**The impact of limited residential address data on health effect analysis of predicted air pollution in a simulation study**

Running title: Limited addresses in air pollution prediction and health effect analysis

Yoon-Bae Jun, Insang Song, Sun-Young Kim\*

Yoon-Bae Jun

Department of Statistics, Seoul National University,

Gwanak-gu, Seoul 08826, Korea.

Insang Song

Department of Geography, University of Oregon, Eugene, Oregon, 97403, United States

**\*Corresponding author**

Sun-Young Kim, PhD

Department of Cancer Control and Population Health

Graduate School of Cancer Science and Policy, National Cancer Center

Goyang-si, Gyeonggi-do, 10408, Korea

Phone: +82-31-920-2745

Email: [sykim@ncc.re.kr](mailto:sykim@ncc.re.kr)

**Abstract**

Recent studies have developed spatially-resolved air pollution predication models that estimate air pollution concentrations at people’s addresses. However, the limited address data derived by unintended designs or confidentiality concerns hampered enjoying the benefit of advanced prediction models. Using the simulated disease (LBW) status and long-term ambient air pollution (PM10) concentrations at pre-selected home addresses obtained from the real world data analysis, we set up eight environmental scenarios and three data availability conditions to investigate the impact of limited use of residential address in practice. We applied various exposure prediction methods based on true exposures at regulatory monitoring sites and the relevant census data. We evaluated health effect estimates of predicted PM10 on LBW status using logistic regression. First, we verified that the UK-based exposure prediction model could result in alleviating under-estimation of relative risk of particulate matter. Second, the negative impact of using partially known residential address could be considerably eliminated with the help of the statistical modeling with the use of finer spatially located census grids. Third, model-based exposure prediction could be more efficient than simple area-averaging even for the research environment in Soul, Korea, 2010. Finally, we will discuss what should be also took into account so as to apply our implications to general air pollution epidemiological studies.

**Key words**: Address, Exposure prediction, Health effect, Long-term exposure, Particulate matter

**Introduction**

Long-term exposure to air pollution was associated with mortality and morbidity in many studies based on large health data such as multi-city or multi-country cohorts.1–4. Accurate assessment of individual exposure to long-term air pollution has been crucial in these studies because individual air pollution measurements are not available given financial and technical constraints. Early studies assessed individual exposure relying on available air pollution monitoring data only. For example, these studies commonly assigned area averages of monitoring data to the people residing in the corresponding areas. Recent studies of advanced approaches enhanced the capacity to assess the association between air pollution and health by representing substantial spatial variability of exposures across study participants. Studies used air pollution prediction models to estimate air pollution concentrations at people’s homes or workplaces as their individual-level exposures 5,6. Specifically, physicochemical models used emissions and meteorology data and estimate air pollution concentrations on the grid 7,8. Statistical models were mostly constructed in pointwise regression including geographical or meteorological characteristics, named land use regression9–11. Additional spatial correlation structure was modelled by using geostatistical techniques such as kriging12,13 and spatial smoothing as applied in generalized additive model14.

The benefit of exposure modelling can be maximized, when complete address data are available. However, it is common that address information is available at the coarse spatial scale such as census tract, zipcode area, and district15,16. Existing cohorts were often not designed to collect full address data at baseline 1,17. This limitation is more common in administrative health data constructed based on census or health insurance. Despites their strength of large and representative populations that allow examining the association at the national or regional scale, address data were restricted given the concerns of confidentiality 18–25. For instance, studies using the U.S. Medicare cohort, the Canadian Census cohort, and Taiwan National Health Insurance Database cohort assessed individual exposure to air pollution based on zipcode-level addresses19,26,27. This limited address data may increase exposure misclassification and affect accuracy and/or precision of health effect estimates. Although there has been increasing application of individual address data for studies of long-term air pollution and health, few studies investigated their impact on health analysis.

This simulation study aimed to understand the impact of the extent of availability for address information on exposure prediction and health effect estimation. To reflect realistic environments, the simulation was designed based on our previous study that investigated the association between long-term exposure to particulate matter less than or equal to 10 micron per diameter (PM10) and low birth weight (LBW) in Seoul, Korea28.

**Methods**

*Overview*

Our simulation procedure consists of three steps. First, we performed exploratory data analyses and obtained parameters for the underlying distributions of PM10 and LBW. Using the underlying distributions and parameters, we generated annual-average PM10 concentrations at all locations including mothers’ homes and regulatory monitoring sites according to eight different environmental scenarios (ES1-ES8) and generated LBW status of mothers. Second, we constructed three data availability conditions (DAC1-DAC3) for PM10 measurements and home addresses of mothers, and estimated individual-level PM10 exposures at mothers’ homes using “true” (observed) concentrations at monitoring sites and nine prediction methods. Finally, we conducted health analysis of LBW for individual-level PM10 concentrations and compared the performance of health effect estimates across different scenarios of data availability, exposure environment, and prediction method.

*Data Analysis and Parameter Acquisition*

We obtained parameters to be applied to our simulation study from the exploratory analysis of air quality regulatory monitoring data for PM10, Geographic Information System (GIS) -based geographic variables, and birth certificate data in Seoul, South Korea, during 2010. Hourly PM10 measurement data in the regulatory monitoring network were obtained from the National Institute of Environmental Research (NIER). The air quality regulatory monitoring network in Seoul includes 25 urban-background and 12 urban-roadside sites in 2010. Urban-background sites are located in heavily populated residential areas with the aim of assessing the population level of exposure; one monitoring site is deployed in each of the 25 districts in Seoul (2016 NIER annual report). In contrast, urban-roadside sites are located next to large and busy major roads for monitoring air pollution affected by traffic emissions. Using hourly measurements, we computed daily averages at each site for the days when the number of hourly measurements is more than 75%. Then, the annual average concentrations were computed at each site where there is at least one daily measurement in each of the 10 months and in any 45 consecutive days29.

Using annual average concentrations of PM10, we fitted variogram model and estimated mean and variance parameters. Three variance parameters include range, partial sill, and nugget that indicate the distance in which spatial correlation exists, spatial variability, and non-spatial variability, respectively30. Five mean parameters were regression coefficients of five GIS-based geographic variables that were most related to fine particulate matter less than or equal to 2.5 micron per diameter (PM2.5) in our previous study for Seoul29. The data sources and computation procedure of GIS-based geographic variables were described elsewhere31. These variables include the length of major roads in a 100 m circular buffer, the proportion of water surface land use in 500 m, the number of construction companies in 1,000 m, the distance to the nearest bus stop, and the number of employees in construction industries in 100 m.

We obtained birth certificate data from the Statistical Geographic Information Service (SGIS) operated by the Statistics Korea. Term LBW babies were defined as singleton live birth between 37 to 41 weeks with less than 2.5 kilograms28. We computed the proportion of LBW cases to the total births (0.01561) and used in the following simulation as the true LBW proportion. The underlying coefficient of PM10 on LBW (0.003275) was obtained from the previous study of the association between PM10 and low birth weight in 549,270 non-employed mothers residing in Seoul, Korea, for 2002-201228.

*Residential Address Assignment*

Because mothers’ home addresses in birth certificate data are available at the district level, we generated the locations of mothers’ homes based on the number of births. Seoul, the Capital of South Korea, is composed of 25 districts (“Si-Gun-Gu”, median area and average population in 2010: 21.59 km2 and 412,520) and 422 neighbourhoods (“Eup-Myun-Dong”: 8.69 km2 and 24,323). There are 16,230 census tracts in Seoul as the smallest census territorial unit nested within neighbourhoods with the median area of 0.02 km2 and average population of 821. To focus on the spatial variation in our simulation, we restricted our study period to a single year in 2010 and selected 46,007 mothers who had birth in 2010. After hypothesizing that census tract centroids are potential home addresses of mothers, we randomly sampled the census tract centroids for the mothers in each district based on the weight of the numbers of live births across neighbourhoods of each district. We treated these locations fixed over the simulations.

|  |
| --- |
| Figure01. Maps of 25 urban-background regular monitoring sites (red) and the government offices of 25 districts (black) in Seoul (a), neighborhood community centers (light blue) and census tracts centroids (light purple) in a red-lined district in (a) (b), and 1-km grid coordinates (green) (c) |

*Exposure Generation*

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

We generated true annual average PM10 concentrations, as true exposures to PM10, at all locations including 46,007 mothers’ homes, 37 air quality regulatory monitoring sites, and the centroids of 25 districts and 16,230 census tracts in Seoul. As shown in the equation (1), we assumed that the true exposure to PM10 at a location s, , follows Gaussian random field, and is partitioned into three components of global mean,, spatial error,, and non-spatial error,. is a linear function of predictors, , characterized by mean parameters, . These predictors include five geographic variables that were most associated with particulate matter concentrations in our previous study29. is a spatial random process characterized by partial sill, , and range, , that represent spatial variability and the distance in which spatial correlation exists, respectively. is a non-spatial random process indicating measurement error and characterized by nugget, , representing non-spatial variability. Consequently, the total variability of the true PM10 annual-average concentrations,, can be partitioned into global trend variability, , partial sill, , and nugget, .

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table01. Spatial structure and related variance parameters of true PM10 concentrations by eight environmental scenarios   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | Environmental Scenario  (ES) | Variance parameter\* | | | Spatial structure\*\* | | | | Nugget ( | Partial sill ( | Range (;m) | Mean  structure | Spatial variability | Non-spatial variability | | ES1 | 1.00 | 30.94 | 5,885 | None | Dominant | Little | | ES2 | 6.86 | 28.98 | 9,609 | None | High | Low | | ES3 | 11.51 | 34.71 | 20,355 | None | Low | High | | ES4 | 22.00 | 13.77 | 27,000 | None | Little | Dominant | | ES5 | 1.00 | 16.98 | 2,524 | Moderate | High | Low | | ES6 | 6.86 | 12.17 | 4,820 | Moderate | Low | High | | ES7 | 1.00 | 10.00 | 1,100 | Dominant | High | Low | | ES8 | 6.86 | 3.60 | 1,004 | Dominant | Low | High |   \*To avoid an optimality issue, we assumed the true partial sill ≥ 1 and the true range parameter ≥ 1 km.  \*\*Identified based on different contribution of three variability components (mean structure, and spatial and non-spatial variability) to total variability |

To represent possibly different spatial structures of true PM10 annual average concentrations, we constructed eight environmental scenarios (ES1-ES8) based on varying contributions of the three components of PM10 to total variability (Table 01, FigureS1). ES1 to ES4 has a constant mean (i.e. ) and different variance parameters of range, partial sill, and nugget which represent different contribution to spatial and non-spatial variability. From ES1 to ES4, the contribution of spatial variability decreases while the contribution of non-spatial variability increases. The other four scenarios (ES5-ES8) include mean structures characterized by five geographic variables (i.e. ) in addition to different combination of variance parameters. Whereas ES5 and ES6 have intermediate contribution of mean structure, ES7 and ES8 have dominant contribution. ES5 and ES7 contain more spatial variability than non-spatial variability, while ES6 and ES8 comprise more non-spatial variability than spatial variability. ES8 was constructed by using the parameters that best represent the data, indicating the most similar environment to Seoul.

*Three Data Availability Conditions*

Table02. Data availability conditions for individual-level PM10 concentrations and home addresses of 46,007 mothers in Seoul, Korea.

|  |  |  |
| --- | --- | --- |
| Data Availability Condition (DAC) | PM10 concentration | Home address |
| DAC1 | Mothers’ homes | Complete |
| DAC2 | Regulatory monitoring sites | Complete |
| DAC3 | Regulatory monitoring sites | Limited to district |

We hypothesized three data availability conditions (DACs) that represent ideal to realistic conditions of data availability for PM10 measurements and home addresses of mothers (Table03). DAC1 assumes that PM10 measurements are fully available to all 46,007 mothers at their homes. For DAC2, complete home address data of mothers are available but PM10 measurements are available only at 37 regulatory monitoring sites. The most limited data availability condition, DAC3, assumes that residential addresses are partially known at the district level and PM10 measurements are also limited to 37 regulatory monitoring sites.

*Outcome Generation*

LBW status of mothers were generated based on the proportion of LBW cases obtained by our exploratory data analysis (see the Data Analysis and Parameter Acquisition section), simulated true PM10 concentrations (see the Exposure Generation section), and the effect estimate of LBW for PM10 obtained by our previous study28. We assumed that LBW probability of a mother, p, follows an inverse logit function of true baseline LBW rate () and the true effect estimate of LBW () for individual PM10 concentration (X), as shown in the equation 2.

|  |  |  |
| --- | --- | --- |
|  | . | (2) |

Then, given LBW probability of each mother, p, we randomly generated LBW status from Bernoulli distribution.

*Exposure Prediction*

We applied nine exposure prediction methods based on five prediction modelling approaches according to three DACs. In DAC1, we used true PM10 concentrations at all mothers’ homes (TE) without applying any prediction methods. In contrast, DAC2 and DAC3 where PM10 measurements are available only at 37 regulatory monitoring sites require exposure prediction approaches that allow us to estimate PM10 concentrations at mothers’ homes based on true PM10 concentrations at monitoring sites.

For DAC2 where PM10 measurements are available only at regulatory monitoring sites but mothers’ home addresses are fully available, we applied four prediction methods commonly used in previous studies with the same data availability conditions. In the nearest monitor (NM) method, we assigned PM10 concentrations measured at the nearest regulatory monitoring site to the home address of each mother. Inverse distance weighted average (IDWA) produces the average concentration across regulatory monitoring sites weighted by inverse squared Euclidean distance from each home. Land use regression (LUR) assumes that concentrations are represented by geographic characteristics of direct or indirect pollution sources. These characteristics were computed as GIS variables and included as covariates in regression equations. Consequently, we computed mothers’ PM10 concentrations by using regression parameters and GIS variables at their homes. Universal kriging (UK) is a geostatistical method which optimally derives interpolated concentrations based on mean structure and spatial correlation. In our application, we built UK model using the same five geographic variables as in LUR. Along with estimated regression and variance parameters and GIS variables at mothers’ addresses, we computed PM10 concentrations of mothers. Out of the 37 regulatory monitoring sites in Seoul, we used underlying PM10 concentrations from 25 urban background sites for NM, IDWA, and area averaging, and from all 37 sites including 12 urban roadside sites for LUR and UK.

In DAC3 with limited PM10 and address data, we applied area averaging (AA) and four modified versions of UK. AA assigns the average concentration across all monitoring sties in an area to all mothers living the same area. Although this approach is also applicable to DAC2, we considered AA application to DAC3 only because area averages would not be a preferred option when address data are fully available. Since the regulatory monitoring network in Seoul has one urban background site in every district, we treated the PM10 concentration at a single site as a special case of AA. In addition to AA, we developed four additional approaches based on UK to compute population-representative exposure. When detailed address data are not available as shown in many administrative cohort data but pointwise prediction models are accessible, alternative approaches could be the aggregation of predictions at many representative points. First, we predicted PM10 concentrations using UK at the government offices of 25 districts and assigned to the mothers living in the same districts (UKD). In addition, to represent population exposure at the fine spatial level, we used UK predictions at the community centers of 422 neighbourhoods and the centroids of 16,230 census tracts, and computed district averages (UKNA and UKCA). Finally, we replaced with UK predictions at 610 1-km grid coordinates to average to the district in order to represent the PM10 distribution over space more than population exposure.

*Health Effect Estimation*

Using true or predicted PM10 and true LBW status of mothers, we estimated health effects of LBW for PM10 using logistic regression. Then, we repeated the whole procedure from exposure generation to health effect estimation 1,000 times. Finally, we computed properties of health effect estimates () over 1,000 simulations to evaluate accuracy of effect estimates depending on different scenarios for data availability, exposure prediction methods, and pollution environments. These properties include bias (), root mean square error (RMSE) (), average of standard error (ASE) (), and coverage probability (CP) (), where if the condition is true, and otherwise. RMSE and ASE indicate the uncertainty of effect estimates on average, while CP presents the probability of including the true effect estimate within estimated 95% confidence intervals. In addition, to explore the impact of data availability and prediction approaches on statistical power to detect the true association, we computed the ratio of the number of simulations that provide significantly positive effect estimates (p-value < 0.05) in DAC2 or DAC3 to the number of significant simulations in DAC1.

**Results**

*True and predicted PM10 concentrations*

Table 04 and FigureS2 summarize true and predicted annual-average PM10 concentrations at home addresses of 46,007 mothers by data availability, environmental scenarios, and prediction methods. Mothers’ PM10 concentrations predicted at their homes in DAC2 and DAC3 were generally similar on average but less variable compared to true concentrations in DAC1. Variability was even smaller in DAC3 with limited address information at the district particularly for averaged exposure of UKNA, UKCA, and UKGA, compared to that in DAC2 with complete addresses. Average PM10 concentration across eight ESs was ranged between 46.90 and 58.55, 47.43 and 60.52, and 47.25 and 59.66 µg/m3 in DAC1, DAC2, and DAC3, respectively (TableS8). The correlation with TE was generally higher in UK and UKCA compared to other prediction methods across all ESs (Pearson correlation coefficient= 0.26-0.70) (TableS7). NM, IDWA, AA and UKD gave good correlation when there is no mean structure with some spatial correlation, but poor correlation otherwise. In contrast, LUR provided high correlation (0.60-0.65) when there is a dominant mean structure in ES7 and ES8, but low correlation less than 0.1 without a mean structure. ES4 that assumes large non-spatial variability particularly showed low correlation across all prediction approaches.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table04. Mean and standard deviation of true and predicted PM10 annual average concentrations at 46,007 home addresses of mothers under three data availability scenarios (DAC1-DAC3), eight environmental scenarios (ES1-ES8), and six exposure prediction methods (AA, UK, UKD, UKNA, UKCA, and UKGA) at the 1st simulation   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | |  | DAC1 | DAC2 | DAC3 | | | | | |  | TE | UK | AA | UKD | UKNA | UKCA | UKGA | | ES1 | 51.16(5.56)\* | 52.27(1.56) | 51.59(4.24) | 51.63(4.53) | 51.61(4.24) | 51.6(4.27) | 51.49(3.97) | | ES2 | 49.8(5.97) | 49.97(1.59) | 49.71(4.13) | 49.91(4.8) | 49.68(4.49) | 49.81(4.38) | 49.49(4.04) | | ES3 | 58.55(6.45) | 60.52(3.43) | 59.56(4.56) | 59.66(4.46) | 59.45(4.63) | 59.60(4.09) | 59.21(4.02) | | ES4 | 46.9(5.11) | 47.52(1.99) | 47.61(2.56) | 47.57(2.29) | 47.42(1.21) | 47.52(1.50) | 47.25(1.04) | | ES5 | 50.91(5.26) | 51.51(4.30) | 51.46(4.16) | 51.51(4.33) | 51.26(2.65) | 51.41(4.07) | 51.19(2.13) | | ES6 | 53.05(4.66) | 53.21(3.45) | 53.32(4.08) | 53.37(3.54) | 53.19(2.07) | 53.22(2.86) | 52.99(2.00) | | ES7 | 51.61(5.81) | 51.97(5.09) | 51.89(4.43) | 51.42(5.33) | 51.22(2.48) | 51.24(4.16) | 51.28(2.98) | | ES8 | 51.19(5.17) | 51.35(4.32) | 51.97(4.29) | 51.34(4.64) | 51.21(1.74) | 51.19(3.14) | 51.13(2.41) |   \* Mean (Standard deviation)  TE: true-exposure; NM: nearest-monitor; IDWA: inverse-distance-weighted-average; LUR: land use regression; UK: universal kriging; AA: area-average UKD: UK predictions at district governmental offices; UKNA: district averages of UK predictions at neighbourhood community centres; UKCA: district averages of UK predictions at census tract centroids; UKGA: district averages of UK predictions at 1-km grid coordinates |

*Health effect estimate properties by three data availability conditions*

Performance of effect estimates of LBW for annual-average PM10 concentrations became worse when data availability of measurements or address data were limited. Table05 shows properties of health effect estimates in four ESs including ES2, ES3, ES5, and ES8 where different exposure environments are more distinct, while Table S6, Figure S4, and Figure S5 show all eight scenarios. Compared to DAC1 that showed gave good performance, bias and RMSE tended to increase in DAC2 and DAC3 with slightly larger increase in DAC3 than in DAC2 depending on the prediction method. Performance varied more across different prediction methods and environmental scenarios in DAC2 than in DAC3. Both DAC2 and DAC3 gave CPs close to 0.95. Ratios of significantly positive simulations compared to those in DAC1 indicating statistical power were generally lower in DAC3 than DAC2 with more variation in DAC2 (Figure S6).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table05. Properties of effect estimates of true and predicted PM10 annual average concentrations on low birth weight over 1,000 simulations by three data availability conditions (DAC1-DAC3), six exposure prediction methods (AA, UK, UKD, UKNA, UKCA, UKGA), and four environmental scenarios (ES2, ES3, ES5, ES8)   |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  | | Bias1) | RMSE2) | ASE3) | CP4) |  | Bias1) | RMSE2) | ASE3) | CP4) | |  | ES2 |  |  |  |  | ES3 |  |  |  |  | | DAC1 | TE | -0.04 | 0.015 | 0.008 | 0.95 | TE | 0.03 | 0.015 | 0.008 | 0.95 | | DAC2 | UK | -0.03 | 0.024 | 0.015 | 0.94 | UK | -0.05 | 0.027 | 0.016 | 0.94 | | DAC3 | AA  UKD  UKNA  UKCA  UKGA | -0.15  -0.08  -0.01  0.00  0.09 | 0.015  0.024  0.033  0.028  0.038 | 0.008  0.014  0.019  0.017  0.024 | 0.94  0.94  0.94  0.94  0.94 | AA  UKD  UKNA  UKCA  UKGA | -0.14  -0.09  0.07  0.05  0.04 | 0.016  0.027  0.041  0.033  0.048 | 0.008  0.016  0.027  0.020  0.036 | 0.95  0.94  0.94  0.95  0.94 | |  | ES5 |  |  |  |  | ES8 |  |  |  |  | | DAC1 | TE | 0.03 | 0.015 | 0.008 | 0.95 | TE | 0.00 | 0.014 | 0.007 | 0.95 | | DAC2 | UK | 0.01 | 0.021 | 0.011 | 0.95 | UK | -0.01 | 0.017 | 0.009 | 0.96 | | DAC3 | AA  UKD  UKNA  UKCA  UKGA | -0.19  -0.16  0.04  0.01  0.05 | 0.018  0.020  0.031  0.023  0.037 | 0.009  0.011  0.017  0.013  0.020 | 0.95  0.94  0.94  0.94  0.95 | AA  UKD  UKNA  UKCA  UKGA | -0.25  -0.24  0.06  0.02  -0.11 | 0.020  0.019  0.041  0.022  0.033 | 0.010  0.010  0.021  0.012  0.017 | 0.95  0.95  0.94  0.96  0.95 |   1) Multiplied by 100  2) Root mean square error  3) Average of standard error  4) Coverage probability with 95% confidence interval |

*Health effect estimate properties by eight exposure prediction methods*

All prediction methods in DAC2 and AA and UKD in DAC3 generally showed negative bias indicating under-estimated health effect estimates: these biases were particularly large for NM, AA, and UKD relying on the observations at single monitoring sites. UK in DAC2 gave lower bias and RMSE than other approaches across all ESs. This good performance was notably prominent in ES5 to ES8 that includes a mean structure. NM tended to provide small RMSE and ASE, but relatively large negative bias, while IDWA gave relatively small bias but large RMSE and ASE. LUR provided good performance only when there is a mean structure in the true environment (ES5 to ES8). Three UK-based district averages showed much smaller bias with either direction and slightly higher RMSE and ASE compared to the other two prediction methods of AA and UKD in DAC3. Among these three approaches, UKCA as the exposure averaged over a large number of population-representative points at the fine spatial scale showed better performance than UKNA and UKGA based on coarse spatial-scale points and spatially-representative points, respectively. Ratios of positive simulation were also generally higher in UK and UK-based district averages.

*Health effect estimate properties by eight environmental scenarios*

Although some prediction approaches showed consistent patterns over all ESs, other approaches gave different performance depending on the mean structure and spatial correlation. UK in DAC2 and UKCA in DAC3 generally gave the smallest bias, relatively small RMSE and ASE, and high ratio of positive simulation across all eight ESs, while NM in DAC2 and AA and UKD in DAC3 consistently showed large bias (Table S6, Figure S4). Larger uncertainty in UKNA and UKGA, possibly resulting from small variability in PM10, was also consistent across ESs. All prediction methods showed large bias and small ratios of positive simulation in ES4. Bias in NM, IDWA, AA, and UKD with predictions by using the observations from single monitors was the largest in ES8 that best represents the real pollution environment in Seoul.

**Discussion**

Our study focused on the impact of limited availability of address data on exposure prediction and health effect estimation compared to complete availability. Specifically, we hypothesized three data availability conditions, and explored their impacts on the performance of health effect estimates depending on different environmental scenarios and exposure prediction methods based on the real-world example of the association between PM10 and LBW. Eight environmental scenarios represent various pollution environments determined by geographical site, type of residence, and/or pollutant compositions in a given area. Furthermore, nine prediction methods exhibit commonly-applied approaches of individual exposure assessment given limited monitoring data. Our study showed that the UK-based method under incomplete data availability can be quite good exposure allocation strategy to achieve similar level of health effect estimation to those under complete data availability.

Our simulation study intended to answer an important research question that can help inference of epidemiological studies of air pollution relying on limited address data of subjects. Even though many existing cohorts and/or administrative health data can provide address information only at the coarse spatial scale, it will be much easier for researchers to get spatially finer located census grids with sufficient population density information. If so, we indicate that the impact of limited address data can be considerably eliminated by the help of the statistical modeling with the use of finer spatially located census grids. One example is provided in Choe et al 201832. This study reported that increased concentrations of PM10 during controlled ovarian stimulation and after embryo transfer were associated with a decreased probability of intrauterine pregnancy. On the other hand, the result showed that increased concentrations of PM10 during the other periods of IVF cycle were also associated with a decreased probability of intrauterine pregnancy, but their statistical significance were not verified from the data. Note that this study used simple area-averaging exposure assessment to individuals who reside in Seoul, Korea. Based on our simulation studies, we can hypothesize that those effect size were likely to be under-estimated and those statistical significance could be improved by the use of kriging-based approach using the same census tract information as our simulation study. Another example is provided in Kim et al 201733. This study found the meaningful association between long-term exposure to PM10 and non-accidental mortality on a national scale in South Korea, but this study also argued that future studies should investigate the impact of exposure misspecification since it used only district-level addresses to assess individual level of measurements. Our simulation study can support the statement that the UK-based exposure assessment in Kim et al 201733 was better way than simple area-averaging to analyse the health effect within the range of data available. Even in different geographic areas, our simulation study can at least provide insights into how we can measure the impact of address use and exposure strategies in the future work.

Our findings generally showed that UK-based approaches gave good performance in health effect estimates consistently across different air pollution environments when individual air pollution measurements are not available. While UK showed better performance compared to other prediction approaches when complete address data are available in DAC2, UK averaging approaches outperformed with individual address data limited to the district in DAC3. One reason is that UK modelled by using both mean and variance structure well represent air pollution conditions at people’s residence even when there is no mean structure34. In addition, employment of population-representative locations in the following averaging process under the unavailability of precise residential addresses possibly minimized uncertainty of model misspecification. Bias was the smallest and also non-systematic as opposed to other prediction methods that gave consistently negative bias. Except for the simple approaches such as AA and NM, RMSE and ASE of UKCA, averaged based on UK predictions at census track centroids, remained the lowest level among the approaches in DAC3, and comparable to those of other prediction approaches in DAC2. It was less likely to detect statistically significant health effect estimates overall in DAC3 but the impact of limited residential address data was not considerable when UKCA was used. However, the benefit of UK-based averaging could be reduced when we use predictions at the locations including those irrelevant to population as shown in UKGA.

All prediction methods except for UK-based averaging generally showed underestimated health effects given limited exposure or address data. This underestimation can be explained by exposure measurement error derived by poor characterization of individual exposure in prediction models35,36. In our simulation, prediction methods heavily relying on a mean structure such as LUR gave greater underestimation when there is no mean structure in true exposure scenarios, while simple prediction approaches using measurements only shown as NM and IDWA gave larger underestimation when there is a mean structure. Prediction methods using a single location based on the nearest monitor (NM), a centralized monitor (AA), or district government office (UKD) also gave larger negative bias than other methods. In addition, poor assessment of individual exposure can result from poor representativeness of prediction points used for averaging. Our study showed larger positive or negative bias in UKGA using grid coordinates than UKCA based on census tract centroids. Previous simulation studies reported that measurement error derived by spatial misalignment between regulatory monitoring sites and people’s residences affected misspecification of the prediction model and resulted in positive or negative bias in health effect analysis37–39. Our finding of larger bias in UKGA suggested possible impact of spatial misalignment between prediction points and people’s residences when complete address data are unavailable and prediction models are applied to many representative points.

The application of our simulation using various environmental scenarios and parameters obtained from real data analyses suggests possible generalization to other pollutants and study areas. Although we focused on PM10 well known as a regional pollutant with relatively weak mean structure and large spatial correlation, we constructed seven environmental scenarios by assuming different spatial structure in addition to the ES8 based on the parameters estimated directly from the regulatory monitoring data in Seoul. This variation of spatial structure possibly represents more local and regional pollutants with stronger impacts of local sources and spatial homogeneity, respectively, and allows us to apply our findings to other pollutants such as fine particles and gaseous pollutants and different regions affected by different meteorology. In addition, our reliance on real-world data can improve the practical applicability of findings in this simulation study. We obtained exposure parameters from the data analyses of regulatory monitoring data in relatively dense monitoring network including more than 40 sites in Seoul with the area of 25 km39.

Our study includes several limitations to be further investigated in future research. First, we used low birth weight and logistic regression. Future studies should confirm whether our suggestions are consistent to different health outcomes and health analysis models. Second, we assumed mothers’ residential addresses are fixed although we sampled these addresses based on population distribution. Future studies should investigate the sensitivity of our findings depending on the true distribution of people’s residences. Finally, density of monitors is another potential factor to nullify some of our findings. The implication might differ according to how much real locations of monitoring sites represent the exposure density over the study region.39–41.

This study showed that prediction models well representing geographic environments and supplemented with population relatively locations when complete individual address data are unavailable improved accuracy and precision in health effect estimates. Our study can help understand the degree of accuracy and precision of health effect estimates and statistical power related to limited address data, and provide a guidance for a preferred approach to improve the performance.

**References**

1. Beelen, R. *et al.* Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. *Lancet* **383**, 785–795 (2014).

2. Cesaroni, G. *et al.* Long term exposure to ambient air pollution and incidence of acute coronary events: prospective cohort study and meta-analysis in 11 European cohorts from the ESCAPE Project. *BMJ* **348**, f7412 (2014).

3. Hoek, G. *et al.* Long-term air pollution exposure and cardio- respiratory mortality: a review. *Environ. Heal.* **12**, 43 (2013).

4. Di, Q. *et al.* Air Pollution and Mortality in the Medicare Population. *N. Engl. J. Med.* **376**, 2513–2522 (2017).

5. Hoek, G. *et al.* A review of exposure assessment methods for epidemiological studies of health effects related to industrially contaminated sites. *Epidemiol. Prev.* **42**, 21–36 (2018).

6. Hoek, G. Methods for Assessing Long-Term Exposures to Outdoor Air Pollutants. *Curr. Environ. Heal. Reports* **4**, 450–462 (2017).

7. Binkowski, F. S. & Roselle, S. J. Models-3 Community Multiscale Air Quality (CMAQ) model aerosol component 1. Model description. *J. Geophys. Res. Atmos.* **108**, (2003).

8. Hanha, S. R. Air Quality Model Evaluation and Uncertainty. *JAPCA* **38**, 406–412 (1988).

9. Brauer, M. *et al.* Estimating Long-Term Average Particulate Air Pollution Concentrations: Application of Traffic Indicators and Geographic Information Systems. *Epidemiology* **14**, 228–239 (2003).

10. Hoek, G. *et al.* A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos. Environ.* **42**, 7561–7578 (2008).

11. Jerrett, M. *et al.* A review and evaluation of intraurban air pollution exposure models. *J. Expo. Sci. Environ. Epidemiol.* **15**, 185–204 (2005).

12. Jerrett, M. *et al.* Spatial Analysis of Air Pollution and Mortality in Los Angeles. *Epidemiology* **16**, 727–736 (2005).

13. Sampson, P. D. *et al.* A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2.5 concentrations in epidemiology. *Atmos. Environ.* **75**, 383–392 (2013).

14. Paciorek, C. J., Yanosky, J. D., Puett, R. C., Laden, F. & Suh, H. H. Practical large-scale spatio-temporal modeling of particulate matter concentrations. *Ann. Appl. Stat.* **3**, 370–397 (2009).

15. Grubesic, T. H. & Matisziw, T. C. On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. *Int. J. Health Geogr.* **5**, 58 (2006).

16. Seong, S. C. *et al.* Cohort profile: the National Health Insurance Service-National Health Screening Cohort (NHIS-HEALS) in Korea. *BMJ Open* **7**, e016640 (2017).

17. Laden, F., Schwartz, J., Speizer, F. E. & Dockery, D. W. Reduction in Fine Particulate Air Pollution and Mortality. *Am. J. Respir. Crit. Care Med.* **173**, 667–672 (2006).

18. Kim, S.-Y. & Song, I. National-scale exposure prediction for long-term concentrations of particulate matter and nitrogen dioxide in South Korea. *Environ. Pollut.* **226**, 21–29 (2017).

19. L., Z. S., Francesca, D., Aidan, M. & M., S. J. Mortality in the Medicare Population and Chronic Exposure to Fine Particulate Air Pollution in Urban Centers (2000–2005). *Environ. Health Perspect.* **116**, 1614–1619 (2008).

20. Carey, I. M. *et al.* Mortality Associations with Long-Term Exposure to Outdoor Air Pollution in a National English Cohort. *Am. J. Respir. Crit. Care Med.* **187**, 1226–1233 (2013).

21. H., F. P. *et al.* Air Pollution and Mortality in Seven Million Adults: The Dutch Environmental Longitudinal Study (DUELS). *Environ. Health Perspect.* **123**, 697–704 (2015).

22. Huss, A., Spoerri, A., Egger, M., Röösli, M. & Group, S. N. C. S. Aircraft noise, air pollution, and mortality from myocardial infarction. *Epidemiology* **21**, 829–836 (2010).

23. Crouse, D. L. *et al.* Ambient PM2.5, O₃, and NO₂ Exposures and Associations with Mortality over 16 Years of Follow-Up in the Canadian Census Health and Environment Cohort (CanCHEC). *Environ. Health Perspect.* **123**, 1180–1186 (2015).

24. Hansell, A. *et al.* Historic air pollution exposure and long-term mortality risks in England and Wales: prospective longitudinal cohort study. *Thorax* **71**, 330–338 (2016).

25. Okayama, A. *et al.* Dietary sodium-to-potassium ratio as a risk factor for stroke, cardiovascular disease and all-cause mortality in Japan: the NIPPON DATA80 cohort study. *BMJ Open* **6**, e011632 (2016).

26. Peters, P. A. *et al.* Data resource profile: 1991 Canadian Census Cohort. *Int. J. Epidemiol.* **42**, 1319–1326 (2013).

27. Jung, C.-R., Lin, Y.-T. & Hwang, B.-F. Ozone, particulate matter, and newly diagnosed Alzheimer’s disease: a population-based cohort study in Taiwan. *J. Alzheimer’s Dis. JAD* **44**, 573–584 (2015).

28. Choe, S.-A., Jang, J., Kim, M. J., Jun, Y.-B. & Kim, S.-Y. Association between ambient particulate matter concentration and fetal growth restriction stratified by maternal employment. *BMC Pregnancy Childbirth* **19**, 246 (2019).

29. Yi, S.-J. *et al.* Association between Exposure to Traffic-Related Air Pollution and Prevalence of Allergic Diseases in Children, Seoul, Korea. *BioMed Research International* (2017).

30. Cressie, N. *Statistics for Spatial Data*. (Wiley-Interscience, 2015).

31. Eum, Y., Song, I., Kim, H.-C., Leem, J.-H. & Kim, S.-Y. Computation of geographic variables for air pollution prediction models in South Korea. *Environ. Health Toxicol.* **30**, (2015).

32. Choe, S. A., Jun, Y. B., Lee, W. S., Yoon, T. K. & Kim, S. Y. Association between ambient air pollution and pregnancy rate in women who underwent IVF. *Hum. Reprod.* **33**, 1071–1078 (2018).

33. Kim, O. J., Kim, S. Y. & Kim, H. Association between long-term exposure to particulate matter air pollution and mortality in a South Korean national cohort: Comparison across different exposure assessment approaches. *Int. J. Environ. Res. Public Health* **14**, (2017).

34. Kim, S. Y., Sheppard, L. & Kim, H. Health effects of long-term air pollution: Influence of exposure prediction methods. *Epidemiology* **20**, 442–450 (2009).

35. Sheppard, L. *et al.* Confounding and exposure measurement error in air pollution epidemiology. *Air Qual. Atmos. Health* **5**, 203–216 (2012).

36. Szpiro, A. A., Sheppard, L. & Lumley, T. Efficient measurement error correction with spatially misaligned data. *Biostatistics* **12**, 610–623 (2011).

37. Szpiro, A. A., Paciorek, C. J. & Sheppard, L. Does more accurate exposure prediction necessarily improve health effect estimates? *Epidemiology* **22**, 680–685 (2011).

38. Szpiro, A. A. & Paciorek, C. J. Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics* **24**, 501–517 (2013).

39. Lee, A., Szpiro, A., Kim, S. Y. & Sheppard, L. Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology. *Environmetrics* **26**, 255–267 (2015).

40. Sellier, Y. *et al.* Health effects of ambient air pollution: Do different methods for estimating exposure lead to different results? *Environ. Int.* **66**, 165–173 (2014).

41. Lepeule, J. *et al.* Maternal exposure to nitrogen dioxide during pregnancy and offspring birth weight: comparison of two exposure models. *Environ. Health Perspect.* **118**, 1483–1489 (2010).

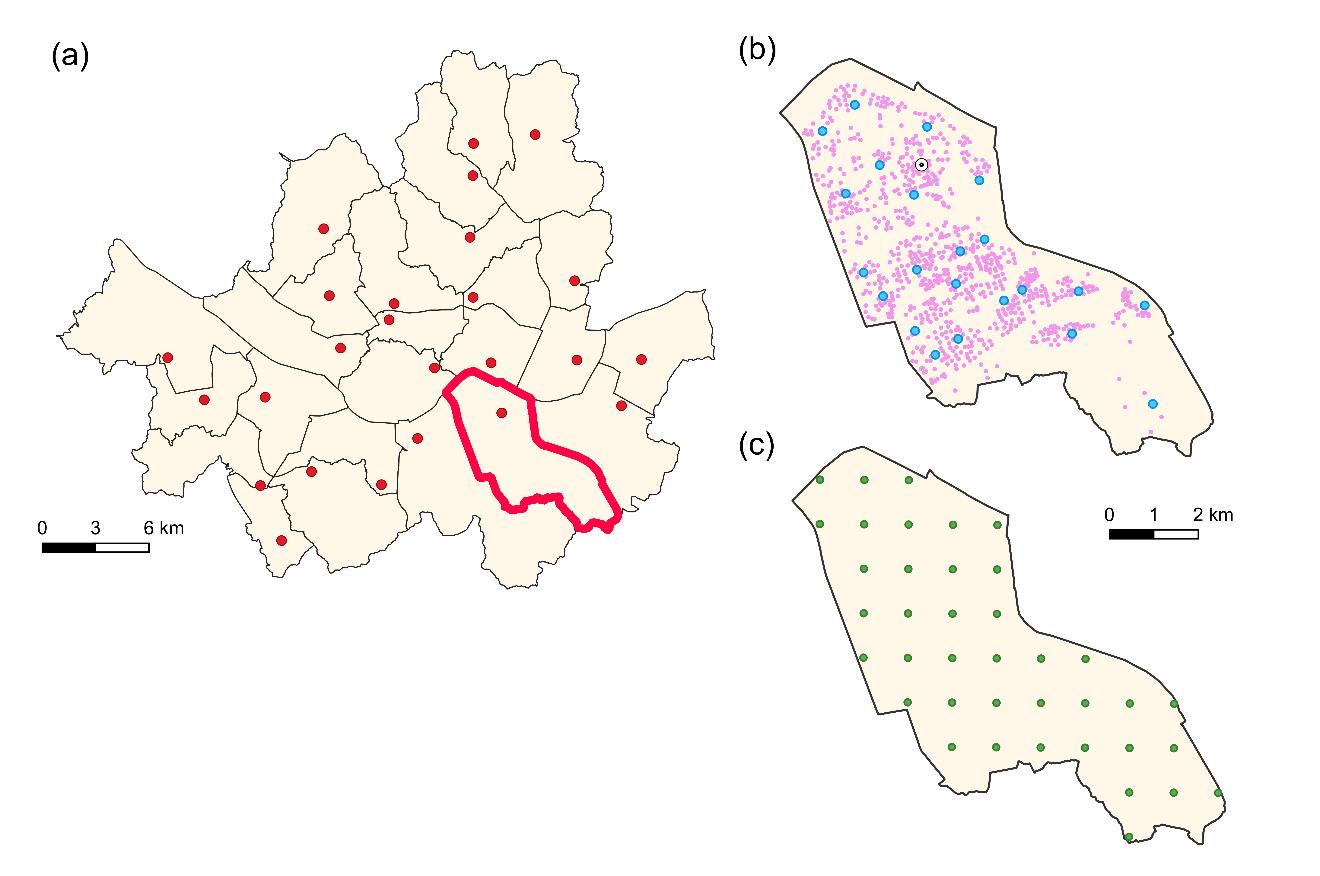
**Tables and Figures**

Figure01. Maps of 25 urban-background regular monitoring sites (red) and the government offices of 25 districts (black) in Seoul (a), neighborhood community centers (light blue) and census tracts centroids (light purple) in a red-lined district in (a) (b), and 1-km grid coordinates (green) (c)

Table01. Spatial characteristics of eight environmental scenarios and their variance parameters used for simulating true PM10 annual average concentrations

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Environmental Scenario  (ES) | Variance parameter | | | Spatial characteristics\* | | |
| Nugget | Partial sill | Range (m) | Mean  structure | Spatial variability | Non-spatial variability |
| ES1 | 1.00 | 30.94 | 5,885 | None | Dominant | Little |
| ES2 | 6.86 | 28.98 | 9,609 | None | High | Low |
| ES3 | 11.51 | 34.71 | 20,355 | None | Low | High |
| ES4 | 22.00 | 13.77 | 27,000 | None | Little | Dominant |
| ES5 | 1.00 | 16.98 | 2,524 | Moderate | High | Low |
| ES6 | 6.86 | 12.17 | 4,820 | Moderate | Low | High |
| ES7 | 1.00 | 10.00 | 1,100 | Dominant | High | Low |
| ES8 | 6.86 | 3.60 | 1,004 | Dominant | Low | High |

\*Characterized by contribution of three variability components (mean structure, and spatial and non-spatial variability) to total variability

Table02. Description of Data Availability Conditions (DAC1-DAC3) for individual-level PM10 concentrations and home addresses of 549,270 mothers residing in Seoul, Korea.

|  |  |  |
| --- | --- | --- |
| Data Availability Condition | PM10 concentration | Home address |
| DAC1 | All mothers | All mothers |
| DAC2 | Regulatory monitoring sites | All mothers |
| DAC3 | Regulatory monitoring sites | Limited to some coarse district level |

Table04. Mean and standard deviation of true and predicted PM10 annual average concentrations at 46,007 home addresses of mothers under three data availability scenarios (DAC1-DAC3), eight environmental scenarios (ES1-ES8), and six exposure prediction methods (AA, UK, UKD, UKNA, UKCA, and UKGA) at the 1st simulation

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | DAC1 | DAC2 | DAC3 | | | | |
|  | TE | UK | AA | UKD | UKNA | UKCA | UKGA |
| ES1 | 51.16(5.56)\* | 52.27(1.56) | 51.59(4.24) | 51.63(4.53) | 51.61(4.24) | 51.6(4.27) | 51.49(3.97) |
| ES2 | 49.8(5.97) | 49.97(1.59) | 49.71(4.13) | 49.91(4.8) | 49.68(4.49) | 49.81(4.38) | 49.49(4.04) |
| ES3 | 58.55(6.45) | 60.52(3.43) | 59.56(4.56) | 59.66(4.46) | 59.45(4.63) | 59.60(4.09) | 59.21(4.02) |
| ES4 | 46.9(5.11) | 47.52(1.99) | 47.61(2.56) | 47.57(2.29) | 47.42(1.21) | 47.52(1.50) | 47.25(1.04) |
| ES5 | 50.91(5.26) | 51.51(4.30) | 51.46(4.16) | 51.51(4.33) | 51.26(2.65) | 51.41(4.07) | 51.19(2.13) |
| ES6 | 53.05(4.66) | 53.21(3.45) | 53.32(4.08) | 53.37(3.54) | 53.19(2.07) | 53.22(2.86) | 52.99(2.00) |
| ES7 | 51.61(5.81) | 51.97(5.09) | 51.89(4.43) | 51.42(5.33) | 51.22(2.48) | 51.24(4.16) | 51.28(2.98) |
| ES8 | 51.19(5.17) | 51.35(4.32) | 51.97(4.29) | 51.34(4.64) | 51.21(1.74) | 51.19(3.14) | 51.13(2.41) |

\* Mean (Standard deviation)

TE: true-exposure; NM: nearest-monitor; IDWA: inverse-distance-weighted-average; LUR: land use regression; UK: universal kriging; AA: area-average UKD: UK prediction at a SGG district centroid; UKNA: district average based on UK predictions at neighbourhood centroids; UKCA: district average based on UK predictions at census tract centroids; UKGA: district average based on UK predictions at 1km-by-1km grid centroids

Table05. Properties of effect estimates of true and predicted PM10 annual average concentrations on low birth weight over 1,000 simulations by three data availability conditions (DAC1-DAC3), some representative exposure prediction methods and their variations (AA, UK, UKD, UKNA, UKCA, UKGA), and four environmental scenarios (ES2, ES3, ES5, ES8)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Bias1) | RMSE2) | ASE3) | CP4) |  | Bias1) | RMSE2) | ASE3) | CP4) |
|  | ES2 |  |  |  |  | ES3 |  |  |  |  |
| DAC1 | TE | -0.04 | 0.015 | 0.008 | 0.95 | TE | 0.03 | 0.015 | 0.008 | 0.95 |
| DAC2 | UK | -0.03 | 0.024 | 0.015 | 0.94 | UK | -0.05 | 0.027 | 0.016 | 0.94 |
| DAC3 | AA  UKD  UKNA  UKCA  UKGA | -0.15  -0.08  -0.01  0.00  0.09 | 0.015  0.024  0.033  0.028  0.038 | 0.008  0.014  0.019  0.017  0.024 | 0.94  0.94  0.94  0.94  0.94 | AA  UKD  UKNA  UKCA  UKGA | -0.14  -0.09  0.07  0.05  0.04 | 0.016  0.027  0.041  0.033  0.048 | 0.008  0.016  0.027  0.020  0.036 | 0.95  0.94  0.94  0.95  0.94 |
|  | ES5 |  |  |  |  | ES8 |  |  |  |  |
| DAC1 | TE | 0.03 | 0.015 | 0.008 | 0.95 | TE | 0.00 | 0.014 | 0.007 | 0.95 |
| DAC2 | UK | 0.01 | 0.021 | 0.011 | 0.95 | UK | -0.01 | 0.017 | 0.009 | 0.96 |
| DAC3 | AA  UKD  UKNA  UKCA  UKGA | -0.19  -0.16  0.04  0.01  0.05 | 0.018  0.020  0.031  0.023  0.037 | 0.009  0.011  0.017  0.013  0.020 | 0.95  0.94  0.94  0.94  0.95 | AA  UKD  UKNA  UKCA  UKGA | -0.25  -0.24  0.06  0.02  -0.11 | 0.020  0.019  0.041  0.022  0.033 | 0.010  0.010  0.021  0.012  0.017 | 0.95  0.95  0.94  0.96  0.95 |

Total 1000 random simulations; 1) sample average of the bias (estimated beta – true beta) multiplied by 100 2) root mean square of beta residuals 3) average of standard error of estimated betas over simulations, 4) coverage probability with 95% confidence interval

**Supplementary Tables and Figures**

FigureS1. Maps of the true simulated annual-average PM10 concentrations by eight environmental scenarios (ESs) in Seoul, Korea

|  |  |  |  |
| --- | --- | --- | --- |
| **ES1** | **ES2** | **ES3** | **ES4** |
|  |  |  |  |
| **ES5** | **ES6** | **ES7** | **ES8** |
|  |  |  |  |

FigureS2. Box-plots of true (TE: true-exposure) and predicted (NM: nearest monitor, IDWA: inverse distance weight averaged, LUR: Land-use regression, AA: area averaged, UKD: UK prediction at district centroid, UKNA: district average of UK prediction at neighborhood centroids, UKCA: district average of UK prediction at census centroid centroids, UKGA: district average of UK prediction at 1-km grid coordinates) annual-average PM10 concentrations at 46,007 home addresses of mothers across eight exposure prediction methods by eight environmental scenarios in the 1st simulation.

|  |  |
| --- | --- |
| Environmental Scenario 1 | Environmental Scenario 2 |
|  |  |
| Environmental Scenario 3 | Environmental Scenario 4 |
|  |  |
| Environmental Scenario 5 | Environmental Scenario 6 |
|  |  |
| Environmental Scenario 7 | Environmental Scenario 8 |
|  |  |

FigureS3. Scatter plots of true (x-axis) and predicted (y-axis) annual-average PM10 concentrations at 46,007 home addresses of mothers in the 1st simulation

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ES1 | ES2 | ES3 | ES4 | ES5 | ES6 | ES7 | ES8 |
| NM |  |  |  |  |  |  |  |  |
| IDWA |  |  |  |  |  |  |  |  |
| LUR |  |  |  |  |  |  |  |  |
| UK |  |  |  |  |  |  |  |  |
| AA |  |  |  |  |  |  |  |  |
| UKD |  |  |  |  |  |  |  |  |
| UKNA |  |  |  |  |  |  |  |  |
| UKCA |  |  |  |  |  |  |  |  |
| UKGA |  |  |  |  |  |  |  |  |

FigureS4. Bar plots of bias (multiplied by 100) across nine prediction methods by eight environmental scenarios

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

FigureS5. Bar plots of root mean square error (blue bar), average of standard error (orange bar), and coverage probabilities (grey point) for eight environmental scenarios.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

TableS6-1. Properties of effect estimates of true and predicted PM10 annual average concentrations on low birth weight over 1,000 simulations by three data availability conditions (DAC1-DAC3), five exposure prediction methods (AA, NM, IDWA, LUR, UK), and four environmental scenarios without mean structures (ES1-ES4)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | PM10 coefficient () | | | |  | PM10 coefficient () | | | |
| Bias1) | RMSE2) | ASE3) | CP4) |  | Bias1) | RMSE2) | ASE3) | CP4) |
|  | ES1 |  |  |  |  | ES2 |  |  |  |  |
| DAC1 | TE | 0.01 | 0.015 | 0.008 | 0.96 | TE | -0.04 | 0.015 | 0.008 | 0.95 |
| DAC2 | NM  IDWA  LUR  UK | -0.11  0.05  -0.20  -0.06 | 0.016  0.025  0.043  0.022 | 0.008  0.013  0.027  0.012 | 0.96  0.96  0.94  0.96 | NM  IDWA  LUR  UK | -0.13  -0.03  -0.33  -0.03 | 0.015  0.025  0.045  0.024 | 0.008  0.013  0.025  0.015 | 0.93  0.95  0.96  0.94 |
| DAC3 | AA  UKD  UKNA  UKCA  UKGA | -0.12  -0.08  0.05  -0.03  0.08 | 0.016  0.022  0.027  0.025  0.032 | 0.008  0.012  0.015  0.014  0.018 | 0.95  0.95  0.96  0.96  0.96 | AA  UKD  UKNA  UKCA  UKGA | -0.15  -0.08  -0.01  0.00  0.09 | 0.015  0.024  0.033  0.028  0.038 | 0.008  0.014  0.019  0.017  0.024 | 0.94  0.94  0.94  0.94  0.94 |
|  | ES3 |  |  |  |  | ES4 |  |  |  |  |
| DAC1 | TE | 0.03 | 0.015 | 0.008 | 0.95 | TE | -0.02 | 0.014 | 0.007 | 0.95 |
| DAC2 | NM  IDWA  LUR  UK | -0.16  -0.05  -0.09  -0.05 | 0.015  0.025  0.047  0.027 | 0.008  0.013  0.026  0.016 | 0.94  0.95  0.96  0.94 | NM  IDWA  LUR  UK | -0.30  -0.25  -0.37  -0.30 | 0.015  0.026  0.045  0.035 | 0.008  0.014  0.027  0.021 | 0.92  0.96  0.93  0.94 |
| DAC3 | AA  UKD  UKNA  UKCA  UKGA | -0.14  -0.09  0.07  0.05  0.04 | 0.016  0.027  0.041  0.033  0.048 | 0.008  0.016  0.027  0.020  0.036 | 0.95  0.94  0.94  0.95  0.94 | AA  UKD  UKNA  UKCA  UKGA | -0.29  -0.19  -0.23  -0.30  -0.17 | 0.015  0.036  0.065  0.045  0.079 | 0.008  0.019  0.038  0.030  0.045 | 0.94  0.96  0.94  0.92  0.95 |

Total 1000 random simulations; 1) sample average of the bias (estimated beta – true beta) multiplied by 100 2) root mean square of beta residuals 3) average of standard error of estimated betas over simulations, 4) coverage probability with 95% confidence interval

TableS6-2 Properties of effect estimates of true and predicted PM10 annual average concentrations on low birth weight over 1,000 simulations by three data availability conditions (DAC1-DAC3), five exposure prediction methods (AA, NM, IDWA, LUR, UK), and four environmental scenarios with mean structures (ES5-ES8)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | PM10 coefficient () | | | |  | PM10 coefficient () | | | |
| Bias1) | RMSE2) | ASE3) | CP4) |  | Bias1) | RMSE2) | ASE3) | CP4) |
|  | ES5 |  |  |  |  | ES6 |  |  |  |  |
| DAC1 | TE | 0.03 | 0.015 | 0.008 | 0.95 | TE | 0.01 | 0.015 | 0.007 | 0.95 |
| DAC2 | NM  IDWA  LUR  UK | -0.16  0.00  -0.03  0.01 | 0.015  0.031  0.024  0.021 | 0.008  0.016  0.013  0.011 | 0.96  0.95  0.95  0.95 | NM  IDWA  LUR  UK | -0.20  -0.11  -0.04  -0.02 | 0.015  0.030  0.025  0.021 | 0.008  0.015  0.013  0.011 | 0.95  0.95  0.95  0.96 |
| DAC3 | AA  UKD  UKNA  UKCA  UKGA | -0.19  -0.16  0.04  0.01  0.05 | 0.018  0.020  0.031  0.023  0.037 | 0.009  0.011  0.017  0.013  0.020 | 0.95  0.94  0.94  0.94  0.95 | AA  UKD  UKNA  UKCA  UKGA | -0.24  -0.19  0.01  0.00  0.03 | 0.018  0.021  0.032  0.024  0.038 | 0.009  0.011  0.018  0.012  0.021 | 0.96  0.95  0.95  0.95  0.95 |
|  | ES7 |  |  |  |  | ES8 |  |  |  |  |
| DAC1 | TE | -0.01 | 0.014 | 0.007 | 0.95 | TE | 0.00 | 0.014 | 0.007 | 0.95 |
| DAC2 | NM  IDWA  LUR  UK | -0.28  -0.23  0.01  0.00 | 0.015  0.036  0.018  0.017 | 0.008  0.019  0.009  0.009 | 0.94  0.94  0.94  0.93 | NM  IDWA  LUR  UK | -0.29  -0.21  -0.01  -0.01 | 0.015  0.037  0.018  0.017 | 0.008  0.019  0.009  0.009 | 0.94  0.94  0.96  0.96 |
| DAC3 | AA  UKD  UKNA  UKCA  UKGA | -0.31  -0.24  0.08  0.02  0.01 | 0.020  0.019  0.039  0.022  0.032 | 0.010  0.010  0.021  0.012  0.017 | 0.94  0.94  0.94  0.94  0.95 | AA  UKD  UKNA  UKCA  UKGA | -0.25  -0.24  0.06  0.02  -0.11 | 0.020  0.019  0.041  0.022  0.033 | 0.010  0.010  0.021  0.012  0.017 | 0.95  0.95  0.94  0.96  0.95 |

Total 1000 random simulations; 1) sample average of the bias (estimated beta – true beta) multiplied by 100 2) root mean square of beta residuals 3) average of standard error of estimated betas over simulations, 4) coverage probability with 95% confidence interval

FigureS6 True-positive rate of significant effect of PM10 for the eight Environmental Scenarios (ES1-ES8)

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

TableS7.Summaries of linear regression fitting for 8 environmental scenarios with 8 exposure modelling strategies

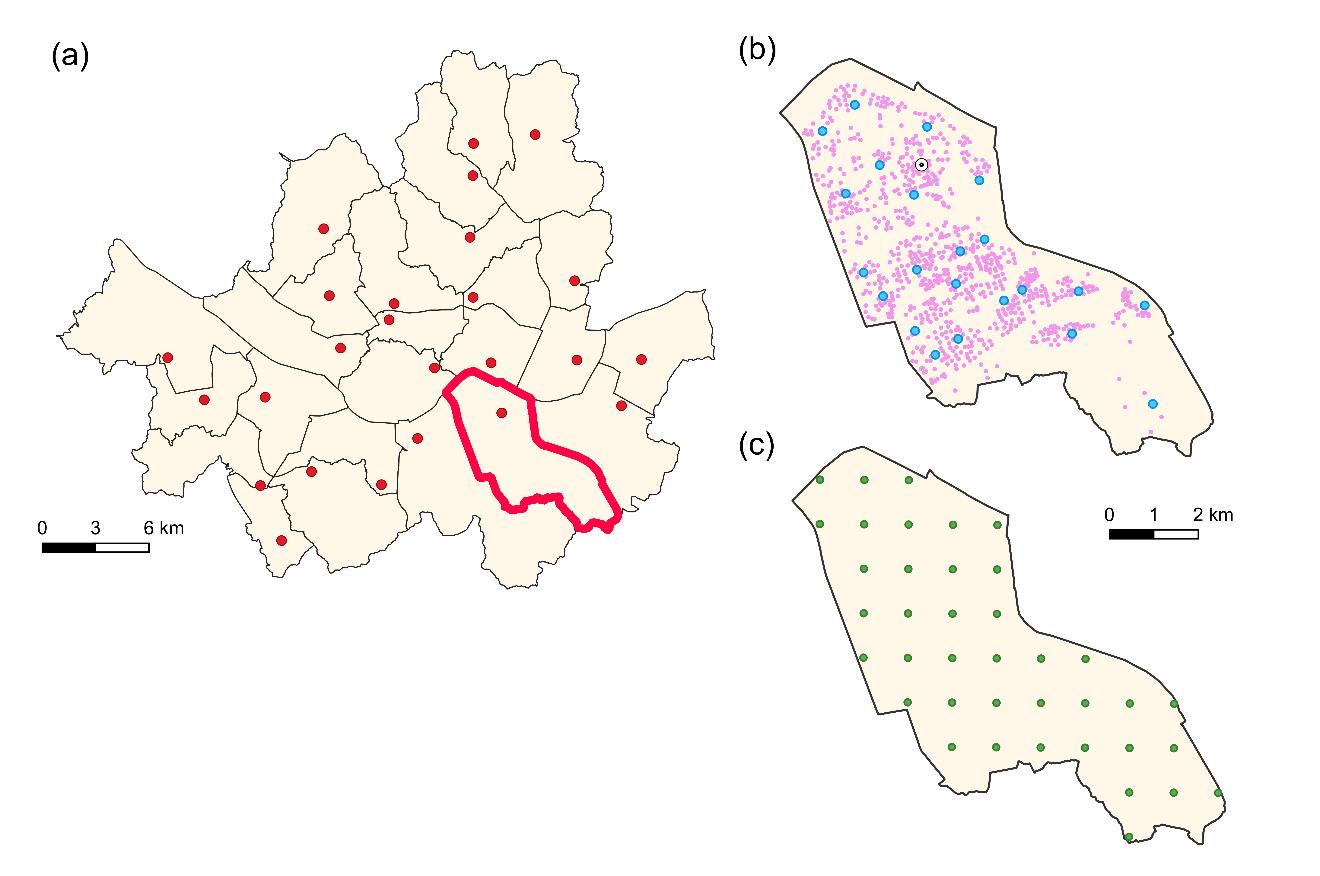
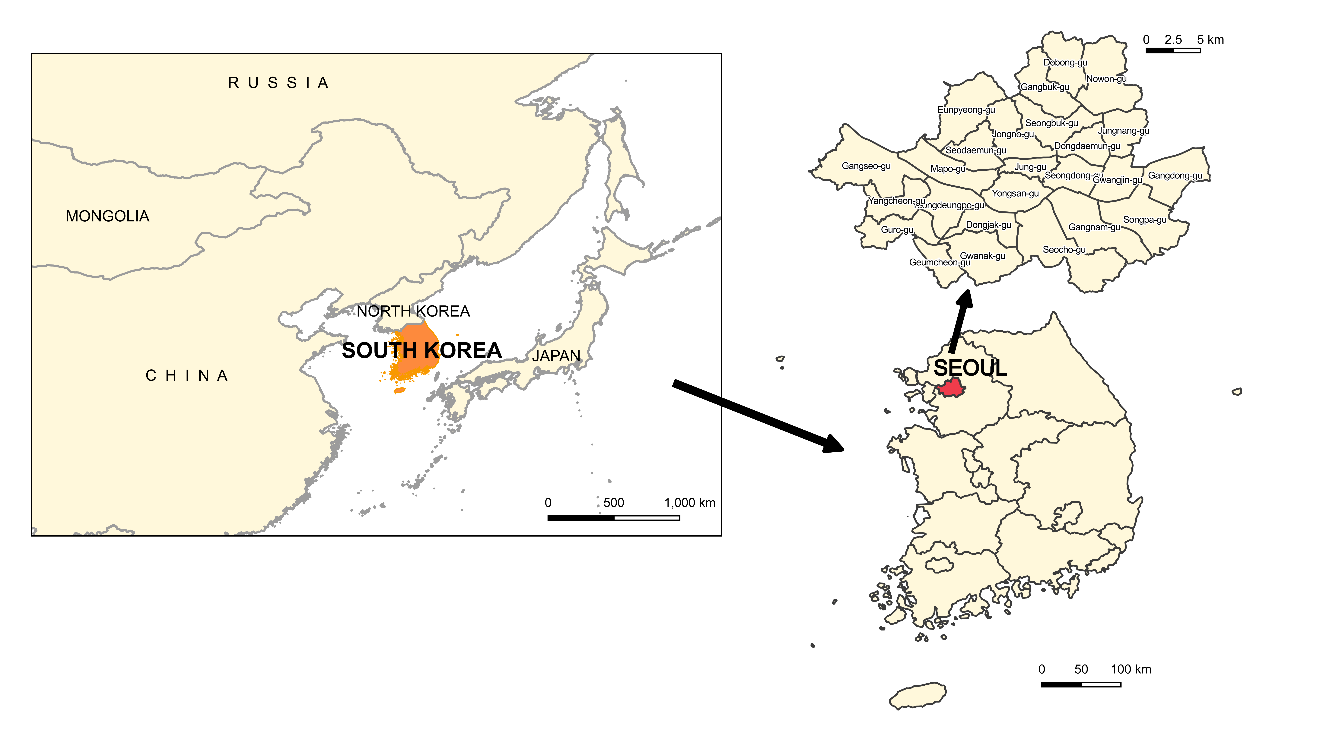
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PLAIN | ES1 | | | ES2 | | | ES3 | | | ES4 | | |
|  | R | Inter | Slope | R | Inter | Slope | R | Inter | Slope | R | Inter | Slope |
| DAC2 |  |  |  |  |  |  |  |  |  |  |  |  |
| NM | 0.623 | 11.692 | 0.776 | 0.488 | 19.517 | 0.609 | 0.517 | 20.833 | 0.657 | 0.036 | 40.796 | 0.141 |
| IDWA | 0.618 | 21.01 | 0.589 | 0.502 | 27.461 | 0.452 | 0.606 | 24.773 | 0.589 | 0.072 | 40.82 | 0.142 |
| LUR | 0.019 | 50.282 | 0.039 | 0.025 | 47.871 | 0.042 | 0.016 | 64.503 | -0.068 | 0.031 | 50.734 | -0.069 |
| UK | 0.697 | 16.783 | 0.681 | 0.467 | 22.562 | 0.549 | 0.456 | 32.29 | 0.467 | 0.001 | 48.26 | -0.015 |
| DAC3 |  |  |  |  |  |  |  |  |  |  |  |  |
| AA | 0.503 | 13.507 | 0.732 | 0.419 | 20.913 | 0.579 | 0.468 | 19.183 | 0.684 | 0.032 | 40.099 | 0.161 |
| UKD | 0.524 | 23.326 | 0.552 | 0.49 | 25.546 | 0.485 | 0.502 | 30.255 | 0.501 | 0.001 | 46.73 | 0.019 |
| UKNA | 0.637 | 20.471 | 0.609 | 0.452 | 24.458 | 0.506 | 0.585 | 27.297 | 0.549 | 0.006 | 46.57 | 0.018 |
| UKCA | 0.639 | 20.15 | 0.615 | 0.473 | 24.701 | 0.504 | 0.465 | 34.3 | 0.432 | 0.003 | 48.226 | -0.015 |
| UKGA | 0.603 | 23.141 | 0.554 | 0.444 | 27.052 | 0.451 | 0.55 | 32.109 | 0.463 | 0.017 | 46.009 | 0.026 |
| NON-PLAIN | ES5 | | | ES6 | | | ES7 | | | ES8 | | |
|  | R | Inter | Slope | R | Inter | Slope | R | Inter | Slope | R | Inter | Slope |
| DAC2 |  |  |  |  |  |  |  |  |  |  |  |  |
| NM | 0.296 | 22.392 | 0.552 | 0.176 | 30.837 | 0.422 | 0.130 | 34.920 | 0.302 | 0.031 | 42.520 | 0.156 |
| IDWA | 0.207 | 39.608 | 0.204 | 0.131 | 41.867 | 0.204 | 0.136 | 43.303 | 0.120 | 0.005 | 48.076 | 0.028 |
| LUR | 0.359 | 26.58 | 0.490 | 0.266 | 32.955 | 0.382 | 0.648 | 15.567 | 0.705 | 0.603 | 18.113 | 0.649 |
| UK | 0.371 | 25.985 | 0.501 | 0.381 | 28.44 | 0.47 | 0.668 | 12.658 | 0.751 | 0.587 | 16.053 | 0.689 |
| DAC3 |  |  |  |  |  |  |  |  |  |  |  |  |
| AA | 0.088 | 37.862 | 0.236 | 0.035 | 43.304 | 0.183 | 0.070 | 41.073 | 0.160 | 0.000 | 49.132 | 0.010 |
| UKD | 0.096 | 38.949 | 0.246 | 0.079 | 40.238 | 0.247 | 0.071 | 41.414 | 0.203 | 0.011 | 47.427 | 0.089 |
| UKNA | 0.219 | 39.225 | 0.236 | 0.189 | 42.954 | 0.193 | 0.287 | 39.422 | 0.229 | 0.169 | 44.122 | 0.138 |
| UKCA | 0.361 | 27.731 | 0.465 | 0.258 | 36.678 | 0.312 | 0.428 | 27.085 | 0.468 | 0.395 | 31.623 | 0.382 |
| UKGA | 0.153 | 43.120 | 0.158 | 0.164 | 43.785 | 0.173 | 0.268 | 37.59 | 0.265 | 0.143 | 42.129 | 0.176 |

TableS8. Mean and standard deviation of true and predicted PM10 annual average concentrations at 46,007 home addresses of mothers under three data availability scenarios (DAC1-DAC3), eight environmental scenarios (ES1-ES8), and eight exposure prediction methods (AA, NM, IDWA, LUR, UK, UKD, UKNA, UKCA, and UKGA) at the 1st simulation

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | DAC1 | DAC2 | | | | DAC3 | | | | |
|  | TE | NM | IDWA | LUR | UK | AA | UKD | UKNA | UKCA | UKGA |
| ES1 | 51.16(5.56)\* | 50.94(5.73) | 51.42(5.47) | 51.14(4.16) | 52.27(1.56) | 51.59(4.24) | 51.63(4.53) | 51.61(4.24) | 51.6(4.27) | 51.49(3.97) |
| ES2 | 49.8(5.97) | 49.75(5.34) | 49.83(5.2) | 49.94(3.8) | 49.97(1.59) | 49.71(4.13) | 49.91(4.8) | 49.68(4.49) | 49.81(4.38) | 49.49(4.04) |
| ES3 | 58.55(6.45) | 59.23(6.45) | 59.27(5.89) | 59.27(4.88) | 60.52(3.43) | 59.56(4.56) | 59.66(4.46) | 59.45(4.63) | 59.6(4.09) | 59.21(4.02) |
| ES4 | 46.9(5.11) | 47.68(4.59) | 47.43(3.81) | 47.48(2.71) | 47.52(1.99) | 47.61(2.56) | 47.57(2.29) | 47.42(1.21) | 47.52(1.5) | 47.25(1.04) |
| ES5 | 50.91(5.26) | 49.87(4.18) | 50.47(5.33) | 49.98(2.35) | 51.51(4.30) | 51.46(4.16) | 51.51(4.33) | 51.26(2.65) | 51.41(4.07) | 51.19(2.13) |
| ES6 | 53.05(4.66) | 53.02(4.57) | 53.22(4.68) | 52.67(2.62) | 53.21(3.45) | 53.32(4.08) | 53.37(3.54) | 53.19(2.07) | 53.22(2.86) | 52.99(2) |
| ES7 | 51.61(5.81) | 49.34(3.51) | 50.51(4.86) | 49.49(1.89) | 51.97(5.09) | 51.89(4.43) | 51.42(5.33) | 51.22(2.48) | 51.24(4.16) | 51.28(2.98) |
| ES8 | 51.19(5.17) | 49.66(4.00) | 50.53(4.58) | 49.53(2.03) | 51.35(4.32) | 51.97(4.29) | 51.34(4.64) | 51.21(1.74) | 51.19(3.14) | 51.13(2.41) |

\* Mean (Standard deviation)

TE: true-exposure; NM: nearest-monitor; IDWA: inverse-distance-weighted-average; LUR: land use regression; UK: universal kriging; AA: area-average UKGA: UK prediction at a SGG district centroid; UKNA: district average based on UK predictions at neighbourhood centroids; UKCA: district average based on UK predictions at census tract centroids



FigureS9. Description of the map of Seoul in South Korea (top panel). Maps of 25 urban-background regular monitoring sites (red) and the government offices of 25 districts (black) in Seoul (a), neighborhood community centers (light blue) and census tracts centroids (light purple) in a red-lined district in (a) (b), and 1-km grid coordinates (green) (c) are illustrated (bottom panel)