Supporting Information

National satellite-based land use regression: NO_2 in the United States

Eric V. Novotny, Matthew J. Bechle, Dylan B. Millet, Julian D. Marshall

Number of pages: 8

Number of Figures: 5

Number of Tables: 5

U.S. Census Block NO_2 Data

Three data files provide LUR-derived NO_2 concentration estimates (ppb): one file ("Read me") describes the data, another file ("Preview") illustrates the semicolon-separated format for the database by providing data for the first 100 Census Blocks in the database, and the last file ("NO2_ByCensusBlock"; file size: 810 MB) provides estimates for all Census blocks in the contiguous United States. All files can be downloaded here http://personal.ce.umn.edu/~marshall/data.php

Equations

Equations for mean error (ME), absolute error (AE), mean bias (MB) and absolute bias (AB):

$$ME = \frac{1}{N}\sum_{i=1}^{N}(C_m - C_o) \quad \text{(S1)}$$

$$AE = \frac{1}{N}\sum_{i=1}^{N}|C_m - C_o| \quad \text{(S2)}$$

$$MB = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{C_m - C_o}{C_o}\right) \quad \text{(S3)}$$

$$AB = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|C_m - C_o|}{C_o}\right) \quad \text{(S4)}$$

where C_m is the modeled average concentration for station i, C_o is the average observed concentration for station i, and N is the number of monitoring stations.

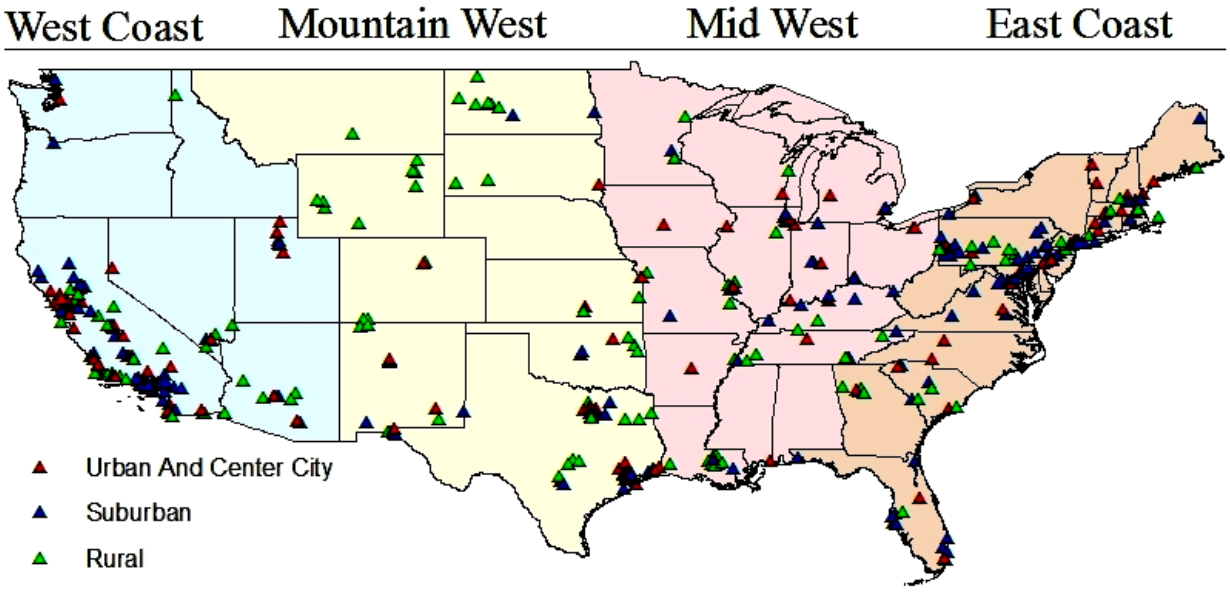

Figure S1. Station locations by region and type.

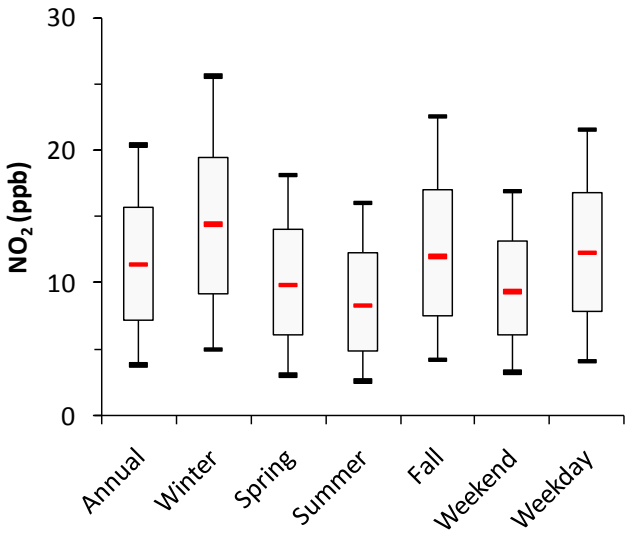

Figure S2. Box plots of year-2006 average NO_2 concentration among the EPA monitors. Inter-quartile ranges are given by the blue boxes; red lines indicate median values. Red lines show median values. Box is the IQR. Black lines outside box are 10^{th} and 90^{th} percentile.

Table S1. Stepwise multiple linear regression analysis for US dataset with OMI NO_2, global dataset without OMI NO_2 and US dataset without OMI NO_2. Parameters are listed in the order in which they were added to the model.

Parameter	Unit	β	Std. Err.	$p>\|t\|$	R^2	IQR	β * IQR	VIF
US Dataset with OMI NO_2								
Intercept	--	3.94	0.47	<0.01				
Impervious (7000m)	%	0.12	0.01	<0.01	0.58	31.2	3.74	2.5
OMI NO_2	ppb	0.92	0.07	<0.01	0.70	3.3	3.04	1.6
Tree canopy (600m)	%	-0.47	0.01	<0.01	0.72	15.1	-7.10	1.2
Major roads (700m)	km	0.30	0.07	<0.01	0.74	2.60	0.78	1.4
Impervious (100m)	%	0.03	8.83E-03	<0.01	0.75	53.7	0.71	2.4
Elevation	km	2.36	0.47	<0.01	0.76	0.27	0.64	1.5
Distance to coast	km	-1.17E-03	3.95E-04	<0.01	0.77	620	-0.73	1.5
Minor roads (100m)	km	2.53	1.14	0.03	0.77	0.27	0.68	1.3
Global Dataset without OMI NO_2								
Intercept		7.2	0.54	<0.01				
Impervious (6000m)	%	0.12	0.02	<0.01	0.55	35.1	4.21	3.8
Major roads (800m)	km	0.23	0.07	<0.01	0.58	3.20	0.74	1.4
Population (10000m)	#	7.54E-04	1.69E-04	<0.01	0.61	1100	0.83	1.8
Tree canopy (1800m)	%	-0.09	0.02	<0.01	0.63	11.1	-1.00	1.2
Distance to coast	km	-2.2E-03	7.73E-04	<0.01	0.64	620	-1.36	1.5
Elevation	km	1.81	0.57	<0.01	0.65	0.27	0.49	1.5
Major roads (10000m)	km	6.16E-03	2.17E-03	<0.01	0.66	270	1.66	4.3
US dataset without OMI NO_2								
Intercept		5.70	0.51	<0.01				
Impervious (7000m)	%	0.13	0.02	<0.01	0.58	31.2	4.06	4.7
Population (700m)	#	4.26E-04	1.15E-04	<0.01	0.61	2000	0.85	1.7
Major roads (300m)	km	0.76	0.27	<0.01	0.63	0.53	0.40	1.2
Tree canopy (500m)	%	-0.04	0.02	<0.01	0.64	14.7	-0.59	1.2
Distance to coast	km	-1.78E-03	4.68E-04	<0.01	0.65	620	-1.10	1.6
Elevation	km	2.24	0.55	<0.01	0.66	0.27	0.60	1.5
Impervious (100m)	%	0.03	9.87E-03	<0.01	0.67	53.7	1.61	2.2
Major roads (10000m)	km	7.35E-03	1.99E-03	<0.01	0.68	270	1.98	3.8

Distance in () is the buffer radius, parameters without a buffer distance were taken at the station locations. IQR is the inter-quartile range for the given parameter, β * IQR is the β coefficient multiplied by the IQR, and VIF is the variance inflation factor to check for multicollinearity.

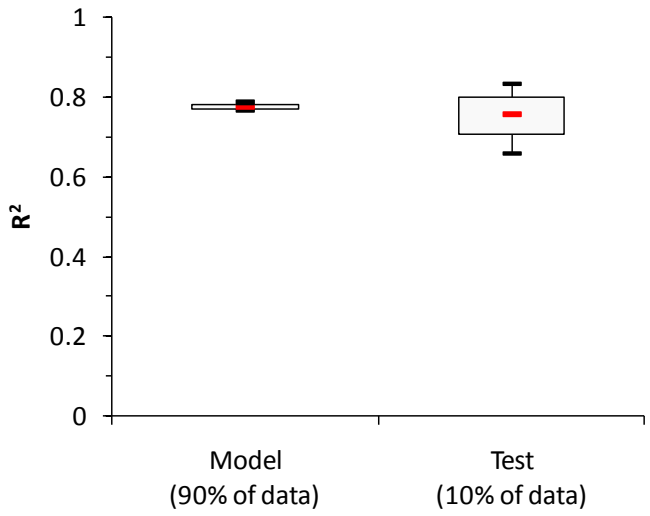

Figure S3. Box plot showing R^2 values between observed and modeled data, for the 90% of data used to create the model (model-building data) and for the remaining 10% (model-testing data) for 500 Monte Carlo simulations. Red lines show median values. Box is the IQR. Black lines outside box are 10^{th} and 90^{th} percentile.

Table S2. Error and bias between the measured values and the model-building and model-testing datasets for the 500 Monte Carlo simulations.

	Model-building	**Model-testing**
Mean error (ppb)	0	0.08
Mean absolute error (ppb)	2.4	2.55
Mean bias (%)	23	25
Mean absolute bias (%)	40	42

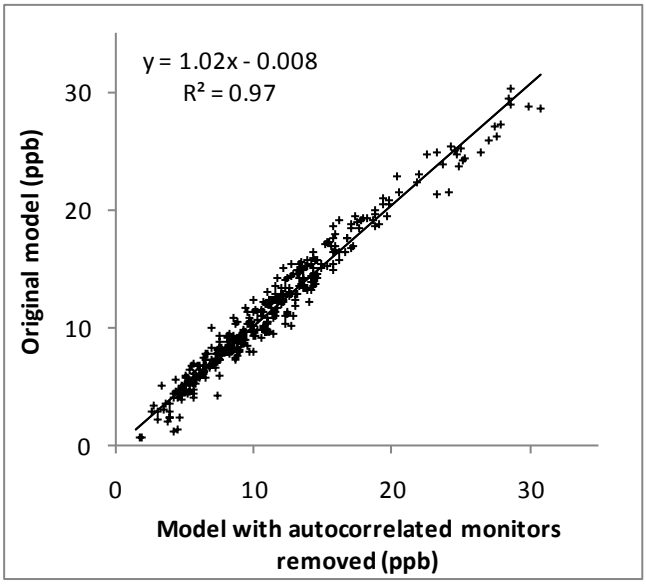

Figure S4. Comparison between the core model (Table 3 in main text) and the autocorrelation-corrected model (same as the core model, but omits 66 stations where the residuals of the models have a statistically significant spatial autocorrelation at the 95% level). We tested spatial autocorrelation of the model residuals by calculating Moran's I using ArcGIS, more information on this topic can be found here:

http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Cluster_and_Outlier_Analysis:_Anselin_Local_Moran%27s_I_(Spatial_Statistics)

Table S3: Range of values for station parameters and independent variables.

	Units	Min	Max	Median (IQR)
Station Parameters				
Distance to major road	km	0.001	40.9	0.44 (0.18 - 1.04)
Annual measured NO_2	ppb	0.3	34.2	11.4 (7.2 - 15.6)
Latitude		25.73	48.64	37.14 (33.55 - 40.61)
Longitude		-124.18	-68.03	-95.08 (-115.34 to -81.16)
Independent Variables				
Impervious (6000m)	%	0	74	22.7 (5.8 - 40.9)
OMI NO_2	ppb	0.2	17.5	2.9 (1.5 - 4.8)
Tree canopy (1000m)	%	0	77	5.4 (2.4 - 10.8)
Major roads (800m)	km	0	22.4	1.56 (0 - 3.18)
Minor roads (100m)	km	0	.77	0.16 (0 - 0.27)
Elevation	km	0	2.36	0.15 (0.03 - 0.30)
Distance to coast	km	0	2,100	156 (29.1 - 651)
Major roads (200m)	km	0	2.84	0 (0 - 0.19)

Table S4. Model results for seasonal regression analysis

	R^2	Adj. R^2	N	SSE	SSR	DFR	F	*p*
Fall	0.74	0.73	358	4966	14030	6	169	<0.001
Spring	0.74	0.73	366	3737	10578	8	129	<0.001
Summer	0.76	0.75	385	3473	10736	5	238	<0.001
Winter	0.76	0.76	345	5047	16206	8	138	<0.001
Weekday	0.78	0.77	361	4010	13973	7	180	<0.001
Weekend	0.75	0.74	363	2856	8373	7	152	<0.001

N is the number of stations used in the analysis, SSE is the sum of squared error, SSR is the sum of squared residuals, DFR is the degrees of freedom, F is the F ratio and P is the significance level of the F ratio.

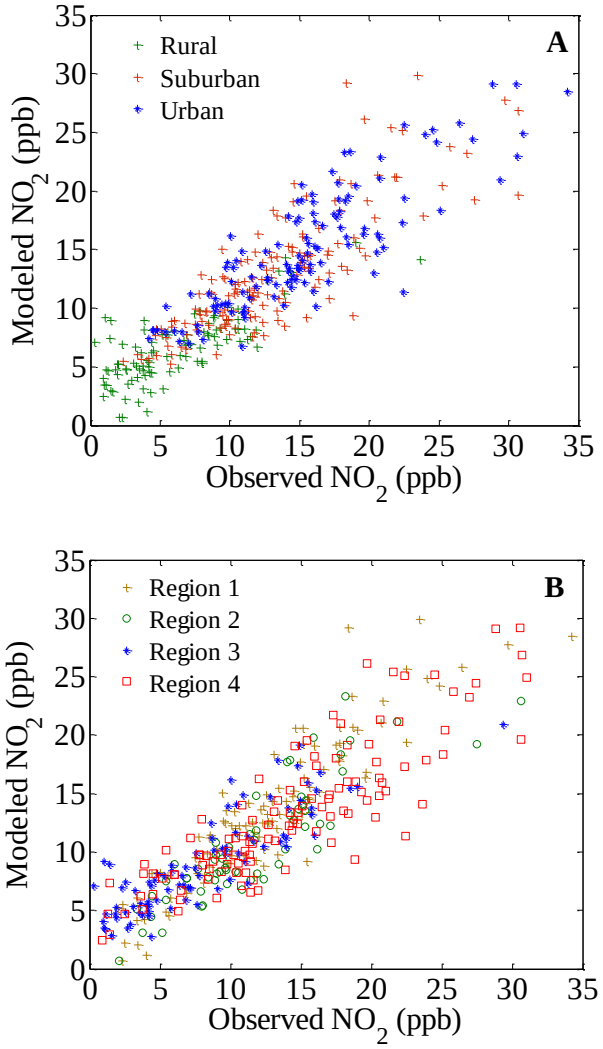

Figure S5. Modeled vs observed results for the annual average model with satellite measurements. Panel A shows the values divided into rural, urban and suburban categories and panel B is divided by regions (Figure S1).

Table S5. Stepwise multiple linear regression for urban, suburban and rural areas. Parameters are listed in the order in which they were added to the model.

Parameter	Unit	β	std. err.	$p>\|t\|$	Partial R^2	IQR	β * IQR	VIF
Model: Urban								
Intercept	--	5.75	1.15	<0.01				
Annual OMI NO_2	ppb	1.31	0.09	<0.01	0.57	3.8	4.98	1.3
Impervious (1800m)	%	0.11	0.02	<0.01	0.68	23.8	2.62	2.4
Elevation	km	3.75	0.77	<0.01	0.74	0.22	0.83	1.2
Major roads (800m)	km	0.19	6.84E-02	0.01	0.76	3.56	0.68	1.2
Tree canopy (6000m)	%	-0.10	0.03	<0.01	0.77	8.77	-0.88	1.3
Minor roads (3000m)	km	-1.59E-02	5.07E-03	<0.01	0.79	80.0	-1.27	2.0
Minor roads (100m)	km	3.94	1.52	0.01	0.80	0.26	1.02	1.1
Model: Suburban								
Intercept		5.55	0.80	<0.01				
Annual OMI NO_2	ppb	0.82	0.09	<0.01	0.49	3.37	2.76	1.4
Impervious (800m)	%	0.05	0.02	0.01	0.60	26.8	1.34	1.7
Major roads (200m)	km	3.50	0.81	<0.01	0.65	0.17	0.60	1.1
Tree canopy (8000m)	%	-0.07	0.02	<0.01	0.68	16.0	-1.12	1.1
Elevation	km	2.96	0.91	<0.01	0.70	0.25	0.74	1.1
Major roads (10000m)	km	7.97E-03	2.36E-03	<0.01	0.72	221	1.76	1.8
Model: Rural								
Intercept		3.17	0.46	<0.01				
Impervious (200m)	%	0.10	0.04	0.03	0.45	8.00	0.80	1.7
Annual OMI NO_2	ppb	1.01	0.22	<0.01	0.52	1.61	1.63	1.9
Canopy (1000m)	%	-0.06	0.02	<0.01	0.60	22.2	-1.33	1.2
Major roads (400m)	km	1.76	0.45	<0.01	0.64	0.14	0.25	1.1
Population (10000m)	#	4.80	1.45	<0.01	0.68	0.13	0.62	2.1

Distance in () is the buffer radius; parameters without a buffer distance were taken at the station locations. IQR is the inter-quartile range for the given parameter, β * IQR is the β coefficient multiplied by the IQR. VIF is the variance inflation factor to check for multicollinearity.