Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Predicting Minor Road Traffic from Major Road Traffic Counts using Geographically Weighted Poisson Regression in Lancashire

Mahardi Arief Tanjung[1], Gari Mauramdha[1*], Jachrizal Sumabrata[1], Malcolm Morgan[2]

[1]Department of Civil and Environmental Engineering, Faculty of Engineering, University of Indonesia,
[2]Institute for Transport Studies, University of Leeds, Woodhouse, Leeds LS2 9JT, United Kingdom
Email: gari.mauramdha01@ui.ac.id

This study addresses the challenge of inadequate traffic count data on minor roads due to their low traffic volume and resource-intensive setup of count stations. The objective is to develop a traffic estimation model for these roads using reproducible code and identifying optimal variables for prediction. Lancashire in 2018 serves as the case study area, with 2020 data validating model predictions during the COVID-19 pandemic. Utilizing secondary data encompassing traffic volume, road network, socioeconomic, and LSOA boundary data, the study processed and analyzed traffic data per vehicle type to assess variable effects on the prediction model. Correlation tests revealed significant relationships between variables, particularly traffic volume per vehicle type, enhancing model performance by 1% to 5%. However, high correlations led to elevated standard errors in models, necessitating variable sorting. Despite minimal performance impact post-sorting, standard error significantly decreased. Moreover, integrating socioeconomic variables on minor roads improved model performance compared to general socioeconomic variables.

Keywords: Traffic Estimation Model, Minor Roads, Reproducible Code.

1. Introduction

Road traffic volume data plays a crucial role in traffic management, influencing various sectors from safety to environmental impact. Studies by Alkhatib et al. (2022) and others have leveraged traffic volume databases to assess intersection management with smart road traffic control systems and forecast future traffic patterns. Furthermore, this data has been instrumental in predicting highway accidents, as demonstrated by Alqatawna et al. (2021), and in modeling environ-mental concerns like noise pollution and greenhouse gas emissions, as highlight-ed by Thakre et al. (2020) and Puliafito et al. (2015) respectively. Despite its significance, the collection of road traffic data presents challenges. The Department for Transport utilizes a combination of manual counts and automatic traffic counters (ATCs), with a significant concentration on major roads. However, with over 216,000 miles of minor roads in Great Britain, data collection becomes re-source-intensive, costly, and time-consuming compared to major roads. Given these challenges, this study aims to develop a reliable model for estimating traffic flow on minor roads. The research focuses on employing reproducible code for data processing and modeling, exploring various variable combinations for traffic flow estimation on minor roads, and evaluating the model's effectiveness under the COVID-19 scenario. Please note that the first paragraph of a section or subsection is not indented.

2. Literature Review

In summary, numerous models and factors have proven successful in researchers' estimations of minor road traffic. Still, there are still gaps in earlier research that have not yet been filled. Several studies demonstrated that various variables had a significant influence on the prediction model, such as road density, road centrality, or distance to the nearest main road. However, there are some variables that may affect the prediction model. Yang et al. (2011) examined the possibility that variables gathered for high-functional-class roads might not be adequate for the local AADT estimate. Therefore, the study compared the influence of socio-economic variables on LSOA with socioeconomic variables on minor roads. In addition, research related to modal split for modelling is still limited, such as Wang, T. et al. (2013), who tried to implement the four-step model in developing a traffic prediction model by ignoring the mode split stage. Therefore, this research tries to analyze how the modal split affects the traffic prediction model. However, the use of excessively large number of variables may increase the complexity of the model without providing significant improvement in performance (Cheng, 1993). Therefore, it is necessary to prioritize the variables to be applied so that the modelling can provide better results with sufficient complexity. The findings of the literature review demonstrated that the use of geospatial models might significantly improve the accuracy of variables used to estimate the volume of minor roads. Therefore, this study also adopted modelling using the GWPR model. In addition, Yu's (2022) attempt to build on Martinez's earlier work to produce reproducible open-source code for Cambridgeshire showed initial success. However, its data modelling showed limitations in predicting traffic, and the traffic prediction results on minor roads were still unreliable. Therefore, this paper will build on Yu's previous work to create a new set of reproducible code based on the data processing and modelling framework of this study to provide better prediction results.

3. Methodology

Lancashire, 2018 was identified as the case study area, and Lancashire, 2020, as the validation of the model predictions developed to determine the credibility and reliability of modelling against the COVID-19 pandemic. The data used in the research is secondary data, which includes traffic volume data, road network data, socioeconomic data, and LSOA boundary data. All data can be obtained online to support one of the research objectives of producing a reproducible code for traffic estimation models on minor roads. Based on the collected data, data processing is carried out where road traffic data is separated into traffic data per vehicle type, which will be used in model development to determine the effect of variables from traffic for each vehicle type on the prediction model. Correlation tests are used to determine the level of relationship between variables that will be used in determining the variables that will be used in the prediction model. Some modelling was developed and compared to find out how good the prediction model developed in this study is.

4. Result

Correlation Test

The correlation test results reveal varying levels of significance in relationships among the independent variables and between the independent and dependent variables, as depicted by the dot sizes in Figure 1; larger dots denote stronger correlations, while smaller ones indicate weaker ones. At a 95% confidence level, certain variables show no correlation with each other, represented by black crosses. Notably, the correlation values between the dependent and independent variables range from 0.07 to 0.43, with the standardized centrality variable exhibiting the highest significance and employment in LSOA showing the least. A potential concern is the high correlation between traffic flow variables for different modes, with values ranging from 0.34 to 1, particularly between total traffic volume and car/taxi traffic volume. This is attributed to cars and taxis dominating the total traffic flow, accounting for up to 79%. Additionally, socioeconomic variables display high correlation, with a significant value of 0.91 observed between car ownership on minor roads and the minor road population count.

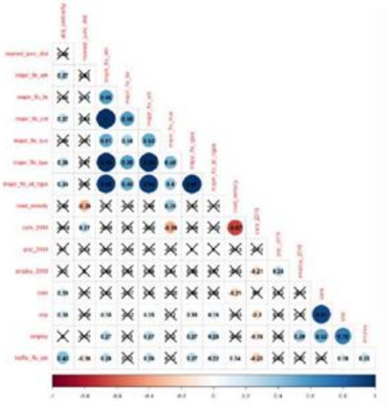

Fig. 1. Variables correlation result

Based on the correlation test, there are five independent variables that have a low significance value for the dependent variable. However, there is a possibility that each variable has a different type of correlation with the dependent variable, and the correlation test used is only to determine the linear relationship between variables (Yu, 2022). In summary, each variable will be used and combined in the estimation model to discover the best prediction model that could be produced.

Modelling Result

The modelling process commenced with the establishment of a base model; wherein pivotal variables derived from previous research were chosen as reliable predictors for minor road traffic. Sequentially, an array of mode split variables was incorporated into this foundational model to gauge each mode's traffic flow's influence on the model's precision in estimating minor road traffic. After this integration, socio-economic variables were infused into the model, leading to a comparative examination between socio-economic variables pertaining to LSOAs and those specific to minor roads. To assess the modelling's effectiveness, its performance was meticulously evaluated by analyzing various parameters and probing into the residuals generated throughout the modelling procedure. Additionally, two intrinsic attributes of minor roads—namely road centrality and distance to the nearest major road, along with traffic flow towards the nearest major road—were employed as the foundational variable combination. This selection, proven effective in forecasting minor road traffic, utilized a logarithmic form for the basic variables. It's worth noting that employing variables in logarithmic form has been demonstrated to yield superior model performance compared to using them in their original numerical state (Yu, 2022).

An analysis of traffic estimation modelling using mode split was conducted by substituting the variable traffic flow on major roads with traffic flow by type of vehicle on major roads.

Table 1. Basic variable with socio economic combination performance

Variable	Indicator		
	AICc	Pesudo R	RMSE
Basic Variable Combination	319175.6	0.501983	2366.421
Car and Taxi	321255.5	0.498737	2374.429
Two Wheels	325013.5	0.492875	2401.595
Bus	331969.7	0.482018	2406.697
LGVS	316933.6	0.505475	2351.994
HGVS	321061.1	0.499032	2364.93

Based on the findings presented in Table 1, the LGVS traffic variable emerges as the most effective in traffic estimation modelling for minor roads, registering a pseudo-R value of approximately 0.505. In contrast, the bus traffic variable exhibits the weakest performance with a pseudo-R index of 0.482. Additionally, the incremental inclusion of each mode split variable results in a modest improvement in modelling performance, ranging from 1% to 5%. Notably, employing individual traffic variables for each mode type in the modelling process, as op-posed to aggregating total traffic variables from major roads, enhances the pseudo-R and AICc indicators by 20% and the RMSE indicator by 10% compared to the performance achieved using the basic variable set. Moreover, incorporating all traffic volume variables can amplify the performance gains to a maximum of 23% beyond the baseline variable combination. However, despite these performance enhancements, the selected modelling variable combination encompasses independent variables with pronounced intervariable correlations, potentially introducing multicollinearity issues that could compromise the modelling results.

Social-economic variables such as road density, population, car ownership, and employment were integrated into the modelling framework to assess their impact on traffic estimation performance when added to the basic variable com-bination. Specifically, a comparative analysis was conducted to distinguish the effects of these socio-economic variables between LSOAs and minor roads. The inclusion of road density led to a 4% enhancement in the traffic estimation performance on minor roads, as reflected by the pseudo-R and AICc values. Furthermore, incorporating population, car ownership, and employment variables resulted in notable improvements: a 0.107 increase in pseudo-R, a 17% reduction in AICc, and a decrease in RMSE by 363,791 or 12%. When juxtaposed with the performance of socio-economic variables based on LSOAs, utilizing those specific to minor roads yielded superior modelling outcomes. The comprehensive integration of all socio-economic variables culminated in a 25% enhancement in the model's predictive capacity over the baseline, boasting a pseudo-R of 0.629, an AICc of 237533.4, and an RMSE of 1934.89. Nonetheless, it's imperative to acknowledge the pronounced inter-variable correlation, particularly between car ownership and population within the socio-economic variables, which may precipitate multicollinearity issues.

Table 2. Basic variable with socio economic combination performance

Variable	1st Qu.	Median	Mean
Basic combination + road density	307283.5	0.520566	2316.559
Basic combination + socio economic in LSOA	264923.5	0.586716	2090.938
Basic combination + socio economic along minor road	250364.9	0.609444	2002.63
Basic combination + all socio economic along minor road	237533.4	0.629495	1934.892

Traffic estimation modelling for minor roads using all independent variables exhibited considerable variability in standard errors, ranging from 15.9% to 77.6%. The centrality variable demonstrated the most reliable performance with a standard error of 0.005, whereas traffic variables on major roads, particularly those related to all vehicle and car/taxi traffic, presented the highest standard errors at 0.799 and 0.627, respectively. Such elevated standard errors likely result from substantial correlations among independent variables, potentially leading to multicollinearity issues that could compromise the model's accuracy (Schwarz et al., 2014). Table 3 indicates that no independent variable consistently influenced the dependent variable in either a positive or negative direction, hinting at potential uncertainties in the modelling outcomes. Residuals, which capture the difference between actual and model-predicted values of the dependent variable, serve as indicators of the model's accuracy. Ideally, these residuals should follow a normal distribution, showing symmetry around zero and lacking discernible patterns. Standardized residuals, obtained by dividing the residual by its standard deviation, provide deeper insights into the distribution of residuals (Yu, 2022). Modelling without considering inter-variable relationships yielded standard residuals ranging from -2.204 to 3.309, with an average close to zero. However, only 187 out of 204 data points, or 92%, could accurately represent actual traffic conditions on minor roads with a 95% confidence level, highlighting that the remaining 8% did not reliably reflect real-world traffic data.

Table 3. Standard residual summary

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8	584	942	1364	1561	43353

When traffic estimation modelling accounts for the correlation between independent variables, it allows for a reduction in the number of variables used. In the modelling of traffic estimation for minor roads, standard variables like centrality, nearest junction distance, traffic flow on major roads for all vehicles, two-wheel vehicles, buses and coaches, as well as population, employment, car ownership in LSOA, and road density were employed. This approach mitigates multi-collinearity, enhancing modelling performance. The best-fit variable combination yielded improved results, reducing the standard error of the traffic estimation modelling to between 0.04 and 0.209. However, akin to the results from the model using all variables, the optimal combination model lacked variables that consistently impacted the model either positively or negatively, leading to uncertain-ty in the outcomes. Furthermore, using a reduced set of variables in the traffic prediction modelling for minor roads resulted in standard residuals that were comparable to those from the model employing all variables, with differences not exceeding 0.4 at maximum and only about 0.2 at minimum.

Table 4. Standard residual summary of best fit model

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.4608	-0.7214	-0.2098	-0.0867	0.2235	3.7551

Based on the significance test, all coefficients used in the best-fit variable combination have a significance value of more than 90%, with the smallest significance value found in the bus traffic flow coefficient, which has a significance value of 96.1%. In addition, only the nearest junction distance coefficient has a significance value of 100% less than the previous study, which contained four coefficients with a significance value of 100% (Yu, 2022). The significant results for each coefficient can be seen in Table 5.

Table 5. Significant test result for prediction model parameter

Coefficient	Number of Significant	Number of non-Significant	Significant Percentages
Intercept	201	3	98.5%
log_std_centrality	202	2	99%
log_nearest_juc_dist	204	0	100%
log_major_flow_all_motor	203	1	99.5%
log_major_flow_two_wheels	197	7	96.6%
log_major_flow_bus	196	8	96.1%
log_pop	197	7	96.6%
log_pop	201	3	98.5%
log_cars_2018	200	4	98%
log_road_density	197	7	96.6%

The modelling with the best-fit combination variable produced traffic predictions on minor roads that could represent traffic conditions based on survey results with a 95% confidence level of 191 points out of a total of 204 traffic count points, or equivalent to 93.6%. The distribution of standard residuals can be seen in Figure 2.

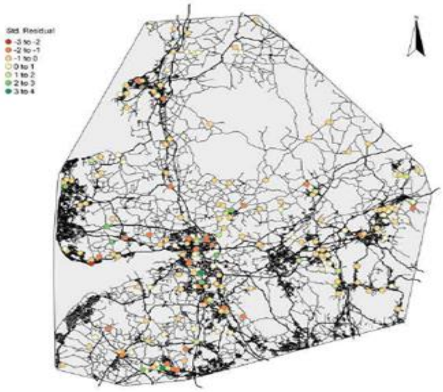

Fig. 2. Standard residuals distribution of best fit prediction model

Result Validation

The validation phase of this research assesses the developed models using data from the same study area but across different years to evaluate their efficacy in predicting minor road traffic under varying circumstances, notably during the COVID-19 pandemic, which significantly altered traffic flow and travel behaviors. Data from Lancashire for the year 2020, a period heavily influenced by pandemic-related policies affecting travel patterns, was chosen for validation. However, the 2020 dataset from Lancashire posed challenges due to limited availability, particularly regarding socio-economic data. To address this, car ownership data for 2020 was estimated based on the fluctuating rates of licensed vehicles in Great Britain during 2019 and 2020. Specifically, while licensed vehicles increased by 1.3% in 2019 (Department for Transport, 2020), they decreased by 0.3% in 2020 (Department for Transport, 2021). Moreover, the 2020 traffic count data from Lancashire exhibited inconsistencies in both distribution and quantity. Despite the reduced number of traffic counts for minor roads in 2020 compared to 2018, the model's pseudo-R values remained consistent between the two years using the chosen variable combinations. Furthermore, the performance metrics of RMSE and AICc in 2020 surpassed those of 2018, as depicted in the accompanying figure.

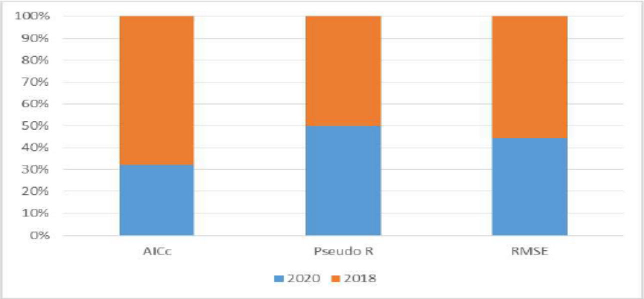

Fig. 3. Performance indicator comparison

The standard residual generated from the modelling also has a better value compared to 2018, with an average value of 0.01289 closer to zero.

Table 6. Standard residual summary in 2020

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.16258	-0.54621	-0.08294	0.01289	0.345342	3.45342

However, it is uncertain what influenced the improved modelling performance in 2020, where the amount of data available and the distribution of traffic count data are quite different from 2018. Nevertheless, modelling with the best-fit combination variable is still capable of predicting traffic on minor roads under COVID-19 pandemic conditions.

Minor Road Estimation

The traffic prediction on the minor road was done using the best modelling approach which resulted in an average traffic on the minor road of 1445 vehicles per day which is only half of the average daily traffic volume of vehicles on the minor road based on the survey report. The most significant difference is in the maximum traffic estimation which can reach 43,353 vehicles per day whereas the maximum daily traffic from the survey results only reaches 18,150 vehicles per day with a total amount of traffic on minor roads reaching more than 20,000 vehicles at 60 points.

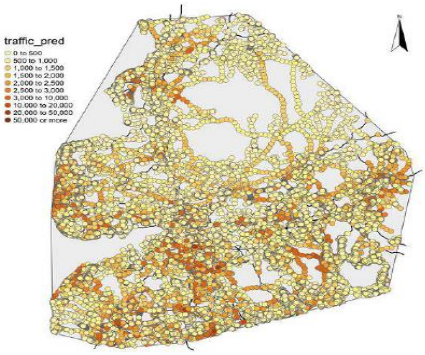

Fig. 4. Minor road prediction results in 2018

Table 7. Comparison of traffic prediction summary and test data in 2018

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Prediction	8	584	942	1364	1561	43353
Test Data	80	494.5	1379	2689.8	3383.2	18150

Meanwhile, the results of traffic estimation in 2020 resulted in traffic predictions on minor roads with an average volume of 1,127 vehicles per day, with the smallest traffic of 13 vehicles per day and the highest traffic reaching 46,396 vehicles per day, while traffic on minor roads reaching more than 20,000 vehicles per day was 9 points.

Please note that the first paragraph of a section or subsection is not indented. The first paragraphs that follow a table, figure, equation etc. does not have an indent, either. Subsequent paragraphs, however, are indented.

5. Conclusion

Analysis of traffic volume data is pivotal for various domains within transport research, encompassing the needs of public transport facilities, road safety enhancements, countermeasure development, trip model calibration, pavement de-sign, and adherence to air quality standards. Despite the prevalence of precise AADT data for high-volume routes, low-volume routes often lack such data. Given that low-volume roads constitute a significant segment of Great Britain's highway system, there's a pressing need for further research to forecast traffic volumes on these routes. This research employs LSOA as a zone boundary, chosen for its accessibility and efficiency in applying the GWPR model. Nonetheless, the heterogeneous sizes of LSOAs, with many encompassing vast areas, can compromise the accuracy of minor road traffic predictions. As Zhong and Hanson (2009) highlighted, smaller zone determinations yield more precise model-ling outcomes for minor road traffic volume predictions. The logarithmic form is instrumental in enhancing the model's efficacy, though its application can be limited to variables with zero values. This study delves into the influence of modal split on minor road traffic estimation modelling, a departure from previous general vehicle traffic-focused studies. Results indicate that incorporating modal split variables augments prediction model performance, albeit with an elevated standard error due to strong correlations among modal split types. Models based on socio-economic variables specific to minor roads outperform those on a broader scale, suggesting potential avenues for incorporating additional variables in future research. Despite successfully predicting traffic for minor roads, this study has limitations. It relies solely on traffic count data from Lancashire County, extrapolated using voronoi diagrams to other case study areas like Bolton, Blackburn, and Wigan, which might exhibit different traffic flow characteristics. The study also overlooks potential discrepancies between estimated and actual traffic count data and doesn't filter out outliers in the available data, a factor that could undermine the prediction model, as noted by Mohamad et al. (1998).

Acknowledgements

This research was funded by a collaboration program between the University of Indonesia and the Ministry of Transportation's HR in implementing the double degree program. In this collaboration, UI through the Department of Civil and Environmental Engineering, Faculty of Engineering, provides a post-graduate program in Civil Engineering for specialization in transportation. As for the foreign universities that will be addressed, the University of Leeds through the Institute for Transport Studies (ITS) which provides master’s programs in various fields of transportation expertise.

References

1. Alkhatib, A.A.A., Maria, K.A., AlZu'bi, S. and Maria, E.A. 2022. Smart Traffic Scheduling for Crowded Cities Road Networks. Egyptian Informatics Journal. 23(4), pp.163-176.
2. Alqatawna, A., Rivas Álvarez, A.M. and García-Moreno, S.S.-C. 2021. Comparison of Multivariate Regression Models and Artificial Neural Networks for Prediction Highway Traffic Accidents in Spain: A Case Study. Transportation Research Procedia. 58, pp.277-284.
3. Apronti, D., Ksaibati, K., Gerow, K. and Hepner, J.J. 2016. Estimating traffic volume on

Wyoming low volume roads using linear and logistic regression methods. Journal of Traffic and Transportation Engineering (English Edition). 3(6), pp.493-506.

4. Das, S. and Tsapakis, I. 2020. Interpretable machine learning approach in estimating traffic volume on low-volume roadways. International Journal of Transportation Science and Technology. 9(1), pp.76-88.
5. Fotheringham, A.S. and Oshan, T.M. 2016. Geographically weighted regression and multicollinearity: dispelling the myth. Journal of Geographical Systems. 18(4), pp.303-329.
6. Khan, S.M., Islam, S., Khan, M.Z., Dey, K.C., Chowdhury, M.A., Huynh, N.N. and Torkjazi, M.R. 2018. Development of Statewide Annual Average Daily Traffic Estimation Model from Short-Term Counts: A Comparative Study for South Carolina. Transportation Research Record. 2672, pp.55 - 64.
7. Lowry, M. 2014. Spatial interpolation of traffic counts based on origin–destination centrality. Journal of transport geography. 36, pp.98-105.
8. Morgan, M., Anable, Jillian, and Lucas, Karen. 2021. A place-based carbon calculator for England. In: Zenodo. Morley, D.W. and Gulliver, J. 2016. Methods to improve traffic flow and noise exposure estimation on minor roads. Environmental pollution (1987). 216, pp.746-754.
9. Munira, S. and Sener, I.N. 2020. A geographically weighted regression model to examine the spatial variation of the socioeconomic and land-use factors associated with Strava bike activity in Austin, Texas. Journal of transport geography. 88, p102865.
10. Puliafito, S.E., Allende, D., Pinto, S. and Castesana, P. 2015. High resolution inventory of GHG emissions of the road transport sector in Argentina. Atmospheric Environment. 101, pp.303-311.
11. Pulugurtha, S.S. and Mathew, S. 2021. Modeling AADT on local functionally classified roads using land use, road density, and nearest nonlocal road data. Journal of Transport Geography. 93, p103071.
12. Schwarz, C., Schwarz, A. and Black, W.C. 2014. Examining the impact of multicollinearity in discovering higher-order factor models. Communications of the Association for Information Systems. 34(1), pp.1191-1208.
13. Selby, B. and Kockelman, K.M. 2013. Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression. Journal of Transport Geography. 29, pp.24-32.
14. Environmental science and pollution research international. 27(30), pp.38311-38320.
15. Wang, J. and Boukerche, A. 2021. Non-parametric models with optimized training strategy for vehicles traffic flow prediction. Computer Networks. 187, p107791.
16. Yeboah, A.S., Codjoe, J. and Thapa, R. 2022. Estimating Average Daily Traffic on Low-Volume Roadways in Louisiana. Transportation Research Record. 2677(1), pp.1732-1740.
17. Yu, Y. 2022. Predicting Minor Road Traffic from Physical and Geodemographic Characteristics. thesis, University of Leeds.