Future Urban Mobility with MaaS

Ensemble-Based Methodology to Identify Optimal Personal Mobility Service Areas Using Public Data

Sangjae Lee\textsuperscript{a}, Seung-oh Son\textsuperscript{b}, Juneyoung Park\textsuperscript{a,b}, and Jaehong Park\textsuperscript{c}

\textsuperscript{a}Dept. of Transportation and Logistic Engineering, Hanyang University, Ansan 15588, Korea
\textsuperscript{b}Dept. of Smart City Engineering, Hanyang University, Ansan 15588, Korea
\textsuperscript{c}Dept. of Highway & Transportation Research, Korea Institute of Civil Engineering & Building Technology, Goyang 10223, Korea

1. Introduction

Public transportation has an important role to form a sustainable metropolis (Murray et al., 1998; Miller et al., 2016). Cities encourage citizens to use public transportation. However, it is difficult to use public transportation in some areas because distance of getting to public transportation is too long in some cities. Compared to other cities, Seoul has higher satisfaction rate with public transportation. There are many cases, however, where short-distance taxis are used because walking to destinations after using public transportation is inconvenient; instead, Personal mobility (PM) devices can be used for these short-distances trip. This study aims to find the optimal PM service area using GIS (Geographic Information System)-based public transportation big data analyses. Variables were generated by collecting socio-economic factors, public transportation data, and geographic data and Extreme gradient boosting and Random forest, which are representative ensemble methods, were used for evaluation. We divided Seoul into a hexagonal grid and developed the optimal PM location service model by creating hexagonal cell data units and analyzing the areas with the models. We found that residential complexes, parks, and near subway stations (all areas with high foot traffic) are best suited for optimal placement. We also determined deployment should be in lower sloped areas. We expect this work to help determine public transportation stop and shared mobility station locations as well as contribute to public transportation demand surveys and accessibility analyses.

ARTICLE HISTORY
Received 29 July 2021
Revised 2 February 2022
Accepted 14 March 2022
Published Online 7 May 2022

ABSTRACT

Public transportation networks are well established in main cities, but there are some inconveniences in using public transportation in some cities. Public transportation is less accessible and walking distance of getting to public transportation is too long in some cities. Compared to other cities, Seoul has a higher satisfaction rate with public transportation. There are many cases, however, where short-distance taxis are used because walking to destinations after using public transportation is inconvenient; instead, Personal mobility (PM) devices can be used for these short-distances trip. This study aims to find the optimal PM service area using GIS (Geographic Information System)-based public transportation big data analyses. Variables were generated by collecting socio-economic factors, public transportation data, and geographic data and Extreme gradient boosting and Random forest, which are representative ensemble methods, were used for evaluation. We divided Seoul into a hexagonal grid and developed the optimal PM location service model by creating hexagonal cell data units and analyzing the areas with the models. We found that residential complexes, parks, and near subway stations (all areas with high foot traffic) are best suited for optimal placement. We also determined deployment should be in lower sloped areas. We expect this work to help determine public transportation stop and shared mobility station locations as well as contribute to public transportation demand surveys and accessibility analyses.
a spatial regression using data mining and public transportation data and found medium- and short-distance travel related to each other while long-distance travel did not correlate. Their work demonstrated that PM can be useful in good weather and for short distances. If operated as a shared service, it can meet the demand for short-distance public transportation and encourage public transportation use. Other related papers investigated additional PM indicators. Choi and Jung (2020) analyzed factors that influencing people to use sharing service of PM. Survey data correlated each factor using multilevel ordered logistic regression models. The survey was conducted under the assumption that shared PM services were used to link to urban railways. 74.1% of bus users and 62.1% of private vehicle and taxi users said they would be willing to use urban railways if shared PM services were available. Lower fares, bike paths, and clearer weather also contributed to positive responses for PM services. In addition, the higher the age and income, the higher the average slope, the more inclined to use. Park et al. (2014) conducted a study to determine whether transit passengers preferred walking or biking to the station. A survey was conducted on commuter train passengers traveling to the train station, and the correlation of the collected indicators was analyzed through the mode choice model. Two binomial logit models were used to find variable statistical significance for walking and biking, travel distance, vehicle possession, and proximity to vehicle roads. Miller et al. (2016) presented criteria for analyzing public transportation sustainability that could affect cities and countries survival. Public transportation system sustainability was assessed using the Composite Sustainability Index (CSI). The CSI is produced by collecting, standardizing, and weighting social, economic and environmental indicators. Referring to this research method, this study attempts to derive the optimal PM placement area by utilizing public transportation indicators, socioeconomic indicators, and geographic conditions.

2. Methodology

2.1 Research Process
The research process of this study is as follows. Based on background and objective of the research in introduction section, this study tried to find optimal PM service area by these three methods. The three methods are hexagonal grid, ensemble methods, and division methods. Based on the data derived through this, the importance of variables for each model was derived, and the data was visualized. Conclusions can be drawn by analyzing the optimal PM placement area derived from the visualization results. A brief flowchart of the analysis procedure is summarized in Fig. 1.

2.2 Hexagonal Grid
This study aims to visualize the prediction model derived from the ensemble methods in the map. This can be visualized through Geographic information system (GIS). GIS is an information system for converting geographic information into computer data and utilizing it efficiently (Chang, 2004). As a case of visualization research through GIS, Choi et al. (2021) analyzed the spatial impacts of social overhead capital (SOC) on housing prices using a geographically weighted regression (GWR) model. The study set spatial units of 200-m and 400-m grids. Uber, an

Fig. 1. Flowchart to Find Optimal PM Service Areas Using Ensemble Method

Fig. 2. Hexagonal Grid
American ride-sharing service company, analyzed the geospatial space using a hexagonal hierarchical grid structure called H3 (Isaac Brodsky, 2018). In this study, a hexagonal grid structure is applied due to its strength of accounting different sizes of urban zone units. Especially, hexagon-based zonal structures allow equal distance between neighboring cells whereas rectangles and other structures have limitations of it. This makes it easier to analyze the influence of neighboring cells.

To select the optimal personal mobility service areas, we divided the analysis area into a hexagonal grid with 100 m radii (Fig. 2). The entire Seoul area was divided into 16605 hexagonal grids, and ID was assigned to each hexagonal cell (Liao et al., 2020; Choi et al., 2021). The hexagonal grid structure makes it easy to compare neighboring cells and uniformly analyze their influence (Isaac Brodsky, 2018). The hexagon cell size was determined by considering walking distances and public transportation accessibility (Curie, 2010).

2.3 Ensemble Method

The prediction model of this study was assessed with ensemble techniques. Ensemble method is one of the well-known machine learning techniques that use computer learning algorithms and can improve predictive accuracy (Opitz and Maclin, 1999). Ensemble method has been widely used in recent transportation research. Zefreh et al. (2020) used random forest model to analyze passengers’ public transport satisfaction. Zhang and Haghani (2015) used a Gradient boosting regression model (GBM) to improve travel time prediction accuracy. Gradient boosting can improve predictive accuracy as it strategically combines additional models while correcting errors from previous models. In addition, GBM’s prediction accuracy was found to be higher than other ensemble models by comparing the random forest and ARIMA models with the mean absolute percent error (MAPE). Cho et al. (2019) used linear models (linear regression, Ridge regression, LASSO regression, Poisson generalization linear regression) and non-linear models (random forest, gradient boosting, support vector machine, extreme gradient boosting) to predict passenger numbers at Seoul subway stations. The collected data were divided into three groups using GMM (Gaussian Mixture Model) cluster analysis. By comparing the Root mean square error (RMSE) for each model, they derived a prediction result for the smallest RMSE for each group. Lee et al. (2020) constructed a LightGBM model for predicting highway accident severity. After comparing the performance of five machine learning models (LightGBM, CatBoost, XGBoost, AdaBoost, and Random Forest) with four indicators (precision, recall, accuracy, F-1 score) through a confusion matrix.

The indicators are predicted using ensemble techniques and statistical software R. Ensemble is a method of combining multiple learning algorithms to improve prediction accuracy. Types of ensemble methods include Bagging and Boosting.

Bagging is a method of creating multiple learning algorithms (decision trees) with randomly extracted data (bootstraps) and aggregating the average of the results generated to create the final model (Dietterich, 2000). The Random Forest model uses the bagging method. The Random Forest model can reduce correlation between models by randomly selecting explanatory variables when generating decision trees (Breiman, 2001). The importance of the variable in the Random Forest is MDA (Mean decrease in accuracy) when the dependent variable is nominal, and the percentage increase in Mean square error (%incMSE) when it is continuous. %incMSE can be calculated by $MSE_{\text{root}}$, $MSE_j$ (Eq. (1)):

$$%\text{incMSE} = \frac{MSE_j - MSE_{\text{root}}}{MSE_{\text{root}}} \times 100\%,$$

$MSE_{\text{root}}$ can be calculated as the mean square error (MSE) of OOB. OOB is the out of bag sample; Data not extracted by bootstrap. $MSE_{\text{root}}$ is shown in Eq. (2):

$$MSE_{\text{root}} = \sum_{i\in \text{OOB}} (y_i - \hat{y}_i)^2 / |\text{OOB}|,$$

where $y_i$ is the actual value in hexagonal cell ID $i$ and $\hat{y}_i$ is the predicted value in hexagonal cell ID $i$.

And $MSE_j$ is the MSE of predicted value by permuting the value of variable $j$. Equation of $MSE_j$ is shown in Eq. (3). (Grömping, 2009):

$$MSE_j = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}.$$  

Boosting is a method that iterates the process of creating a decision tree and creating a new model based on model accuracy weighting (Dietterich, 2000). Among various boosting methods, we used Extreme Gradient Boosting. Extreme Gradient Boosting (XGB) minimizes errors through loss functions and objective functions. The loss function represents the difference between the predicted value and the actual value. Regression analysis utilizes the Square Error to generate the loss function. The objective function measures the complexity of the model and can increase model accuracy by minimizing the loss function and the objective function (Chen and Guestrin, 2016). Gain determines the importance of Extreme Gradient Boosting variables. It is a numerical value of the information acquisition at the decision tree’s node branch point (variable). When $G$ is the sum of the gradient loss functions, $H$ is the sum of the Hessian loss functions, and $\gamma$ is the overfitting regulation parameter that controls the tree’s complexity (default 0, 20 when there are many variables), the gain score is shown in Eq. (4) (Chen and Guestrin, 2016):

$$Gain = \frac{1}{2 \alpha} \left[ \frac{G_{\text{H}}}{H_{\text{L}} + \lambda} + \frac{G_{\text{H}}}{H_{\text{R}} + \lambda} \right] - \gamma,\quad (4)$$

$L$ is the data divided to the left of the Leaf node and $R$ to the right.

In this study, Random Forest and Extreme Gradient Boosting models were used to analyze the results. Fig. 3 shows the flowchart of the Random forest model (Fig. 3(a)) and the Extreme gradient boosting model (Fig. 3(b)).

Divide the collected data into groups with and without public
bike stations. Data groups with public bike stations (“bike” > 0) are designed to produce predictive models through ensemble methods. Using this model, expected use numbers for areas without public bicycle stations (“bike” = 0) can be predicted. The higher the predicted value, the more suitable it is for PM. Flowchart of analysis modelling is shown Fig. 4.

2.4 Division Method
Natural break is a method used to divide visualized data by tier. Natural break is an algorithm that optimizes data values into natural classes according to the number of divided groups (Jenks, 1967; McMaster, 1997):
1. Specify number of groups and randomly specify each group’s center point.
2. Find nearest center point for each data set and assign group.
3. Recalculate assigned group’s center point.
4. Repeat steps 2 through 3 until optimized.

When visualizing GIS, natural breaks help minimize variance within the same group, maximize variance between classes, and provide an intuitive and clear view of the data compared to other methods (Chen et al., 2013).

3. Data
Previous studies have collected data on population, number of lanes, bus stops, etc. (Cho et al., 2019) to predict subway passenger numbers. PM usage intention studies show that land gradient and income levels effect PM usage (Choi and Jung, 2020). Based on this, we collected data that we expected to affect the PM use. First, we designated the dependent variable as the number of public bicycle users with characteristics most similar to PM (Zhu et al., 2020). And we collected related public transportation data (number of people using the subway, number of bus routes). The number of people using the subway and the number of bus routes can be used as an alternative variable that can describe the convenience of the public transportation (Radzimski and Dzięcielski, 2021). Also, we collected data such as population counts, foot traffic counts, land slopes, and income information (Cho et al., 2019; Choi and Jung, 2020). The data were collected from “Seoul Open Data” and “Data Portal” on a monthly basis from September to November 2019.

Table 1 shows descriptive statistics of variable. Variables that used in this study is the number of public bicycle users (“bike”), population (“pop”), foot traffic (“actpop”), number of people getting on and off the subway (“subway”), number of bus routes (“bus”), land slope (“slope”), income levels (“income”). The number of public bicycle users (“bike”) was calculated by summing the number of monthly public bicycle users by public bike station. Then, matched to the hexagonal cells corresponding
to public bicycle station coordinates. Population (“pop”) and foot traffic (“actpop”) were matched to corresponding hexagon cells by dividing the number of residents and foot traffic per dong (dong; smallest level of city division in South Korea) by the number of hexagon cells per dong. The number of people getting on and off the subway (“subway”) was calculated by summing to the number of passengers per month. Total population was differentially distributed around the hexagonal cells corresponding to subway station coordinates. The number of bus routes (“bus”) was determined using the number of routes passing by bus stops. Land slope (“slope”) and income levels (“income”), available as geographic information data, were matched with the hexagonal cells at corresponding locations.

4. Results and Discussion

4.1 Model Variable Importance, Precision Accuracy

Figure 5 shows the variable importance of Random forest (RF) (Fig. 5(a)) and Extreme gradient boosting model (XGB) (Fig. 4(b)). The importance of Random Forest variables was highest in the order of ‘slope’, ‘subway’, ‘actpop’, ‘pop’, ‘income’, and ‘bus’. The importance of Extreme Gradient Boosting variables was highest in the order of ‘subway’, ‘actpop’, ‘slope’, ‘pop’, ‘income’, and ‘bus’.

Random Forest and Extreme Gradient Boosting models’ prediction accuracy was compared based on the Root Mean Square Error (RMSE) and the coefficient of determination ($R^2$) (Cho et al., 2019). RMSE is the standard deviation of the predicted value, where the smaller the RMSE, the higher the regression model’s predicted accuracy. $R^2$ is the residual sum of squares (SSR) divided by the total sum of squares (SST; sum of squared deviations). This represents the variable’s explanatory power. Equation of RMSE and coefficient of determination is shown in Eqs. (5) and (6):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}},$$  
(5)

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y - \bar{y})^2}.$$  
(6)

As shown in Table 2, Random forest (RF) RMSE is 583.4824 and $R^2$ is 0.455. Extreme gradient boosting (XGB) RMSE is 495.4897 and $R^2$ is 0.709. Therefore, the Extreme gradient boosting model has statistically higher prediction accuracy than Random forest model.

Table 3 shows descriptive statistics of predictions of Extreme gradient boost model (XGB) and Random forest model (RF). Both Extreme Gradient Boosting and Random Forest models show no significant difference in mean, but the Extreme Gradient Boosting model variance is larger. Both models are statistically significant, but with better model descriptive accuracy and greater variance, the visualization of the Extreme Gradient Boosting model is expected to be more pronounced.
4.2 Visualization

The results of visualizing the entire Seoul area through the XGB model is shown in Fig. 6. The results illustrate the number of bicycles rented in red and the model’s predicted value in blue. Values increase as the color darkens. Five regions with the highest prediction value were selected and analyzed (presented in Table 4); the regions were Yeoui-dong, Jamsil 2-dong, Hwayang-dong, Seowon-dong, and Banpo 4-dong.

4.2.1 Yeouinaru, National Assembly Station, Yeoui-dong

Figure 7 shows the result of visualization in Yeouinaru, National Assembly Station, Yeoui-dong. Public bicycle stations were placed at Yeouinaru Station and nearby schools. Predictions were high for parks near the station. The business district near the National Assembly Station have a high foot traffic, and a high income level, so the number of PM users in the region is expected to be high.

4.2.2 Jamsil Sports Complex, Jamsil-dong

As shown in Fig. 8, Jamsil 2-dong has high predicted values for parks in front of schools, residential areas, and sports complexes. Since the sports complex has high foot traffic and is wide and flat, we expect many leisure PM operations.

4.2.3 Konkuk University Station, Hwayang-dong

As shown in Fig. 9, Hwayang-dong predicted high traffic near Konkuk University Station and Konkuk University. There are many narrow alleys with concentrations of nearby restaurants and houses, so it is expected that PM will be used a lot if placed near the subway station.

4.2.4 Sillim Station, Seowon-dong

As shown in Fig. 10, In front of Sillim Station in Seowon-dong, commercial districts are concentrated in narrow areas. Areas with high prediction values have lower slopes than neighboring

---

Table 3. Descriptive Statistics of Predictions

|          | XGB               | RF                |
|----------|-------------------|-------------------|
| Mean     | 1098.895          | 1093.772          |
| Std. Err | 6.676984          | 3.597032          |
| Median   | 1058.619          | 985.7642          |
| Mode     | 1062.342          | 985.7642          |
| Std. Dev | 860.3987          | 463.5149          |
| Variance | 740286            | 214846.1          |
| Min      | −1800.06          | 341.6733          |
| Max      | 12434.01          | 5561.668          |
| Observation | 16605          | 16605             |

Table 4. Maximum Prediction Value by Administrative dong

| Rank | Administrative dong | Maximum prediction value (Monthly number of users) |
|------|---------------------|--------------------------------------------------|
| 1    | Yeoui-dong          | 12434.01074                                     |
| 2    | Jamsil 2-dong       | 10882.83594                                     |
| 3    | Hwayang-dong        | 10405.56836                                     |
| 4    | Seowon-dong         | 8953.405273                                     |
| 5    | Banpo 4-dong        | 7945.851074                                     |
| 6    | Shinlim-dong        | 7699.766113                                     |
| 419  | Hongje 2-dong       | 1143.926025                                     |
| 420  | Sangbong 1-dong     | 1105.709473                                     |
| 421  | Songcheon-dong      | 1083.062866                                     |
| 422  | Mok 3-dong          | 1070.649536                                     |
| 423  | Jangwi 3-dong       | 1016.612183                                     |

---

Fig. 6. Ensemble Model Visualization (Seoul)
areas, making it suitable for PM placement.

4.2.5 Central City Terminal, Banpo-dong
As shown in Fig. 11, Banpo 4-dong showed high prediction areas near department stores and residential in front of the terminal. Since there are many apartment complexes, the population density is high, and department stores and bus terminals mean there is high foot traffic.
5. Conclusions

The previous study found that it is difficult to perform detailed analyses by dividing the city into units (Mollalo et al., 2020). Also, it was found to be difficult to compare and analyze neighboring cells using a square grid structure and to intuitively visualize divisions using colors or three-dimensional renderings (Enticott et al., 2020; Lodha and Verma, 2000; Choi et al., 2021). In this study, we used a hexagonal grid to facilitate comparative analyses between neighboring cells, and to visualize single color saturation via natural interval breaks to understand the results more intuitively. Previous studies predicted values and derived variable importance through linear and nonlinear models for future development direction studies. In this study, specific PM service areas are visualized using a hexagonal grid through ensemble models. This allows an intuitive view of the optimal PM service deployment and estimates demand by deriving the expected number of users. This paper aims to find the optimal PM service areas using GIS-based analysis of public transportation big data. Location variables were generated by collecting socio-economic factor data, public transportation data, and geographic data. The model was derived using a shared bicycle ensemble method since they are most similar characteristically to PM devices. The model was developed by dividing Seoul into a hexagonal grid and entering the data into hexagonal cells.

In this study, we analyzed the optimal personal mobility placement areas using a hexagonal grid structure and ensemble models. The ensemble models’ prediction accuracy was compared using RMSE and R2; the Extreme Gradient Boosting model was selected for visualization. The hexagonal grid was used to visualize the area, and the geographical characteristics of the PM introduction were analyzed based on five administrative-dongs with high model prediction values. Optimal placement areas often appear near subway stations.

Fig. 10. Sillim Station, Seowon-dong

Fig. 11. Central City Terminal, Banpo-dong
Residential areas and parks, which have high foot traffic, are also good placement areas. This is because subway location and high foot traffic are important variables. Next, this study shows it is useful to deploy PM devices in areas with lower slopes than areas with higher slopes. This reflects public bicycle stations placed on flat land rather than on sloping areas. High placement variables are also located in the vicinity of metropolitan transportation facilities. PM device placement in these areas will increase connectivity with other modes of public transportation such as buses and subways.

Limitation of this study is that the analyzed data was from areas with high public transport accessibility; areas with low public transport accessibility were not considered. These areas should also be considered to improve mobility of the city. To solve this problem, select areas with low accessibility to public transportation first, and apply this study’s results to that area (Litman, 2008). This study also did not distinguish use characteristics by time or weather condition, so it did not take into account the difference between commuting time and the rest of the day. Also did not consider whether the weather was bad or good. We expect more efficient operations are possible if the service condition is considered (Barr, 2018). And the model is designed for bicycles similar to PM devices, which can produce results different from actual PM characteristics. Using PM data, the optimal region can be derived more accurately. Future studies can subdivide the hexagonal cells into areas with smaller radii, leading to more accurate PM placement. Our model can also help determine new public transport stop locations and be used for public transport demand surveys and accessibility analyses.

Acknowledgments

This research was supported by research project “Development of Sustainable MaaS (Mobility as a Service) 3.0+ Technology in Rural Areas” funded by the Korea Institute of Civil Engineering and Building Technology (KICT).

ORCID

Sangjae Lee https://orcid.org/0000-0003-1089-3220
Seung-oh Son https://orcid.org/0000-0002-7442-6907
Juneyoung Park https://orcid.org/0000-0002-1598-3367
Jaehong Park https://orcid.org/0000-0001-9167-4542

References

Barr S (2018) Personal mobility and climate change. Wiley Interdisciplinary Reviews: Climate Change 9(5):e542, DOI: 10.1002/wcc.542
Breiman L (2001) Random forests. Machine Learning 45(1):5-32, DOI: 10.1023/A:1010933404324
Campbell AA, Cherry CR, Ryerson MS, Yang X (2016) Factors influencing the choice of shared bicycles and shared electric bikes in Beijing. Transportation Research Part C: Emerging Technologies 67:399-414, DOI: 10.1016/j.trc.2016.03.004
Chang K-T (2004) Introduction to geographic information systems. McGraw-Hill Higher Education, Boston, MA, USA
Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, August 13-17, San Francisco, CA, USA, DOI: 10.1145/2939672.2939785
Chen J, Yang S, Li H, Zhang B, Lv J (2013) Research on geographical environment unit division based on the method of natural breaks (Jenks). The International Archives of the Photogrammetry. Remote Sensing and Spatial Information Sciences 3:47-50, DOI: 10.5194/isparchive-XL-4-W3-47-2013
Cho S, Kim B, Kim N, Song J (2019) A study on the number of passengers using the subway stations in Seoul. The Korean Journal of Applied Statistics 32(1):111-128, DOI: 10.5351/KJAS.2019.32.1.111
Choi Y, Jeung I, Park J (2021) Comparative analysis of spatial impact of living social overhead capital on housing price by residential type. KSCE Journal of Civil Engineering 25(3):1056-1065, DOI: 10.1007/s12205-021-1250-z
Choi MH, Jung HY (2020) A study on the influencing factor of intention to use personal mobility sharing services. Journal of Korean Society of Transportation 38(1):1-13, DOI: 10.7470/jkst.2020.38.1.001
Currie G (2010) Quantifying spatial gaps in public transport supply based on social needs. Journal of Transport Geography 18(1):31-41, DOI: 10.1016/j.jtrangeo.2008.12.002
Dietterich TG (2000) An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning 40(2):139-157, DOI: 10.1023/A:1007607513941
Dowling R, Irwin J, Faulks I, Howitt R (2015) Use of personal mobility devices for first-and-last mile travel: The Macquarie-Ryde trial. In: Cameron I, Haworth N, McIntosh L (eds) Proceedings of the 2015 Australasian road safety conference (ARSC2015). Australasian College of Road Safety (ACRS), Pearce, Australia
Enticott G, Ward K, Ashton A, Brunton L, Broughan J (2020) Mapping the geography of disease: A comparison of epidemiologists’ and field-Level experts’ disease maps. Applied Geography 126
Grömping U (2009) Variable importance assessment in regression: Linear regression versus random forest. The American Statistician 63(4): 308-319, DOI: 10.1198/tast.2009.08199
Isaac Brodsky (2018) H3: Uber’s Hexagonal hierarchical spatial index. Uber Engineering, Retrieved June 27, 2018, https://eng.uber.com/h3/
Jenks GF (1967) The data model concept in statistical mapping. The American Statistician 21(3):136-144, DOI: 10.1080/00031306.1967.1047020
Lee HM, Jeon GS, Jang JA (2020) Predicting of the severity of car traffic accidents on a highway using light gradient boosting model. The Journal of the Korea Institute of Civil Engineering 33(1):5-17, DOI: 10.1559/15230743120190704-10702, DOI: 10.1016/j.ensoft.2020.104702
Litman T (2008) Evaluating accessibility for transportation planning. Victoria Transport Policy Institute, Victoria, BC, Canada
Lodha SK, Verma AK (2000) Spatio-temporal visualization of urban crimes on a GIS grid. Proceedings of the 8th ACM international symposium on advances in geographic information systems, November 6-11, Washington DC, USA, DOI: 10.1145/355274.355300
McMaster R (1997) In memoriam: George F. Jenks (1916-1996). The American Statistician 51(4):246-247, DOI: 10.1080/00031305.1997.10476338
McMaster R (1997) In memoriam: George F. Jenks (1916-1996).
Miller P, de Barros AG, Kattan L, Wirasinghe SC (2016) Public...
transportation and sustainability: A review. *KSCE Journal of Civil Engineering* 20(4):1076-1083, DOI: 10.1007/s12205-016-0705-0

Mollalo A, Vahedi B, Rivera KM (2020) GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Science of the Total Environment* 728:138884, DOI: 10.1016/j.scitotenv.2020.138884

Murray AT, Davis R, Stimson RJ, Ferreira L (1998) Public transportation access. *Transportation Research Part D: Transport and Environment* 3(5):319-328, DOI: 10.1016/S1361-9209(98)00010-8

Opitz D, Maclin R (1999) Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research* 11:169-198, DOI: 10.1613/jair.614

Park S, Kang J, Choi K (2014) Finding determinants of transit users’ walking and biking access trips to the station: A pilot case study. *KSCE Journal of Civil Engineering* 18(2):651-658, DOI: 10.1007/s12205-014-0073-6

Radzimski A, Dzięcielski M (2021) Exploring the relationship between bike-sharing and public transport in Poznań, Poland. *Transportation Research Part A: Policy and Practice* 145:189-202, DOI: 10.1016/j.tra.2021.01.003

Zefreh MM, Hussain B, Sipos T (2020) In-depth analysis and model development of passenger satisfaction with public transportation. *KSCE Journal of Civil Engineering* 24(10):3064-3073, DOI: 10.1007/s12205-020-1871-7

Zhang Y, Haghani A (2015) A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies* 58:308-324, DOI: 10.1016/j.trc.2015.02.019

Zhu R, Zhang X, Kondor D, Santi P, Ratti C (2020) Understanding spatio-temporal heterogeneity of bike-sharing and scooter-sharing mobility. *Computers, Environment and Urban Systems* 81:101483, DOI: 10.1016/j.compenvurbsys.2020.101483