SUPPLEMENT TO "STATIC AND ROVING SENSOR DATA FUSION FOR SPATIO-TEMPORAL HAZARD MAPPING WITH APPLICATION TO OCCUPATIONAL EXPOSURE ASSESSMENT"

By Guilherme Ludwig[*], Tingjin Chu[†,¶], Jun Zhu[*], Haonan Wang[‡] and Kirsten Koehler[§]

University of Wisconsin-Madison[*], *Renmin University of China*[†], *Colorado State University*[‡], *and Johns Hopkins University*[§]

APPENDIX A: TUNING PARAMETER SELECTION IN CASE STUDY

A.1. Uncertainty Quantification. To quantify the uncertainty of estimated parameters and evaluate the overall prediction accuracy of the models, we employ a leave-one-sensor-out cross-validation procedure. For each static sensor $i = 1, \ldots, n_{\mathrm{S}}$, we remove the corresponding data $\boldsymbol{y}_{\boldsymbol{s}_i}$, and fit the model with the remaining data $\boldsymbol{y}^{-i}$.

To obtain confidence intervals and standard errors for $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\sigma}^2$, a cross-validated empirical distribution of the respective parameters can be used. Let $\hat{\alpha}$ denote a generic parameter estimate. By removing the data for one static sensor at a time, we obtain n_{S} different estimates $\hat{\alpha}^{(-i)}$ for $i = 1, \ldots, n_{\mathrm{S}}$. The standard error of $\hat{\alpha}$ is obtained by

$$\text{s.e.}(\hat{\alpha}) = \left\{ n_{\mathrm{S}}^{-1} \sum_{i=1}^{n_{\mathrm{S}}} (\hat{\alpha}^{(-i)} - \hat{\alpha})^2 \right\}^{1/2}.$$

In addition, the $(1-\alpha)100\%$ confidence interval can be constructed by the $(\alpha/2)$th and $(1-\alpha/2)$th empirical quantiles from the empirical cumulative distribution function

$$\widehat{F}(\eta) = n_{\mathrm{S}}^{-1} \sum_{i=1}^{n_{\mathrm{S}}} \mathcal{I}\{\hat{\alpha}^{(-i)} \le \eta\},$$

where $\mathcal{I}(\cdot)$ is the indicator function.

To obtain the mean squared prediction error (MSPE), the model fitted without $\boldsymbol{y}_{\boldsymbol{s}_i}$ is used to produce a predicted value $\hat{\boldsymbol{y}}_{\boldsymbol{s}_i}^{(-i)}$ and thus,

$$\text{MSPE} = n_{\mathrm{S}}^{-1} \sum_{i=1}^{n_{\mathrm{S}}} \sum_{k=1}^{p_i} \{y_{\boldsymbol{s}_i}(t_k) - \hat{y}_{\boldsymbol{s}_i}^{(-i)}(t_k)\}^2.$$

[¶] Tingjin Chu is the corresponding author (E-mail address: tingjin_chu@outlook.com).

A similar approach is taken to tune the smoothness of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\varphi}}$ estimators by searching over a grid of K and ζ values for K_0 and ζ_0 that minimize the MSPE. It is of interest to find optimal tuning parameters without resorting to cross validation, although tuning principal functional components automatically is an open problem by itself (Ramsay and Silverman, 2005, p.179).

A.2. Tuning parameter for case study. A preliminary step in the model fitting is to determine the number of basis functions K, the tuning parameter ζ, and the number of functional principal components L. We employed the leave-one-static-sensor-out algorithm to estimate the MSPE. More specifically, we considered a grid of values for the number of deterministic spline basis functions ($K = 4, 8, 12, 16$ and 32), the number of temporal principal components $L = 2, 3, 4$, and the principal components smoothing parameters $\zeta = 0, 1, 10, 100, 10^3, 10^4$, and 10^5. The results for the inhomogeneous variance case are given in Table 1. It also displays estimates $\hat{\sigma}^2_{\mathrm{S}}$ and $\hat{\sigma}^2_{\mathrm{R}}$. There is a modest improvement from incorporating a larger L number of components, particularly at lower smoothness values for both the deterministic mean function and the random spatio-temporal effects, with diminishing returns. The deterministic spline smoothness is optimal at mid-range values for K (either 8 or 12) and the random effects do not need to be smoothed (a value of $\zeta = 0$ seems best). The minimum MSPE corresponds to $K = 12$, $\zeta = 0$, and $L = 4$. However, we decided to go with $L = 3$ functional components, as the corresponding MSPE (11.8) is close to the minimum (11.2).

There are two remarks to be made. First, the choice of smoothing reflects on the estimates of σ^2_{S} and σ^2_{R}. The estimate of σ^2_{S} changes the most as a function of the tuning parameter ζ when $\zeta > 10^2$. Second, the fixed and random temporal effects compete with each other. For example, consider the MSPE when $L = 3$ for Table 1. A small K (yielding a smooth deterministic component) minimizes MSPE when ζ is small (yielding a rough stochastic component). Conversely, allowing 32 basis functions for the deterministic spline requires more smoothing of the stochastic component ($\zeta \approx 10^4$) to minimize the MSPE.

Table 1

Choosing the best tuning parameters for the STDF algorithm: estimates of MSPE, σ^2_{S} and σ^2_{R} based on choice of smoothness for deterministic trend (df), functional principal component smoothness (ζ) and number of principal components (L).

	K	MSPE							$\hat{\sigma}^2_{\mathrm{S}}$							$\hat{\sigma}^2_{\mathrm{R}}$						
ζ		0	1	10	10^2	10^3	10^4	10^5	0	1	10	10^2	10^3	10^4	10^5	0	1	10	10^2	10^3	10^4	10^5
$L = 2$	4	14.2	14.6	16.4	18.6	18.0	18.3	28.2	3.5	3.6	3.9	4.4	4.5	4.9	8.8	1.3	1.4	1.3	1.0	1.1	1.5	10.4
	8	12.0	12.0	12.8	21.4	17.4	18.1	30.6	2.1	2.1	2.2	2.7	3.9	4.3	7.7	1.9	2.1	2.4	2.8	1.1	1.5	9.1
	12	13.1	12.6	12.8	19.3	16.5	16.9	31.6	2.0	2.0	2.2	2.7	3.2	3.7	7.0	1.9	2.1	2.2	1.8	1.0	1.6	8.7
	16	13.8	14.2	16.6	34.6	16.5	17.2	32.0	1.5	1.5	1.6	2.1	3.2	3.7	6.6	1.1	1.2	1.5	1.9	0.9	1.5	7.7
	32	37.5	37.1	35.8	33.7	32.8	29.7	33.7	1.5	1.5	1.6	1.9	2.6	3.7	5.1	0.4	0.4	0.4	0.4	0.7	5.0	7.3
$L = 3$	4	12.8	13.0	13.0	13.6	17.0	18.1	27.6	2.2	2.3	2.4	2.8	4.1	4.8	8.7	1.4	1.5	1.6	1.6	1.1	1.5	10.4
	8	12.1	12.0	12.6	19.6	17.2	17.7	30.1	1.6	1.6	1.8	2.4	3.8	4.2	7.6	1.4	1.5	1.7	2.6	1.0	1.5	9.1
	12	11.8	11.9	12.4	18.6	16.1	16.3	31.1	1.6	1.6	1.7	2.5	3.1	3.6	6.9	1.2	1.4	1.6	1.7	0.9	1.6	8.6
	16	12.9	13.0	15.2	34.5	16.1	16.7	31.4	1.3	1.3	1.5	2.0	3.1	3.6	6.5	1.0	1.1	1.3	1.8	0.8	1.5	7.7
	32	37.6	37.1	35.5	33.0	32.3	29.5	32.6	1.4	1.4	1.5	1.8	2.4	3.6	5.0	0.3	0.3	0.3	0.4	0.6	5.1	7.3
$L = 4$	4	11.5	11.4	11.8	14.0	17.0	18.0	27.6	1.7	1.8	1.9	2.5	4.1	4.8	8.7	1.5	1.7	1.9	2.1	1.0	1.5	10.3
	8	11.4	11.5	12.4	19.3	17.3	17.6	30.1	1.5	1.5	1.7	2.3	3.7	4.2	7.6	1.2	1.4	1.5	2.5	0.9	1.5	9.1
	12	11.2	11.4	12.2	18.0	16.3	16.3	31.1	1.4	1.4	1.6	2.4	3.1	3.6	6.9	1.1	1.3	1.4	1.5	0.8	1.5	8.6
	16	12.7	13.0	15.1	34.4	16.2	16.7	31.4	1.3	1.3	1.4	1.9	3.1	3.6	6.5	0.9	1.0	1.2	1.8	0.7	1.4	7.6
	32	38.0	37.7	36.3	33.3	32.9	29.4	32.5	1.3	1.3	1.4	1.8	2.4	3.6	5.0	0.3	0.3	0.3	0.3	0.3	5.1	7.2

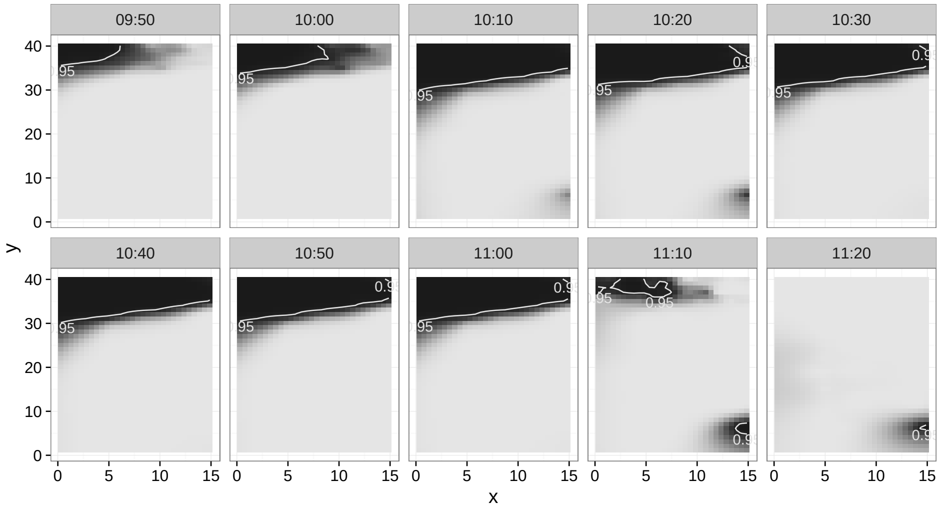

Fig A.1. *Probability of noise exceeding 85* dB *map; darker areas indicate higher probability, and the contour lines at 0.95 probability are included. Each panel corresponds to a point in time from 9:50* am *to 11:20* am *at 10-minute intervals.*

We tuned the parameters for smoothness in the homogeneous variance case similarly. In this case, the estimated MSPE values were close, so we decided to keep the same choices for K, ζ and L for comparison. We also fitted the model without using roving sensors at all and kept the choices of tuning parameters and number of components.

A.3. Risk maps. In addition to maps displaying the intensity of a hazard, an informative representation of the hazard can be made by plotting the probability of a hazard exposure exceeding a threshold. The marginal probabilities can be obtained by computing the z-scores using the hazard map and standard error maps, and evaluating the probability of exceedance. To illustrate, Figure A.1 shows the probability that the hazard levels exceed 85 dB, with contour lines at 0.95.

APPENDIX B: SIMULATION STUDY

Our purpose in this simulation is to study the quality of mapping, the effect of fusing roving and static sensor data in terms of prediction, and to compare our methodology with some of the existing approaches (Koehler and Peters, 2013). There are two models for the deterministic component $\mu_{\boldsymbol{s}}(t)$, three models for the covariance functions of $\eta_{\boldsymbol{s}}(t)$ and six configurations of sensors: a combination of either 6 or 18 static sensors, and either 0, 1 or 2

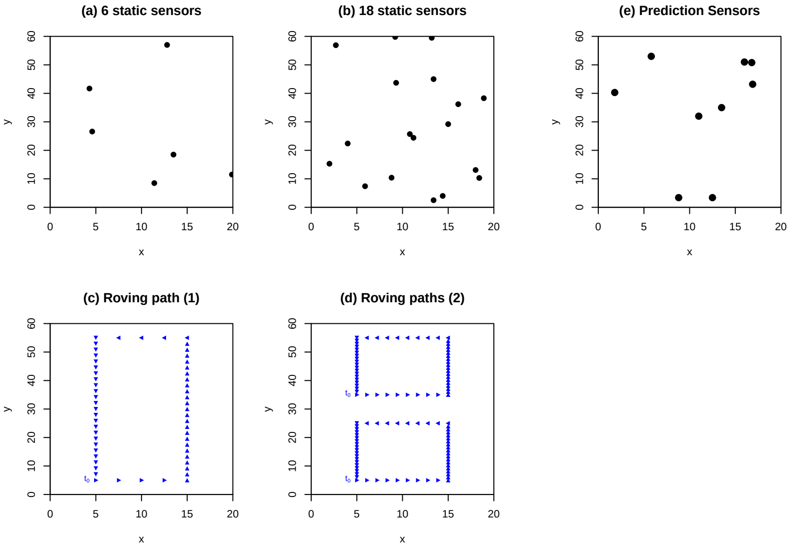

FIG A.2. *Sensor configurations used in the simulation. (a) and (b): Static sensors, either 6 or 18. (c) and (d): Roving sensors, either 0 (no picture), 1 (top) or 2 (bottom). (e): Static points for which the mean squared prediction error (MSPE) was calculated.*

roving sensors. The location of the sensors (roving, static and those reserved for MSPE calculation) are displayed in Figure A.2; the static sensor and prediction sensor locations were once generated randomly, but fixed across independent experiments. We included three scenarios for the measurement error variances $(\sigma^2_{\mathrm{S}}, \sigma^2_{\mathrm{R}})$ of the static and roving sensors: the first in which they are equal, a second in which $\sigma^2_{\mathrm{R}} = \sigma^2_{\mathrm{S}}/4$ and a third in which $\sigma^2_{\mathrm{R}} = 4\sigma^2_{\mathrm{S}}$. Each sensor was observed for 60 units of time with complete observations for all sensors, and each experiment was replicated, independently, 200 times.

In all scenarios our STDF method was fitted with the same tuning parameters. The number of basis functions used for the deterministic mean function is 3, and the number of basis functions for the random spatio-temporal effects part is 10, with smoothness parameter $\zeta = 0$. We also evaluated the effect of fitting models with homogeneous and inhomogeneous variances, the former of which is denoted by STDFh.

The mean functions used were either linear in time and space, with

$$\mu_A(\boldsymbol{s}, t) = 40 - s_x + s_y/2 - t/5,$$

or linear in space, nonlinear in time, with

$$\mu_B(\boldsymbol{s}, t) = 40 - s_x + s_y/2 - t/5 + 8\sin(2\pi t/60).$$

We allowed the deterministic time to be (potentially) nonlinear by including a spline term in the least squares detrending step.

For the spatio-temporal components, we used $\varepsilon_{\boldsymbol{s}}(t) \sim N(0, 1^2)$, $\varepsilon_{\boldsymbol{r}}(t) \sim N(0, \gamma_k 1^2)$, $\gamma_k = 1/4, 1, 4$, and $\eta_{\boldsymbol{s}}(t) \sim GP(0, \sigma(\boldsymbol{s}, t, \boldsymbol{s}', t'))$, for three choices of covariance function $\sigma(\boldsymbol{s}, t, \boldsymbol{s}', t')$:

(I) Independent case, in which $\eta(\boldsymbol{s}, t) = 0$.
(S) Spatially correlated case, in which $\sigma_\eta(\boldsymbol{s}, t, \boldsymbol{s}', t') = \sigma_\eta^2 e^{-\|\boldsymbol{s}-\boldsymbol{s}'\|/\theta_{\boldsymbol{s}}} \delta(t = t')$, where δ is the indicator function.
(ST) Spatio-temporal case, with $\sigma_\eta(\boldsymbol{s}, t, \boldsymbol{s}', t') = \sigma_\eta^2 e^{-\|\boldsymbol{s}-\boldsymbol{s}'\|/\theta_{\boldsymbol{s}}} e^{-|t-t'|/\theta_t}$.

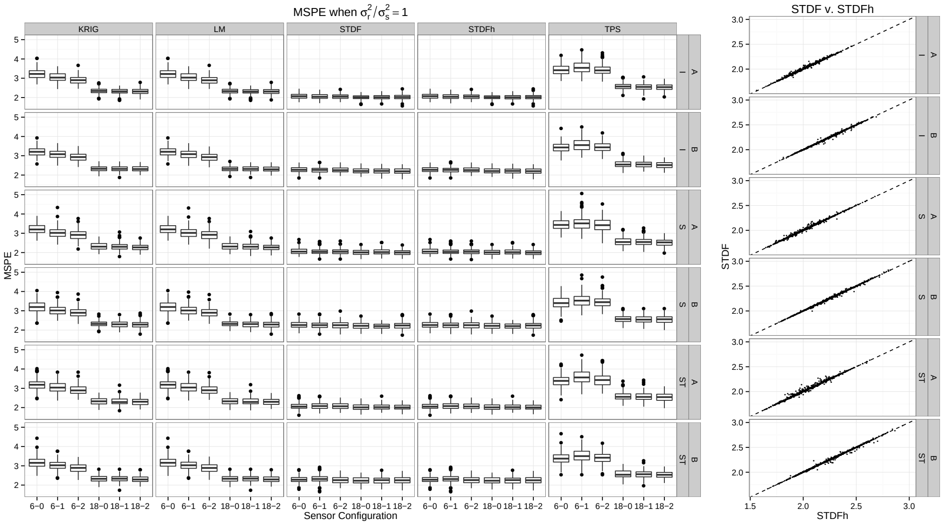

Fig A.3. *Left: Comparison of MSPE averaged in space and time in simulation study; x-axis is labeled according to static-roving sensor combination; strips in the top part of the panel indicate the model fitted, and strips in the right part of the panel indicate which deterministic mean and covariance model was used. In all cases,* $\sigma^2_{\mathrm{R}} = \sigma^2_{\mathrm{S}}$. *Right: Scatterplot of the MSPE results for the STDF model and the STDFh (STDF homogeneous) model. The observed ratio of MSPEs is* 0.9950 *on average.*

The traditional methods considered here were universal kriging (UK), thin-plate spline (TPS), and linear regression (LM) at transects in time. More specifically, we fix points in time and fit the traditional methods to the corresponding observations. Predictions are made based only on the observations at the same time point. The first two were fitted using the default specification in the `fields` R package (Nychka, Furrer and Sain, 2014). That is, for each fixed time, each of the methods was applied. The reported mean squared prediction errors (MSPE) are averages in time.

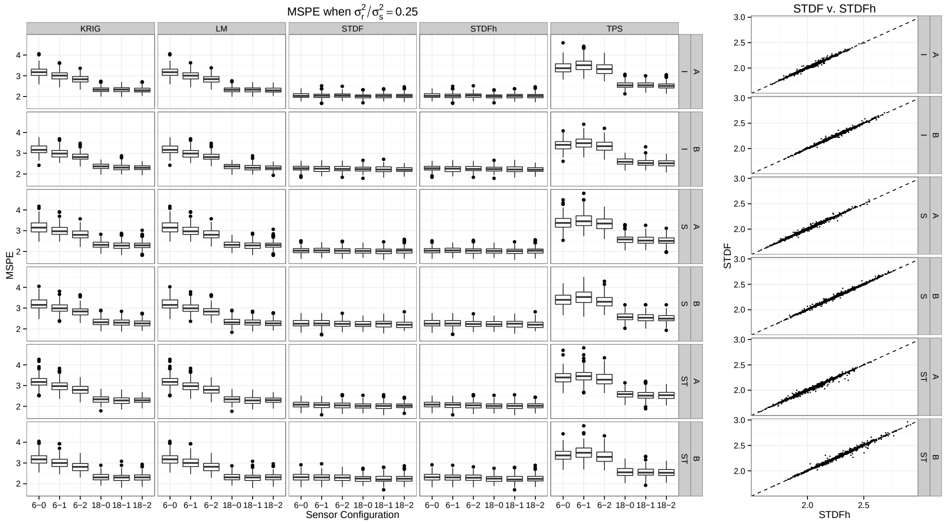

FIG A.4. *Left: Comparison of MSPE averaged in space and time in simulation study; x-axis is labeled according to static-roving sensor combination; strips in the top part of the panel indicate the model fitted, and strips in the right part of the panel indicate which deterministic mean and covariance model was used. In all cases,* $\sigma_{\mathrm{R}}^2 = 0.25\sigma_{\mathrm{S}}^2$. *Right: Scatterplot of the MSPE results for the STDF model and the STDFh (STDF homogeneous) model. The observed ratio of MSPEs is* 0.9994 *on average.*

The MSPE results are compared across models in Figures A.3, A.4 and A.5, with the former corresponding to the case when $\sigma_{\mathrm{S}}^2 = \sigma_{\mathrm{R}}^2$, the second with the case when $0.25\sigma_{\mathrm{S}}^2 = \sigma_{\mathrm{R}}^2$ and the latter when $4\sigma_{\mathrm{S}}^2 = \sigma_{\mathrm{R}}^2$. The results show that the STDF method outperforms the traditional methods in basically all scenarios. The results also reveal that our method is robust when roving sensors have variance greater than the static sensors (Figure A.5). The STDF method has similar performance for the homogeneous and inhomogeneous specifications when the measurement error variances are the same for static and roving sensors. The inhomogeneous case performs better when the roving sensor variances are larger, as shown in the right scatterplot panel for Figure A.5.

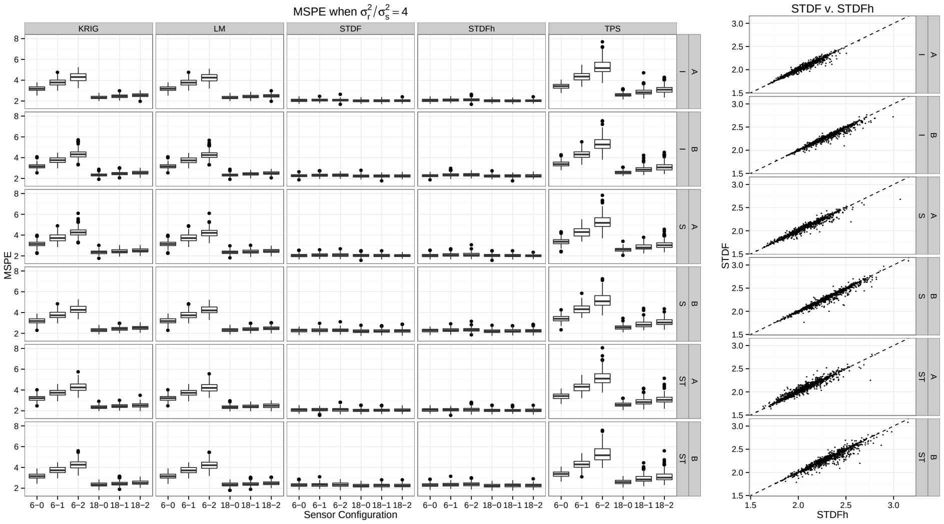

FIG A.5. *Left: Comparison of MSPE averaged in space and time in simulation study; x-axis is labeled according to static-roving sensor combination; strips in the top part of the panel indicate the model fitted, and strips in the right part of the panel indicate which deterministic mean and covariance model was used. In all cases,* $\sigma_{\mathrm{R}}^2 = 4\sigma_{\mathrm{S}}^2$. *Right: Scatterplot of the MSPE results for the STDF model and the STDFh (STDF homogeneous) model. The observed ratio of MSPEs is* 0.9471 *on average.*

REFERENCES

KOEHLER, K. A. and PETERS, T. M. (2013). Influence of analysis methods on interpretation of hazard maps. *Annals of Occupational Hygiene* **57** 558–570.

NYCHKA, D., FURRER, R. and SAIN, S. (2014). fields: Tools for spatial data R package version 7.1.

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*. Springer, New York.

GUILHERME LUDWIG
DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN-MADISON
MADISON, WI 53706
E-MAIL: gvludwig@stat.wisc.edu

TINGJIN CHU
CENTER FOR APPLIED STATISTICS AND
INSTITUTE OF STATISTICS AND BIG DATA
RENMIN UNIVERSITY OF CHINA
BEIJING 100872, CHINA
E-MAIL: tingjin_chu@outlook.com

JUN ZHU
DEPARTMENT OF STATISTICS AND DEPARTMENT OF ENTOMOLOGY
UNIVERSITY OF WISCONSIN-MADISON
MADISON, WI 53706
E-MAIL: jzhu@stat.wisc.edu

HAONAN WANG
DEPARTMENT OF STATISTICS
COLORADO STATE UNIVERSITY
FORT COLLINS, CO 80523
E-MAIL: wanghn@stat.colostate.edu

KIRSTEN KOEHLER
DEPARTMENT OF ENVIRONMENTAL HEALTH SCIENCES
JOHNS HOPKINS BLOOMBERG SCHOOL OF PUBLIC HEALTH
BALTIMORE, MD 21205
E-MAIL: kkoehle1@jhu.edu