Supplemental Materials

Estimating Spatiotemporal Variability of Ambient Air Pollutant Concentrations with A Hierarchical Model

Lianfa Li[1, 2], Jun Wu[1§], Jo Kay Ghosh[3], Beate Ritz[4]

[1] Program in Public Health, College of Health Sciences, University of California, Irvine, USA

[2] State Key Lab of Resources and Environmental Information Systems, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, China

[3] Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, USA

[4] Department of Epidemiology, School of Public Health, University of California, Los Angeles, USA

* Corresponding author: Program in Public Health & Department of Epidemiology, Anteater Instruction & Research Bldg (AIRB) # 2034, University of California, Irvine CA 92697-3957. Tel: 949-824-0548, Fax: 949-824-0529, email: junwu@uci.edu

Contents

1. Overview

This paper introduces a hierarchical model to predict spatiotemporal variability of nitrogen dioxide (NO_2) and nitrogen oxides (NO_x) concentrations in the urban area of Southern California by combining high temporal resolution data from routine monitoring stations with high spatial resolution data from investigator-initiated episodic measurements. Our approach had an improvement for estimation of spatiotemporal variability of NO_2 and NO_x concentrations in the Los Angeles region and this has meaningful indications for studies of short-term health effects.

2. Materials and Methods

A hierarchical two-stage model was designed to estimate the spatiotemporal variablity of NO_2 and NO_x concentrations.

2.1. Episodic measurements from two field campaigns

1) Measurements Collected by University of California, Los Angeles (UCLA): NO_2 and NO_x samples were collected using passive Ogawa samplers (Ogawa & Company USA, Inc., Pompano Beach, FL) in two continuous weeks in a warm season (September 16-October 1, 2006) and a cold season (February 10-25, 2007) in Los Angeles County, California. Each sampler was deployed for a two-week period. The sampling locations were selected using a location-allocation algorithm that maximized the potential variability in measured pollutant concentrations and the spatial distribution of the targeted study population (Su et al., 2009). There were a total of 161 valid samples of measurements in each season. We also conducted additional measurements co-located at 14 SCAQMD stations.

2) Measurements Collected by University of California, Irvine (UCI): Residential outdoor samples of NO_2 and NO_x were collected using passive Ogawa samplers (Ogawa &

Company USA, Inc., Pompano Beach, FL) for two weeks in the warm season (July 10-18 and July 24-August 1) and the cool season (November 13-21 and December 4-12) in 2009 in south Los Angeles County and Orange County, California. Sampling sites were outdoor homes of subjects who participated in an air pollution and pregnancy outcome study funded by the National Institute of Environmental Health Sciences; i.e. participants who agreed to allow us to conduct outdoor sampling at their homes. There were a total of 32 valid measurements in each sampling week. We again co-located sampling at 11 SCAQMD stations. Sampling sites of both the UCLA and UCI episodic measurements are not shown in Figure 1 to protect the confidentiality of human subjects.

Systematic bias is possible for the measurements taken by the active samplers at the SCAQMD sites and the passive samplers used in the UCLA and UCI field campaigns. We adjusted for such potential systematic bias by converting all measurements from passive samplers to the equivalent values of the active samplers based on the co-located measurements from the SCAQMD sites (Supplemental Materials Table S1 gives specific adjustment coefficients by linear regression for both the UCLA and the UCI measurements).

2.2. Roadway classification

We obtained roadway data for the study region from the ESRI StreetMap™ North America 9.3 (http://www.esri.com). This dataset included 2003 TeleAtlas® street polylines, which - as we previously demonstrated – are more accurate than TIGER 2000-based streets (Wu et al., 2005). We calculated total roadway length within different buffer sizes around each sampling site and classified roadways into four categories based on the U.S. Census Feature Class Code (U.S. Census Bureau, 1993): primary highways, typically interstates, with limited access (A1); primary roads without limited access, non-interstate roads (A2); smaller, secondary or connecting roads, usually with more than two lanes (A3); and local, neighborhood and rural roads, usually with a single lane of traffic in each direction (A4). We calculated the shortest distance from sampling sites to each roadway type.

2.3. Decomposition of independent temporal basis functions

Theoretically, the temporal basis functions represent the general temporal changes in different dimensions of concentrations within the study domain. Average weekly concentrations (N=5225= 209 weeks ×25 stations) were first log-transformed and normalized (mean of 0 and variance of 1), and then used to construct the independent temporal basis functions (smoothed EOFs). As a technique of principle component analysis, singular value decomposition (SVD) was used to generate the independent temporal basis functions. The SVD decomposition was performed with:

$$\boldsymbol{Y} = \boldsymbol{U\Sigma V}^* \quad [\text{A1}]$$

where $\boldsymbol{Y} = (\vec{\boldsymbol{y}}(1), \dots, \vec{\boldsymbol{y}}(n))$ and $\vec{\boldsymbol{y}}(u) = (y_{u1}, \dots, y_{um})^T$, the *m*x*n* normalized matrix of the logarithmic transformation of routine observed concentrations, *n* is the number of sites and *m* is the total number of time slices. $\boldsymbol{U} = (\vec{\boldsymbol{u}}(1), \dots, \vec{\boldsymbol{u}}(m))$ and $\vec{\boldsymbol{u}}(k) = (u_{k1}, \dots, y_{km})^T$, *m*x*m* real matrix whose columns represents temporal basis series (called left singular vectors), Σ is an *m*x*n* diagonal matrix with nonnegative real numbers (called singular values) on the diagonal representing the variance explained by each temporal basis series, *V* is an *n*×*n* real matrix of right singular vectors (*V** is the conjugate transpose of *V*) (Marcus and Minc, 1968). We used penalized thin plate splines of $\vec{\boldsymbol{u}}(i)$ in GAM (Duchon, 1977) to model smoothed EOF, $f_i(t)$ in equation [2] of this paper. The derivative based thin plate spline penalty was used to measure wiggliness and chose a good degree of freedom between data fitting and smoothness. A good degree of freedom should make the smoothed curves not over-fitting the data but approximate the truly piece of temporal trend (Szpiro et al., 2010). So its selection, while optimizing the fit, should penalize wiggliness. We used Wood and

Augustin (2002)'s integrated approach of model selection and automatic smoothing parameter selection with generalized cross-validation (GCV) used to determine the smoothness parameters..

2.4. Procedure of interpolation for UCLA measurements

We used the 25 routine measurements and the linear spatiotemporal model to derive the ratios of two continuous weekly concentrations for UCLA's measurements respectively for 2006 and 2007 (09/16/06-10/01/06 and02/10/07-02/25/07). The steps are described as followed.

(1) Similar to Section 3.1 of the paper, the 25 routine measurements of time series were used to construct the temporal basis functions that were smoothed using GAM to represent the temporal trends of the temporal basis functions for the study domain;

(2) 1 or 2 traffic-related covariates as emission sources or proxy to emission sources and wind speeds as a dispersion factor were extracted for the locations of the routine measurements. Due to few samples (just 25 samples from the SCAQMD routine stations), we just used 2 or 3 covariates to avoid over-fitting in linear regression. For emission sources, distance-weighted roadway length, AADT or traffic land-use statistics was calculated around the measurement sites using the optimal buffering distance (see Section 2.3.1 of the paper for determination of buffering distance). Since we just had few samples (25), only a covariate of emission factor with the highest Pearson's

correlation and p-value<0.1 was used with long-term average wind speed in the linear model.

(3) A simplified maximum likelihood method without incorporation of spatial autocorrelation was used to estimate the spatially varying coefficients for the temporal basis functions:

$$\hat{\Psi} = \mathrm{argmax}_{\Psi} p(Y_{ut};\ \Psi) \quad \text{[A2]}$$

where $Y_{ut}=\{y_{ut}\}$ is set of the observations from the log-transformed measured values of concentrations at location u and time slice t, Ψ is the coefficients to be estimated, mainly the linear coefficients of spatial covariates to calculate $\beta_{i.}$; $p(Y_{ut};\ \Psi)$ is the density for Y_{ut}. Due to too few spatial samples (25) and independence of temporal basis functions, we did not consider temporal and spatial dependence in the model different from Szpiro et al.'s method (2010) for solving β_{iu}, in [A1]. Thus the likelihood function was simplified as:

$$\mathrm{Ln}L = -(n/2)\ln(2\pi) - (n/2)\ln(\sigma^2) - (\sigma^2/2)\sum(Y_{ut} - \mu_Y(\Psi)) \quad \text{[A3]}$$

in [A3], $\sum_Y(\Psi)$ was simplified as $(\sigma^2)^n$ without consideration of spatial autocorrelation.

(4) Based on the linear coefficients solved above ($\hat{\Psi}$), we could get an initial estimates for mean concentrations at the UCLA sampling locations, u^* at particular time slices, t=$\{t_{11},t_{12},t_{21},t_{22}\}$, assuming t_{11} and t_{12} for 09/16/06-09/23/06 and 09/24/06-10/01/06 as well as t_{21} and t_{22} for 02/10/07-02/17/07 and 02/18/07-02/25/07.

$$\hat{Y}^*_{u^*t} = \mathrm{E}(\hat{Y}^*_{u^*t} | Y_{ut}; \Psi = \hat{\Psi}) \qquad [A3]$$

where $\hat{Y}^*_{u^*t}$ is the concentration to be interpolated for UCLA's measurement site, u^*.

(5) Based on [A3], we got the estimates for four time slices and then estimated their ratio along with two continuous weeks.

$$r_1 = \hat{Y}^*_{u^*t_{11}} / \hat{Y}^*_{u^*t_{12}} \quad \text{or} \quad r_2 = \hat{Y}^*_{u^*t_{21}} / \hat{Y}^*_{u^*t_{22}} \qquad [A4]$$

(6) According to the ratios and the observed bi-weeks means, we directly estimated the values at each of the four time slices (t_{11},t_{12},t_{21},t_{22}) for UCLA's episodic sites.

2.5. Selection of covariates

There are two steps for selection of the effective covariates using GAM to correlate spatial covariates to temporal basis functions:

(1) First, correlation analysis and scatter plots were made to remove the irrelevant covariates whose correlation with β_i was less than 0.1 and the scatter plots did not suggest a clear pattern. The factors selected were regarded as an initial pool of regressors for next selection.

(2) Then, multicollinearity of independent covariates and their statistical significance were examined. To avoid multicollinearity, we used variance inflation factors (VIFs) to divide the covariates into several groups: a) one group of weakly correlated covariates (VIF<10); the following 2 type of groups of remaining highly correlated covariates (VIF⩾10): b) traffic-related groups including shortest distances to different types of roadways (A1, A2, A3, or A4), distance-weighted roadway length, traffic land-use, weighted AADT; c) land-use group including different types of land-use.

(3) Next, backward-selection was iteratively conducted until the optimal set of covariates selected. Specifically, we selected one covariate at a time from each group of the highly correlated covariates and combined them with all of the weakly correlated covariates to construct a combination of covariates for predicting β_i. All of the covariates were tested in the model. R^2 was used to backward-select the covariates in each combination: the covariates with p values $\geqslant 0.1$ were removed until R^2 remained the same, improved, or decreased least when all possible combinations of the remaining covariates were considered. Finally, the covariate combination with the maximum R^2 was selected as optimal regressors.

Table S1. Correlation between average biweekly measured values of collocated monitoring sites of UCI/UCLA and SCAQMD and their linear regression coefficients

Source	Pollutant	Number of collocated locations	Correlation coefficient	Parameters	
				Slope	Intercepts
UCI	NO_2	14	0.98	0.88	5.20
			0.99	0.94	3.56
			0.996	0.58	12.91
			0.95	0.65	7.74
	NO_x	14	0.96	0.69	8.53
			0.95	0.69	5.46
			0.96	1.22	-13.41
			0.97	0.72	14.38
UCLA	NO_2	11	0.94	0.68	4.43
			0.95	1.00	0.29
	NO_x	11	0.98	0.80	2.38
			0.97	0.81	12.05

Table S2. Important notation and symbols

Symbol	Meaning
y_{ut}	Log-transformed concentrations at time slice, t and location, u; $\hat{y}_{ut}$ is the estimated value for y_{ut} by the model.
μ_{ut}	Mean trend value at time slice, t and location u. It represents the seasonal trend. $\hat{\mu}_{ut}$ is the estimate by model.
ε_{ut}	Spatiotemporal residual at time slice, t and location, u. $\hat{\varepsilon}_{ut}$ is the estimate.
β_{iu}	i^{th} spatially varying coefficient for the i^{th} temporal basis function. $\hat{\beta}_{iu}$ is the estimate.
$\beta_{i.}$	Set of the i^{th} spatially varying coefficient across all the locations.
$f_i(t)$	i^{th} temporal basis function. $f_0(t)$ is the constant function.
$\varepsilon_{u.}$	Set of spatiotemporal residual at u across all the time slices.
$\hat{\hat{\beta}}_{iu}(X)$	Estimate of the mean for β_{iu} modeled using set of spatial covariates (X)
$\hat{\varepsilon}_{ius}(Z)$	Estimate of the spatial residual at u for $\hat{\beta}_{iu}$, $Z \in Nb(u)$
$\hat{\varepsilon}_{iun}$	Random residual at u, with normal distribution, $\hat{\varepsilon}_{iun} \sim N(0,1)$
$E(\hat{\beta}_{iu})$	Same as $\hat{\hat{\beta}}_{iu}(X)$
$g\left(E(\hat{\beta}_{iu})\right)$	Link function for normal distribution between regression equation and $E\left(\hat{\beta}_{iu}\right)$.
$s_w(w\_s_u, w\_c_u)$	Smooth function for wind speeds of both directions at u. w_s_u indicates wind speed along with south-north, w_c_u indicates along with west-east.
$s_j(x_u^j)$	Smooth functions for other local covariates.
γ_k	Linear coefficient for the k^{th} linear regressor, $x_{.}^{k}$.
$\Theta(\phi, \sigma^2, \tau)$	Variogram parameters (range ϕ, partial sill σ^2 and coefficient τ)
$\hat{\varepsilon}_{us}(u_i)$	Estimate of the spatial residual at the neighboring location, u_i,
$\hat{\varepsilon}_{rs}(u_i)$	Estimate of the regional residual or total variation of $\beta_{i.}$.
$\sum_{ij}$	Variogram output of matrix.
$\lambda_{u_i}^{u}$	Optimal weights for $\hat{\varepsilon}_{us}$, estimated by cokriging.
$\lambda_{u_i}^{r}$	Optimal weights for $\hat{\varepsilon}_{rs}$, estimated by cokriging.

Note: For the covariate defined above, if a cap sign like ˆ is added on the top of a covariate, it indicates the estimated or predicted value for this covariate.

3. Results

Table S3. Statistics of fit parameters by the routine time series and sporadic samples

Statistics	NO_2			NO_x		
	β_0	β_1	β_2	β_0	β_1	β_2
Min	0.59	-3.05	-15.03	0.69	-16.29	-12.03
Max	3.46	15.60	12.5	4.32	2.05	15.17
Mean	2.90	5.72	0.06	3.51	-9.22	0.80
Variance	0.13	11.11	19.14	0.20	8.3	17.96
Interquartile range (IQR)	0.37	2.15	2.77	0.42	2.88	5.89
Median	3.01	5.20	-0.26	3.62	-9.27	1.00

Table S4. Variogram coefficients of Spatial Residuals for β_0, β_1 and β_2

Statistics	NO_2			NO_x		
	β_0	β_1	β_2	β_0	β_1	β_2
Model	Stable	Exp*	Stable	Stable	Exp*	Stable #
Model parameter	0.82	-	0.2	0.2	-	2
Range	7.4 km	0.9 km	2.2 km	5.5 km	1.0 km	2.5km
Partial sill 1***	0.013	4.5	3.8	0.016	1.2	2.4
Partial sill 2	0.009	2.8	1.7	0.02	1.1	1.7
Partial sill 3	0.04	4.4	3.2	0.05	1.8	2.4

Note: *: Exp: exponential variaogram model. #: Stable model (Johnston et al., 2003). **: partial sill 1 is for local residuals; partial sill 2 is for covariance between local residuals and regional residuals; partial sill 3 for regional residuals. Nugget was assumed to be 0 given the strong spatial autocorrelation of the residuals (Szpiro et al., 2010).

Table S5. Sensitivity analysis for interpolation of UCLA samples

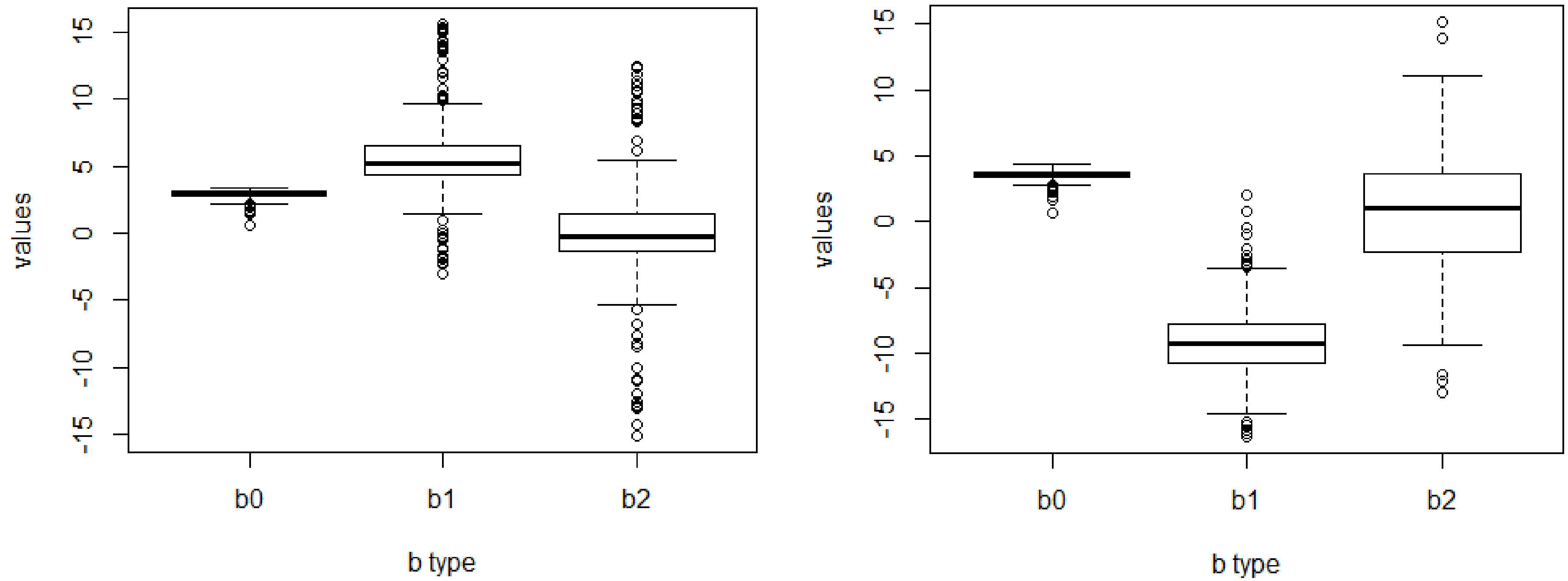
Description	Type	R^2 for time series(1)	R^2 for long-time averages(2)
Method 1(3)	**NO_2**	**0.84**	**0.89**
	NO_x	**0.81**	**0.77**
Method 2	NO_2	0.82	0.87
	NO_x	0.63	0.75
Method 3	NO_2	0.83	0.87
	NO_x	0.86	0.72

Note: Note: (1) R-Square between the predicted values of all the time series for the routine stations and the temporal trends based on their observed values; (2) R-Square between the averages of the predicted values over 4 years for the 25 routine stations and averages of their observed values over 4 years; (3) Results from the original paper.

Table S6. Sensitivity analysis for outliers of episodic samples

Description	Type	Thresholds	Number of samples[1]	R^2 for time series	R^2 for long-time averages
1. No thresholds set to remove outliers	NO_2	No outliers removed	32+161+31=224	0.74	0.84
	NO_x	No outliers removed	32+161+31=224	0.74	0.67
2. Lower and upper inner fences [2]	NO_2	β_0: [-0.2,3.7], β_1: [-2.3,6.8], β_2: [-12.2,4.5]	14+154+31=199	0.88	0.89
	NO_x	β_0: [1.0,4.7], β_1: [-15.8,1.8], β_2: [-8.1,9.4]	13+160+30=203	0.87	0.70
3. Lower and upper outer fences [3]	NO_2	β_0: [-0.2,4.2], β_1: [-2.4,12.3], β_2: [-12.3,8.4]	28+157+31=216	0.84	0.89
	NO_x	β_0: [0.7,5.0], β_1: [-17.7,3.7], β_2: [-9.7,11]	26+161+31=218	0.81	0.77

Note: 1. number of UCI samples+ number of UCLA samples+ number of SCAQMD sites=total number of samples; 2. Lower and upper inner fences defined as Q1-1.5*IQR and Q3+1.5*IQR where Q1 and Q3 are respectively the first and third quartiles, and IQR is inter-quartile range. If a sample's β_i is smaller than lower upper inner fence or bigger than upper inner fence, the sample will be removed from the dataset; 3. Similarly, lower and upper outer fences defined as Q1-3*IQR and Q3+3*IQR. This was what we have used for the main results.



a. NO_2 b. NO_x

Figure S1. Box plots of spatially varying coefficients for NO_2 and NO_x (b_0-β_0; b_1-β_1; b_2=β_2)

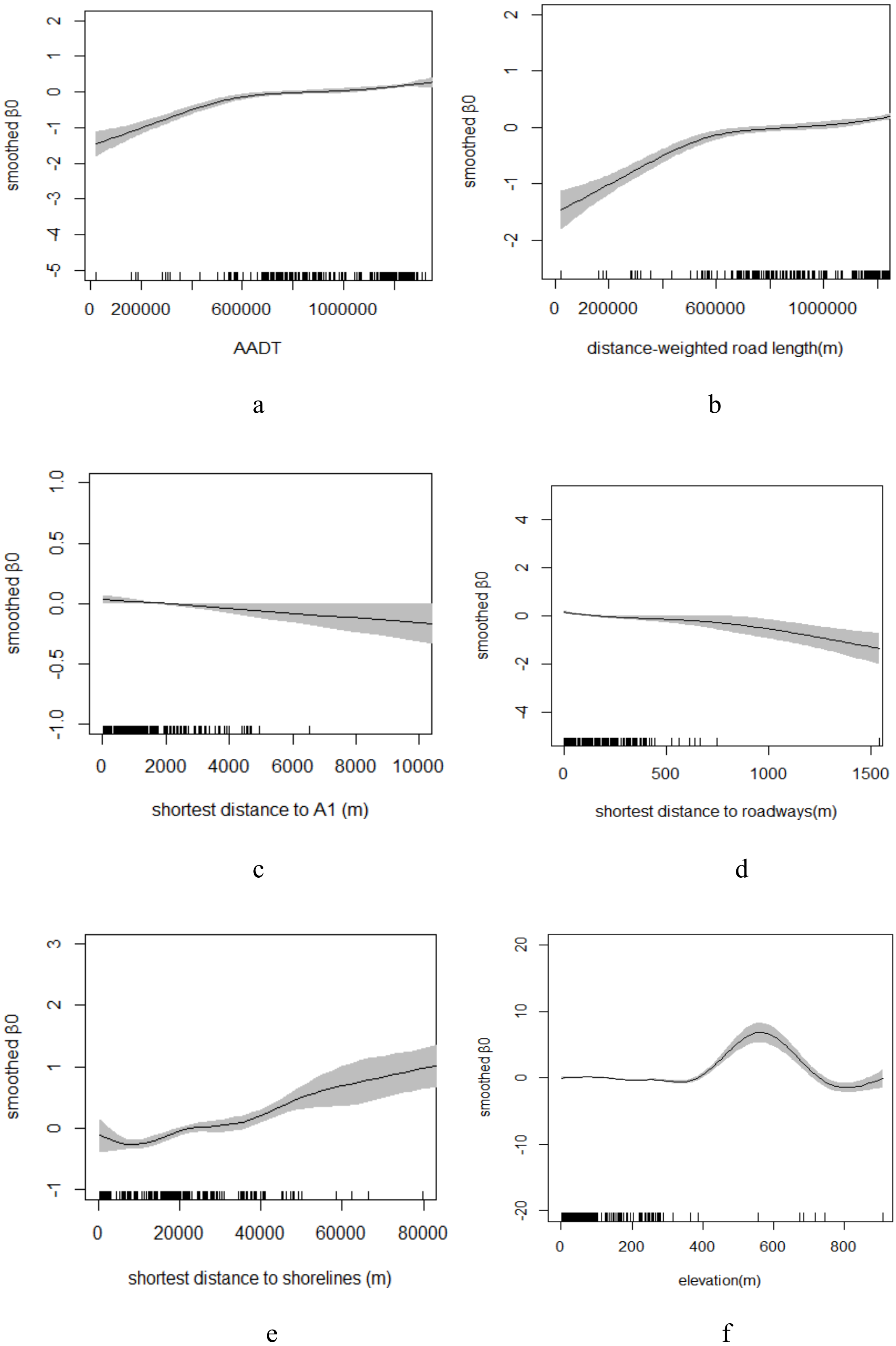
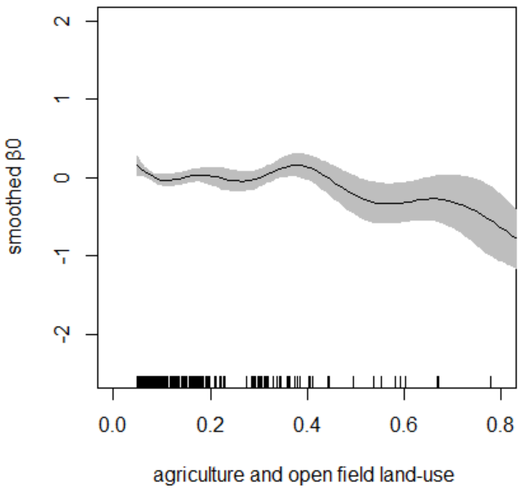
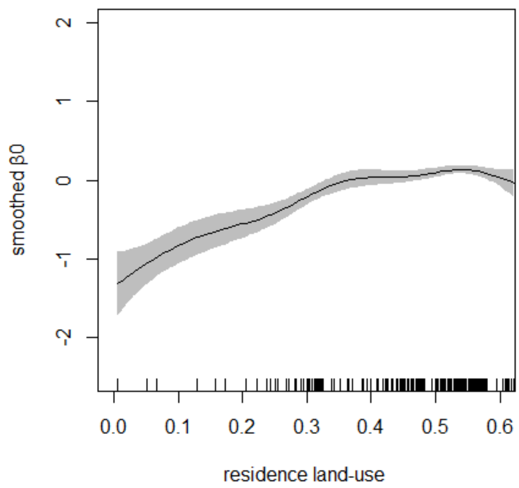
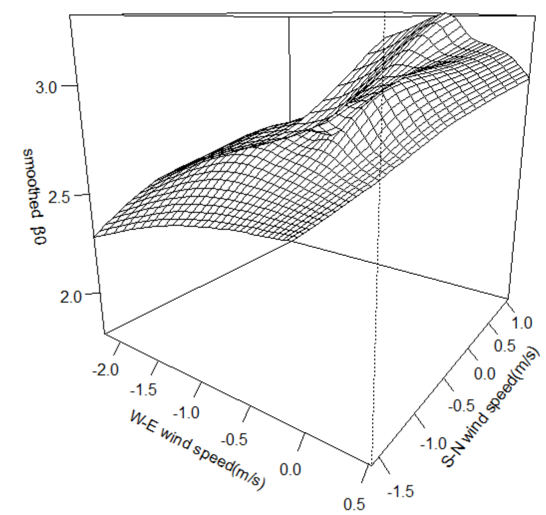


Figure S2. Non-linear relationship between β_0 and local covariates with the 95% confidence intervals for NO_2 by GAM (to be continued)



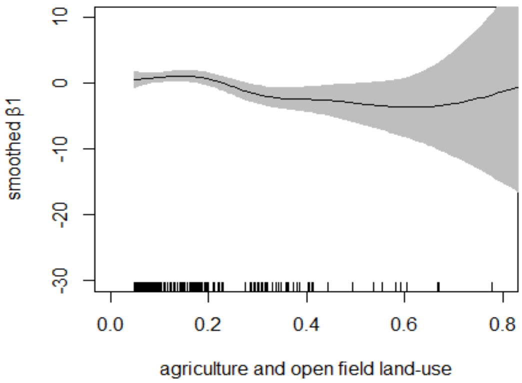


g h

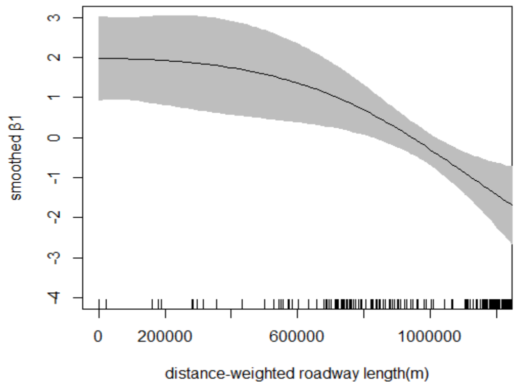


i (wind speeds in west-east (W-E) and south-north (S-N))

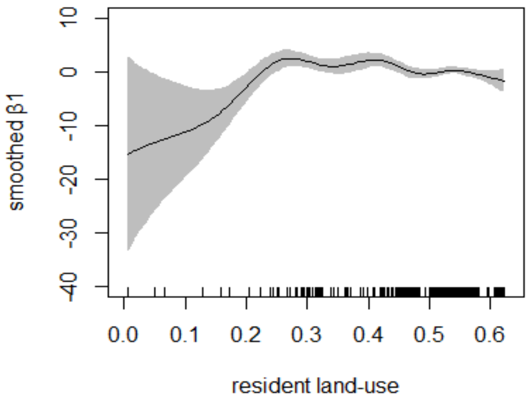
Figure S2. Non-linear relationship between β_0 and local covariates with the 95% confidence intervals for NO_2 by GAM (continued)



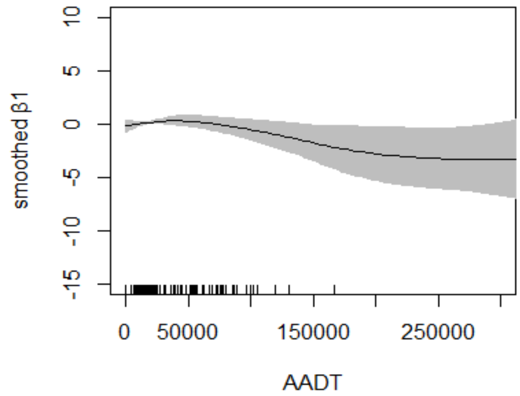
a



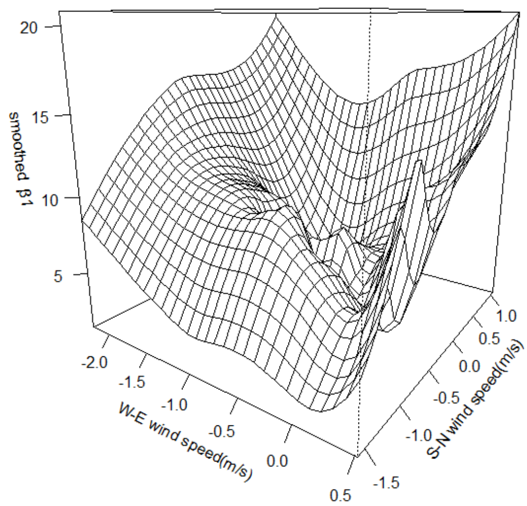
b



c

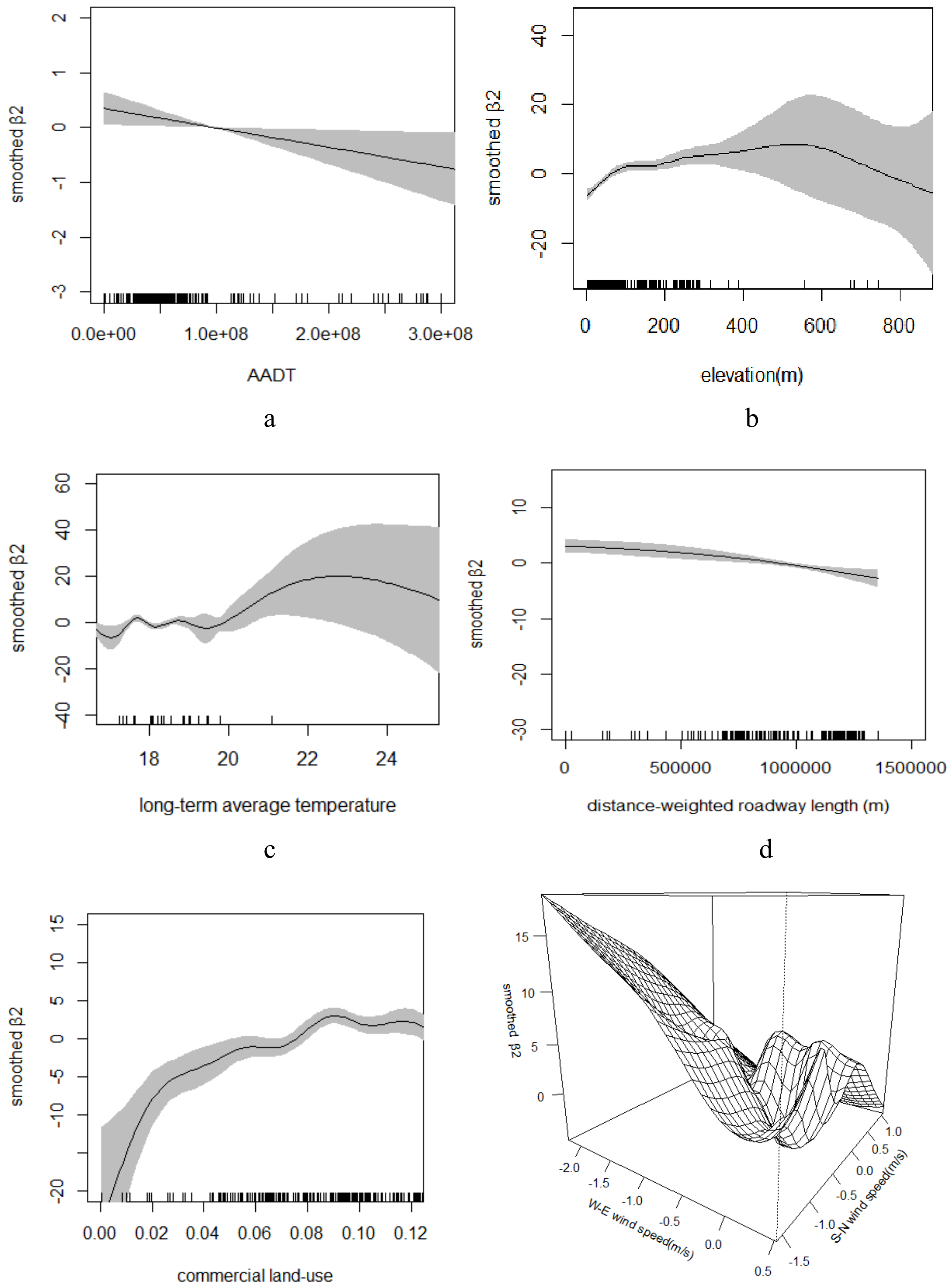


d



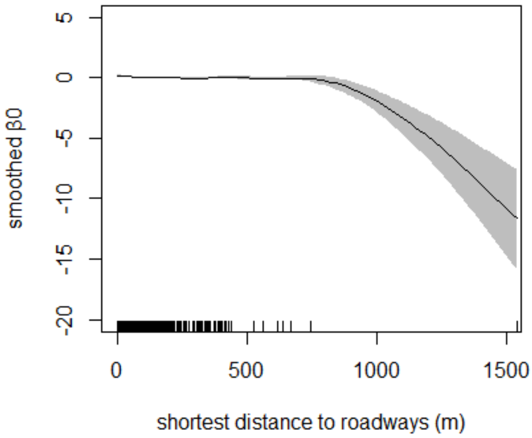
e (wind speeds in W-E and S-N)

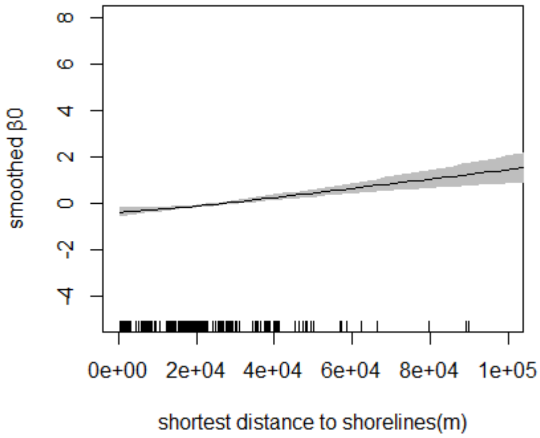
Figure S3. Non-linear relationship between β_1 and local covariates with the 95% confidence intervals for NO_2 by GAM



e f (wind speeds in W-E and S-N)

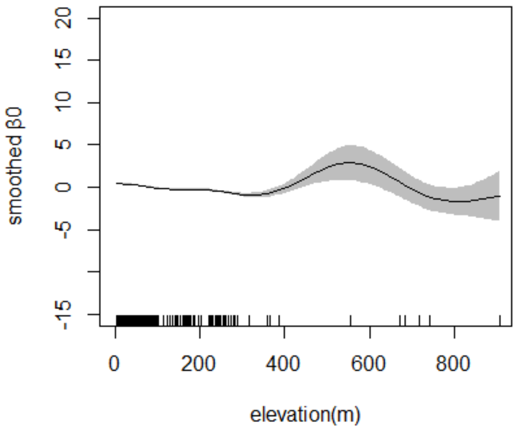
Figure S4. Non-linear relationship between β_2 and local covariates with the 95% confidence intervals for NO_2 by GAM

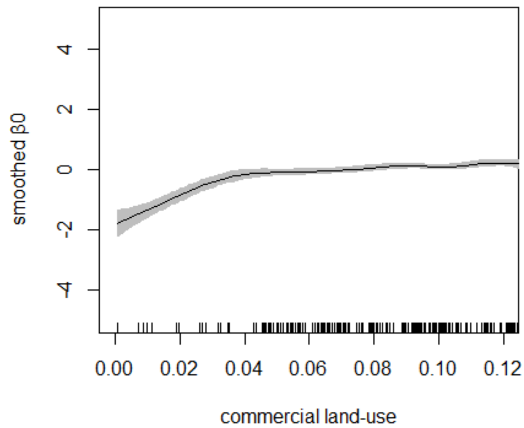




a

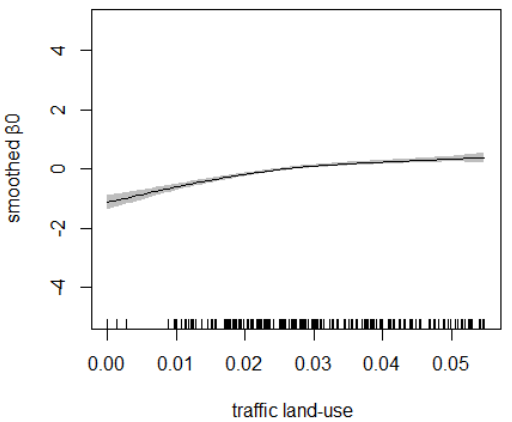
b

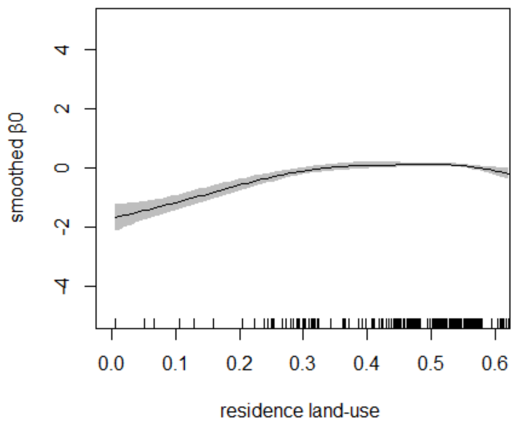




c

d

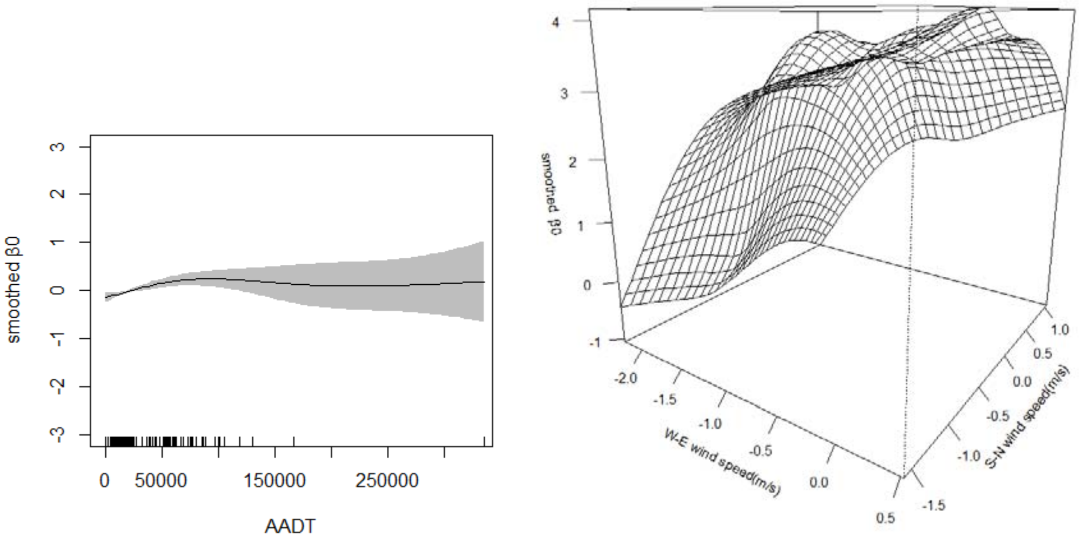




e

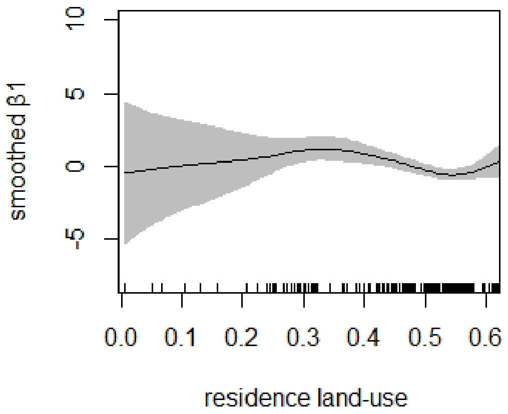
f

Figure S5. Non-linear relationship between β_0 and local covariates with the 95% confidence intervals for NO_x by GAM (to be continued)

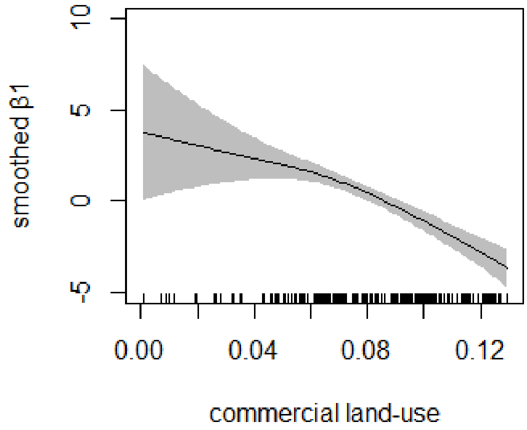


g h (wind speeds in W-E and S-N)

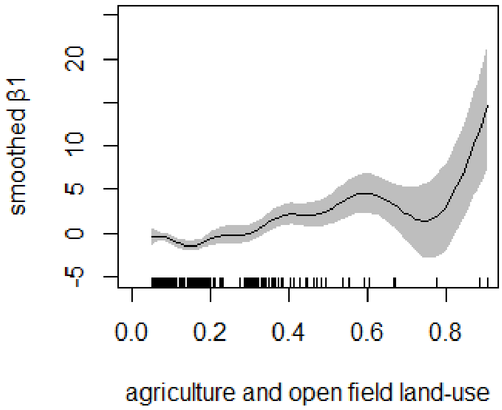
Figure S5. Non-linear relationship between β_0 and local covariates with the 95% confidence intervals for NO_x by GAM (continued)



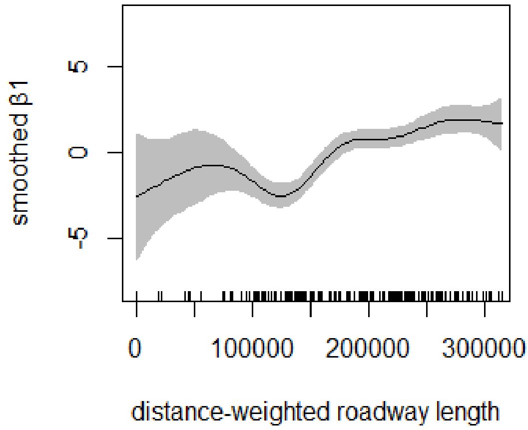
a



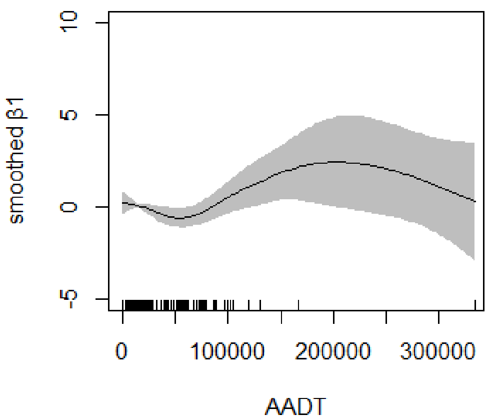
b



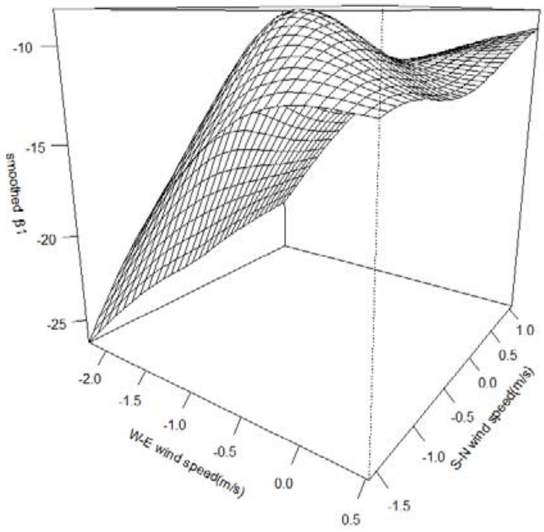
c



d



e



f (wind speeds in W-E and S-N)

Figure S6. Non-linear relationship between β_1 and local covariates with the 95% confidence intervals for NO_x by GAM

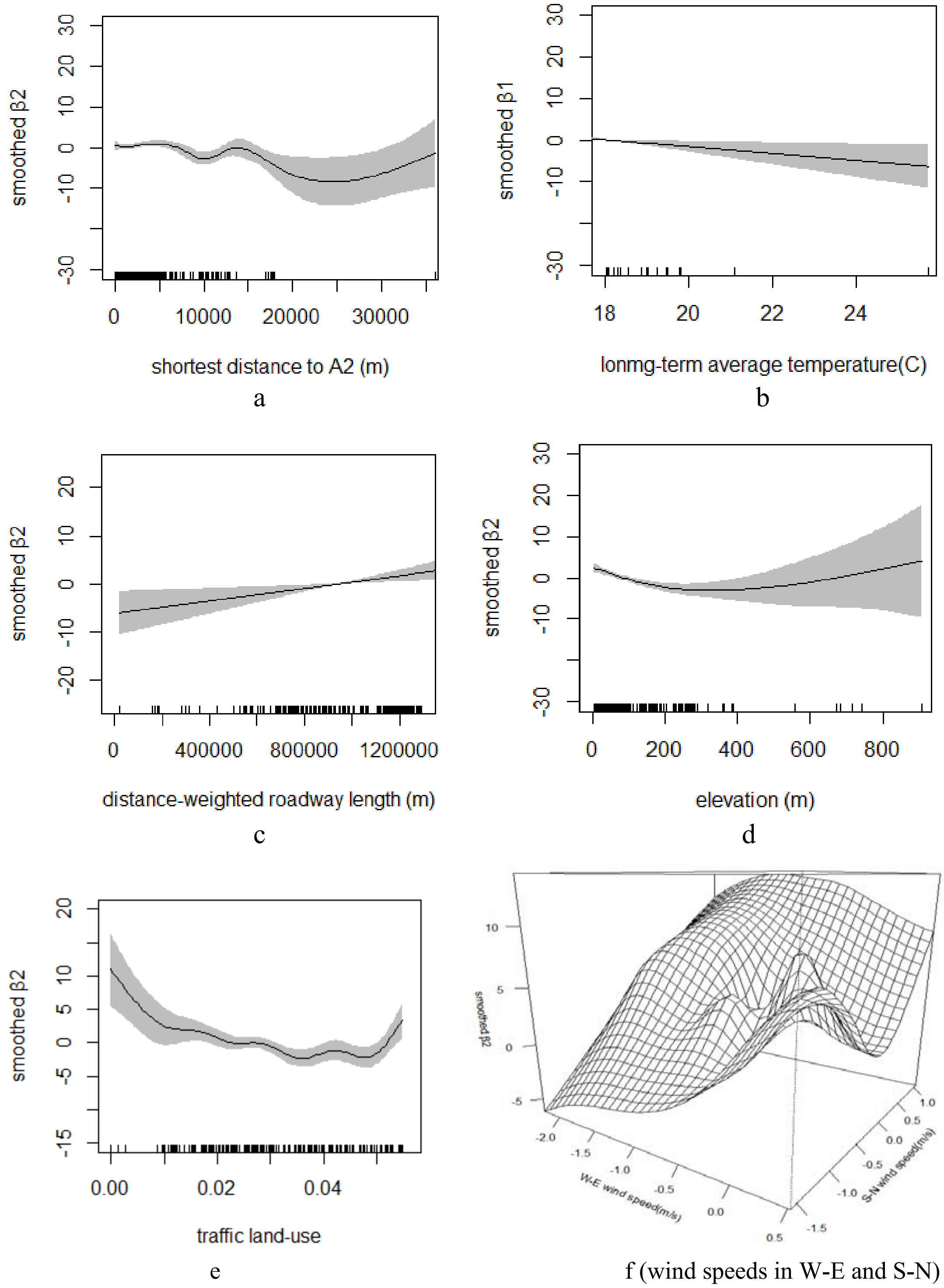
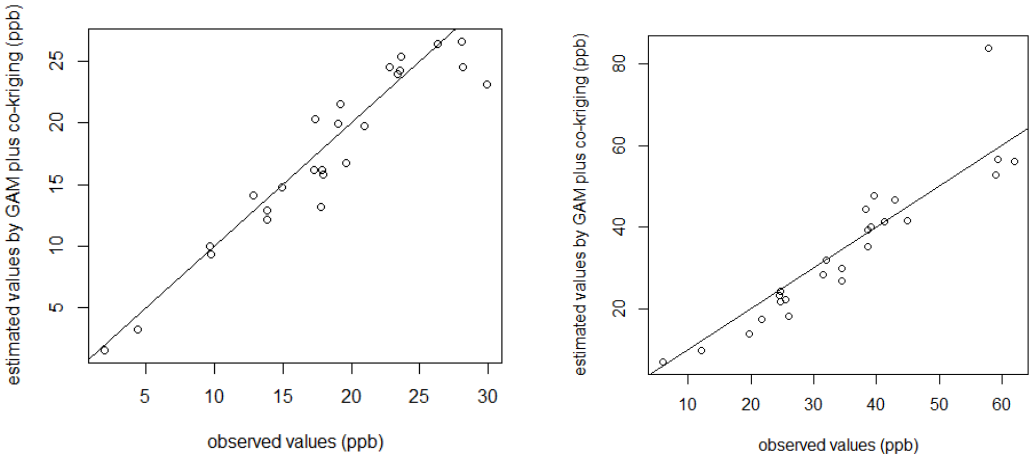


Figure S7. Non-linear relationship between β_2 and local covariates with the 95% confidence intervals for NO_x by GAM



a. NO_2 (R^2=0.89) b. NO_x (R^2=0.77)

Figure S8. Average values of long-term concentrations: observed values vs. estimated values

4. References

Duchon, J., 1977. Splines minimizing rotation-invariant semi-norms in solobev spaces, in: Schemp, W., Zeller, K. (Eds.), Lecture Notes in Mathematics: Construction Theory of Functions of Several Variables. Springer, Berlin, pp. 85-100.

Johnston, K., Hoef, M.J., Krivoruchko, K., Lucas, N., 2003. ArcGIS 9: Using ArcGIS Geostatistical Analyst, in: ESRI (Ed.).

Marcus, M., Minc, H., 1968. Elementary Linear Algebra. The MacMillan Company, NY.

Su, G.J., Jerrett, M., Beckerman, B., Wilhelm, M., Ghosh, K.J., Ritz, B., 2009. Predicting traffic-related air pollution in Los Angeles using a distance decay regression selection strategy Environmental Research 109, 657-670.

Szpiro, A.A., Sampson, D.P., Sheppard, L., Lumley, T., Adar, D.S., Kaufman, D.J., 2010. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. Environmetrics 21, 606-631.

U.S. Census Bureau, 1993. A Guide to State and Local Census Geography. Association of Public Data User, Princeton, NJ.

Wood, S.N., Augustin, N.H., 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecological Modelling 157, 157-177.

Wu, J., Funk, T.H., Lurmann, F.W., Winer, A.M., 2005. Improving Spatial Accuracy of Roadway Networks and Geocoded Addresses. Transactions in GIS 9, 585-601.