Geospatial Modeling for Cancer Prevention and Control

Andrew Lawson
Department of Public Health sciences
MUSC

NCI Geospatial Approaches to Cancer Control
2016

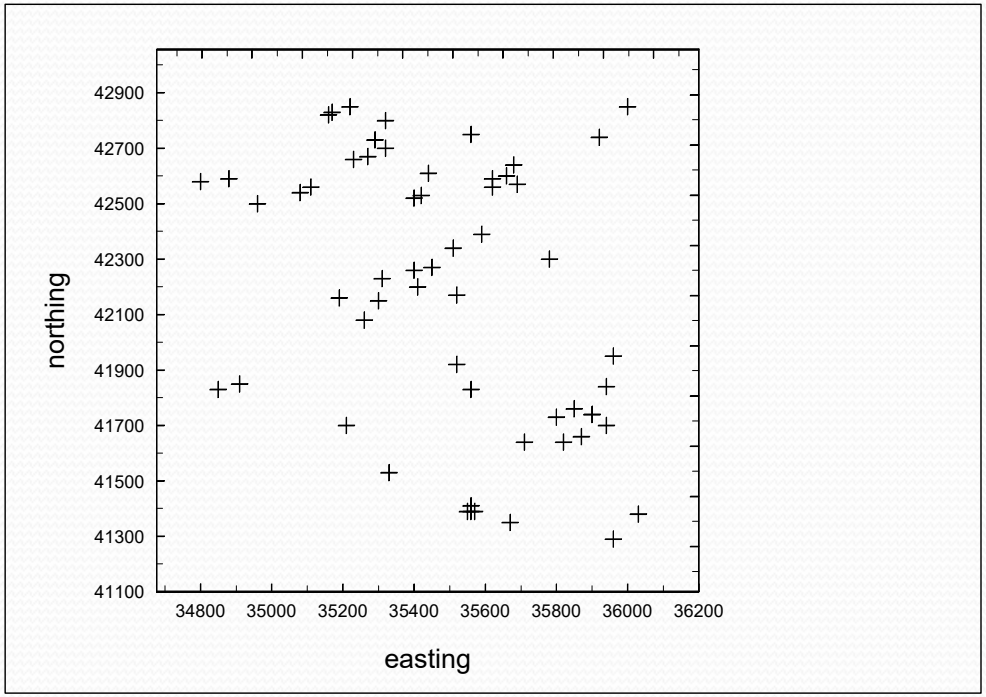
Small area health data characteristics

- Small area health data is characterized by discrete outcomes:
  - Incident/prevalent counts in arbitrary regions (zip codes, census tracts, ZCTAs, counties etc)
  - Indicator flag for individual outcome (late/early stage breast cancer; competing risks: larynx versus lung cancer) which is geo-referenced
  - Categorical individual outcomes: stages of cancer, multiple competing risks (osphageal, larynx, lung) which are geo-referenced

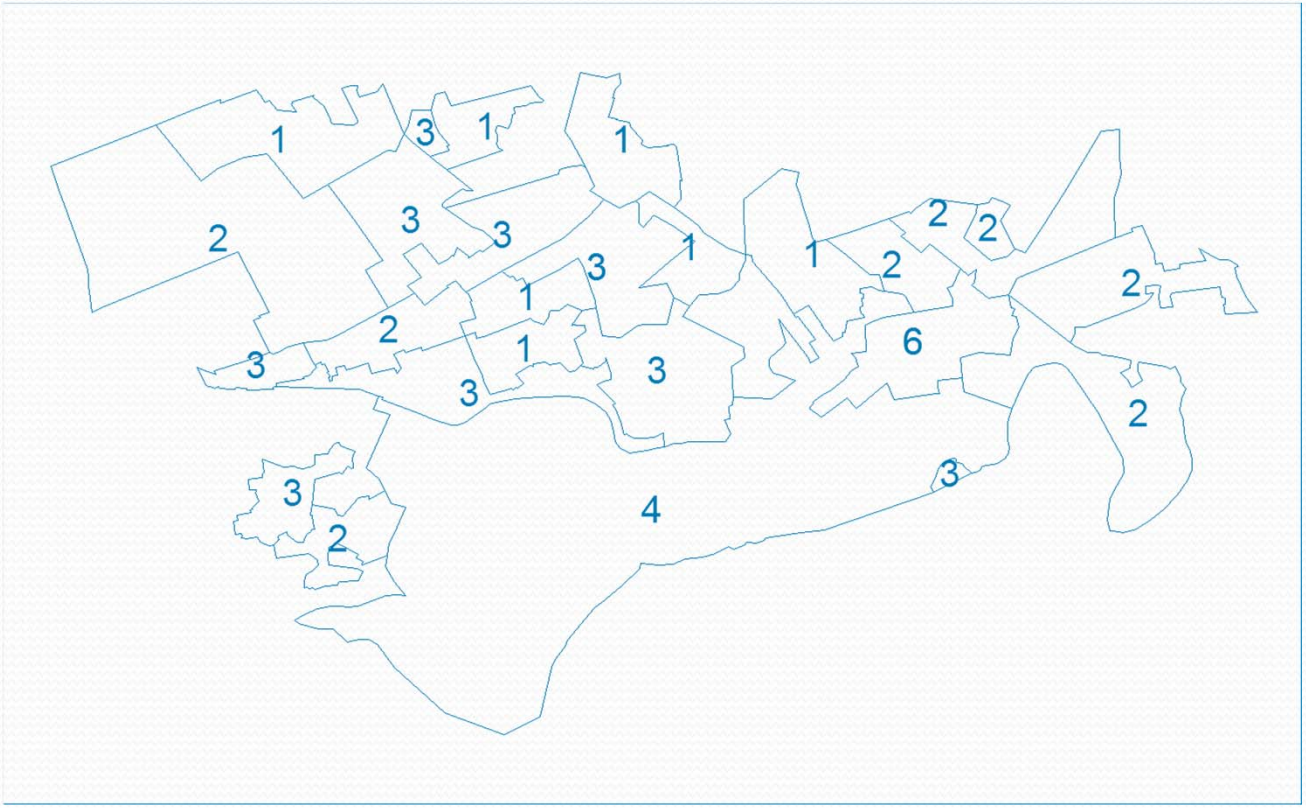
When is geo-referencing important?

- Environmental effects on health are often highly localised and depend crucially on exposure at or near residential address
- Residential address is often assumed to be a surrogate for exposure itself.
- Health services availability/access could be geographically constrained
- Genetic effects could be evident in homogeneous and non-dynamic populations
- Behavioral factors could be neighborhood specific

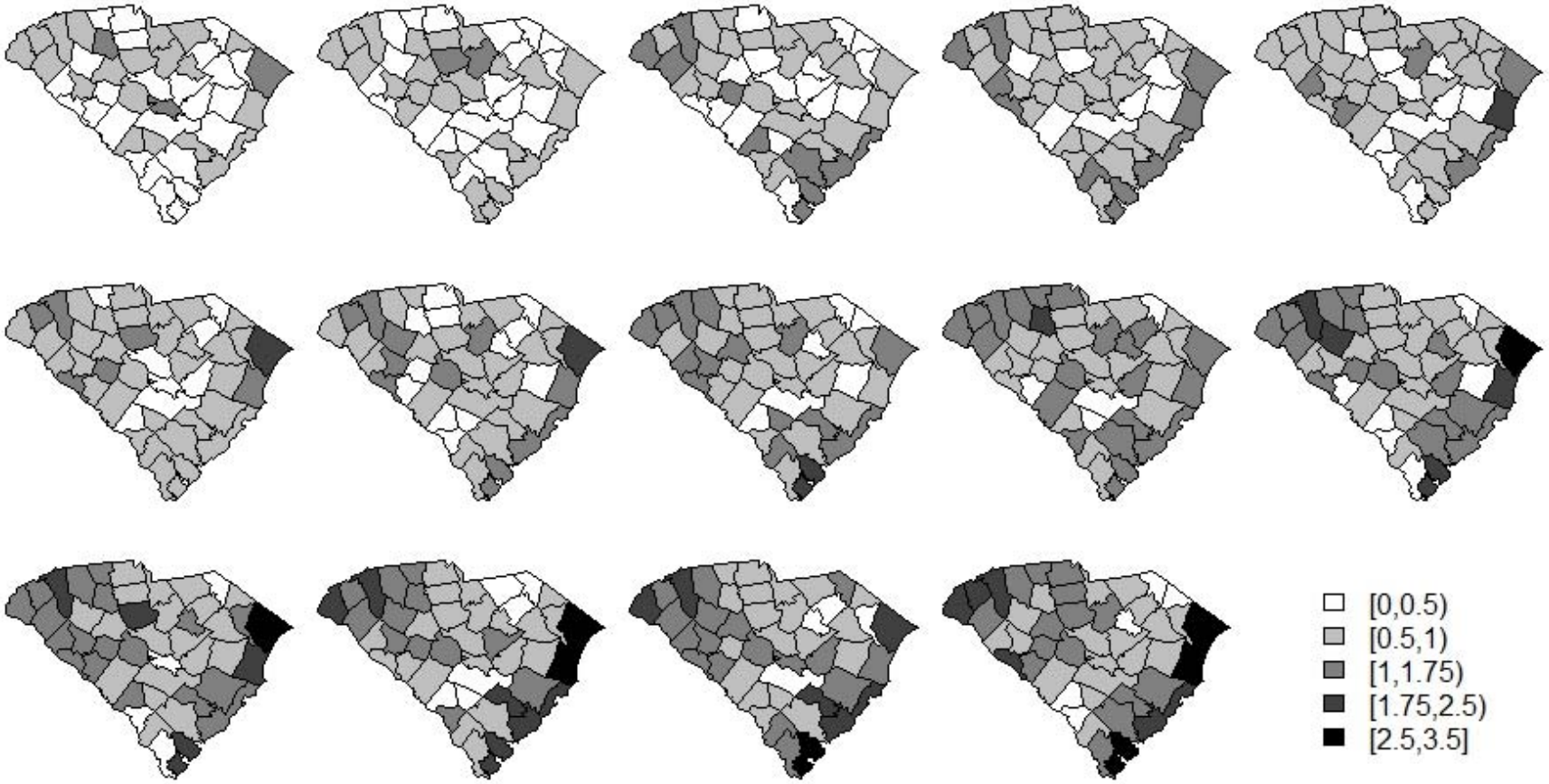
Larynx cancer case incidence in NW England 1973 – 1984 (residential addresses)



26 Census tracts in Falkirk, Central Scotland: counts of respiratory cancer deaths 1978 - 1983

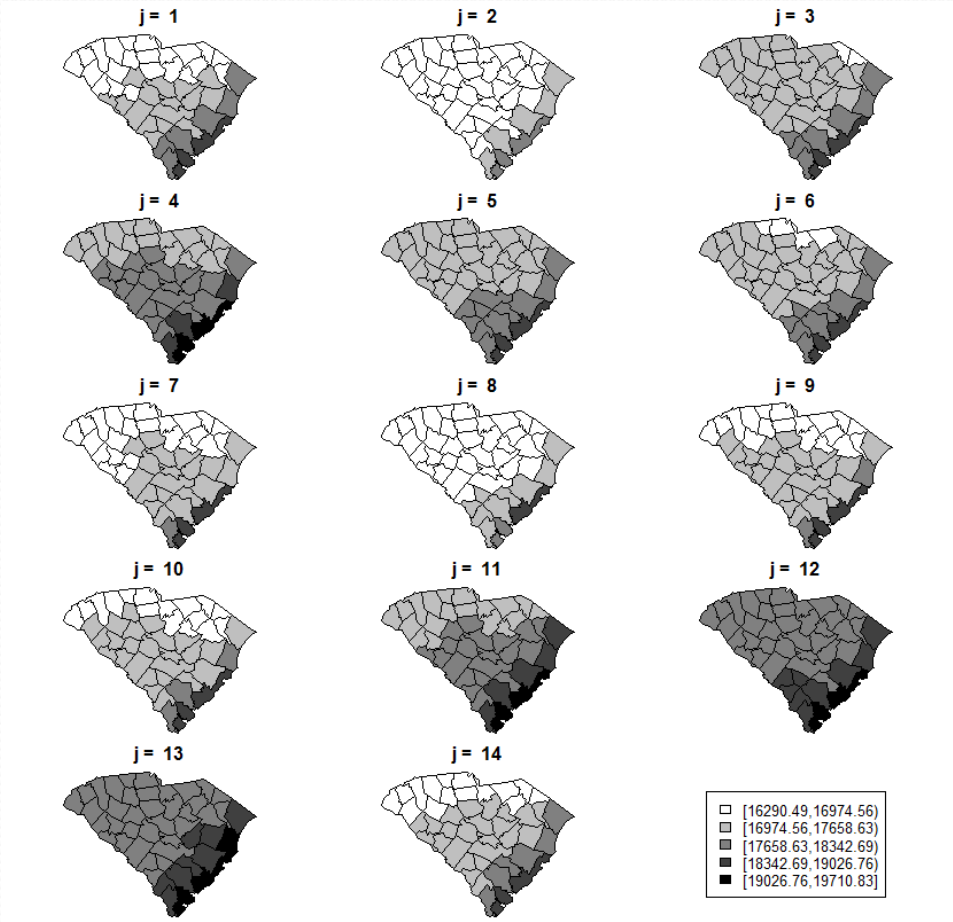


14 years of melanoma incidence in South Carolina (SIRs) 1996 - 2009



NCI Geospatial Approaches to Cancer Control
2016

Sunlight (average annual daily Kjl/m^2)



NCI Geospatial Approaches to Cancer Control 2016

Geospatial information is becoming much finer in resolution

- Google earth:
  - Geo-referencing of addresses is easy
  - street view is now used in studies of neighborhood characteristics
- Phone apps with GPS can be used to locate actions by participants in behavioral studies
  - Interventions
  - Crowd sourcing
- Portable personal air quality monitors
- Social media:
  - Twitter ?
  - Facebook?

Statistical Modeling issues

- Inferential issues must arise when assessing disease risk.
- Commercial GIS systems are limited in that they largely do not provide good flexible inferential tools, especially if temporal variation is to be studied.
- Modeling of risk is preferable to testing
- Why is testing not useful?
  - Limited focused application
  - cant be adjusted easily for predictor effects
  - Relies on large sample results
  - Lack of flexibility
  - Cant assess latent structure
- Estimation/modeling is better

Bayesian Modeling of diseased risk

- Bayesian approaches to geospatial modeling of risk are very flexible
- They assume a natural hierarchical model for risk
- They allow flexible specification of spatial correlation
- They allow complex structures to be included at different levels of the hierarchy
- Contextual effects can be easily incorporated within models
- Neighborhoods can be considered to be contextual

Bayesian Modeling of diseased risk

- Bayesian models allow the specification of prior distributions for parameters and hence allow great flexibility of both to allow a broad range of parameter values or a proscribed range.
- The incorporation of measurement error, missingness, regression variable selection and model selection, and joint modeling of dependencies between different diseases can be straight forwardly made.
- Latent unobserved structure can be modelled more easily in this context (SEMS, ME, hidden Markov models)

Multivariate modeling of small area cancer incidence: an example

- Lung and bronchus cancer (LBCa)
- Oral cavity and pharynx cancer (OCPCa)
- Melanoma (MCaS)
  - Some commonalities:
    - Behavioral impacts could be common for LBCa and OCPCa
    - Sunlight exposure could impact both MCaS and OCPCa
    - Unobserved/unknown etiological links?
  - Observed predictors
- Observed for 46 counties of South Carolina over 14 years

Predictors

- ST: average daily sunlight, unemployment rate , percent under poverty line
- S: proportion of population with health insurance, radon (county average in home), population percentage African American
- T: rainfall (annual average)

Risk models

- We assume that the log relative risk is to be modelled
- In previous work we have found that for spatio-temporal modeling mixtures of components are useful prescriptions:
- For k diseases:

$$y_{ijk} \sim Pois(e_{ijk}\theta_{ijk})$$

$$\log(\theta_{ijk}) = \alpha_k + \sum_h p_h M^h_{ijk}$$

$$\sum_h p_h = 1;\ 0 < p_h < 1$$

Components

- Spatial:

$$M_{ik}^{S} = X_i' \boldsymbol{\beta}_k^{S} + v_{ik} + u_i$$

- Spatio-temporal:

$$M_{ijk}^{ST} = X_{ij}' \boldsymbol{\beta}_{jk}^{ST} + X_j' \boldsymbol{\beta}_k^{T} + \gamma_j + \phi_{ijk}$$

- Assume sharing of uncorrelated spatial and random walk effects

Prior distributions

$$v_{ik} \sim N\left(0, \tau_{vk}^{-1}\right) \qquad u_i \sim ICAR(\tau_u^{-1}) \qquad \gamma_j \sim N\left(\gamma_{j-1}, \tau_\gamma^{-1}\right) \qquad \beta \sim N\left(0, \tau_\beta^{-1}\right)$$

Precision prior distributions: assumed sd-uniform prior distributions with fixed range i.e. $\tau_*^{-1/2} \sim U(0, C)$

These are reasonably weakly informative (c.f. PC priors)

Computational Considerations

- Model selection strategies can be computationally intensive.
- We considered working with Laplace approximation but it lacked the necessary flexibility.
- We consider McMC as a suitable computational paradigm.
- To reduce computation time, we used the `R` package `snowfall`.
- Currently exploring R-NIMBLE for speedups with CAR approximations

Fitted Models

Model	Definition	Formulation	Mixture parameter
F1	Uncorrelated linkage between the spatial and ST components of the models	$\log \theta_{ijk} = a_{0k} + p_{ik} M_{ik}^{S} + (1 - p_{ik}) M_{ijk}^{ST}$	$\text{logit}\, p_{ik} = z_{ik} + \alpha_{ik}$ $z_{ik} \sim Norm(0, \tau_{zk}^{-1})$ $\alpha_{ik} \sim Norm(0, \tau_{\alpha k}^{-1})$
F2	Spatial structure via a conditional autoregressive (CAR) distribution on the mixture parameter		$\text{logit}\, p_{ik} = z_{ik} + \alpha_{ik}$ $z_{ik} \sim CAR(\tau_{zk}^{-1})$ $\alpha_{ik} \sim Norm(0, \tau_{\alpha k}^{-1})$
F3	Mixture parameter varies across space and time, but the correlation remains only spatial.	$\log \theta_{ijk} = a_{0} + p_{ijk} M_{ik}^{S} + (1 - p_{ijk}) M_{ijk}^{ST}$	$\text{logit}\, p_{ijk} = z_{ijk} + \alpha_{ijk}$ $\alpha_{ijk} \sim Norm(0, \tau_{\alpha k}^{-1})$ $z_{ijk} \sim CAR(\tau_{zjk}^{-1})$
F4	Mixture parameter varies across space and time with spatial and temporal correlation.		$\text{logit}\, p_{ijk} = (z_{ijk} + w_{jk})/2 + \alpha_{ijk}$ $z_{ijk} \sim CAR(\tau_{zk}^{-1})$ $w_{jk} \sim RW(1)(\tau_{wjk}^{-1})$ $\alpha_{ijk} \sim Norm(0, \tau_{\alpha k}^{-1})$

Some fitted model variants with mixture sharing

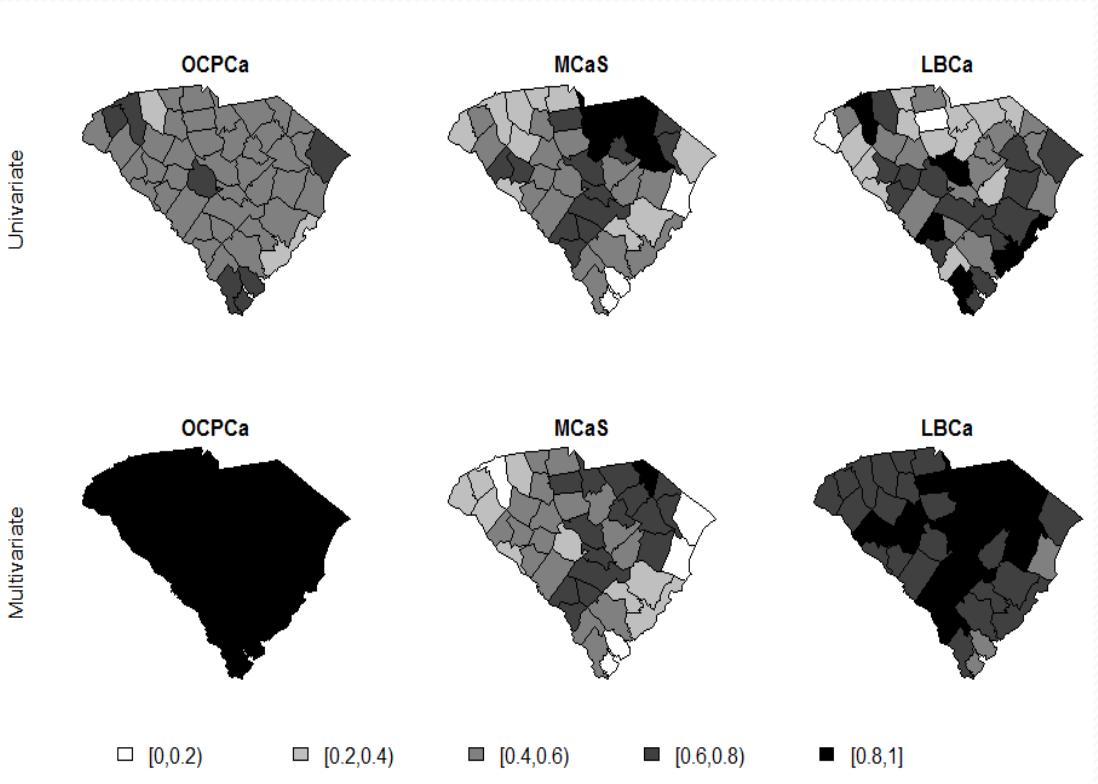
Disease	Fitted Model	Univariate		Multivariate	
		$WAIC$	pD_{WAIC}	$WAIC$	pD_{WAIC}
All	F1	12295.14	824.78	11780.97	477.71
	F2	11413.71	387.91	**11749.35**	457.21
	F3	11448.8	449.03	11764.26	541.69
	F4	**11404.57**	395.76	11782.37	493.18
	KH	11665.43	518.40	11832.07	560.68
OCPCa	F1	3212.38	102.74	3450.55	131.32
	F2	3204.27	95.03	**3446.10**	130.56
	F3	**3196.91**	101.39	3471.25	163.55
	F4	**3197.54**	95.96	3604.93	201.53
	KH	3223.60	105.27	3482.81	163.04
MCaS	F1	3850.89	195.90	3815.31	182.49
	F2	**3785.76**	165.16	3848.12	190.28
	F3	3815.19	202.68	3846.18	218.29
	F4	3791.33	175.25	**3757.62**	160.52
	NK	3941.36	251.21	3857.35	233.16
LBCa	F1	5231.87	526.14	4515.12	163.90
	F2	4423.68	127.72	4455.13	136.38
	F3	4436.70	144.96	4446.84	159.85
	F4	**4415.70**	124.55	**4419.83**	131.13
	KH	4500.47	161.92	4491.91	164.48

Notes

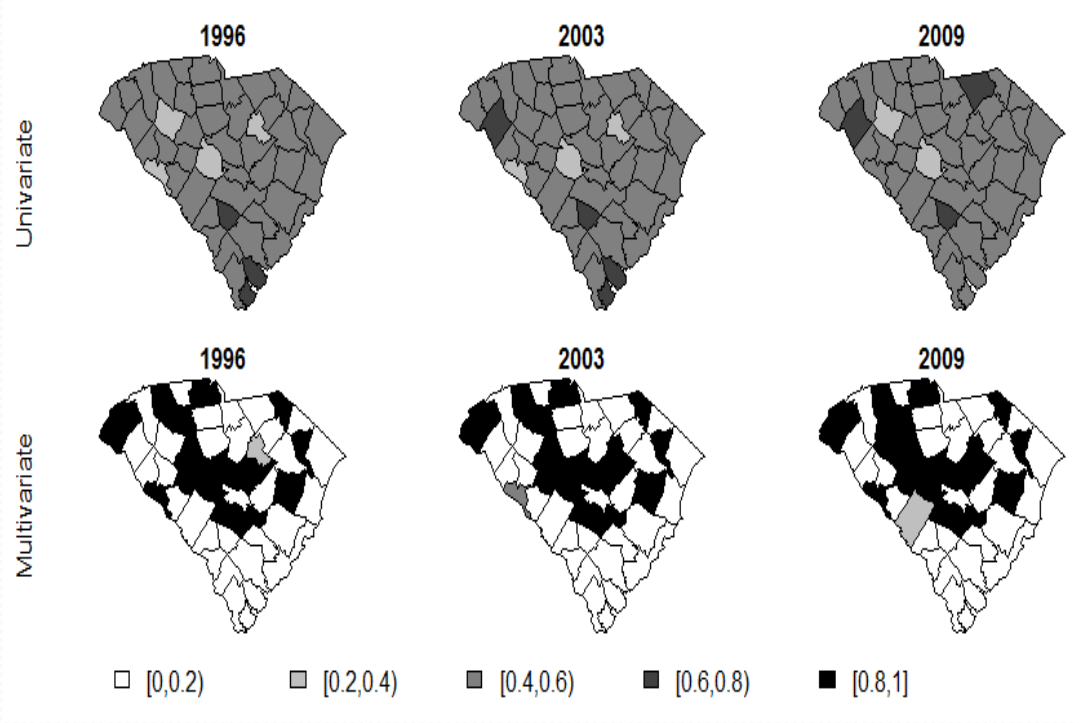
- Conventional KH model never favored for these data either as univariate or multivariate fits.
- Lung and bronchus favors a more complex F4 model
- Melanoma also favors F4 for the multivariate (which bests the univariate fit)
- Oral cavity and pharyngial is different in that it shows F2 is favored for the multivariate fit where there is no temporal mixing dependence. There is in fact limited temporal variation in this disease.

Mixture probability estimates for F2

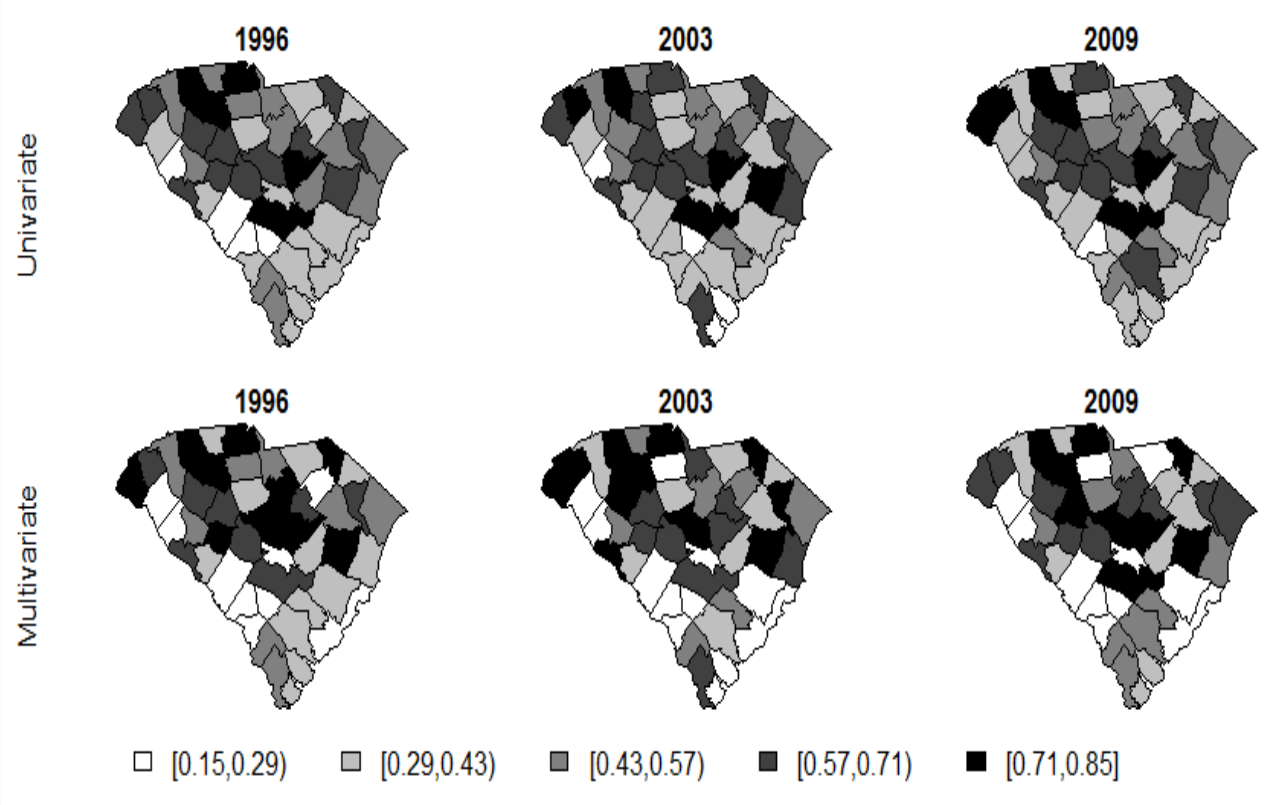
- Spatial only mixing
  - Relatively uniform for OCPCa
  - MCaS least affected across U or M
  - LBCa: little change in multivariate but more variable as a univariate



Univariate and multivariate OCPCa F4 mixture parameter estimates for the first, a central, and the last years.

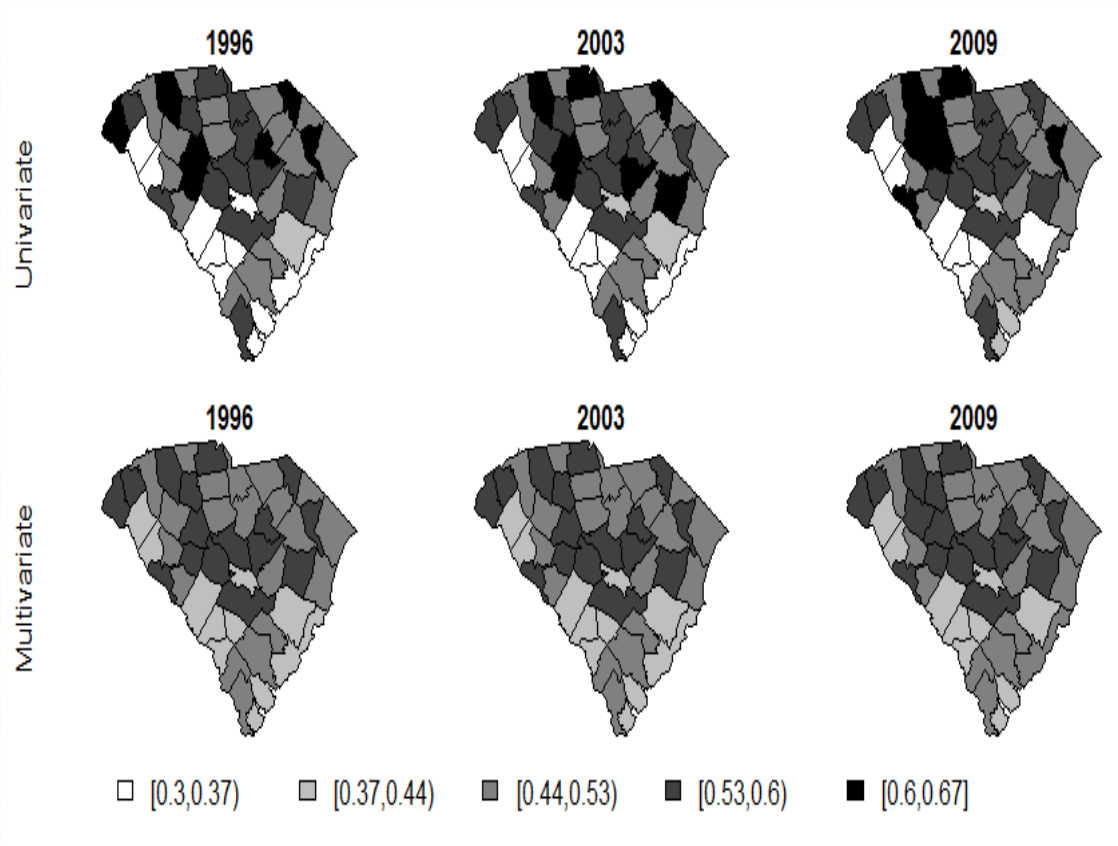


MCaS F4 fit: mixture parameter estimates



LBCa F4 fit: mixture parameter estimates

- LBCa: univariate and multivariate different.
  - Sharing impacts more



Predictor model comparison

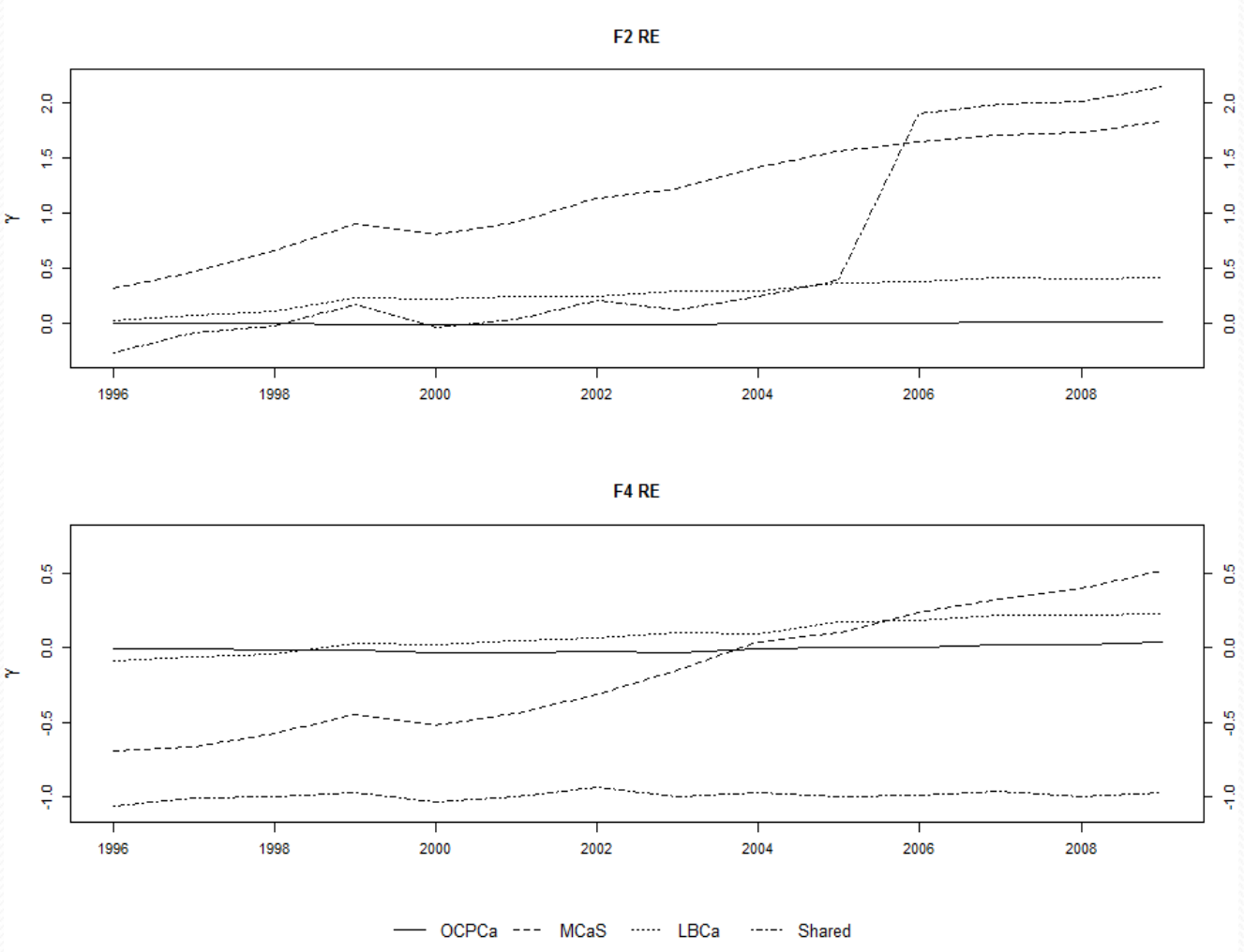
Model			Univariate		Multivariate	
			$WAIC$	pD_{WAIC}	$WAIC$	pD_{WAIC}
Knorr-Held		All	11665.43	518.40	11832.07	560.68
		OCPCa	3223.60	105.27	3482.81	163.04
		MCaS	3941.36	251.21	3857.35	233.16
		LBCa	4500.47	161.92	4491.91	164.48
F1	RE	All	12295.14	824.78	11780.97	477.71
		OCPCa	3212.38	102.74	3450.55	131.32
		MCaS	3850.89	195.90	3815.31	182.49
		LBCa	5231.87	526.14	4515.12	163.90
	PRED	All	11639.88	441.45	11846.23	489.72
		OCPCa	3239.50	90.27	3436.49	135.27
		MCaS	3889.96	206.49	3871.37	185.06
		LBCa	4510.42	144.69	4538.36	169.38
F2	RE	All	11413.71	387.91	**11749.35**	457.21
		OCPCa	3204.27	95.03	3446.10	130.56
		MCaS	**3785.76**	165.16	3848.12	190.28
		LBCa	4423.68	127.72	4455.13	136.38
	PRED	All	11520.61	443.91	11885.63	487.34
		OCPCa	3241.74	105.89	**3431.50**	131.08
		MCaS	3799.99	185.99	3959.95	203.01
		LBCa	4478.88	152.03	4494.18	153.26
F3	RE	All	11448.8	449.03	11764.26	541.69
		OCPCa	**3196.91**	101.39	3471.25	163.55
		MCaS	3815.19	202.68	3846.18	218.29
		LBCa	4436.70	144.96	4446.84	159.85
	PRED	All	11453.95	469.19	11797.07	565.01
		OCPCa	3211.40	109.45	3507.38	179.42
		MCaS	3806.47	200.19	3838.37	215.73
		LBCa	4436.08	159.55	4451.33	169.86
F4	RE	All	**11404.57**	395.76	11782.37	493.18
		OCPCa	**3197.54**	95.96	3604.93	201.53
		MCaS	3791.33	175.25	**3757.62**	160.52
		LBCa	**4415.70**	124.55	**4419.83**	131.13
	PRED	All	11458.04	447.79	11939.70	591.29
		OCPCa	3239.07	112.20	3732.72	263.59
		MCaS	3793.16	193.44	3761.95	177.77
		LBCa	4435.81	142.15	4445.03	149.93

Predictor versus REs

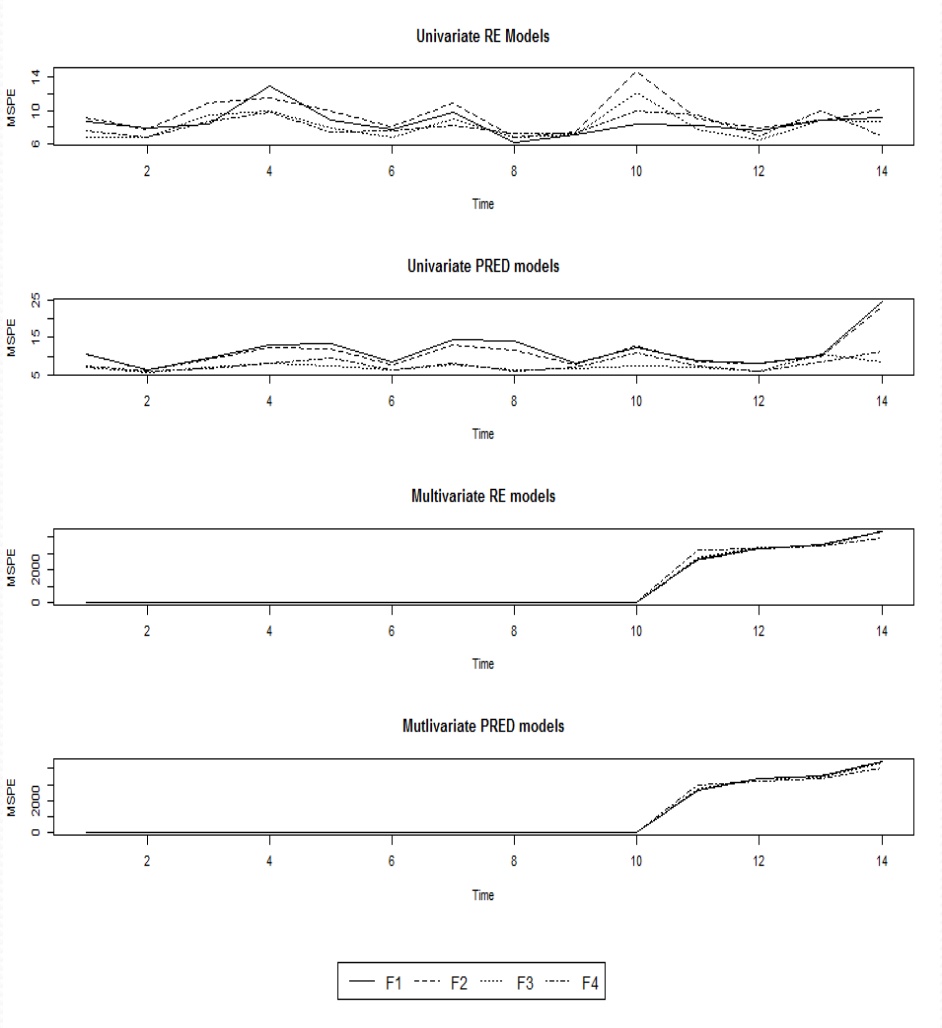
- Overall there are few situations where the predictor models acheive better performance than RE models
  - Exception is OCPCa for F2 where the multivariate model fits better with predictors.
- Other results: both LBCa and MCaS have better RE fits although for MCaS there is equivocal results.
- Notable that over time the OCPCa is relatively constant but the other diseases show some increasing trend
  - This leads to a jump in the shared component around 2005

Temporal Profiles: multivariate models

Temporal random walk profiles for the 3 diseases and shared component under the multivariate model with fitted models F2 and F4. The jump in the shared component is largely due to



MSPE results for each year for OCPCa



Epidemiologic conclusions

- OCPCa: for univariate F_3 or F_4, while simpler F_2 favored for mulitvariate
- LBCa: F_4 was strongly favored for univariate and multivariate
- MCaS: F_2 (univariate) and F_4 (multivariate)

- Overall:
  - OCPCa does not vary temporally and the mixing parameters show limited spatial structure, whereas the other diseases show differences.
  - Only for OCPCa do the predictor models fit better than RE models
  - Shared models perform well for LBCa and MCaS, but the sharing displays a jump due to the differences in temporal behavior of the OCPCa and the other cancers

Melanoma and sunlight

- Can we assess the ecological association between MCaS and sunlight ?
- We have 14 years of melanoma incidence in counties of SC (46) and
- Sunlight intensity (kjl/m^2) in counties over years
- We can assume a spatial only or a spatio-temporal relationship
- We will allow for unobserved confounding with random effects

Models

- Spatial

$$\log(\theta_i) = \beta_0 + v_i + u_i \text{ or}$$

$$\beta_0 + \beta_{sun} x_i + v_i + u_i$$

- Spatio-temporal

$$\log(\theta_{ij}) = \beta_0 + \beta_{sun} x_i + v_i + u_i + \gamma_j + \psi_{ij}$$

Results

Model	DIC	pD	WAIC	wW
Spatial (2004)				
BYM	286.4	24.65	287.5	18.56
1+sun+BYM	284.2	23.79	284.0	17.17
β_{sun} (median)	0.0004	(0.0001, 0.0006)		
ST models				
1996-2014				
Knorr-Held*	3865.5	294.8	3847.5	212.8
+sun	3864.8	288.8	3854.6	214.8
β_{sun} (median)	0.0001	(0.0001, 0.0003)		
-sun - interaction	4278.8	58.08	4380.79	141.16
+sun -interaction	4279.3	57.67	4389.54	146.25
β_{sun} (median)	0.0000	(-0.0001, 0.0001)		

Notes

- Inclusion of sunlight does show some marginal significance
- Temporal range can affect these marginal effect estimates
- Interactions are important
  - Removal decreases the goodness of fit by a large amount
- Also sunlight fails to be significant when interaction is removed
  - Increased noise tends to mask the significance

Public Health impact

- Public health is carried out in a geo-spatial context
  - Resource allocation based on risk estimates
  - Analysis of local clustering
  - Local environmental exposure links to disease outcomes.

Cancer Control and Prevention?

- Interventions can be targeted geo-spatially
  - Important if neighborhood effects are important
- Health Services can be targeted at catchments defined by risk
  - Disease clusters can be addressed
- Resource allocation is usually based on SIRs for different regions and so directly relates to cancer risk
- Prevention:
  - Targeted to populations at most risk which can be geo-spatially defined