PREDICTION OF REAL-TIME TRAIN ARRIVAL TIMES ALONG THE SWEDISH SOUTHERN MAINLINE

KAH YONG TIONG[1], ZHENLIANG MA[2] & CARL-WILLIAM PALMQVIST[1]
[1]Department of Technology and Society, Lund University, Sweden
[2]Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, Sweden

ABSTRACT
Real-time train arrival time prediction is crucial for providing passenger information and timely decision support. The paper develops methods to simultaneously predict train arrival times at downstream stations, including direct multiple output liner regression (DMOLR) and seemingly unrelated regression (SUR) models. To capture correlations of prediction equations, two bias correction terms are tested: (1) one-step prior prediction error and (2) upstream prediction errors. The models are validated on high-speed trains operation data along the Swedish Southern Mainline from 2016 to 2020. The results show that the DMOLR model slightly outperforms the SUR. The DMOLR's prediction performance improves up to 0.32% and 24.03% in term of RMSE and R^2 respectively when upstream prediction errors are considered.
Keywords: train arrival time predictions, direct multiple output liner regression, seemingly unrelated regression.

1 INTRODUCTION

The accurate prediction of train movements is critical to ensure the quality and reliability of railway transport. It provides passengers with reliable decision support, allowing them to take proactive actions to mitigate the impact of train delays. Passengers, for example, can make alternate travel plans if they are informed in advance of any train delays.

In passenger information systems, real-time information is utilised to generate a continuous prediction of train arrival times for multiple downstream stations at arbitrary prediction times. The predictions are updated in real time periodically or when the train arrives at a station. However, most studies developed data-driven approaches to predict train arrival times at the next train station [1]–[5]. The paper aims to model the train arrival time prediction at multiple station and arbitrary times and explores prediction approaches and prediction bias correction methods.

The remaining of the paper is organized as follows: Section 2 examines existing literature on the train delay prediction. Section 3 introduces the model formulation and prediction methodologies. Section 4 presents a case study of the high-speed train (HSR) line in Sweden. Section 5 concludes the findings of the paper and discusses future research directions.

2 LITERATURE REVIEW

With the growing availability of data in the rail industry, a number of studies have been conducted to predict the train delay in real-time. This task is well suited to data-driven methods, which include statistical models, machine learning models, and deep learning models. For statistical methods, Gorman [6] used linear regression to identify factors that contribute to railroad congestion delays. According to Jiang et al. [7], semi-parametric models outperform linear models, weibull distributions, binomial logistic regression, and random forest alone while maintaining interpretability.

Machine learning methods are becoming increasingly popular due to their ability to handle high-dimensional data and nonlinear relationships between dependent and explanatory variables. For example, Oh et al. [8] developed a real-time dwell time prediction model for

dynamic railway timetables by combining support vector regression (SVR), multiple linear regression, and RF techniques using data from real-time metro operation and smartcard data. Huang et al. [9] adopted Kalman filter to update SVR prediction using real-time information to ensure accurate running time predictions under unexpected situations. Considering passengers have different destinations, Tiang et al. [10] developed an arrival times prediction model for any downstream stations rather than only the next station using direct multi-output light gradient boosting machine. The works of Barbour et al. [11], Li et al. [12], Wen et al. [13] and Gao et al. [14] demonstrated the superiority of RF over other data-driven methods in predicting train events.

In recent years, deep learning methods have also been widely used to predict train delays. For example, Wen et al. [2] and Mou et al. [15] utilised long short-term memory to predict the arrival delay time of the train at the next station. Oneto et al. [16] employed extreme learning machines (ELM) together with train operation data and weather data to build a dynamic train delay prediction model. Li et al. [3] and Bao et al. [17] utilised particle swarm optimization algorithm to optimize hyperparameter of ELM when predicting train delay for real-time train dispatching. To comprehensively account for the temporal and spatial dependence between multiple trains and routes, Zhang et al. [18] proposed a train spatio-temporal graph convolutional network to predict the collective cumulative effect of train delays.

Despite the fact that numerous efforts have been made to develop train delay prediction models using advanced and complex algorithms, little work has been done to improve the model through bias or error adjustment. To address this gap, we propose a prediction framework that uses seemingly unrelated regression (SUR) to account for systematic model bias and prediction residual correlations due to unobserved predictors, as well as bias correction modules to improve model prediction.

3 METHODOLOGY

3.1 Problem formulation

The goal of this paper is to predict train arrival times for multiple downstream stations at arbitrary times. This allows passengers to keep track of their own journey between two points: origin and destination.

Consider a train line with multiple stations $\{i = 0, 1, 2, \ldots, N\}$, with N being the last station. Given the train is located at the current station i at time t, predict the train arrival times at the downstream stations $\{a_{i+1}, a_{i+2}, \ldots, a_N\}$. The prediction model receives real-time train operation data based on the current train location, allowing downstream train arrival time prediction based on the most up-to-date information. When the train arrives at station $i + 1$, predictions for the train arrival times at the downstream stations $\{a_{i+2}, a_{i+3}, \ldots, a_N\}$ are simultaneously updated. The prediction problem is represented as a multi-output regression conditional on the train location.

$$\widehat{\mathbf{y}} = f(X \mid i), \tag{1}$$

where $\widehat{\mathbf{y}}$ is the predicted train arrival time for all the downstream stations; i is the current train station; X is the predictor variables presented in Section 3.4.

Fig. 1 shows the prediction framework.

3.2 Prediction methods

In order to make train arrival time predictions for multiple downstream stations, multiple regression models are generated simultaneously. In this study, linear regression with two

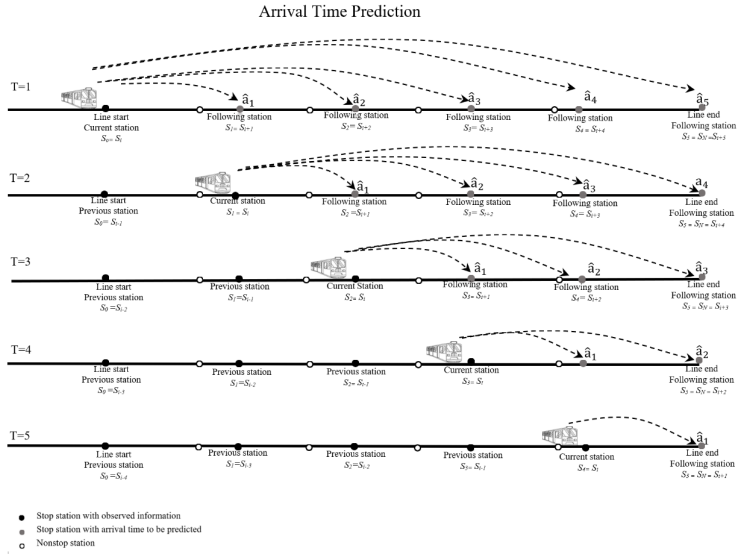


Figure 1: Prediction framework.

different frameworks, that is (1) direct multiple output linear regression (DMOLR) and (2) seemingly unrelated regression (SUR), are proposed to train these regression models.

The DMOLR model train the multiple linear regression equations independently with corresponding sets of explanatory variables. The predictor variables at current station i is inputted into $N - i$ number of regression models to generate prediction for downstream train arrival times $\widehat{a}_{i+1}, \widehat{a}_{i+2}, \ldots, \widehat{a}_N$. The prediction outputs are independent of one another in DMOLR, neglecting contemporaneous correlations between prediction equation error terms.

SUR, on the other hand, is a system of linear regressions that accounts for prediction errors that are correlated across equations [19]. Because there may be correlations between the shared unobserved characteristics, SUR's use of correlated errors terms across multiple regression models may increase the likelihood of prediction. For detailed information on SUR, see Washington et al. [20] and Brownlee [21] for more information on direct multiple output regression.

3.3 Bias correction modules

This section introduces two iterative correction approaches: (1) one-step before prediction error correction and (2) upstream prediction error correction. The purpose of this step is to improve the model's performance by leveraging the real-time observed information (prediction errors at previous stations) along with the predictors in Section 3.4.

The model utilises real time information $X_{i,i}$ at current station i where the train currently located to make prediction for all downstream station $\widehat{y} = \widehat{a}_{i+1,i}, \widehat{a}_{i+2,i}, \ldots, \widehat{a}_{N,i}$. The actual train arrival time $a_{i+1,i+1}$ is observed when the train arrives at the next station $i + 1$, and the prediction error $E_{i+1,i+1}$ can be calculated using eqn (2).

$$E_{i,i} = \widehat{a}_{i,i-1} - a_{i,i}, \tag{2}$$

where $E_{i,i}$ is the prediction error at station i when the train is at station i; $\widehat{a}_{i,i-1}$ is the predicted train arrival times for station i when the train is at station $i-1$; $a_{i,i}$ is the actual train arrival times for station i when the train is at station i.

For one-step before prediction error correction, only prediction error at current station i where the train currently located, $E_{i,i}$, is utilised as a predictor together with predictors $X_{i,i}$ introduced in Section 3.4. For example, in Fig. 2(a), at $T=3$, the train arrives at station S_2, the model predicts the arrival times for downstream stations $\{S_3, S_4, S_5\}$ to be $\{\widehat{a}_{3,2}, \widehat{a}_{4,2}, \widehat{a}_{5,2}\}$. At $T=4$, the train arrives at station S_3, the actual arrival times $a_{3,3}$ of station S_3 is observed, the prediction error $E_{3,3}$ is computed and used as predictor with $X_{3,3}$ to make prediction $\{\widehat{a}_{4,3}, \widehat{a}_{5,3}\}$.

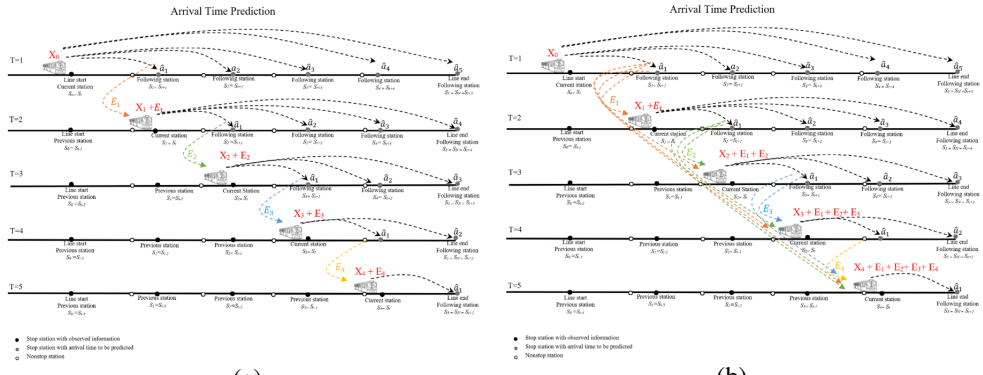
For upstream prediction errors correction, all previously determined prediction errors together with prediction errors at the current station, $\{E_{1,1}, \ldots, E_{i-1,i-1}, E_{i,i}\}$, are utilised as predictors together with $X_{i,i}$ as predictors. For example, in Fig. 2(b), at $T=4$, the train arrives at station S_3, the prediction error $E_{3,3}$ is computed and used together with previously computed prediction errors $\{E_{1,1}, E_{2,2}\}$ and $X_{3,3}$ as predictors to make prediction for $\{\widehat{a}_{4,3}, \widehat{a}_{5,3}\}$.

3.4 Predictor variables

The train operation data is used to define predictor variables utilised in this study. Six common factors that influence train event prediction are selected as potential predictors:

1. The arrival delay $\Delta a_{i,k}$ of the train k at current station i
2. The departure delay $\Delta d_{i,k}$ of the train k at current station i
3. The actual dwell time $dw^r_{i,k}$ of the train k at current station i
4. The scheduled headway $hw^s_{i,k}$ of the train k with previous train $k-1$ at current station i
5. The actual headway $hw^r_{i,k}$ of the train k with previous train $k-1$ at current station i
6. The scheduled running time $r^s_{i,k}$ of the train k from current station i to the next station $i+1$

Y represents the times required for the train to arrive at the targeted stations given the current station where the train is currently located. All the variables in this study are measured in minutes.



(a) (b)

Figure 2: (a) One-step before prediction error correction; and (b) Upstream prediction errors correction.

To avoid multicollinearity, these variables are first analysed and filtered using Pearson Correlation. $\Delta a_{i,k}$ is excluded since it is highly correlated to $\Delta d_{i,k}$ (with Pearson correlation coefficient (PCC) of 0.96) and has slightly lower importance with Y than $\Delta d_{i,k}$ (with PCC of 0.11 vs. 0.18). Thus, the predictor variables $\Delta a_{i,k}, dw^r_{i,k}, hw^s_{i,k}, hw^r_{i,k}, r^s_{i,k}$ are chosen for the final regression model.

4 CASE STUDY

4.1 Data

The train arrival time prediction is applied to the northbound direction of the long-distance HSR along the Swedish Southern Mainline, Sweden, from Lund C station to Linköpings C station (red dashed line with yellow dots in Fig. 3).

We focus on four HSR (train service No. 530, 538, 542, and 546) due to the data availability throughout the time span from December 2016 to December 2020. However, the train headways are calculated by taking into account all types of trains that use the line.

As trains travel along the route toward the final station N, the explanatory variables used to predict arrival times will change in time and space. To model train moving downstream closer toward N, separate datasets with different sets of explanatory variables that reflect different current railway traffic condition in time and space are used to train the prediction models. Since this study area has six stations, five separate datasets are prepared.

4.2 Experimental setting

The data is cleaned before training by removing observations with missing values or the extreme observation that exceeding two standard deviations from the central. The StandardScaler function in the scikit-learn library is then employed to normalized continuous

Figure 3: Swedish Southern Mainline.

variables into [0, 1]. The scikit-learn library's MultiOutputRegressor class extends the linear regression model to build MOLR whereas SUR models are implemented using "systemfit" package in R.

The R-squared is used to determine the model's goodness of fit in this study. The model fits better if R-squared is close to one. The root-mean-square error (RMSE) is used to evaluate the performance of prediction models in terms of estimation errors, as shown in eqn (3).

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(\hat{y}_k - y_k)^2}, \tag{3}$$

where y_k and $\hat{y}_k$ respectively represent the actual and predicted arrival times, recorded in minutes. The closer RMSE are to zero, the better the performance of the model.

4.3 Results

First, we compare the performance of SUR and DMOLR in terms of R^2 and RMSE. The SUR is found to not have an obvious advantage compared with DMOLR since DMOLR has a slightly lower RMSE (0.05% lower than DMOLR) and no difference in R^2 on overall. This indicates that taking into account the prediction residual correlations has no significant impact on the subsequent stations' train arrival time prediction. Thus, the direct multi-output regression model structure is used for further analysis.

Then, we attempt to measure the effectiveness of the bias correction approaches in enhancing the prediction accuracy of the prediction models. The % Improvement in RMSE and R^2 is determined through the comparison between DMOLR with correction and the baseline model, that is DMOLR without any correction using eqn (4).

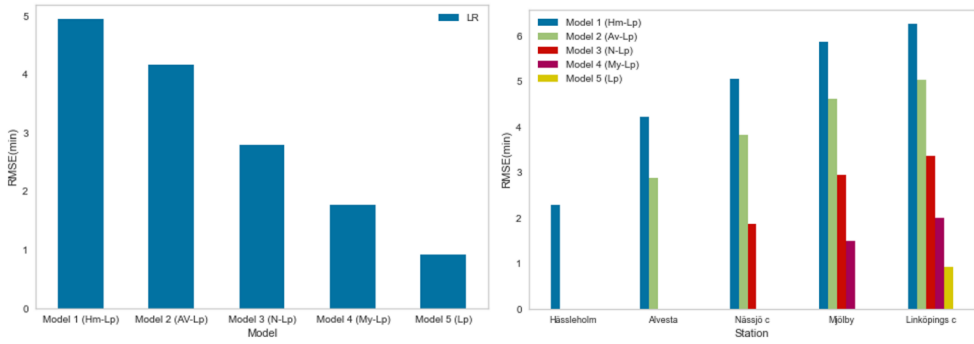
$$\text{Improvement } (\%) = \left(\frac{A_i - B_i}{B_i}\right) \times 100\%, \tag{4}$$

where A_i is the model prediction performance with correction at station i station, B_i is the model prediction performance without correction at station i (baseline model). It is important to note that improvement in RMSE indicates a lower RMSE and opposite is true for R^2.

Table 1 shows the comparison results of prediction models with or without bias correction modules. It shows that the DMOLR model with one-step before prediction error correction performs better than that without correction. This finding gives evidence that iterative prediction error adjustment using real-time information enables the model to constantly adjust itself, thus giving a slightly better prediction effect. DMOLR with upstream prediction errors correction via exhibits a slightly more notable improvement effect over the other models in

Table 1: Percentage improvement using different bias correction modules where baseline model is DMOLR without correction.

Model ($S_{i+1} - S_N$)	Improvement one-step before prediction error correction		Upstream prediction errors correction	
	% Improvement in RMSE	% Improvement in R^2	% Improvement in RMSE	% Improvement in R^2
Model 1 (Hm-Lp)	0.00%	0.00%	0.00%	0.00 %
Model 2 (AV-Lp)	1.34%	5.22%	0.02%	0.09%
Model 3 (N-Lp)	0.00%	0.50%	0.17%	20.90%
Model 4 (My-Lp)	0.09%	0.36%	0.15%	0.62%
Model 5 (Lp)	0.27%	20.34%	0.32%	24.03%

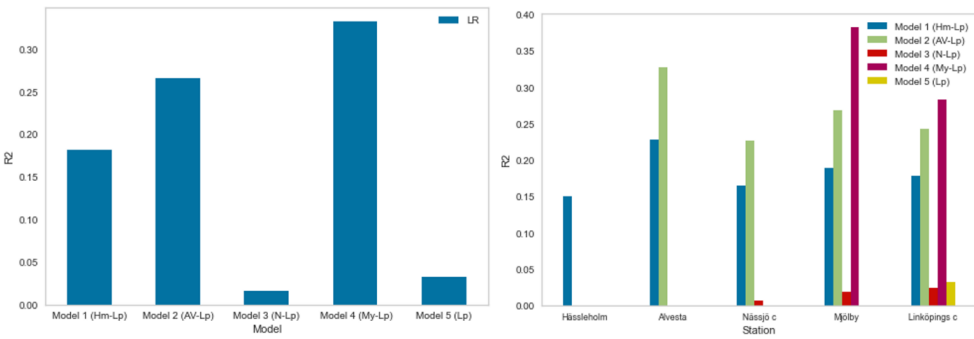


(a) (b)

Figure 4: (a) Average RMSE; and (b) RMSE across stations.

Table 1, showing the importance of both historical and real-time information in enhancing the performance of the models.

On the basis of previous analysis, the performance of DMOLR with upstream prediction error correction is evaluated from other angles, in each space horizon and each evaluation metric. Fig. 4(a) demonstrates average RMSE when moving closer toward the last station, Linköpings c (i is closer to N). We can notice that model 1 (Hm-Lp), which needs to predict the arrival times for more number of stations from i to N, has the largest prediction error with an average RMSE of 4.94 min. This is reasonable since it has more uncertainty and fluctuations in railway traffic conditions when the distance between i and N is longer. Fig. 4(b) explores the changes in RMSE and MAE of DMOLR at each station when the trains move along the route. Linköpings c has the largest prediction errors since the prediction of arrival times is a cumulative measurement. We can also observe that the prediction performance of DMOLR at a station improves with the train moving closer to it since more relevant real-time information that can capture the actual dynamics of traffic is used to make predictions. However, the R^2 values of these models as shown in Fig. 5 are relatively low. This might be attributed to the use of less representative predictor variables.



(a) (b)

Figure 5: (a) Average R2; and (b) R2 across stations.

5 CONCLUSION

The paper aims to develop and explore ways to improve train arrival time prediction for multiple downstream stations at arbitrary times in passenger railway systems. Specifically, this paper focuses on two main objectives. First, it examines whether contemporaneous correlations between the error terms across regression equations should be taken into account when developing prediction models. Second, the effectiveness of bias correction modules in enhancing the prediction model performance is assessed.

The case study findings shows that the train arrival time prediction is independent of the prediction residual correlations since the performance of SUR and DMOLR are comparable across stations. The train arrival prediction models taking into consideration of upstream prediction errors correction is found to be an effective in improving the prediction in downstream stations. Future studies will improve the model by exploring deep learning methods that could automatically capture complex interactions of variables. The low R^2 of the model indicates the need to include more predictor variables, such as historical train operational data, passenger data, weather conditions, infrastructure information, to improve the model's overall representative power.

ACKNOWLEDGEMENT

This work was funded by the Swedish Transport Administration, Grant Number TRV2018/139443.

REFERENCES

[1] Barbour, W., Mori, J.C.M., Kuppa, S. & Work, D.B., Prediction of arrival times of freight traffic on us railroads using support vector regression. *Transportation Research Part C: Emerging Technologies*, **93**, pp. 211–227, 2018.

[2] Wen, C., Mou, W., Huang, P. & Li, Z., A predictive model of train delays on a railway line. *Journal of Forecasting*, **39**(3), pp. 470–488, 2020.

[3] Li, Y., Xu, X., Li, J. & Shi, R., A delay prediction model for high-speed railway: an extreme learning machine tuned via particle swarm optimization. *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, pp. 1–5, 2020.

[4] Li, Z., Huang, P., Wen, C., Tang, Y. & Jiang, X., Predictive models for influence of primary delays using high-speed train operation records. *Journal of Forecasting*, **39**(8), pp. 1198–1212, 2020.

[5] Huang, P., Wen, C., Fu, L., Peng, Q. & Tang, Y., A deep learning approach for multi-attribute data: A study of train delay prediction in railway systems. *Information Sciences*, **516**, pp. 234–253, 2020.

[6] Gorman, M.F., Statistical estimation of railroad congestion delay. *Transportation Research Part E: Logistics and Transportation Review*, **45**(3), pp. 446–456, 2009.

[7] Jiang, S., Persson, C. & Akesson, J., Punctuality prediction: Combined probability approach and random forest modelling with railway delay statistics in Sweden. *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, IEEE, pp. 2797–2802, 2019.

[8] Oh, Y., Byon, Y.J., Song, J.Y., Kwak, H.C. & Kang, S., Dwell time estimation using real-time train operation and smart card-based passenger data: A case study in Seoul, South Korea. *Applied Sciences*, **10**(2), p. 476, 2020.

[9] Huang, P., Wen, C., Fu, L., Peng, Q. & Li, Z., A hybrid model to improve the train running time prediction ability during high-speed railway disruptions. *Safety Science*, **122**, 104510, 2020.

[10] Tiong, K.Y., Ma, Z.L. & Palmqvist, C.W., Real-time train arrival time prediction at multiple stations and arbitrary times. *25th International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2022. (Paper submitted.)
[11] Barbour, W., Samal, C., Kuppa, S., Dubey, A. & Work, D.B., On the data-driven prediction of arrival times for freight trains on us railroads. *21st International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, pp. 2289–2296, 2018.
[12] Li, Z., Wen, C., Hu, R., Xu, C., Huang, P. & Jiang, X., Near-term train delay prediction in the Dutch railways network. *International Journal of Rail Transportation*, **9**(6), pp. 520–539, 2021.
[13] Wen, C., Lessan, J., Fu, L., Huang, P. & Jiang, C., Data-driven models for predicting delay recovery in high-speed rail. *4th International Conference on Transportation Information and Safety (ICTIS)*, IEEE, pp. 144–151, 2017.
[14] Gao, B., Ou, D., Dong, D. & Wu, Y., A data-driven two-stage prediction model for train primary-delay recovery time. *International Journal of Software Engineering and Knowledge Engineering*, **30**(7), pp. 921–940, 2020.
[15] Mou, W., Cheng, Z. & Wen, C., Predictive model of train delays in a railway system. *RailNorrköping 2019. 8th International Conference on Railway Operations Modelling and Analysis (ICROMA)*, Norrköping, Sweden, 17–20 Jun., pp. 913–929, 2019.
[16] Oneto, L. et al., Dynamic delay predictions for large-scale railway networks: Deep and shallow extreme learning machines tuned via thresholdout. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, **47**(10), pp. 2754–2767, 2017.
[17] Bao, X., Li, Y., Li, J., Shi, R. & Ding, X., Prediction of train arrival delay using hybrid ELM-PSO approach. *Journal of Advanced Transportation*, **2021**, 2021.
[18] Zhang, Y., Li, R., Guo, T., Li, Z., Wang, Y. & Chen, F., A conditional Bayesian delay propagation model for large-scale railway traffic networks. *Australasian Transport Research Forum, ATRF 2019-Proceedings*, 2019.
[19] Nasri, A. & Zhang, L., Multi-level urban form and commuting mode share in rail station areas across the United States: A seemingly unrelated regression approach. *Transport Policy*, **81**, pp. 311–319, 2019.
[20] Washington, S., Karlaftis, M., Mannering, F. & Anastasopoulos, P., *Statistical and Econometric Methods for Transportation Data Analysis*, Chapman and Hall/CRC, 2020.
[21] Brownlee, J., How to develop multi-output regression models with Python, 2020.