Applied Network Science

RESEARCH Open Access

Exploring the association between network centralities and passenger flows in metro systems

Athanasios Kopsidas[1*], Aristeides Douvaras[2] and Konstantinos Kepaptsoglou[1]

*Correspondence:
akopsidas@mail.ntua.gr

[1] School of Rural and Surveying Engineering, National Technical University of Athens, 9 Iroon Polytechniou Str, Zografou Campus, 15773 Athens, Greece
[2] Athens University of Economics and Business, 47A Evelpidon Str. & 33 Lefkados Str., Athens, Greece

Abstract

Network science offers valuable tools for planning and managing public transportation systems, with measures such as network centralities proposed as complementary predictors of ridership. This paper explores the relationship between different cases of passenger flows at metro stations and network centralities within both metro and alternative public transport (substitute) networks; such an association can be useful for managing metro system operations when disruptions occur. For that purpose, linear regression and non-parametric machine learning models are developed and compared. The Athens metro system is used as a testbed for developing the proposed methodology. The findings of this study can be used for deriving medium-term ridership estimates in cases of metro disruptions, as the proposed methodology can support contingency plans for both platform and rail track disruptions.

Keywords: Network theory, Centrality, Metro system, Passenger flows, Linear regression, Machine learning, XGBoost

Introduction

Metro systems are crucial in metropolitan areas since they offer fast, convenient, and reliable transportation services. Information on ridership (often represented by passenger flows between metro stations) is essential for efficient planning and management of metro operations, such as the development of timetables, the allocation of resources, and so on. For that purpose, several econometric and machine/deep learning models have been developed for predicting passenger flows in metro systems (Han et al. 2021). The associated problem of estimating metro passenger flows is usually approached from a spatial–temporal perspective, while the temporal component is treated as a time-series (Ou et al. 2020). Econometric time-series models are usually based on autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average (SARIMA) (Zheng et al. 2020), while machine learning and deep learning-based efforts mainly include neural networks and support vector machine models (Sun et al. 2015; Guo et al. 2019; Li et al. 2019; Yang et al. 2021). Machine learning techniques have

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

also been applied to long-term passenger flow estimates (Toqué et al. 2017) and to ridership predictions of new stations and lines (Gong et al. 2020; Wang et al. 2022a, b).

Complex network theory (CNT) has been used for the analysis of traveler flows, including passenger flows in public transportation systems (Zhang et al. 2022). Since there exist correlations between topological properties and human activities, CNT can be effectively used for that purpose (Leung et al. 2011). Indeed, network properties can be useful passenger flow predictors, and network-based models can provide fast and efficient decision models for that purpose (Luo et al. 2020). In this context, complex network characteristics of road networks were recognized by the literature as determinants of traffic flow: their degree of association depends on network structure (Zhao and Zhao 2016), topological and angular distance-based centralities seem to be more appropriate for predictions than metric distance-based ones (Omer and Jiang 2015), and multiple centralities can demonstrate better performance than single-variable ones (Pun et al. 2019). Betweenness is the most frequently used centrality for traffic flow predictions, but most researchers agree that it is not sufficient in its conventional form (Kazerani and Winter 2009b; Leung et al. 2011; Gao et al. 2013; Ye et al. 2016). As such, several modifications of betweenness centrality have been proposed (Kazerani and Winter 2009a; Galafassi and Bazzan 2013; Puzis et al. 2013; Henry et al. 2019; Zhang and Chen 2020; Cogoni et al. 2023). Additional centralities used as traffic flow predictors include closeness (Jayasinghe and Sano 2017), degree, PageRank (Zhao et al. 2017), and stress (Lowry 2014). Similarly, conventional or modified betweenness is the most common centrality used as a flow predictor in the cases of cycling (Cooper 2017; Chan and Cooper 2019; Hochmair et al. 2019) and pedestrian flows (Agryzkov et al. 2019; Cooper et al. 2021; Sevtsuk 2021).

In public transportation, the association between network properties and passenger flows has been investigated, along with the possibility of using the former as efficient passenger flow predictors. Jayasinghe and Munshi (2014) used closeness, straightness, and betweenness centralities and developed linear regression models to estimate boardings and alightings at public transport stops. Senousi et al. (2022) explored the association between conventional and modified centralities with passenger flows in public transport networks, concluding that centralities can be used as passenger flow predictors only when specific network representations are concerned. Wang et al. (2022a, b) explored factors affecting bus ridership considering spatial autocorrelation and found that betweenness centralities within bus and road networks can be significant determinants of bus ridership on specific days and timeslots. Dai et al. (2022) used degree, betweenness, and closeness centralities to investigate the spatial relationship between bus line and temporal bus flow networks and found some time-dependent differences among them. Liu et al. (2022) used machine learning models to explore the transfer ridership between bus and metro, with network density and closeness centrality being the most influential factors. When it comes to metro systems, He et al. (2019) suggested that network properties such as degree and betweenness centralities are associated with ridership at metro stations. Luo et al. (2020) showed that both infrastructure and service level-based network centralities can be exclusively used to estimate passenger flows in public transport systems. In this direction, Kopsidas et al. (2023) identified passenger flow predictors not only among the centralities within the network of metro infrastructures but also in their substitute network, which is the network of alternative public transport options that connect metro stations.

Although CNT-based models can incorporate nontrivial and dynamic properties of transportation networks and thus bring crucial information to light (Zhang et al. 2022), their usage for predictions is still limited. Their accuracy, albeit increasing, is still questionable, and the most efficient predictors are yet to be found. This is because the number of measures that can be used is vast since there are endless capabilities in metro network specifications. Physical, traffic, or service networks can be formed with stations, lines, intersections, etc. as nodes, any form of possible interaction like direct or indirect connections as edges, attributes such as passenger flows and travel times as weights, as well as different network representations (L-space, P-space, etc.) (Lin and Ban 2013). At the same time, although CNT measures have been incorporated into predictive models, they are still complementary elements. Only recently has the potential of exclusively network-based models been investigated.

In this direction, this study aims to further explore CNT-based predictors of passenger flows in metro systems, with focus on fast and reliable estimations in cases of metro disruptions. For this purpose, econometric and machine learning models are developed to associate network centralities with passenger flows at metro stations. Both the options of total flows and origin–destination (OD) flows are investigated since the information extracted from each case is uniquely valuable for disruption management. OD flows account for passenger departures and arrivals from/to metro stations and are mostly related to platform disruptions. Total flows are the sum of OD flows plus passthroughs, i.e., passengers who pass through a metro station without alighting, and are mostly related to rail track disruptions. Indeed, disruptions of the rail track infrastructure near a metro station would affect all passengers at that station, who would either board, alight, or pass through. On the other hand, disruptions of the station platform would affect only the passengers who were willing to board or alight the metro system, but not the ones passing through the station. On this occasion, the disruption could be bypassed by utilizing the closest metro station, but on the occasion of rail track disruptions, it would lead to network segmentation.

The contribution of this work is that it extends the concept of extracting passenger flow predictors from network characteristics by making a clear distinction between total passenger flows and OD flows. To the authors' knowledge, although the association of the former with metro and substitute network centralities was already explored by Kopsidas et al. (2023), it is the first time that metro/substitute network-based predictors of the latter are explored in this paper. The structure of this paper is as follows: the methodology is given in the next section. The results emerging from the application of the methodology to a real-world metro network (the one in Athens, Greece) are subsequently presented, followed by a discussion on them. Last, the conclusions of the study are offered in the final section.

Methodology

Overview

The factors associated with passenger flows are selected among network centralities within the metro and substitute networks. The graph of the metro network G is an unweighted, directed, L-space representation of the metro infrastructure, with stations as nodes and connections between any two consecutive stations of the same line as edges. The graph of the substitute network G_s, as defined by Kopsidas and Kepaptsoglou (2022), is a weighted, directed P-space representation of the alternative options (usually bus routes) connecting any two metro stations outside the metro system. The nodes of the substitute network also correspond to stations, while the edges are alternative public transport travel options between them, weighted by their performance in terms of time. Alternative option performance is estimated as the reciprocal of the difference between the trip duration of the best alternative option and the duration needed for the same trip within the metro system. The idea behind exploring the correlation between centralities within the substitute network and passenger flows is based on the assumption of a two-way causality vector: (i) more passengers can reach a metro station to make a transfer between bus and metro when more efficient alternative bus services reach this station; (ii) more alternative services departing from a metro station can be an incentive for passengers to use the metro as a first mode for reaching their destination. On both occasions, passenger flows at metro stations are expected to increase.

Centrality measures

The centrality measures used in this study are unweighted degree, closeness, and betweenness, as well as weighted degree (also called strength) and weighted betweenness. The degree of a node i in unweighted graphs measures the number of edges that i belongs to, and it is calculated by Eq. (1):

$$C_i^D = \sum_j^n e_{ij} \tag{1}$$

where e_{ij} denotes the edge formed by the nodes i and j, and n is the total number of nodes in the network. As degree is highly influenced by the size of the network, it is usually normalized by dividing by $n - 1$. In directed graphs, node degree is also equal to the sum of node indegree and outdegree. In weighted graphs, weighted degree or strength is calculated as the sum of the weights of the edges i belongs to, and it is given by Eq. (2):

$$S_i = \sum_j^n w_{ij} \tag{2}$$

where w_{ij} denotes the weight of an edge e_{ij}.

Closeness centrality indicates the proximity of a node to all the other nodes of a network, and it is calculated by Eq. (3):

$$C_i^C = \frac{n-1}{\sum_{i \neq j \in N} d_{ij}} \tag{3}$$

where d_{ij} is the distance from any other node j to node i.

In addition, betweenness centrality expresses the proportion of the total shortest paths of the network that pass through a node i, and it is given by Eq. (4):

$$C_i^B = \sum_{s \neq i \neq t \in N} \frac{\sigma_{st}^i}{\sigma_{st}} \tag{4}$$

where σ_{st} is the number of shortest paths between any nodes s and t, and σ_{st}^i is the number of the shortest paths passing through node i. In weighted graphs, the calculation of shortest paths takes into account the weight and not the number of network edges.

Models

Two linear regression models are developed to explore the potential of network centralities as predictors of total passenger flows and OD flows, respectively. It is noted that in both cases, the share of each station in the total flows is considered, not the respective absolute values. That is, the passenger flows of a station are divided by the sum of the respective flows of all stations. This way, the proposed methodology can be applied to metro systems with different exogenous ridership determinants (e.g., population) by multiplying the estimated shares with the total passenger flows of each system. Henceforth, passenger flows will correspond to the share of a station with respect to total passenger flows.

The dependent variable of the first model is total passenger flows (TPF), that is, the sum of inflows and outflows of a metro station. All boardings, alightings, and pass-throughs are incorporated into this measure. The degree, betweenness and closeness of the nodes in the metro network are used as independent variables, along with the strength of the nodes in the substitute network and a dummy variable I of station importance.

The model specification is presented in Eq. (5):

$$TPF_i = \beta_0 + \beta_1 C_i^D + \beta_2 C_i^B + \beta_3 C_i^C + \beta_4 S_i + \beta_5 I_i + \varepsilon_i, \quad i = 1, \ldots, n \tag{5}$$

where TPF_i is the total passenger flows of station *i*, C_i^D, C_i^B, and C_i^C are the degree, betweenness, and closeness centralities, respectively, of station i within the metro network, S_i is the strength of the same station within the substitute network, I_i denotes station importance, ε_i is the error of the estimation i, β_0 is the intercept of the model, and β_{1-5} are the coefficients of the independent variables.

When it comes to the OD flow model, the sum of only boardings and alightings constitutes the model's dependent variable (ODF). Node strength and weighted betweenness centrality within the substitute network, along with the dummy variable of importance, are used as covariates. The specification of the second model is presented in Eq. (6):

$$ODF_i = \beta'_0 + \beta'_1 S_i + \beta'_2 C_i^{WB} + \beta'_3 I_i + \varepsilon'_i, \quad i = 1, \ldots, n \tag{6}$$

where ODF_i is the OD flows of station *i*, S_i and C_i^{WB} denote the strength and weighted betweenness centrality, respectively, of station i within the substitute network, I_i represents the station's importance, β'_{0-3} are the intercept and covariate coefficients of the second model, and ε'_i is the estimation error.

It is noted that a different variation of the substitute network is used for calculating node weighted betweenness centrality because, when calculating shortest paths, the higher the weight of an edge, the longer the path. As such, options of higher performance would be misleadingly related to longer paths. For this reason, the edge weights within the substitute network denote travel times (or inverse route performance) for the calculation of weighted betweenness. However, the conventional definition is used for the calculation of node strength.

Finally, on top of regression models, the same independent variables are fed to a machine learning model based on the Extreme Gradient Boosting algorithm (XGBoost) to investigate possible non-linearities and evaluate the relative performance of the linear models.

Evaluation metrics

The accuracy of the models is evaluated through appropriate statistical metrics, including mean absolute error (MAE), mean absolute percentage error (MAPE), root mean square error (RMSE), and normalized root mean square error (NRMSE). The models are developed in training sets, consisting of 75% of the initial datasets. The evaluation process takes place in test sets comprising 25% of the total data. The metrics are analytically given in Eqs. (7–10).

$$MAE = \sum_{i=1}^{n} \frac{|\widehat{y_i} - y_i|}{n} \tag{7}$$

$$MAPE = \sum_{i=1}^{n} \frac{\frac{|\widehat{y_i} - y_i|}{y_i}}{n} \tag{8}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{|\widehat{y_i} - y_i|^2}{n}} \tag{9}$$

$$NRMSE = \frac{RMSE}{\overline{y}} \tag{10}$$

where y_i denotes the observed passenger flows at station i, $\widehat{y_i}$ the predicted flows, and $\overline{y}$ the average of the observed values.

Results

Application

To test the applicability and performance of the above-described methodology, the Athens, Greece, metro system is used as a testbed. The Athens metro system (Fig. 1) consists of three lines that intersect at key transfer stations such as Syntagma, Omonia, Monastiraki, and Attiki, enabling easy transfers between them. At the time of the analysis, the network consisted of 60 operational stations, serving thousands of daily commuters and tourists, and offering access to archaeological sites, universities, and commercial areas. Stations like Syntagma and Monastiraki, located at the heart of the city and near major attractions such as the Acropolis, experience increased passenger activity. In general, the metro system serves as a backbone for the city's public transportation network since it is integrated with other public transport modes (e.g., buses, tram), and its stations often serve as transit hubs, promoting easy transfers between different modes of transportation. Furthermore, the system is operational from early morning until late evening with

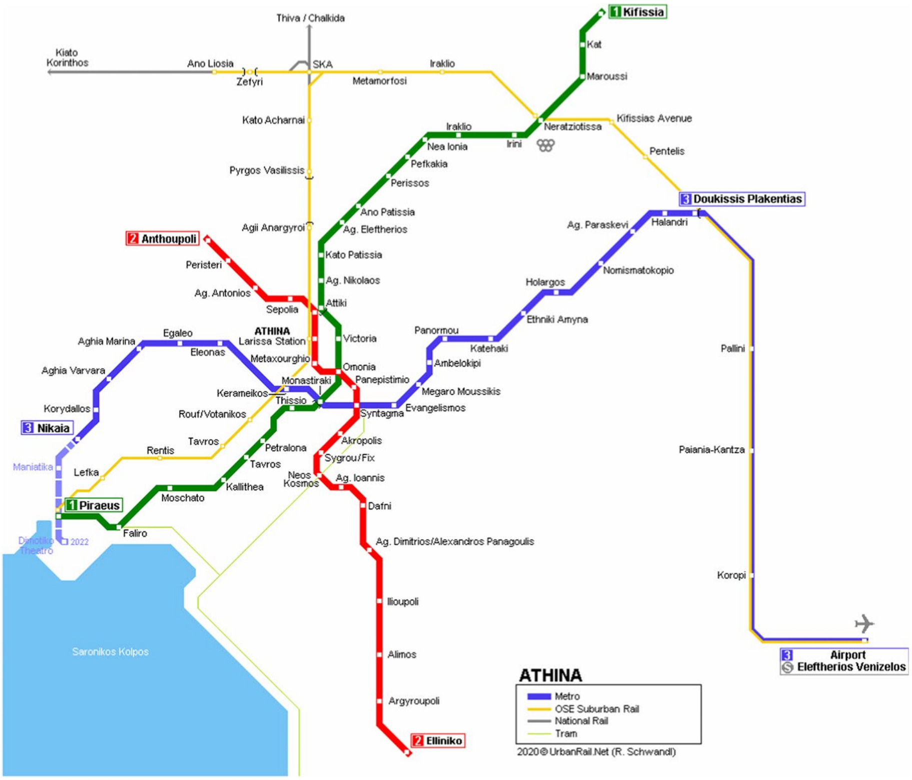

Fig. 1 The Athens metro system (www.urbanrail.net/eu/gr/athens/athens.htm. Accessed 14 July 2022)

headways of about 5–10 min, while higher frequencies reaching one train per 3 min are encountered during peak hours.

Fare collection data are used for the estimation of total daily passenger flows and OD flows, while travel times between stations are extracted from Google Maps. The data on passenger flows and the performance of the alternative (substitute) public transport options correspond to a working day, representative of the system's activity. The graph of the Athens metro network G, consisting of 60 nodes and 122 edges, and the graph of the substitute network G_s, consisting of 60 nodes and 649 edges, are presented in Fig. 2. Moreover, trip data for the Athens metro system are presented in Table 1 and Fig. 3. In Table 1, both inter-line and intra-line daily ridership is reported, while Fig. 3 illustrates the five stations with the largest total passenger flows and OD flows.

Total flow model

The first linear regression model, presented in Table 2, explores the association between total passenger flows and different centrality measures within both networks. A positive correlation between all independent variables and total passenger flows is observed, while the level of statistical significance of the covariates is at least 0.05 (also 0.01 in most cases). In addition, the standardized coefficients suggest that betweenness centrality in the metro network and strength in the substitute network are the most influential predictors of total passenger flows. According to the model, stations of higher degree, betweenness and closeness in the network of metro infrastructure, higher strength in

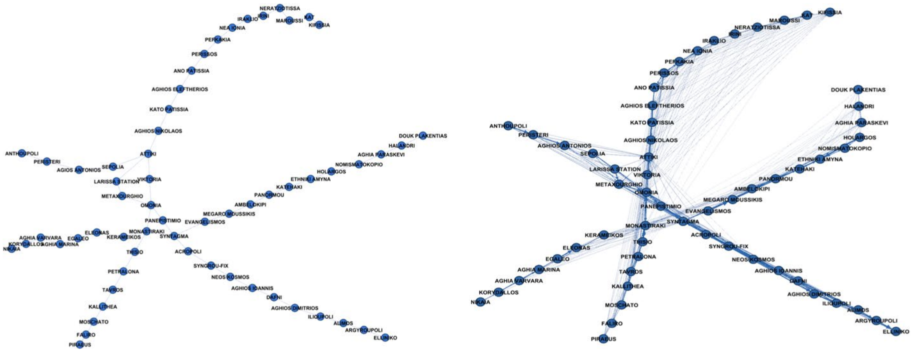

Fig. 2 Graphs of the metro (left) and substitute (right) networks

Table 1 Daily ridership data for the Athens metro system

	Internal			**Transfers**			**Total**
	Line 1	**Line 2**	**Line 3**	**Lines 1–2**	**Lines 1–3**	**Lines 2–3**	
Stations	21	17	18	2	1	1	60
OD flows	99,950	97,908	89,833	28,961	34,061	52,789	403,502
Total flows	1,964,300	2,045,332	1,918,772	–	–	–	5,928,404

Table 2 Total passenger flow regression model

Variable	**Attribute**	**Coefficient (standardized)**	***p*-value**	**VIF**
Intercept		− 0.012	0.001*	
Degree	Node degree centrality within the metro network	0.247 (0.229)	0.008*	2.946
Betweenness	Node betweenness centrality within the metro network	0.027 (0.311)	0.001*	3.189
Closeness	Node closeness centrality within the metro network	0.084 (0.200)	0.014	2.630
Sub_strength	Node strength within the substitute network	0.347 (0.299)	0.000*	2.509
Important	1 if a station is important, 0 else	0.004 (0.148)	0.006*	1.128
R^2	0.911	Durbin-Watson statistic	1.912	

*0.01 level of statistical significance

the substitute network, as well as important stations, are expected to carry more passengers in terms of boardings, alightings, and pass-throughs. For instance, important stations are related to a share of total passenger flows that is 0.4% higher than that of regular stations. It is noted that a station is considered important, and the dummy variable takes the value of 1 when at least one of the following conditions is fulfilled (0 in any other case):

i. The station is a transfer station between the metro and other rail networks (tram, suburban rail, etc.).
ii. The station is in large commercial areas.
iii. The station is located near academic institutions.

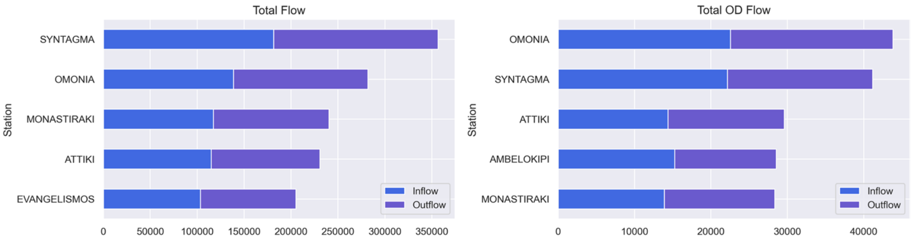

Fig. 3 The five stations with the largest total passenger flows (left) and OD flows (right) in the Athens metro system

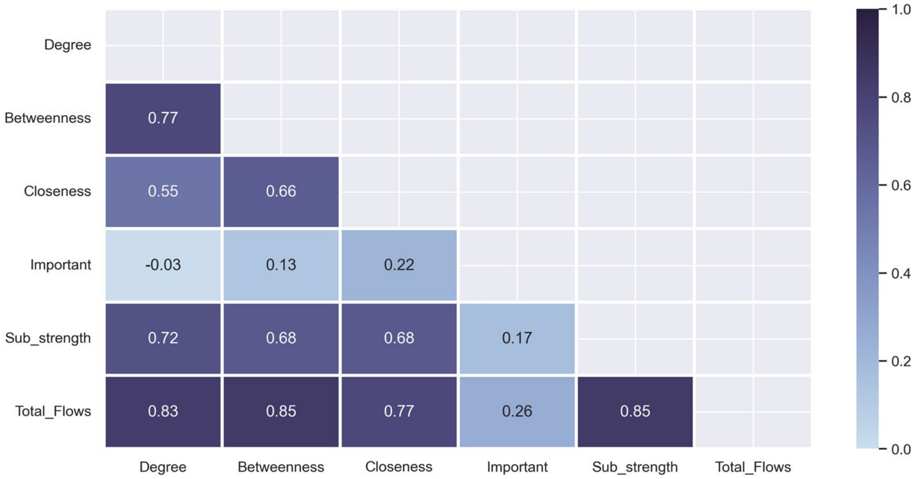

Fig. 4 Pearson correlation matrix for total flow model variables

When it comes to model fit, R^2 is equal to 0.911, indicating that a large amount of total variance is explained by the model. As for the linear model assumptions, first, there is no evidence of strong multicollinearity since variable inflation factors (VIF) are relatively low and Pearson correlation coefficients do not exceed 0.80 (Fig. 4). Moreover, the Durbin-Watson statistic (1.912) indicates that there is no strong evidence of autocorrelation (a value near 2 indicates no autocorrelation). Furthermore, the Breusch-Pagan test suggests that there is no strong evidence of heteroscedasticity (p-value $= 0.06 > 0.05$). Last, the Shapiro–Wilk test (more appropriate for small samples) shows that the errors are normally distributed (p-value $= 0.948 > 0.05$). Therefore, a linear model is indeed appropriate for the association of total passenger flows with centrality measures within the metro and substitute networks as predictors.

In addition, an XGBoost machine learning model is also developed and fed with the same explanatory variables, so that valuable comparisons can be made about feature importance and model accuracy. Tree-based models, such as XGBoost, are accepted as state-of-the-art for analyzing tabular data, yielding sound predictions with low computational cost (Grinsztajn et al. 2022). As such, the XGBoost algorithm can be utilized as a benchmark model based on which the performance of the regression model is assessed. The model's hyperparameters are set after a Grid search k-fold cross validation hyperparameter tuning (k = 5) as follows: "colsample_bytree" = 0.2, "eta" = 0.0005, "max_depth" = 1, "n_estimators" = 1100, and "subsample" = 0.05. F-scores, and SHAP values (SHapley Additive exPlanations) are used to evaluate the feature importance in

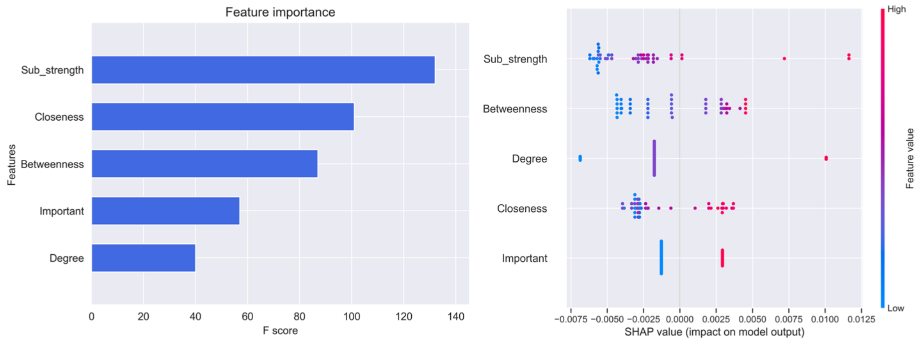

Fig. 5 F-score feature importance (left) and SHAP values (right) of the total flow XGBoost model

Table 3 Evaluation metrics for total flow models

	MAE	MAPE (%)	RMSE	NRMSE (%)
Regression	0.002422	28.18	0.003121	18.73
XGBoost	0.002469	15.74	0.003425	19.12

the XGBoost model, as depicted in Fig. 5. These values are widely used for measuring feature importance. Technically, the former count how many times a variable is used for splitting a decision tree (Chen et al. 2019), and the latter indicate the impact of the features on individual predictions (Spadon et al. 2019). Model F-scores suggest that node strength in the substitute network is the most important feature in this model, followed by closeness and betweenness in the metro network, while SHAP values suggest that node strength in the substitute network and betweenness in the metro network are the most influential factors. These findings are rather like those of the linear model since metro betweenness and substitute strength are validated as two of the most critical model components.

The evaluation metrics of the models are presented in Table 3. According to them, both models demonstrate very satisfying levels of accuracy, and therefore, they can also be used to make predictions of total passenger flows at stations. The accuracy of the linear model, albeit lower, is comparable to XGBoost's, providing evidence that the linear models are appropriate for the purpose of this study.

OD flow model

The second model explores the association between OD passenger flows and centrality measures within both metro and substitute networks. In this case, only departures and arrivals from/to metro stations are concerned, i.e., boardings and alightings. The model is presented in Table 4. All the explanatory variables are again positively correlated with the dependent variable. This means that stations with higher centrality measures are expected to be the origin or destination for more passengers than stations with lower centralities. Interestingly, only centralities within the substitute network can be significant determinants of OD flows (considering at least a 0.05 level of statistical

Table 4 OD passenger flow regression model

Variable	Attribute	Coefficient (standardized)	*p*-value	VIF
Intercept		0.008	0.004*	
Sub_strength	Node strength within the substitute network	0.330 (0.303)	0.048	2.574
Sub_ weighted_ betweenness	Node weighted betweenness centrality within the substitute network	0.057 (0.481)	0.003*	2.588
Important	1 if a station is important, 0 else	0.005 (0.195)	0.046	1.038
R^2	0.645	Durbin-Watson statistic	1.884	

*0.01 level of statistical significance

significance), along with the dummy variable of station importance, per Table 4. On the contrary, centralities within the metro network cannot stand as significant covariates, highlighting the different underlying mechanisms of this kind of passenger flows. As for feature importance, the standardized beta coefficients suggest that weighted betweenness in the substitute network is by far the most influential determinant of OD flows, followed by strength in the same network.

The fit of this model is quite lower than the previous model's ($R^2 = 0.645$), indicating that there are additional factors, other than centrality measures, that significantly contribute to OD flow generation. The assumptions of the linear model can be validated as follows: Low VIF values and acceptable Pearson correlation coefficients (Fig. 6) validate the absence of strong multicollinearity. The Durbin-Watson statistic (1.884) lies in an acceptable range (between 1.5 and 2.5 as a rule of thumb). There is no statistically significant evidence of heteroscedasticity based on the Breusch-Pagan test (p-value $= 0.107 > 0.05$). Last, the errors are normally distributed according to the Shapiro–Wilk test (p-value $= 0.117 > 0.05$). It can thus be inferred that a linear model associating network centralities with OD passenger flows can be valid, but its performance in terms of the amount of total variance explained is relatively low due to the existence of other significant determinants (probably non-network-based) omitted by the model.

The corresponding XGBoost model's hyperparameters are: "colsample_bytree" = 0.2, "eta" = 0.0005, "max_depth" = 1, "n_estimators" = 150, and "subsample" = 0.095. F-scores and SHAP values are again used to evaluate feature importance (Fig. 7). Both F-scores and SHAP values suggest that the weighted betweenness centrality of the stations in the substitute network is the most important feature, followed by strength and station importance. The results are similar to those of the regression, and therefore, there is evidence that centrality measures within the substitute network are significant explanatory variables of OD passenger flows.

Finally, the evaluation metrics for the OD flow models are presented in Table 5; both the linear and XGBoost models demonstrate decent accuracy but are not as satisfying as the ones of the total flow model. The XGBoost seems to provide more accurate predictions, but the linear model's accuracy is not far from it. Both models can thus be used to associate centralities within the substitute network with OD flows, but it would not be safe to use those models for predictions.

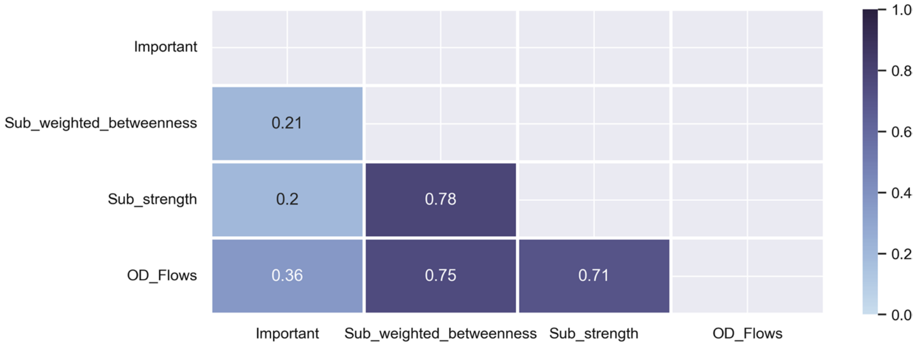

Fig. 6 Pearson correlation matrix for OD flow model variables

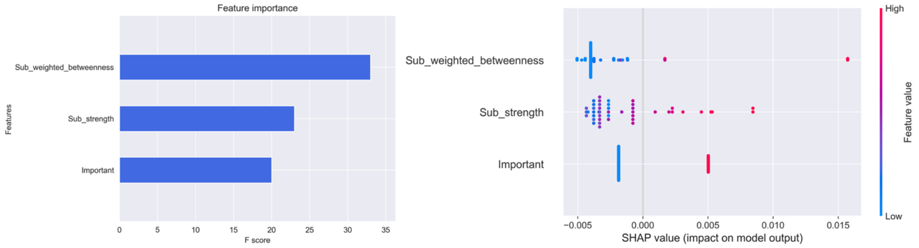

Fig. 7 F-score feature importance (left) and SHAP values (right) of the OD flow XGBoost model

Table 5 Evaluation metrics for OD models

	MAE	MAPE (%)	RMSE	NRMSE (%)
Regression	0.004415	55.82	0.005417	42.38
XGBoost	0.004446	45.21	0.005284	34.81

Discussion

The models developed for associating passenger flows with the centralities within the metro network (network of infrastructure) and the substitute network (alternative public transport options) validate the existence of high correlations among those measures and highlight the potential of network-based predictive models, as first suggested by Luo et al. (2020). Although causation is not implied, the existence of correlation can be valuable regarding fast, cost-efficient, and reliable predictions during the operation of a metro system. In particular, measures such as node degree, betweenness and closeness within the metro network, and node strength in the substitute network can be appropriate predictors of total passenger flows at stations (all boardings, alightings, and pass-throughs). Moreover, node centralities within the substitute network, such as strength and weighted betweenness, can be used as predictors of origin–destination passenger flows (only boardings and alightings). Interestingly, node centralities within the metro network do not seem to be adequate predictors of OD flows. This finding is in line with the literature suggesting that conventional centralities, such as betweenness, are not appropriate predictors of traffic flows (Kazerani and Winter 2009a, b; Gao et al. 2013; Ye et al. 2016). As such, further complex network formulations need to be considered when searching for centralities that can be used as OD flow predictors.

However, the results emerging from model application suggest that the case is not the same for both types of passenger flows when it comes to model performance. The variance of total passenger flows among metro stations can be explained to a great extent by centralities within metro and substitute networks. At the same time, the evaluation metrics for model predictive accuracy are rather satisfying and thus indicate that network-based models can be used for predictions of total passenger flows. On the other hand, the amount of variance of OD flows explained by the proposed centrality-based model is significantly less, and the evaluation metrics are not good enough to support a satisfying predictive accuracy. Evidently, there are additional significant factors that affect the distribution of OD flows rather than network structure. Demographic characteristics, such as place of residence, work area, place of education, etc. are essential factors affecting travelers' origin and destination at a micro-level, as well as population and building density, land use, etc. affect total departures and arrivals from/to metro stations at a macro-level. For instance, He et al. (2019) suggested that, except for network structure, land use, socioeconomics, and intermodal transport accessibility are also significant determinants of metro ridership. As such, centrality measures cannot be solely used for OD flow predictions at this point, but they must be combined with other appropriate socio-economic variables instead.

But what are the most appropriate centralities to begin with? The findings of this study suggest that node strength and weighted betweenness centrality within the substitute network can be the most appropriate predictors of OD flows among centralities within the metro and substitute networks. In fact, the network of alternative public transport options, such as bus routes, can provide valuable insight about the volumes of departures and arrivals at metro stations due to a reverse engineering association. Since bus route design, in terms of frequency, capacity, and coverage, has already incorporated determinants of travel demand, the substitute network, which accounts for alternative route performance, succeeds in capturing information about metro OD flows related to the same socio-economic determinants. This finding is in line with the literature, highlighting the superiority of modified centrality measures over conventional ones (Ye et al. 2016; Senousi et al. 2022). The same is also supported by the fact that it is weighted and not conventional betweenness, which is significantly correlated with OD flows.

On the contrary, the findings suggest that total passenger flows can be effectively described by centralities within metro and substitute networks. This measure is different because it also includes pass-throughs among metro stations. For a better comprehension of why these are highly correlated with network structure, one can imagine total passenger flows in a metro network as the equivalent of liquids in a system of tubes. It is reasonable that those flows are influenced by the exact structure of the network itself, just as the flows of liquids are influenced by the structure of tubes. Hence, the strong association between centrality measures within the metro network and total passenger flows can be justified in the same way. As for the node strength within the substitute network, not only is it related to total passenger flows, but it is also the most important predictor among the centralities. A similar reverse-engineering justification, like for OD flows, can be proposed to explain this association.

In this study, more complex machine learning models (XGBoost) are also developed along with regression models to evaluate the relative performance of the latter through a constructive comparison. The results suggest that both models can be used for this reason as well as for making predictions when appropriate. Although the accuracy of the XGBoost models is higher, the difference is not big enough to exclude statistical models, which may be more appropriate for small datasets and more convenient for researchers. In fact, when small datasets are concerned, more complex machine learning models cannot unfold their full potential, and thus, statistical models can be equally reliable. However, the proposed methodology can also be scalable through more complex machine learning techniques.

Total passenger flows and OD flows are treated separately in this study. According to the results, different explanatory variables are significant for each type, and different model fits are encountered. This suggests that there are different mechanisms behind the birth of each flow type. From a disruption management perspective, there are also distinct implications attached to each flow type. On the one hand, rail track disruptions at/near metro stations would create network segmentation. That is, both the upstream and downstream passenger flows would be entirely disrupted and each of the new segments could not be reached by the other. Practically, a trip would be violently terminated at the point of disruption. Evidently, all metro passengers at the disrupted station would be affected by this situation since they could not board, alight, or pass through that station. On the other hand, station platform disruptions would affect only the exact station, that is, only a node of the network, but the rest of the network would remain unharmed. On such occasions, only passengers who would be willing to board or alight at this station would be affected. The passengers passing through the station would continue their trip freely since the rail would operate normally. Total passenger flows would be affected in the first case, but only OD passenger flows would be affected in the second case. As such, the determinants of each flow type must be researched separately so that predictions can be customized depending on the needs of the operator.

The policy implications of this study mainly focus on ridership estimations in cases of disruption. Through network theory, valuable information about metro system operation can be effectively captured. CNT-based models can be much faster and more economical for making reliable ridership estimations. Public transport operators are expected to be supported by such models during daily operations management, especially when disruptions occur. In cases of disruptions, ridership estimates are essential for assessing the potential impacts of them, as well as for designing, sizing, and budgeting mitigation and contingency plans. For instance, platform disruptions at metro stations could be addressed by enhancing the capacity of bus networks serving the nearest operational stations, or rail track disruptions could be faced by bus-bridging services connecting the segmented parts of the metro network. The different possibilities for handling disruptions highlight the importance of treating different flow types separately.

As far as study limitations are concerned, the passenger flows considered in the analysis are daily, representative of a workday. Their static nature constitutes a limitation of this work since time-dependent dynamic data on passenger flows would further enhance the analysis and be expected to provide more valuable insights. The size of the available dataset is also a limitation, since larger datasets (corresponding to longer time periods) would provide safer conclusions. Last, daily data are appropriate for medium-term estimations, for example, when disruptions of 1–3 days are concerned. Narrowing down the temporal horizon to smaller timeslots would increase the applicability of the method, including short-term estimations as well.

Conclusions

In this study, network theory-based determinants of metro ridership are explored among centralities within the network of infrastructure (metro network) and the network of alternative public transport options (substitute network). For this purpose, both linear regression and non-parametric machine learning models are developed. Regarding metro ridership, total passenger flows and origin–destination flows are treated separately since they exhibit distinct characteristics and have different implications. According to modeling results, a centrality-based model can be rather accurate for total passenger flow prediction, but for OD flows, the accuracy is limited. Node degree, betweenness and closeness centralities within the metro network, node strength within the substitute network, as well as a dummy variable of station importance, are significant explanatory variables of total passenger flows. On the other hand, node strength and weighted betweenness centrality within the substitute network, along with station importance, are significant covariates regarding the OD flow models. Metrics of feature importance indicate that node strength and weighted betweenness centralities in the substitute network are the most influential predictors of total and OD passenger flows, respectively. Furthermore, according to the evaluation metrics, the performance of regression models can be like the XGBoost models for passenger flow estimation in metro systems. The fact that only centralities within the substitute network are significant OD flow predictors highlights the importance of utilizing the concept of the substitute network in similar modeling efforts. In addition, the fact that different variables are found to be significant for each passenger flow type indicates different patterns and validates the need for separate handling. The findings of this study can facilitate public transport operators when fast, cost-efficient, and reliable ridership estimations in metro systems are needed, especially in cases of disruptions. The proposed methodology is optimized for ridership estimations regarding both rail track and platform disruptions. For future research, the association between public transport ridership and additional network theory measures should be explored to discover the most appropriate network-based ridership predictors. Network-based predictive models are still at an early stage, and thus, further research in the field is necessary. Dynamic travel demand data, as well as data from different metro systems, should also be utilized for the calibration and enhancement of the proposed methodology.

Abbreviations

CNT	Complex network theory
OD	Origin–destination
TPF	Total passenger flows
ODF	Origin–destination passenger flows
XGBoost	Extreme Gradient Boosting
MAE	Mean absolute error
MAPE	Mean absolute percentage error
RMSE	Root mean square error
NRMSE	Normalized root mean square error
VIF	Variable inflation factors
SHAP	SHapley Additive exPlanations

Acknowledgements
Not applicable.

Author contributions

AK contributed to the conception and design of this work, data analysis, visualization and interpretation, and writing of the manuscript. AD contributed to the data analysis, visualization and interpretation. KK contributed to the data acquisition, general supervising and revision of the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the Basic Research Program, PEVE 2021, National Technical University of Athens.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 24 February 2023 Accepted: 16 August 2023
Published online: 22 September 2023

References

Agryzkov T, Tortosa L, Vicent JF (2019) A variant of the current flow betweenness centrality and its application in urban networks. Appl Math Comput 347:600–615

Chan EYC, Cooper CH (2019) Using road class as a replacement for predicted motorized traffic flow in spatial network models of cycling. Sci Rep 9(1):1–12

Chen M, Liu Q, Chen S, Liu Y, Zhang CH, Liu R (2019) XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system. IEEE Access 7:13149–13158

Cogoni M, Busonera G, Versaci F (2023) Estimating peak-hour urban traffic congestion. In: International conference on complex networks and their applications, Springer, Cham, pp 541–552

Cooper CH (2017) Using spatial network analysis to model pedal cycle flows, risk and mode choice. J Transp Geogr 58:157–165

Cooper CH, Harvey I, Orford S, Chiaradia AJ (2021) Using multiple hybrid spatial design network analysis to predict longitudinal effect of a major city centre redevelopment on pedestrian flows. Transportation 48(2):643–672

Dai T, Ding T, Liu Q, Liu B (2022) Node centrality comparison between bus line and passenger flow networks in Beijing. Sustainability 14(22):15454

Galafassi C, Bazzan AL (2013) Analysis of traffic behavior in regular grid and real world networks. In: Proceedings of the fifth international workshop on emergent intelligence on networked agents (WEIN)

Gao S, Wang Y, Gao Y, Liu Y (2013) Understanding urban traffic-flow characteristics: a rethinking of betweenness centrality. Environ Plann B Plann Des 40(1):135–153

Gong Y, Li Z, Zhang J, Liu W, Yi J (2020) Potential passenger flow prediction: a novel study for urban transportation development. In: Proceedings of the AAAI conference on artificial intelligence, Vol. 34, No. 04, pp 4020–4027

Grinsztajn L, Oyallon E, Varoquaux G (2022) Why do tree-based models still outperform deep learning on typical tabular data? Adv Neural Inf Process Syst 35:507–520

Guo J, Xie Z, Qin Y, Jia L, Wang Y (2019) Short-term abnormal passenger flow prediction based on the fusion of SVR and LSTM. IEEE Access 7:42946–42955

Han Y, Peng T, Wang C, Zhang Z, Chen G (2021) A hybrid glm model for predicting citywide spatio-temporal metro passenger flow. ISPRS Int J Geo-Inf 10(4):222

He Y, Zhao Y, Tsui KL (2019) Geographically modeling and understanding factors influencing transit ridership: an empirical study of Shenzhen metro. Appl Sci 9(20):4217

Henry E, Bonnetain L, Furno A, El Faouzi NE, Zimeo E (2019) Spatio-temporal correlations of betweenness centrality and traffic metrics. In: 2019 6th international conference on models and technologies for intelligent transportation systems (MT-ITS), IEEE, pp 1–10

Hochmair HH, Bardin E, Ahmouda A (2019) Estimating bicycle trip volume for Miami-dade county from Strava tracking data. j Transp Geogr 75:58–69

Jayasinghe A, Sano K (2017) Estimation of annual average daily traffic on road segments: network centrality-based method for metropolitan areas. In: Transportation research board annual meeting compendium of papers, No. 17-03141

Jayasinghe A, Munshi T (2014) 'Centrality measures' as a tool to identify the transit demand at public transit stops; a case of Ahmedabad city, India. Int J 2(7):1063–1074

Kazerani A, Winter, S (2009a) Modified betweenness centrality for predicting traffic flow. In: Proceedings of the 10th international conference on geocomputation, Sydney, Australia, November 30–December, Vol. 2

Kazerani A, Winter S (2009b) Can betweenness centrality explain traffic flow. In: 12th AGILE international conference on geographic information science, pp 1–9

Kopsidas A, Douvaras A, Kepaptsoglou K (2023) Extracting metro passenger flow predictors from network's complex characteristics. In: International conference on complex networks and their applications, Springer, Cham, pp 529–540

Kopsidas A, Kepaptsoglou K (2022) Identification of critical stations in a metro system: a substitute complex network analysis. Physica A 596:127123

Leung IX, Chan SY, Hui P, Lio P (2011) Intra-city urban network and traffic flow analysis from GPS mobility trace. arXiv preprint http://arxiv.org/abs/1105.5839
Li H, Wang Y, Xu X, Qin L, Zhang H (2019) Short-term passenger flow pre-diction under passenger flow control using a dynamic radial basis function network. Appl Soft Comput 83:105620
Lin J, Ban Y (2013) Complex network topology of transportation systems. Transp Rev 33(6):658–685
Liu D, Rong W, Zhang J, Ge YE (2022) Exploring the nonlinear effects of built environment on bus-transfer ridership: take shanghai as an example. Appl Sci 12(11):5755
Lowry M (2014) Spatial interpolation of traffic counts based on origin–destination centrality. J Transp Geogr 36:98–105
Luo D, Cats O, van Lint H (2020) Can passenger flow distribution be estimated solely based on network properties in public transport systems? Transportation 47(6):2757–2776
Omer I, Jiang B (2015) Can cognitive inferences be made from aggregate traffic flow data? Comput Environ Urban Syst 54:219–229
Ou J, Sun J, Zhu Y, Jin H, Liu Y, Zhang F, Wang X (2020) STP-TrellisNets: Spatial-temporal parallel TrellisNets for metro station passenger flow pre-diction. In: Proceedings of the 29th ACM international conference on information and knowledge management, pp 1185–1194
Pun L, Zhao P, Liu X (2019) A multiple regression approach for traffic flow estimation. IEEE Access 7:35998–36009
Puzis R, Altshuler Y, Elovici Y, Bekhor S, Shiftan Y, Pentland A (2013) Augmented betweenness centrality for environmentally aware traffic monitoring in transportation networks. J Intell Transp Syst 17(1):91–105
Senousi AM, Liu X, Zhang J, Huang J, Shi W (2022) An empirical analysis of public transit networks using smart card data in Beijing, China. Geocarto Int 37(4):1203–1223
Sevtsuk A (2021) Estimating pedestrian flows on street networks: revisiting the betweenness index. J Am Plann Assoc 87(4):512–526
Spadon G, de Carvalho AC, Rodrigues-Jr JF, Alves LG (2019) Reconstructing commuters network using machine learning and urban indicators. Sci Rep 9(1):1–13
Sun Y, Leng B, Guan W (2015) A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system. Neurocomputing 166:109–121
Toqué F, Khouadjia M, Come E, Trepanier M, Oukhellou L (2017) Short & long term forecasting of multimodal transport passenger flows with machine learning methods. In: 2017 IEEE 20th international conference on intelligent transportation systems (ITSC), IEEE, pp 560–566
Wang J, Li Y, Jiao J, Jin H, Du F (2022a) Bus ridership and its determinants in Beijing: a spatial econometric perspective. Transportation 50(2):383–406
Wang K, Wang P, Huang Z, Ling X, Zhang F, Chen A (2022b) A two-step model for predicting travel demand in expanding subways. IEEE Trans Intell Transp Syst 23(10):19534–19543
Yang X, Xue Q, Yang X, Yin H, Qu Y, Li X, Wu J (2021) A novel prediction model for the inbound passenger flow of urban rail transit. Inf Sciences 566:347–363
Ye P, Wu B, Fan W (2016) Modified betweenness-based measure for prediction of traffic flow on urban roads. Transp Res Rec 2563(1):144–150
Zhang X, Chen M (2020) Enhancing statewide annual average daily traffic estimation with ubiquitous probe vehicle data. Transp Res Rec 2674(9):649–660
Zhang M, Huang T, Guo Z, He Z (2022) Complex-network-based traffic network analysis and dynamics: a comprehensive review. Phys a: Stat Mech Appl. https://doi.org/10.1016/j.physa.2022.128063
Zhao PX, Zhao SM (2016) Understanding urban traffic flow characteristics from the network centrality perspective at different granularities. Int Arch Photogramm Rem Sens Spat Inf Sci 41:263–268
Zhao S, Zhao P, Cui Y (2017) A network centrality measure framework for analyzing urban traffic flow: a case study of Wuhan, China. Physica A 478:143–157
Zheng Z, Ling X, Wang P, Xiao J, Zhang F (2020) Hybrid model for predicting anomalous large passenger flow in urban metros. IET Intel Transp Syst 14(14):1987–1996

Publisher's Note

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

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com