11-30-2018

Concentrations of criteria pollutants in the contiguous U.S., 1979 – 2015: Role of model parsimony in integrated empirical geographic regression

Sun-Young Kim
University of Washington - Seattle Campus, puha0@uw.edu

Matthew Bechle
University of Washington, matthew.bechle@gmail.com

Steve Hankey
Virginia Tech, hankey@vt.edu

Elizabeth (Lianne) A. Sheppard
University of Washington, sheppard@uw.edu

Adam A. Szpiro
University of Washington, aszpiro@uw.edu

See next page for additional authors

Authors
Sun-Young Kim, Matthew Bechle, Steve Hankey, Elizabeth (Lianne) A. Sheppard, Adam A. Szpiro, and Julian D. Marshall

Concentrations of criteria pollutants in the contiguous U.S., 1979 – 2015: Role of model parsimony in integrated empirical geographic regression

Sun-Young Kim (1,2), Matthew Bechle (3), Steve Hankey (4), Lianne Sheppard (2,5), Adam A. Szpiro (5), Julian D. Marshall (3)

1) Department of Cancer Control and Population Health, Graduate School of Cancer Science and Policy, National Cancer Center, Goyang-si, Gyeonggi-do, Korea

2) Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, WA, USA

3) Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, USA

4) School of Public and International Affairs, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

5) Department of Biostatistics, University of Washington, Seattle, WA, USA

Corresponding author: Sun-Young Kim

Department of Cancer Control and Population Health

Graduate School of Cancer Science and Policy

National Cancer Center, Goyang-si, Gyeonggi-do, Korea

Tel: +82-31-920-2745

Fax: +82-2-920-2759

Acknowledgments: This publication was developed as part of the Center for Clean Air Climate Solution (CACES), which was supported under Assistance Agreement No. R835873 awarded by the U.S. Environmental Protection Agency (EPA). It has not been formally reviewed by EPA. The views expressed in this document are solely those of authors and do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this publication. Additional support was provided by the National Research Foundation of Korea grants funded by the Ministry of Education and the Ministry of Science and ICT in South Korea (2013R1A6A3A04059017 and 2018R1A2B6004608) and the National Cancer Center in South Korea (NCC-1810220-01).

Competing financial interests

The authors declare they have no actual or potential competing financial interests.

ABSTRACT

BACKGROUND: National- or regional-scale prediction models that estimate individual-level air pollution concentrations commonly include hundreds of geographic variables. However, these many variables may not be necessary and parsimonious approach including small number of variables may achieve sufficient prediction ability. This parsimonious approach can also be applied to most criteria pollutants. This approach will be powerful when generating publicly available datasets of model predictions that support research in environmental health and other fields.

OBJECTIVES: We aim to (1) build annual-average integrated empirical geographic (IEG) regression models for the contiguous U.S. for six criteria pollutants, for all years with regulatory monitoring data during 1979 – 2015; (2) explore the impact of model parsimony on model performance by comparing the model performance depending on the numbers or variables offered into a model; and (3) provide publicly available model predictions.

METHODS: We compute annual-average concentrations from regulatory monitoring data for PM_{10}, $PM_{2.5}$, NO_2, SO_2, CO, and ozone at all monitoring sites for 1979-2015. We also compute ~900 geographic characteristics at each location including measures of traffic, land use, and satellite-based estimates of air pollution and landcover. We then develop IEG models, employing universal kriging and summary factors estimated by partial least squares (PLS) of independent variables. For all pollutants and years, we compare three approaches for choosing variables to include in the model: (1) no variables (kriging only), (2) a limited number of variables chosen by forward selection, and (3) all variables. We evaluate model performance using 10-fold cross-validation (CV) using conventional randomly-selected and spatially-clustered test data.

RESULTS: Models using 3 to 30 variables generally have the best performance across all

pollutants and years (median R^2 conventional [clustered] CV: 0.66 [0.47]) compared to models with no (0.37 [0]) or all variables (0.64 [0.27]). Using the best models mostly including 3-30 variables, we predicted annual-average concentrations of six criteria pollutants for all Census Blocks in the contiguous U.S.

DISCUSSION: Our findings suggest that national prediction models can be built on only a small number (30 or fewer) of important variables and provide robust concentration estimates. Model estimates are freely available online.

Keywords: Air pollution, Cohort, Exposure Assessment, Geographic Covariates, Population Exposure

INTRODUCTION

Regulatory monitors and project-based monitoring campaigns typically provide air pollution measurements that are limited in space and time. Air pollution prediction approaches are a cost-effective approach to estimate fine-scale exposures to air pollution. Recent population-level studies of air pollution have relied on empirical models to estimate long-term concentrations of outdoor air pollution based largely on observation-driven geostatistical approaches (Eeftens et al. 2012; Keller et al. 2015; Kim et al. 2017) or hybrid approaches that incorporate satellite-based observations of air quality and theory-based mechanistic models with geostatistical approaches (Ma et al. 2014; van Donkelaar et al. 2016). These model predictions are used to assess population-level characteristics of air pollution, such as health effects (Beleen et al. 2015; Kaufman et al. 2016), the burden of disease (Fann et al. 2017; Xie et al. 2016), and exposure disparities (Clark et al. 2017; Hajat et al. 2016).

Empirical models for air pollution are generally developed using a large suite of input data often including hundreds of geographic covariates (e.g., traffic, population, land use) with the goal of predicting concentrations at locations lacking monitoring data (Hoek et al. 2008). More recently, studies have included regional estimates of air pollution from mechanistic models (Lindstrom et al. 2015) and satellite-based air pollution measurements such as tropospheric nitrogen dioxide (NO_2) column abundance and Aerosol Optical Depth (AOD) (Chu et al. 2016; Hoek 2017). These regional air pollution estimates are particularly useful for national- or global-scale prediction where air pollution measurements are sparse over large areas (Bechle et al. 2015; Di et al. 2016; Larkin et al. 2017; Novotny et al. 2011; van Donkelaar et al. 2015; Young et al. 2016).). To incorporate and prioritize information from the many hundreds of predictor variables, studies typically employ regression-based statistical techniques such as variable selection,

shrinkage, and dimensional-reduction (Eeftens et al 2012; Mercer et al. 2011; Sampson et al. 2013) or other artificial intelligence approaches (Beckerman et al. 2013; Di et al. 2016)

Computational demands of building, testing, and applying models that include hundreds of variables are large, especially for national-scale models. Yet, there is little guidance in the literature regarding the added benefit of using many variables versus more parsimonious models. Furthermore, although some national-scale models exist for some years and pollutants for $PM_{2.5}$, PM_{10}, NO_2, or ozone (Sampson et al. 2013; Young et al. 2017), empirical models do not currently exist for most criteria pollutants in a unified framework across all years with regulatory monitoring data in the U.S. This article aims to address both of those gaps. Specifically, we develop, test, and compare full versus parsimonious national models for annual average concentrations of six criteria pollutants and for all years with available monitoring data during 1979 – 2015. We test the hypothesis that model performance is better with more variables than with a smaller number of intentionally selected variables. Then, we select the best performing models to generate concentration estimates for all residential Census Block centroids in the contiguous U.S. for all modeling years with the goal of making our model predictions available freely online. We refer our model to "Integrated Empirical Geographic" (IEG) regression models to indicate key characteristics of the model: many "integrated" input datasets from land use, satellite-derived measures, and emission estimates; "empirical" modeling approach to characterize data-driven relationships rather than based on theory of physics and chemistry; and, "geographic" features in data and modeling technique.

METHODS

Regulatory Monitoring Data for Criteria Pollutants

We downloaded daily or hourly measurements of six criteria pollutants including PM_{10},

$PM_{2.5}$, NO_2, SO_2, CO, and O_3 at all Air Quality System (AQS) monitoring sites for all available years from 1979 through 2015 via the U.S. Environmental Protection Agency (EPA) AQS data repository (Figure S1). NO_2, SO_2, and O_3 are available for the entire period (1979–2015); CO, PM_{10}, and $PM_{2.5}$ are available starting in 1990, 1988, and 1999, respectively. For PM_{10} and $PM_{2.5}$, we use data from the Federal Reference Method (FRM) and Integrated Monitoring of Protected Visual Environments (IMProVE) networks.

We compute annual averages for all pollutants (except ozone) at sites that meet our inclusion criteria, as follows. We compute 24-hour averages for monitors with 18 or more valid hourly measurements in that day, and then compute annual averages at sites with a minimum number of operating days (244 days for daily/hourly measurements, 61 days for 1-in-3 day measurements, and 41 days for 1-in-6 day measurements) during a year and no more than 45 consecutive days without a measurement. For ozone, we use the daily maximum of the 8-hour moving average for days with 18 or more operating hours during the day and compute an ozone season average from May through September. All IEG regression modeling is done after applying square root transformation to all pollutant concentrations to meet normality assumption.

Geographic Variables

We consider >900 geographic variables, as independent variables for our IEG models, in eight categories: traffic, population, land-use, elevation, vegetation, industrial emissions, and satellite air pollution estimates (Table S1). To reflect changes of land use characteristics over time, we obtained the two types of land use variables from ground-based datasets generated in 1970s and 1980s, and satellite and aerial imagery in 2006. The variables are computed as summaries within buffer areas between 50 meter and 15 kilometers (0.05, 0.1, 0.15, 0.3, 0.4, 0.5, 0.75, 1, 1.5, 3, 5, 10, and 15 km) depending on the variable and/or as distance to the closest

feature. We exclude variables with little spatial variability (e.g., same values at the 10th and 90th percentiles) or few unique values, reducing the number of variables to an average of ~350 for a given pollutant and year.

Traffic variables are distance to the nearest road and sum of road lengths within eleven circular buffers (0.05, 0.1, 0.15, 0.3, 0.4, 0.5, 0.75, 1, 1.5, 3, and 5 km) based on TeleAtlas data (http://www.teleatlas.com/OurProducts/MapData/Dynamap/index.htm). Population variables are the number of people in twenty circular buffers (0.5, 0.75, 1, 1.5, 3, 5, 10, and 15 km), based on year-2000 U.S. Census population (http://arcdata.esri.com/data/tiger2000/tiger_download.cfm). Land use variables in 1970s and 1980s are percent of areas for various land use characteristics such as residential, industrial, commercial, and agriculture land use identified by the U.S. Geological Survey (http://water.usgs.gov/GIS/dsdl/ds240/index.html) in circular buffer areas. Land cover variables based on satellite imagery in 2006 are percent of areas for land use characteristics such as developed high and low density obtained from the Multi-Resolution Land Cover Characteristics (MRLC) Consortium (http://www.mrlc.gov/index.php) in circular buffer areas. Elevation is the absolute elevation measurement at a given location and relative elevation compared to elevation in a circular buffer areas, calculated from national elevation dataset (http://nationalmap.gov/elevation.html). Vegetation variables are normalized difference vegetation index computed from satellite imagery (http://glcf.umd.edu/data/ndvi/) in circular buffer areas. Emission variables are the total amount emission estimates in circular buffer areas based on national emission inventory data (http://www.epa.gov/ttn/chief/net/2002inventory.html).

We obtain and compute annual satellite-based estimates of air pollution concentrations for $PM_{2.5}$, NO_2, SO_2, CO, and formaldehyde (HCHO) (Table S2); details on the specific steps are in the Supplemental Materials. The net result is satellite-derived annual, ground-level estimates

for $PM_{2.5}$ (1998-2014; 0.1° × 0.1° grid) (van Donkelaar et al., 2016), NO2 (2005-2015; 0.1° × 0.1° grid); SO2 (2005-2016; 0.25° × 0.25° grid); and CO (2001-2016; 0.25° × 0.25° grid), and a multiyear average for ground-level concentrations for HCHO (2005-2016; 0.25° × 0.25° grid).

Modeling Approach

Our approach builds on a universal kriging framework, described elsewhere (e.g., Bergen et al. 2012; Sampson et al. 2013; Young et al. 2016; Young et al. 2016), that partitions annual average concentrations into two components (Banjinee et al. 2004): variance and mean. The variance component is modeled using three parameters: range (the distance at which spatial correlation exists), partial sill (spatial variability), and nugget (non-spatial variability). The mean component includes two or three dimension-reduced summary predictors estimated using partial least squares (PLS) from the covariates offered. The mean component is equivalent to the linear regression model often referred to as LUR with PLS data-reduction.

To investigate the role of model parsimony, we purposefully select via forward selection a specific number of variables to offer the PLS; we investigate how model performance varies depending on the number of variables offered. The number of variables offered ranges from zero (i.e., no variables – a kriging only approach) to the full covariate database, with several intermediate values (e.g., 5-variable models, 20-variable models). For example, the 20-variable model would involve forward selection to select the best 20 variables, followed by PLS data-reduction to identify two or three PLS components comprised of those 20 variables, and regression modeling based on those two or three PLS components.

We hypothesize that adding more variables would improve the model and the performance diminishes as more variables are added. In that case, there may be a “point of diminishing returns”: a number of variables for which adding more variables yields little

additional benefit.

Model Evaluation

We evaluate models using two types of 10-fold cross-validation (CV): conventional and spatially clustered (Young et al. 2016). For conventional CV, we randomly divide all monitoring sites into 10 groups. Then, we select one group as the hold-out sites, develop models using the remaining data, and predict air pollution concentrations at hold-out sites. This process is repeated separately for each of the 10 groups to create a pseudo-independent test data set. Spatially clustered CV is similar except that the 10 groups are spatial clusters identified using k-means clustering (Figure S2) (Young et al. 2016). Conventional CV reflects model performance at a random location, whereas clustered CV reflects model performance far from a monitor. For dense monitor networks, such as $PM_{2.5}$ in the U.S., conventional CV may be more representative of model performance where most people live.

CV statistics include root-mean-square error (RMSE) and MSE-based R-squared (R^2). The MSE-based R^2 is calculated as 1 minus the ratio of MSE to data variance, whereas a conventional R^2 is calculated as the squared correlation coefficient. Conventional R^2 assesses agreement between predictions and observations about the regression line; MSE-based R^2 instead assesses agreement about the 1:1 line (Keller et al. 2015; Kim et al. 2016). To allow for comparison across different pollutants, we also compute standardized RMSE (i.e., RMSE divided by the mean concentration across all sites). For each pollutant and year, the "best" and "worst" models are identified based on R^2 and standardized RMSE from both conventional and clustered CV.

Sensitivity Analyses

To investigate the contribution of each category of variables (see above and Table S1),

we develop models that separately exclude each category of variables and compare the model performance between the models.

In addition, we conduct the following three sensitivity analyses to examine the impact of our methodological choices on model performance. To shed light on whether our results of best and worst models are sensitive to a type of CV approach, we compute CV statistics in one CV using the best models chosen by the other CV. That is, conventional CV is recomputed for the best models chosen by clustered CV, whereas clustered CV is recomputed for the best models chosen by conventional CV. To assess the impact of forward selection during model-building, we replace forward selection with random selection and compare the model performance of the same numbers of variables selected at random to that of our original forward selection. To test our model evaluation focusing on estimation of regression and covariance parameters in universal kriging, we expand our CV to include forward selection and estimation of PLS predictors as well as parameter estimation as a more conservative evaluation. We apply these three sensitivity analyses to limited examples: two pollutants for NO_2 and $PM_{2.5}$ and one year in 2000.

Lastly, we test the robustness of ozone models to other ozone averaging approaches: annual and summer season (May-September) summaries of ozone using 24-hour means, 8-hour means, and 1-hour maximum.

Prediction

Using the best models for each pollutant and year, we predict annual average concentrations for the ~7 million residential Census Block centroids in the contiguous U.S. with nonzero population. Then, we compute population-weighted averages at various geographic scales (Census Block Groups, Census Tracts, Counties, States, and contiguous U.S.) based on

2010 Census boundaries.

RESULTS

Summary of Monitoring Data

Means and standard deviations of annual average concentrations at AQS monitoring sites decrease over time for all pollutants (Table 1, Figure S3). During 1980 to 2010, average concentrations decrease almost 6-fold for SO_2 (from 12.7 to 2.2 ppb) but only 14% for ozone (from 52.0 to 45.8 ppb). For ozone, the 10^{th} percentile concentration decrease less than 2% over 30 years (from e7.8 to 37.2 ppb). From 2000 to 2010, reductions for $PM_{2.5}$ and PM_{10} are 39% and 28%, respectively.

IEG Model Performance by Number of Variables

Different from our hypothesis, adding more variables did not consistently improve model performance, especially for clustered CV (Figure S4). For all pollutants and for both CV approaches, models using 3-30 variables generally show higher R^2 and lower standardized RMSE than models using no or all variables (Table 2 and Figure 1).

The no-variable (i.e., kriging-only) models were generally the lowest-performing (Figure S5). Selecting best-performing models generally was consistent among metrics (MSE-R^2, standardized RMSE), and model performance of the best model is typically robust to selection using clustered versus conventional CV (Figure S6).

IEG Model Performance by CV

CV results consistently indicate better model performance using conventional CV than using clustered CV (Table 2, Figure 1), indicating poor performance when there are no monitors in the vicinity. Considering all pollutants and years, median R^2 and standardized RMSE, based on conventional CV, for the best models are 0.66 (interquartile range [IQR]: 0.57–0.83) and 0.23

(0.13–0.31), respectively. Analogous values for clustered CV are median R^2 of 0.47 (0.31–0.65) and standardized RMSE of 0.27 (0.19–0.38). Median (IQR) R^2 and standardized RMSE for the worst models are 0.57 (0.44–0.67) and 0.32 (0.18–0.39) for conventional CV, and 0 (0–0.01) and 0.47 (0.31–0.62) for clustered CV.

IEG Model Performance by Pollutant

Parsimonious models for $PM_{2.5}$ and NO_2 show generally good performance using conventional CV: median R^2 (standardized RMSE) of the best models are 0.86 (0.13) for $PM_{2.5}$ and 0.87 (0.21) for NO_2 (Table 2, Figure 1). Analogous results using clustered CV are 0.65 (0.20) for $PM_{2.5}$, 0.80 (0.24) for NO_2. For NO_2, differences in model performance between "best" and "no variable" models are larger for clustered CV than for conventional CV (median R^2 for the best/no-variable model: 0.87/0.61 (conventional CV) versus 0.80/0.00 (clustered CV)). That finding indicates the substantial benefit of having variables in the model when there are no monitors nearby and indicates that the kriging-only NO_2 model offers nearly zero information far from monitors. In contrast, for SO_2, ozone, and PM_{10}, differences between "best" and "no variable" models were modest for conventional CV (median R^2 for best/no-variable models: 0.59/0.57 [SO_2], 0.75/0.72 [ozone], 0.59/0.49 [PM_{10}]). Analogous differences were larger for clustered CV (0.27/0.00 [SO_2], 0.47/0.35 [ozone], 0.32/0.00 [PM_{10}]). Overall, for both CV approaches, NO_2 and $PM_{2.5}$ yield better models than other pollutants (Figure 2). CO shows moderate model performance regardless of the number of variables (median R^2 for best models: 0.47 [conventional CV] and 0.44 [clustered CV]). Over time, model performance tended to improve for ozone and PM_{10}, decline for SO_2 and CO, and remain relatively unchanged for $PM_{2.5}$ and NO_2.

Selected Variables

Investigation of covariates by category chosen via forward selection (Figure 3) reveals that satellite air pollution estimates are almost always selected in the top 5 variables across all pollutants and years; urban or rural land use is consistently selected in the top 10 variables. Impervious surface and traffic are often selected for NO_2, whereas emissions and/or elevation are common for SO_2 and ozone, respectively. Models with the top 30 variables include almost all categories except population and emissions, depending on the year and pollutant.

Sensitivity Analyses

In our sensitivity analysis of re-computing CV statistics based on conventional or clustered CV for the best and worst models based on the other CV approach, the selection of best models reduced numbers of variables and worst models mostly without any variables were consistent.

The three sensitivity analyses conducted on NO_2 and $PM_{2.5}$ for 2000 indicate the following. First, model performance is highly degraded when satellite variables are not included (Figure S7), especially for clustered CV. The inclusion of land use variables becomes important as models include larger numbers of variables. Second, when variable selection is random rather than via forward selection, model performance is noticeably reduced (Figure S8). However, even with random selection of variables, the improvement in performance for models with all variables relative to models with ~30 variables is small when using conventional CV. Thus, we find that even using a subset of randomly selected variables can yield models that are comparable to the "all variable" models. Third, when we shift the CV procedure to make it broader to include the entire model-building endeavor rather than only variable selection and universal kriging, results generally show consistent patterns as with the core results, for clustered CV (Figure S9). With conventional CV, shifting the CV procedure reduced the difference in

model performance between parsimonious models and “all variables” models.

Sensitivity analyses involving alternative metrics of ozone concentration revealed that our original approach using 8-hour moving averages of the summer season shows the best performance (Table S3).

Model application

Predicted annual-average concentrations throughout the U.S. (Figures 4 and S10), generated using “best” models, reflect the decreasing concentrations. The extent of temporal change and the spatial patterns vary by pollutant. Population-weighted averages of annual average concentrations at Census Block centroids show similar means and narrow variability compared to those at monitoring sites (Table 3). Predicted concentrations for all Block Groups, Tracts, Counties, and States in the contiguous U.S. are publicly and freely available online at URL-to-be-added-upon-acceptance.

DISCUSSION

We built and tested IEG models for six pollutants for all years with national monitoring data during 1979 – 2015 in the contiguous U.S.; results for “best-performing” models are publicly available online. We explore systematically the role of parsimony: how model performance changes when models are built using more or fewer variables.

A common assumption would be that parsimonious models will under-perform relative to “all variable” models: a more-variable model is always better. Thus, we hypothesized that adding more variables would always improve model-performance, though potentially with diminishing returns at some point. Results here indicate that our hypothesis was not hold. Our findings indicate that parsimonious models outperform or perform as well as “all variable” models. We find good model performance using a relatively small numbers of variables (between

3 and 30 variables); satellite-derived estimates of air pollution and of land cover were common variables in the IEG models generated here.

An important motivator for this research question is the considerable effort and computational intensity of tabulating hundreds of geographic variables; that effort and computational intensity is a barrier to widespread development and usage of national IEG models. This limitation impacts the feasibility of subsequent analyses in epidemiology, exposure assessment, environmental justice, and other fields. As the spatial domain for air pollution exposure models and health analyses is expanded to national or global scales (Bechle et al. 2015; Di et al. 2016; Larkin et al. 2017; Novotny et al. 2011; van Donkelaar et al. 2015; Young et al. 2016), data and processing requirements will grow as additional input data are needed to improve prediction ability. Our approach reveals which predictive variables are most important for generating parsimonious models that outperform all-variable models; as future studies investigate similar questions, the results could help guide future IEG modeling.

Model performance varied by pollutant, with better performance for $PM_{2.5}$, NO_2, and ozone than for CO, SO_2, and PM_{10}. All models benefited from introducing at least a small number of geographic covariates. Model performance is similar for kriging-only as for IEG "best models" in the following cases: using conventional CV, for SO_2 and to some extent for ozone; using clustered CV, for none of the pollutants (though among pollutants the "best" IEG / kriging-only gap is smallest for ozone). In general, kriging-only models deliver much of the total value of the IEG model with conventional CV but deliver zero or near-zero value with clustered CV.

Differences in model performance may reflect differences in chemistry and physics of the pollutant, spatial patterns of emissions, quality of input data, correlation with land uses, availability of relevant satellite data, a design of monitoring network (number of monitors and

their placement). For example, the gap between kriging-only and "best" IEG is larger for NO_2 than for $PM_{2.5}$, reflecting the time scale for formation of secondary $PM_{2.5}$; spatial patterns are more homogeneous for $PM_{2.5}$ than for NO_2; and, number of monitors is ~3× larger for $PM_{2.5}$ than NO_2. The extant monitoring network is designed for regulatory purposes: mainly, to test for compliance with National Ambient Air Quality Standards (NAAQS). As the use of IEG models grows, EPA or others could consider utility to IEG models (e.g., monitoring in locations with a variety of land uses) as an additional goal.

The slightly worse performance of the models using all variables as compared to models using some variables was not anticipated. This performance could potentially be explained by the fact that we treated the selected variables as fixed and did not include the selection process in our model evaluation methodology, however, this finding held when we included forward selection and estimation of PLS predictors in our evaluation in addition to estimation of regression and covariance parameters. Similar model performance between the models using limited and full sets of covariates in conventional CV may represent possible over estimation of prediction ability in our original evaluation approach. However, consistently better performance with reduced numbers of variables, shown in clustered CV, indicates good prediction ability of a parsimonious approach in areas without monitors. This finding also highlights the importance of clustered CV when evaluating observation-driven models. In addition, the degraded model performance with the same numbers of randomly-selected variables supports our conclusion that a small subset of important variables can be sufficiently predictive for annual average air pollution concentrations compared to the full set of variables.

Our results highlight the importance of satellite data for IEG (Hoek 2017); satellite data are selected as one or more of the top five variables consistently across all pollutants and years.

The most commonly selected satellite estimates were satellite $PM_{2.5}$ for $PM_{2.5}$ models and HCHO for ozone models. Considering IEG model performance when a category of variables is excluded, the performance decline is greater excluding satellite data than excluding other data, especially with clustered CV.

A common concern for IEG models such as those generated here is that the range of values for independent variables might differ at monitoring locations relative to prediction locations where people live (Szpiro et al 2011; Szpiro and Paciorek 2013). Monitors may be located in areas where few people live and may not be able to represent people's exposures. If that were the case, then for locations where values for the independent variables are outside the range of values at monitoring stations, one could censor those values or offer a data-quality flag. However, when we compared the distribution of geographic variables between monitoring sites and Census Block centroids, for 95% and 98% of ~900 variables, the standard deviation for monitoring sites are smaller than 2.5 and 5 times standard deviation for Census Block centroids. This finding suggests that the range of values at Block centroids are similar to the range across monitoring sites, suggesting reasonably good spatial alignment between monitoring and prediction locations in our work. Because our models use estimated PLS predictors instead of direct measures of variables, extreme values of a few variables are less likely to impact model predictions (Kim et al. 2016).

Our study has several limitations to motivate future research. We consider only spatial aspects of IEG models and use many temporally-fixed geographic variables (exceptions include satellite-derived estimates of air pollution concentrations, and land use variables are for the 1970s and 2006). Future work could build national, publicly available models with finer temporal resolution than here (i.e., better than annual-averages) and could test model parsimony

with respect to temporal models or spatiotemporal models. Future studies could add variables that represent geographic characteristics changing over time. Future studies could investigate parsimony for IEG models using other modeling approaches, such as neural network or random forest. Satellite air pollution estimates employed here for NO_2, SO_2, and HCHO are tropospheric column abundance, rather than ground-level estimates. Previous studies have shown that IEG models improvements from satellite-derived estimates of air pollution are similarly for column-total as for ground-level estimates (cite); future work could test that finding for SO_2 and HCHO. The present research employed emission estimates, which are an input into chemical transport models (CTMs), and prior research has included CTM as an input to IEG model-building. Future research could test the role of model parsimony in IEGs that incorporate CTM output. Future research on IEG models could potentially include national datasets on traffic volumes, vehicle fleet composition, Google point-of-interest data, urban form from Landsat imagery, and recently-launch satellites. We hypothesize that such datasets would improve IEG model performance, though recognizing that because the IEG models already have many inputs (including satellite-based estimates of air pollution concentrations), new datasets may or may not improve model performance appreciable.

In summary, this study provides important findings on cost-effective approaches for national-scale air pollution prediction. Results indicate that national IEG model performance can be similar or better if built on only a small number of purposely-selected covariates from hundreds, relative to models build using all of those variables. Our model predictions for the contiguous US are freely available online, at URL-to-be-added-upon-acceptance.

REFERENCES

Banerjee S, Carlin BP, Gelfand AE, 2004. Basics of point-referenced data models. In: Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall/CRC Press, Boca Raton, FL, pp. 21-68.

Bechle M, Millet DB, Marshall JD. 2015. National Spatiotemporal Exposure Surface for NO2: Monthly Scaling of a Satellite-Derived Land-Use Regression, 2000–2010. Environ Sci Technol 49 (20):12297–12305.

Beckerman BS, Jerrett M, Serre M, Martin R, Lee SJ, van Donkelaar A, Roo Z, Su J, Bunett R. 2013. A Hybrid Approach to Estimating National Scale Spatiotemporal Variability of PM2.5 in the Contiguous United States. Environ Sci Technol 47(13):7233-7241.

Beelen R, Raaschou-Nielsen O, Stafoggia M, Andersen ZJ, Weinmayr G, Hoffmann B, et al. 2014. Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicenter ESCAPE project. Lancet 383(9919):785e795.

Bergen S, Sheppard L, Sampson PD, Kim SY, Richards M, Vedal S, et al. 2013. A national prediction model for components of PM2.5 and measurement error corrected health effect inference. Environ. Health Perspect 121(9):1017e1025.

Boersma KF, Eskes HJ, Dirksen RJ, van der A RJ, Veefkind JP, Stammes P, Huijnen V, Kleipool QL, Sneep M, Claas J, Leitao J, Richter A, Zhou Y, Brunner D. 2011. An improved retrieval of tropospheric NO2 columns from the Ozone Monitoring Instrument. Atmos Meas Tech 4:1905-1928.

Chu Y, Liu Y, Li, X, Liu Z, L H, Lu Y, Mao Z, Chen X, Li N, ren M, Liu F, Tian L, Zhu Z, Xiang H. 2016. A review on predicting ground PM2.5 concentration using satellite aerosol optical depth. Atmosphere 7(129):1-25.

Deeter MN, Edwards DP, Francis GL, Gille JC, Martinez-Alonso S, Worden HM., Sweeney C. 2017. A Climate-scale Satellite Record for Carbon Monoxide: The MOPITT Version 7 Product. Atmos Meas Tech 10:2533-2555.

De Smedt I, Müller JF, Stavrakou T, van der AR, Eskes H, Van Roozendael M. 2018. Twelve years of global observations of formaldehyde in the troposphere using GOME and SCIAMACHY sensors. Atmos Chem Phys 8(16):4947−4963.

Di Q, Kloog I, Koutrakis P, Lyapustin A, Wang Y, Schwartz J. 2016. Assessing PM2.5 exposures with high spatiotemporal resolution across the continental United States. Environ Sci Technol 50:4712-4721.

Di Q, Wang Y, Zanobetti A, Wang Y, Koutrakis P, Choirat C, Dominici F, Schwartz JD. 2017. Air Pollution and Mortality in the Medicare Population. N Engl J Med 376(26):2513-2522.

Eeftens M, Beelen R, de Hoogh K, Bellander T, Cesaroni G, Cirach M, et al. 2012. 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 46(20):11195-11205.

Fann N, Kim SY, Olives C, Sheppard L. Estimated changes in life expectancy and adult mortality resulting from declining PM2.5 exposures in the contiguous United States:1980–2010. Environmental Health Perspectives 2017:125;9.

Hajat A, Diez-Roux AV, Adar SD, Auchincloss AH, Lovasi GS, O'Neill MS, Sheppard L, Kaufman JD. 2013. Air pollution and individual and neighborhood socioeconomic status: evidence from the multi-ethnic study of atherosclerosis (MESA). Environ Health Perspect 121(11-12_:1325-1333.

Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fischer P, et al. 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos Environ 42(33): 7561e7578.

Hoek G. 2017. Methods for Assessing Long-Term Exposures to Outdoor Air Pollutants. Curr Environ Health Rep 4(4):450-462.

Hogrefe C, Pouliot G, Wong D, Torian A, Roselle S, Pleim J, Mathur R, 2015: Annual application and evaluation of the online coupled WRF-CMAQ system over North America under AQMEII phase 2. Atmos Environ 115:683-694.

Kaufman JD, Adar SD, Barr RG, Budoff M, Burke GL, Curl CL, et al. 2016. Association between air pollution and coronary artery calcification within six metropolitan areas in the USA (the Multi-Ethnic Study of Atherosclerosis and Air Pollution): a longitudinal cohort study. Lancet 13(10045):696e704.

Keller JP, Olives C, Kim SY, Sheppard L, Sampson PD, Szpiro AA, et al. 2015. A unified spatiotemporal modeling approach for prediction of multiple air pollutants in the multi-ethnic study of atherosclerosis and air pollution. Environ. Health Perspect 123(4):301e309.

Kim SY, Sheppard L, Bergen S, Szpiro AA, Sampson PD, Kaufman JD, et al. 2016. Prediction of fine particulate matter chemical components for the Multi-Ethnic Study of Atherosclerosis cohort: a comparison of two modeling approaches. J Exp Sci Environ Epidemiol. 26(5):520e528.

Kim SY, Olives C, Sheppard L, Sampson PD, Larson TV, Keller JP, Kaufman JD. 2017. Historical Prediction Modeling Approach for Estimating Long-Term Concentrations of PM2.5 in Cohort Studies before the 1999 Implementation of Widespread Monitoring. Environ Health Perspect 125(1):38-46.

Kelly C. 2007, OMI/Aura Formaldehyde (HCHO) Total Column 1-orbit L2 Swath 13x24 km V003, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), 10.5067/Aura/OMI/DATA2015.

Larkin A, Geddes JA, Martin RV, Xiao Q, Liu Y, Marshall JD, Brauer M, Hystad P. 2017. Global Land Use Regression Model for Nitrogen Dioxide Air Pollution Environ Sci Technol 51 (12):6957–6964.

Lindstrom J, Szpiro AA, Sampson PD, Oron AP, Richards M, Larson TV, et al., 2013. A flexible spatio-temporal model for air pollution with spatial and spatiotemporal covariates. Environ Ecol Stat 21:411e433.

Ma Z, Hu X, Huang L, Bi J, Liu Y. 2014. Estimating Ground-Level PM2.5 in China Using Satellite Remote Sensing. Environ Sci Technol 48 (13):7436–7444.

Mercer LD, Szpiro AA, Sheppard L, Lindström J, Adar SD, Allen RW, Avol EL, Oron AP, Larson T, Liu LJ, Kaufman JD. 2011. Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NOx) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). Atmos Environ 45(26):4412-4420.

Novotny EV, Bechle M, Millet DB, Marshall JD. 2011. National Satellite-Based Land-Use Regression: NO2 in the United States Environ. Sci. Technol 45 (10):4407–4414.

OMI Science Team. 2012. OMI/Aura Level 2 Sulphur Dioxide (SO2) Trace Gas Column Data 1-Orbit subset Swath along CloudSat track 1-Orbit Swath 13x24 km, Edited by GES DISC, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC) (https://disc.gsfc.nasa.gov/datacollection/OMSO2_CPR_003.html).

Sampson PD, Richards M, Szpiro AA, Bergen S, Sheppard L, Larson TV, et al. 2013. A regionalized national universal kriging model using partial least squares regression for estimating annual PM2.5 concentrations in epidemiology. Atmos Environ 75:383e392.

Szpiro AA, Paciorek CJ, Sheppard L. 2011. Does more accurate exposure prediction necessarily improve health effect estimates? Epidemiology 22(5);680-685.

Szpiro AA and Paciorek CJ. 2013. Measurement error in two-stage analyses, with application to air pollution epidemiology. Environmetrics 24;501-517.

van Donkelaar A, Martin RV, Brauer M, Hsu NC, Kahn RA, Levy RC, Lyapustin A, Sayer AM, Winker DM. 2016. Global Estimates of Fine Particulate Matter using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors. Environ Sci Technol 50(7):3762-3772.

Yin P, Brauer M, Cohen A, Burnett RT, Liu J, Liu Y, Liang R, Wang W, Qi J, Wang L, Zhou M. 2017. .Long-term Fine Particulate Matter Exposure and Nonaccidental and Cause-specific Mortality in a Large National Cohort of Chinese Men. Environ Health Perspect 7;125(11):117002.

Young MT, Sandler DP, DeRoo LA, Vedal S, Kaufman JD, London SJ. 2014. Ambient air pollution exposure and incident adult asthma in a nationwide cohort of U.S. women. Am J Respir Crit Care Med 15;190(8):914-921.

TABLES

Table 1. Summary statistics of annual average concentrations for six criteria air pollutants across regulatory monitoring sites in the contiguous U.S. for 1980, 1990, 2000, and 2010

Pollutant	Year	N	Percentile					Mean	SD
			10	25	50	75	90		
NO_2	1980		8.4	16.9	24.8	34.4	49.8	26.5	15.2
(ppb)	1990		6.0	11.4	17.8	24.9	31.8	18.9	10.5
	2000		5.4	9.7	15.5	20.3	26.2	15.6	8.2
	2010		2.8	5.2	9.1	13.1	17.3	9.6	5.6
SO_2	1980		3.3	6.5	10.5	15.6	23.9	12.7	10.4
(ppb)	1990		1.6	3.6	7.1	9.7	13.5	7.3	4.8
	2000		1.4	2.4	4.2	6.2	9.0	4.7	2.9
	2010		1.0	1.1	1.6	2.8	4.2	2.2	1.6
Ozone	1980		37.8	45.8	52.0	59.4	66.2	52.0	11.3
(ppb)	1990		39.3	44.9	49.3	54.1	59.4	49.3	7.8
	2000		39.6	44.7	50.1	54.8	58.4	49.4	7.4
	2010		37.2	41.6	46.6	51.0	53.7	45.8	6.8
CO	1990		0.48	0.66	0.95	1.26	1.67	1.02	0.48
(ppm)	2000		0.33	0.41	0.54	0.76	0.99	0.62	0.28
	2010		0.29	0.31	0.33	0.38	0.46	0.35	0.10
PM_{10}	1990		18.6	23.1	27.6	33.6	39.6	29.0	10.0
($\mu g/m^3$)	2000		12.9	18.1	22.7	27.2	35.3	23.8	10.3
	2010		8.8	13.6	18.0	22.4	27.9	18.6	8.3
$PM_{2.5}$	2000		6.8	10.1	12.8	15.5	17.1	12.5	4.1
($\mu g/m^3$)	2010		4.4	7.2	9.5	11.3	12.5	9.0	3.0

Table 2. Cross-validation (CV) statistics[a] for the IEG regression models developed here, by pollutant, year, and level of model parsimony (zero variables / between 3 and 30 variables / all variables)

		Conventional CV						Clustered CV					
		Standardized RMSE*			R^2			Standardized RMSE*			R^2		
Pollutant	Year	0	3-30	All	0	3-30	All	0	3-30	All	0	3-30	All
NO_2	2000	0.33	0.19	0.20	0.61	0.87	0.85	0.60	0.23	0.29	0.00	0.82	0.70
(ppb)	2010	0.39	0.23	0.25	0.56	0.84	0.81	0.64	0.33	0.33	0.00	0.68	0.68
SO_2	2000	0.39	0.38	0.39	0.60	0.63	0.61	0.62	0.47	0.51	0.00	0.44	0.32
(ppb)	2010	0.64	0.63	0.65	0.29	0.31	0.28	0.79	0.65	0.72	0.00	0.26	0.10
O_3	2000	0.07	0.07	0.07	0.76	0.78	0.78	0.11	0.10	0.11	0.45	0.55	0.51
(ppb)	2010	0.06	0.06	0.06	0.81	0.82	0.81	0.11	0.10	0.11	0.44	0.51	0.44
CO	2000	0.37	0.32	0.34	0.33	0.50	0.43	0.47	0.35	0.43	0.00	0.42	0.12
(ppm)	2010	0.25	0.23	0.25	0.17	0.28	0.20	0.28	0.24	0.28	0.00	0.23	0.00
PM_{10}	2000	0.31	0.27	0.28	0.50	0.60	0.59	0.45	0.37	0.39	0.00	0.27	0.20
($\mu g/m^3$)	2010	0.34	0.29	0.30	0.41	0.57	0.56	0.47	0.37	0.39	0.00	0.33	0.26
PM_{25}	2000	0.16	0.12	0.13	0.77	0.86	0.85	0.30	0.21	0.22	0.15	0.59	0.53
($\mu g/m^3$)	2010	0.17	0.13	0.13	0.73	0.85	0.84	0.31	0.19	0.20	0.14	0.70	0.64

[a] Standardized RMSE is the root mean square error (RMSE) divided by average concentration. Values are shown for three levels of model parsimony: for models with zero variables (i.e., kriging only), denoted with "0"; the median among all "parsimonious" models, i.e., those developed with between 3 and 30 variables, denoted "3-30"; and for full models with all variables, denoted "all".

Table 3. Summary statistics of population-weighted annual average concentrations for the contiguous US, by pollutant and decadal year, based on Census Block centroids, using "best" IEG model predictions

Pollutant	Year	N	Percentile					Mean	SD
			10	25	50	75	90		
NO_2 (ppb)	1980		7.4	12.1	19.9	27.9	36.8	21.3	11.7
	1990		6.1	8.4	12.9	19.0	26.9	15.2	9.2
	2000		5.6	7.8	11.8	16.7	23.2	13.3	7.5
	2010		3.3	4.7	7.2	10.8	15.8	8.5	5.1
SO_2 (ppb)	1980		3.4	5.8	8.9	12.5	16.6	9.6	5.3
	1990		2.0	3.0	4.6	7.0	9.2	5.3	3.0
	2000		1.8	2.2	3.1	4.4	6.1	3.6	1.8
	2010		0.9	1.2	1.5	2.0	2.5	1.6	0.7
Ozone (ppb)	1980		39.0	45.4	51.3	57.3	63.6	51.1	9.6
	1990		39.6	44.8	48.6	52.4	56.8	48.5	6.5
	2000		40.2	43.9	49.0	53.6	57.1	48.5	6.7
	2010		37.7	43.1	46.6	49.6	52.2	45.6	6.0
CO (ppm)	1990		0.33	0.43	0.61	0.86	1.19	0.69	0.35
	2000		0.29	0.35	0.43	0.55	0.74	0.48	0.20
	2010		0.23	0.28	0.31	0.35	0.39	0.31	0.07
PM_{10} ($\mu g/m^3$)	1990		19.8	22.7	25.9	30.2	36.8	27.5	7.9
	2000		15.7	18.8	22.0	25.4	30.8	22.9	6.8
	2010		12.8	15.2	18.3	21.5	24.1	18.4	4.6
$PM_{2.5}$ ($\mu g/m^3$)	2000		8.6	10.7	12.9	15.2	16.7	12.9	3.4
	2010		6.3	7.9	9.6	10.8	12.1	9.4	2.2

FIGURE CAPTIONS

Figure 1. Standardized root mean square errors and R^2s of the national prediction models including no, some, and all variables from conventional and clustered cross-validation for 1979-2015 over the contiguous U.S. by six criteria air pollutants; best models determined by each of the two types of cross-validation as one of the some-variables models

Figure 2. Scatter plots of standardized root mean square errors and R^2s from the best national prediction models across six criteria air pollutants in 2000 over the contiguous U.S. by conventional and clustered cross-validation

Figure 3. Categories of top 30, 10 and 5 geographic and satellite variables chosen by forward selection in the national prediction models of six criteria air pollutants for 1979-2015 over the contiguous U.S.

Figure 4. Maps of Census Block Group population-weighted mean predicted annual average concentrations for $PM_{2.5}$, NO_2, and ozone from the best national prediction models mostly including 3-30 variables for 2000 and 2010 in the contiguous U.S.

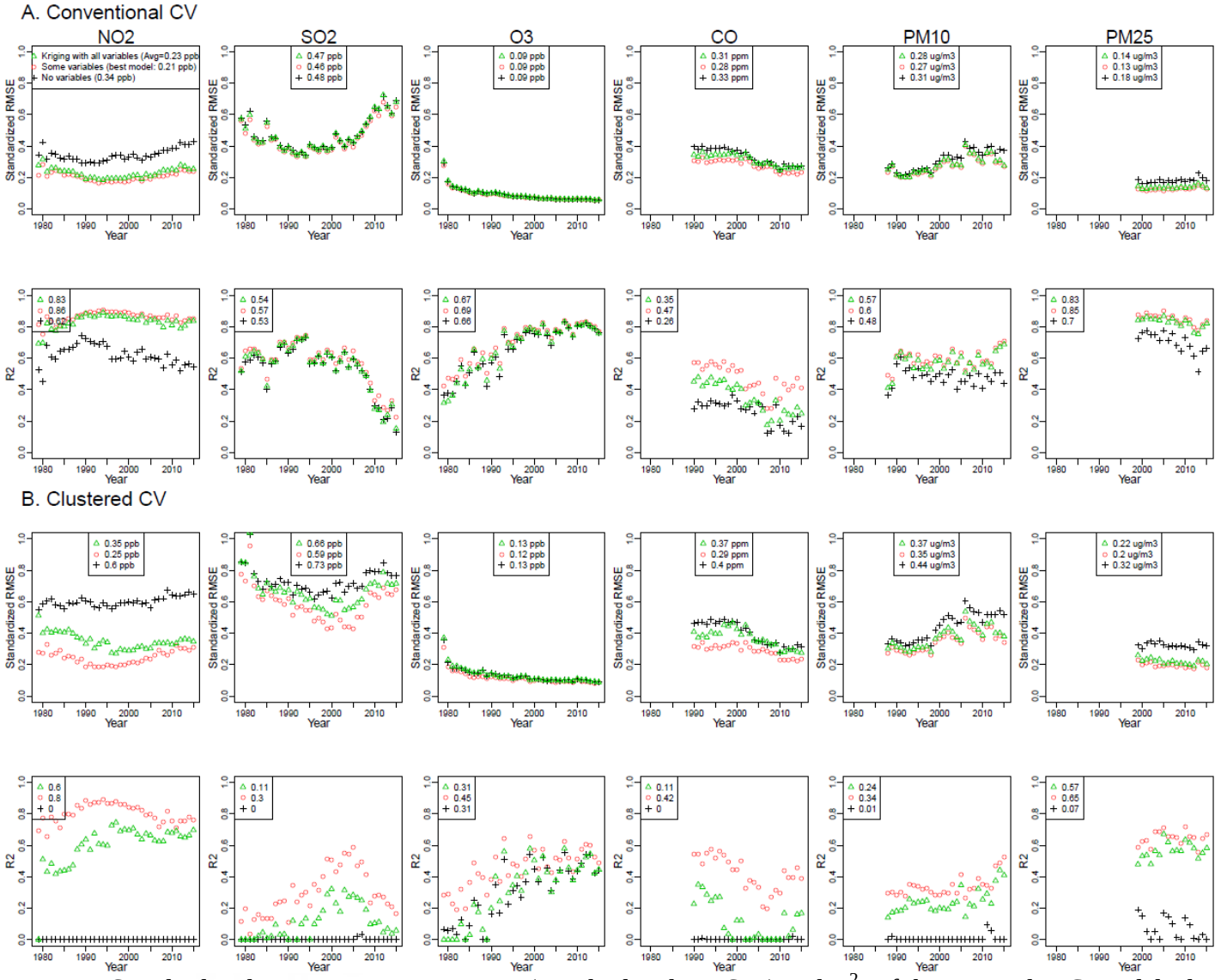

Figure 1. Standardized root mean square errors (standardized RMSEs) and R^2s of the national IEG models during 1979-2015 using conventional CV and clustered CV. Terminology here is the same as in Table 2.

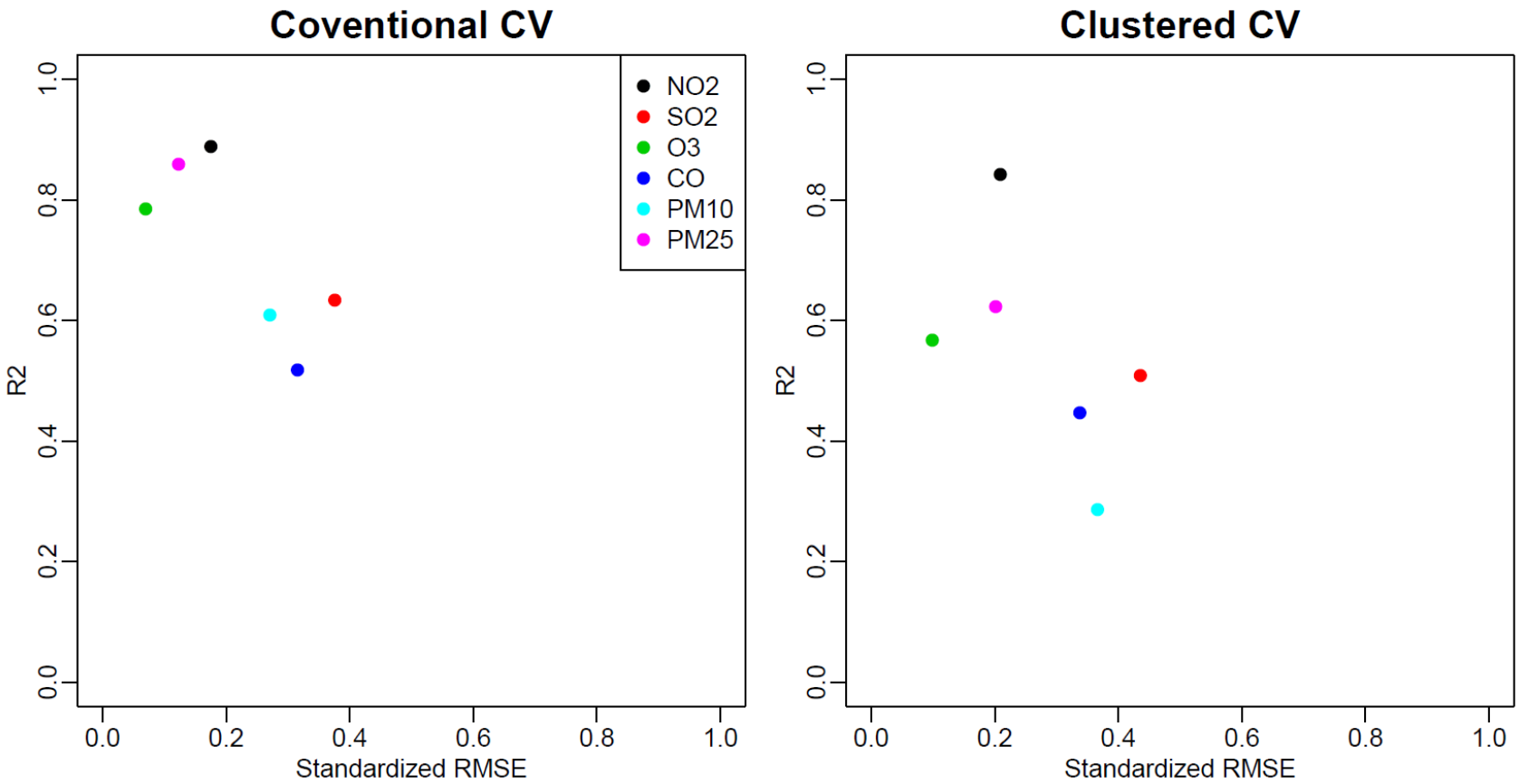

Figure 2. Standardized RMSEs and R^2s from "best" IEG models, for the contiguous US in 2000, for conventional and clustered CV, by pollutant.

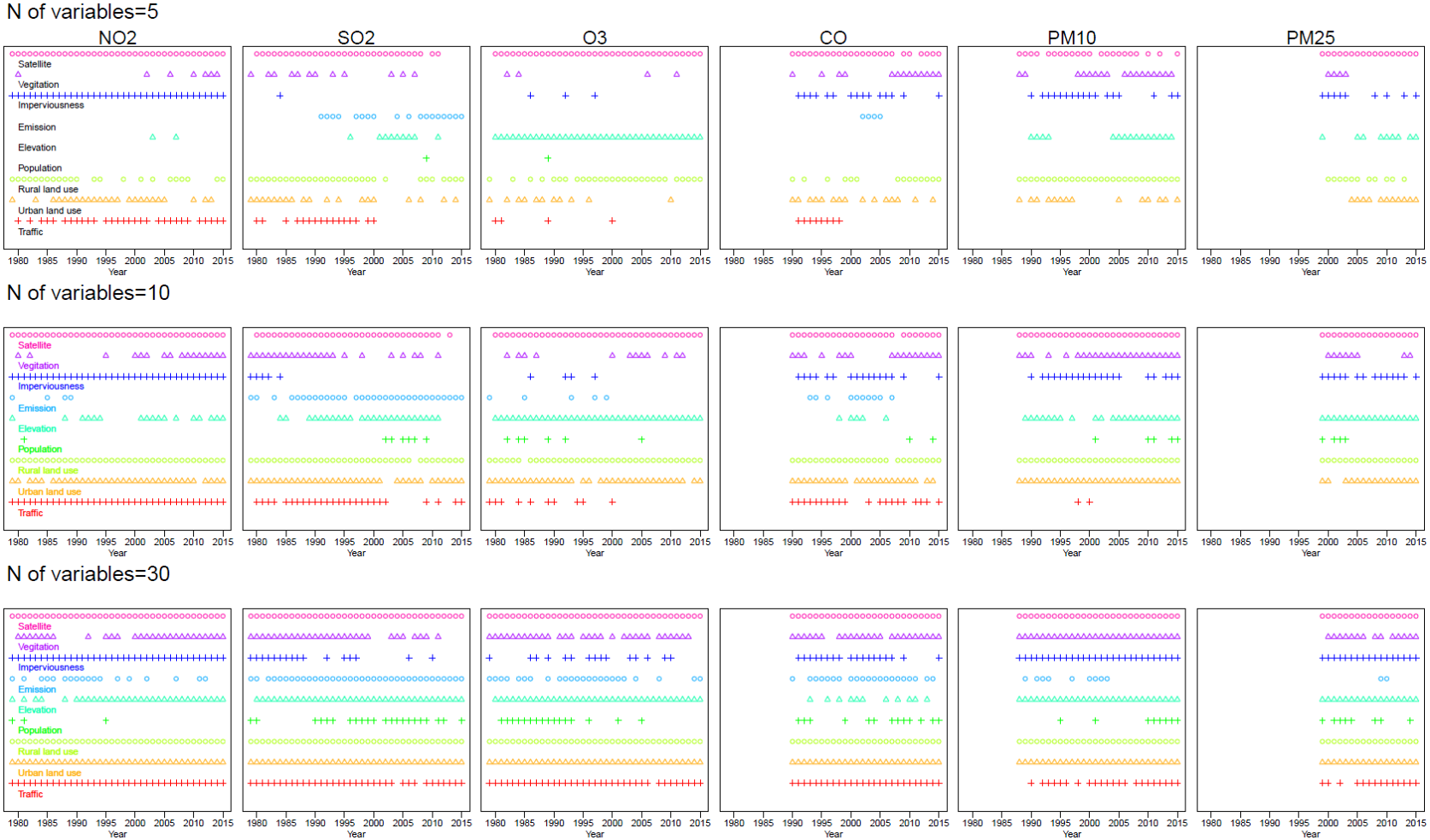

Figure 3. Categories of variables chosen by forward selection for national IEG models, by year, pollutant, and number of variables in the model.

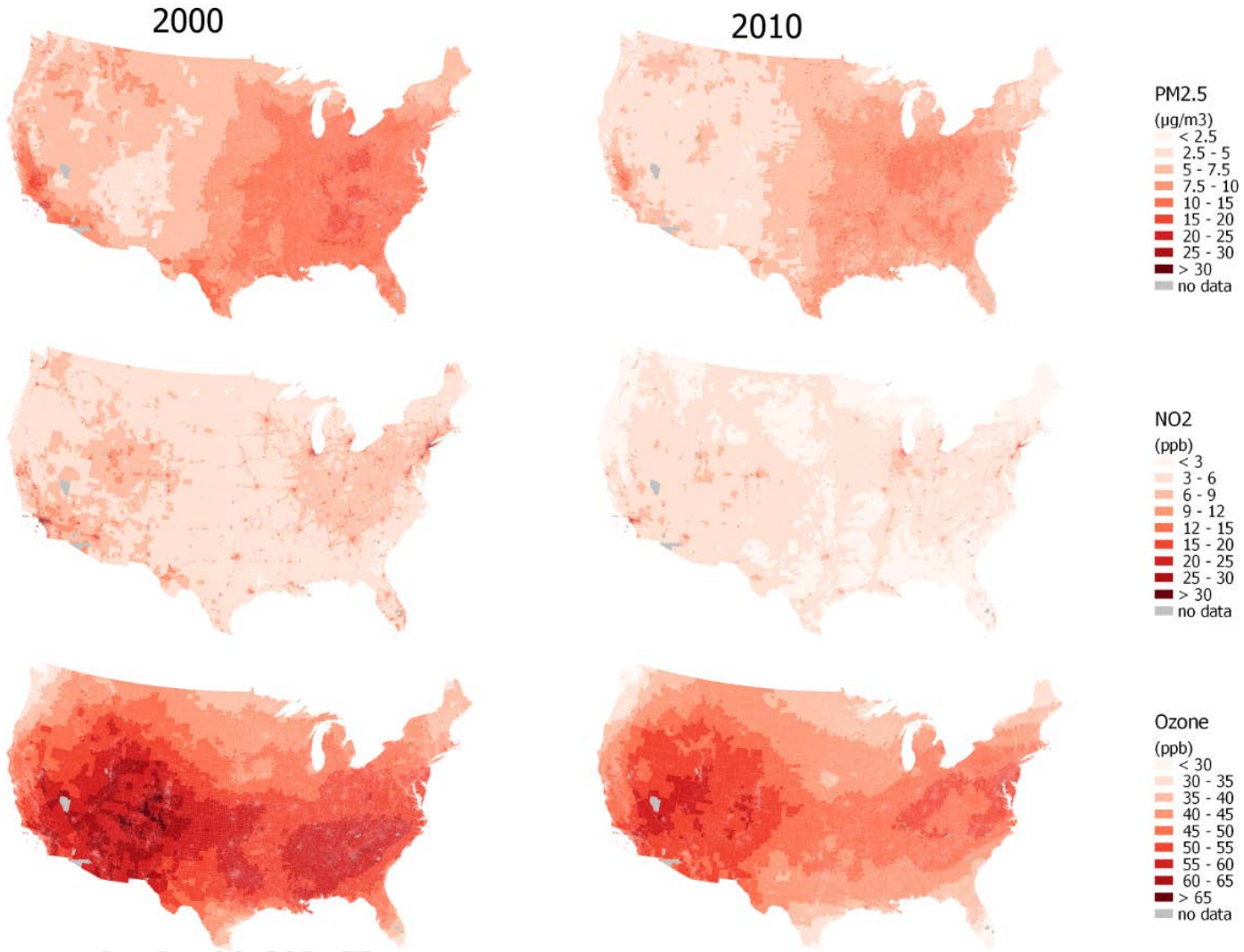

Figure 4. Population-weighted mean predicted annual-average concentrations, by pollutant and year, for Census Block Groups in the contiguous US, using the "best" IEG regression model developed here.

SUPPLEMENTAL MATERIAL

Concentrations of criteria pollutants in the contiguous U.S., 1979 – 2015: Role of model parsimony in integrated empirical geographic regression

Sun-Young Kim, Matthew Bechle, Steve Hankey, Lianne Sheppard, Adam A. Szpiro, Julian D. Marshall.

Data generation and processing for satellite air pollution estimates

Briefly, five aerosol optical depth (AOD) satellite retrievals (from several instruments and retrieval algorithms) are combined with (1) satellite-based measurements of vertical aerosol profiles, (2) modeled AOD and ground-level $PM_{2.5}$ from a global chemical transport model (GEOS-Chem), and (3) ground-based AOD measurements from the aerosol robotic network (AERONET) to estimate annual ground-level $PM_{2.5}$ on a 0.1° grid (van Donkelaar et al., 2016). We also obtain daily L2 (i.e., processed data at native instrument resolution) surface-level CO multispectral (combined near infrared and thermal infrared) retrievals (v7) from the Measurements of Pollution in The Troposphere (MOPITT) sensor on the National Aeronautics and Space Administration (NASA)'s Terra satellite for years 2001-2016 (Deeter et al., 2017). For each year, daily surface-level CO measurements are screened for missing data and solar zenith angle (SZA) >80°, then oversampled onto a 0.25° × 0.25° grid. Oversampling is an averaging method for satellite data that takes advantage of overlapping pixels when temporally averaging measurements at native resolution; all pixels falling within a circular buffer centered on each grid cell are averaged to that grid cell. Tropospheric NO_2, SO_2, and HCHO are derived from daily measurements from the Ozone Monitoring Instrument (OMI) onboard the NASA Earth Observing System (EOS)-Aura satellite. We obtain daily L2 (native instrument resolution) tropospheric NO_2 retrievals (DOMINOv2) for years 2005-2015 from the Tropospheric Emission Monitoring Internet Service (www.temis.nl) (Boersma et al., 2011), along with daily L2 tropospheric HCHO for 2005-2016 and daily L3 (pre-gridded product) tropospheric SO_2 retrievals from NASA's Goddard Earth Sciences Data and Information Services Center (GES-DISC) for years 2005-2016 (Chance 2007; OMI Science Team 2012). In addition to annual

averages for tropospheric NO_2 and SO_2, we compute 3-year averages for NO_2 and long-term average for HCHO. Daily L2 NO_2 and HCHO data are screened for missing data, flags (including "row anomaly" flags: see http://projects.knmi.nl/omi/research/product/rowanomaly-background.php), SZA > 60°, cloud fraction >40%, and surface albedo >30%. For each year and for 3-year averages, screened daily tropospheric NO_2 are oversampled onto a 0.1° × 0.1° grid. HCHO is more difficult to detect from space, owing to a lower signal-to-noise ratio and spectral interference from other molecules in the same fitting window (De Smedt et al., 2008); we therefore oversample screened daily tropospheric HCHO for the entire 12 year period (2005-2016) onto a 0.25° × 0.25° grid. Daily gridded 0.25° × 0.25° L3 SO_2 data are screened for data flags (including "row anomaly") and temporally averaged to annual averages. Using the gridded products described above, we assign to each target location annual or long-term averages of daily observations (Table S2). For the years before or after the satellite data are available, we use the average of the 3 closest years.

Table S1. List of geographical variables and satellite air pollution estimates

Category	Measure	Variable description[b]
Traffic	Distance to the nearest road[a]	Any road, A1, intersection
	Sum within buffers of 0.05-15 km	A1, A2+A3, truck route, intersections
Population	Sum within buffers of 0.5-3 km	Population in block groups
Land use (Urban)	Percent within buffers of 0.05-15 km	Urban or Built-Up land (residential, commercial, industrial, transportation, urban)
		Developed low, medium, and high density
		Developed open space
Land use (Rural)	Percent within buffers of 0.05-15 km	Agricultural land (cropland, groves, feeding)
		Rangeland (herbaceous, shrub)
		Forest land (deciduous, evergreen, mixed)
		Water (streams, lakes, reservoirs, bays)
		Wetland
		Barren land (beaches, dry salt flats, sand, mines, rock)
		Tundra
		Perennial snow or Ice
Position	Coordinates	Longitude, latitude
Source	Distance to the nearest source[a]	Coastline
		Commercial area
		Railroad
		Railyard
		Airport
		Major airport
		Large port
Emissions	Sum of site-specific facility emissions within buffers of 3-30 km	$PM_{2.5}$
		PM_{10}
		CO
		SO_2
		NO_X
Vegetation	Quantiles within buffers of 0.5-10 km	Normalized Difference Vegetation Index (NDVI)
Imperviousness	Percent within buffers of 0.05-5 km	Impervious surface value
Elevation	Elevation above sea levels	Elevation value
	Counts of points above or below a threshold within buffers of 1-5 km	
Satellite estimate	Estimates in a grid	$PM_{2.5}$
		NO_2
		CO
		SO_2
		HCHO

a. Distances calculated to spatial features are truncated at 25 km

b. See the Multi-Ethnic Study of Atherosclerosis and Air pollution (MESA Air) Data Organization and Operating Procedures (DOOP) for data sources for these variables (https://www.uwchscc.org/MESAAP/Documents/MESAAirDOOP.pdf).

Table S2. Available years of satellite estimates for air pollution and metrics used for national prediction models

Pollutant	Year	Metric	
		Years with data	Years without data
NO_2	2005-2015	Annual average	3-year average for 2005-2007 before 2005
		3-year average	3-year averages for 2005-2007 before 2006
			3-year average for 2013-2015 after 2014
SO_2	2005-2016	Annual average	3-year average for 2005-2007 before 2005
CO	2001-2016	Annual average	3-year average for 2001-2003 before 2001
HCHO	2005-2016	12-year average	12-year average for 2005-2016 for all years
$PM_{2.5}$	1998-2014	Annual average	3-year average for 1998-2000 before 1998
			3-year average for 2012-2014 after 2014

Table S3. Cross-validation (CV) statistics of ozone national prediction models including no some and all geographic variables and/or satellite estimates for four metrics by summer and all seasons

					Conventional CV			Clustered CV		
Season	Metric	N of sites	Year	N of variables	Median	25%	75%	Median	25%	75%
All	24-hr mean	213-479	1979-1986	0	0.38	0.26	0.47	0.00	0.00	0.00
			1990-2015	3-30*	0.73	0.68	0.75	0.63	0.55	0.66
				All	0.64	0.59	0.67	0.40	0.25	0.50
	8-hr max			0	0.57	0.45	0.65	0.14	0.03	0.19
				3-30	0.70	0.64	0.74	0.46	0.32	0.54
				All	0.67	0.59	0.71	0.25	0.11	0.40
	8-hr mean			0	0.60	0.48	0.67	0.17	0.07	0.23
				3-30	0.71	0.66	0.77	0.50	0.36	0.57
				All	0.69	0.61	0.74	0.32	0.16	0.45
	1-hr max		1980-1986	0	0.57	0.44	0.63	0.18	0.06	0.29
			1990-2015	3-30	0.67	0.58	0.73	0.43	0.23	0.50
				All	0.63	0.52	0.69	0.18	0.00	0.35
Summer (May-Sep)	24-hr mean	232-916	1979-2015	0	0.50	0.46	0.57	0.08	0.00	0.14
				3-30	0.69	0.62	0.73	0.46	0.29	0.52
				All	0.61	0.55	0.64	0.28	0.14	0.36
	8-hr max			0	0.72	0.65	0.77	0.34	0.17	0.46
				3-30	0.73	0.68	0.78	0.41	0.24	0.52
				All	0.73	0.65	0.78	0.36	0.16	0.47
	8-hr mean			0	0.72	0.65	0.78	0.32	0.16	0.47
				3-30	0.74	0.66	0.79	0.41	0.23	0.52
				All	0.73	0.64	0.79	0.36	0.15	0.48
	1-hr max			0	0.73	0.57	0.77	0.36	0.22	0.47
				3-30	0.74	0.59	0.78	0.39	0.27	0.54
				All	0.74	0.58	0.78	0.32	0.18	0.48

* Summaries of the highest R^2s on each year

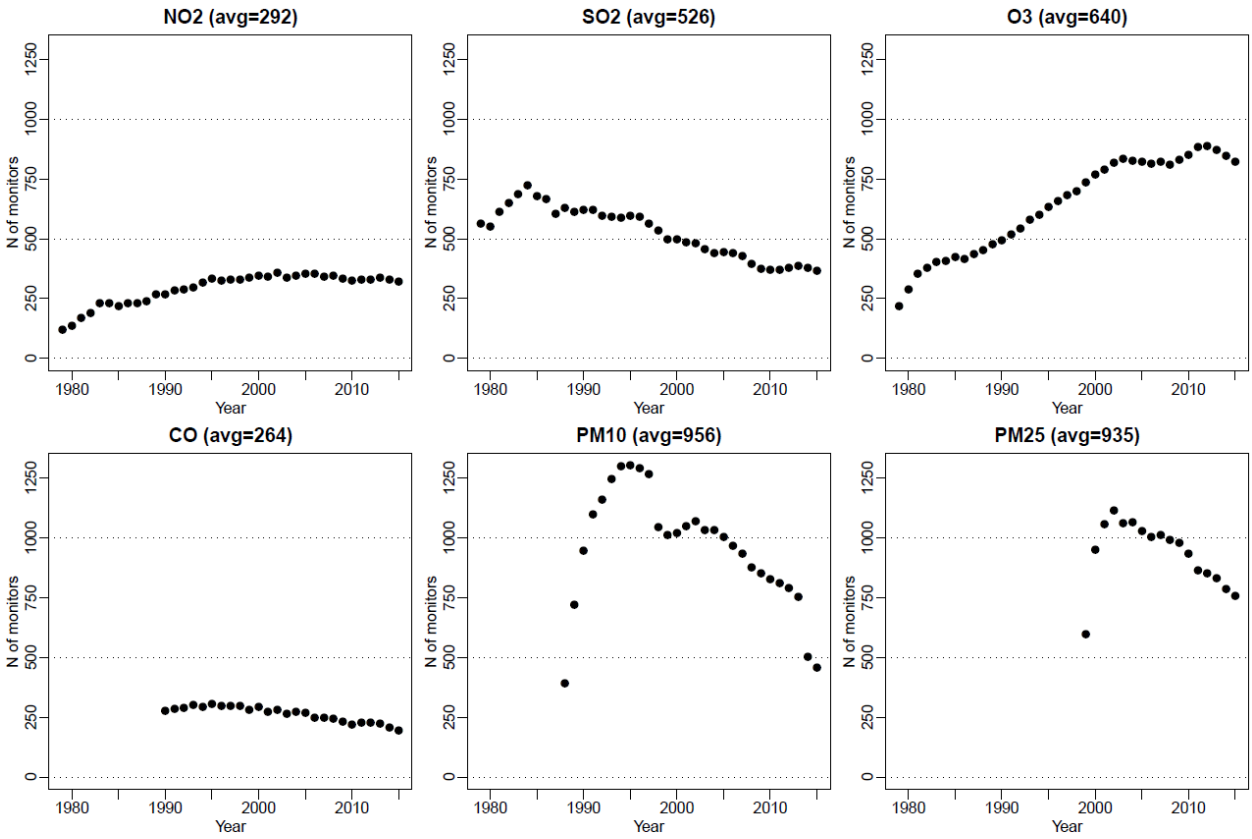

Figure S1. Numbers of regulatory monitoring sites that meet our site inclusion criteria for computing representative annual average concentrations of six criteria air pollutants for 1979-2015 in the continental U.S.

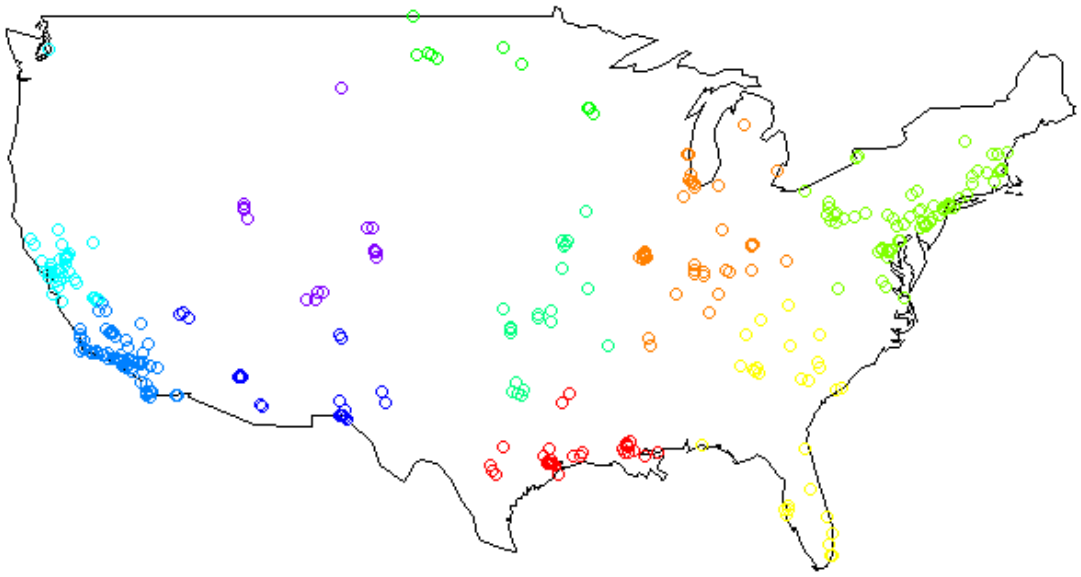

Figure S2. Map of 10 spatial clusters of 345 NO_2 regulatory monitoring sites in 2000 determined by k-means clustering

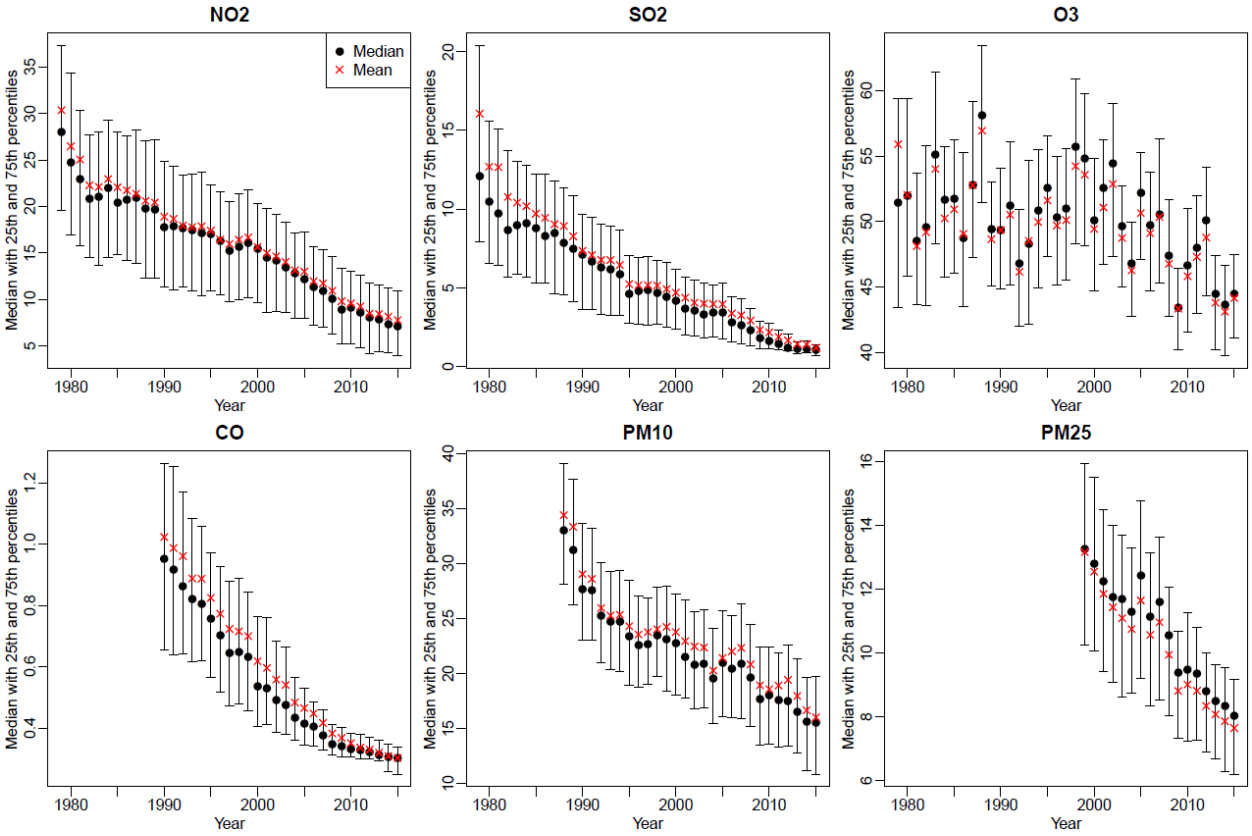

Figure S3. Quantile-based plots of annual average concentrations of six criteria air pollutants across all regulatory monitoring sites for 1979-2015 in the contiguous U.S.

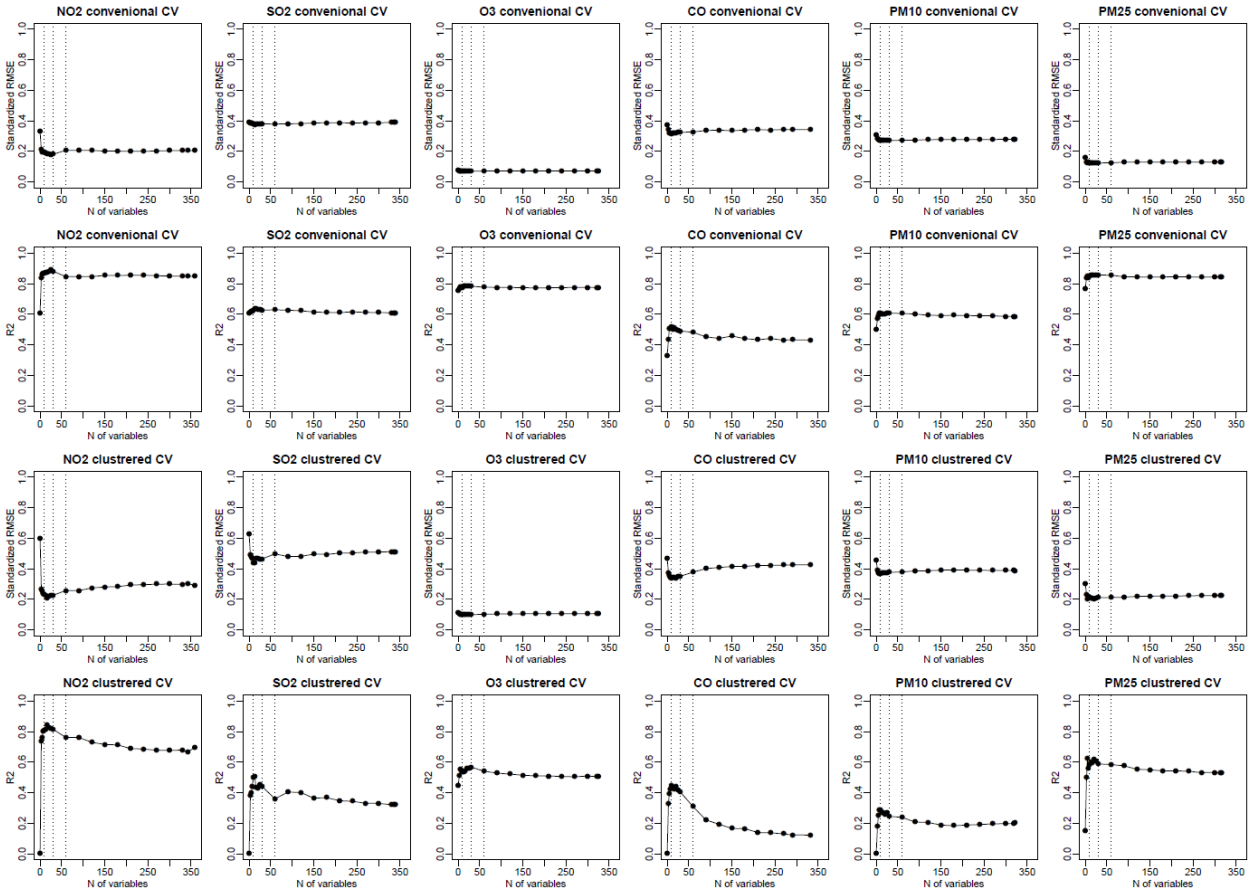

Figure S4. The relationship between numbers of variables and cross-validation (CV) statistics from national prediction models of six criteria air pollutants in 2000 by conventional and clustered cross-validation (vertical lines for 10, 30, and 60)

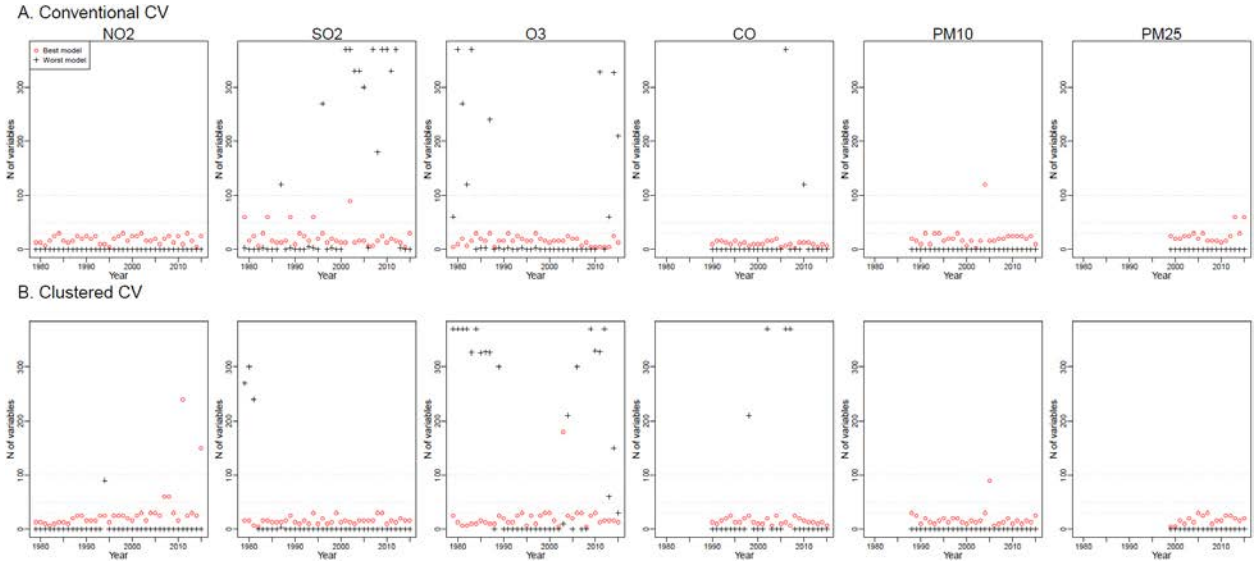

Figure S5. Numbers of variables selected for the best and worst national prediction models with the highest and lowest cross-validated R^2s, respectively (which was also the model with the lowest and highest standardized root mean square error), by pollutant and CV type (conventional CV, clustered CV). For ease of reading, figures include horizontal lines for y-axis values of 30, 50, and 100.

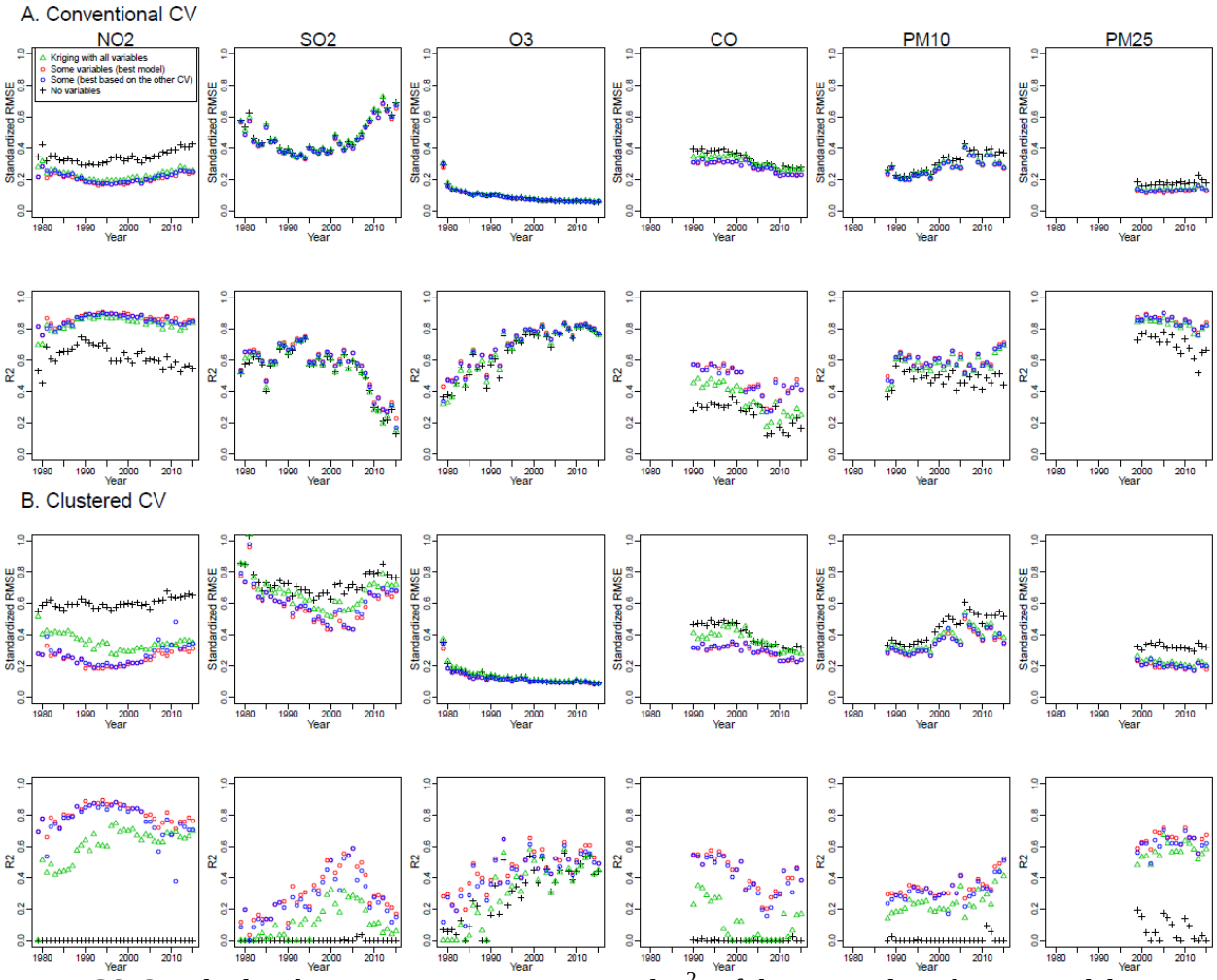

Figure S6. Standardized root mean square errors and R^2s of the national prediction models including no variables, some variables (i.e., between 3 and 30 variables), and all variables from conventional and clustered cross-validation, by year and pollutant, for the contiguous U.S. "Best" models determined by two types of cross-validation as one of the some-variables models.

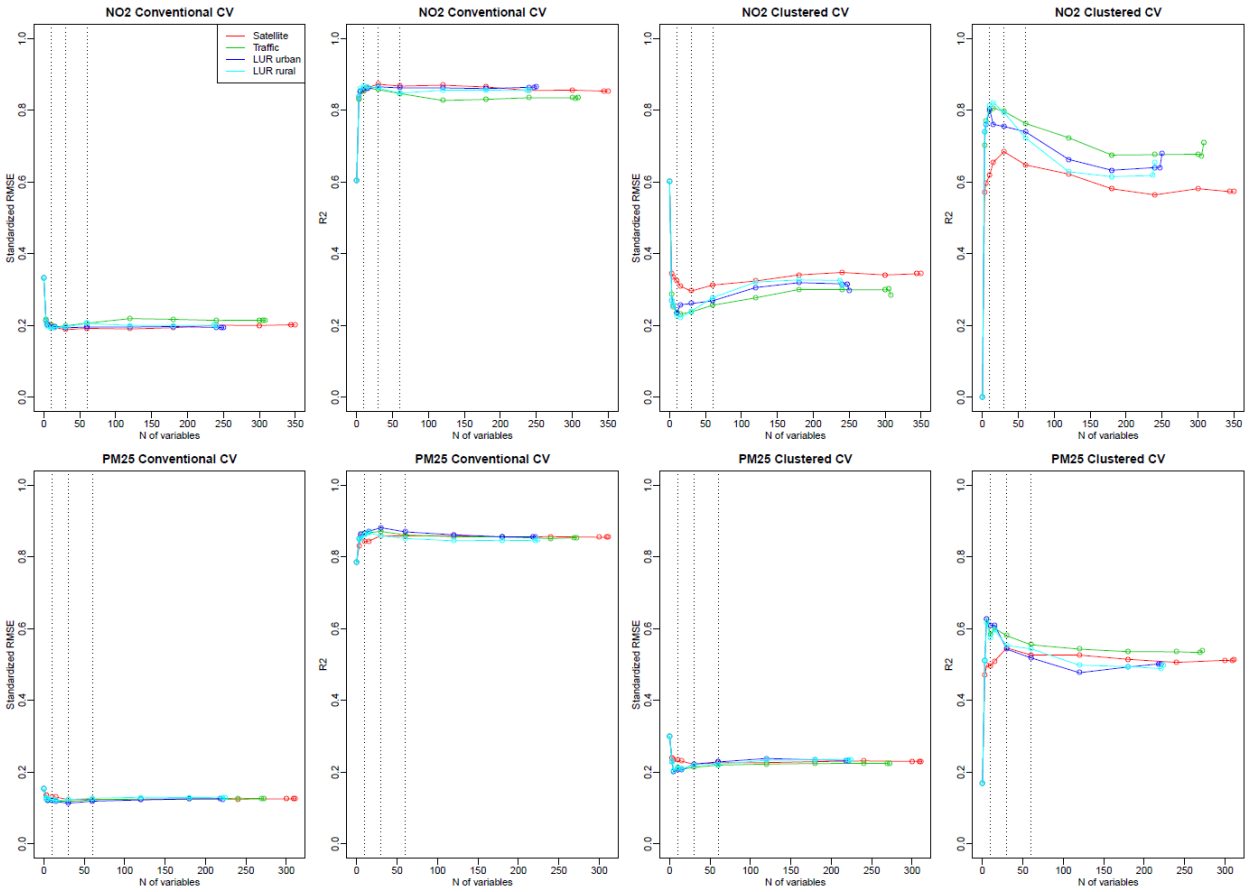

Figure S7. The relationship between numbers of variables and cross-validation (CV) statistics from national prediction models of NO_2 and $PM_{2.5}$ in 2000 by exclusion of a different category of geographic variables and satellite air pollution estimates by conventional and clustered cross-validation. For ease of reading, vertical lines are shown at x-axis values of 10, 30, and 60.

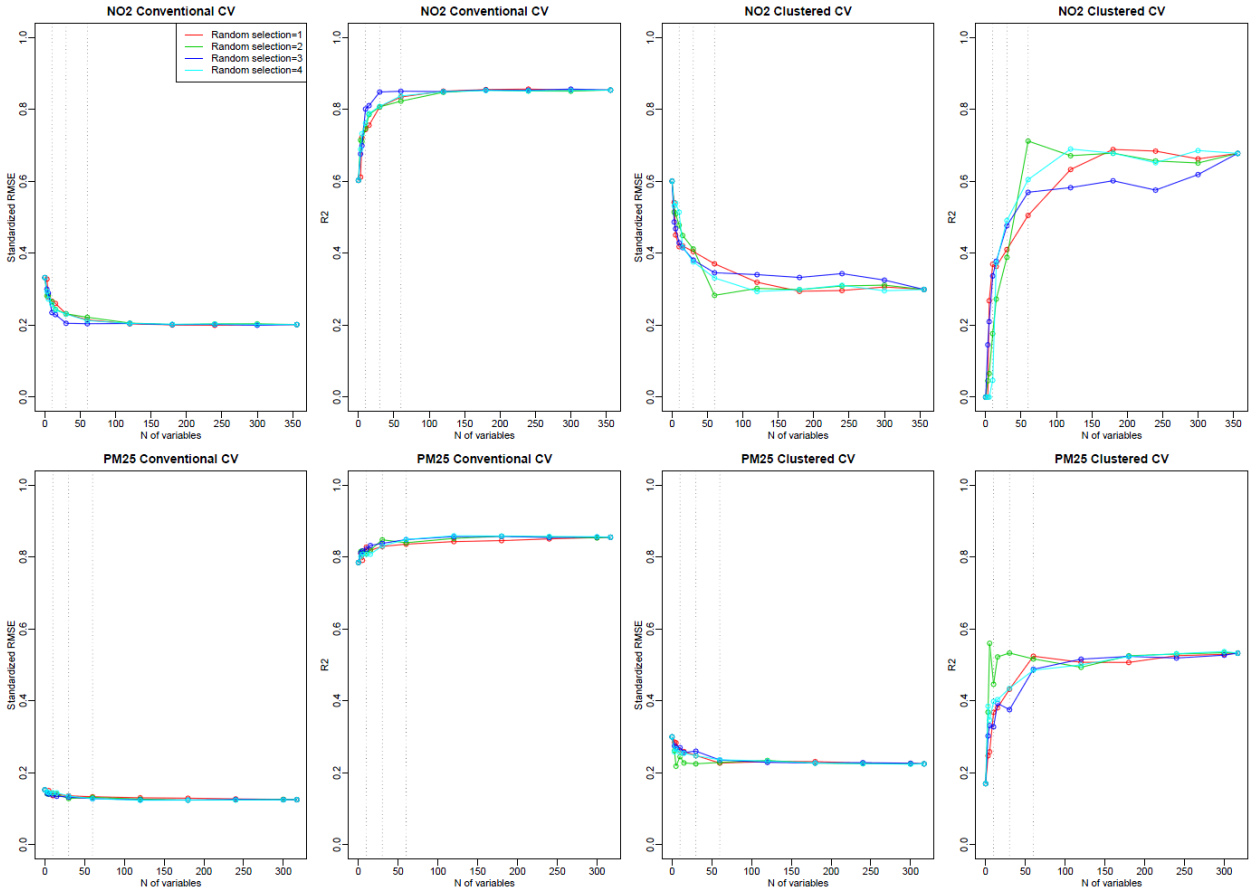

Figure S8. The relationship between numbers of randomly selected variables and cross-validation (CV) statistics from national prediction models of NO_2 and $PM_{2.5}$ in 2000 by conventional and clustered cross-validation. For easy of viewing, vertical lines are shown at x-axis values of 10, 30, and 60

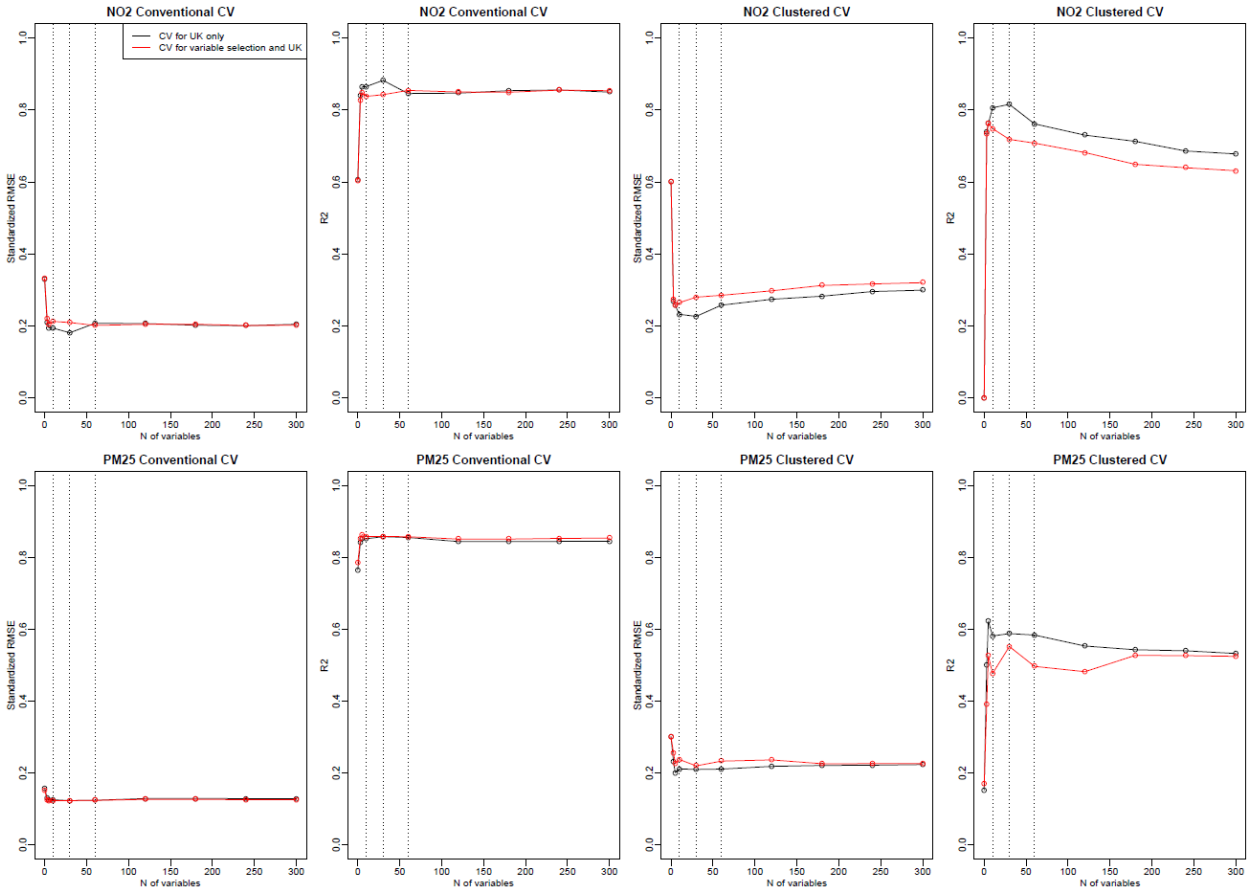

Figure S9. The relationship between numbers of variables and cross-validation (CV) statistics including forward selection, estimation of PLS predictors, and parameter estimation in national prediction models of NO_2 and $PM_{2.5}$ in 2000 by conventional and clustered CV. Vertical lines shown for x-axis values of 10, 30, and 60.

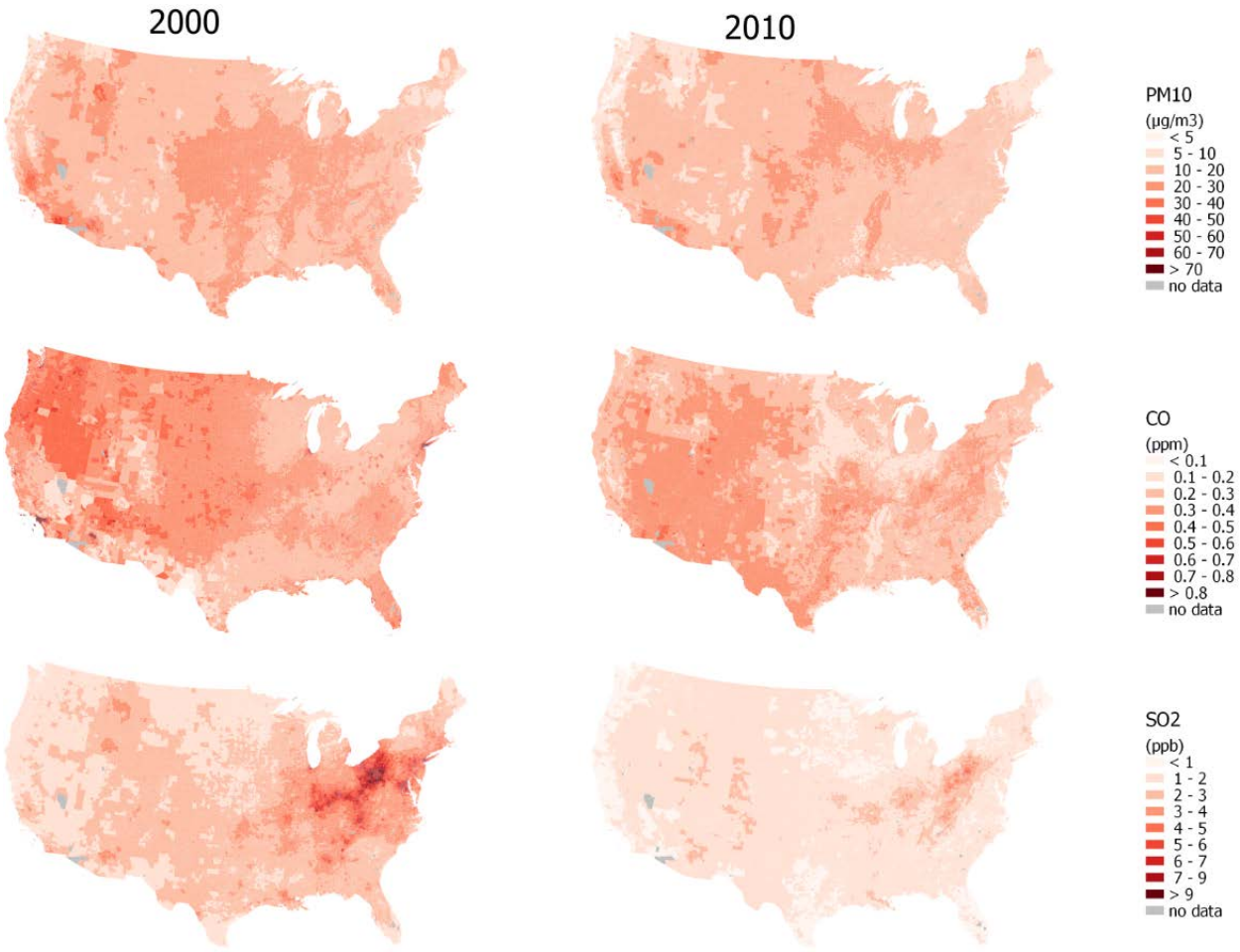

Figure S10. Maps of predicted annual averages of PM_{10}, CO, and SO_2 from the best national prediction models mostly including 3-30 variables for 2000 and 2010 in the contiguous U.S.