JOURNAL OF ADVANCED TRANSPORTATION
J. Adv. Transp. 2016; **50**:769–785
Published online 4 April 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/atr.1374

Estimating vehicle miles traveled (VMT) in urban areas using regression kriging

Seheon Kim[1], Dongjoo Park[1]*, Tae-Young Heo[2], Hyunseung Kim[1] and Dahee Hong[3]

[1] *Department of Transportation Engineering, University of Seoul, 90 Jeonnong-dong, Dongdaemun-gu, Seoul, South Korea*
[2] *Department of Information and Statistics, Chungbuk National University, 52 Naesudongro, Heungdeok-gu, Cheongju, South Korea*
[3] *Center for Transport Database, The Korea Transport Institute, 2311 Daehwa-dong, Ilsanseo–gu, Goyang, South Korea*

SUMMARY

The recent increase in demand for performance-driven and outcome-based transportation planning makes accurate and reliable performance measures essential. Vehicle miles traveled (VMT), the total miles traveled by all vehicles on roadways, has been utilized widely as a proxy for traffic impact assessment, vehicle emissions, gasoline consumption, and crashes. Accordingly, a number of studies estimate VMT using diverse data sources. This study estimates VMT in the urban area of Bucheon, South Korea, by predicting the annual average daily traffic for unmeasured locations using spatial interpolation techniques (i.e., regression kriging and linear regression). The predictive performance of this method is compared with that of the existing Highway Performance Monitoring System (HPMS) method. The results show that regression kriging could provide more accurate VMT estimates than the HPMS method and linear regression, especially with a small sample size. Copyright © 2016 John Wiley & Sons, Ltd.

KEY WORDS: vehicle miles traveled (VMT); Highway Performance Monitoring System (HPMS); regression kriging; spatial interpolation

1. INTRODUCTION

The recent increase in the deployment of Intelligent Transportation Systems (ITS) technologies in urban areas has reinforced the necessity of adopting performance-based transportation decision-making. The USA's newly enacted transportation law (MAP-21) establishes the requirements for both performance-based transportation decision-making and performance measures for congestion reduction and system reliability [1]. Vehicle miles traveled (VMT), the total miles traveled by all vehicles on roadways, is one of the most important performance measures as it can be utilized as a strong proxy for traffic impact assessment, vehicle emissions, gasoline consumption, and crashes [2, 3]. It has been widely used in transport planning [4], travel demand analysis [5–7], traffic crash analysis [8–10], and energy consumption analysis [11–13].

Given the demand for accurate and stable VMT estimates, a number of methods and data sources have been proposed. These can be broadly classified into traffic-count-based methods and non-traffic-count-based methods [14]. Non-traffic-count-based methods utilize non-traffic data such as, but not limited to, household activity surveys, odometer recordings, and fuel sales. Kweon and Kockelman [15] employ a nonparametric regression (NPR) to identify variations in household VMT using the 1995 Nationwide Personal Transportation Survey, which includes data on vehicle ownership, housing type, household income, public transit availability, and residential area (i.e. residential or urban). The results show that the goodness-of-fit of the NPR is substantially improved relative to that of ordinary least squares, although it requires time-consuming tasks and significantly larger sample sizes for

*Correspondence to: Dongjoo Park, Professor, Department of Transportation Engineering, University of Seoul, 90 Jeonnong-dong, Dongdaemun-gu, Seoul, South Korea. E-mail: djpark@uos.ac.kr additional control variables. Pathomsiri *et al.* [16] develop an econometric model for estimating VMT in multivehicle households and apply it to the 2001 National Household Transportation Survey. Erlbaum [17] applied the fuel-sales-based method to estimate VMT. White [18] and Greene [19] used odometer reading data to estimate annual VMT by driver. However, the previously mentioned methods, based on household activity surveys, fuel sales, and odometer reading data, are too resource-intensive and costly to perform regularly. Also, these methods reflect personal travel and do not reflect the total number of vehicles on the road.

Traffic-count-based VMT estimation methods are currently the most commonly performed and preferred method because they are based on actual data for vehicle movement [20]. The predictive accuracy of traffic-count-based methods depends on the quality and coverage of traffic count data, given that the length of all road network sections is known. Thus, it can be said that if traffic counts were available for all roads in a network, the VMT estimate would be the most accurate measure of vehicle movement. Traffic counts are, however, only available at segments of road networks where there is a count station because of the relatively high cost of the stations. For urban areas, traffic count data are collected less frequently than in other areas because of the complexity of measurement [21]. For this reason, the Highway Performance Monitoring System (HPMS) method of VMT estimation, a representative traffic-count-based method performed in the USA, simply extrapolates the VMT of a sample section into other sections, so long as the other sections are in the same strata of traffic volume group and road functional system [20]. Despite the advantages of being a relatively simple and quick procedure, the HPMS method has been criticized by its accuracy and sample size demand needed to achieve the required precision level. Moreover, stratification of sampling by traffic volume group requires knowledge of the traffic count information on all roads [14]. A number of recent studies attempt to overcome these limitations by using various data sources. Teng and Wang [3] examine the applicability of ITS, such as cameras and loop detectors, in estimating VMT instead of using the HPMS short-term counts. Their study tests the conditions under which ITS daily traffic count data can be used to replace HPMS short-term counts in terms of a threshold number of missing ITS data. Zhang and He [22] use the global positioning system (GPS) and other supplemental data sources to estimate VMT. Their results suggest that GPS-based surveys are feasible for VMT estimation on the different functional classes, including local roads where the ground-truth data are scarce. Blei *et al.* [23] also explore the VMT estimation methods using GPS data. In the context of the deficiencies in the existing HPMS methodology, this study investigates the improvement of predictive accuracy by using a state-of-the-art interpolation technique instead of simply extrapolating the sample section VMT into area-wide VMT. In addition, the stratification for the sampling process could be replaced by using non-volume data (e.g., functional class).

Interpolation techniques can be classified into three categories: statistical analysis (e.g., multiple linear regression), geostatistical interpolation (e.g., ordinary kriging (OK) and universal kriging (UK)), and hybrid techniques (e.g., regression kriging (RK) and kriging with external drift) [24, 25]. A number of studies employ interpolation techniques to improve predictions of traffic volume at unmeasured locations based on limited data. Many researchers introduce a regression model to predict current-year annual average daily traffic (AADT), assuming that one or more auxiliary variables are correlated with traffic volume. They develop a multiple-regression model including socioeconomic variables such as population, automobile ownership, household income, and employment as predictors [26–28]. Lam and Xu [29] examine the estimation accuracy of AADT between a regression model and a neural network approach from short-period traffic count data; their results suggest that the latter performed better. Lam *et al.* [30] compare four models—auto-regressive integrated moving-average, neural network, NPR and Gaussian maximum likelihood—to predict hourly traffic flows in Hong Kong, suggesting that the NPR model is likely to account for unexpected changes more effectively. However, these regression models have been criticized because of their unreasonable assumption that all observations are independently drawn from a certain probability distribution, even for places for which spatial correlation exists [31]. Improved geostatistical interpolation techniques, based on the assumption that distributed objects are spatially correlated, have been applied in order to provide more reliable predictions for the missing data. Eom *et al.* [31] apply UK to estimate AADT and suggest that the spatial regression model outperformed the ordinary regression. Wang and Kockelman [32] use OK to spatially interpolate the AADT values for unmeasured sites using the AADT estimates at urban saturation

DOI: 10.1002/atr

traffic count stations in Texas. Selby and Kockelman [33] use UK to predict the traffic count for Texas and compare the results with those from a geographically weighted regression and spatial regression techniques. In addition, they suggest that the simple Euclidean distance-based method predicted as well as the network-based method, implying that the latter's complexity is not necessary in real-world applications. RK involves various combinations of linear regressions and kriging, and it is a suitable technique for predicting a primary variable when the explanatory variables are available at all locations and correlated with the target variable [34]. The RK method is defined as follows. Linear regression models are constructed prior to OK on the regression residuals. OK is performed on the residuals of the selected linear regression models. RK gives more reliable results when the auxiliary information explains a significant amount of the variation in the target variable [25, 35]. Because the latter applies to the present study, the RK method is used as the interpolation method.

In light of the deficiencies of current VMT estimation methods, this study proposes a traffic-count-based VMT estimation method for urban areas by adopting interpolation techniques (i.e., RK) and by comparing the prediction accuracy with that of the HPMS method. Moreover, given that the sampling strategy is a critical factor affecting the predictive accuracy of both the interpolation and HPMS methods, the comparison analysis considers both sample size and strata classification. This comparison analysis can provide a better understanding of measuring VMT in more robust and consistent ways.

The remainder of this study is organized as follows. In the next section, the theoretical foundations of the VMT estimation methods are presented. Section 3 describes the data and experimental design. Section 4 compares the results of the spatial interpolation using RK with those obtained from the existing HPMS method. The final section summarizes the main conclusions, implications, and extensions of this study.

2. METHODOLOGY

This section provides the VMT estimation methods that implement RK and the existing HPMS procedure. Although both of these methods are traffic-count-based, a methodological difference exists between them. In order to acquire the value of VMT where traffic count data do not exist, the HPMS method extrapolates the sample section's VMT into area-wide VMT using an expansion factor, whereas the RK spatially interpolates the traffic volume (e.g. AADT) at unmeasured locations.

2.1. *Dividing roadway into unit links*

In general, geostatistical methods are applied to point-referenced spatial data, where the target variable is measured at specific locations [31]. Given that the VMT value, which uses linear referenced roadway segment data (e.g. length), is an indicator of roadway system performance, it is necessary to split the roadway into individual links. Each individual link should contain a non-overlapping roadway segment and homogeneous traffic flow characteristics. Each link therefore experiences the same level of traffic volume. This individual link is referred to as a "unit link" in this study. Each unit link has its own representative, auxiliary variables and spatial coordinates. Similarly, in the HPMS method, a sampling unit is generated for roadways based on the geospatial intersection of five key data items (AADT, functional systems, urban code, through lanes, and facility type), where their respective values are homogenous along a given roadway [20]. This sampling unit is called a Table of Potential Samples (TOPS). In this study, in order to divide roadways into unit links, arterial roads are split into unit links at each signalized intersection. The expressways and urban expressways are divided into unit links at each junction or interchange.

2.2. *HPMS method*

This study provides a simple overview of the HPMS VMT calculation process for urban areas. The detailed calculation process is contained in the HPMS field manual, which includes the calculation of statewide and universal expansion factors [20]. In order to estimate VMT in urban areas, HPMS samples are chosen from the TOPS and stratified into a set of functional classes and AADT volume groups. Stratification is performed to improve the precision of the estimates without significantly increasing the sample size [20]. The required sample size for each stratum is estimated using Equation (1):

DOI: 10.1002/atr

$$n = \frac{\left(\frac{Z^2c^2}{d^2}\right)}{1 + \left(\frac{1}{N}\right)\left(\left(\frac{Z^2c^2}{d^2}\right) - 1\right)} \tag{1}$$

where n is the required sample size, Z is the value of the standard normal critical value for an alpha confidence level, c is the AADT coefficient of variation from a state's AADT data, d is the desired precision rate, and N is the TOPS or population stratum size (the number of TOPS sections available for sampling in each stratum).

Sample section VMT estimates are obtained by multiplying the sample section AADT to the section length. Expansion factors are then applied to these estimates in order to extrapolate the sample section data to area-wide (or city-wide) VMT estimates. The expansion factors are obtained with Equation (2):

$$EF_{ij} = \frac{(\text{Total length of all road segments})_{ij}}{(\text{Total length of the sample sections})_{ij}} \tag{2}$$

where EF_{ij} is the expansion factor for AADT volume group i in functional class j. The combined length of all road segments for each AADT volume group in each functional class is divided by the total length of the sample sections in each AADT volume group of each functional class.

The area-wide VMT is calculated by applying an expansion factor, as shown in Equation (3):

$$\text{DVMT} = \sum_i \sum_j \sum_k \text{DVMT}_{ijk} \times EF_{ij} \tag{3}$$

where DVMT is the area-wide daily VMT estimate in urban areas, DVMT_{ijk} is the daily VMT estimate for sample section k in group i for class j, and EF_{ij} is the expansion factor for group i in class j.

2.3. Regression kriging

Let $Z(s)$ represent the realization of a random process in two-dimensional Euclidean space with spatial location $(x, y) \in s$. In order to accurately model spatially correlated data, the random process $Z(s)$ should be considered a stationary random process. A spatial random field is a real valued stochastic process $\{Z(s) : s \in D \subset R^2\}$, where D, the area of interest, is a fixed subset of R^2. The feasibility of statistical inference on single realization of a random field as well as construction of optimal predictors is based on a notion of some form of stationarity [36]. Assumptions of stationarity allow values at different places to be different realizations of the property [37]. There are two types of stationarity assumptions for random fields. One is the second-order stationarity satisfying $E[Z(s)] = \mu$ and $\text{Cov}[Z(s) - Z(s + h)] = C(h)$, while the other is the intrinsic stationarity satisfying $E[Z(s)] = \mu$ and $\text{Var}[Z(s) - Z(s + h)]/2 = \gamma(h)$, where C and γ, called the covariance function and semivariogram, respectively, depend only on distance h. A set of measurements $\{Z(s_1), \ldots, Z(s_n)\}$ at known locations can be obtained. A spatial structure model for the random field $Z(s)$ may be modeled as in Equation (4):

$$Z(s) = X(s)^T\beta + \varepsilon(s), \text{ for } s \in D \tag{4}$$

where $X(s)^T\beta$ is the large-scale variation or mean function and $\varepsilon(s)$ is the small-scale stochastic variation.

Semivariogram analysis is used for descriptive analysis in this study. The spatial structure of the data is investigated using the semivariogram. This structure is also used for predictive applications, in which the semivariogram is fitted to a theoretical model, parameterized, and used to predict the regionalized variable at other unmeasured points. Estimating the mean function $X(s)^T\beta$ and the covariance structure of $\varepsilon(s)$ for each s in the area of interest is the first step in both analyzing the spatial variation and prediction. One commonly used measure of spatial dependence is the semivariogram. The estimated semivariogram provides a description of how the data are correlated with distance. The factor

1/2 in $\gamma(h)$ indicates it is a semivariogram, and $2\gamma(h)$ is the variogram. Thus, the semivariogram function measures half the average squared difference between pairs of data values separated by a given distance, h, which is known as the lag [38]. Assuming that the process is stationary, the semivariogram is defined in Equation (5):

$$\gamma(h) = \frac{1}{2N_h} \sum_{N(h)} \left[z(s_i) - z(s_j)\right]^2 \tag{5}$$

where $N(h)$ is the set of all pairwise Euclidean distances $i - j = h$, N_h is the number of distinct pairs in $N(h)$, and $z(s_i)$ and $z(s_j)$ are the value at spatial location i and j, respectively, and $\gamma(h)$ is the estimated semivariogram value at distance h.

The main purpose of semivariogram analysis is to construct a semivariogram that accurately estimates the autocorrelation structure of the underlying stochastic process. The semivariogram has three important parameters: the nugget, sill, and range. The nugget is the sub-grid-scale variation or measurement error and is indicated graphically by the intercept of the semivariogram. The sill is the value of the semivariance as the lag (h) goes to infinity, and it is equal to the total variance of the data set. The range is a scalar that controls the degree of correlation between data points (i.e., the distance at which the semivariogram reaches its sill). As shown in Figure 1, the shape of semivariogram is typically characterized in terms of the nugget, sill, and range.

When a valid empirical estimate of the semivariance is obtained, it is then necessary to select a type of theoretical semivariogram model based on that estimate. Commonly used theoretical semivariogram shapes increase monotonically as a function of distance. The most appropriate semivariogram model is chosen by plotting the empirical semivariogram and comparing it with various theoretical models. In this study, the following three parametric semivariogram models are tested: exponential, Gaussian, and spherical. These models are given by the following equations:

$$\text{Exponential model: } \gamma(h) = \theta_0 + \theta_1\left\{1 - \exp\left(-\frac{3h}{\theta_2}\right)\right\}, \tag{6}$$

$$\text{Gaussian model: } \gamma(h) = \theta_0 + \theta_1\left\{1 - \exp\left(-3\left(\frac{h}{\theta_2}\right)^2\right)\right\}, \text{ and} \tag{7}$$

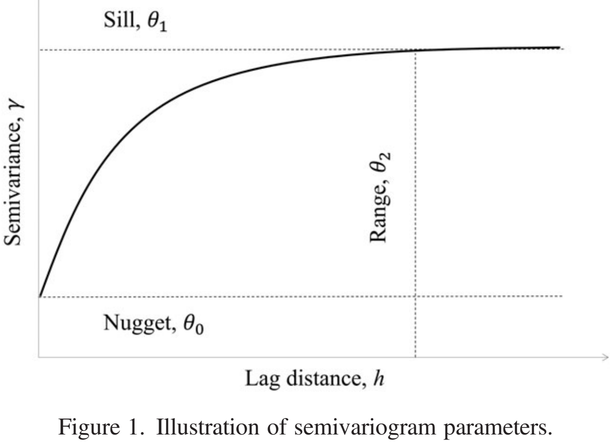


Figure 1. Illustration of semivariogram parameters.

DOI: 10.1002/atr

$$\text{Spherical model: } \gamma(h) \;=\; \begin{cases} \theta_0 \;+\; \theta_1\left\{1 \;-\; \exp\left(\dfrac{3h}{2\theta_2} - \dfrac{1}{2}\left(\dfrac{h}{\theta_2}\right)^3\right)\right\}, \; 0 \leq h \leq \theta_2 \\ \theta_0 \;+\; \theta_1, \; h > \theta_2 \end{cases} \tag{8}$$

where h is a spatial lag, θ_0 is the nugget, θ_1 is the spatial variance (also referred to as the sill), and θ_2 is the spatial range. The nugget, sill, and range parameters of the theoretical semivariogram model can be fit to the empirical semivariogram $\gamma(h)$ by minimizing the nonlinear function. When fitting a semivariogram model, if we consider the empirical semivariogram values as our "observations" and try to fit a model to them as a function of the lag distance h, the ordinary least squares function is as given by $\sum_h [\hat{\gamma}(h) \;-\; \gamma(h \;:\; \theta)]^2$, where $\gamma(h : \theta)$ denotes the theoretical semivariogram model and $\theta = (\theta_0, \theta_1, \theta_2)$ is a vector of parameters.

Regression kriging computes the parameters θ and β separately. The parameters β in the mean function are estimated by the least squares method. The residuals are then computed, and their parameters in the semivariogram are estimated by various estimation methods, such as least squares or a likelihood function. Prediction of RK at a new location s_0 can be performed separately using a regression model to predict the mean function and a kriging model of prediction residuals and then adding them back together as in Equation (9):

$$Z(s_0) \;=\; \sum_{k=0}^{\mathrm{n}} \beta_k X_k(s_0) \;+\; \sum_{i=0}^{\mathrm{n}} \lambda_i(s_0)\, \varepsilon(s_i) \tag{9}$$

where $s_i = (x_i, y_i)$ is the known location of the ith sample, x_i and y_i are the coordinates, β_k is the estimated regression model coefficient, λ_i represents the weight applied to the ith sample (determined by the variogram analysis), $\varepsilon(s_i)$ represents the regression residuals, and $X_1(s_0) \ldots X_n(s_0)$ are the values of the explanatory variables at a new location s_0. The weight λ_i is chosen such that the prediction error variance is minimized, yielding weights that depend on the semivariogram [34]. More details about the kriging weight λ_i follow immediately [39].

The main objective is to predict $Z\,(s)$ at a known location s_0, given the observations $\{Z\,(s_1), Z\,(s_2), \ldots, Z\,(s_3)\}'$. For simplicity we assume $E\{Z\,(s)\} = 0$ for all s. We briefly outline the derivation of the widely used kriging predictor. Let the predictor be of the form $\hat{Z}\,(s_0) \;=\; \lambda'\, Z(s)$, where $\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}'$. The objective is to find weights λ, such that

$$Q(s_0) \;=\; E[\lambda'\, Z(s) - Z(s_0)]^2 \tag{10}$$

is a minimum. By minimizing $Q(s_0)$ with respect to λ, it can be shown that $\hat{Z}\,(s_0) \;=\; \sigma'(s_0,\, s)\overset{-1}{\sum} Z(s)$, where $\sigma'\,(s_0, s) = E(Z(s_0)\, Z(s))$, and $\sum = E[Z(s)\, Z'\,(s)]$ is the covariance matrix. The minimum of $Q(s_0)$ is min $Q(s_0) \;\;=\; \sigma^2 - \sigma'(s_0,\, s)\overset{-1}{\sum}\sigma(s_0,\, s)$. Note that $Q(s_0)$ can be rewritten in terms of the variogram by applying

$$\sigma(s_0,\, s) \;=\; \sigma^2 1 - \frac{1}{2}\Gamma(s_0,\, s) \tag{11}$$

where $\Gamma(s_0, s)$ is the corresponding matrix of variograms. We can thus rewrite $Q(s_0)$ given in Equation (10) as

$$Q(s_0) \;=\; -\frac{1}{2}\lambda'\,\Gamma\,\lambda + \;\lambda'\,\Gamma(s_0,\, s). \tag{12}$$

$Q(s_0)$ is now minimized with respect to λ, subject to the constraint $\lambda'\;1 = 1$ (accounting for the unbiasedness of the predictor $\hat{Z}\,(s_0)$), by noting that

$$Q'(s_0) = Q(s_0) - m(\lambda' 1 - 1). \tag{13}$$

Differentiating $Q'(s_0)$ with respect to λ yields

$$\lambda' = \left(\gamma + 1\frac{1 - 1'\Gamma^{-1}\gamma}{1'\Gamma^{-1}1}\right)\Gamma^{-1}. \tag{14}$$

2.4. Vehicle miles traveled calculation

When the traffic volumes of unknown unit links are available for all roadways using RK, daily VMT (DVMT) can be calculated by multiplying the unit link AADT by the centerline mileage of a unit link as shown in Equation (15):

$$\text{DVMT} = \sum_{i=1}^{n}(\text{AADT}_i \times L_i), \tag{15}$$

where DVMT is the daily VMT estimate, AADT_i is the annual average daily traffic for unit link i, L_i is the centerline mileage of a unit link i, and n is the total number of unit links in the study area.

3. DATA AND EXPERIMENTAL DESIGN

3.1. Data description

The data used for the study were collected in 2011 in Bucheon (near Seoul), South Korea. The population of Bucheon is around 889 500, and it covers an area of 53.4 km^2. The data utilized in this study are taken from only the top three functional classes (expressways and urban expressways, principal arterials, and minor arterials) of the roadway. They run along the centerline of the roadway for approximately 80 km, as shown in Table I.

This study uses two types of data: traffic count data and road information. In the RK, traffic count is the dependent variable, and road information is used as explanatory variables. The traffic count data for principal arterials and minor arterials are obtained from short-period traffic counts during the four months from June 2011 to September 2011. In order to estimate AADT at arterials, short-period traffic counts data are factored up using a set of factors. The hourly, monthly, and weekly factors are developed based on the permanent traffic counts [40]. The AADT data for expressways and urban expressways are provided annually by the Korea Institute of Construction Technology. The AADT data were obtained at 127 points on unit links among the 150 total unit links as described in Figure 2 and were considered as actually observed value for spatial interpolation approach in this study. Because the AADT data have large values and their distributions are right-skewed, log-transformed AADT values are used in the regression analysis. This removes the skewness, and the log-transformed data are symmetrically distributed, as shown in Figure 3.

Table II shows the subsets of road information obtained from the GIS database, the Korea Transport Database, provided by the Korea Transport Institute. Road information is available for 150 total unit

Table I. Road network in study area.

Road functional class	Total length (km)	Number of unit links	Average unit link length (km)
Expressways and urban expressways	8.85	5	1.77
Principal arterials	19.20	31	0.62
Minor arterials	51.26	114	0.45
Total	79.31	150	0.53

DOI: 10.1002/atr

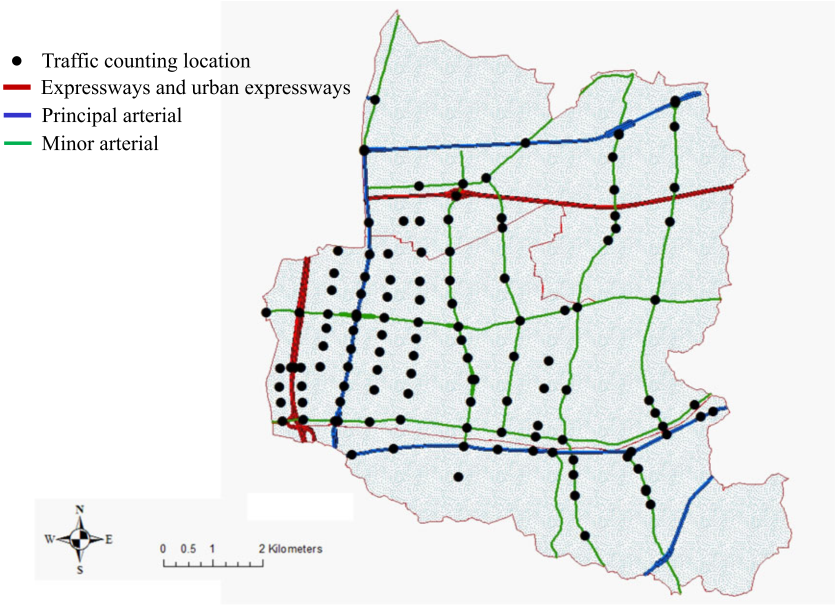

● Traffic counting location
Expressways and urban expressways
Principal arterial
Minor arterial
N
W
E
S
0 0.5 1 2 Kilometers

Figure 2. Traffic counting locations and roadways for VMT estimation in Bucheon.

links and includes spatial coordinates, number of lanes, speed limit, density of signalized intersection, road functional class, land use type, and contact with the major arterials. The density of signalized intersection is measured as the number of signalized intersections within 1 km of an arterial corridor. Contact with the major arterials is a binary variable indicating whether the major arterial crosses the unit link or not. The categorical variables road functional class and land use type are divided into sub-classes and converted to indicator variables characterizing each unit link. Road functional class 1 and land use type 4 are used for the benchmark case in the regression analysis.

3.2. Experimental design

A set of scenarios for the estimation method and sampling strategy (e.g., strata classification and sample size) were set for the comparison analysis, as shown in Figure 4. First, this study applies three VMT estimation methods: HPMS, linear regression, and RK. Second, in order to improve the representativeness of the sample by reducing the sampling error, stratified sampling is used when applying the Monte Carlo method to estimate population statistics with 10 different random seeds. For the strata classification, road functional class is compared with the results stratified by AADT group. In HPMS, stratification by a defined set of AADT groups for each road functional class is performed to improve

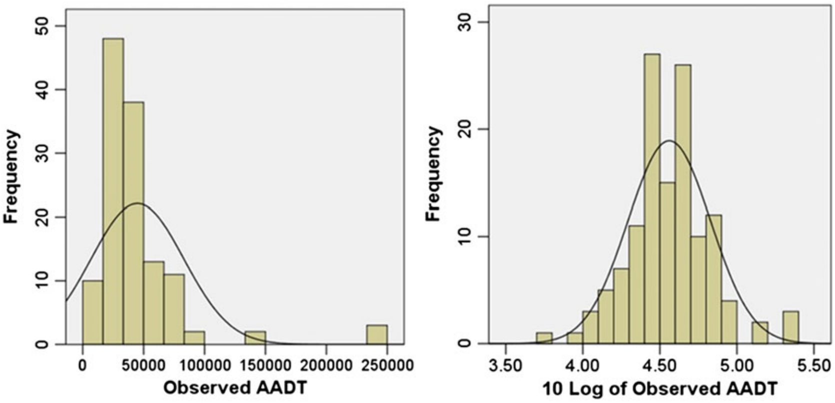


Figure 3. Histogram of original and log-transformed AADT.

DOI: 10.1002/atr

Table II. Summary statistics of variables used in the study.

Variables	Description	Mean	Standard deviation	Min	Max
Traffic count data 10 log_AADT *Road information*	- log-transformed AADT value	4.56	0.27	3.80	5.39
X coordinate	- X coordinate (m)	292 774	1858	296 006	289 064
Y coordinate	- Y coordinate (m)	545 295	2214	550 363	540 967
Number of lanes (two-way)	- 1: 2-lane, 2: 4-lane, 3: 6-lane, 4: 8-lane, 5: more than 10-lane	3.07	1.33	1	5
Speed limit	- 1: 60 km/h, 2: 70 km/h, 3: 80 km/h, 4: 90 km/h, 5: more than 90 km/h	2.58	0.74	1	5
Density of signalized Intersection	- 1: uninterrupted roads, 2: 0.0 ~ 0.3 unit/km, 3: 0.3 ~ 0.7 unit/km, 4: 0.7 ~ 1.0 unit/km, 5: 1.0 ~ 2.0 unit/km, 6: 2.0 ~ 4.0 unit/km, 7: more than 4.0 unit/km	2.53	1.23	1	7
Road functional class 1	- Minor arterials	0.21	0.40	0	1
Road functional class 2	- Major arterials	0.03	0.18	0	1
Road functional class 3	- Expressways and urban expressways	0.35	0.48	0	1
Land use type 1	- Commercial area	0.21	0.41	0	1
Land use type 2	- Residential area	0.44	0.50	0	1
Land use type 3	- Industrial area	0.13	0.33	0	1
Land use type 4	- Miscellaneous	0.22	0.41	0	1
Contact with the major arterials	- 0: no, 1: yes	0.76	0.43	0	1

Note: The Transverse Mercator (TM) coordinate system KATECH was used in this study.
The number of available unit links for traffic count data and road information is 127 and 150, respectively.

DOI: 10.1002/atr

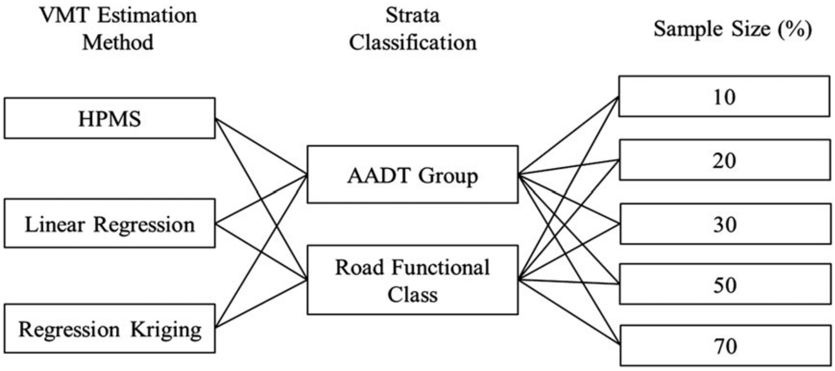


Figure 4. Various scenarios for comparison analysis.

the precision of the estimates without significantly increasing the sample size. However, if there is not information for AADT in all samples, stratification might not possible. Therefore, this study uses the road functional class for stratification and compares the results with the results of AADT group stratification. Regarding the strata classification, this study examined two, namely, AADT group and road functional class:

- AADT group (vehicle/day): 0–20 000 under, 20 000–35 000 under, 35 000–55 000 under, 55 000–85 000 under, and 85 000–250 000 under
- Road functional class: expressways and urban expressways and major arterials and minor arterials

Third, the relative performance of each scenario is examined according to the five sample sizes: 10%, 20%, 30%, 50%, and 70%. Accordingly, this study evaluates 30 scenarios ((3 VMT estimation methods) × (2 strata classifications) × (5 sample sizes)) for VMT estimation.

In order to assess the performance of each scenario, the mean absolute percentage error (MAPE; Equation (16)) is estimated for 10 iterations with different random seed.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - \hat{A}_i}{A_i}\right| \times 100 \tag{16}$$

In Equation (16), n is the total number of iterations, A_i is the actual VMT value, and $\hat{A}_i$ is the estimated VMT value

It is assumed that the actual value, A_i, is derived from the maximum available sample (127 observed traffic counts of 150 total unit links) for each scenario because traffic count data does not exist for all unit links—in other words, there is no actual VMT value. The estimated VMT value and the MAPE are the result of the average of ten iterations, using different random seeds in the sampling process. The coefficient of variation of the VMT estimate is a measure of relative variability within each scenario. The log-transformed AADT values from RK are reverse transformed for the comparison analysis, and the VMT was calculated by multiplying the centerline mileage of the corresponding unit link.

4. RESULTS AND DISCUSSION

The results of RK, linear regression, geostatistical analysis, and the VMT estimation using the maximum sample size are presented. The procedure to implement RK first uses the linear regression model to predict the mean function and then uses a kriging model for prediction. The performances of the HPMS method and RK for the various sample sizes and strata classifications are examined. Many researchers show that regression kriging outperforms the linear regression method for prediction [41–43].

4.1. Regression kriging

4.1.1. Linear regression result

The stepwise process with a $p < 0.05$ was used for the first step of RK. Table III shows the selected explanatory variables and their associated coefficients for the maximum sample size (i.e. 127 points over the all 150 points). In this data set, the intercept, number of lanes, speed limit, density of signalized intersection, two road functional classes, and one land use type are included. The model yields an adjusted R^2 of 0.718 and R^2 of 0.731, respectively. The estimated coefficients offer interesting insights into the effects of auxiliary variables to the transformed AADT. The coefficient for functional class 3 (expressways and urban expressways relative to major arterials) is 0.344, the largest of all functional classes, indicating that the (log transformed) AADT at expressways and urban expressways is 34.4% higher than at minor arterials. The coefficient for land use type 1 (commercial area relative to miscellaneous area) is 0.142 and has the most significant effect among land use types. As expected, the number of lanes and speed limit are positively related to AADT. The coefficients for geometric design of the road and the characteristics of traffic flow are statistically significant. According to the increase of number of signals within 1 km, the interrupted flow under signal control on the road has the negative effect on AADT.

4.1.2. Geostatistical analysis

The residual semivariogram is estimated with the spatial trend removed. The residuals from the linear regression are used to construct the empirical semivariogram, which is fitted with three different models (exponential, Gaussian, and spherical). Table IV lists the estimate of the semivariogram parameters. In all models the nugget is a relatively large fraction of the total sill. The nugget-to-sill ratio (NSR) is generally used to quantify the importance of the random component and provides a quantitative estimation of the spatial dependence [44]. The results suggest that unexplained variation dominates after the feature-space effects are removed. The range can be interpreted as the distance between two points where spatial autocorrelation exists. For distances greater than or equal to the range, spatial correlation is effectively zero [39]. The variograms yield spatial autocorrelations within the ranges of about 1.3 and 2.8 km for the residuals from the exponential model and spherical model, respectively.

In order to determine the model with the best fit, the weighted sum of squares of error (WSSE) and the Akaike Information Criterion (AIC) are used as the evaluation criteria, with lower values indicating

Table III. Selected explanatory variables and coefficient of linear regression model.

Selected variable	Estimate	Standard Error	*t*-Stat	*p*-value
Intercept	4.233	0.106	40.041	<0.0001
Number of lanes	0.077	0.011	6.767	<0.0001
Speed limit	0.092	0.025	3.732	<0.0001
Density of signalized intersection	−0.041	0.014	−2.998	0.003
Road functional class 2	0.079	0.037	2.154	0.033
Road functional class 3	0.344	0.107	3.201	0.002
Land use type 1	0.142	0.030	4.753	<0.0001
R^2		0.731		
Adjusted R^2		0.718		

Table IV. Parameter estimates of residual semivariogram.

	Parameters estimate				Criteria	
Model	Nugget, θ_1	Sill, θ_2	Range (m), θ_3	NSR (θ_1/ θ_2)	WSSE	AIC
Exponential	0.0962	0.1176	1282.21	0.8183	13.8351	6.8093
Gaussian	0.1013	0.1144	1515.24	0.8853	14.2035	7.1510
Spherical*	0.0875	0.1121	2813.91	0.7812	11.0647	3.9045

*Selected semivariogram model.

DOI: 10.1002/atr

a better fit with the data. Test results for the three models supported the spherical semivariogram, as indicated by its smaller WSSE and AIC. The spherical variogram was used in kriging to predict AADT. Figure 5 shows the empirical variogram and the log-transformed residual semivariogram fitted with the spherical model. Inference from the estimated range parameter shows that the spatial autocorrelation effect of traffic volume tapers off at approximately 2800 m in the study area.

4.1.3. Estimating vehicle miles traveled

After estimating the AADT for unknown unit links using RK, the VMT in the study area was calculated by multiplying the estimated AADT by the centerline mileage of the unit links. Table V shows the centerline mileage and daily VMT estimations by road functional class. Figure 6 illustrates the predicted daily VMT in the study area. The VMT estimation result in the study area is 4 226 192 vehicle-kilometer per day on a total of 79.4 km of roadway centerline miles. The results of the VMT estimation on expressways and urban expressways had the highest proportion of VMT because of their numbers of lanes and high traffic volumes.

4.1.4. Model validation

In order to assess the performance of RK, this study uses observations to spatially interpolate AADT values and then compares these estimates to the observed AADT values for remaining observations. In this method, some measurement points are removed and the traffic volume at that point is predicted by using the remaining points. The predicted and measured values are compared at the removed points. The test sets were sampled at identical proportions for each sample stratum. The averages of the MAPE are 20.01% and 24.34% for the 20% and 40% sample size test data, respectively. Figure 7 presents scatter plots of the observations at non-sampled unit links and the corresponding estimated values. The results show strong correlation between the predicted and observed values. However, some overpredictions occur at AADT 50 000 vehicles per day and underprediction occurs at AADT

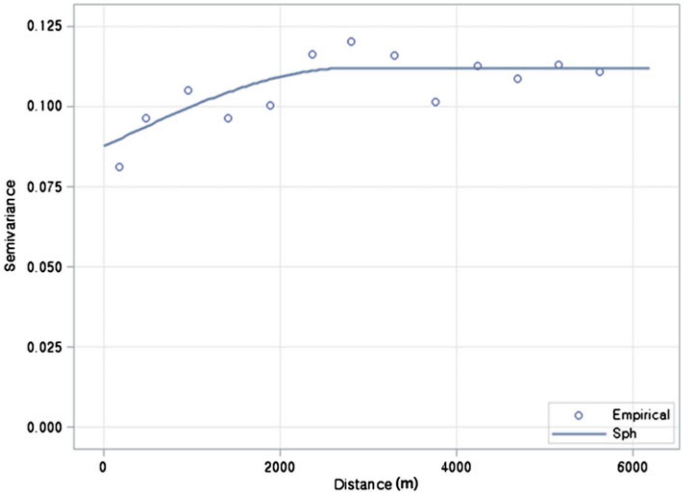


Figure 5. Log-transformed residual semivariogram fitted with the spherical model.

Table V. Centerline mileage and DVMT estimates by road functional class.

	Expressways and urban expressways	Major arterials	Minor arterials	Total
Centerline mileage (km)	8.9 (11.2%)	19.2 (24.2%)	51.3 (64.6%)	79.4 (100.0%)
Daily VMT (vehicle-km/day)	1 563 188 (37.0%)	1 016 598 (24.1%)	1 646 406 (39.0%)	4 226 192 (100.0%)

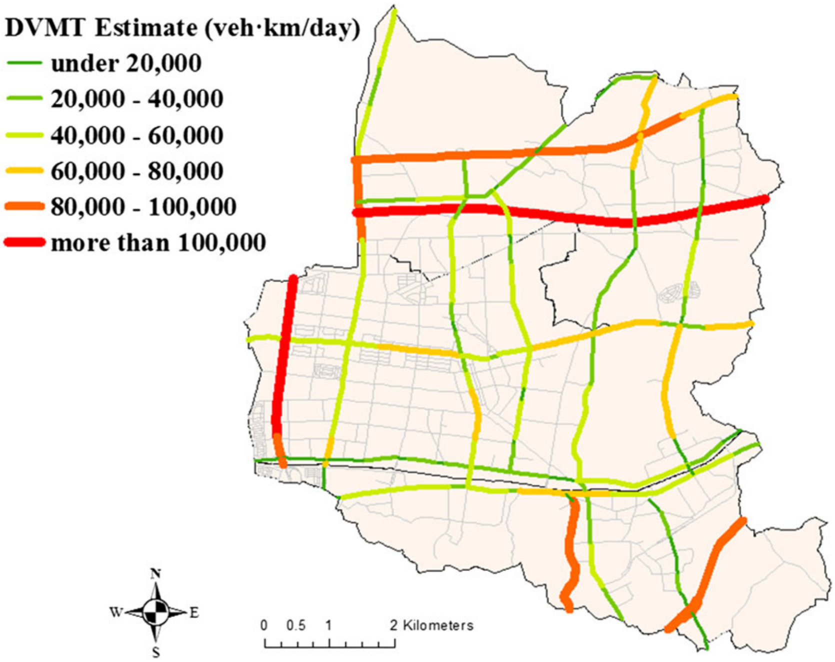


Figure 6. Daily VMT estimates in the study area.

250 000 vehicles per day for the 20% sample size test data. In general, RK provides more accurate predictions at unit links with an AADT below 100 000 vehicles per day.

4.2. Result comparison

The MAPE of RK is smaller than those of HPMS and the linear regression model for most cases. These results validate that RK outperformed the other two methods, which is consistent with previous studies [41–43].

Table VI and Figure 8 show the results from the different VMT estimation methods, strata classification, and sample size. The MAPE of the VMT estimate decreases as the sample size increases, with a change from 10% to 30% in the smaller sample sizes.

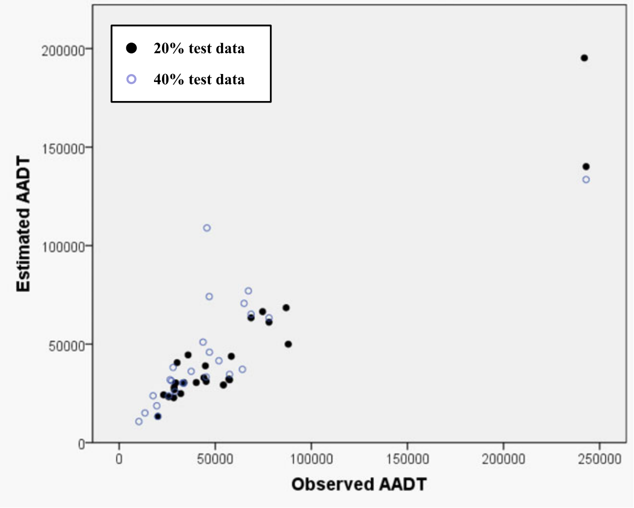


Figure 7. Scatter plot of observed AADT and estimated AADT (20% and 40% test data).

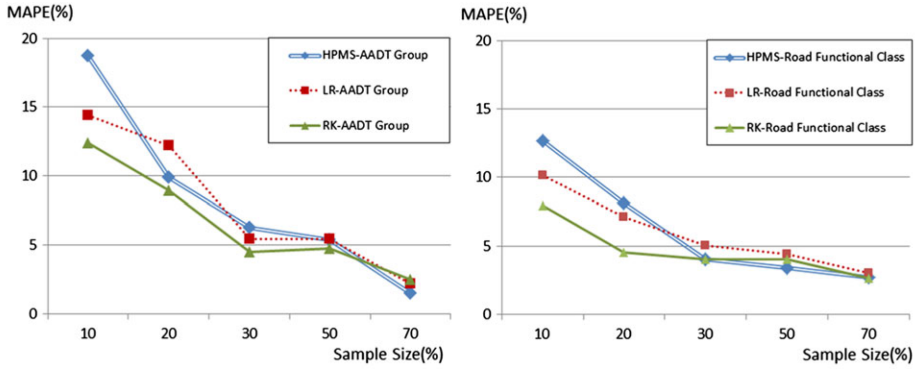
DOI: 10.1002/atr

Table VI. Errors in estimation and coefficient of variation by scenario: VMT estimation method, strata classification, and sample size.

VMT estimation method	Strata classification	Sample size (%)	Actual VMT value* (vehicle-km/day)	Average of VMT estimate (vehicle-km/day)	MAPE (%)	Coefficient of variation
HPMS	AADT group	10	4 216 635	3,694,261	18.7	0.162
		20		3 800 152	9.9	0.050
		30		4 014 278	6.3	0.059
		50		4 293 109	5.4	0.066
		70		4 184 096	1.5	0.014
	Road functional class	10	4 238 723	4 475 873	12.7	0.120
		20		4 385 453	8.1	0.079
		30		4 066 827	4.1	0.017
		50		4 211 102	3.4	0.042
		70		4 277 789	2.7	0.033
Linear regression	AADT group	10	4 231 819	3 729 381	14.4	0.109
		20		3 714 899	12.2	0.061
		30		4 410 537	5.4	0.057
		50		4 391 354	5.4	0.058
		70		4 291 834	2.2	0.031
	Road functional class	10	4 231 819	4 140 022	9.7	0.109
		20		4 513 079	4.4	0.024
		30		4 422 502	5.0	0.049
		50		4 412 866	4.4	0.045
		70		4 392 427	3.1	0.032
Regression kriging	AADT group	10	4 226 192	3 744 817	12.4	0.084
		20		3 860 664	9.0	0.053
		30		4 374 283	4.5	0.048
		50		4 342 836	4.8	0.048
		70		4 310 395	2.5	0.030
	Road functional class	10	4 226 192	4 144 376	9.6	0.110
		20		4 414 317	4.6	0.028
		30		4 397 072	4.0	0.019
		50		4 336 879	4.0	0.036
		70		4 265 373	2.7	0.030

*The actual VMT value was estimated at the maximum available sample size for each scenario.

The functional class-based strata classification performed relatively well, perhaps as a result of a larger number of samples in a single stratum because of the smaller number of classifications (i.e., three for the road functional class compared with five for the AADT group). The VMT estimation performance for all scenarios improves as the sample size increases. In addition, the stabilized coefficient of variation of the VMT estimate indicates that the VMT estimation result could be more reliable because of the increase in sample size (with some exceptions in the road functional class scenario). It is



(a) MAPE by AADT group

(b) MAPE by road functional class

Figure 8. MAPE by various scenarios: (a) MAPE by AADT group, (b) MAPE by road functional class.

DOI: 10.1002/atr

expected that these exceptions come from a small number of sample unit links in the stratum of expressways and urban expressways, which accounts for one third of the total VMT estimates.

The comparison analysis yields the following results. Regarding the VMT estimation method, RK could provide more accurate VMT estimates than the existing HPMS method and linear regression, especially at low sample sizes. In terms of the strata classification, when the unit link samples were stratified by three road functional classes, the MAPE of the VMT estimation was decreased, meaning that it could provide more accurate results. In summary, the following methods regarding VMT estimation method are recommended:

- VMT estimation method: Regression kriging
- Strata classification: Road functional class

5. CONCLUSIONS

To meet the present demand for performance-based transportation planning and budget distribution, VMT, a strong performance measure, is estimated with different methods based on various data sources. This study aims to develop a practical procedure to provide spatial distribution of traffic volume and calculate the VMT. This study suggests a novel VMT estimation method for urban areas using RK to obtain more stable and reliable results. The performance of RK assessed by cross validation was satisfactory and outperformed the traditional regression method and HPMS results. This yields the conclusion that RK is a practical procedure that can be applied to the prediction of the spatial distribution of traffic volume.

The findings in this study can be summarized as follows. First, it is shown that RK can provide more accurate VMT estimates than the HPMS method, especially at lower sample sizes. Second, in terms of strata classification, the road functional class performs relatively better than the AADT group, perhaps as a result of having a larger number of samples in a single stratum. Third, the MAPE of the VMT estimate decreases according to the increase in the sample size regardless of the VMT estimation method. This study is significant because it examines settings for RK (i.e., unit link separation, strata classification, and sample size) to estimate VMT in urban areas and applies these methods to estimate VMT in Bucheon, South Korea, as a case study. The proposed method has important advantages in estimating VMT where traffic count data are scarce. This VMT estimation method, based on a small number of observations, holds promise for transportation agencies that require reliable VMT estimate for performance-based transportation decision-making where the available data are sparse and budgets are limited.

A number of issues remain to be addressed in future research. First, it is necessary to collect sufficient traffic count data on collector and local roads. Because of the limitations of traffic-count-data, this study estimates the VMT for only the top three road functional classes. Second, to represent the city-wide VMT value, it is necessary to collect additional auxiliary data (e.g., socio-economic data and socio-demographic data) for reliable spatial regression modeling.

ACKNOWLEDGEMENTS

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2014R1A2A2A01006906). In addition, Statistical analysis and interpretation were completed with the assistance of the Statistical Analysis and Consulting Center in Chungbuk National University.

REFERENCES

1. Pu W, Meese AJ. Using new data sources to meet MAP-21 requirements for performance-based planning: national capital region's experience in monitoring congestion and reliability. *Proceedings of the 92nd Annual Meeting of the Transportation Research Board, Washington, D.C. United States* 2013.
2. Moore AT, Staley SR Poole RW. The role of VMT reduction in meeting climate change policy goals. *Transportation Research Part A* 2010; **44**(8): 565–574.
3. Teng H, Wang N. Estimating vehicle miles traveled combined with ITS data. *Transportation Planning and Technology* 2011; **34**(8): 777–794.

4. Zhang L, He X, Lu Y, Krause C Ferrari N. Are we successful in reducing vehicle miles traveled in air quality nonattainment areas? *Transportation Research Part A* 2014; **66**: 280–291.
5. Rentziou A, Gkritza K Souleyrette RR. VMT, energy consumption, and GHG emissions forecasting for passenger transportation. *Transportation Research Part A* 2012; **46**(3): 487–500.
6. Paul A. Land-use-accessibility model: a theoretical approach to capturing land-use influence on vehicular flows through configurational measures of spatial networks. *International Journal of Urban Sciences* 2012; **16**(2): 225–241.
7. Barr LC. Testing for the significance of induced highway travel demand in metropolitan areas. *Transportation Research Record: Journal of the Transportation Research Board* 1706; **2000**: 1–8.
8. Kumarage AS, Weerawardana J. System cost-based multi-criteria analysis for urban transport solutions. *International Journal of Urban Sciences* 2013; **17**(2): 212–225.
9. Kim DG, Lee Y. Modelling crash frequencies at signalized intersections with a truncated count data model. *International Journal of Urban Sciences* 2013; **17**(1): 85–94.
10. Ma L, Yan X Weng J. Modeling traffic crash rates of road segments through a lognormal hurdle framework with flexible scale parameter. *Journal of Advanced Transportation* 2015. DOI:10.1002/atr.1322.
11. Kim D, Lee J. Application of neural network model to vehicle emissions. *International Journal of Urban Sciences* 2010; **14**(3): 264–275.
12. Ko J, Park D, Lim H Hwang I. Who produces the most CO2 emissions for trips in the Seoul metropolis area? *Transportation Research Part D* 2011; **16**(5): 358–364.
13. Park D, Kim NS, Park H Kim K. Estimating trade-off among logistics cost, CO2, and time: a case study of container transportation systems in Korea. *International Journal of Urban Sciences* 2011; **16**(1): 85–98.
14. Kumapley RK, Fricker JD. Review of methods for estimating vehicle miles traveled. *Transportation Research Record: Journal of the Transportation Research Board* 1996; **1551**: 59–66.
15. Kweon YJ, Kockelman K. Nonparametric regression estimation of household VMT. *Proceedings of the 83th Annual Meeting of the Transportation Research Board*, *Washington*, *D.C. United States* 2004.
16. Pathomsiri S, Haghani A Schonfeld PM. Vehicle miles traveled in multivehicle households. *Transportation Research Record: Journal of the Transportation Research Board* 1926; **2005**: 198–205.
17. Erlbaum NS. *Estimated County Level Vehicle Miles of Travel*Data Services Bureau Report Planning Division, New York State Department of Transportation: Albany, 1989.
18. White SB. On the use of annual vehicle miles of travel estimates from vehicle owners. *Accident Analysis & Prevention* 1976; **8**(4): 257–261.
19. Greene DL. Long-run vehicle travel prediction from demographic trends. *Transportation Research Record* 1987; **1135**: 1–9.
20. Federal Highway Administration (FHWA). *Highway Performance Monitoring System Field Manual*United States Department of Transportation, Federal Highway Administration: Washington D. C., 2013.
21. Chow AHF, Santacreu A, Tsapakis I, Tanasaranond G Cheng T. Empirical assessment of urban traffic congestion. *Journal of Advanced Transportation* 2014; **48**(8): 1000–1016.
22. Zhang L, He X. Feasibility and advantages of estimating local road vehicle miles traveled on basis of global positioning system travel data. *Transportation Research Record: Journal of the Transportation Research Board* 2013; **2399**: 94–102.
23. Blei A, Kawamura K, Javanmardi M, Mohammadian A. Evaluation of VMT estimation methods using GPS travel survey data. *Proceedings of the 94th Annual Meeting of the Transportation Research Board*, *Washington*, *D.C. United States* 2015.
24. Bishop TFA, McBratney AB. A comparison of prediction methods for the creation of field-extent soil property maps. *Geoderma* 2001; **103**(1): 149–160.
25. Zhu Q, Lin HS. Comparing ordinary kriging and regression kriging for soil properties in contrasting landscapes. *Pedosphere* 2010; **20**(5): 594–606.
26. Neveu AJ. Quick response procedures to forecast rural traffic. *Transportation Research Record* 1987; **944**: 47–53.
27. Mohamad D, Sinha KC, Kuczek T Scholer CF. Annual average daily traffic prediction model for county roads. *Transportation Research Record: Journal of the Transportation Research Board* 1617; **1998**: 69–77.
28. Zhao F, Chung S. Contributing factors of annual average daily traffic in a Florida county: exploration with geographic information system and regression models. *Transportation Research Record: Journal of the Transportation Research Board* 1769; **2001**: 113–122.
29. Lam WHK, Xu J. Estimation of AADT from short period counts in Hong Kong—a comparison between neural network method and regression analysis. *Journal of Advanced Transportation* 2000; **34**(2): 249–268.
30. Lam WHK, Tang YF, Chan KS Tam ML. Short-term hourly traffic forecasts using Hong Kong annual traffic census. *Transportation* 2006; **33**(3): 291–310.
31. Eom JK, Park MS, Heo TY Huntsinger LF. Improving the prediction of annual average daily traffic for nonfreeway facilities by applying a spatial statistical method. *Transportation Research Record: Journal of the Transportation Research Board* 1968; **2006**: 20–29.
32. Wang X, Kockelman KM. Forecasting network data: Spatial interpolation of traffic counts using Texas data. *Transportation Research Record: Journal of the Transportation Research Board* 2009; **2105**: 100–108.
33. Selby B, Kockelman KM. Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression. *Journal of Transport Geography* 2013; **29**: 24–32.

34. Hengl T, Heuvelink GBM, Stein A. Comparison of kriging with external drift and regression-kriging. Technical Report, International Institute for Geoinformation Science and Earth Observation (ITC), Enschede, Netherlands, 2003.
35. Hengl T, Heuvelink GBM Rossiter DG. About regression-kriging: from equations to case studies. *Computers & Geosciences* 2007; **33**(10): 1301–1315.
36. Ligas M, Kulczycki M. Simple spatial prediction-least squares prediction, simple kriging, and conditional expectation of normal vector. *Geodesy and Cartography* 2010; **59**(2): 69–81.
37. Webster R, Oliver MA. *Geostatistics for environmental scientists*. John Wiley & Sons: Chichester, 2001.
38. Matheron G. Principles of geostatistics. *Economic Geology* 1963; **58**(8): 1246–1266.
39. Cressie N. *Statistics for Spatial Data*. John Wiley & Sons: Hoboken, NJ, 1993.
40. Tsapakis I, Schneider WH, Nichols AP Haworth J. Alternatives in assigning short-term counts to seasonal adjustment factor groupings. *Journal of Advanced Transportation* 2014; **48**(5): 417–430.
41. Beelen R, Hoek G, Pebesma E, Vienneau D, de Hoogh K Briggs DJ. Mapping of background air pollution at a fine spatial scale across the European Union. *Science of the Total Environment* 2009; **407**(6): 1852–1867.
42. Pearce JL, Rathbun SL, Aguilar-Villalobos M Naeher LP. Characterizing the spatiotemporal variability of PM 2.5 in Cusco, Peru using Kriging with external drift. *Atmospheric Environment* 2009; **43**(12): 2060–2069.
43. Mercer LD, Szpiro AA, Sheppard L, *et al*. Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NO x) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Atmospheric Environment* 2011; **45**(26): 4412–4420.
44. Di Virgilio N, Monti A Venturi G. Spatial variability of switchgrass (Panicum virgatum L.) yield as related to soil parameters in a small field”. *Field Crops Research* 2007; **101**(2): 232–239.

DOI: 10.1002/atr