Benchmarking the Performance of Urban Rail Transit Systems: A Machine Learning Application

Farah A. Awad[1]*, Daniel J. Graham[2], Laila AitBihiOuali[3], Ramandeep Singh[4], and Alexander Barron[5]

[1] Postgraduate Researcher, Transport Strategy Centre, Centre for Transport Studies, Department of Civil Engineering, Imperial College London, London, United Kingdom
[2] Professor, Transport Strategy Centre, Centre for Transport Studies, Department of Civil Engineering, Imperial College London, London, United Kingdom
[3] Assistant Professor, Department of Civil Engineering, University of Southampton, Southampton, United Kingdom
[4] Postdoctoral Research Associate, Transport Strategy Centre, Centre for Transport Studies, Department of Civil Engineering, Imperial College London, London, United Kingdom
[5] Associate Director, Transport Strategy Centre, Centre for Transport Studies, Department of Civil Engineering, Imperial College London, London, United Kingdom

SHORT SUMMARY

Urban rail transit systems operate in heterogenous environments. Distinguishing between inherent performance and the role of efficiencies due to differing environmental and system-specific characteristics is challenging. This study provides a data-driven benchmarking method which accommodates heterogeneity in operational performance among urban rail systems. Using an international dataset of 36 metros in year 2016, operators are clustered into peer groups through clustering algorithms based on operational characteristics. ANOVA and post-hoc tests are then applied to explore variations between clusters. Finally, efficiency performance benchmarking is conducted through Data Envelopment Analysis. Our clustering results corroborate to the natural geographic grouping of the systems. Moreover, our results show that the use of an aggregated index is inadequate to represent the operator's overall quality-of-service. Finally, results show that clustering operators into groups based on similarities in their operational characteristics would introduce more meaningful benchmarks for best practices as they are more likely to be attainable.

Keywords: benchmarking, cluster analysis, DEA, performance evaluation, urban rail

1. INTRODUCTION

Benchmarking practices allow for the comparison of performance between operators which is important for identifying service dimensions with poor performance and best practices to improve it. Urban rail transit (URT) systems are typically monopolies with no systems operating in the same area that could allow for a comparison of performance. Hence, international benchmarking practices are necessary. However, their heterogeneity in terms of operational characteristics and external influences makes it difficult to recognize inherent performance. There are many elements of URT performance. For instance, financial, environmental, and operational performance. Inherent performance is not observed. However, different key performance indicators (KPIs) measure some combination of inherent performance and environmental or exogenous advantage.

Financial performance is normally estimated as a measure of efficiency where the output of the service is represented through measures of the supplied service, or the demand. However, operational characteristics of the service, which are directly related to the service quality perceived by the passengers, also vary among operators. URT Quality-of-service is complex, fuzzy, and is composed of many indicators with many interactions among them. Therefore, the assumption of homogeneity of service quality, which is conventional in public transit efficiency studies, may provide misleading results and unrealistic best practice setting as operators which have vast differences in their operational characteristics are compared together.

Moreover, the efficiency of URT is affected by a range of internal and external influences, for instance, the scale and density of operations (Anupriya et al. 2020; Graham 2008), population density (Lobo and Couto, 2016), economic vibrancy (Graham 2008; Lobo et al. 2016), and the governance structure (Jain *et al.*, 2008). Therefore, distinguishing between inherent performance of firms and the role of efficiencies due to differing environmental and system-specific characteristics poses a methodological challenge.

Efficiency studies of public services fall under four main categories (Estruch-juan et al. 2020): partial efficiency measures (e.g. Santos *et al.*, 2010; Tsai and Mulley, 2013; Allen, 2014), average efficiency measures (e.g. Anupriya *et al.*, 2020), econometric frontier methods (e.g. Lobo and Couto, 2016), and non-parametric frontier methods (e.g. Graham, 2008; Jain *et al.*, 2008). Although some of these measures control for the effects of certain exogenous variables, withal they result in an aggregate analysis across all operators that provide different levels of service and operate under varied conditions. The objective of this study is to propose a systemic data-driven approach to accommodate heterogeneity in URT by reducing the influence of exogenous factors and operational characteristics on efficiency metrics. While variance in performance is necessary for benchmarking, by forming logical groupings of operators we can seek to reduce the influence of heterogeneity that is not related to inherent performance.

Cluster Analysis (CA) (Driver et al. 1932) is an unsupervised machine learning method which identifies patterns based on similarities within the dataset. CA has been used in the public transit literature to create peer groups. For instance, transit operators have been clustered based on characteristics of the operating environment (Arndt et al. 2011), and based on traditional drivers of productivity such as service supply, network length, and city characteristics (Karlaftis et al. 2002; Ripplinger 2010; Zemp et al. 2011). Moreover, on a city/country level, CA was applied to obtain peer groups based on indicators of sustainability of their transit systems (Alonso et al. 2015; Persia et al. 2016; Shen et al. 2020), based on ridership influencing factors (Ederer et al. 2019), and based on their accessibility (Hawas et al. 2016). In a significant contribution by Fielding *et al.* (1985), bus operators were classified based on similar operational indicators including speed, peak-to-base ratio, number of peak vehicles, and total vehicle miles. The findings shed light on the importance of contextualizing benchmarks used for performance comparisons based on operational characteristics. We note that indicators used to classify systems have mainly focused on traditional productivity inputs and characteristics of the operating environment. However, operational characteristics to classify transit operators have rarely been used. Moreover, cross-country analysis for public transit systems has only been done on a macro-level using aggregated city/country data.

In this study, we apply machine learning clustering algorithms to group URT operators into peer groups, based indicators of operational characteristics to undertake like-for-like comparisons, and to investigate the correlations between dimensions of the service and characteristics of the system to better understand differences in performance. As far as the authors are aware, this is the first study on URT which applies clustering algorithms to create peer groups for performance comparisons which bridges a major gap in the benchmarking literature by accounting for the heterogeneity and interdependencies among different service dimensions.

2. METHODOLOGY

The methodological framework is demonstrated in Figure 1. It includes four main steps which are detailed below.

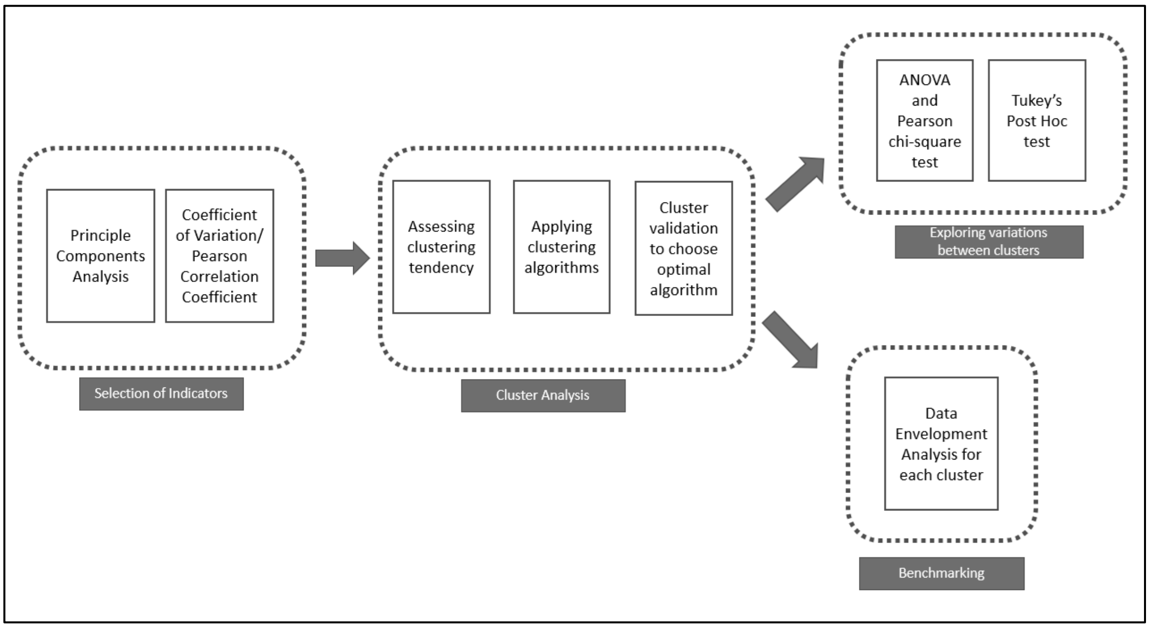
Principle Components Analysis
Coefficient of Variation/ Pearson Correlation Coefficient
Selection of Indicators
Assessing clustering tendency
Applying clustering algorithms
Cluster validation to choose optimal algorithm
Cluster Analysis
ANOVA and Pearson chi-square test
Tukey's Post Hoc test
Exploring variations between clusters
Data Envelopment Analysis for each cluster
Benchmarking

Fig. 1. Proposed Framework for Research Design

Indicators

To choose representative indicators of operational characteristics, we first apply Principal Components Analysis (PCA) (Hotelling 1933). PCA is a factor analysis method which applies projection methods on multivariate data to produce orthogonal linear combinations of the original variables. We retain principal components explaining at least 50% of the variance in the data. Next, indicators with loadings having an absolute value of at least 0.25 are retained. Finally, four indicators are chosen after referring to the Coefficient of Variation and Pearson Correlation Coefficients as the goal is to find indicators that have enough variation to generate distinct clusters while providing enough information about different dimensions of the service.

Cluster Analysis

In CA, unlabelled data is grouped into distinct clusters where similarity within a cluster is maximized, while similarity between clusters in minimized. Three clustering algorithms are applied: K-means, K-medoids, and agglomerative hierarchical clustering. The optimal algorithm is chosen based on the Silhouette coefficient and Dunn index.

In K-means, cluster centroids are represented by the means. Clusters are randomly initialized, then the mean is continuously updated through an iterative process where the objective is to minimize the total within-cluster variation. The within-cluster variation of cluster (C_I) is defined as follows:

$$W(C_I) = \sum_{i \in C_I}(i - \mu_I)^2 \tag{1}$$

Where i is an object assigned to cluster C_I, and μ_I is the mean of the objects assigned to that cluster. Hence, the total within cluster variation is defined as follows:

$$TW = \sum_{k=1}^{K} W(C_k) \tag{2}$$

K-medoids may be considered as a robust extension to the K-means algorithm where representative objects (*medoids*) are assigned as centroids rather than means. The objective is to minimize the sum of the distances (DS) of the objects to their medoids which can be represented as follows:

$$DS = \sum_{k=1}^{K} \sum_{i \in C_I} \mathrm{d}(i, M_I) \tag{3}$$

Where d is the distance between an object assigned to cluster C_I and the medoid M_I of that cluster. Hence, each medoid is swapped with each non-medoid object while computing the objective function. This process is continued until the objective function can no longer be minimized.

In agglomerative algorithms, the dissimilarity matrix is computed first. Then, a linkage function uses the distances provided by the dissimilarity matrix to cluster all objects into pairs based on the minimum distance between each two clusters. In a sequential process, clusters are grouped into larger clusters until all objects are grouped into one cluster. Ward's minimum variance linkage method calculates the distance (D) between two clusters C_I and C_J as the minimum total within-cluster variance by merging the clusters with minimum distance at each step as follows:

$$D\left(C_I C_J\right) = \frac{N_I . N_J}{N_I + N_J}(\mu_I - \mu_J)^2 \tag{4}$$

Where N_I and N_J are the number of objects in clusters C_I and C_J, respectively; and μ_I and μ_J are the means of clusters C_I and C_J, respectively.

Exploring Variations between Clusters

We explore the differences between clusters in terms of operational characteristics and exogenous influences. The analysis of variance (ANOVA) is applied to test whether the differences in features are statistically significant between clusters. The features which are found to have significant differences between clusters are further analyzed using the Tukey's post-hoc test (Tukey 1949). This test identifies which clusters are different from the others based on the studentized range distribution. It compares the means of all possible pairs of groups which identifies distinct features of each cluster.

Efficiency Performance Benchmarking using DEA

We use Data Envelopment Analysis (DEA) to obtain efficiency scores of URT operators relative to the performance of identified best-performers. Two models are estimated and compared, the first model estimates efficiency scores using the whole dataset, while the second model estimates efficiency scores for one cluster only. The DEA specification can be represented as follows:

$$\min \quad \psi$$

$$\text{s.t} \begin{cases} \sum_k \lambda_k X_k \ \leq \ \psi X_o \\ \sum_k \lambda_k Y_k \ \geq \ \psi Y_o \\ \sum_k \lambda_k \ = \ 1 \\ \lambda \ \geq 0 \end{cases} \tag{5}$$

where Y is a vector of outputs; X is a vector of inputs; $K, k = 1, \dots, N$ represents the set of units; ψ is the efficiency score for unit o, and λ is an N×1 vector of constants.

We use three inputs and one output to reflect the production process of URT systems. The inputs include labor (total number of staff hours); fleet (total number of cars); and network length (km), while the output is the supplied service measured as car kilometers.

3. RESULTS AND DISCUSSION

The dataset used in this study consists of 36 international URT systems in year 2016 which are shown in Figure 2. The data comes from the Community of Metros (COMET®), a global urban railway benchmarking group managed by the Transport Strategy Centre (TSC) at Imperial College London. Due to the sensitivity of the data, a confidentiality agreement requires the results to be presented in an anonymized form. Therefore, each URT is referred to by an ID indicating the geographical region to which the system operates in and a numbering which is set randomly.

In the first step, four indicators of operational characteristics were chosen as inputs for CA: capacity utilization (proxy for crowding) measured as passenger km/ capacity km; rate of fatalities (proxy for safety risk) measured as fatalities/ billion passenger km; peak-period headway measured as average seconds between cars; and the average scheduled commercial speed in km/h. Next, out of three clustering algorithms, Ward's hierarchical clustering method was found to be the most optimal, which resulted in four urban rail operator peer groups with similar operational characteristics. The dendrogram demonstrating the hierarchical classification is shown in Figure 3, and the Silhouette plot is shown in Figure 4. The first cluster contains 17 operators, all of which operate in the Asia-Pacific region (As) or in Latin America (LAm) except for one European (Eu) metro. The second cluster has 10 members which are European or North American (NAm) except for one Latin American operator. The third cluster contains one Asian-Pacific and two European members. Finally, the fourth cluster has 6 members, 5 of which are European or North American while one member is Asian-Pacific.

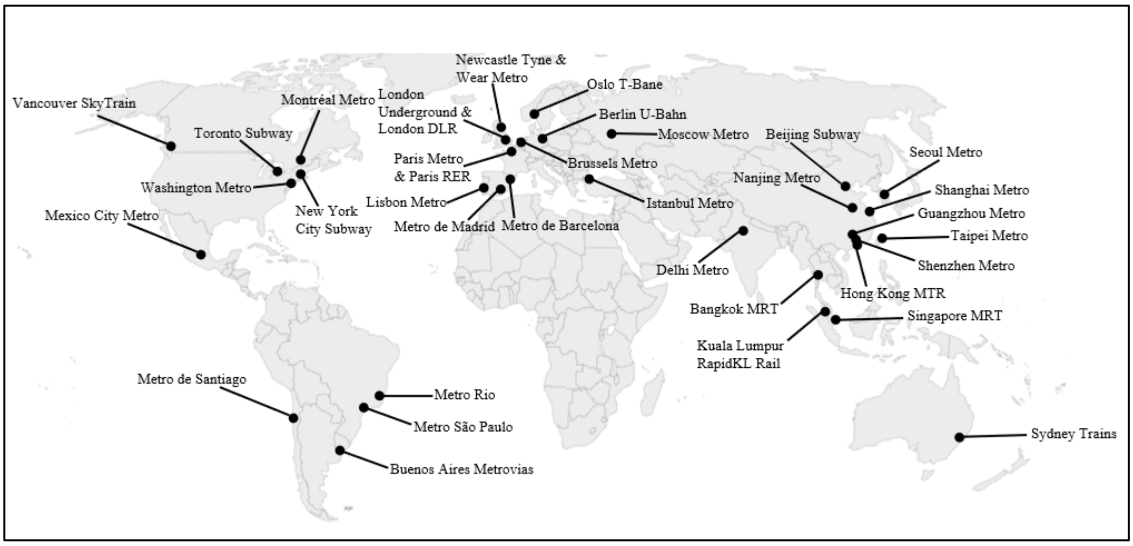

Fig. 2. Geographic Distribution of URT Dataset

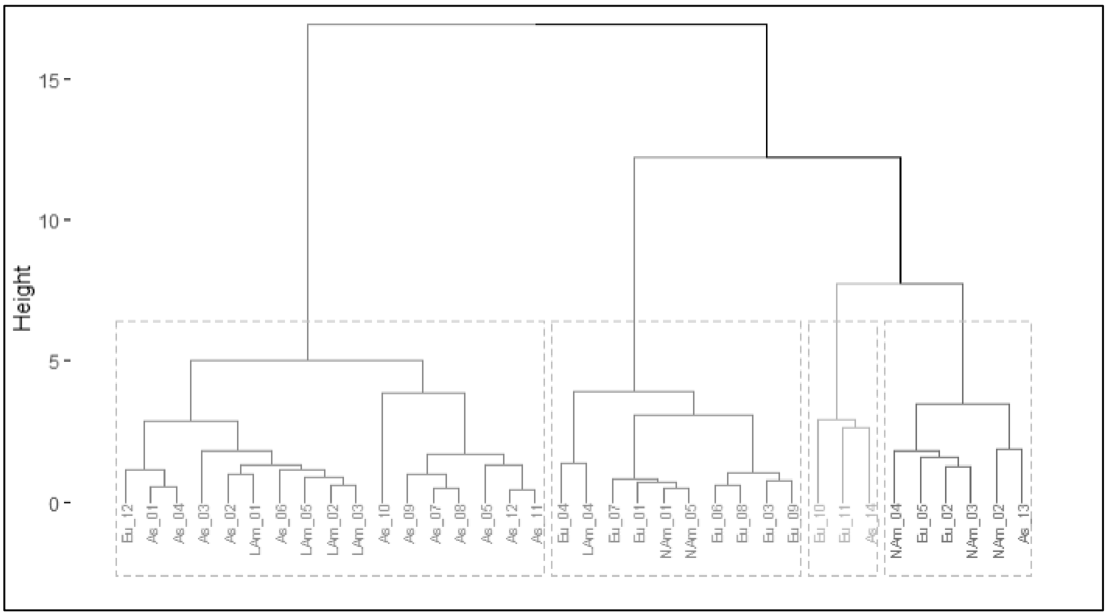

Fig. 3. Cluster Dendrogram of URT Operators

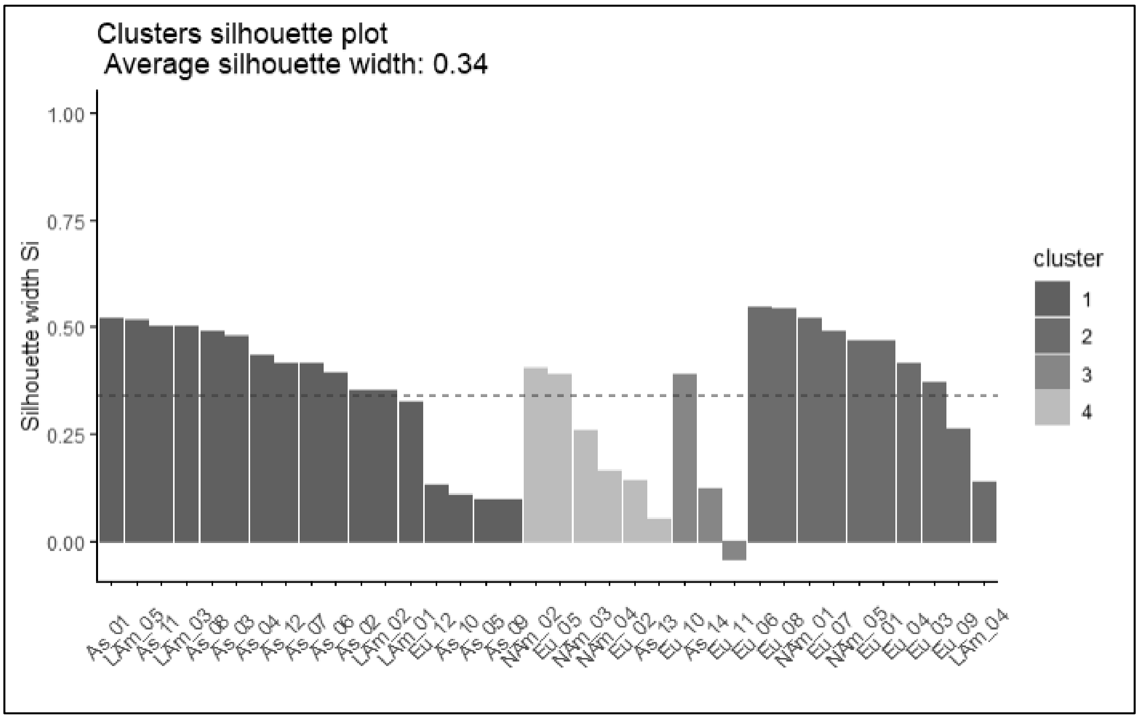

Fig. 4. Hierarchical Clustering Silhouette Plot

To better understand the distinct features of each cluster, First, we apply ANOVA and Post-hoc tests on the 4 indicators of operational characteristics used as inputs in the CA. Then we apply the tests for other indicators of performance and exogenous influences which are shown in Figure 5.

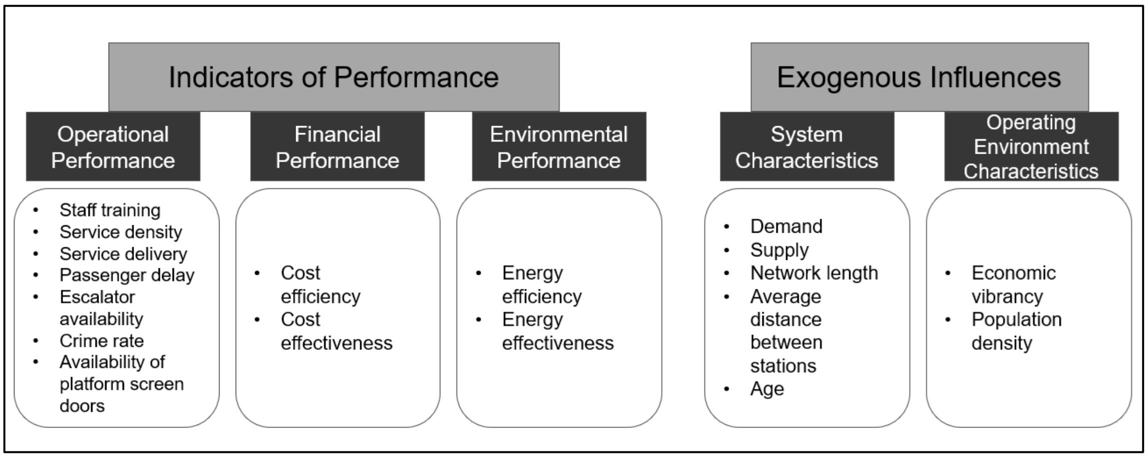

Figure 5: Variables used to Explore Variations between Clusters

Test results show that Cluster 1 contains newer URT systems with higher crowding; higher safety; and have more station facilities. Systems in this cluster also operate in regions with lower economic vibrancy; have higher demand; and provide higher service density. Cluster 2 contains systems which operate in lower speed and have shorter distances between stations. Systems in Cluster 3 have lower crowding; lower safety; and lower peak-period headway. Finally, systems in Cluster 4 have higher speeds and higher passenger delay rates.

Furthermore, energy efficiency and cost efficiency do not significantly differ between clusters. We note however, that Cluster 1 has a significantly higher energy effectiveness and cost effectiveness. This is likely due to its higher passenger load. Nevertheless, the network size and the supplied level of service do not differ between groups. This may indicate that the increase in the resources required to accommodate an increase in demand is less than proportional to the increase in ridership. Results also show that operational characteristics which reflect on the quality-of-service experienced by the user, does not seem to be correlated with traditional drivers of productivity. Therefore, using financial indicators as the sole indicators of performance may be misleading as they would give an incomplete picture of performance.

In the final step, two DEA models are estimated to explore the significance of creating contextualized peer groups in efficiency studies. The first model estimates efficiency scores using the whole dataset, while the second model includes only members of Cluster 1. Results from the two models are shown in Table 1. We find that by comparing operators within a cluster, efficiency scores are higher, and more efficient units are detected. This may indicate that part of the inefficiencies obtained in the first model were due to differences in operational practices or technologies, which would be difficult to alter. Therefore, comparing operators with similar operational characteristics would be more useful for detecting best practices that can be applied by inefficient operators to improve their productivity.

4. CONCLUSIONS

We present an improved method for transit performance benchmarking by analytically constructing operator peer groups with similar operational characteristics to enable like-for-like comparisons, which has not been done so far in the urban rail literature. Our results show that benchmarking members of a cluster with similar operational characteristics generates higher efficiency scores, and more efficient units are detected. This indicates that benchmarking operators with high heterogeneity gives misleading results as exogenous influences are interpreted as inefficiency. Therefore, comparing operators with similar operational characteristics would be more useful for detecting best practices that can be applied by inefficient operators to improve their efficiency. The analysis of variations between clusters emphasize that the use of a single aggregated index is inadequate to represent the operator's overall quality-of-service. Moreover, the clustering results corroborate to the natural geographic grouping of the URT systems, which draw attention to the importance of considering heterogeneity in terms of cultural and city-specific characteristics. This methodology addresses future transport needs as it provides a reference for urban rail operators by examining the outcomes of best performers in each cluster and using the lessons learned from other systems. This will enable them to make informed decisions on service planning and resource allocation which can provide substantial time and money savings.

Table 1. DEA efficiency scores for members of Cluster 1

ID	Efficiency Scores (relative to the whole dataset)	Efficiency Score (relative to members of Cluster 1)
As_01	0.760	0.760
As_02	0.960	0.977
As_03	0.839	0.843
As_04	1.000	1.000
As_05	0.785	0.817
As_06	0.753	1.000
As_07	0.866	0.972
As_08	0.852	0.852
As_09	0.723	0.813
As_10	1.000	1.000
As_11	0.687	0.709
As_12	0.755	1.000
LAm_01	0.859	0.866
LAm_02	0.857	0.909
LAm_03	1.000	1.000
LAm_05	0.833	1.000
Eu_12	1.000	1.000

ACKNOWLEDGMENT

We thank the members of the COMET consortia for providing their data for this research.

REFERENCES

Allen, DW. 2014. "Economies of Scale in Operating Costs for LRT and Streetcars." In *Compendium of Papers of the 93rd Transportation Research Board Annual Meeting*, Washington, DC.

Alonso, Andrea, Andrés Monzón, and Rocío Cascajo. 2015. "Comparative Analysis of Passenger Transport Sustainability in European Cities." *Ecological Indicators* 48: 578–92.

Anupriya et al. 2020. "Understanding the Costs of Urban Rail Transport Operations." *Transportation Research Part B: Methodological* 138: 292–316.

Arndt, Jeffrey, Suzie Edrington, Matthew Sandidge, and Luca Quadrifoglio. 2011. 7 *Peer Grouping and Performance Measurement to Improve Rural and Urban Transit in Texas*

(No. FHWA/TX-11/0-6205-1).

Driver, Harold Edson, and Alfred Louis Kroeber. 1932. “Quantitative Expression of Cultural Relationships.” *Berkeley: University of California Press* 31(4).

Ederer, David et al. 2019. “Comparing Transit Agency Peer Groups Using Cluster Analysis.” *Transportation Research Record* 2673(11): 505–16.

Estruch-juan, Elvira, Enrique Cabrera Jr, and Maria Molinos-Senante. 2020. “Are Frontier Efficiency Methods Adequate to Compare the Efficiency of Water Utilities for Regulatory Purposes?” *Water* 12(4).

Fielding, Gordon J., Mary E. Brenner, and Katherine Faust. 1985. “Typology for Bus Transit.” *Transportation Research Part A: General* 19(3): 269–78.

Graham, Daniel J. 2008. “Productivity and Efficiency in Urban Railways: Parametric and Non-Parametric Estimates.” *Transportation Research Part E: Logistics and Transportation Review* 44(1): 84–99.

Hawas, Yaser E., Mohammad Nurul Hassan, and Ammar Abulibdeh. 2016. “A Multi-Criteria Approach of Assessing Public Transport Accessibility at a Strategic Level.” *Journal of Transport Geography* 57: 19–34.

Hotelling, H. 1933. “Analysis of a Complex of Statistical Variables into Principal Components.” *Journal of Educational Psychology* 24(6): 417–41. (July 24, 2021).

Jain, Priyanka, Sharon Cullinane, and Kevin Cullinane. 2008. “The Impact of Governance Development Models on Urban Rail Efficiency.” *Transportation Research Part A: Policy and Practice* 42(9): 1238–50.

Karlaftis, Matthew G., and Patrick McCarthy. 2002. “Cost Structures of Public Transit Systems: A Panel Data Analysis.” *Transportation Research Part E: Logistics and Transportation Review* 38(1): 1–18.

Lobo, António, and António Couto. 2016. “Technical Efficiency of European Metro Systems: The Effects of Operational Management and Socioeconomic Environment.” *Networks and Spatial Economics* 16(3): 723–42.

Persia, Luca, Ernesto Cipriani, Veronica Sgarra, and Eleonora Meta. 2016. “Strategies and Measures for Sustainable Urban Transport Systems.” *Transportation Research Procedia* 14: 955–64.

Ripplinger, David. 2010. “Classifying Rural and Small Urban Transit Agencies.” *Transportation Research Record* (2145): 100–107.

Santos, Joana, Pedro Simões, Álvaro Costa, and Rui Cunha Marques. 2010. *Efficiency of the Portuguese Metros. Is It Different from Other European Metros?* Munich, Germany.

Shen, Yongjun, Qiong Bao, and Elke Hermans. 2020. “Applying an Alternative Approach for Assessing Sustainable Road Transport: A Benchmarking Analysis on Eu Countries.” *Sustainability (Switzerland)* 12(24): 1–16.

Tsai, Chi-hong Patrick, and Corinne Mulley. 2013. “Benchmarking the Efficiency Performance of International Metro Systems.” *Proceedings of the Eastern Asia Society for Transportation Studies* 9.

Tukey, John W. 1949. “Comparing Individual Means in the Analysis of Variance.” *Biometrics* 5(2): 99.

Zemp, Stefan, Michael Stauffacher, Daniel J. Lang, and Roland W. Scholz. 2011. “Classifying Railway Stations for Strategic Transport and Land Use Planning: Context Matters!” *Journal of Transport Geography* 19(4): 670–79.