Space-Time Areal Mixture Model: Relabeling Algorithm and Model Selection Issues

Md Monir Hossain, PhD
Assistant Professor of Pediatrics
Division of Biostatistics and Epidemiology

Acknowledgement

Collaborator:

Andrew B Lawson (MUSC, Charleston)

Russell S Kirby (USF, Tampa)

Bo Cai (USC, Columbia)

Jihong Liu (USC, Columbia)

Jungsoon Choi (South Korea)

NIH Grant:

NCI: R03 (08/06-08/08) (PI: Hossain)

NHLBI: R21 (06/09-04/12) (PI: Lawson, Cai, Hossain)

Earlier Works

Areal model (Stat in Med, 2006; EES, 2005):

- Local-likelihood cluster (LLC) model
  - Compared with the BYM model
  - Results: For detecting clusters of low and medium risk areas, LLC models signal better than BYM model
- Cluster detection diagnostics

Earlier Works

Spatio-temporal Areal model ((EES, 2012; EES, 2010) :

- Space-time local-likelihood cluster (LLCST) model
- Space-time mixture of Poisson (MPST) model
- Space-time cluster detection diagnostics

- Space-time stick-breaking process (SBPST)
  - Compared with the SREST model

Motivation

With the growing popularity of using spatial mixture model in cluster analysis, using model selection criteria to find the most parsimonious model is an established technique.

Label-switching is an inherent problem with the mixture models and a variety of relabeling algorithms have been proposed over the decade.

We used a space-time mixture of Poisson regression model with homogeneous covariate effects

The results are illustrated for real and simulated datasets.

The objective is to aware the researcher that if the purpose of statistical modeling is to identify the clusters, applying the relabeling algorithm to the best fitted model may not generate the optimum labeling.

Space-Time Mixture of Poisson Model

$$o_{it} \sim \sum_{l=1}^{L} \omega_{itl} Poisson\left(\theta_l e_{it} \exp\left(\eta_{it}\right)\right)$$

$$\eta_{it} = \beta_{i0} x_{it0} + \beta_{i1} x_{it1} + \cdots + \beta_{ip} x_{itp}$$

$$0 < \omega_{itl} < 1, \text{ and } \sum_{l=1}^{L} \omega_{itl} = 1$$

The weights are determined by the unobserved/hidden allocation variable:

$$\omega_{itl} = p\left(Z_{it} = l\right)$$

$$\omega_{itl} = \exp\left(h_{itl} / \phi\right) / \sum_{l=1}^{L} \exp\left(h_{itl} / \phi\right)$$

$$h_{itl} = \kappa_{il} + \gamma_{tl}$$

Model Selection Criteria

Deviance information criteria (Spiegelhalter et al., 2002):

$$\mathrm{DIC}_M = \overline{D(\Theta_M)} + p_M$$

Deviance information criteria (Celeux et al., 2006):

$$\mathrm{DIC}_{3M} = \overline{D(\Theta_M)} + \left[\overline{D(\Theta_M)} + 2\log \hat{p}(\mathbf{o} \mid \Theta_M) \right]$$

Mean square predictive error (Gelman and Ghosh, 1998):

$$\mathrm{MSPE}_M = \sum_{g=1}^{G} \sum_{i=1}^{n} \sum_{t=1}^{T} \left(O_{it} - O_{it}^{(\mathrm{g}, M)} \right)^2 \Big/ nTG$$

Relabeling Algorithm

Posterior similarity matrix:

$$\pi_{ijt} = \Pr\left(Z_{it} = Z_{jt} \mid \mathbf{o}\right) \approx \frac{1}{G}\sum_{g=1}^{G} I\left(Z_{it}^{(g)} = Z_{jt}^{(g)}\right)$$

$$E\left[\left(Z_{it} = Z_{jt}\right) \mid \mathbf{o}\right] = \pi_{ijt} = E\left[I\left(Z_{it} = Z_{jt}\right) \mid \mathbf{o}\right]$$

Viewed as a similarity matrix of posterior expected clustering.

Regarded as a similarity matrix for unknown true clustering.

Binder's loss:

$$L\left(\mathbf{Z}_t', \mathbf{Z}_t\right) = \sum_{i<j} l_1 \cdot I\left(Z_{it}' \neq Z_{jt}'\right) I\left(Z_{it} = Z_{jt}\right) + l_2 \cdot I\left(Z_{it}' = Z_{jt}'\right) I\left(Z_{it} \neq Z_{jt}\right)$$

$$E\left[L\left(\mathbf{Z}_t', \mathbf{Z}_t\right) \mid \mathbf{o}\right] = \sum_{i<j} \left| I\left(Z_{it}' = Z_{jt}'\right) - \pi_{ijt} \right|$$

$$\hat{\mathbf{Z}}_t = \arg\min_{g} E\left[L\left(\mathbf{Z}_t', \mathbf{Z}_t\right) \mid \mathbf{o}\right]$$

Relabeling Algorithm

Lau and Green (2007) criteria:

$$E\left[L\left(\mathbf{Z}_t', \mathbf{Z}_t\right) \mid \mathbf{o}\right] = \sum_{i<j} I\left(Z_{it}' = Z_{jt}'\right)\left(1-\pi_{ijt}\right)$$

Posterior expected adjusted Rand (PEAR):

$$AR\left[\mathbf{Z}_t', E\left(\mathbf{Z}_t \mid \mathbf{o}\right)\right] = \frac{\sum_{i<j} I\left(Z_{it}' = Z_{jt}'\right)\pi_{ijt} - \sum_{i<j} I\left(Z_{it}' = Z_{jt}'\right)\sum_{i<j}\pi_{ijt} \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_{i<j} I\left(Z_{it}' = Z_{jt}'\right) + \sum_{i<j}\pi_{ijt}\right] - \sum_{i<j} I\left(Z_{it}' = Z_{jt}'\right)\sum_{i<j}\pi_{ijt} \Big/ \binom{n}{2}}$$

$$\hat{\mathbf{Z}}_t = \arg\max_g AR\left[\mathbf{Z}_t', E\left(\mathbf{Z}_t \mid \mathbf{o}\right)\right]$$

Variable Selection

Kuo and Mallick (1998):

$$\eta_{ij} = \beta_{i0} x_{ij0} + I_1 \beta_{i1} x_{ij1} + \cdots + I_p \beta_{ip} x_{ijp}$$

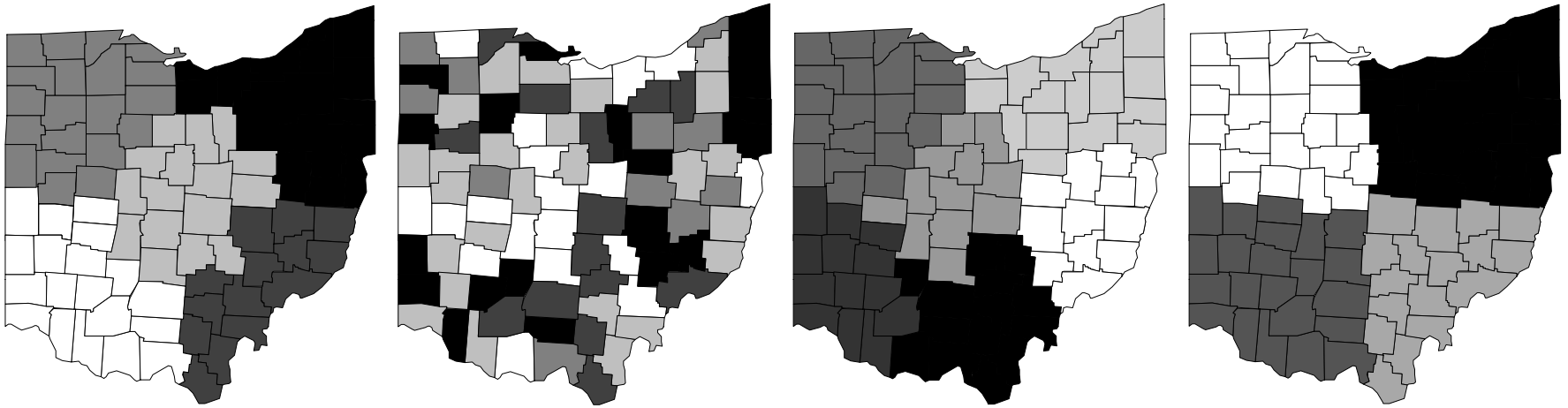
Covariate being selected or not is independent of covariate effect a priori:

$$p\left(I_q \beta_{iq}\right) = p\left(I_q\right) p\left(\beta_{iq}\right)$$

Simulated Data Example

Simulation-1: 2 clusters

Simulation-2: 5 clusters



$$x_{itq} \sim \text{Uniform}(0,1); i = 1,\ldots,n,\ t = 1,\ldots,T, \text{ and } q = 1,\ldots,3$$

Ohio geography: Expected lung cancer (year: 1968-88)

Realization: 100

Simulated Data Example

Simulation-1

Indicator variable for covariate effect	Posterior mean	Standard error
1	0.09778	0.22036
2	0.72674	0.20037
3	0.98873	0.01051

Simulation-2

Indicator variable for covariate effect	Posterior mean	Standard error
1	0.67662	0.00937
2	0.83869	0.01011
3	0.95003	0.01030

Simulated Data Example

Simulation-1

Covariate	DIC	DIC3	MSPE	Binder's loss	Lau & Green loss	Maximum PEAR
Intercept only	11618.51	11944.69	189.7469	1174.993	1173.514	0.37789
Intercept and 1	11619.55	11953.45	185.6524	1093.262	1091.638	0.39438
Intercept and 2	11986.08	12277.73	214.2977	337.1096	336.5138	0.57292
Intercept and 3	**11541.8**	**11873.84**	**173.6958**	221.3878	220.0612	0.545087
Intercept, 1 and 2	11573.86	11915.37	180.5074	291.3908	291.0096	0.4914627
Intercept, 1 and 3	11615.46	11958.0	174.4494	199.1183	198.8899	0.5548457
Intercept, 2 and 3	11580.64	11920.02	175.5993	**137.805**	**137.7711**	**0.6375331**
Full model	11564.58	11906.14	175.4762	175.4762	167.6591	0.4951442

Simulated Data Example

Simulation-2

Covariate	DIC	DIC3	MSPE	Binder's loss	Lau & Green loss	Maximum PEAR
Intercept only	**13089.64**	**13635.52**	**422.1604**	1237.826	1084.311	0.2069
Intercept and 1	14840.08	15964.27	1221.702	1030.294	991.5413	0.3707
Intercept and 2	15512.65	16973.48	1487.344	873.979	851.8798	0.4744
Intercept and 3	16293.81	17920.75	1981.996	**873.1445**	**850.3804**	**0.4700**
Intercept, 1 and 2	15574.62	16992.66	1523.662	901.3975	877.3205	0.4447
Intercept, 1 and 3	15287.31	16695.92	1358.119	877.1459	853.2936	0.4658
Intercept, 2 and 3	15235.93	16609.79	1388.221	878.995	854.6614	0.4707
Full model	14131.80	15034.43	970.9007	1091.618	1037.593	0.3477

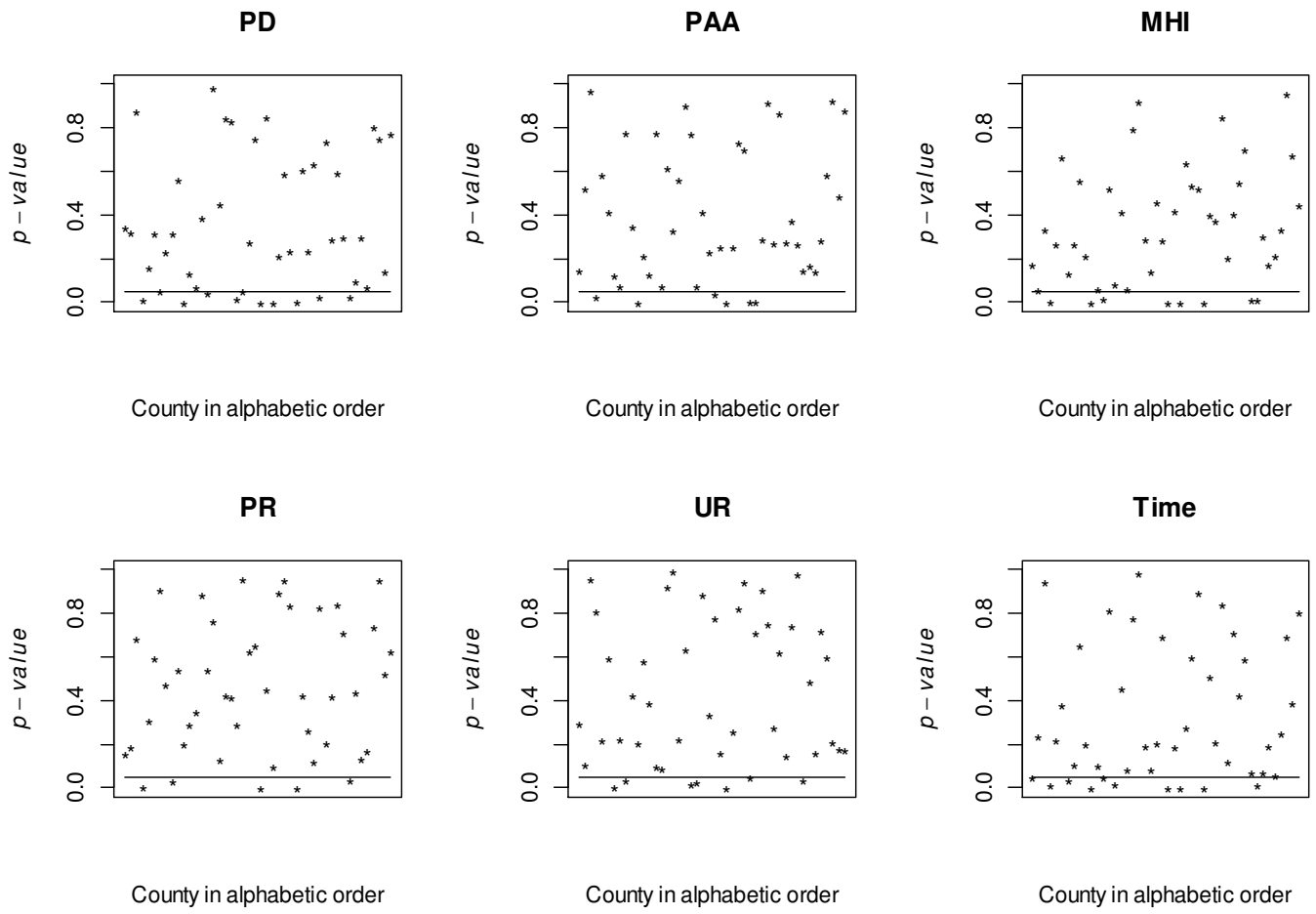
Real Data Example

South Carolina low birth weight data (Year: 1997-2007)

Component	DIC	DIC_3	MSPE
2	3660.9	3783.7	270.8
3	3620.1	3734.1	252.8
4	3646.5	3771.2	251.9
5	**3609.0**	**3720.7**	**241.4**
6	3633.9	3754.6	253.8
7	3650.0	3780.6	250.9

Indicator variable for covariate effect	Posterior mean	Standard error
PD	0.2347	0.4238
PAA	0.1835	0.3870
MHI	0.4565	0.4981
PP	0.1862	0.3892
UR	0.1797	0.3839
Time	0.5883	0.4921

Real Data Example



Real Data Example

South Carolina low birth weight data (Year: 1997-2007)

Covariate	DIC	DIC_3	MSPE	Binder's loss	Lau & Green loss	Maximum PEAR
Intercept only	**3568.6**	**3662.4**	**223.7**	170.78	168.59	0.607
Intercept, MHI	3597.0	3700.4	232.2	**105.62**	**102.43**	**0.780**
Intercept, time	3572.2	3663.0	227.7	497.71	497.10	0.037
Intercept, MHI, time	3582.2	3677.7	232.9	177.38	177.30	0.645
Full model	3609.0	3720.7	241.4	306.64	265.10	0.270

Conclusions

Previously, Best et al. (2005) showed that the spatial model with convolution prior (e.g. Besag et al., 1991) overestimate the risk surface for the high risk areas and the best model selected by the DIC is not always able to select the right clusters.

We used a space-time mixture of Poisson regression model with homogeneous covariate effects.

We designed two simulation studies with smaller and larger numbers of clusters, and with common covariate effects. The covariates are generated with stronger to weak levels of spatial correlation.

In our simulated and real datasets, we observed that model selection criteria do not indicate to the right cluster model.

Conclusions

Limitations:

Simultaneous estimation of the number of clusters and variable selection for spatial data

Improvement the DIC performance:

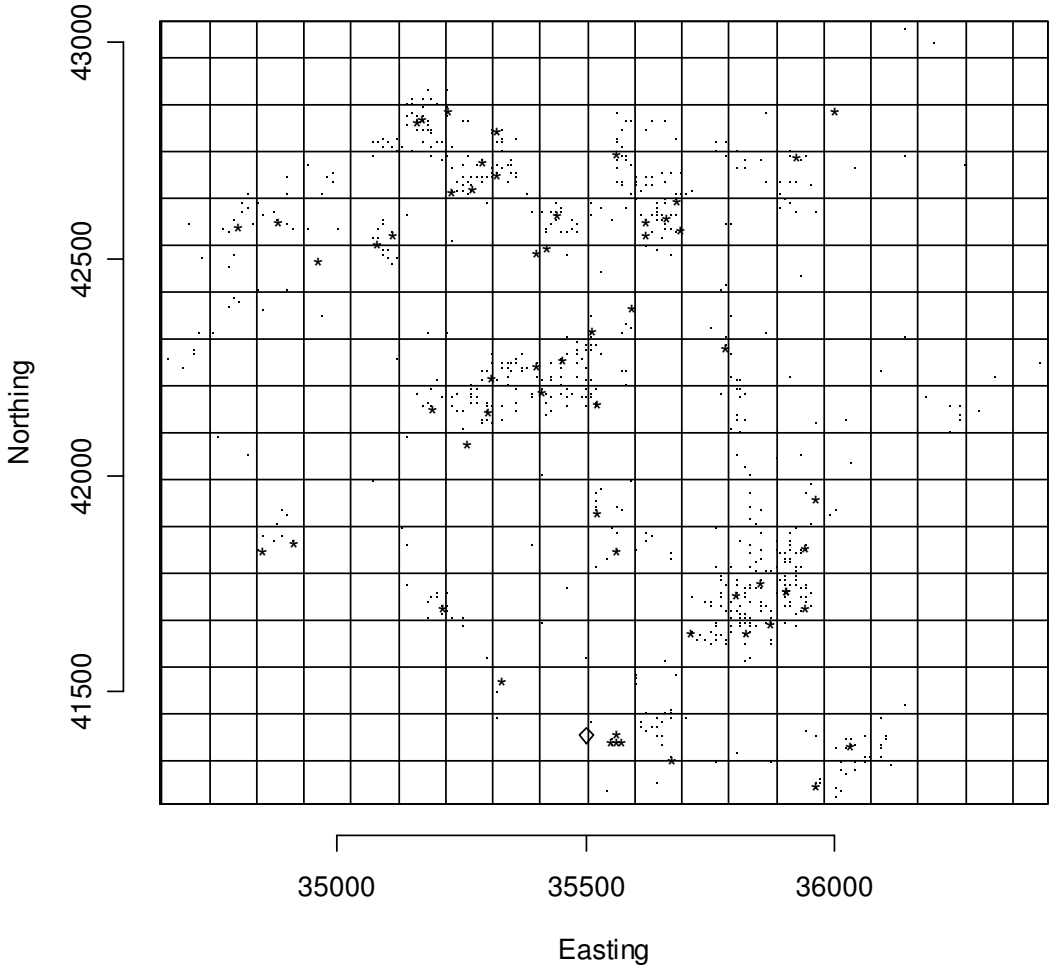
- **Variational Bayes approach (McGrory and Titterington, 2007)**
- **Viewing DIC as an approximate penalized loss function (Plummer, 2008)**

Model based relabeling algorithm where the label for each observation is chosen by maximizing the classification probability (Yao, 2012)

Earlier Works

Lancashire larynx cancer (58 cases) and lung cancer (978 controls)

Recorded in Chorley and South Ribble Health Authority during 1974-83



Earlier Works

Point process modeling (CSDA, 2010):

- Approximate likelihood
  - Berman-Turner (BT) model
  - Conditional logistic (CL) model
  - Binomial mesh (BM) model
  - Poisson mesh (PM) model

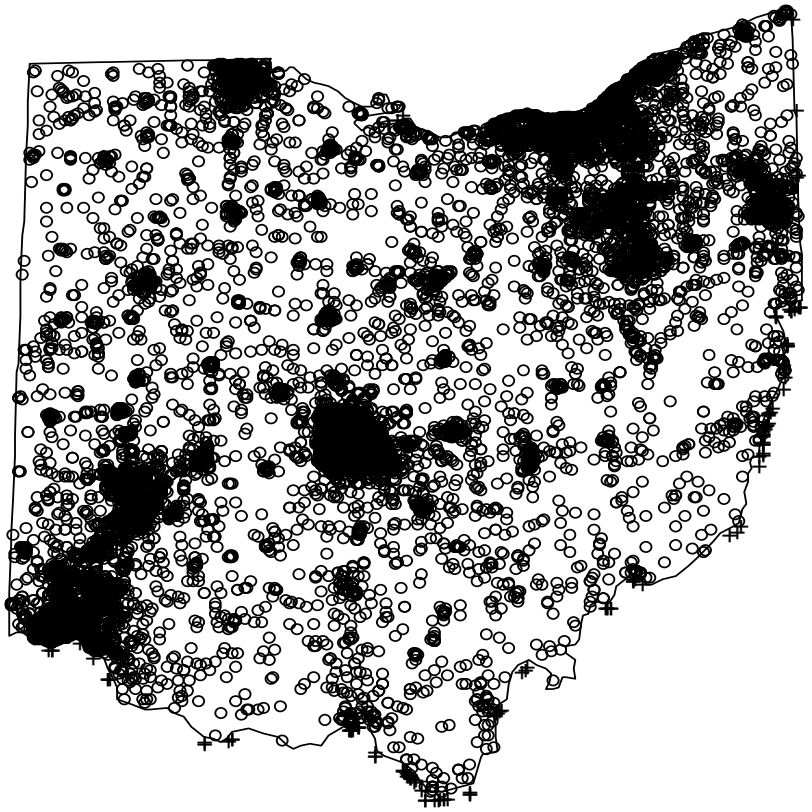
Earlier Works

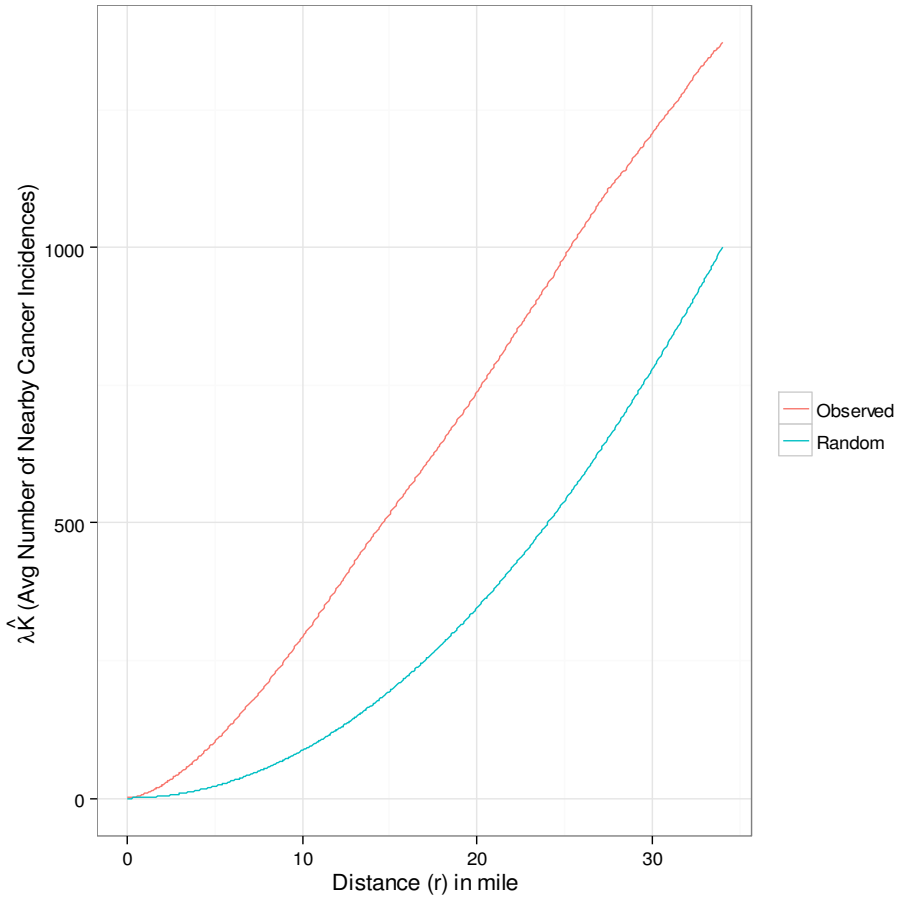
Point process modeling:

Parameter	BT method		CL model	PM model		BM model	
	$\phi = 0.0001$	$\phi = 0.001$		20X20	30X30	20X20	30X30
β_0	-0.641 (-4.112, 1.474)	-0.648 (-4.018, 1.489)	-0.739 (-4.249, 1.582)	-0.868 (-5.027, 1.672)	-0.843 (-4.874, 1.755)	-0.915 (-5.255, 1.872)	-0.773 (-4.965, 2.087)
β_1	-0.00078 (-0.00383, 0.002161)	-0.00078 (-0.00384, 0.00221)	-0.00077 (-0.00432, 0.00252)	-0.00050 (-0.00388, 0.00307)	-0.00048 (-0.00388, 0.00311)	-0.00048 (-0.00398, 0.00317)	-0.00059 (-0.00424, 0.00313)
$\log(\rho)$	-3.155 (-3.752, -2.758)	-3.161 (-3.684, -2.767)	-3.374 (-4.076, -2.885)	-3.200 (-3.737, -2.808)	-3.215 (-3.755, -2.804)	-3.145 (-3.705, -2.717)	-3.173 (-3.726, -2.756)
σ_u	0.390 (0.127, 0.891)	0.411 (0.095, 0.932)	0.771 (0.241, 1.475)	0.165 (0.004, 0.473)	0.141 (0.001, 0.522)	0.216 (0.009, 0.633)	0.242 (0.002, 0.728)
σ_v	0.050 (0.025, 0.087)	0.0331 (0.005, 0.071)	0.224 (0.078, 0.510)	0.529 (0.111, 1.305)	0.637 (0.095, 1.440)	0.537 (0.014, 1.342)	0.647 (0.090, 1.632)

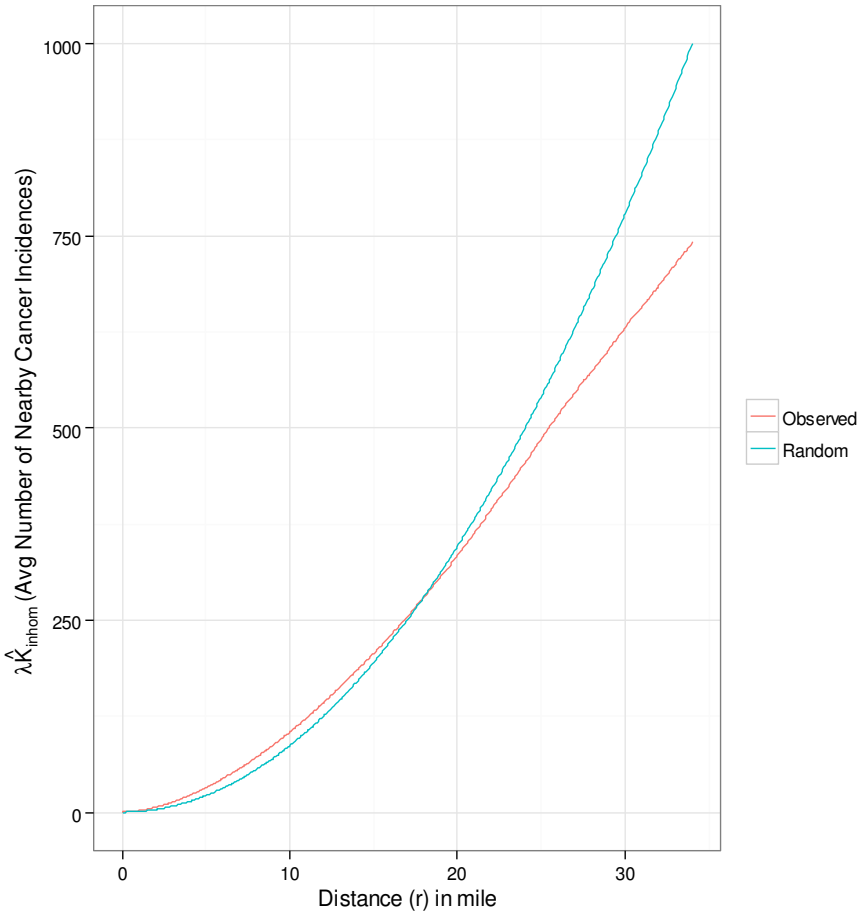
Current Works (Childhood Cancer)

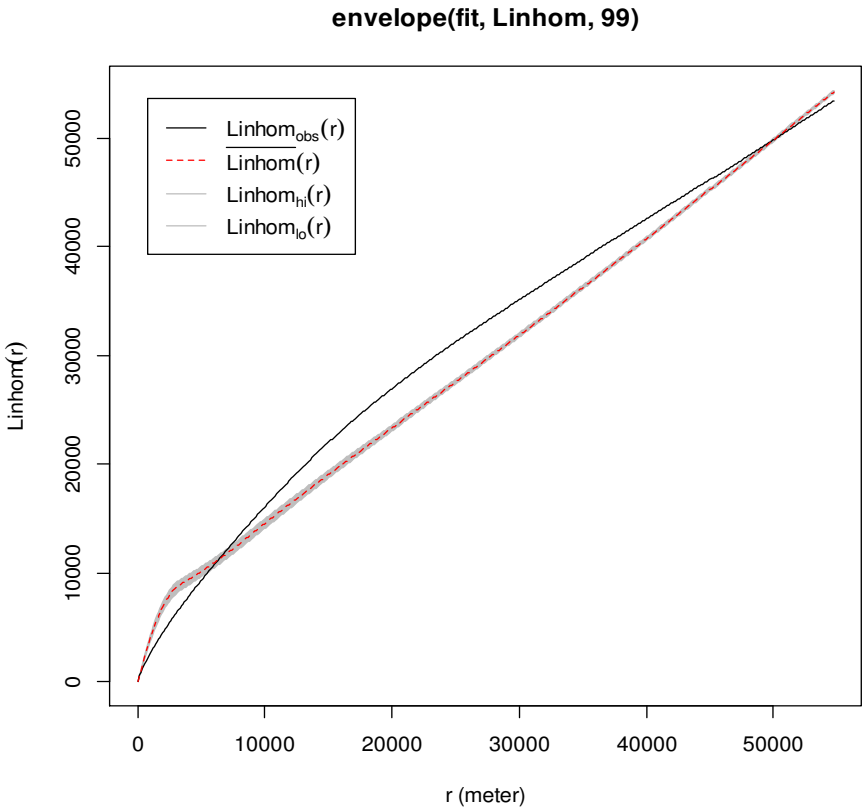
Childhood Cancer (age: 0-24, year: 1996-2009)





Current Works (Childhood Cancer)





Thanks!