Web-based Supplementary Materials for
"Spatial cluster detection for weighted outcomes using cumulative geographic residuals"

by
Andrea J. Cook, Yi Li, David Arterburn, and Ram C. Tiwari

Web Appendix

Asymptotic Equivalence of $Z_{loc}(x_1, x_2|b)$ and $\hat{Z}_{loc}(x_1, x_2|b)$, given the observed data, for the Cumulative Weighted Geographic Residual

We will begin by first showing the asymptotic distribution of $Z_{loc}(x_1, x_2|b)$ assuming the model for the data is,

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + e_i \quad \text{where } e_i \overset{ind}{\sim} (0, \sigma^2/w_i).$$

Consider the following one-term Taylor series expansion of $Z_{loc}(x_1, x_2|b)$ at $\boldsymbol{\beta}$,

$$\begin{aligned} Z_{loc}(x_1, x_2|b) &= \tfrac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i(x_1, x_2|r_i, s_i, b)\hat{e}_i \\ &= \tfrac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i(x_1, x_2|r_i, s_i, b)\left(Y_i - \mathbf{X}_i\boldsymbol{\beta}\right) \\ &\quad -\tfrac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i(x_1, x_2|r_i, s_i, b)\mathbf{X}_i\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\ &= \tfrac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i(x_1, x_2|r_i, s_i, b)e_i \\ &\quad -\tfrac{1}{n} \left(\sum_{i=1}^{n} W_i(x_1, x_2|r_i, s_i, b)\mathbf{X}_i\right)\left[\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right] + o_p(1). \end{aligned} \tag{1}$$

where $e_i = Y_i - \mathbf{X}_i\boldsymbol{\beta}$.

First, we will prove $Z_{loc}(x_1, x_2|b)$ is tight and then establish its asymptotic distribution. Throughout this proof we assume that X_i, r_i, and s_i are bounded. Start with the first part of $Z_{loc}(x_1, x_2|b)$, that is, let $P_{1i}(x_1, x_2|b) = W_i(x_1, x_2|r_i, s_i, b)e_i$. Note that $P_{1i}(x_1, x_2|b)$ is two monotone functions element-wise in (x_1, x_2), because $P_{1i}(x_1, x_2|r_i, s_i, b) = f(x_1|s_i, b) *$

$g(x_2|r_i,b)$ and $f(x_1|s_i,b) = I(x_1 < b \leq x_1 + b)$ and $g(x_2|r_i,b) = I(x_2 < b \leq x_2 + b)\hat{e}_i w_i$ are monotone functions on x_1 and x_2, respectively. Therefore, the processes $\{P_{1i}(x_1,x_2|b); i = 1,\ldots,n\}$ are "manageable" (Pollard, 1990 and Billias et al., 1997). It then follows from the functional central limit theorem, that $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} P_{1i}(x_1,x_2|b)$ goes to a zero-mean Gaussian distribution as $n \to \infty$, since $E(e_i) = 0$, and is tight.

For the second part we start by defining the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. It has been shown that, for the general exponential family, as long as the conditional mean of Y_i, $E(Y_i;\mathbf{X}_i)$, is correctly linked to $\mathbf{X}_i$ through a link function $g(\cdot)$ with

$$g\{E(Y_i;\mathbf{X}_i)\} = \mathbf{X}_i\boldsymbol{\beta},$$

that,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N_p(\mathbf{0}, \mathbf{A}^{-1}(\boldsymbol{\beta})\mathbf{B}(\boldsymbol{\beta})\mathbf{A}^{-1}(\boldsymbol{\beta}))$$

where

$$\mathbf{A}(\boldsymbol{\beta}) = p\lim_{n\to\infty}\left[\frac{1}{n}\mathbf{I}(\boldsymbol{\beta})\right],$$

$\mathbf{B}(\boldsymbol{\beta}) = p\lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n} U_{i\beta}(\boldsymbol{\beta})U_{i\beta}^T(\boldsymbol{\beta})\right]$ (Liang and Zeger, 1986) and $p\lim$ denotes the limit in probability (if it exists).

Since $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a zero-mean Gaussian distribution and $1/n\sum_{i=1}^{n} W_i(x_1,x_2|r_i,s_i,b)\mathbf{X}_i$ is bounded, then the second term of (1) is also tight, proving $Z_{loc}(x_1,x_2|b)$ is tight.

Further, since $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is asymptotically equivalent to, $\mathbf{A}^{-1}(\boldsymbol{\beta})\frac{1}{\sqrt{n}}U_\beta(\boldsymbol{\beta})$, then $Z_{loc}(x_1,x_2|b)$ is asymptotically equivalent to,

$$\begin{aligned}
\tilde{Z}_{loc}(x_1,x_2|b) &= \tfrac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_i(x_1,x_2|b) \\
&= \tfrac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i(x_1,x_2|b)e_i + \tfrac{1}{\sqrt{n}}\left(\tfrac{1}{n}\sum_{i=1}^{n} W_i(x_1,x_2|b)\mathbf{X}_i\right)\mathbf{A}^{-1}(\boldsymbol{\beta})U_\beta(\boldsymbol{\beta}) \\
&= \tfrac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[W_i(x_1,x_2|b)e_i + \left(\tfrac{1}{n}\sum_{i=1}^{n} W_i(x_1,x_2|b)\mathbf{X}_i\right)\mathbf{A}^{-1}(\boldsymbol{\beta})U_{i\beta}(\boldsymbol{\beta})\right].
\end{aligned}$$

For fixed (x_1, x_2), $\tilde{Z}_{loc}(x_1, x_2|b)$ is a sum of n independent and identically distributed zero-mean random vectors. By the multivariate central limit theorem, the finite-dimensional distributions of $\tilde{Z}_{loc}(x_1, x_2|b)$ are asymptotically zero-mean gaussian, implying the same for $Z_{loc}(x_1, x_2|b)$. This fact, together with the tightness of $Z_{loc}(x_1, x_2|b)$, implies that $Z_{loc}(x_1, x_2|b)$ converges weakly to a zero-mean Gaussian process with covariance function $E(\Psi_1(x_{1a}, x_{2a}|b)\Psi_1(x_{1b}, x_{2b}|b))$ at $((x_{1a}, x_{2a}|b), (x_{1b}, x_{2b}|b))$ as $n \to \infty$.

Next we will establish the weak distribution of $\hat{Z}_{loc}(x_1, x_2|b)$ which is defined as

$$\hat{Z}_{loc}(x_1, x_2|b) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[W_i(x_1, x_2|b)\hat{e}_i + \nu(x_1, x_2|b)\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})U_{i\beta}(\hat{\boldsymbol{\beta}}) \right] G_i, \tag{2}$$

where

$$\nu(x_1, x_2|b) = -\sum_{i=1}^{n} W_i(x_1, x_2|b)\partial\mu/\partial\boldsymbol{\beta} = -\sum_{i=1}^{n} W_i(x_1, x_2|b)\mathbf{X}_i.$$

$\mathbf{I}(\boldsymbol{\beta}) = -\partial U_\beta/\partial\boldsymbol{\beta}$ and G_i $(i = 1, \ldots, n)$ are independent mean 0 and variance 1 random variables that are also independent of $(Y_i, \mathbf{X}_i, s_i, r_i)$. Conditional on the data $\{(Y_i, \mathbf{X}_i, r_i, s_i), i = 1, \ldots, n\}$, the only random components in $\hat{Z}_{loc}(x_1, x_2|b)$ are $(G_1, \ldots, G_n)$. Thus, it follows from the multivariate central limit theorem that, conditional on the data, the finite-dimensional distributions of $\hat{Z}_{loc}(x_1, x_2|b)$ are asymptotically zero-mean normal. Since $\hat{Z}_{loc}(x_1, x_2|b)$ consists of monotone functions, $P_{2i}(x_1, x_2|b) = [W_i(x_1, x_2|b)\hat{e}_i + \nu(x_1, x_2|b)\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})U_{i\beta}(\hat{\boldsymbol{\beta}})]$, in (x_1, x_2), which are manageable, the functional central limit theorem implies that $\hat{Z}_{loc}(x_1, x_2|b)$ is tight. Define

$$\hat{\Psi}_i(x_1, x_2|b) = W_i(x_1, x_2, b)\hat{e}_i + \left(\frac{1}{n} \sum_{i=1}^{n} W_i(x_1, x_2|r_i, s_i, b)\mathbf{X}_i \right) \left[\frac{1}{n}\mathbf{I}(\hat{\boldsymbol{\beta}}) \right]^{-1} U_{i\beta}(\hat{\boldsymbol{\beta}}).$$

The conditional covariance function of $\hat{Z}_{loc}(x_1, x_2|b)$ at $((x_{1a}, x_{2a}|b), (x_{1b}, x_{2b}|b))$ is,

$$\frac{1}{n}\sum_{n=1}^{n}\hat{\Psi}_i(x_{1a}, x_{2a}|b)\hat{\Psi}_i(x_{1b}, x_{2b}|b)$$

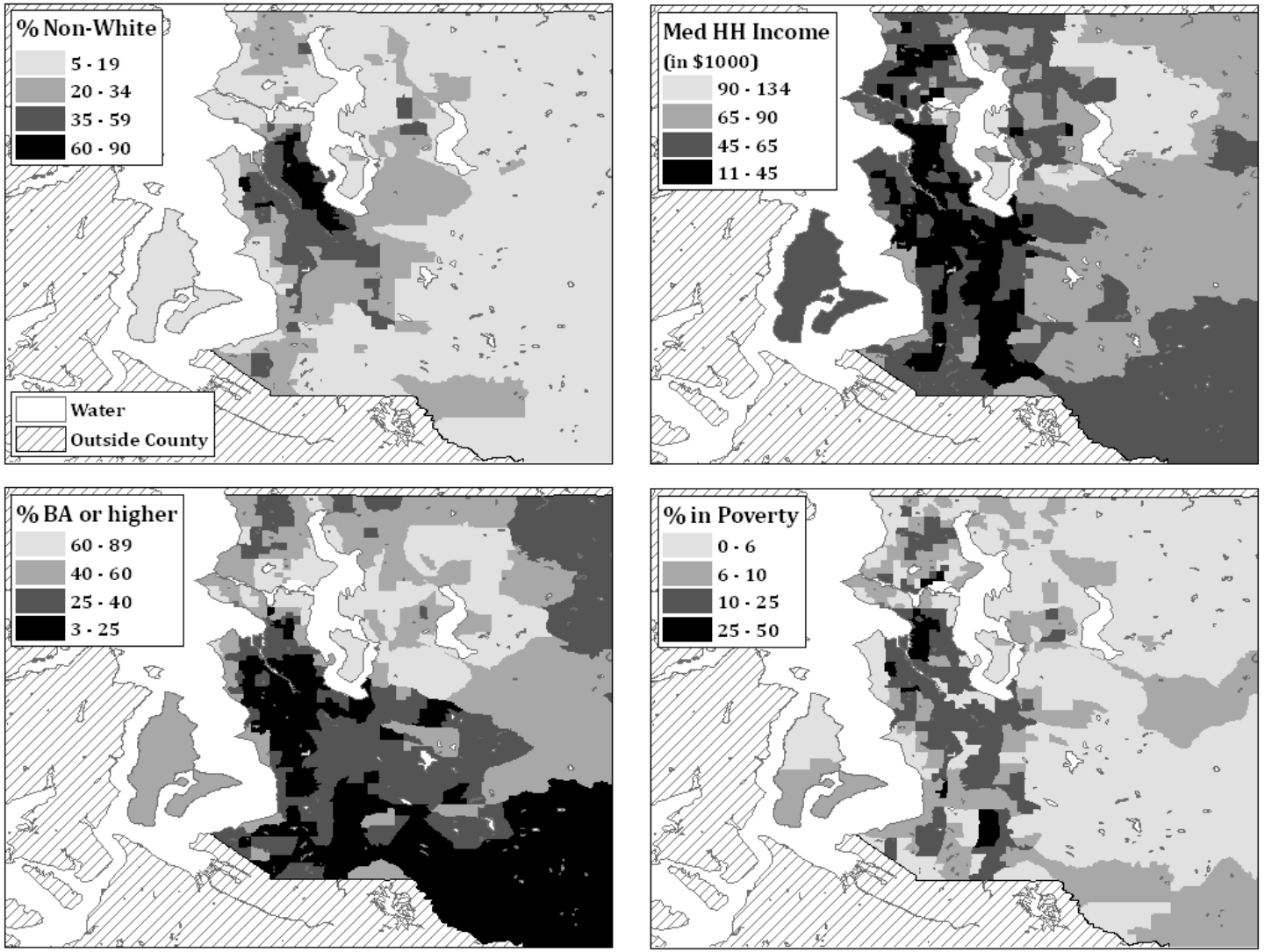
which converges to $E(\Psi_1(x_{1a}, x_{2a}|b)\Psi_1(x_{1b}, x_{2b}|b))$, the deterministic limiting covariance function of $Z_{loc}(x_1, x_2|b)$, as $n \to \infty$ by the law of large numbers, given $\hat{\Psi}_i(x_{1a}, x_{2a}|b)$ are *iid* and existence of $\left[\frac{1}{n}\mathbf{I}(\hat{\boldsymbol{\beta}})\right]^{-1}$. Therefore, $Z_{loc}(x_1, x_2|b)$ and $\hat{Z}_{loc}(x_1, x_2|b)$ converge to the same limiting zero-mean Gaussian process.

References

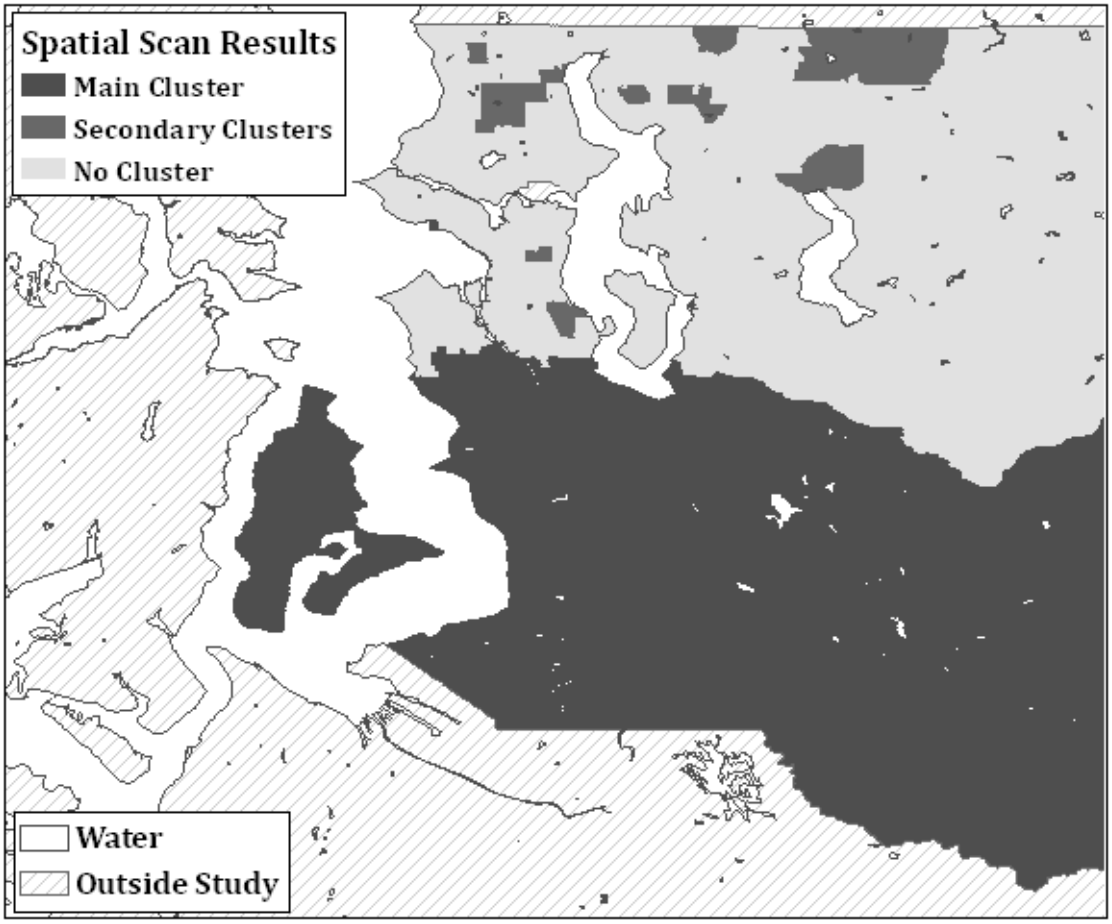
Billias, Y., Gu, M. and Ying, Z. (1997). Towards a general asymptotic theory for cox model with staggered entry. *The Annals of Statistics* **25**, 662–682.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Pollard, D. (1990). *Empirical Processes: Theory and Applications.* NSF-CBMS Regional Conference Series in Probability and Statistics 2, Institute of Mathematical Sciences, Hayward, CA.



Web Figure 1. Assessing spatial clustering of female BMI cluster in King County WA. Part A displays the raw mean census tract BMI and Part B and C display displays the areas with statistically significant spatial clustering from the unadjusted analyses and then adjusting for all individual and area-level covariates.



Web Figure 2. Assessing spatial clustering of female BMI cluster in King County WA. Apply the unweighted spatial scan assuming a normal model and no covariate adjustment using SaTScan^{TM} software (www.satscan.org)

Web Table 1

Type I error calculations of the Weighted Cumulative Geographic Residual Test for different number of regions and weights.

ν	σ	Type I error
1	1	0.060
1	5	0.052
1	10	0.043
1	20	0.052
1	30	0.060
10	1	0.049
10	5	0.049
10	10	0.050
10	20	0.048
10	30	0.052
20	1	0.069
20	5	0.054
20	10	0.044
20	20	0.051
20	30	0.046
30	1	0.052
30	5	0.041
30	10	0.063
30	20	0.053
30	30	0.065

Model Framework: $Y_i \sim N(0,1)$ with weight $w_i \sim \nu + \sigma * Uniform(0,1)$ and $i = 1, \ldots, N$

Type I error=$\frac{1}{1000} \sum_{j=1}^{1000} I(P - val_j < 0.05)$

Web Table 2

Type I error calculations of the Weighted Cumulative Geographic Residual Test when weights are variable.

N	weight	Type I error	N	weight	Type I error
25	1	0.029	100	1	0.040
25	20	0.016	100	20	0.055
25	40	0.034	100	40	0.042
25	60	0.020	100	60	0.029
25	80	0.037	100	80	0.042
36	1	0.035	121	1	0.040
36	20	0.038	121	20	0.050
36	40	0.033	121	40	0.049
36	60	0.032	121	60	0.044
36	80	0.034	121	80	0.050
49	1	0.043	144	1	0.045
49	20	0.041	144	20	0.059
49	40	0.043	144	40	0.047
49	60	0.049	144	60	0.045
49	80	0.033	144	80	0.043
64	1	0.039	169	1	0.041
64	20	0.042	169	20	0.048
64	40	0.051	169	40	0.050
64	60	0.034	169	60	0.046
64	80	0.054	169	80	0.049
81	1	0.052	225	1	0.046
81	20	0.042	225	20	0.040
81	40	0.046	225	40	0.039
81	60	0.061	225	60	0.040
81	80	0.049	225	80	0.058

Model Framework: $Y_i \sim N(0,1)$ with weight w_i and $i = 1, \ldots, N$

Type I error=$\frac{1}{1000} \sum_{j=1}^{1000} I(P - val_j < 0.05)$

Web Table 3

Power calculations of the Cumulative Geographic Residual Test for adjusted and unadjusted area-level covariate analyses when spatial clustering does not exist independent of area-level covariate and outcome relationship (c=0)

	β	Unadjusted Power	Adjusted Power
Moderate	-2	0.047	0.052
Dependence	-1	0.047	0.056
$\gamma = 0.5$	0	0.039	0.045
	1	0.066	0.060
	2	0.104	0.048
	3	0.126	0.035
	4	0.146	0.060
Strong	-2	0.058	0.052
Dependence	-1	0.068	0.036
$\gamma = 1.0$	0	0.053	0.049
	1	0.214	0.055
	2	0.330	0.043
	3	0.364	0.049
	4	0.373	0.037

Stage 1: Unadjusted Power is the proportion of statistically significant clusters when $E(Y_i|Z_i^*) = \beta\gamma Z_i^*$

Stage 2: Adjusted Power is the proportion of statistically significant clusters when $E(Y_i|Z_i^*, X_i) = \beta X_i$