- AIGNER, D., LOVELL, C. A. K. & SCHMIDT, P. (1977) Formulation and estimation of stochastic frontier production function models. *J. Econometrics*, **1**, 21–37.
- ALEXANDER, C. A., BUSCH, G. & STRINGER, K. (2003) Implementing and interpreting a data envelopment analysis model to assess the efficiency of health systems in developing countries. *IMA J. Manag. Math.*, **14**, 49–63.
- ANUPRIYA, G., J. D., CARBO, J. M. & ANDERSON, R. J. (2019) A dynamic panel modelling approach for analysis of urban rail transport costs. *Working Paper, Transport Strategy Centre, Imperial College London*.
- BOGETOFT, P. (2012) *Performance Benchmarking: Measuring and Managing Performance*. New York, NY: Springer.
- CAMP, R. C. (1989) *Benchmarking: The Search for Industry Best Practices that Lead to Superior Performance*. Milwaukee, WI: ASQC Quality Press.
- CANAVAN, S., GRAHAM, D. J., MELO, P. C., ANDERSON, R. J., BARRON, A. S. & COHEN, J. M. (2015) Impacts of moving-block signaling on technical efficiency: application of propensity score matching on urban metro rail systems. *Transport. Res. Record J. Transport. Res. Board*, **2543**, 68–74.
- CHARNES, A., COOPER, W. W. & RHODES, E. (1978) Measuring the inefficiency of decision making units. *Eur. J. Oper. Res.*, **2**, 429–444.
- DARAIO, C. & SIMAR, L. (2005) Introducing environmental variables in nonparametric frontier models: a probabilistic approach. *J. Prod. Anal.*, **24**, 93–121.
- DARAIO, C. & SIMAR, L. (2007a) *Advanced Robust and Nonparametric Methods in Efficiency Analysis: Methodology and Applications*. New York, NY: Springer.
- DARAIO, C. & SIMAR, L. (2007b) Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach. *J. Prod. Anal.*, **28**, 13–32.
- DIGGLE, P. J., HEAGERTY, P., LIANG, K.-Y. & ZEGER, S. L. (2002) *Analysis of longitudinal data*. Oxford: Oxford University Press.
- FRANCIS, G. & HOLLOWAY, J. (2007) What have we learned? Themes from the literature on best-practice benchmarking. *Int. J. Manag. Rev.*, **9**, 171–189.
- GRAHAM, D. J. (2008) Productivity and efficiency in urban railways: para-metric and non-parametric estimates. *Transport. Res. Part E Log. Transport. Rev.*, **44**, 84–99.
- GREENE, W. H. (2005) Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *J. Econometrics*, **126**, 269–303.
- IMBENS, G. W. & RUBIN, D. B. (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY, USA: Cambridge University Press.
- IMBENS, G. W. & WOOLDRIDGE, J. M. (2009) Recent developments in the econometrics of program evaluation. *J. Econ. Lit.*, **47**, 5–86.
- KAPLAN, R. & NORTON, D. (1992) The balanced scorecard-measures that drive performance. *Harvard Business Rev.*, **70**, 71–79.
- KUO, C. (2011) Cost efficiency estimations and the equity returns for the us public solar energy firms in 1990–2008. *IMA J. Manag. Math.*, **22**, 307–321.
- LEIBFRIED, K. H. J. & MCNAIR, C. J. (1992) *Benchmarking: A Tool for Continuous Improvement*. New York, NY: McGraw-Hill.
- MCCULLOCH, C. E., SEARLE, S. R. & NEUHAUS, J. (2008) *Generalized, Linear, and Mixed Models*, 2nd edn. New York: Wiley.
- MEEUSEN, W. & BROECK, J. (1977) Efficiency estimation from cobb-douglas production functions with composed error. *Internat. Econom. Rev.*, **18**, 435–444.
- MORIARTY, J. P. & SMALLMAN, C. (2009) En route to a theory of benchmarking. *Benchmarking*, **16**, 484–503.
- PEARL, J., GLYMOUR, M. & JEWELL, N. (2016) *Causal Inference in Statistics: A Primer*. Chichester, UK: Wiley.

- PETERS, J., JANZING, D. & SCHÖLKOPF, B. (2017) *Elements of Causal Inference-Foundations and Learning Algorithms. Adaptive Computation and Machine Learning Series*. Cambridge, MA, USA: The MIT Press.
- SCHMIDT, A. M., MOREIRA, A. R. B., HELFAND, S. M. & FONSECA, T. C. O. (2009) Spatial stochastic frontier models: accounting for unobserved local determinants of inefficiency. *J. Prod. Anal.*, **31**, 101–112.
- SINGH, R., HÖRCHER, D., GRAHAM, D. J. & ANDERSON, R. J. (2020) Decomposing journey times on urban metro systems via semiparametric mixed methods. *Transport. Res. Part C Emerg. Technol.*, **114**, 140–163.
- Transport for London (2008). Inter station database. Accessed at www.whatdotheyknow.com/request/distance_between_adjacent_underg.
- Transport for London (2015). Tap counts and network origin-destination data (Oct 2015 sample). Accessed at passenger.data.tfl.gov.uk.
- Transport for London (2018). 2018 MTT Link Frequency. Accessed at crowding.data.tfl.gov.uk.
- TSIATIS, A. A. (2006) *Semiparametric Theory and Missing Data*. Berlin: Springer.
- TSIONAS, E. G. (2002) Stochastic frontier models with random coefficients. *J. Appl. Econom.*, **17**, 127–147.
- TSOLAS, I. E. (2010) Assessing performance in greek bauxite mining by means of frontier estimation methodologies. *IMA J. Manag. Math.*, **21**, 253–265.
- TSOLAS, I. E. (2011) Relative profitability and stock market performance of listed commercial banks on the athens exchange: a non-parametric approach. *IMA J. Manag. Math.*, **22**, 323–342.

