TRANSPORT PROBLEMS 2026 Volume 21 Issue 1
PROBLEMY TRANSPORTU DOI: 10.20858/tp.2026.21.1.18

Keywords: daily station ridership; peak-hour demand; station-level passenger flow; peak-hour asymmetry; demand modeling; ordinary least squares; random Forest

Ihor TARAN[1]*, Mateusz SZARATA[2], Kamil KOWALSKI[3]

THE RELATIONSHIP BETWEEN PEAK-HOUR PASSENGER DEMAND AND DAILY STATION RIDERSHIP: A CASE STUDY OF THE RZESZÓW AGGLOMERATION

Summary. Passenger rail transport is a key component of regional mobility systems, where the reliable assessment of station-level passenger volumes is essential for timetable planning and service evaluation. This study analyses the relationship between peak-hour passenger demand and daily station ridership using empirical data from the regional railway network of the Podkarpackie Voivodeship in Poland. The analysis is based on field measurements of passenger boardings and alightings collected during a month-long campaign covering the entire regional railway network. For each station, total daily passenger exchange as well as morning and evening peak-hour demand were recorded. The study examines whether peak-hour demand levels and their asymmetry can serve as a practical proxy for daily station ridership. Two modeling approaches were applied: multivariate linear regression estimated by ordinary least squares and a random forest model to capture potential non-linear effects. The results show that daily station ridership is primarily associated with morning peak-hour demand, while evening peak demand plays a secondary but stable role. Greater peak asymmetry is linked to lower daily passenger exchange, whereas territorial affiliation has only a marginal influence once peak demand is considered. The findings demonstrate that peak-hour measurements provide a data-efficient and practically applicable basis for approximately estimating daily station ridership, thus supporting timetable planning, service adjustments, and preliminary evaluations of demand changes in regional railway systems.

1. INTRODUCTION

Rail passenger transport is one of the key components of regional transport systems, particularly in areas with a clearly polycentric settlement structure. Under such conditions, the effectiveness of network operation is determined not only by infrastructure parameters and the level of service but also by the characteristics of passenger demand, which develops at individual stations and along specific sections of the network.

One of the most representative indicators of the performance of a passenger rail system is station passenger flow, which is the total number of passengers who board and alight the system over the course of a day. This indicator reflects both the overall level of rail use and the daily structure of demand, in which – as practice shows – the morning and afternoon peak hours play a key role, as they are associated with everyday commuting to workplaces and educational institutions.

In the Podkarpackie Voivodeship, the analysis of passenger flows is further complicated by the pronounced spatial and functional heterogeneity of the network. The region is characterized by the

[1] Rzeszów University of Technology, Department of Road and Bridges; Powstanców Warszawy Ave. 12, 35-029 Rzeszów, Poland: e-mail: i.taran@prz.edu.pl; orcid.org/0000-0002-3679-2519
[2] Rzeszów University of Technology, Department of Road and Bridges; Powstanców Warszawy Ave. 12, 35-029 Rzeszów, Poland: e-mail: matsza@prz.edu.pl; orcid.org/0000-0003-0227-2811
[3] Rzeszów University of Technology, Department of Road and Bridges; Powstanców Warszawy Ave. 12, 35-029 Rzeszów, Poland: e-mail: k.kowalski@prz.edu.pl; orcid.org/0000-0002-1653-7756
* Corresponding author. e-mail: i.taran@prz.edu.pl

presence of several strong urban centers (in particular, Rzeszów, Przemyśl, and Stalowa Wola), alongside an extensive network of medium-sized and small towns served by rail services of varying intensities. As a result, even stations located on the same railway line and served by comparable rolling stock may exhibit markedly different passenger-demand volumes and temporal profiles.

This study is based on the results of comprehensive field surveys and data from automated systems, covering passenger flows across the rail network of the Podkarpackie Voivodeship, which were collected as part of a separate research report [1]. The compiled material includes information on the numbers of passengers boarding and alighting at stations, as well as the intraday distribution of demand, which enables a description of the current pattern of network use and a transition to a more formal analysis.

The article focuses on the relationship between peak-hour passenger demand and daily passenger flow at the level of individual stations. This approach allows peak-period measurements to be treated as a compact and readily obtainable proxy for the overall demand structure, and it provides a starting point for moving from descriptive statistics to modeling.

The study addresses three interrelated research tasks: (1) assessing the relationship between a station's daily passenger flow and passenger loads during the morning and afternoon peak periods, including their asymmetry; (2) developing a model for forecasting daily passenger flow based on a minimal set of field measurements; and (3) verifying the stability of the results and analyzing the sensitivity of the conclusions to changes in peak-period loads, as well as discussing potential practical applications – in timetable planning, in evaluating the effects of changes in train service frequency, and in prioritizing development measures.

Two methodologically distinct approaches were employed to achieve the above objectives: linear regression estimated using the ordinary least squares method, as a transparent and interpretable baseline tool, and a random forest model, which allows non-linearities and interactions between factors to be taken into account.

2. LITERATURE REVIEW

Transport Research on passenger flows at individual railway and metro stations is well-established. This is reflected in a large body of literature examining the factors that shape passenger demand, its spatial and temporal patterns, and its links to infrastructure characteristics and the station's surrounding environment. More recent work has emphasized identifying determinants of station-level ridership using both classical regression models and machine-learning methods, as well as the development of direct demand modeling approaches for new and existing rail-based transport systems [2].

The foundations of this research stream lie in studies in which a station's daily passenger turnover is modeled as a function of land-use characteristics, built-up density, job accessibility, and transport network parameters. For example, Zhu et al. [3] proposed a Bayesian negative binomial regression framework to assess the factors shaping ridership at railway stations while accounting for spatio-temporal heterogeneity, showing that the effects of individual variables differ substantially by time of day and station location. Similar conclusions regarding the need to analyze time intervals separately were presented by Wu et al. [4], who treated morning and afternoon demand as distinct operating regimes of a station.

A significant part of the literature consists of studies using local regression models that allow for spatial and temporal variation in coefficients. In particular, Ma et al. [5] applied geographically and temporally weighted regression to analyze passenger demand, showing that global models tend to smooth out local effects and underestimate the importance of territorial heterogeneity. An extension of this approach was proposed by He et al. [6], who introduced adaptive local regularization to improve estimation stability under high multicollinearity among explanatory factors.

Related studies based on smart-card data and points-of-interest (POI) datasets have also shown that accounting for spatio-temporal heterogeneity improves the explanatory power of models compared with ordinary least squares (OLS) and basic geographically weighted regression approaches, and that the direction and strength of factor effects can differ substantially between boardings and alightings as well

as across time intervals [7]. In parallel, a research stream has focused on more precise definitions of distance metrics and geographic weighting (including network-distance measures), enabling a more accurate representation of station catchment areas and strengthening the identification of local effects [8]. More recent studies using multiscale geographically weighted regression further indicate that the spatial ranges of influence of individual variables (spatial scales) are also heterogeneous and depend on the temporal context (weekdays vs. weekends, peak vs. off-peak), which provides additional evidence of the limitations of single-scale global models [9, 10]. At the same time, these methods require large data volumes and strong spatio-temporal representativeness, which constrains their applicability in practical transport planning tasks.

In addition to regression models with global and local coefficients, the literature also considers panel-data approaches and multilevel models, which make it possible to separate within-station and between-station variability in passenger flows. These models are used to examine the temporal stability of demand determinants and to reduce estimation bias arising from unobserved time-invariant effects. In practice, however, applying panel models and fixed-effects approaches requires long time series and consistent, comparable measurements – conditions that are often difficult to meet in the case of limited or one-off station surveys [11]. Recent research has demonstrated that modeling and simulation frameworks can be constructed directly from sensor-based system-state data, reducing reliance on extensive exogenous variables while preserving analytical robustness [12].

As computational methods have advanced, ensemble machine-learning algorithms – particularly gradient boosting and random forest algorithms – have been increasingly applied. Related studies in railway-system analysis have shown that data-driven quantitative models based on detailed operational measurements allow for the identification of non-linear effects and system-level optimization potential, complementing traditional demand-focused analyses [13]. Studies by Gu and Dou [14] and Shao et al. [15] showed that using GBDT enables the identification of non-linear and threshold effects of land use and transport accessibility on passenger demand that are not captured by linear models. These works highlight the presence of saturation effects, whereby the marginal impact of factors decreases once a certain level of intensity or load is exceeded.

At the same time, a key limitation of most machine-learning-based approaches remains the limited interpretability of their results. In many studies, the analysis is reduced to assessing variable importance, which does not allow for a direct interpretation of the sign and functional form of each factor's effect on passenger demand. In response, more recent work has proposed interpretable ML approaches based on SHAP values and partial dependence plots. For example, Yang et al. [16] showed that the non-linear relationships of factors shaping ridership exhibit pronounced temporal variability, and that their interpretation requires sensitivity analysis of the model rather than reliance on global goodness-of-fit measures alone.

A separate issue discussed in the literature concerns the extent to which passenger-demand models can be reused across different cities, lines, and time periods. It has been shown that even when a model achieves good fit within a single metropolitan area, transferring it to another context (e.g., a different network structure, station profile, or demand regime) may degrade predictive performance and alter the relative importance of factors. This, in turn, increases the importance of model interpretability and stability for practical applications [17, 18].

Particular attention should be given to the use of peak-period loads as a proxy for daily passenger turnover. Yu et al. [19] showed that the time of maximum passenger activity at an individual station may differ substantially from the citywide peak and proposed peak-deviation indicators to support the correct interpretation of peak-hour data. Further developments of this approach suggest that accounting for the variability of peak-period factors and their asymmetry can reduce systematic bias in estimating station-level peak-hour ridership and increase the usefulness of results for station design and operations [18]. Nevertheless, most studies focus either solely on daily aggregates or analyze peak hours without explicitly modeling their relationship with daily passenger turnover.

The literature review points to several persistent methodological limitations of existing approaches. First, many models rely on extensive sets of exogenous variables, which limits their applicability when field data availability is constrained. This limitation has also been highlighted in applied transport-modeling studies, which emphasize the need for parsimonious modeling approaches capable of

supporting route and demand analysis under restricted data conditions [20]. Second, linear models, despite their high interpretability, are unable to capture saturation effects and non-linear demand responses. Third, contemporary ML models often operate as "black boxes," without in-depth sensitivity analysis or scenario-based interpretation. Finally, the relationship between peak-period loads (AM/PM) and a station's daily passenger turnover is rarely treated as a standalone modeling objective.

In this context, the present study addresses the identified gap by proposing a minimalist yet informative approach in which daily passenger turnover is modeled solely on the basis of peak-period measurements and their asymmetry. Combining linear regression with a random forest model makes it possible, on the one hand, to ensure interpretability and formal hypothesis testing, and, on the other hand, to identify non-linear effects and assess the sensitivity of daily demand to changes in peak loads. This approach complements existing research and is intended to explain not only the phenomenon but also the practical use of the models in transport planning tasks.

3. RESEARCH METHODOLOGY AND RESULTS

The basis for the modeling was the result of on-site measurements of passenger flows at railway stations in the Podkarpackie Voivodeship collected during a month-long campaign of comprehensive passenger-traffic surveys covering the entire regional rail network. In report [1], three key variables were recorded for each station:

Q_D – the overall daily passenger volume (i.e., the total number of boarding and alighting within a 24-hour period);

Q_{AM} – the number of passengers served by the station during the morning commuting peak;

Q_{PM} – the number of passengers served by the station during the afternoon commuting peak.

The morning and evening peak periods correspond to the identified one-hour intervals of maximum network load (6:15–7:15 for the morning peak and 15:15–16:15 for the afternoon peak), determined based on 15-minute passenger-flow observations. These indicators constitute the direct outcome of observations, without any aggregation or interpretation. They reflect the actual intensity of station use by the population throughout a day. In addition, each station was assigned to a specific functional urban area (FUA) (Fig. 1).

These areas represent territorial functional systems centered around medium-sized and large cities such as Mielec, Tarnobrzeg, Stalowa Wola, Dębica, Krosno, Jasło, Sanok, Przemyśl, Jarosław-Przeworsk, the Rzeszów ROF, and others. In theory, the FUA classification may capture spatial variation in mobility patterns; however, the actual effect of this factor is verified within the models.

The data were extracted from a report [1] in which each table corresponded to a single FUA and listed the stations belonging to that cluster. All tables were then merged into a single dataset, with each row representing an individual station.

Table 1 below provides a small, representative snapshot of the final database, limited to the following columns: FUA name and station name, Q_D, Q_{AM}, Q_{PM}.

Table 1

A sample of the final dataset

FUA	Station	Q_D	Q_{AM}	Q_{PM}
FUA Dębica	Dębica	1710	165	195
FUA Dębica	Dębica Wschodnia	198	20	25
FUA Jarosław-Przeworsk	Jarosław	2464	215	260
FUA Jarosław-Przeworsk	Pełkinie	184	18	23
FUA Jasło	Jasło	870	95	105

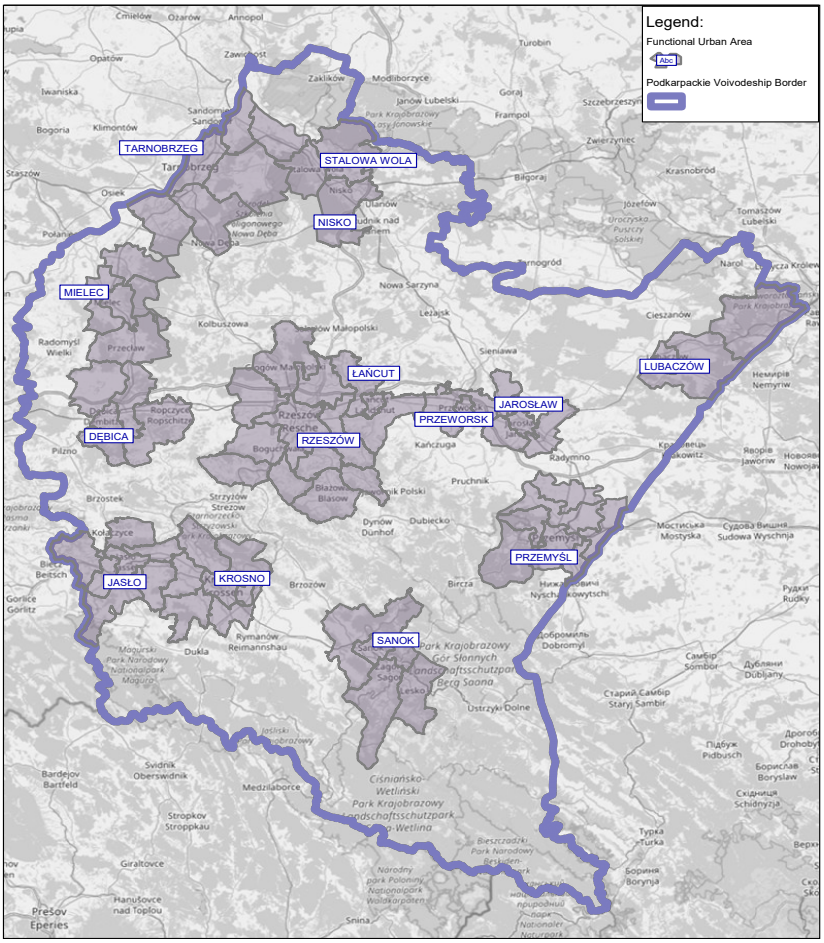

Fig. 1. FUA boundary

Examining the absolute values of Q_{AM} and Q_{PM} is not sufficient to understand how a station operates. It is also essential to consider the intra-day demand structure (i.e., the degree to which passenger loads are evenly distributed between the morning and evening peak periods). To this end, the peak-load asymmetry ratio (ρ) is introduced:

$$\rho = \frac{Q_{AM}}{Q_{PM}} \tag{1}$$

This ratio allows stations to be differentiated by their usage profile:
ρ = 1 – the station operates symmetrically, with demand evenly distributed;
ρ > 1 – a pronounced morning peak combined with a weak evening peak (predominantly one-way trips);
ρ < 1 – activity is concentrated in the evening hours.

This asymmetry ratio (ρ) provides a measure of a station's stability as a node in day-to-day mobility: balanced stations typically perform a broader range of functions and play a more significant role within the network. When calculating the asymmetry ratio (ρ), particular attention was paid to cases where Q_{PM} equaled zero. Such stations were either excluded from the regression analysis or treated as a separate category to avoid division-by-zero effects. This approach helped prevent distortions and artificial effects in the modeling results.

The input dataset includes station passenger volume during the morning and evening peak periods (Q_{AM} and Q_{PM}), the peak-load asymmetry ratio (ρ), and the categorical FUA variable describing each station's territorial affiliation. Total daily passenger volume Q_D was used as the dependent variable. The categorical FUA variable was subsequently transformed into a set of binary indicator variables (one-hot encoding), which makes it possible to account for territorial heterogeneity across stations without violating the assumptions of linear models.

As a result, the final set of variables was constructed entirely from directly measured survey data and did not require any external information. This methodological approach allows an assessment of the extent to which a station's daily passenger volume can be approximated based on peak-period loads and their asymmetry without claiming to fully explain the mechanisms underlying demand formation.

The study tests the hypothesis that passenger-volume levels recorded at stations during the morning and evening peak periods (Q_{AM}, Q_{PM}) capture a substantial share of the information about their total daily load and can therefore be used to describe and provide an approximate assessment of the variability of the Q_D indicator within a given system and study period without relying on extended sets of external factors. Although Q_{AM}, Q_{PM}, and Q_D were obtained within the same survey campaign, modeling is applied to examine the structure and stability of the relationships between peak and daily indicators, and to evaluate whether peak-period measurements can serve as a proxy for daily load in practical applications.

Naturally, the model could be extended by incorporating additional factors such as population density, train service frequency, competition from bus transport, distance to the metropolitan core, and other attributes of the station's surroundings. The aim of this article is not to build the most comprehensive model possible but to demonstrate the self-sufficiency of field-measurement data. If peak passenger-volume indicators can adequately describe and predict daily passenger volume, this would imply that they provide a robust and interpretable basis for further analyses and for supporting planning decisions. Expanding the set of variables is treated as a natural direction for future research rather than as a prerequisite for obtaining a valid and substantively coherent baseline model.

The first modeling stage applied in this study is multivariate linear regression estimated using the OLS method. Despite the rapid development of machine-learning methods, OLS remains a fundamental and widely accepted tool for the quantitative analysis of transport demand and is extensively used in station-level research.

The use of linear regression at the initial stage was motivated by several of its methodological advantages. First, it offers a high degree of transparency and interpretability. Namely, coefficient estimates make it possible to directly identify the sign and relative strength of individual effects and to assess how changes in peak-period passenger volumes and their relationship translate into the daily indicator.

A linear model provides a convenient framework for structural hypothesis testing. In particular, it enables a separate assessment of the contribution of temporal demand characteristics (the morning and evening peak periods, as well as their asymmetry) and territorial factors related to station type. It also allows for the verification of whether these effects operate independently.

Finally, OLS serves as a reference level for subsequent comparisons with more complex non-linear models. A material improvement in performance when moving to machine-learning algorithms is interpreted as evidence of non-linear relationships, saturation effects, or interactions between factors that are not captured by a linear model specification. As a result, multivariate linear regression functions as the structural backbone of the study, formalizing the key relationships and providing an interpretable basis for further analysis of results obtained using the random forest model. Within the context of this research, linear regression is used as a baseline tool to formally pose the problem of identifying which components of peak loads and which station characteristics contribute most to shaping daily passenger volume. This approach enables a shift from descriptive analysis to a quantitative testing of hypotheses regarding the role of the morning and evening peak periods, their asymmetry, and the station's territorial affiliation.

In order to achieve the stated goal, a formal linear model was introduced that links a station's daily passenger volume to a set of explanatory variables. This allows the transition from a conceptual framework to a strict mathematical formulation. The linear model is expressed in the standard form:

$$Q_{D_i} = \beta_0 + \beta_1 \cdot Q_{AM_i} + \beta_2 \cdot Q_{PM_i} + \beta_3 \cdot \rho_i + \sum_{k=1}^{K} \gamma_k \cdot FUA_{ik} + \varepsilon_i \quad (2)$$

where: Q_{D_i} – the i-th station's daily passenger volume; Q_{AM_i}, Q_{PM_i} – the morning- and evening-peak passenger volumes at station i; $\rho_i = Q_{AM_i}/Q_{PM_i}$ – asymmetry ratio; FUA_{ik} – binary indicator variables denoting whether a station belongs to the k-th FUA; β_0 – intercept; $\beta_1, \beta_2, \beta_3$ – regression coefficients

associated with Q_{AM_i}, Q_{PM_i}, and ρ_i, respectively; γ_k – estimates of systematic adjustments to daily passenger volume for stations in the k-th FUA relative to the baseline category; ε_i – error term.

Constructing the model in this way makes it possible to unambiguously separate the contribution of the temporal demand structure (Q_{AM}, Q_{PM}, ρ) from the territorial factor, represented by the FUA binary indicator variables, thereby giving the analyzed relationship a clear and interpretable structure.

The model was tested on the raw data for the set of analyzed stations. Each observation in the dataset corresponds to an individual station, and all model estimations and performance metrics were calculated across the full sample of surveyed stations. Estimation yielded a set of coefficients (Tables 2 and 3) describing the effects of peak-period loads, their asymmetry, and the station's territorial affiliation on daily passenger volume. In the tables below, the linear regression coefficients are reported in two groups. The first group contains the parameters for the quantitative variables capturing the relationship between daily passenger volume and peak-period volumes and their asymmetry. The second group contains the coefficients for the binary FUA variables, reflecting systematic differences in daily passenger volume associated with the station's functional setting.

Table 2

Estimated coefficients of the linear regression model

Variable	Coefficient	Estimate
Q_{AM}	β_1	8.1827
Q_{PM}	β_2	3.6591
ρ	β_3	-12.2217

Table 2 presents the coefficients of the linear regression model for the quantitative variables describing station peak-period loads and their asymmetry. These coefficients capture the relationship common to all stations between daily passenger volume and the indicators of the morning and evening peak periods, as well as the intra-day demand structure. The coefficient values are interpreted as the change in daily passenger volume resulting from a one-unit increase in a given variable while all other factors are held constant. The stochastic term ε represents the portion of a station's daily passenger volume that is not explained by peak-period loads or their asymmetry. It captures the effects of off-peak travel, transfer and through movements, the local characteristics of station operations, and measurement errors. Empirically, ε corresponds to the model residual, understood as the difference between the observed daily passenger volume and the value estimated from the regression equation.

Table 3

Linear regression coefficients for binary FUA variables

FUA	Coefficient
FUA Dębica	-73.7501
FUA Jarosław-Przeworsk	-77.2107
FUA Jasło	36.0933
FUA Krosno	38.6694
FUA Mielec	-5.3737
FUA Przemyśl	172.7085
FUA Sanok	12.2022
FUA Stalowa Wola	46.2859
FUA Tarnobrzeg	57.3780
FUA Rzeszów	-207.0028

Table 3 presents the coefficients for the binary FUA variables, which reflect a station's affiliation with a specific functional context. These coefficients constitute systematic adjustments to the level of daily passenger volume relative to the baseline FUA category, which is not included in the table. One FUA category was omitted from the model specification and serves as the reference (baseline) group in

order to avoid perfect multicollinearity. Their values indicate by how much, on average, the daily passenger turnover of a station located in a given functional setting differs from that of the baseline FUA, assuming identical values of peak-period loads and their asymmetry. Fig. 2 provides a visual representation of the FUA coefficient estimates reported numerically in Table 3, facilitating their visual interpretation in terms of relative magnitude and direction.

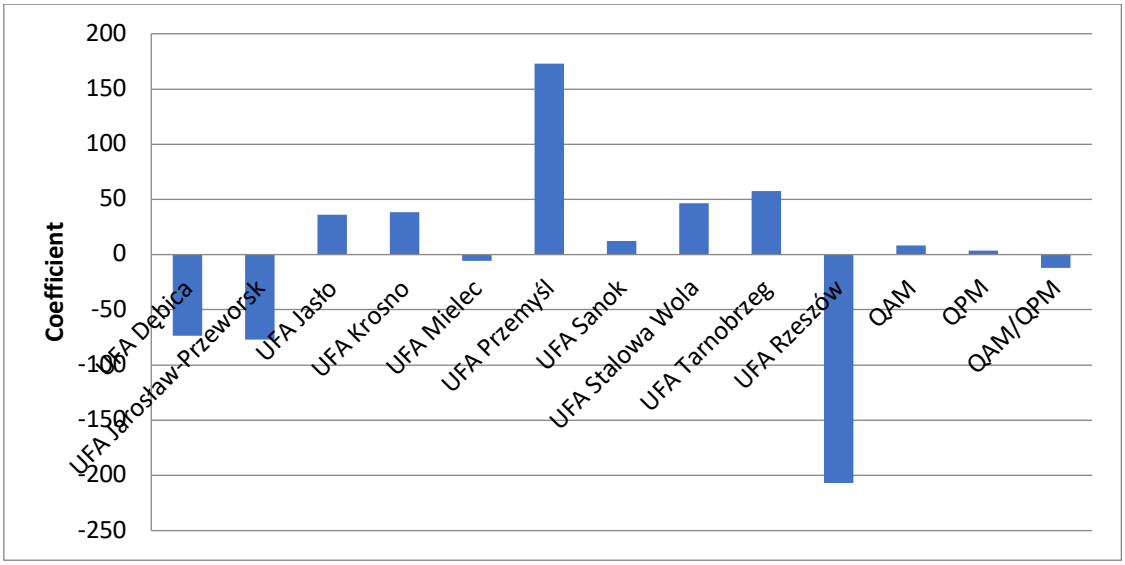

Fig. 2. Linear regression coefficients

The coefficient values of the linear regression model presented in Table 3 enable a substantive interpretation of the following identified relationships between the station's daily passenger volume and the characteristics of peak-period loads:

1. The coefficient for Q_{AM} is $\beta_1 = 8.18$. This means that an increase of one passenger in the morning passenger volume is statistically associated with an increase in the daily passenger volume, indicating the key role of the morning period in the structure of observed demand rather than a direct causal relationship.
2. The impact of the evening peak is nearly 2.3 times weaker than that of Q_{AM} ($\beta_2 = 3.66$), which can be explained by the fact that in the evening hours, passengers disperse in many directions, resulting in less concentrated evening flows.
3. Flow asymmetry reduces the level of daily demand ($\beta_3 = -12.22$). The stronger the disturbance of the ρ indicator, the lower the overall daily passenger volume. The dominance of a single time interval in the demand structure reflects a limited functional role of the station within the daily mobility system and is associated with a lower level of daily passenger volume.
4. Territorial effects (FUA) are secondary in nature. Most γ_k coefficients for the FUA variables take small values or exhibit low statistical significance. Clear effects occur only for a few FUA categories, such as Przemyśl and Rzeszów ROF, that are related to their specific role within the network.

Consequently, within the dataset used and the adopted operationalization of the territorial context, a station's daily passenger volume is primarily associated with the magnitude of the morning and evening peak periods, while being located within an FUA contributes only a limited corrective effect.

The OLS estimation results indicate that a station's daily passenger volume can be described to a considerable extent as a linear combination of peak-period loads. At the same time, the linear specification of the model does not allow for the inclusion of potential nonlinear effects and interactions between factors, which are characteristic of large stations and network hubs. For this reason, the subsequent part of the analysis employs a random forest model, enabling the robustness of the linear findings to be verified and the impact of peak-period loads on daily passenger volume to be examined in more depth.

The linear regression model made it possible to identify the basic structure of how peak-period loads affect a station's daily passenger volume and demonstrated that, in general terms, the observed relationship is close to linear. However, the linear approach assumes an equal contribution of an additional passenger in the morning or evening peak regardless of the station load level, and it does not account for possible interactions among factors.

In real transport systems, increases in peak-period loads are often associated with saturation effects. At large stations and hubs, additional passenger inflows encounter capacity constraints of the infrastructure and timetable parameters; as a result, the increase in daily passenger volume attributable to a unit increase in peak demand gradually decreases. Such behavior cannot be represented in a linear model.

To account for potential nonlinear effects and interactions among factors, a random forest model – an ensemble algorithm based on decision trees—was additionally applied in the study. To ensure comparability of results, we trained the random forest model on the same set of variables as the linear regression model. In addition to predicted values of daily passenger volume, the model also produces estimates of relative feature importance, characterizing the contribution of individual factors to reducing the prediction error.

A key advantage of the random forest model is that it does not require an a priori specification of the functional form of the relationship; the algorithm automatically identifies nonlinearities and combinations of features. In this study, the random forest model is used to verify the robustness of the conclusions obtained from the linear model and as a tool for further scenario-based sensitivity analysis of daily passenger volume with respect to changes in peak-period loads.

The variable importance values computed using the random forest model are presented in Table 4. The importance measures are normalized values produced by the algorithm; they are summed to 1 across all variables and reflect their relative contributions to prediction error reduction. In this context, feature importance refers to the relative contribution of each variable across the ensemble of decision trees. Therefore, the reported percentages do not represent shares in passenger volume or explained variance (R^2) but instead indicate the relative contribution of each predictor to the model's overall predictive performance.

Table 4

Variable importance measures in the random forest model

Variable	Importance measure
FUA Dębica	0.002684529
FUA Jarosław-Przeworsk	0.000396932
FUA Jasło	0.0000231889
FUA Krosno	0.00000507009
FUA Mielec	0.000118388
FUA Przemyśl	0.007886806
FUA Sanok	0.0000837088
FUA Stalowa Wola	0.000299678
FUA Tarnobrzeg	0.000144752
FUA Rzeszów	0.005950322
Q_{AM}	**0.560504441**
Q_{PM}	**0.412180976**
ρ	0.009721207

Based on the results presented in Table 4, several conclusions were formulated. First, the variable importance analysis in the random forest model reveals a clear and stable hierarchy of factors that, in terms of structure, is consistent with the results of the linear regression. The dominant determinant of daily passenger volume remains the morning peak (Q_{AM}), which accounts for approximately 56% of the model's total explanatory power. This means that more than half of the information on a station's daily load is captured by the magnitude of the morning peak flow. Even without assuming a linear functional

form, the model confirms the OLS findings regarding the key role of the morning period: knowledge of Q_{AM} allows the station's daily passenger volume to be reconstructed with high accuracy.

Second, the evening peak (Q_{PM}) is the second most important factor, accounting for about 41% of the total importance. Although its contribution is clearly lower than that of the morning peak, it remains a significant component of the model and reflects the characteristics of the spatiotemporal structure of return trips. Taken together, the morning and evening peaks account for more than 97% of the total feature importance, underscoring the pivotal role of peak periods in shaping daily passenger volume.

Third, the peak-load asymmetry ratio (ρ) makes a limited but stable contribution (< 1% of the total importance). Despite the small magnitude of this effect, the feature is consistently used by the model and serves as a corrective indicator, allowing differences between station types to be considered. Peak asymmetry is not an independent driver of demand; however, it improves the accuracy of station characterization in borderline cases.

Fourth, a station's territorial affiliation, represented by the set of FUA indicator variables, has a minimal effect on the model results, as their combined importance does not exceed 3%. This points to the secondary role of geographic context compared with actual peak flows and confirms that a station's daily passenger volume is not determined by functional-area membership alone but primarily by the actual scale of morning and evening activity.

Based on these conclusions, the importance structure obtained using the random forest model almost fully reproduces the logic of the linear regression coefficients, which constitutes an important methodological finding. Two fundamentally different approaches – OLS and a nonlinear ensemble model – independently confirm the same hierarchy of factors, strengthening the robustness and credibility of the conclusions. At the same time, the random forest model complements the linear model by enabling the identification of saturation effects and differences in station behavior at varying load levels, which was subsequently leveraged in the sensitivity analysis.

Developing regression and nonlinear models makes it possible to describe the current structure of passenger demand. However, this alone does not answer the key practical question in transport planning: how does the daily passenger volume of a station change as a result of an intentional modification of the service offer? In rail transport, such changes are typically associated with adjustments to peak-period operations, such as introducing or withdrawing train services, reallocating capacity, or changing service frequency.

The sensitivity analysis is based on a controlled variation of the morning peak passenger volume (Q_{AM}) while keeping the remaining station characteristics constant – namely, the evening peak (Q_{PM}), the peak-load asymmetry indicator (ρ), and territorial affiliation (FUA). This experimental design reflects a situation in which the station's infrastructure conditions and spatial context remain unchanged, whereas changes in demand during the morning peak are achieved by modifying the timetable or service offer parameters.

For each specified change in Q_{AM}, the predicted value of daily passenger volume is computed using the trained random forest model. This makes it possible to determine how daily passenger volume responds to increases or decreases in morning peak demand and to assess the presence of linear response ranges, saturation effects, and asymmetry in the system's response.

From a technical perspective, the experiment was organized as follows:

1. A representative baseline combination of features was selected – conventionally, a "typical station" or the average station profile within a given group. The vector of variables ($Q_{AM\ base}$, $Q_{PM\ base}$, ρ_{base}, FUA) was adopted as the reference point.
2. For the Q_{AM} value (parameter), a range of increments was defined ΔQ_{AM}

$$\Delta Q_{AM} \in [-50, -40, \ldots, 0, \ldots, 40, 50], \tag{3}$$

That is, scenarios were considered in which the morning peak flow either decreased or increased by several dozen passengers.

3. For each value, ΔQ_{AM}, a new value of the morning peak was calculated using the formula

$$Q_{AM\ new} = Q_{AM\ base} + \Delta Q_{AM}, \tag{4}$$

while the remaining parameters remained unchanged.

4. The resulting set ($Q_{AM\,new}$, $Q_{PM\,base}$, ρ_{base}, FUA) was fed into the trained random forest model, which was used to compute the forecast of daily passenger volume, $Q_D{}^{RF}(\Delta Q_{AM})$.
5. Based on the obtained results, a response curve was constructed linking the predicted daily passenger volume to changes in the morning peak:

$$Q_D{}^{RF} = f(\Delta Q_{AM}) \tag{5}$$

The response curve obtained from the computations of predicted daily passenger volume was visualized in the form of a plot (Fig. 3).

The resulting response curve shows a clear departure from strictly linear behavior. Near the baseline morning peak value, the relationship between changes in Q_{AM} and daily passenger volume remains close to linear, indicating a comparable response of daily demand to moderate changes in the morning service offer.

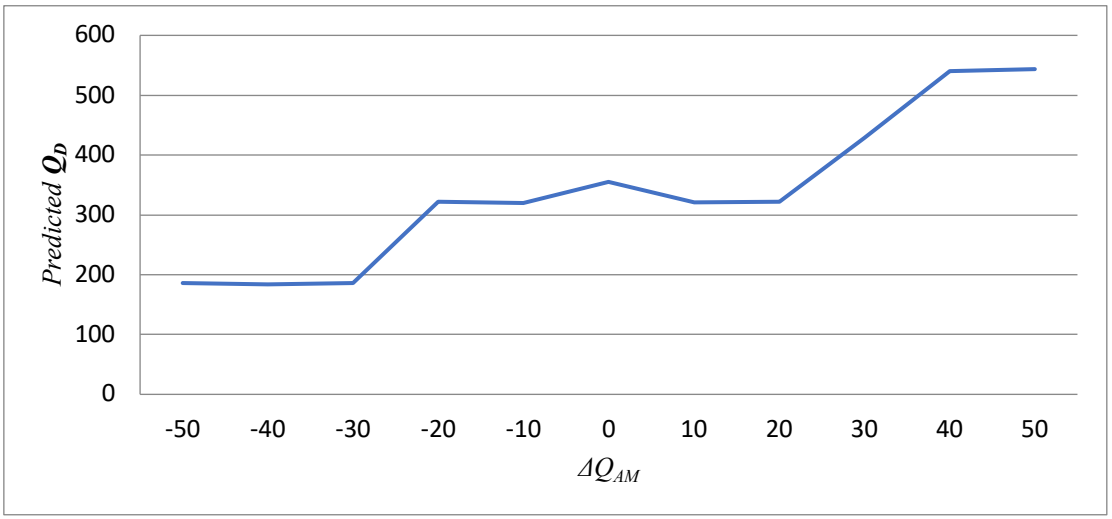

Fig. 3. The relationship between predicted daily passenger volume and changes in the morning peak ΔQ_{AM}

As Q_{AM} increases further, the slope of the curve gradually decreases, reflecting a saturation effect: each additional passenger in the morning peak generates an increasingly smaller increment in daily passenger volume. This indicates the presence of structural constraints typical of large stations and hubs, where simply expanding peak-hour service becomes a less effective instrument for stimulating demand.

When Q_{AM} decreases, daily passenger volume also declines, but the system's response appears to be smooth. Even with a moderate reduction in the morning service offer, a baseline core of demand is maintained, and the losses in daily passenger volume are less proportional than would be implied by the linear model.

The sensitivity analysis naturally complements the results of the earlier stages of the study. It confirms the dominant role of the morning peak in shaping daily passenger volume and indicates that a linear approximation is adequate only within a limited range of loads. The identified saturation effects explain the differences between the linear regression and the random forest model for stations with high peak flows and highlight the advantage of the nonlinear approach when analyzing extreme scenarios.

From a practical perspective, the sensitivity analysis makes it possible to move from describing the demand structure to assessing the consequences of specific managerial decisions. The resulting response curves can be used to preliminarily screen measures related to changes in the morning service offer, to evaluate the effects of introducing or reducing services, and to perform further economic analyses that link changes in daily passenger volume with revenue and cost levels. In this sense, the sensitivity analysis transforms the random forest model from an analytical tool into an applied component of transport planning support.

However, sensitivity analysis assesses model behavior under changes in input parameters and does not reveal the models' comparative predictive accuracy. In the next chapter, the OLS and random forest

models are validated using standard performance metrics: root mean square error (RMSE), mean absolute error (MAE) and coefficient of determination (R^2). The results are summarized in Table 5.

Table 5

Model quality metrics

Model	RMSE	MAE	R^2
OLS	289.62	119.63	0.9722
Random Forest	426.68	82.25	0.9397

The validation results indicate that the linear regression and the random forest model behave fundamentally differently depending on the scale of a station's daily passenger volume. The linear model exhibits a lower RMSE and a higher coefficient of determination, which indicates its ability to reproduce the overall dispersion structure of daily passenger volume well, especially for stations with high Q_D values. In practice, OLS explains the vast majority of variability in daily demand, as it correctly describes the behavior of major network hubs.

At the same time, the random forest model achieves a lower MAE, indicating higher predictive accuracy for most stations with moderate passenger flows. For a "typical station," the random forest model produces a smaller average error than linear regression, reflecting its ability to better fit local data characteristics.

The combination of a lower RMSE for OLS and a lower MAE for the random forest model is unusual, yet cognitively informative. It suggests that the random forest model performs poorly when predicting extremely high values of daily passenger volume, where a few large errors substantially increase RMSE. At the same time, for most stations with medium and low flows, the nonlinear model provides an accurate approximation of the observed values.

4. DISCUSSION

Overall, the differences between the models can be interpreted as follows: the linear regression describes the system well regarding global relationships and the behavior of the largest hubs, whereas the random forest model more effectively models the network, thus providing higher accuracy for typical forecasts. In the context of this study, this underscores that the surprisingly high performance of a simple linear regression stems from the near-linearity of the fundamental relationships in the dataset, while the advantages of the nonlinear model become apparent primarily when modeling stations with moderate passenger flows, for which the accuracy of typical predictions is crucial.

The numerical quality metrics were complemented by a visual verification of the fit between the actual and predicted daily passenger volume for the random forest model. The results of the scatter plot analysis of observed versus predicted values indicate that, for most stations with low and medium Q_D values, the random forest predictions cluster close to the line of perfect agreement, which corresponds to the low MAE and high accuracy of typical forecasts. At the same time, for stations with extremely high daily passenger volume (Przemyśl and the Rzeszów ROF), the dispersion of predictions increases, leading to a few large errors and, consequently, a higher RMSE compared with the linear model. No significant systematic bias in the forecasts was identified, as the deviations are local and concentrated in the tails of the distribution.

A combined analysis of the numerical metrics and the visual verification of predictions allows several key conclusions to be formulated. In the task considered, the linear regression does not perform any worse than the random forest model and, in terms of some global performance measures (RMSE and R^2), even achieves better results. This points to the near-linear nature of the relationship between daily passenger volume and the characteristics of station peak loads.

At the same time, the random forest model remains a useful analytical tool. It provides higher prediction accuracy for typical stations with moderate passenger flows and enables the identification of

nonlinear effects, in particular, the saturation effect at large hubs, which was leveraged in the sensitivity analysis. Thus, the combined use of OLS and random forest models proves methodologically justified: the linear model offers an interpretable view of global relationships, while the nonlinear model refines the system's behavior from a local and scenario-based perspective.

5. CONCLUSIONS

Real-world field data from passenger-flow surveys conducted by the authors on the railway network of the Podkarpackie Voivodeship were used in this study. This ensures the empirical credibility of the results and distinguishes the study from analyses based on aggregated statistics or modeled data.

The analysis showed that a station's daily passenger volume is primarily determined by the magnitude of the morning peak flow, while the evening peak plays a secondary yet stable role. This finding is methodologically robust and is confirmed by the linear regression and the nonlinear random forest model.

It was established that, for most stations with moderate passenger flows, the relationship between daily passenger volume and peak-period loads is nearly linear, which explains the high explanatory power of the OLS model. At the same time, saturation effects emerged at the largest network hubs, whereby further increases in the morning service diminished the increases in daily demand. These nonlinearities are correctly identified by the random forest model and are reflected in the sensitivity analysis results.

Model validation yielded a nontrivial yet substantively important outcome: linear regression better captures the global structure of passenger demand and the behavior of large stations, whereas the random forest model provides higher predictive accuracy for typical stations with moderate flows. The joint use of both models is justified, as it makes it possible to retain interpretability while also identifying the limits of effectiveness of managerial interventions.

The conclusions have direct practical relevance for transport planning. For stations with a medium load level, increasing the morning service offer can be an effective instrument for stimulating daily mobility, whereas for large hubs, more complex solutions are required to go beyond a simple increase in the number of services.

References

1. Chmielewski, J. & Szarata, M. & Siwowski, T. et al. *Final Report from the Study Development of a Regional Transport Model for the Podkarpackie Voivodeship*. Rzeszów University of Technology. Rzeszów. 2025.
2. Iseki, H. & Liu, C. & Knaap, G. The determinants of travel demand between rail stations: A direct transit demand model using multilevel analysis for the Washington D.C. Metrorail system. *Transportation Research Part A: Policy and Practice*. 2018. Vol. 116. P. 635-649.
3. Zhu, Y. & Chen, F. & Wang, Z. & Deng, J. Spatio-temporal analysis of rail station ridership determinants in the built environment. *Transportation*. 2019. Vol. 46(6). P. 2269-2289.
4. Wu, H. & Lee, J (Brian). & Levinson, D. The node-place model, accessibility, and station level transit ridership. *Journal of Transport Geography*. 2023. Vol. 113. No. 103739.
5. Ma, X. & Zhang, J. & Ding, C. & Wang, Y. A geographically and temporally weighted regression model to explore the spatiotemporal influence of built environment on transit ridership. *Computers, Environment and Urban Systems*. 2018. Vol. 70. P.113-124.
6. He, Y. & Zhao, Y. & Tsui, KL. An adapted geographically weighted LASSO (Ada-GWL) model for predicting subway ridership. *Transportation*. 2021. Vol. 48(3). P. 1185-1216.
7. Chen, E. & Ye, Z. & Wang, C. & Zhang, W. Discovering the spatio-temporal impacts of built environment on metro ridership using smart card data. *Cities*. 2019. Vol. 95. P. 102359.
8. Li, D. & Zang, H. & He, Q. Assessing rail station accessibility based on improved two-step floating catchment area method and map service API. *Sustainability*. 2022. Vol. 14(22). No. 15281.

9. Kang, W. & Oshan, T.M. Scale and correlation in multiscale geographically weighted regression (MGWR). *Journal of Geographical Systems*. 2025. Vol. 27(3). P. 399-424.
10. Oshan, T. & Li, Z. & Kang, W. et al. A Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. *ISPRS International Journal of Geo-Information*. 2019. Vol. 8(6). No. 269.
11. Diaz-Gutierrez, J. & Ranjbari, A. How effective are fixed-effects models in fixing the transit supply-demand bidirectional interaction? *Transportation Letters*. 2025.Vol. 17(7). P. 1199-1212.
12. Szántó, N. & Fischer, S. & Monek, G.D. A Novel Method for Simulation Model Generation of Production Systems Using PLC Sensor and Actuator State Monitoring. *Journal of Sensor and Actuator Networks*. 2025. Vol. 14(3). No. 55.
13. Fischer, S. & Hermán, B. & Sysyn, M. et al. Quantitative analysis and optimization of energy efficiency in electric multiple units. *Facta Universitatis, Series: Mechanical Engineering*. 2025. Vol. 23(2). P. 351-375.
14. Gu, Y. & Dou, M. Nonlinear and threshold effects on station-level ridership: insights from disproportionate weekday-to-weekend impacts. *ISPRS International Journal of Geo-Information.* 2024. Vol. 13(10). No. 365.
15. Shao, Q. & Zhang, W. & Cao, X. et al. Threshold and moderating effects of land use on metro ridership in Shenzhen: Implications for TOD planning. *Journal of Transport Geography*. 2020. Vol. 89 No.102878.
16. Yang, L. & Peng, Y. & Chen, J. et al. Temporal variations in the non-linear relationships between metro ridership and the built environment: insights from interpretable machine learning using four-year data. *Intelligent Transportation Infrastructure*. 2024. Vol. 3. No. liae023.
17. Patni, S. & (Siva) Srinivasan, S. Competing or complimentary: Modeling transit ridership at route-level considering inter-route interdependencies. *Travel Behaviour and Society*. 2024. Vol. 36. No. 100815.
18. Zhao, Y. & Wei, J. & Li, H. & Huang, Y. Predicting station-level peak hour ridership of metro considering the peak deviation coefficient. *Sustainability*. 2024. Vol. 16(3). No. 1225.
19. Yu, L. & Cong, Y. & Chen, K. Determination of the peak hour ridership of metro stations in Xi'an, China using geographically-weighted regression. *Sustainability*. 2020. Vol. 12(6). No. 2255.
20. Sabraliev, N. & Abzhapbarova, A. & Nugymanova, G. et al. Modern aspects of modeling of transport routes in Kazakhstan. *NEWS of National Academy of Sciences of the Republic of Kazakhstan*. 2019. Vol. 2(434). P. 62-68.

Received 01.05.2024; accepted in revised form 06.03.2026