EHT Environmental Health and Toxicology

Open Access

Volume: 30, Article ID: e2015010, 14 pages
http://dx.doi.org/10.5620/eht.e2015010

• Special Topic

eISSN: 2233-6567

Computation of geographic variables for air pollution prediction models in South Korea

Youngseob Eum[1,2], Insang Song[1], Hwan-Cheol Kim[3], Jong-Han Leem[3], Sun-Young Kim[4]

[1]Department of Geography, Seoul National University, Seoul, Korea; [2]Department of Geography, State University of New York at Buffalo, NY, USA; [3]Department of Occupational and Environmental Medicine, Inha University School of Medicine, Incheon; [4]Institute of Health and Environment, Seoul National University, Seoul, Korea

Recent cohort studies have relied on exposure prediction models to estimate individual-level air pollution concentrations because individual air pollution measurements are not available for cohort locations. For such prediction models, geographic variables related to pollution sources are important inputs. We demonstrated the computation process of geographic variables mostly recorded in 2010 at regulatory air pollution monitoring sites in South Korea. On the basis of previous studies, we finalized a list of 313 geographic variables related to air pollution sources in eight categories including traffic, demographic characteristics, land use, transportation facilities, physical geography, emissions, vegetation, and altitude. We then obtained data from different sources such as the Statistics Geographic Information Service and Korean Transport Database. After integrating all available data to a single database by matching coordinate systems and converting non-spatial data to spatial data, we computed geographic variables at 294 regulatory monitoring sites in South Korea. The data integration and variable computation were performed by using ArcGIS version 10.2 (ESRI Inc., Redlands, CA, USA). For traffic, we computed the distances to the nearest roads and the sums of road lengths within different sizes of circular buffers. In addition, we calculated the numbers of residents, households, housing buildings, companies, and employees within the buffers. The percentages of areas for different types of land use compared to total areas were calculated within the buffers. For transportation facilities and physical geography, we computed the distances to the closest public transportation depots and the boundary lines. The vegetation index and altitude were estimated at a given location by using satellite data. The summary statistics of geographic variables in Seoul across monitoring sites showed different patterns between urban background and urban roadside sites. This study provided practical knowledge on the computation process of geographic variables in South Korea, which will improve air pollution prediction models and contribute to subsequent health analyses.

Keywords: Air pollution, Cohort study, Exposure prediction, Geographical information system

Correspondence: Sun-Young Kim
1 Gwanak-ro, Gwanak-gu, Seoul 08826,
Korea
Tel: +82-2-880-2759
Fax: +82-2-747-7662
E-mail: puhao@snu.ac.kr

Received: July 10, 2015
Accepted: October 22, 2015
Published online: October 23, 2015

This article is available from: http://e-eht.org/

Introduction

Cohort studies have investigated associations between long-term exposures to air pollution and various health outcomes based on the spatial contrasts of exposures and outcomes of individuals [1,2]. A significant challenge in these studies is that individual measurements of air pollution are not available. Some recent studies have developed exposure prediction approaches to estimate individual-level concentrations of air pollution [3,4]. Geographic variables, computed by using geographic information system (GIS), were often included in such models as predictors for air pollution concentrations at given locations. These variables represented potential sources of air pollution such as traffic and population in surrounding areas and improved the characterization of fine-scale variability of air pollution [3]. Large cohort studies in North America and Europe computed hundreds of variables to represent possible sources [5,6].

Interest has increased in estimating air pollution exposures based on cohort studies in South Korea [7,8]. However, most exposure prediction approaches have relied only on air pollution monitoring data without incorporating geographic variables, possibly due to logistical constraints [9]. The computation of geographic variables requires extended knowledge of available data sources and GIS techniques.

In the present study, we aimed to develop a list of geographic variables based on their relationships with air pollution reported in previous studies. Subsequently, we intended to explore available data sources, to combine all data to a single GIS database and to compute the finalized variables at 294 regulatory air pollution monitoring sites in South Korea. Specifically, we considered various geographic variables as input data for statistical exposure prediction models that have been largely used in cohort studies of air pollution. Our work focused on 2010 data owing to the large number of monitoring sites and availability of the most recent census data [10]. The data integration and variable computation were performed by using ArcGIS version 10.2 (ESRI Inc., Redlands, CA, USA).

Conceptual Background of Geographic Variables

We explored eight categories of geographic variables as potential sources of air pollution. The choice of eight categories was based largely on two large cohort studies, the European Study of Cohorts for Air Pollution Effects [5] and the Multi-ethnic Study of Atherosclerosis and Air Pollution [6], which examined particulate matter (PM) less than or equal to 10 μm and 2.5 μm in diameter (PM_{10} and $PM_{2.5}$, respectively) in addition to black carbon, nitrogen oxide (NO_X), nitric oxide (NO), and nitrogen dioxide (NO_2) air pollution. The eight categories include traffic, demographic characteristics, land use, transportation facilities, physical geography, emissions, vegetation, and altitude (Table 1).

Table 1. List of geographic variables in eight categories with their data sources and types of data

Category[a]	Variable	Source	Type of data (data format)
Traffic	Distance to the nearest roads (all roads, MR1, and MR2)	KTDB	Road network (line)
	Sum of road lengths (all roads, MR1, and MR2)[b]		
	Number of registered vehicles	KOSIS	Vehicle registration (table)
Demographic characteristics	Number of people	SGIS	Census (table)
	Number of households		
	Numbers of housing buildings by a type of residence and by a constructed year		
	Numbers of companies and employees by a type of business		
Land use	Proportions of residential, industrial, commercial, cultural, transportation, public facility, agricultural, forest, grassland, wetland, bare ground, and water areas	EGIS	Land cover map (polygon)
Transportation facilities	Distances to the nearest railroad and subway station	SGIS	Railroad and subway stations (point)
	Distance to the nearest bus stop	Biz-GIS	Bus stop (point)
	Distance to the nearest air port	ODP	Airport (table)
	Distance to the nearest major port	SP-IDC	Port (table)
Physical geography	Distance to river	SGIS	River (polygon)
	Distance to coastline	NSIC	Coastline (line)
	Distance to the military demarcation line	SGIS	Administrative boundary (polygon)
Emissions	Proportions of major pollutants (CO, NO_X, SO_X, TSP, PM_{10}, VOC, and NH_3)	NIER	Emission estimates (table)
Vegetation	Annual summary (average, minimum, and maximum) of NDVI	IIS	Satellite image (raster)
	Median value in August for previous, current and following years		
Altitude	Absolute elevation	USGS	Digital Elevation Data (raster)
	Proportion of concentric elevation points above or below 20 or 50 m		

MR1, major road 1; MR2, major road 2, TSP, total suspended particle; CO, carbon monoxide; NO_X, nitrogen oxides; SO_X, sulfur oxides; NH_3, ammonia; VOC, volatile organic compounds; NDVI, Normalized Difference Vegetation Index; KTDB, Korean Transport Database; KOSIS, Korean Statistical Information Service; SGIS, Statistical Geographic Information Service; EGIS, Environmental Geographical Information Service; ODP, open data portal; SP-IDC, Shipping and Port Integrating Data Center; NSIC, National Spatial Information Clearinghouse; NIER, National Institute of Environmental Research; IIS, Institute of Industrial Science, University of Tokyo; USGS, United States Geological Survey.

[a] Different buffer sizes by category: traffic, 25, 50, 100, 300, 500, and 1000 m; demographic characteristics and land use, 50, 100, 300, 500, 1000, and 5000 m; emissions, 3, 15, and 30 km.

[b] Sum of road lengths were computed for three methods: single central lines of roads, road lines multiplied by numbers of lanes, and road lines multiplied by numbers of lanes and line widths.

Each variable was calculated at regulatory monitoring sites by using one of two metrics: the distance to a feature or a buffer summary statistic (e.g., sum or proportion) of a feature. A buffer is a circular feature that indicates that the air pollution concentration measured at the central point of the buffer is influenced by probable sources at a given distance. In this study, we used three sets of different buffer sizes. The buffer radii for traffic variables were 25, 50, 100, 300, 500, and 1000 m, whereas larger radii of 100, 300, 500, 1000, and 5000 m were applied to non-traffic variables for demographic characteristics and land use categories [5,11,12]. In addition, the largest buffer radii of 3, 15, and 30 km were adopted for emission variables [13].

Traffic

Because traffic is considered a major source of air pollutants such as $PM_{2.5}$ and NO_X, associated with health endpoints in epidemiological studies, geographic variables reflecting traffic have been widely used in exposure prediction models in many cohort studies [14-17]. Vehicles emit various air pollutants from exhaust and non-exhaust factors such as diesel/gasoline engines, brake wear, and road surface wear. The amount of traffic estimated on each road for a given time period could be the best metric for representing vehicle emission. However, traffic volume data are not generally available [12]. Instead, the proximity to the nearest roads and lengths of surrounding roads are frequently adopted as proxies for traffic volume. Road lengths are summed within specific buffer sizes. In particular, NO_X, NO, and NO_2 have been positively associated with decreasing distances to major roads and increasing lengths of roads in surrounding areas [3].

Demographic Characteristics

Densities of population and households implying human activities related to heating, cooking, and transportation would result in increasing air pollution concentrations. The increasing number or density of residents and households in given buffer areas tended to be related to the elevated levels of air pollution concentrations in Europe, Canada, the US, and Taiwan [5,12,18-20].

Land Use

The types of land use such as residential, urban, and green areas have been used as significant predictors for PM and NO_X/NO_2 air pollution in North America and Europe [3]. A study in New York City used a vegetative land use variable to explain the variation in $PM_{2.5}$ concentrations [17]. In addition, a study in Taiwan using five land use categories found that high proportions of urban green and natural areas in given buffers were associated with decreased NO_X concentrations, whereas a high proportion of the low density residential area contributed to increased NO_2 [12].

Transportation Facilities

Public transportation depots such as airports, ports, subway stations, and bus stops could affect increased air pollution concentrations owing to the high emission from transportation equipment and facilities. For example, a large number of concentrations of ultrafine PM and $PM_{2.5}$ was found to be related to aircraft takeoff [21]. $PM_{2.5}$ concentrations attributable to aviation emissions were shown to decrease with increasing distances from airports in the UK [22]. Moreover, a study in Italy reported high concentrations of NO_2 near a port area, possibly attributed to vessel traffic emission [23].

Physical Geography

Natural geographical features may also affect air pollution. For example, the proximity to water bodies affecting air flow in river valleys and oceans could be related to air pollution. The remoteness to a coastline has been associated with increasing NO_2 in San Diego, California, the US [24] and decreasing PM_8 in Shizuoka Prefecture, Japan [25].

Emissions

Emissions derived from various sources for major pollutants could be related to air pollution concentrations. Many countries have provided emission estimates of major pollutants from pollution sources such as roads, transportation, industry, and agriculture [26, 27]. To improve exposure prediction models, some studies in the US and Italy have included emission estimates for primary pollutants [6,28].

Vegetation

Normalized Difference Vegetation Index (NDVI), one of the most frequently used vegetation indices, measures the density of green vegetation on land by using satellite images. This index is computed by determining the reflectance values on a target area of the earth's surface in the visible red (RED) and near-infrared (NIR) bands, as shown in equation 1 [29].

$$\frac{(\text{NIR-RED})}{(\text{NIR+RED})} \qquad (1)$$

High NDVI values imply abundant vegetation and may be negatively associated with air pollution levels. Studies of exposure prediction models in the US have used medians and quantiles of 16-day composite NDVI values over one year and seasonal values for high-vegetation and low-vegetation seasons [6].

Altitude

Atmospheric pressure and circulation changes by elevation could affect movements of air pollutants and air pollution concentrations at a given location. Although altitude has been negatively associated with $PM_{2.5}$ concentrations in four European cities of substantially varied altitudes [5], one study in Taipei, Taiwan, with relatively homogenous altitudes across monitoring sites excluded altitude variables in the final exposure prediction model for NO_X and NO_2 [12].

Data Acquisition

Locations of Regulatory Monitoring Sites

The Ministry of Environment in South Korea established regulatory air quality monitoring networks to monitor the conditions of air pollution and attain the air quality standards since 1980s [30, 31]. We obtained the addresses and coordinates of the 294 monitoring sites operated nationwide in 2010 from the Annual Report of Ambient Air Quality in Korea 2010 [31]. When address and coordinates did not match, we used the address and extracted coordinates from Google Maps.

Road Networks

Road network data for 2010 were obtained from the Korean Transport Database (KTDB) of the Korea Transport Institute (http://www.ktdb.go.kr). The shapefile, a popular data file format in GIS software, for road networks is composed of more than 100000 line segments and corresponding characteristics of each road segment such as a road name, speed limit, direction, and number of lanes. KTDB classifies all roads in South Korea into eight types including national highway, metropolitan city highway, general national road, metropolitan city road, government-financed provincial road, provincial road, district road, and highway link lamp. In addition, the monthly data for the number of registered vehicles in 2011 for a district administrative unit with a median area of 391 km^2 in 2010, known locally as a si–gun–gu, were downloaded from the Korean Statistical Information Service website of the Statistics Korea (http://kosis.kr/eng/); 2010 is the earliest year with available data.

Census

We obtained 2010 census tabular data for a census territorial unit, as a census output area, known as a jipgegu with a median area of 0.02 km^2 in 2010, from the Statistical Geographic Information Service (SGIS) of the Statistics Korea (http://sgis.kostat.go.kr). For each jipgegu, the numbers of residents, households, housing buildings, companies, and employees were available as a total value and by classification of gender and age, type and construction year of houses, and type of business. The type of house was classified on the basis of the size and height of the housing building, whereas the type of business was based on the seven categories of the Korean standard industrial classification. We also obtained a shapefile of jipgegus from the SGIS.

Land Cover Map

Land cover maps were obtained from the Environmental Geographic Information Service of the Korea Ministry of Environment (http://egis.me.go.kr). Land surface images captured by satellites were converted into land cover maps consisting of areas with various land surface characteristics categorized by using image-adjustment and image-classification algorithms. To cover the entire country through 2007, 814 maps were created; 150 maps mostly the Seoul metropolitan areas were updated in 2009. To use the most recently updated maps, we selected the 2007 maps and replaced 150 areas with the updated 2009 data. These land cover maps consisted of 7, 22, and 41 classes for high, medium, and low spatial levels, respectively. The seven high-level classes included urbanized and built areas, agricultural areas, forest areas, grasslands, wetlands, bare ground, and waters (Table 2). We used the high-level classes except for the urbanized and built area class. For such class, we replaced by seven medium-level classes reflecting specific land use characteristics possibly associated with air pollution in complex urban settings.

Emissions

We downloaded the tabular data for emissions, created by the National Institute of Environmental Research, from the National Air Pollutants Emission website (http://airemiss.nier.go.kr/

Table 2. Classification of land surface map data

High spatial level	Medium spatial level
Urbanized and built area	*Residential area*[a] *Industrial area* *Commercial area* *Cultural, sport, recreation area* *Transportation area* *Public facility area*
Agricultural area	Rice paddy Field Cultivated field under structure Orchard Other cultivated field
Forest area	Broad-leaved forest Coniferous forest Mixed stand forest
Grassland	Natural grassland Artificial grassland
Wetland	Inland wetland Coastal wetland
Bare ground	Natural bare ground Other bare ground
Waters	Inland water Ocean water

[a] Six out of seven high-level classes and six medium-level classes of the urbanized and built area (bold and italic) were used in our study.

main.jsp). The emission data contain annual emission estimates of seven pollutants including carbon monoxide, NO_X, sulfur oxide, total suspended particulates, PM_{10}, volatile organic compounds, and ammonia aggregated on a 1-km grid by point, line, and area sources with grid coordinates indicating the bottom left corner of each grid cell.

Normalized Difference Vegetation Index

For NDVI, we downloaded satellite images from the Institute of Industrial Science (IIS), University of Tokyo (http://webmodis.iis.u-tokyo.ac.jp/). The IIS provides cloud- and shadow-free images captured by Aqua/Terra Moderate Resolution Imaging Spectroradiometer (MODIS) satellites over Asia every 10 days. The 10-day composite images were created by using the Enhanced Second Minimum composition method, which selects a pixel with second minimum reflectance in the red channel as a representative value during a 10-day period to avoid pixels shadowed by clouds [32]. The spatial resolution of these raster image data was approximately 428 m for each cell. The NDVI pixel image values were between 0 and 255.

Other data

Shapefile data for bus stops were downloaded from the Biz-GIS website (http://www.biz-gis.com/XsDB/). This website, operated by the Biz-GIS company, provides spatial data in South Korea in shapefile formats for various spatial features such as apartments, hospitals, banks, and retail establishments. The addresses and coordinates of airports generated by the Korea Airports Corporation were obtained from an open data portal (https://www.data.go.kr) that offers data generated by various public agencies. Coastline data generated by the Korea Hydrographic and Oceanographic Administration were obtained from the National Spatial Information Clearinghouse (https://www.nsic.go.kr/ndsi/). We obtained shapefile data for railroad stations, subway stations, urbanized areas, rivers, and boundaries of si–gun–gu from the SGIS. The Shuttle Radar Topography Mission (SRTM) 30 m × 30 m digital elevation model (DEM) data for South Korea was downloaded from the EarthExplorer interface of the United States Geological Survey website (http://earthexplorer.usgs.gov).

Data Integration

To compute the geographic variables, we integrated all data obtained from the various sources into the one GIS database. This integration included the transformation of different coordinate systems to a single system. Spatial data generated from the various organizations were on different coordinate systems or no information was provided for coordinate systems. This limitation created difficulties in displaying all available data on the same map, which is necessary for computing geographic variables. Thus, we adopted the Korean Central Belt 1995 coordinate system, which is most commonly used in our GIS data (Table S1).

In addition, we combined tabular, vector, and raster data into the database. The data for monitoring sites, census, registered vehicles, and emissions were provided in tabular formats. Shapefiles for road networks and administrative boundaries included vector data representing point, line, and polygon features. Image files for satellite data contained raster data, which display data values on grid cells.

Variable Computation

Traffic

We computed the distance to the nearest road, the sum of road lengths in a buffer space, and the number of registered vehicles at each regulatory monitoring site. The distance to the nearest road was computed as the minimum Euclidian distance between the line of the road and each monitoring site. For the sum of road lengths, we created various sizes of traffic buffers, selected road segments within each buffer, and aggregated the lengths of all selected road segments. For the number of registered vehicles, we calculated the annual average of registered cars from monthly data for each si–gun–gu. Then, we linked this tabular data to the shapefile of si–gun–gu and identified the number of vehicle registrations in the si–gun–gu that included a target monitoring site.

The traffic variables of road networks were computed for each of the three categories of roads including all roads, major road 1 (MR1), and major road 2 (MR2). MR1 was defined by national highways and metropolitan city highways, whereas MR2 included MR1 as well as local roads with more than six lanes (Figure 1). We created the category of MR2 owing to the limited number of MR1 roads. The total length of all roads was 90816 km in South Korea, whereas lengths of MR1 and MR2 were 8128 and 12194 km which are 8.95 and 13.43 % of all roads, respectively.

In addition, we incorporated the numbers of lanes and road widths to the sum of road length variables. Road networks in the KTDB are represented by single line features of the centerlines of roads without considering lanes and widths. The KTDB road network data contains attributes of the number of lanes. We assigned road width values depending on highway/non-highway, speed limit, and urban/non-urban areas based on the Administrative Rule on the Structure and Installation of Road (Table S2) [33]. Information on highway/non-highway and

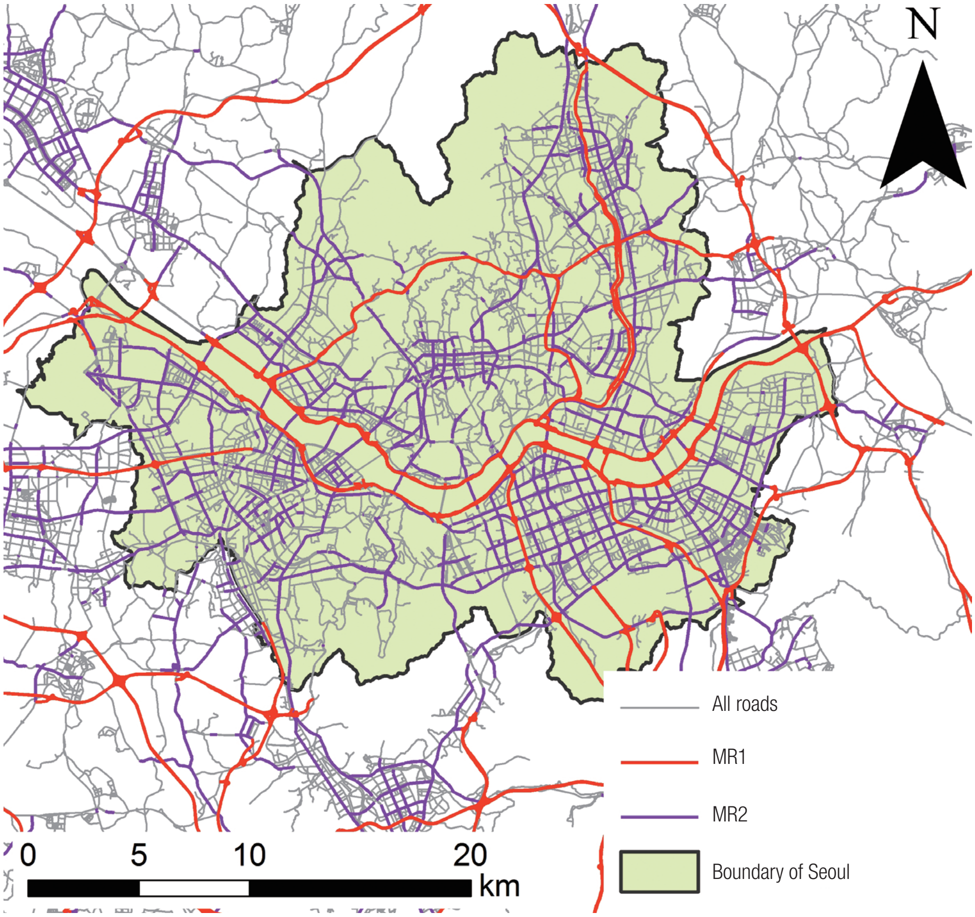

Figure 1. Road networks in Seoul, Korea. MR1, major road 1; MR2, major road 2.

speed limit were given in the KTDB road network data. For urban/non-urban areas, we defined the urban area that overlapped with the urbanized area shapefile obtained from the SGIS (Figure S1). Thus, the road segments intersecting with urbanized areas were considered as roads in the urban areas; others were defined as non-urban roads.

Demographic Characteristics

We computed the numbers of total population and households, numbers of housing buildings by construction years and house types, and numbers of companies and employees by business types in a non-traffic buffer. After linking the tabular census data to the administrative boundary shapefiles by using the jipgegu identifier, we selected jipgegus intersecting with a non-traffic buffer and aggregated the numbers of residents, households, housing buildings, companies, or employees within intersected jipgegus with the weight of jipgegu sizes (Figure 2). For example, equation 2 shows the total population within a given buffer i (P_i), derived by using total population in a jipgegu j (P_j) and an areal weight as the ratio of the area of the intersected jipgegu with the buffer (A_{ij}) to the area of the jipgegu (A_j).

$$P_i = \Sigma_{j=1}^{N} \frac{P_j \times A_{ij}}{A_j} \quad (2)$$

Land Use

A land use variable was computed as the proportion of the areas for a type of a land use to the area of a given non-traffic buffer. For each type of the 12 land use classes within a buffer area, we selected the polygons of each land use within the buffer and computed the proportions of the selected areas of the land use to the buffer area.

Transportation Facilities

The distances of a monitoring site to the nearest transportation depots such as railroad stations, subway stations, bus stops, airports, and major ports were calculated. We defined major ports by ports that accommodate more than 10000 vessels per year based on the statistics from the Shipping and Port Integrated Data Center. Ten out of 31 ports were identified as major ports in South Korea.

Physical Geography

The minimum distances to rivers, coastlines, and the border be-

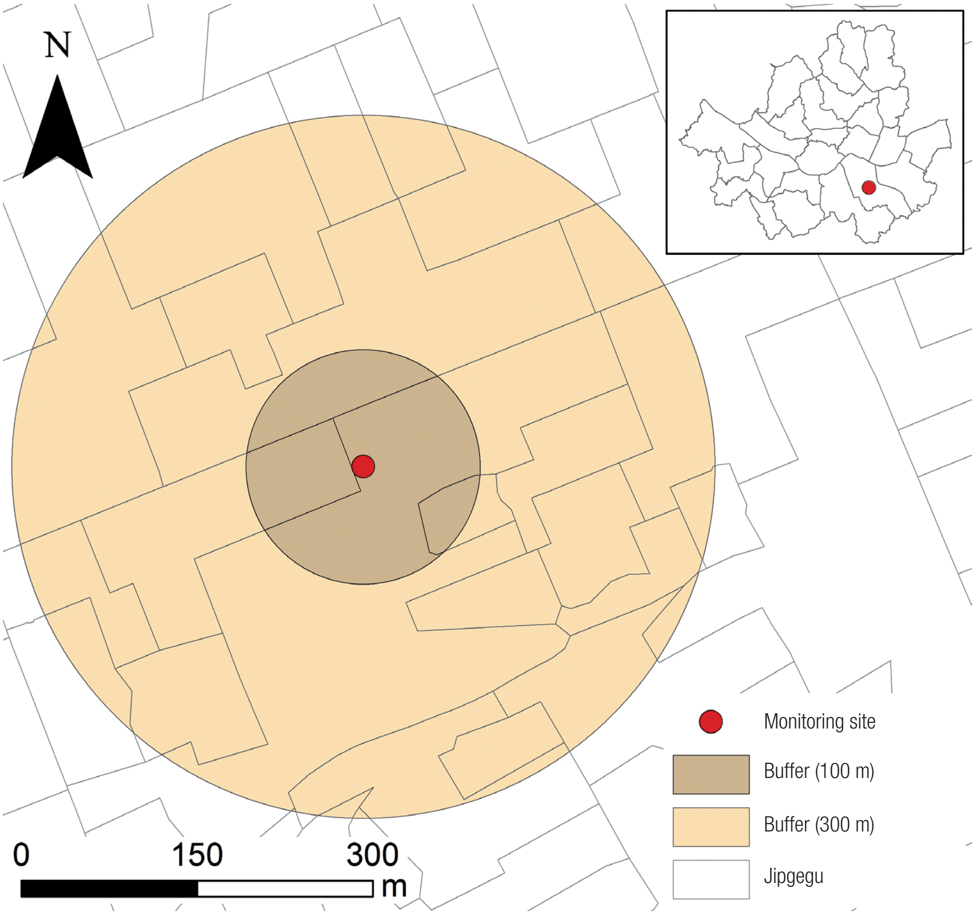

Figure 2. Map of 100 and 300 m buffers and nearby jipgegus of a regulatory monitoring site in Seoul, Korea.

tween North and South Korea were computed. To produce the borderline, we combined all jipgegu polygons into a single polygon, converted the polygon into a line feature representing the outer boundary of South Korea, and extracted the northern end.

Emissions

By using the emission tabular data at 106070 1-km national grid coordinates in South Korea, we aggregated emission estimates from point, line, and area sources at each coordinate. Then, we created 1-km grid-shaped polygons based on grid coordinates, and assigned emission estimates at the grid points to corresponding polygons. Air pollutant emissions in grid polygons were accumulated within 3, 15, and 30 km buffers with areal weight as previously described in the demographic characteristics.

Normalized Difference Vegetation Index

We computed the average, minimum, and maximum of 36 10-day composite MODIS NDVI data during 2010 for each grid to avoid seasonal variation and to estimate spatially representative values. In addition, the median for August during 2009, 2010, and 2011 was computed for reflecting the lushest vegetation in South Korea. Finally, we extracted the NDVI values at each monitoring site from the grid in which the monitoring site is located.

Altitude

By using the 30 m × 30 m SRTM DEM raster image data, we assigned the elevation value in a grid cell to each monitoring site included in the cell. In addition, we computed the relative elevation as the proportion of concentric cells in which the elevation values are above or below threshold elevations of 20 and 50 m, respectively, compared with that at a monitoring site. The concentric cells refer to the DEM grid cells on a 30 m-wide donut-shaped polygon 1 or 5 km from the monitoring site.

Summary of Computed Variables

On the basis of the 313 geographic variables computed at 294 regulatory monitoring sites in South Korea, we examined the relationships between traffic variables and the summary statistics of selected variables to verify our computation results and to

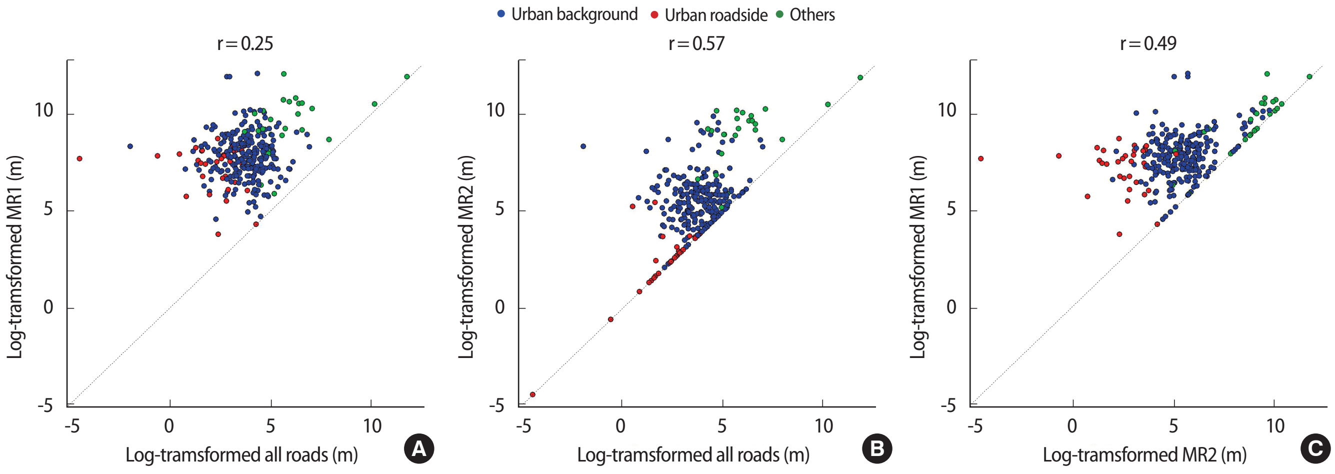

Figure 3. Scatter plots between all roads and major road 1 (MR1) (A), between all roads and major road 2 (MR2) (B), and between MR1 and MR2 (C) across 294 monitoring sites in South Korea.

Table 3. Summary statistics of selected geographic variables by 25 urban background and 12 urban roadside regulatory monitoring sites in Seoul

Category	Variable[a]	Type	Urban background (n=25)				Urban roadside (n=12)			
			Min	Max	Mean	SD	Min	Max	Mean	SD
Traffic	Distance to the nearest road (m)	All roads	7	392	79	77	2	73	21	20
		MR1	219	3647	1469	913	44	3347	1466	1207
		MR2	47	709	261	210	2	226	41	61
	Sum of road length (km)[1]	All roads	0	2.7	1.4	0.6	1.2	4.3	2.3	0.9
		MR1	0	0.5	0.0	0.1	0.0	2.3	0.3	0.7
		MR2	0	1.4	0.4	0.4	0.6	2.6	1.2	0.6
	Sum of road length×lane×width (1000 m^2)[1]	All roads	0	39.1	17.3	9.5	17.1	46	30.8	7.6
		MR1	0	5.4	0.2	1.1	0	22.8	2.6	6.8
		MR2	0	31.4	8.9	8.4	11	34	22.7	7.3
Demographic characteristics	No. of people[1]		4	13900	6624	4042	602	7717	2915	2024
	No. of employees[1]	Construction	0	971	173	233	0	1317	334	444
		Lodging and restaurant	0	2040	376	432	12	1772	815	573
Land use	The proportion of land use (%)[1]	Residential	0	93	39	25	1	85	25	22
		Forestry	0	49	5	11	0	29	3	8
Physical geography	Distance to the nearest river (m)		158	3861	1109	829	51	2805	1368	924
Emissions	PM_{10} (1,000 μg/m^3)[2]		479	1031	688	148	515	983	704	109
Vegetation	Annual mean NDVI		141	167	148	6	140	155	144	4
Altitude	Altitude (m)		14	91	35	17	19	35	27	6
	Proportion of 5 km concentric elevation points (%)	Above 20 m	0	78	18	22	0	17	7	6
		Below 20 m	7	81	35	19	0	9	0	0

Min, minimum; Max, maximum; SD, standard deviation; MR1, major road 1; MR2, major road 2; NDVI, Normalized Difference of Vegetation Index; PM_{10}, particulate matter less than or equal 10 μm in diameter.

[a] Geographic variables calculated within different sizes of buffers: buffer radii of 300 m ([1]) and 3 km ([2]).

provide insight into distributions of the geographic variables. Our presentation of descriptive statistics was restricted to Seoul and focused on the comparison by two types of regulatory monitoring sites including 25 urban background and 12 urban roadside sites.

Figure 3 shows the relationships of natural log-transformed distances to the nearest road among all roads, MR1, and MR2 across regulatory monitoring sites in South Korea. For most monitoring sites, the nearest road was a local road rather than an MR1 or MR2. For one urban roadside site, a MR1 was the nearest road. This site is the only regulatory monitoring site located on a metropolitan city highway according to its address, indicating that our computation was accurate. For about 20 % of the sites, MR2 roads were the nearest roads.

Table 3 gives summary statistics of the two types of sites in Seoul. The urban roadside sites were located closer to all roads or MR2s than urban background sites and were more surrounded by these roads. The differences of the sums of road lengths within 300 m buffers between the two sites types increased when lanes and widths were applied. More than a half of the sites did not have an MR1 within 300 m. More residents lived within 300 m from urban ambient sites than urban roadside sites, and fewer workers were employed in construction, lodging, and restaurant businesses. The average proportion of residential areas within 300 m from urban roadside sites (25 %) was lower than that within 300 m from urban background sites (39 %). Emission estimates for PM_{10} within 3 km were consistent between the two types of sites (0.7 g/m^3). The monitoring sites were located in relatively flat areas with less than 40 % of 5 km distant points below or above 20 m.

Discussion

We demonstrated the computation process of 313 geographic variables at air pollution regulatory monitoring sites in the eight categories of possible pollution sources in South Korea. The computed variables reflected the geographic characteristics in South Korea and showed the different patterns between urban background and urban roadside sites.

For characterizing spatial variability of air pollution, we computed a large suite of geographic variables for the development of statistical exposure prediction models that rely largely on geographic variables. Previous studies included geographic variables chosen by model selection [34] or a few summary variables estimated by dimension reduction techniques [35] into statistical models. Whereas land use regression includes geographic variables only [3], universal kriging also incorporates the spatial correlation structure [36]. In addition to statistical methods, other studies have developed air quality models such as photochemical models and dispersion models based on the chemical and physical atmospheric processes of air pollution. These models used limited geographic variables of traffic, population, or emissions, and other input data such as meteorology. The different approaches resulted in inconsistent model performance in air pollution prediction [37-39] and varying health effect estimates in subsequent health analyses using predicted individual-level air pollution concentrations [40,41]. Studies comparing model performance between land use regression and dispersion models have generally showed large spatial variability in land use regression and large temporal variability in dispersion models [37-39]. Large spatial variability of air pollution is particularly important for assessing health effects of long-term exposures in cohort studies which largely rely on spatial contrasts. Because our ultimate goal in computing geographic variables lies in health analysis rather than the identification of air pollution distribution, we focused on statistical exposure prediction models and presented a large set of geographic variables.

Our descriptive statistics of geographic variables provided insights into data handling and model building in future studies of exposure prediction models. We found some extreme values particularly for distance variables. Some regulatory monitoring sites were located substantially far from national highways, airports, and ports. Variables with large variability resulting from extreme values could affect model selection and exposure prediction. Future studies should exclude or truncate such variables. We used small sizes of buffers including 25 m to represent the fine-scale spatial variability in a metropolitan city with high density. However, these small buffer variables may not provide meaningful or accurate values; few large roads were detected within 25 m and the distance to large roads based on central lines could contain errors similar to the buffer size. Future studies of exposure modeling approaches need to carefully consider the inclusion of these small-buffer variables.

We presented about 300 geographic variables that require future updates. Recent studies additionally included outputs of dispersion models and air quality models [42] or air pollution estimates determined by satellites [43] as predictors. Future studies need to introduce data sources and variable computation of these variables for South Korea. In addition, new data sources that have not been explored in previous studies mostly performed in North America and Europe may be available. These data could explain complex air pollution environments in other regions including densely populated metropolitan cities.

This study contributes to future studies of exposure prediction and health analyses. Our previous study in South Korea used ordinary kriging to predict air pollution concentrations based solely on spatial correlation without geographic variables. Findings showed poor model performance and suggested the inclusion of geographic variables to improve model performance [44]. An extended set of geographic variables could help explain the spatial variability of air pollution in complex urban environments in large and dense metropolitan cities in South Korea. A previous simulation study also showed that improved air pollution predictions tended to give less biased and more precise health effect estimates [45]. High quality exposure predictions incorporating geographic variables would clarify the association of air pollution and health in South Korea.

Conclusion

The computation of extended geographic variables provides an opportunity for developing exposure prediction models that characterize heterogeneity of air pollution over space. This study will help future research utilize geographic variables for the development of prediction models and provide air pollution estimates with fine-scale spatial variability. Such air pollution predictions will allow subsequent health analyses based on individual exposure estimates in South Korea.

Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A6A3A04059017).

Conflict of Interest

The authors have no conflicts of interest with material presented in this paper.

ORID

Jong-Han Leem *http://orcid.org/0000-0003-3292-6492*
Sun-Young Kim *http://orcid.org/0000-0002-7110-3395*

References

1. Pope CA 3rd, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, et al. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA 2002;287(9):1132-1141.
2. Laden F, Schwartz J, Speizer FE, Dockery DW. Reduction in fine particulate air pollution and mortality: extended follow-up of the Harvard Six Cities study. Am J Respir Crit Care Med 2006;173(6):667-672.
3. Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fischer P, et al. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos Environ 2008;42(33):7561–7578.
4. Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, et al. A review and evaluation of intraurban air pollution exposure models. J Expo Anal Environ Epidemiol 2005;15(2):185-204.
5. Eeftens M, Beelen R, de Hoogh K, Bellander T, Cesaroni G, Cirach M, et al. Development of Land Use Regression models for PM(2.5), PM(2.5) absorbance, PM(10) and PM(coarse) in 20 European study areas; results of the ESCAPE project. Environ Sci Technol 2012;46(20):11195-11205.
6. Keller JP, Olives C, Kim SY, Sheppard L, Sampson PD, Szpiro AA, et al. A unified spatiotemporal modeling approach for predicting concentrations of multiple air pollutants in the multi-ethnic study of atherosclerosis and air pollution. Environ Health Perspect 2015;123(4):301-309.
7. Kim E, Park H, Hong YC, Ha M, Kim Y, Kim BN, et al. Prenatal exposure to PM10 and NO2 and children's neurodevelopment from birth to 24 months of age: Mothers and Children's Environmental Health (MOCEH) study. Sci Total Environ 2014;481:439-445.
8. Lim YH, Kim H, Kim JH, Bae S, Park HY, Hong YC. Air pollution and symptoms of depression in elderly adults. Environ Health Perspect 2012;120(7):1023-1028.
9. Son JY, Lee JT, Kim H, Yi O, Bell ML. Susceptibility to air pollution effects on mortality in Seoul, Korea: a case-crossover analysis of individual-level effect modifiers. J Expo Sci Environ Epidemiol 2012;22(3):227-234.
10. Yi SJ, Kim NK, Kim H, Kim SY. Exploration and application of regulatory monitoring data to development of the long-term PM10 exposure prediction models. In: Korean Society of Environmental Health and Toxicology. Proceedings of the 2014 Conference of Korean Society of Environmental Health and Toxicology; 2014 May 29; Gwangju, Korea. Seoul: Korean Society of Environmental Health and Toxicology: 2014, p. 418-419 (Korean).
11. European Study of Cohorts for Air Pollution Effects. ESCAPE exposure assessment manual; 2010 [cited 2015 10 20]. Available from: http://www.escapeproject.eu/manuals/ESCAPE_Exposure-manualv9. pdf.
12. Lee JH, Wu CF, Hoek G, de Hoogh K, Beelen R, Brunekreef B, et al. Land use regression models for estimating individual NOX and NO2 exposures in a metropolis with a high density of traffic roads and population. Sci Total Environ 2014;472:1163-1171.
13. Multi-ethnic Study of Atherosclerosis and Air Pollution. Data organization and operating procedures; 2013 [cited 2015 Oct 20]. Available from: http://depts.washington.edu/mesaair/MESAAirDOOP.pdf.
14. Beelen R, Hoek G, van den Brandt PA, Goldbohm RA, Fischer P, Schouten LJ, et al. Long-term effects of traffic-related air pollution on mortality in a Dutch cohort (NLCS-AIR study). Environ Health Perspect 2008;116(2):196-202.
15. Hoek G, Brunekreef B, Goldbohm S, Fischer P, van den Brandt PA. Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. Lancet 2002;360(9341):1203-1209.
16. Lung SC, Hsiao PK, Wen TY, Liu CH, Fu CB, Cheng YT. Variability of intra-urban exposure to particulate matter and CO from Asian-type community pollution sources. Atmos Environ 2014;83:6–13.
17. Ross Z, Jerrett M, Ito K, Tempalski B, Thurston GD. A land use regression for predicting fine particulate matter concentrations in the New York City region. Atmos Environ 2007;41(11): 2255–2269.
18. Gilbert NL, Goldberg MS, Beckerman B, Brook JR, Jerrett M. Assessing spatial variability of ambient nitrogen dioxide in Montréal, Canada, with a land-use regression model. J Air Waste Manag Assoc 2005;55(8):1059-1063.
19. Morgenstern V, Zutavern A, Cyrys J, Brockow I, Gehring U, Koletzko S, et al. Respiratory health and individual estimated exposure to traffic-related air pollutants in a cohort of young children. Occup Environ Med 2007;64(1):8-16.
20. Smith L, Mukerjee S, Gonzales M, Stallings C, Neas L, Norris G, et al. Use of GIS and ancillary variables to predict volatile organic compound and nitrogen dioxide levels at unmonitored locations. Atmos Environ 2006;40(20):3773–3787.
21. Zhu Y, Fanning E, Yu RC, Zhang Q, Froines JR. Aircraft emissions and local air quality impacts from takeoff activities at a large International Airport. Atmos Environ 2011;45(36):6526–6533.
22. Yim SH, Stettler ME, Barrett SR. Air quality and public health im-

pacts of UK airports. Part II: impacts and policy assessment. Atmos Environ 2013;67:184-192.

23. Lonati G, Cernuschi S, Sidi S. Air quality impact assessment of at-berth ship emissions: case-study for the project of a new freight port. Sci Total Environ 2010;409(1):192-200.
24. Ross Z, English PB, Scalf R, Gunier R, Smorodinsky S, Wall S, et al. Nitrogen dioxide prediction in Southern California using land use regression modeling: potential for environmental health analyses. J Expo Sci Environ Epidemiol 2006;16(2):106-114.
25. Kashima S, Yorifuji T, Tsuda T, Doi H. Application of land use regression to regulatory air quality data in Japan. Sci Total Environ 2009;407(8):3055-3062.
26. European Environment Agency. EMEP/EEA air pollutant emission inventory guidebook 2013 [cited 2015 Oct 20]. Available from: http://www.eea.europa.eu/publications/emep-eea-guidebook-2013.
27. US Environmental Protection Agency. 2014 NEI plan [cited 2015 Oct 20]. Available from: http://www3.epa.gov/ttn/chief/net/2014nei_files/2014_nei_plan.pdf.
28. Rosenlund M, Forastiere F, Stafoggia M, Porta D, Perucci M, Ranzi A, et al. Comparison of regression models with land-use and emissions data to predict the spatial distribution of traffic-related air pollution in Rome. J Expo Sci Environ Epidemiol 2008;18(2):192-199.
29. Rashed T, Jürgens C. Remote sensing of urban and suburban areas. London: Springer; 2010, p. 219-244.
30. Korea Ministry of Environment. Guidelines for installation and management of national air quality monitoring networks Seoul: Korea Ministry of Environment; 2011, p. 3-60 (Korean).
31. Korea National Institute of Environmental Research; Korea Ministry of Environment. Annual report of ambient air quality in Korea. Seoul: Korea Ministry of Environment; 2010, p. 97-273, 461-466. (Korean).
32. Takeuchi W, Yasuoka Y. Development of cloud and shadow free compositing technique with MODIS QKM. In: American Society of Photogrammetry and Remote Sensing. Proceedings of the 2006 Conference of American Society of Photogrammetry and Remote Sensing; 2006 May 1; Reno, NV, USA. Bethesda: American Society of Photogrammetry and Remote Sensing: 2006, p. 1-8.
33. Ministry of Government Legislation. Administrative rule on the structure and installation of road [cited 2015 Oct 15]. Available from: http://www.law.go.kr/LSW/lsInfoP.do?lsiSeq=138166#0000 (Korean).
34. Su JG, Jerrett M, Beckerman B. A distance-decay variable selection strategy for land use regression modeling of ambient air pollution exposures. Sci Total Environ 2009;407(12):3890-3898.
35. Olvera HA, Garcia M, Li WW, Yang H, Amaya MA, Myers O, et al. Principal component analysis optimization of a PM2.5 land use regression model with small monitoring network. Sci Total Environ 2012;425:27-34.
36. Sampson PD, Richards M, Szpiro AA, Bergen S, Sheppard L, Larson TV, et al. A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2.5 concentrations in epidemiology. Atmos Environ (1994) 2013;75:383-392.
37. Beelen R, Voogt M, Duyzer J, Zandveld P, Hoek G. Comparison of the performances of land use regression modelling and dispersion modelling in estimating small-scale variations in long-term air pollution concentrations in a Dutch urban area. Atmos Environ 2010; 44(36):4614–4621.
38. Gulliver J, de Hoogh K, Fecht D, Vienneau D, Briggs D. Comparative assessment of GIS-based methods and metrics for estimating long-term exposures to air pollution. Atmos Environ 2011;45(39):7072–7080.
39. Marshall JD, Nethery E, Brauer M. Within-urban variability in ambient air pollution: comparison of estimation methods. Atmos Environ 2008;42(6):1359–1369.
40. Sellier Y, Galineau J, Hulin A, Caini F, Marquis N, Navel V, et al. Health effects of ambient air pollution: do different methods for estimating exposure lead to different results? Environ Int 2014;66:165-173.
41. Wu J, Wilhelm M, Chung J, Ritz B. Comparing exposure assessment methods for traffic-related air pollution in an adverse pregnancy outcome study. Environ Res 2011;111(5):685-692.
42. Lindström J, Szpiro AA, Sampson PD, Oron AP, Richards M, Larson TV, et al. A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. Environ Ecol Stat 2014;21(3):411-433.
43. Kloog I, Nordio F, Coull BA, Schwartz J. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal PM2.5 exposures in the Mid-Atlantic states. Environ Sci Technol 2012;46(21):11913-11921.
44. Kim SY, Yi SJ, Eum YS, Choi HJ, Shin H, Ryou HG, et al. Ordinary kriging approach to predicting long-term particulate matter concentrations in seven major Korean cities. Environ Health Toxicol 2014;29:e2014012.
45. Kim SY, Sheppard L, Kim H. Health effects of long-term air pollution: influence of exposure prediction methods. Epidemiology 2009;20(3):442-450.

Table S1. Coordinate systems used in different data sources

Data source	Geographical coordinate system	Projection	Origin		Added values to origin		Scale factor
			Longitude	Latitude	Easting	Northing	
KTDB	Bessel 1841	TM	128	38	400000	600000	0.9999
SGIS	Bessel 1841	TM	128.00289	38	200000	500000	1
IIS	WGS 1984	-	-	-	-	-	-
EGIS	ITRF 2000	TM	127	38	200000	500000	1

KTDB, Korean Transport Database; SGIS, Statistical Geographic Information Service; IIS, Institute of Industrial Science, University of Tokyo; EGIS, Environmental Geographical Information Service; WGS, World Geodetic System; ITRF, International Terrestrial Reference System; TM, Transverse Mercator

Table S2. Minimum widths of roads depending on type of road, speed limit, and type of area in the Administrative Rule on the Structure and Installation of Road

Type of road	Speed limit (km/h)	Type of area	
		Non-urban	Urban
Highway		3.50	3.50
Non-highway	≥80	3.50	3.25
	≥70	3.25	3.00
	≥60	3.25	3.00
	<60	3.00	3.00

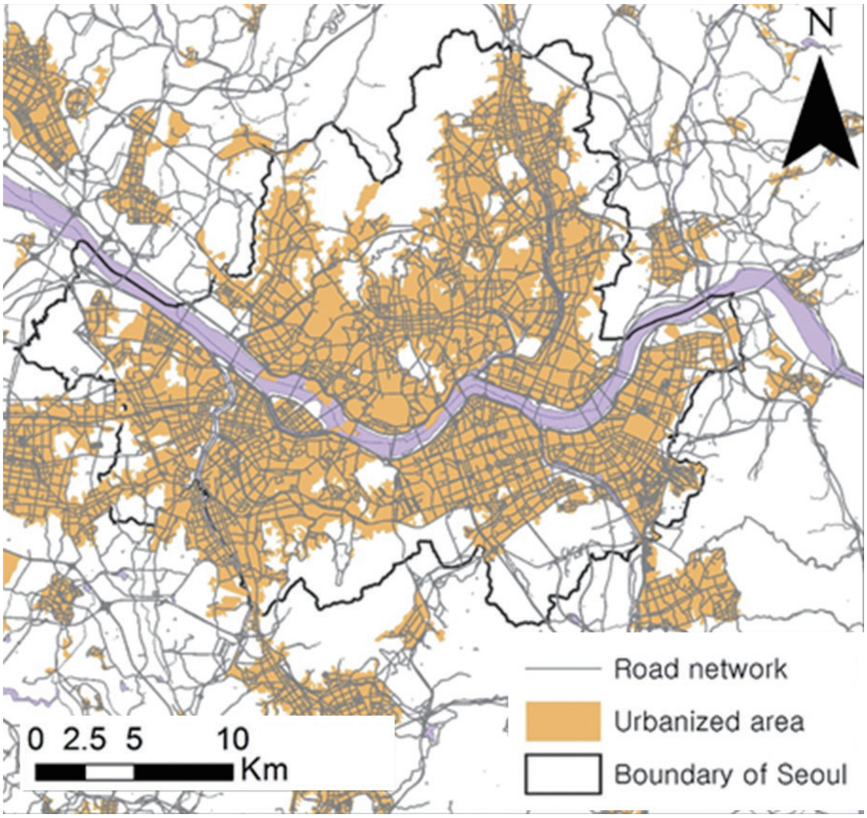

Figure S1. Road networks and urbanized areas in Seoul, Korea.