Environmental Science & Technology

Article

pubs.acs.org/est

Satellite-Based NO_2 and Model Validation in a National Prediction Model Based on Universal Kriging and Land-Use Regression

Michael T. Young,[*,†] Matthew J. Bechle,[‡] Paul D. Sampson,[⊥] Adam A. Szpiro,[§] Julian D. Marshall,[‡] Lianne Sheppard,[§,∥] and Joel D. Kaufman[†,∥]

[†]Department of Epidemiology, University of Washington 4225 Roosevelt Way NE, Seattle, Washington 98105, United States

[‡]Civil & Environmental Engineering, University of Washington, Wilcox 268, Seattle, Washington 98195, United States

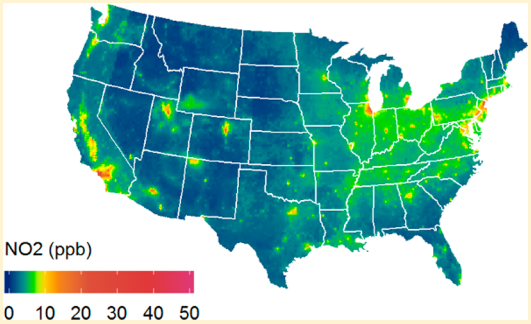
[§]Department of Biostatistics, University of Washington 1705 NE Pacific Street, Seattle, Washington 98195, United States

[∥]Department of Environmental and Occupational Health Sciences, University of Washington 1959 Pacific Street, Seattle, Washington 98195, United States

[⊥]Department of Statistics, University of Washington B313 Padelford Hall, Northeast Stevens Way, Seattle, Washington 98195, United States

Supporting Information

ABSTRACT: Epidemiological studies increasingly rely on exposure prediction models. Predictive performance of satellite data has not been evaluated in a combined land-use regression/spatial smoothing context. We performed regionalized national land-use regression with and without universal kriging on annual average NO_2 measurements (1990−2012, contiguous U.S. EPA sites). Regression covariates were dimension-reduced components of 418 geographic variables including distance to roadway. We estimated model performance with two cross-validation approaches: using randomly selected groups and, in order to assess predictions to unmonitored areas, spatially clustered cross-validation groups. Ground-level NO_2 was estimated from satellite-derived NO_2 and was assessed as an additional regression covariate. Kriging models performed consistently better than nonkriging models. Among kriging models, conventional cross-validated R^2 (R^2_{cv}) averaged over all years was 0.85 for the satellite data models and 0.84 for the models without satellite data. Average spatially clustered R^2_{cv} was 0.74 for the satellite data models and 0.64 for the models without satellite data. The addition of either kriging or satellite data to a well-specified NO_2 land-use regression model each improves prediction. Adding the satellite variable to a kriging model only marginally improves predictions in well-sampled areas (conventional cross-validation) but substantially improves predictions for points far from monitoring locations (clustered cross-validation).



1. INTRODUCTION

Epidemiological research is a critical component underpinning regulatory standards for air quality, and increasingly relies on statistical air pollution models. While older research relied on a few monitors in population centers, this approach provides insufficient detail about within-urban spatial variability.[1] Estimating the health effects of air pollutants often requires multiregion cohort studies with large sample sizes and requires characterizing contrasts both within and between regions.

To accommodate health studies over multiple regions, it often is desirable to develop a single model that can predict to an entire country[2−7] or continent and be efficiently used in multiple studies. Ideally these models can predict concentrations accurately even in areas without ground-level monitoring data. Imperfect predictions can cause bias and variability in health effect estimates,[8] thus motivating the development of improved models that minimize exposure estimation error. A standard approach to building prediction models for air pollution exposure is land-use regression, where geographic covariates are inputs to a linear model predicting a specific air pollutant.[9−11,1] Use of remote sensing data derived from satellite sources has been previously been demonstrated to add value when incorporated into a nonkriging land-use regression approach.[2] However, research assessing the performance of satellite data in the context of a spatial kriging model is limited, particularly for modeling NO_2.

Here we develop multiple LUR approaches for predicting annual averages of ambient nitrogen dioxide (NO_2) in the contiguous United States. Our approach incorporates spatial regression (universal kriging), as described elsewhere.[12] Like with conventional land-use regression (which does not include a kriging component), we rely on model parameters estimated from observed data at regulatory monitoring locations. For the

Received: October 16, 2015
Revised: February 17, 2016
Accepted: February 29, 2016
Published: February 29, 2016

DOI: 10.1021/acs.est.5b05099
Environ. Sci. Technol. 2016, 50, 3686−3694

base model, we incorporate a large suite of geographic covariates into the regression component via partial least-squares regression dimension reduction.

The present analysis focuses on assessing the added predictive performance of incorporating satellite NO_2 data into the model, as well as a comparison between universal kriging and nonkriging (i.e., land-use regression) approaches. For model assessment, we sought to not only evaluate overall prediction accuracy, but specifically evaluate prediction accuracy at points located far from monitoring sites. Model assessment in this context—prediction to areas with poor monitoring coverage—required the use of an alternate validation design, spatially clustered cross-validation. This approach (which is detailed in methods Section 2.7) complements conventional cross-validation by evaluating prediction performance in a monitoring network that is spatially clustered and has large gaps in monitoring coverage such as the Air Quality System (AQS) network.

To summarize the overall predictive capability of different modeling approaches, we present out-of-sample cross-validated R^2 estimates (calculated using the RMSE to reflect bias also; see Section 2.7) for each model in each year. Specifically, we compare performance between R^2 estimates for models that incorporated the satellite variable and for models without satellite data. We additionally compare models with and without kriging. For each model, we calculated cross-validated R^2 estimates using two different cross-validation designs: conventional random cross-validation groups and spatially clustered cross-validation groups. We conclude with a conceptual discussion of the sensitivity of conventional k-fold cross-validation statistics to the conditions of the monitoring network and how this validation design may not be the most informative estimate of model performance when predicting to cohort locations.

2. MATERIALS AND METHODS

2.1. Modeling Extent and Regions. We developed statistical models for the contiguous United States (i.e., excluding Alaska and Hawaii). In order to more flexibly model NO_2, we subdivided the country into three ad hoc modeling regions ("East", "Mountain West", and "West Coast")[12] shown in Figure 1. The goal of using separate modeling regions was to allow different modeling parameters for different areas of the countries. These regions were chosen to reflect terrain differences (specifically ruggedness) which might have an effect on kriging parameters. These regions have been used previously and shown to be improve predictive performance in a similar context.[12] Modeling parameters differed by region as explained in the statistical methods section below.

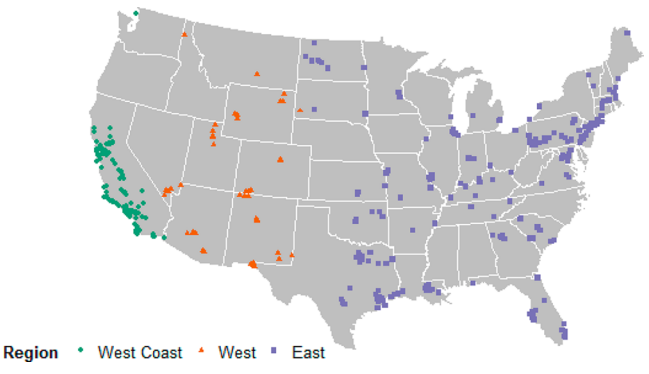


Figure 1. Year 2006 NO_2 monitoring locations. EPA Air Quality System (AQS) monitoring locations in year 2006 meeting our completeness criteria. Point color indicates the region to which each monitor belongs.

2.2. Monitoring Data. We averaged daily NO_2 concentrations from the AQS network to the calendar year at each monitoring site (1990–2012). To avoid seasonality bias, we excluded sites that had an inconsistent time-series in a year. Specifically, we excluded sites in years with fewer than 244 monitored days (i.e., 69% of days) or with a sequential gap of 45 days or greater. In order to stabilize the residual variance, we transformed the monitoring data to the square root scale.[12,13]

2.3. Geographic Covariates. We used a geographic information system (GIS) to calculate covariates representing land-use characteristics that could predict spatial variability in air pollutants. We calculated 801 geographic covariates which fall into two categories: *proximity variables*, measures of proximity to a nearest feature, and *buffer variables*, measures of geographic features within a given radius. Proximity variables include measures of distance to major roadways (specifically roads classified as A1, A2, and A3 according to the census feature class) as well as distance to commercial zones, airports, ports, railroads, and railyards. Buffer variables measure the sum (area, length, or count) or average of various features within a given radius. For every feature, we estimated several buffer variables of different radii. Specific buffer variables included measurements of the area of each USGS land-use classification (both 1990 and 2000), and the average Normalized Difference Vegetation Index (NDVI) value, sum of roadways lengths (A1, A2, and A3), to name a few. For a more detailed sample of the covariates we used, see the Supporting Information and Sampson el al. or Mercer et al.[12,14]

We calculated the values of each geographic variable at monitoring locations using ArcMap version 10.2. We then excluded variables with undesirable qualities for modeling. Exclusion criteria were based on the distribution of each variable calculated at monitoring locations. Specifically, we excluded variables where 80% or more of the values were the same, and variables with z-score values greater than 10 in the monitoring data set. Removing variables with outliers in the monitoring data set (i.e., based on z-score), avoids modeling with potentially unreliable variables which might lead to occasional extreme predictions in the prediction data set. We additionally removed land use variables whose maximum observed buffer area is less than 10% to avoid modeling with rare (low-percentage) land-use types that might lead to a less stable model. Because the set of monitoring locations differed by year, we applied this variable exclusion process separately for each year, and therefore the number of variables remaining after exclusions differed by year (range: 345–370). Out of the 801 geographic covariates we calculated using GIS, 418 of these variables were used in at least one model after exclusions.

2.4. Dimension Reduction using Partial Least Squares. Due to the large number of correlated geographic variables, we used a dimension reduction technique to simplify the model selection process. We chose partial least-squares (PLS) because it is specifically designed to deal with large sets of collinear variables and maximizes covariance between the predictors and outcome while minimizing overfitting via dimension reduction.[15] Incorporating PLS regression into the prediction approach both avoids overfitting of models, and most effectively

DOI: 10.1021/acs.est.5b05099
Environ. Sci. Technol. 2016, 50, 3686–3694

chooses predictors from a high dimensional geographic database.[15] We performed PLS separately for each modeling year. We considered models including the first 2, 3, 4, and 5 PLS components, and then selected the optimal number of PLS components via cross-validated performance as described in Section 2.7.

2.5. Satellite Data. The satellite data source and preprocessing methods have been described previously.[16,17] Briefly, total tropospheric NO_2 was measured via satellite images from the Ozone Monitoring Instrument (OMI) on the Aurora satellite. Images were taken with a resolution of 13 × 24 km^2 at nadir between 12:00−15:00 local time. Columnar satellite data from 2005 to 2007 were averaged to the calendar year and processed to estimate ground-level annual average estimates using the GEOS-Chem reactive transport model.[16,18] For the present analyses, we calculated annual averages for all three years at each monitoring location using a raster-to-point conversion technique (Moore neighborhood average). We combined the three calendar year averages into a single variable estimating the average satellite-measured ground-level NO_2 for 2005 to 2007. We then transformed the satellite data to the square-root scale to match the scale of the outcome variable.

2.6. Statistical Model. The statistical approach for spatial prediction has been described previously.[12] Briefly, we used universal kriging which combines linear regression with a distance-based spatial smoothing model for the covariance between points. The benefit of this spatial smoothing approach over conventional land-use regression is that it leverages correlation in the regression residuals to improve predictions. Universal kriging was chosen because it is a likelihood-based method which allows simultaneous estimation of smoothing parameters and regression parameters (whereas other methods smooth on the residuals using a two-staged approach). We modeled each year of NO_2 separately such that PLS components, regression coefficients, and residual spatial parameters differed by year, thereby allowing for flexibility to model changes over time. We modeled annual average NO_2 separately by region such that each region had its own regression coefficients and variograms to account for spatially varying differences in geographic factors across the United States. A previous analysis showed improved performance using this regionalized approach.[12]

To determine the added performance attributable to the satellite variable, we modeled each year with and without this variable. We included the satellite variable directly as a spatial covariate in the model rather than combined with the other geographic covariates as input to the PLS procedure. This approach allowed us to directly compare model performance due to the inclusion of the satellite variable in particular.

The formula for the models without the satellite variable can be written as

$$Y(s) = \beta_{0,j} + \beta_{1,j} \times X(s) + \varepsilon(s)_j$$

where $\cdot(s)$ indicates a variable is modeled or calculated as a function of space. $Y(s)$ denotes the square root of annual average NO_2 values at monitoring locations. $X(s)$ denotes the matrix of spatially varying PLS components. $\varepsilon(s)$ indicates spatially varying residuals, modeled using universal kriging.[12] Subscripts j indicate separate parameters for each region.

The models with the satellite variable are

$$Y(s) = \beta_{0,j} + \beta_{1,j} \times X(s) + \beta_{2,j} \times Z(s) + \varepsilon(s)_j$$

where $Z(s)$ denotes the spatially varying square root of ground-level estimated satellite data averaged over 2005 through 2007.

To determine the added performance from kriging, we modeled each year with and without kriging. The "without kriging" models are land-use regressions that did not model spatial dependence in the residuals and instead treated the residuals as independent and identically distributed with a constant variance.

2.7. Cross-Validation and Model Assessment. We used cross-validation for two purposes: (1) to compare performance among *model types* (there were four model types per year: with kriging and satellite; with kriging but not satellite; with satellite but not kriging; and, with neither satellite nor kriging); (2) to select the optimal number of PLS components for each year and model type. As mentioned above, each model year and type was fit four times (using 2, 3, 4, or 5 PLS components); we selected the optimal number of PLS components separately for each model type in order to ensure fairness when comparing model types that would otherwise have different amounts of flexibility (i.e., comparing a kriging model with a nonkriging model). We defined the optimal number of PLS components based on the aggregate performance over all years for each model type. Specifically, we averaged R^2 values across modeling years separately for each combination of model type and for each number of PLS components, then, for each modeling type, selected the number of PLS components with the highest R^2 averaged over all years.

We performed cross-validation on the entire modeling approach including variable exclusions, PLS regression, and universal kriging. To facilitate more precise comparison between models with and without satellite data and between models with and without kriging, cross-validation groups (of the same cross-validation approach) were identical for models of the same year.

We used two methods to assess prediction performance: conventionally designed cross-validation (20-fold) and spatially clustered cross-validation (20-fold). Both approaches estimate out-of-sample model performance; they differ in how the training and test data sets are sampled.

For conventional cross-validation, we divided monitoring locations randomly into 20 groups across regions. The set of monitors in one group (the *test set (5%)*) was excluded and then the modeling approach was performed on the remaining monitors (the *training set (95%)*). The model fit on the training set was used to make predictions to monitors in the test set, resulting in out-of-sample predictions at those monitoring locations. This process was repeated 20 times, until all groups served as the test set once, thereby creating out-of-sample predictions for all monitoring locations.

We additionally performed cross-validation using groups that were selected to be spatially clustered in a design we refer to as spatially clustered cross-validation. We determined clusters using a k-means algorithm in the R Project for Statistical Computing *stats* package (specifically the kmeans function)[19] using a prespecified number of clusters for each region based on the distribution of monitors by region such that spatially clustered cross-validation groups were contained entirely within regions. For more detail on this approach, see the Supporting Information.

We used spatially clustered cross-validation to assess model performance in areas far from monitoring locations. Calculating prediction validity at unmonitored locations is not possible, so we designed our approach to simulate prediction at

DOI: 10.1021/acs.est.5b05099
Environ. Sci. Technol. 2016, 50, 3686−3694

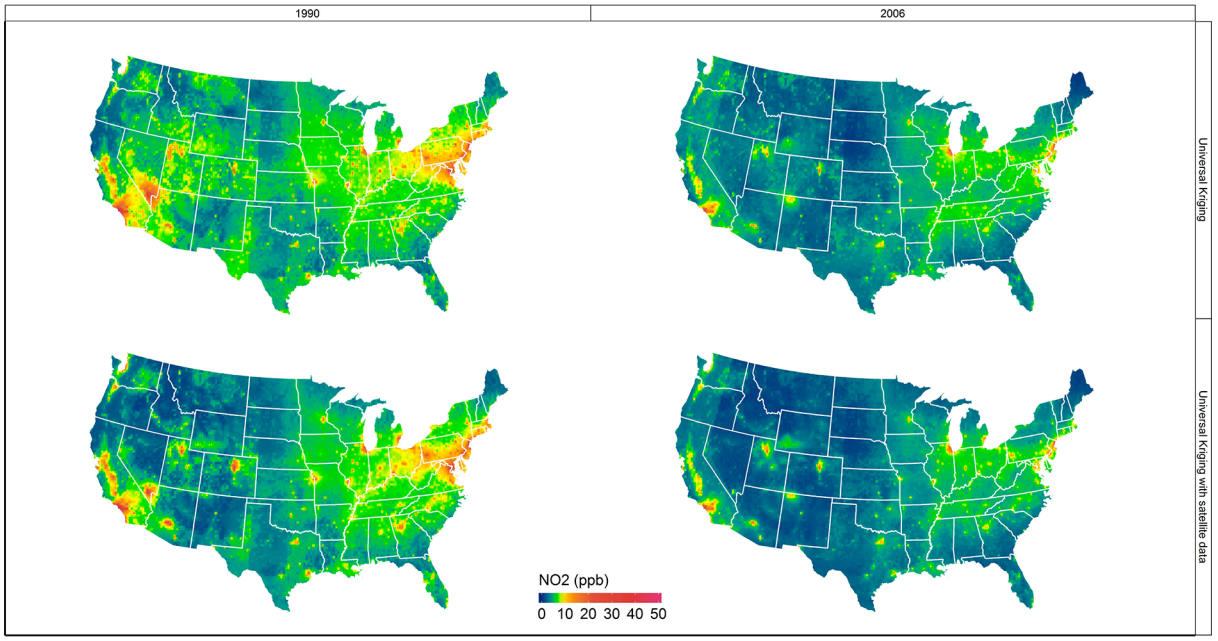


Figure 2. National predictions smoothed from estimates at grid locations spaced 25 km apart. Predicted annual average 1990 and 2006 NO_2 values on a 25 × 25 km national grid illustrate differences in the spatial distribution over time and between models with and without satellite data. All predictions shown are from models with 2 PLS components.

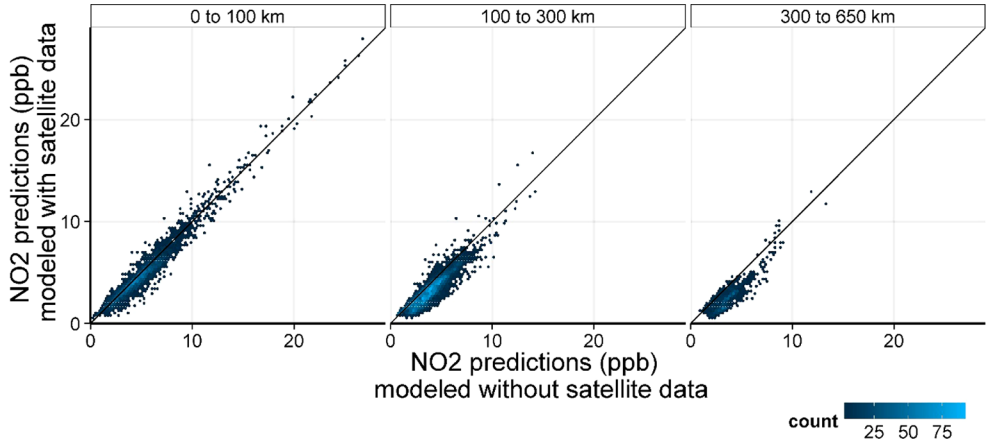


Figure 3. Year 2006 national predictions on a 25 km grid comparing models with and without satellite data, stratified by distance to nearest monitoring site. Color represents point density (to represent overplotting). Points below the identity line represent predictions that were lower in the satellite model than comparable predictions in the model without satellite data.

unmonitored locations by predicting to left-out groups of spatially clustered monitors. The resulting predictions represent model performance when extrapolating to unmonitored regions such as the large areas lacking monitors in the AQS network (e.g., see Figure 1).

We used the predictions from conventional and spatially clustered cross-validation to calculate model performance statistics. Specifically, we squared the predictions at monitoring locations (to back-transform from the square-root scale) such that validation statistics represent prediction accuracy on the native scale. We chose R^2 based on the root mean squared error of prediction (RMSEP) as our primary performance metric;[20] this value is computed as 1-(RMSEP^2/Var(Obs)) and differs from the R^2 computed as the squared correlation. Specifically, the RMSEP-based R^2 value incorporates bias such that biased models have lower R^2 values compared to the correlation-based R^2 value which only reflects precision. Since each year has four models (with and without satellite data; kriging and non-kriging) and two cross-validation designs (conventional and spatially clustered), we calculated eight R^2 values per year.

RESULTS

3.1. Data. From 1990 through 2012, the number of monitors meeting the completeness criteria varied between 271 and 361. Figure 1 shows the locations of NO_2 monitors in 2006; the substantial clustering of AQS sites observed there is consistent across the modeling years. The interquartile range of the monitoring data in 2006 was 8.1 ppb.

3.2. Modeling Predictions and Assessment. Figure 2 shows predictions from kriging models (with and without satellite NO_2) mapped on a 25 × 25 km^2 national grid in selected years (1990, 2006). Predicted NO_2 values tend to be higher in urban (versus rural) areas, and to some extent for rural areas in the East (compared to West and Mountain West).

DOI: 10.1021/acs.est.5b05099
Environ. Sci. Technol. 2016, 50, 3686−3694

Subtle differences are notable in the national grid predictions between models with and without satellite data. Figure 3 demonstrates these differences by comparing national grid predictions between kriging models with and without satellite NO_2 data in the year 2006, showing that the largest differences between model predictions are at locations farthest from monitoring locations. Typically, incorporating satellite data led to further decreases to already-low predictions. The interquartile range of the year 2006 kriging model national grid predictions was 18.0 ppb for the satellite model and 15.9 ppb for the nonsatellite model. Model predictions at monitoring locations (using conventional random cross-validation) had interquartile ranges consistent with the IQR of the measured monitoring data: 8.2 ppb for the satellite model and 8.3 ppb for the nonsatellite model.

Selected models included between 2 and 4 PLS components, with fewer for kriging models (range: 2−3 PLS components) than for nonkriging models (3−4 components) (Table 1). Satellite models have a slight trend of fewer PLS components compared to nonsatellite models (both satellite and nonsatellite models have a range of 2−4 components).

Table 1. Average R^2 Values (1990-2012) Comparing All Model Types[a]

		conventional CV		clustered CV	
satellite covariate	kriging	# PLS comp	avg. R^2	# PLS comp	avg. R^2
yes	no	3	0.81	4	0.72
no	no	4	0.72	4	0.59
yes	yes	2	0.85	2	0.74
no	yes	2	0.84	3	0.64

[a]Cross-validated MSE-based R^2 values were averaged over all modeling years (1990−2012) models with and without satellite data and with and without kriging for both conventional cross-validation and clustered cross-validation. Models with between 2 and 5 PLS components were considered, but only the set of models with the highest average R^2 values are shown.

For the kriging models, the difference between average conventional cross-validated R^2 values comparing satellite and nonsatellite models is only 0.01 (0.85 versus 0.84) (Table 1). For the nonkriging models, the difference in averages is 0.09 (0.81 versus 0.72). Figure 4a shows a comparison of conventional cross-validated R^2 model performance by year. In kriging models without satellite data, R^2 values ranged from 0.78 to 0.88. In kriging models with satellite data, values ranged from 0.80 to 0.89. In most years, the difference in performance between kriging models with and without satellite data is small with the satellite models performing slightly better (range: 0.00−0.03). Models without kriging had lower R^2 values and more improvement in performance with satellite data: satellite model R^2 values ranged from 0.72 to 0.85 and nonsatellite model R^2 values ranged from 0.61 to 0.77.

R^2 values are lower for all models under the clustered cross-validation approach compared to conventional cross-validation. Differences in model performance between models with and without satellite data are also larger under clustered cross-validation. According to the spatially clustered cross-validation approach, the difference in average R^2 values between satellite kriging models and nonsatellite kriging models is 0.10 (0.74 versus 0.64). For the nonkriging models, the difference in clustered R^2 averages is 0.13 (0.72 versus 0.59). Comparing clustered cross-validation across years, R^2 values range from 0.66 to 0.80 for kriging models with satellite NO_2 and from 0.50 to 0.70 for kriging models without satellite NO_2 (Figure 4b). Under the spatially clustered cross-validation design, test set monitors tend to be far from training monitors. Specifically, the average distance between a test-set monitor and its nearest training-set neighbor for the clustered cross-validation design ranged from between 235 and 281 km by year compared to between 30 and 50 km by year for the conventional cross-validation design (also see Table S3 in the Supporting Information).

4. DISCUSSION

To model annual average ambient NO_2 concentrations in the contiguous U.S., we developed a PLS land-use regression that included or omitted kriging, and included or omitted satellite-derived estimates. This modeling approach performed consistently well over a 23-year period with conventional cross-validation R^2 values ranging from 0.78 to 0.88 in kriging models without satellite data.

Since we modeled NO_2 concentrations on the annual average scale, model performance is a reflection of spatial variability in long-term averages or "chronic exposure". The U.S. EPA's National Ambient Air Quality Standards for NO_2 include two averaging times: a 1 h standard and an annual average standard. NO_2 models that predict long-term spatial contrasts are important for longitudinal cohort studies designed to investigate progression of chronic health outcomes; by comparison, models that predict short-term temporal contrasts are more appropriate for panel studies of biomarkers or acute mortality/morbidity studies such as time series studies.[21] Modeling each type of exposure contrast has its uses, but spatial contrasts can be more challenging to predict accurately due to the phenomenon of spatial misalignment.[22] Therefore, the high overall predictive accuracy of this approach is particularly notable. However, a feature of the modeling approach used here is that it was designed for use on long-term average observations, and would not be the best approach for estimating exposure on the acute (daily) time-scale because it is not designed to deal with seasonal or short-term correlation over time.

We chose to model each year separately; this approach allows for more flexible prediction of secular changes in NO_2 concentrations by independently estimating changing relationships between land-use regression covariates and monitored NO_2 concentrations over time.

4.1. KRIGING

Models with kriging performed better than models without kriging. This difference in performance existed despite the fact that we allowed the optimal number of PLS components to be different between model types (i.e., allowing nonkriging models to use more geographic covariate components). This suggests that kriging models predict variability in NO_2 that cannot be predicted by the land-use data. Kriging models even performed slightly better than their land use regression alternatives under the clustered cross-validation design, suggesting that spatial smoothing is effective even at relatively large distances (average distances between test and training monitors were 235−281 km).

4.2. Satellite Data. Using a conventional 20-fold cross-validation design there were only minor differences between

DOI: 10.1021/acs.est.5b05099
Environ. Sci. Technol. 2016, 50, 3686−3694

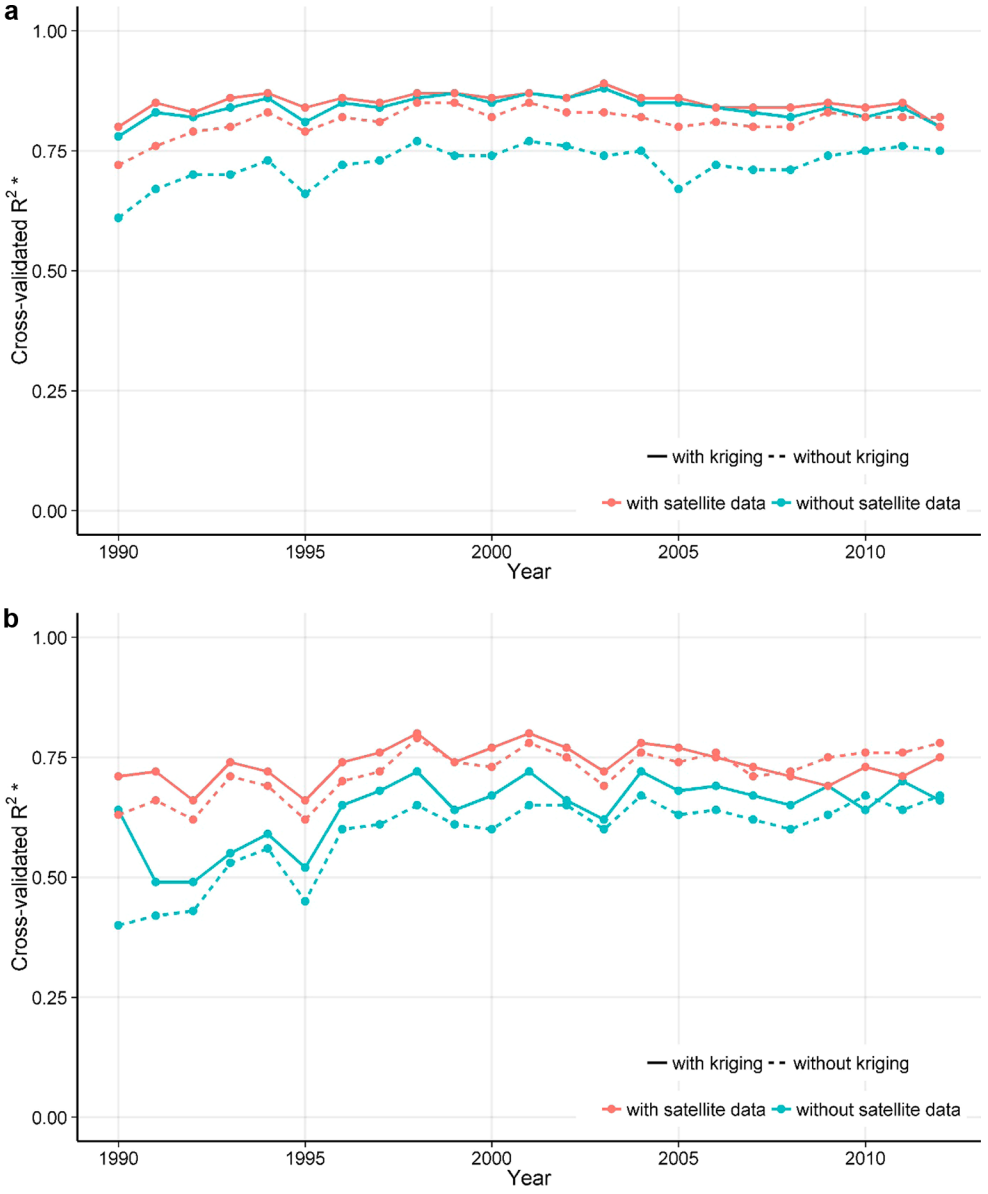


Figure 4. (a) R^2 values by year, estimated using conventional cross-validation. (b) R^2 values by year, estimated using spatially clustered cross-validation. Cross-validated R^2 values (conventional and clustered) comparing models with and without satellite data and models with and without kriging. The number of PLS components in each model is the same for each line—that is, for each combination of modeling characteristics (with and without satellite data and with and without kriging), but differs between lines. See Table 1 for the number of components chosen for each combination of modeling characteristics. *R^2 values are calculated as 1-(RMSEP^2/Var(Obs))

the R^2 values of the kriging models with and without satellite data (avg. cross-validated R^2 values 0.85 versus 0.84, respectively). The satellite data only marginally improved model performance by conventional standards, although even this modest improvement may be enough to justify its inclusion in future analyses. More notably, for both kriging and nonkriging models, adding satellite NO_2 data to the models substantially improved performance under the spatially clustered cross-validation design. This improvement occurred even though we allowed for flexibility in the optimal number of geographic components, suggesting that satellite data provide performance improvements that cannot be replaced by additional geographic covariates when it is necessary to extrapolate spatially.

A potential limitation of the data was our use of NO_2 satellite data restricted to years 2005−2007. However, model performance improvements by incorporating satellite data were not restricted to those years, indicating that the 2005−2007 average satellite data is providing predictive spatial information that is not limited to that time-period.

Performance metrics based on spatially clustered cross-validation may broadly represent the validity of our modeling approach under the less-than-ideal circumstance of predicting to locations far from AQS monitors, as may be the case in a national cohort study given the geographic range of participants and the substantial gaps in the monitoring network (Figure 1). Therefore, despite only minor improvements in conventional cross-validation performance, the substantially improved spatially clustered cross-validation performance of the satellite model over the nonsatellite model may indicate that the satellite data are particularly good at predicting to far-from-monitor locations compared to our other geographic covariates.

DOI: 10.1021/acs.est.5b05099
Environ. Sci. Technol. 2016, 50, 3686−3694

Given the nationally consistent relationship between the satellite data and ground-level NO_2 concentrations,[23,24] it is unsurprising that this variable leads to better performance far from monitoring locations.

Incorporating satellite data into future models may improve health effect estimation in subsequent epidemiologic analyses through a reduction in measurement error.[8,20] To characterize how satellite data specifically affected estimates, we compared predictions at national grid locations. Comparing predictions from models with and without satellite data, Figure 3 indicates that incorporating satellite data into the model tended to result in small decreases in predicted concentrations at national grid locations far from monitoring stations compared to the nonsatellite predictions, which tended to have low values to begin with. Predictions from the satellite model may therefore have a larger range in estimated exposure. If these predictions are more accurate, then the increased exposure contrasts might lead to better inference in epidemiologic studies.[20] Additionally, participants living far from AQS monitors may be characteristically different from those living closer to monitors, meaning that less exposure misclassification may reduce bias in the health effects analysis.

4.3. Kriging versus Satellite Land-Use Regression. We observed that, using conventional cross-validation, given a choice between kriging and land-use regression with satellite data, kriging performed slightly better (kriging land-use regression without satellite data versus nonkriging land-use regression with satellite data: avg. R^2 0.84 versus 0.81). However, with clustered cross-validation, we observed the opposite trend such that satellite data performed better than kriging (avg. R^2 0.72 versus 0.64). The daily spatial resolution of satellite data is relatively coarse, with pixels being approximately 13 × 24 km^2 (at the annual average, ∼12 km resolution). This satellite data cannot provide information on spatial variability at a scale smaller than this. Kriging is comparatively better suited to modeling smaller-scale variability given a dense monitoring network, and it is therefore reasonable that the kriging would out-perform satellite data when predicting to proximal monitors (as in conventional cross-validation). On the other hand, the satellite data provides relatively nationally consistent predictive information, whereas a limitation of the smoothing component of kriging is that it models small-scale variability only in well-monitored regions. These approaches therefore complement each other. The highest model performance, regardless of cross-validation approach, was observed in models with both kriging and satellite data, so we ultimately conclude that a combination of both satellite data and kriging is superior for many applications.

R^2 values calculated using spatially clustered cross-validation were consistently lower compared to those of the conventional cross-validation design. Reduced performance under clustered cross-validation compared to conventional cross-validation was observed not only in kriging models but in land-use regression without kriging. It has been previously shown that land-use regression models have reduced performance in areas on which the model has not been fit, particularly if those areas differ in land-use type or topography,[9] and that prediction error in land-use regression is not uniform even across the range of monitored data.[25] Similarly, the predictive performance of kriging models is also spatially varying[8,26] (with the advantage that kriging can produce standard errors of estimation which model such uncertainty as a function of space). In either case of universal kriging or nonkriging land-use regression, clustered cross-validation represents extrapolation to prediction locations in unmonitored regions, so it is expected that models would demonstrate reduced performance relative to the conventional random cross-validation designs.

4.4. Cross-Validation. The difference in performance between cross-validation designs implies that conventional cross-validation may provide more limited insight into model performance than is typically appreciated. Nonrandom spatial selection of cross-validation groups has been implemented previously,[27] including at least once in the air pollution literature,[28] but relatively little research has explored the sensitivity of the estimation of model performance to validation designs in the context of air pollution modeling. In air pollution applications it has been previously shown that leave-one-out (LOO) cross-validation overestimates performance compared to random k-fold cross-validation designs in the context of land-use regression,[25] and more generally it has been shown that leave-one-out cross-validation may incorrectly estimate performance.[29] Similarly, our results indicate that conventional k-fold cross-validation designs appear to overestimate predictive performance for locations far from the monitors. Under conventional cross-validation design (k-fold or LOO), even though predictions are technically made to "out-of-sample" observations, the spatial averaging and regression coefficients are estimated from nearby observations (with spatially correlated errors), leading to model assessment using unrepresentatively similar observations that may not reflect the amount of extrapolation in prediction to participant locations. While conventional cross-validation may be appropriate for estimating prediction to points that are representative of areas near monitors, clustered cross-validation is more useful for assessing model performance when predicting to unsampled areas far from monitors, which may occur when estimating exposure for some residential locations in epidemiologic cohorts. The limitations of conventional cross-validation may be particularly notable in a spatially clustered monitoring network such as the AQS; in this scenario, due to the mismatch between the distribution of residential locations and the distribution of monitors, random cross-validation may only represent predictive accuracy for a relatively small subset of participants living in areas covered by monitors. Alternate validation techniques such as spatially clustered cross-validation may therefore be necessary in spatial modeling to fully characterize performance.

A limitation of this modeling approach is that it relies on regular observations across the modeling period, and therefore would not be well-suited to model annually a pollutant like ozone which may not have year-round monitoring at some sites. Additionally, any regionalization approach that relies on one-model-per-region carries the disadvantage of spatial discontinuities. While this is a limitation that may cause more variable predictions near the region borders, the approach does provide for model flexibility that has been shown to improve predictions overall. The strengths of our partial-least-squares modeling approach are its regional-flexibility, multiple years of modeling, and good predictive performance. A previous version of this model has been used in epidemiologic analysis[30] and this advanced modeling approach is appropriate for epidemiological analyses in a variety of U.S. cohorts due to its national coverage.

In conclusion, we found that kriging models had superior performance to nonkriging models and that satellite-based models had superior performance to nonsatellite models. Satellite data may be especially useful for predicting in locations

DOI: 10.1021/acs.est.5b05099
Environ. Sci. Technol. 2016, 50, 3686−3694

far from the monitoring network. The spatially clustered validation design in particular showed performance increases with the satellite data. Our findings emphasize the importance of careful consideration of validation methods and spatially varying predictive performance in air pollution modeling.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.est.5b05099.

Additional information on the map figures, geographic covariates, cross-validation groups and additional tables and figures: Figure S1: Partial Least Squares loadings for the PLS components in 1990 and 2006; Figure S2a: Conventional Cross-validation groups (year 2006); Figure S2b: Conventional Cross-validation groups (year 2006); Figure S3a: MEAN Bias by year, estimated using conventional cross-validation; Figure S3b: MEAN Bias by year, estimated using clustered cross-validation Figure S4: Prediction Errors (Clustered Cross-validation); Figure S4: Prediction Errors (Clustered Cross-validation) Figure S5: Modeling Approach; Table S1: Number of NO_2 monitors per year by modeling region; Table S2a: Geographic Covariate types Table S2a: Geographic Covariate types; Table S2b: Land-use categories (2006) ; Table S3: Distance between a test-set monitor and its nearest training-set neighbor Table S4: Regression Coefficients for the Universal Kriging model in 1990 and 2006; Table S5: Kriging Parameters by year (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: myoung3@uw.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Supported by the National Institute of Environmental Health Sciences (NIEHS) F31 Predoctoral Fellowship (1F31ES025096-01), the NIEHS Biostatistics, Epidemiologic, and Bioinformatic Training in Environmental Health Training Grant (T32ES015459), and the Environmental Protection Agency (RD831697, MESA Air).

■ REFERENCES

(1) Hoek, G.; Eeftens, M.; Beelen, R.; Fischer, P.; Brunekreef, B.; Boersma, K. F.; Veefkind, P. Satellite NO2 data improve national land use regression models for ambient NO2 in a small densely populated country. *Atmos. Environ.* **2015**, *105*, 173−180.

(2) Vienneau, D.; de Hoogh, K.; Bechle, M. J.; Beelen, R.; van Donkelaar, A.; Martin, R. V.; Millet, D. B.; Hoek, G.; Marshall, J. D. Western European land use regression incorporating satellite- and ground-based measurements of NO2 and PM10. *Environ. Sci. Technol.* **2013**, *47* (23), 13555−13564.

(3) Knibbs, L. D.; Hewson, M. G.; Bechle, M. J.; Marshall, J. D.; Barnett, A. G. A national satellite-based land-use regression model for air pollution exposure assessment in Australia. *Environ. Res.* **2014**, *135*, 204−211.

(4) Gulliver, J.; de Hoogh, K.; Hansell, A.; Vienneau, D. Development and back-extrapolation of NO2 land use regression models for historic exposure assessment in Great Britain. *Environ. Sci. Technol.* **2013**, *47* (14), 7804−7811.

(5) D. Vienneau, K. de H. Comparison of land-use regression models between Great Britain and the Netherlands. *Atmos. Environ.* **2010**, *44* (5), 688−696.

(6) Beelen, R.; Hoek, G.; Pebesma, E.; Vienneau, D.; de Hoogh, K.; Briggs, D. J. Mapping of background air pollution at a fine spatial scale across the European Union. *Sci. Total Environ.* **2009**, *407* (6), 1852−1867.

(7) Kim, S.-Y.; Sheppard, L.; Kim, H. Health Effects of Long-term Air Pollution: Influence of Exposure Prediction Methods. *Epidemiology* **2009**, *20* (3), 442−450.

(8) Jerrett, M.; Arain, A.; Kanaroglou, P.; Beckerman, B.; Potoglou, D.; Sahsuvaroglu, T.; Morrison, J.; Giovis, C. A review and evaluation of intraurban air pollution exposure models. *J. Exposure Anal. Environ. Epidemiol.* **2005**, *15* (2), 185−204.

(9) Briggs, D. J.; Collins, S.; Elliott, P.; Fischer, P.; Kingham, S.; Lebret, E.; Pryl, K.; van Reeuwijk, H.; Smallbone, K.; Van Der Veen, A. Mapping urban air pollution using GIS: a regression-based approach. *Int. J. Geogr. Inf. Sci.* **1997**, *11* (7), 699−718.

(10) Lebret, E.; Briggs, D.; Van Reeuwijk, H.; Fischer, P.; Smallbone, K.; Harssema, H.; Kriz, B.; Gorynski, P.; Elliott, P. Small area variations in ambient NO 2 concentrations in four European areas. *Atmos. Environ.* **2000**, *34* (2), 177−185.

(11) Marshall, J. D.; Nethery, E.; Brauer, M. Within-urban variability in ambient air pollution: Comparison of estimation methods. *Atmos. Environ.* **2008**, *42* (6), 1359−1369.

(12) Sampson, P. D.; Richards, M.; Szpiro, A. A.; Bergen, S.; Sheppard, L.; Larson, T. V.; Kaufman, J. D. A Regionalized National Universal Kriging Model using Partial Least Squares Regression for Estimating Annual PM2.5 Concentrations in Epidemiology. *Atmos. Environ.* **2013**, *75*, 383.

(13) Li, L.; Wu, J.; Hudda, N.; Sioutas, C.; Fruin, S. A.; Delfino, R. J. Modeling the Concentrations of On-Road Air Pollutants in Southern California. *Environ. Sci. Technol.* **2013**, *47* (16), 9291−9299.

(14) Mercer, L. D.; Szpiro, A. A.; Sheppard, L.; Lindström, J.; Adar, S. D.; Allen, R. W.; Avol, E. L.; Oron, A. P.; Larson, T.; Liu, L.-J. S.; et al. 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.* **2011**, *45* (26), 4412−4420.

(15) Abdi, H. Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2* (1), 97−106.

(16) Novotny, E. V.; Bechle, M. J.; Millet, D. B.; Marshall, J. D. National Satellite-Based Land-Use Regression: NO2 in the United States. *Environ. Sci. Technol.* **2011**, *45* (10), 4407−4414.

(17) Bechle, M. J.; Millet, D. B.; Marshall, J. D. Remote sensing of exposure to NO 2: satellite versus ground-based measurement in a large urban area. *Atmos. Environ.* **2013**, *69*, 345−353.

(18) Bey, I.; Jacob, D. J.; Yantosca, R. M.; Logan, J. A.; Field, B. D.; Fiore, A. M.; Li, Q.; Liu, H. Y.; Mickley, L. J.; Schultz, M. G. *Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation.* **2001**, *106*, 23073.

(19) R. Core Team *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2015.

(20) Szpiro, A. A.; Paciorek, C. J.; Sheppard, L. Does More Accurate Exposure Prediction Necessarily Improve Health Effect Estimates? *Epidemiology* **2011**, *22* (5), 680−685.

(21) Pope, C. A.; Dockery, D. W. Health Effects of Fine Particulate Air Pollution: Lines that Connect. *J. Air Waste Manage. Assoc.* **2006**, *56* (6), 709−742.

(22) Gryparis, A.; Paciorek, C. J.; Zeka, A.; Schwartz, J.; Coull, B. A. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics* **2009**, *10* (2), 258−274.

(23) Lamsal, L. N.; Martin, R. V.; Van Donkelaar, A.; Celarier, E. A.; Bucsela, E. J.; Boersma, K. F.; Dirksen, R.; Luo, C.; Wang, Y. Indirect validation of tropospheric nitrogen dioxide retrieved from the OMI satellite instrument: Insight into the seasonal variation of nitrogen

DOI: 10.1021/acs.est.5b05099
Environ. Sci. Technol. 2016, 50, 3686−3694

oxides at northern midlatitudes. *J. Geophys. Res.* **2010**, *115* (D5).10.1029/2009JD013351
(24) Lamsal, L. N.; Martin, R. V.; van Donkelaar, A.; Steinbacher, M.; Celarier, E. A.; Bucsela, E.; Dunlea, E. J.; Pinto, J. P. Ground-level nitrogen dioxide concentrations inferred from the satellite-borne Ozone Monitoring Instrument. *J. Geophys. Res.* **2008**, *113* (D16), D16308.
(25) Johnson, M.; Isakov, V.; Touma, J. S.; Mukerjee, S.; Özkaynak, H. Evaluation of land-use regression models used to predict air quality concentrations in an urban area. *Atmos. Environ.* **2010**, *44* (30), 3660−3668.
(26) Mulholland, J. A.; Butler, A. J.; Wilkinson, J. G.; Russell, A. G.; Tolbert, P. E. Temporal and spatial distributions of ozone in Atlanta: regulatory and epidemiologic implications. *J. Air Waste Manage. Assoc.* **1998**, *48* (5), 418−426.
(27) Brenning, A. Spatial prediction models for landslide hazards: review, comparison and evaluation. *Nat. Hazards Earth Syst. Sci.* **2005**, *5* (6), 853−862.
(28) Lindström, J.; Szpiro, A. A.; Sampson, P. D.; Oron, A. P.; Richards, M.; Larson, T. V.; Sheppard, L. A Flexible Spatio-Temporal Model for Air Pollution with Spatial and Spatio-Temporal Covariates. *Environ. Ecol. Stat.* **2014**, *21* (3), 411−433.
(29) Bengio, Y.; Grandvalet, Y. No unbiased estimator of the variance of k-fold cross-validation. *J. Mach. Learn. Res.* **2004**, *5*, 1089−1105.
(30) Young, M. T.; Sandler, D. P.; DeRoo, L. A.; Vedal, S.; Kaufman, J. D.; London, S. J. Ambient Air Pollution Exposure and Incident Adult Asthma in a Nationwide Cohort of U.S. Women. *Am. J. Respir. Crit. Care Med.* **2014**, *190* (8), 914−921.

DOI: 10.1021/acs.est.5b05099
Environ. Sci. Technol. 2016, 50, 3686−3694