Hierarchical Modeling for Large Univariate Areal Data

Abhi Datta

Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland

abhidatta.com @dattascience

Areal data

County	Cases	Deaths
Allegany	6,432	200
Anne Arundel	36,655	537
Baltimore City	41,282	885
Baltimore County	52,271	1273
Calvert	3,736	72
Caroline	2,063	21
Carroll	7,736	210
Cecil	5,080	122
Charles	9,158	159
Dorchester	2,382	44
Frederick	17,324	278
Garrett	1,854	60
Harford	12,864	241
Howard	16,326	214
Kent	1,144	41
Montgomery	64,399	1392
Prince George's	74,851	1322
Queen Anne's	2,607	38
St. Mary's	5,248	116
Somerset	2,414	34
Talbot	1,911	35
Washington	12,599	256
Wicomico	6,859	145
Worcester	3,295	91

Figure: COVID-19 cases and deaths in Maryland counties[1]

- Each datapoint is associated with a region like state, county, municipality etc.
- Usually a result of aggregating point level data

[1] https://coronavirus.maryland.gov/#Vaccine

Areal data analysis

- Visualization using chloropleth maps
- Exploratory measures of spatial association (Moran's I)
- Spatial disease mapping
  - Identify factors (covariates) associated with the disease
  - Identify spatial pattern, if any, and smooth spatially
  - Inference is often restricted only to the given set of regions

Chloropleth maps using `sf` package

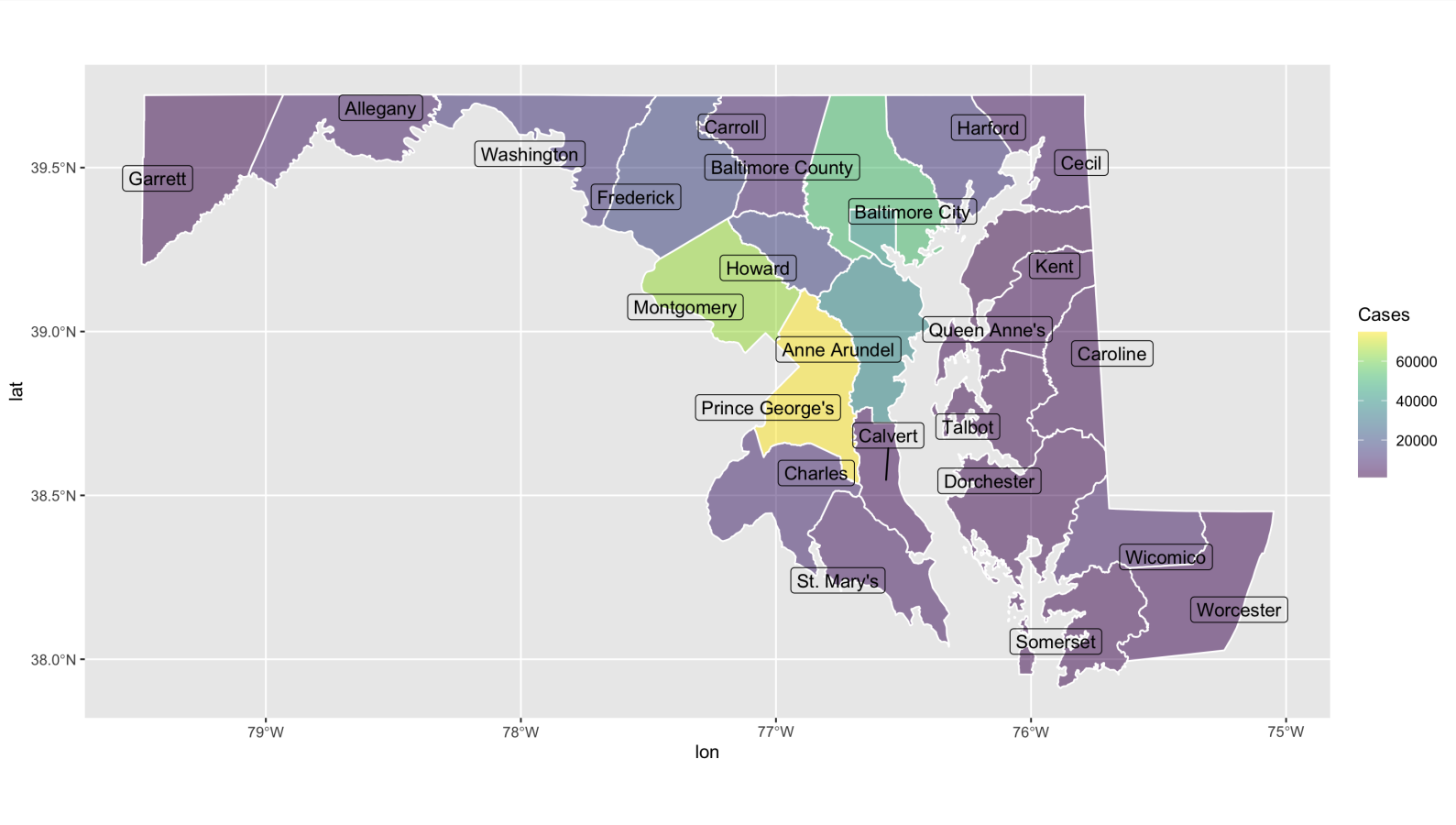


Figure: COVID-19 cases in Maryland counties

Chloropleth maps using `sf` package

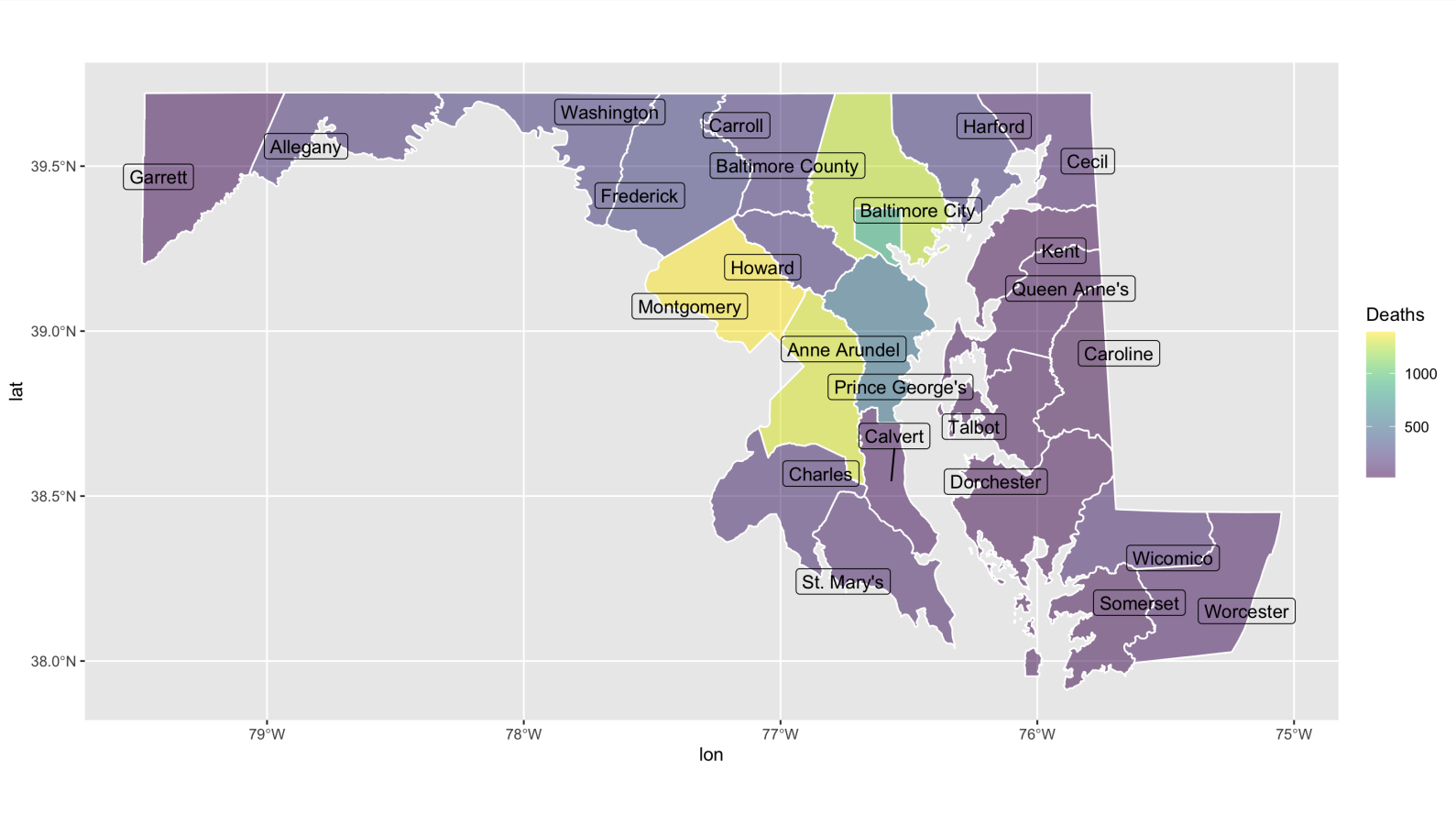


Figure: COVID-19 deaths in Maryland counties

Moran's I

- Areal data: Outcome y_i for region i, $i = 1, \ldots, k$
- To investigate spatial patterns in areal data we need some notion of *distance or proximity* between units
- Let $W = (w_{ij})$ denote some proximity matrix.
- Some possible choices for W
  - $w_{ij} = 1$ iff regions i and j share a border
  - w_{ij} inversely related to intercentroidal distance
  - $w_{ij} = 1$ iff unit j is one of K-nearest neighbors of i (in terms of the intercentroidal distances)
  - Intercentroidal distances use representation of the whole area by a single point
  - Nearest-neighbors will lead to an asymmetric W
  - w_{ii} is customarily set to be 0

Moran's I

$$I = \frac{\frac{1}{\sum_{i \neq j} w_{ij}} \sum_{i \neq j} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\frac{1}{n}\sum_i (y_i - \bar{y})^2}$$

- Spatial analog of lagged autocorrelation coefficient for time-series
- Example: If $w_{ij} = 1$ iff regions i and j share a border, I gives the spatial correlation coefficient between neighboring regions (analogous to autocorrelation coefficient at lag 1)
- If Y_i's ar iid (i.e., no spatial correlation), then $E(Y_i) = -\frac{1}{n-1}$ asymptotically (Agresti, 2002, Chap 14)
- Higher value of I corresponds to high correlation (but I does not strictly lie in $[-1, 1]$)
- `moran.test` function of the `spdep` package]calculates I

Spatial disease mapping

- At unit (region) i, we observe response y_i and covariate x_i, for $i = 1, \ldots, k$
- For continuous (normally-distributed) data, spatial linear mixed model $y_i = x_i'\beta + w_i + \epsilon_i$
- w_i is some spatially smooth random effect (or function on the areal space)
- $\epsilon_i \overset{\text{iid}}{\sim}$ normal errors
- Penalized least squares $\arg\min_{\beta,w} \sum_i \|y_i - x_i'\beta - w_i\|^2 + P(w)$ for some roughness penalty P for $w = (w_1, w_2, \ldots, w_k)'$
- Once again $P(w)$ will rely on some notion of spatial proximity on the areal space

Spatial disease mapping

- Penalized least squares $\arg\min_{\beta,w} \sum_i \|y_i - x_i'\beta - w_i\|^2 + P(w)$ for some roughness penalty P for w
- Common choice $P(w) = \tau_w w' Q w$ for some smoothing matrix Q
- Equivalent to
  $\arg\max_{\beta,w} \prod_{i=1}^n N(y_i \mid x_i'\beta + w_i) \times N(w \mid 0, \tau_w^{-1} Q^{-1})$
- τ_w controls the degree of regularization
- Q can also be parameterized, i.e., $Q = Q(\rho)$ for some parameter(s) ρ

GLM for Spatial disease mapping

- Disease data is often presence-absence (binary) or counts
- $g(E(y_i)) = x_i'\beta + w_i$ where $g(\cdot)$ denotes a suitable link function

Hierarchical areal model:

$$\prod_{i=1}^{k} p_1(y_i|x_i'\beta + w_i) \times N^{-1}(w \,|\, 0, \tau_w Q(\rho)) \times p_2(\beta, \tau_w, \rho)$$

- Notation: $N^{-1}(m, Q)$ denotes normal distribution with mean m and precision (inverse covariance) Q
- p_1 denotes the functional form of the density corresponding to the link $g(\cdot)$

How to model $Q(\rho)$

- Choice of $Q(\rho)$ should enable spatial smoothing
- One possibility: Represent each region by a single point and use Gaussian Process covariance i.e.
  $Q(\rho)_{ij}^{-1} = C(m(i), m(j) \mid \rho)$
- Many possible choices to map the region i into a Euclidean coordinate $m(i)$
- Is it appropriate to represent a large area with a single point?
- Also GP approach is computationally very expensive
- Alternate approach: Represent spatial information in terms of a graph depicting the relative orientation of the regions

CAR models

- Conditional autoregressive (CAR) model (Besag, 1974; Clayton and Bernardinelli, 1992)
- Areal data modeled as a graph or network: V is the set of vertices (regions)
- $i \sim j$ if regions i and j share a common border
- CAR model: $P(w) = \tau_w \sum_{i \sim j} (w_i - w_j)^2$
- Penalizes large differences in w between neighboring regions

CAR models

- Adjacency matrix $A = (a_{ij})$ such that $a_{ij} = I(i \sim j)$
- $D = \text{diag}(n_1, \ldots, n_k)$ where n_i is the number of neighbors of i
- $\tau_w \sum_{i \sim j} (w_i - w_j)^2 = \tau_w w' Q w$ where $Q = D - A$
- CAR prior: $w \sim N^{-1}(0, \tau_w Q)$
- Equivalent to the full conditionals:

$$w_i \mid w_{-i} \sim N^{-1}\left(\frac{1}{n_i} \sum_{j \mid i \sim j} w_j, \tau_w n_i\right)$$

CAR models

- ICAR prior is an improper distribution as $(D - A)1 = 0$ (ICAR) (Improper or intrinsic CAR)
- ICAR can be still used as a prior for random effects
- Spectral decomposition: $Q = P\Lambda P'$ where $P = (1, p_1, \ldots, p_k)$ are eigen-vectors of Q, $\Lambda = \text{diag}(0, \lambda_2, \ldots, \lambda_k)$ are eigen-values
- If $z = P'w$ or $w = Pz$ we have $w'Qw = z'\Lambda z$ and

$$g(E(y)) = X\beta + w, P(w) = w'Qw \equiv$$

$$g(E(y)) = X\beta + Pz, P(z) = \sum_{i=2}^{k} \lambda_i z_i^2 \equiv$$

$$g(E(y)) = X\beta + 1z_1 + \sum_{i=2}^{k} p_i z_i, P(z) = \sum_{i=2}^{k} \lambda_i z_i^2$$

- w needs to be centered if using an intercept in the model

Proper CAR

- Proper CAR model $Q(\rho) = D - \rho A$
  - $\rho = 1 \Rightarrow$ ICAR
  - $\rho < 1 \Rightarrow$ Proper distribution with added parameter flexibility
- Equivalent to the full conditionals:

$$w_i \mid w_{-i} \sim N^{-1}\left(\frac{\rho}{n_i} \sum_{j \mid i \sim j} w_j, \tau_w n_i\right)$$

CAR troubles

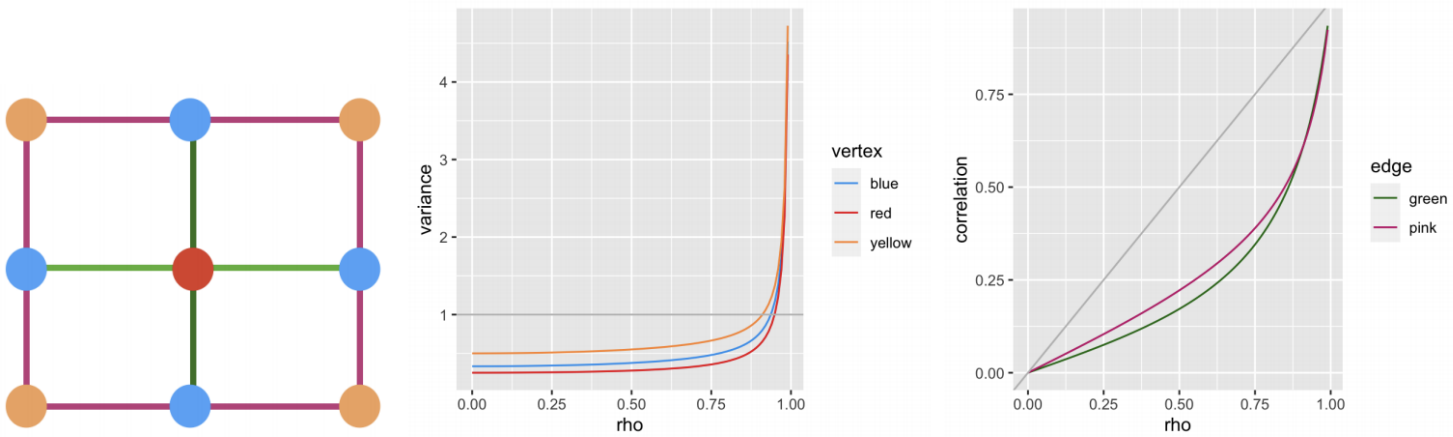


Figure 3: Unequal variances and correlations in the CAR model: (left) A 3×3 grid graph with vertices and edges grouped and colored by symmetry, (middle) CAR-induced variances for each vertex group as a function of ρ, (right) CAR-induced neighbor-pair correlations for each edge group as a function of ρ.

SAR models

- Simultaneous Autoregressive (SAR) model (Whittle, 1954)
- Instead of taking the conditional route, SAR model proceeds by simultaneously modeling the random effects

$$w_i = \rho \sum_{i \neq j} b_{ij} w_j + \epsilon_i \text{ for } i = 1, 2, \ldots, k$$

- $\epsilon_i \overset{ind}{\sim} N^{-1}(0, \tau_i)$ are errors independent of w
- A common choice is to define $b_{ij} = I(i \sim j)/n_i$
- Joint distribution: $w \sim N^{-1}(0, (I - \rho B)' F (I - \rho B))$, $B = (b_{ij})$ and $F = diag(\tau_1, \tau_2, \ldots, \tau_k)$
- $\rho = 1 \Rightarrow$ Improper distribution

Interpretation of ρ in proper CAR and SAR models

- Calibration of ρ as a correlation, e.g., (as reported in Banerjee et al. 2014)

$$\rho = 0.80 \text{ yields } 0.1 \leq \text{Moran's } I \leq 0.15,$$
$$\rho = 0.90 \text{ yields } 0.2 \leq \text{Moran's } I \leq 0.25,$$
$$\rho = 0.99 \text{ yields Moran's } I \leq 0.5$$

- So, used with random effects, scope of spatial pattern may be limited

Interpretation of ρ in proper CAR and SAR models

- ρ cannot be interpreted as correlation between neighboring w_i's (Wall, 2004; Assuncao and Krainski, 2009)

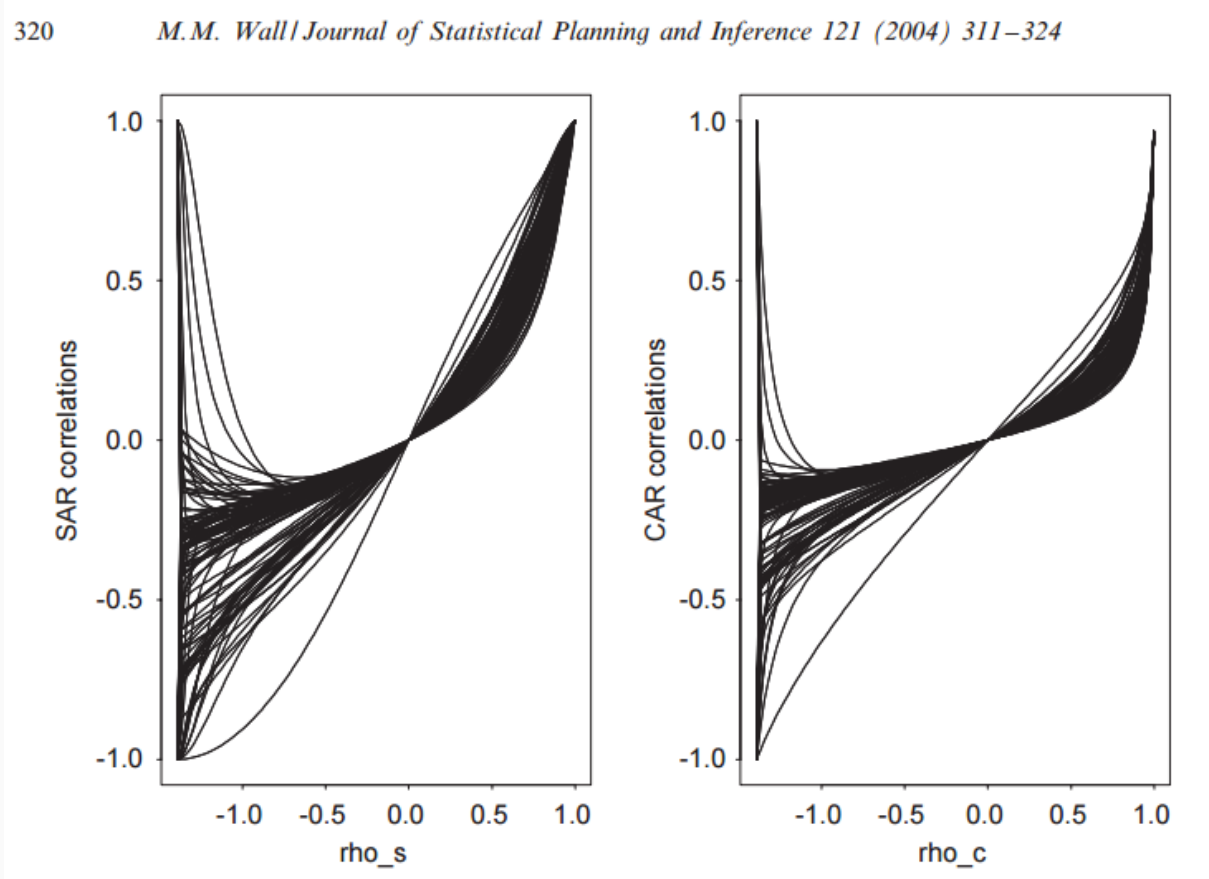


Figure: Neighbor pair correlations as a function of ρ for proper CAR and SAR models over the graph of US states

SAR model and Cholesky factors

- General SAR model:

$$w_i = \sum_{i \neq j} b_{ij} w_j + \epsilon_i \text{ for } i = 1, 2, \ldots, k$$

- $w \sim N^{-1}(0, (I - B)' F (I - B))$ where $F = diag(\tau_1, \tau_2, \ldots, \tau_k)$
- Only proper when $I - B$ is invertible which is not guaranteed for arbitrary B
- SAR is essentially modeling the precision matrix through the Cholesky factor $I - B$

SAR model and Cholesky factors

- General SAR model:

$$w_i = \sum_{i \neq j} b_{ij} w_j + \epsilon_i \text{ for } i = 1, 2, \ldots, k$$

- $w \sim N^{-1}(0, (I - B)' F (I - B))$ where $F = diag(\tau_1, \tau_2, \ldots, \tau_k)$
- Only proper when $I - B$ is invertible which is not guaranteed for arbitrary B
- SAR is essentially modeling the precision matrix through the Cholesky factor $I - B$
- Cholesky factors are not unique
- We can always choose a lower triangular Cholesky factor

New model

$$
\begin{aligned}
w_1 &= \epsilon_1 \\
w_2 &= b_{21} w_1 + \epsilon_2 \\
w_3 &= b_{31} w_1 + b_{32} w_2 + \epsilon_3 \\
&\vdots \\
w_k &= b_{k1} w_1 + b_{k1} w_2 + \ldots + b_{k,k-1} w_{k-1} + \epsilon_k
\end{aligned}
$$

- $B = (b_{ij})$ is now a strictly lower triangular matrix.

New model

- Advantages of lower triangular B:
  - $w \sim N^{-1}(0, (I-B)'F(I-B))$ is a proper distribution for any choice of lower triangular B
  - $\det(L'FL) = \prod_{i=1}^{n} \tau_i$ where $F = diag(\tau_1, \ldots, \tau_k)$ and $L = I - B$
  - $w'L'FLw = \tau_1 w_1^2 \ + \sum_{i=2}^{k} \tau_i (w_i - \sum_{\{j<i\}} w_j b_{ij})^2$
  - Likelihood $N^{-1}(w \,|\, 0, (I-B)'F(I-B))$ can be computed using $O(k+s)$ flops where s denotes the sparsity (number of non-zero entries) of B.
  - Even if k is large, evaluation of likelihood is fast if each region only shares border with a few others

Choice of B and F

- How to specify B and F?
- Sparsity of B is desirable
- If data had replicates for each region, there is large literature on fully data driven estimation of sparse Cholesky factors (Wu and Pourahmadi, 2003; Huang et al., 2006; Rothman et al., 2008; Levina et al., 2008; Wagaman and Levina, 2009; Lam and Fan, 2009)
- Unfortunately many areal datasets lack replication

Choice of B and F

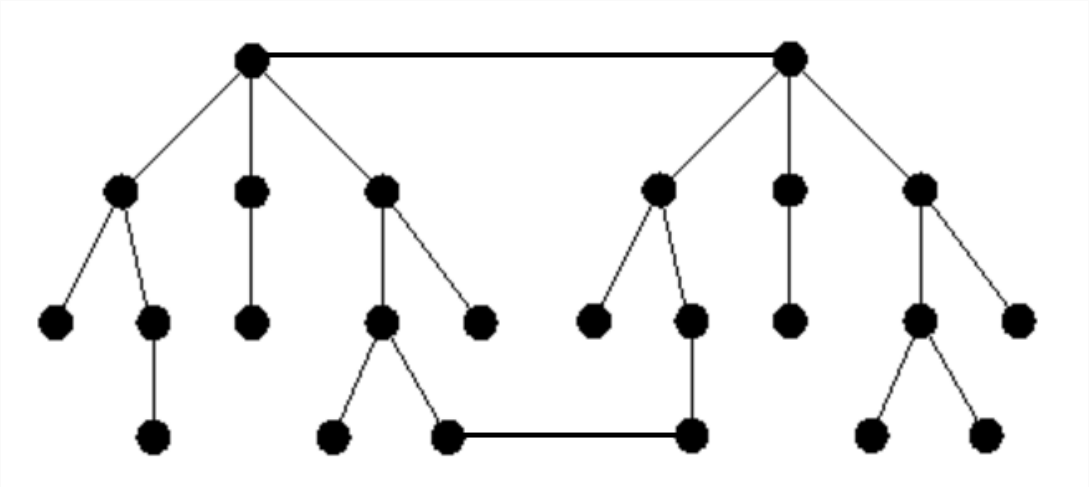
- How to specify B and F?
- Sparsity of B is desirable
- Like in NNGP set $b_{ij} = 0$ for j outside neighbor sets $N(i)$
  - Pros: For graphs neighbor sets are naturally chosen: $N(i) = \{j \,|\, j \sim i, j < i\}$
  - Cons: There is no covariance function on arbitrary graphs from which we can obtain non-zero b_{ij}'s and F

Autoregressive models on trees

- $D = (d_{ij})$ is the shortest distance matrix on the graph
- If the graph was a tree (no loops), then $\rho^D = (\rho^{d_{ij}}$ is then a valid *autoregressive* correlation matrix (AR(1) model on a tree, Basseville et al., 2006).
- Areal graphs are loopy and are not usually trees

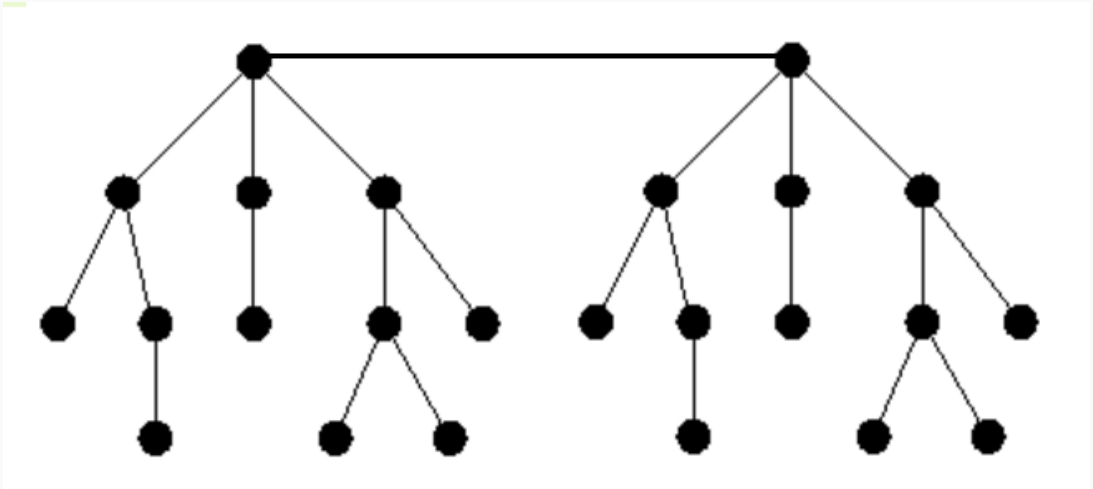
Embedded spanning trees

- EST of a graph $\mathscr{G}$ is a subgraph of $\mathscr{G}$ which is a tree and spans all the vertices of $\mathscr{G}$
- Sudderth (2002) approximated covariances on loopy graphs with spanning trees
- Many choices of EST's; finding best one is computationally infeasible
- EST's will leave out edges of $\mathscr{G}$ and can produce large errors (Sudderth, 2002)



Embedded spanning trees

- EST of a graph $\mathscr{G}$ is a subgraph of $\mathscr{G}$ which is a tree and spans all the vertices of $\mathscr{G}$
- Sudderth (2002) approximated covariances on loopy graphs with spanning trees
- Many choices of EST's; finding best one is computationally infeasible
- EST's will leave out edges of $\mathscr{G}$ and can produce large errors (Sudderth, 2002)



Local embedded spanning trees

- Embedded spanning trees (EST) of a graph G is a subgraph of G which is a tree and spans all the vertices of G
- Note that to specify $w_i = \sum_{j \in N(i)} b_{ij} w_j + \epsilon_i$ we only need a joint distribution on $\{i\} \cup N(i)$
- Let G_i denote the subgraph of G which includes vertices $\{i\} \cup N(i)$ and the edges among them
- The subgraph T_i of G_i which only contains the edges $\{i \sim j \,|\, j \in N(i)\}$ is an embedded spanning tree of G_i
- Use the local embedded spanning trees T_i to specify the b_{ij}'s and τ_i

Directed acyclic graph autoregressive (DAGAR) model[2]

- AR_i denotes the $AR(1)$ distribution on T_i
- Solve for b_{ij} and τ_i such that $E_{AR_i}(w_i \mid w_{N(i)}) = \sum_{j \in N(i)} b_{ij} w_j$ and $\tau_i = 1/Var_{AR_i}(w_i \mid w_{N(i)})$
- No edge is left out !

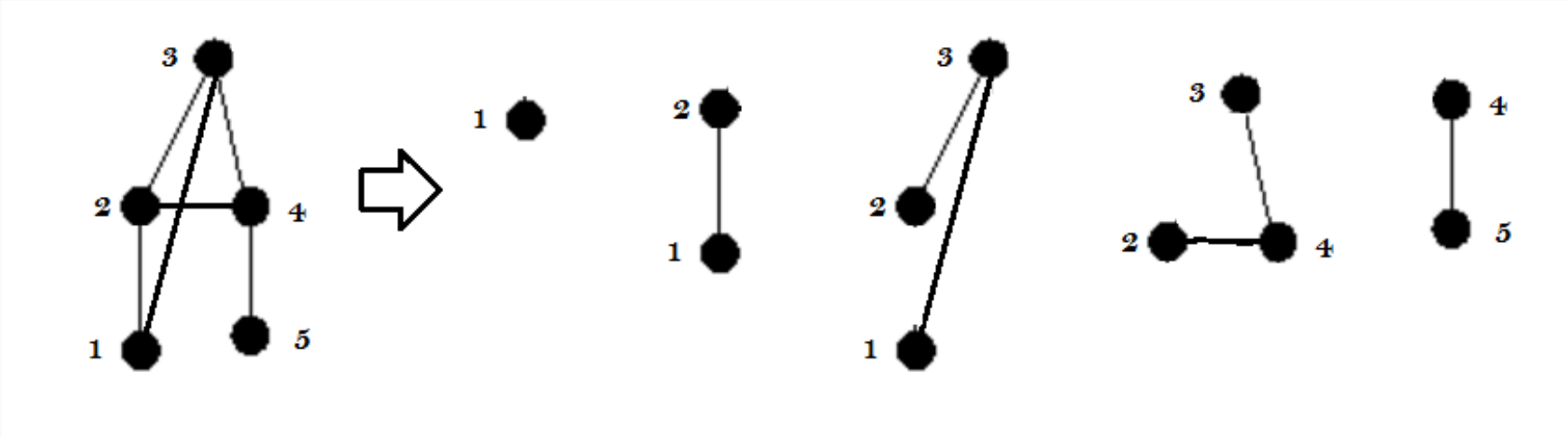


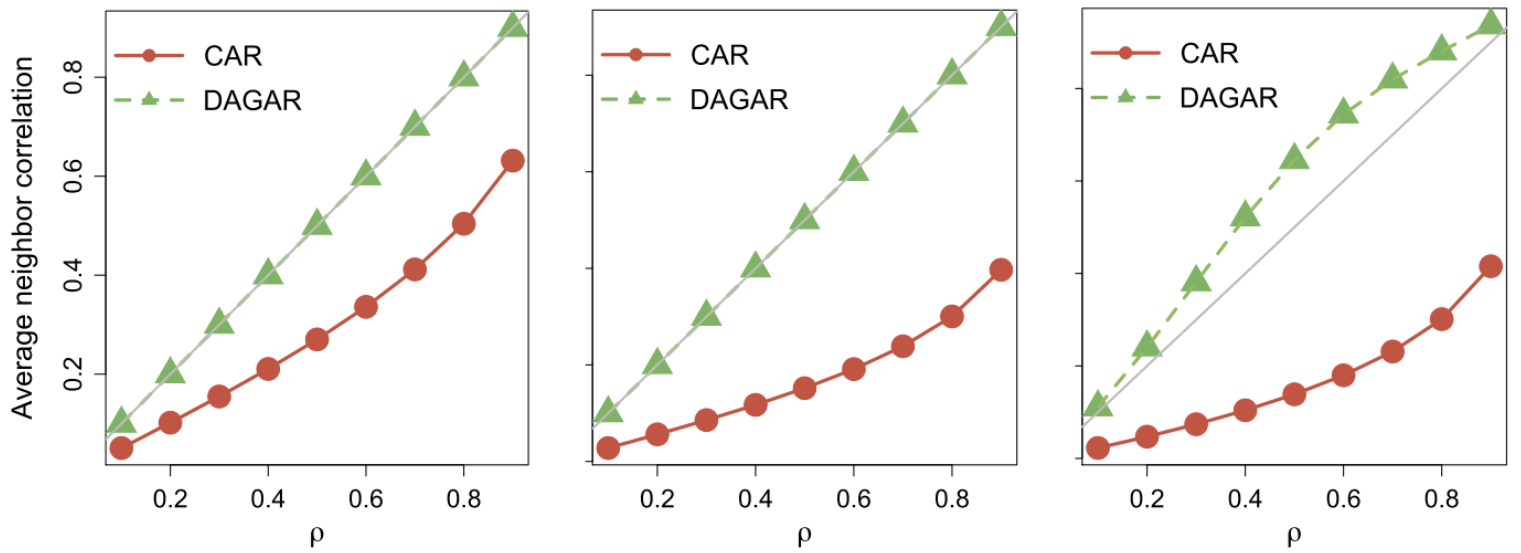
Figure: Decomposing a graph into a sequence of embedded spanning trees

[2]Datta et al. (2019), Spatial Disease Mapping Using Directed Acyclic Graph Auto-Regressive (DAGAR) Models, Bayesian Analysis 14(4): 1221-1244

Properties of DAGAR models

- $b_{ij} = b_i = \rho/(1 + (|N(i)| - 1)\rho^2)$
- $\tau_i = (1 + (|N(i)| - 1)\rho^2)/(1 - \rho^2)$
- $\det(Q_{DAGAR}) = \prod_{i=1}^{k} \tau_i$
- Positive definite for any $0 \leq \rho \leq 1$
- Interpretability of ρ:
  - If the graph is a tree, then DAGAR model is same as the AR(1) model on the tree i.e. correlation between d^{th} order neighbors is ρ^d for $d = 1, 2, \ldots$
  - If the graph is a closed two-dimensional grid, then each neighbor pair correlation is ρ
- $p_{DAGAR}(w)$ can be stored and evaluated using $O(e + k)$ flops where e is the total number of neighbor pairs

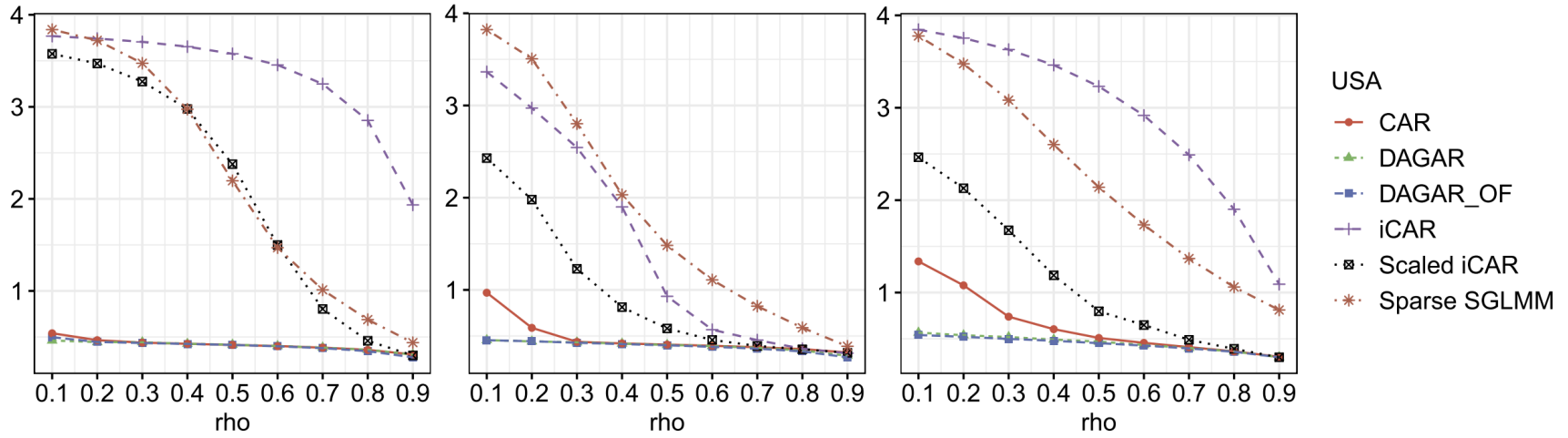
Interpretation of ρ



(a) Path graph of length 100 (b) 10×10 grid (c) 48 contiguous US states

Figure 1: Average neighbor pair correlations as a function of ρ for proper CAR and DAGAR model. The solid gray line represents $x = y$ line.

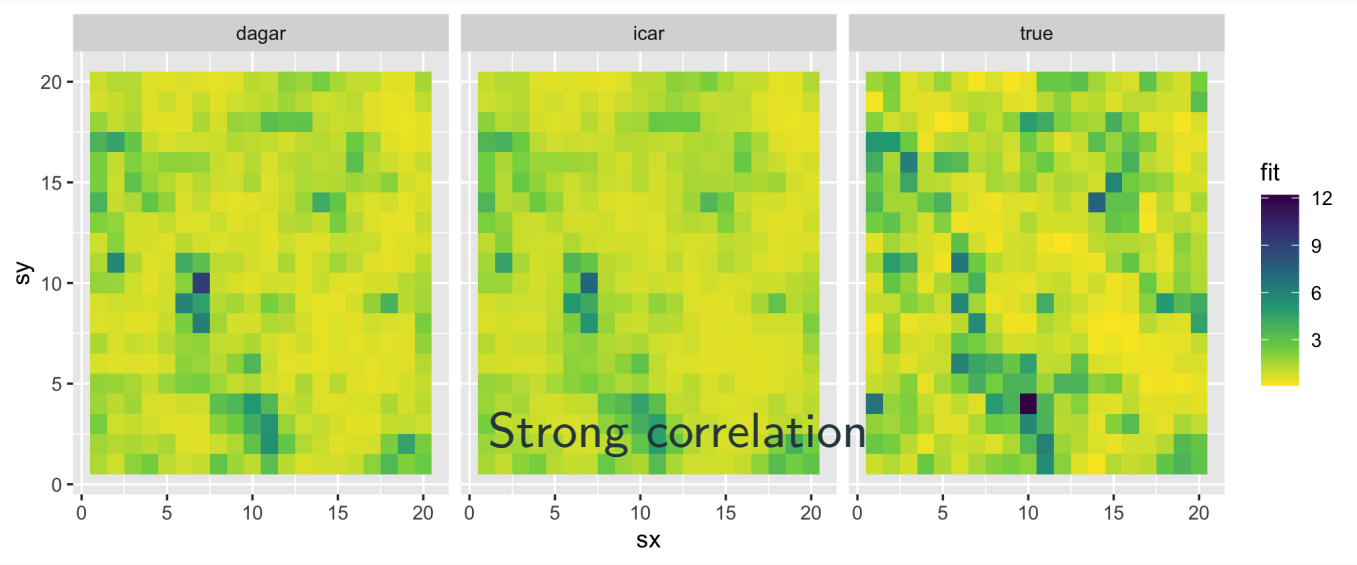
Estimation of w

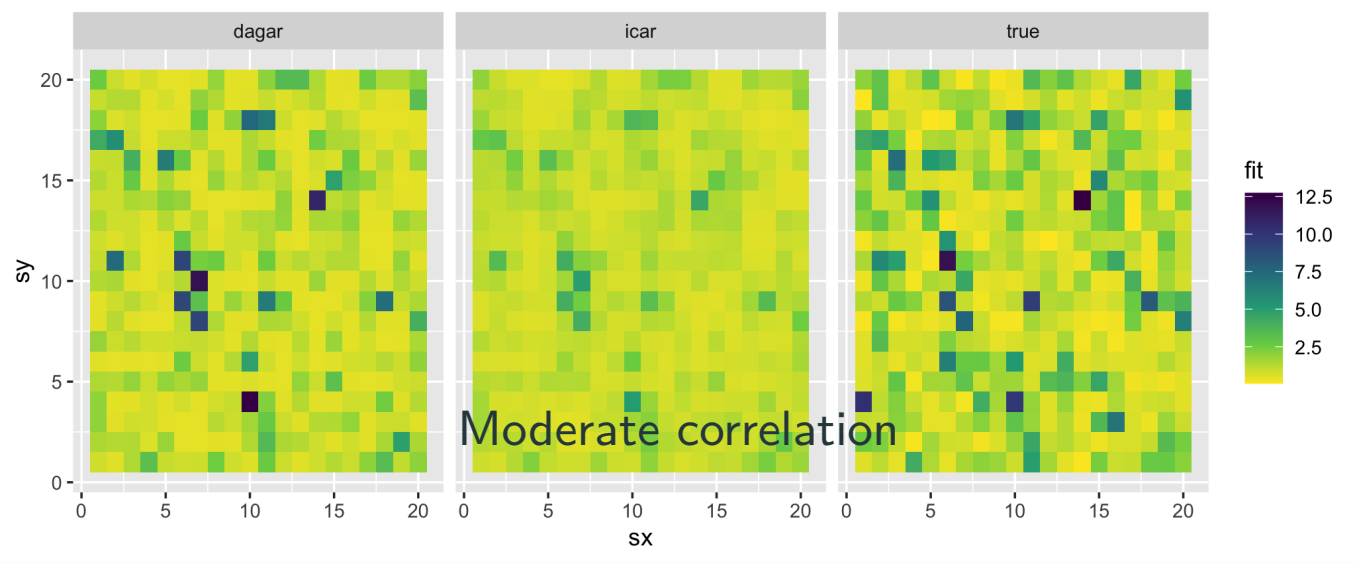


(a) Path (b) Grid (c) USA

Figure 3: MSE as a function of the true ρ (x-axis) for the simulation data analysis using data generated from an exponential GP.

Simulations





Simulations

Table: Strong correlation

Model	RMSE(w)	WAIC
DAGAR	0.59	1247
ICAR	0.59	1262

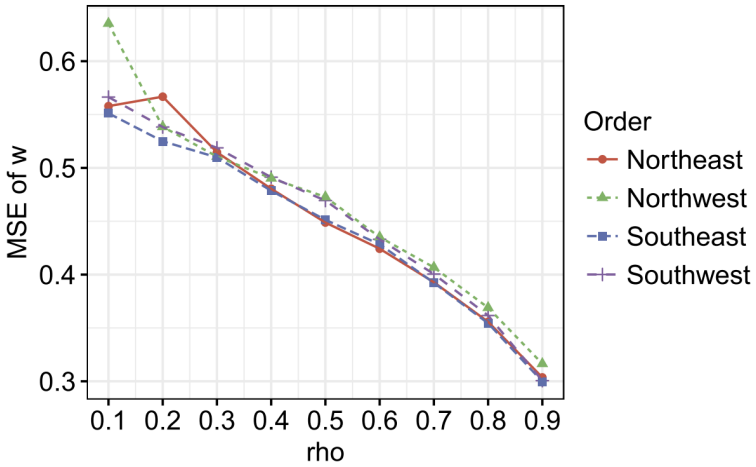
Table: Moderate correlation

Model	RMSE(w)	WAIC
DAGAR	0.70	1234
ICAR	0.74	1268

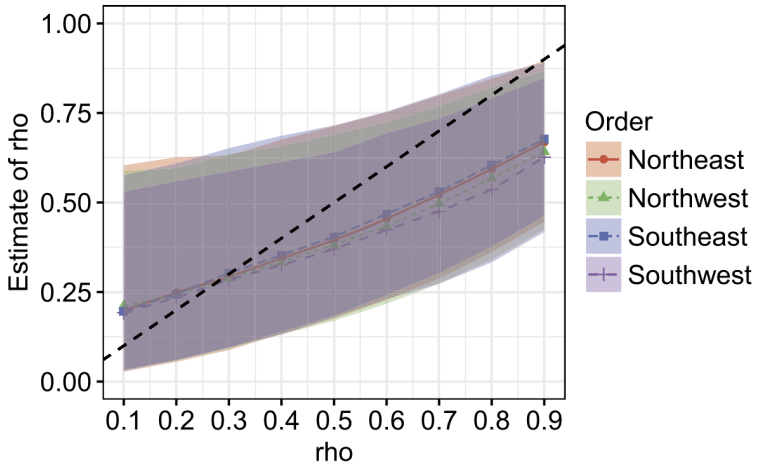
Dependence on ordering

- DAGAR model depends on the ordering of the regions when decomposing into local trees
- We can define a DAGAR model for every ordering
- Spatial regions do not have natural ordering
- How to choose the ordering?
- Model averaging over orderings ? Too many possibilities ($k!$)
- Coordinate based orderings were used in Datta et al., 2016; Stein, 2004; Vecchia, 1988

Dependence on ordering



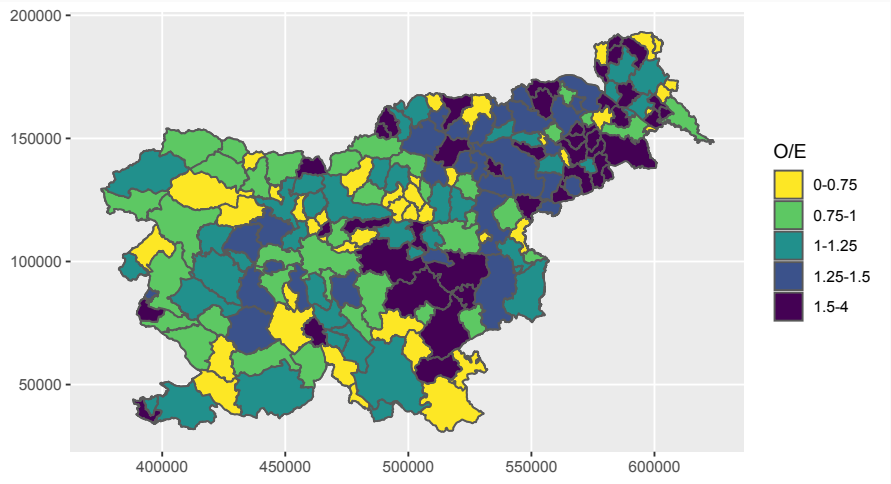
(a) USA: MSE



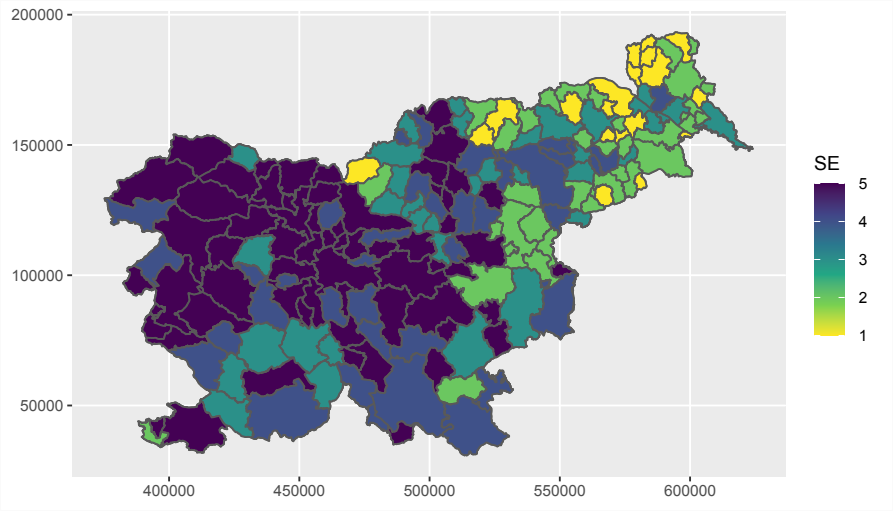
(b) USA: Estimate and confidence bands of ρ

Figure 6: MSE (left) and estimates and confidence bands of ρ (right) as a function of the true ρ (x-axis) for four different orderings of the DAGAR model.

Slovenia stomach cancer data



Standardized cancer incidence

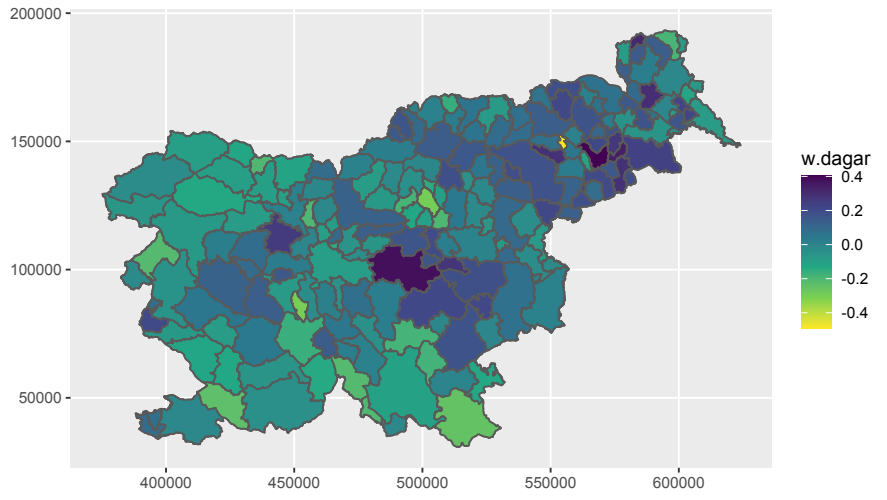


Socio-economic score

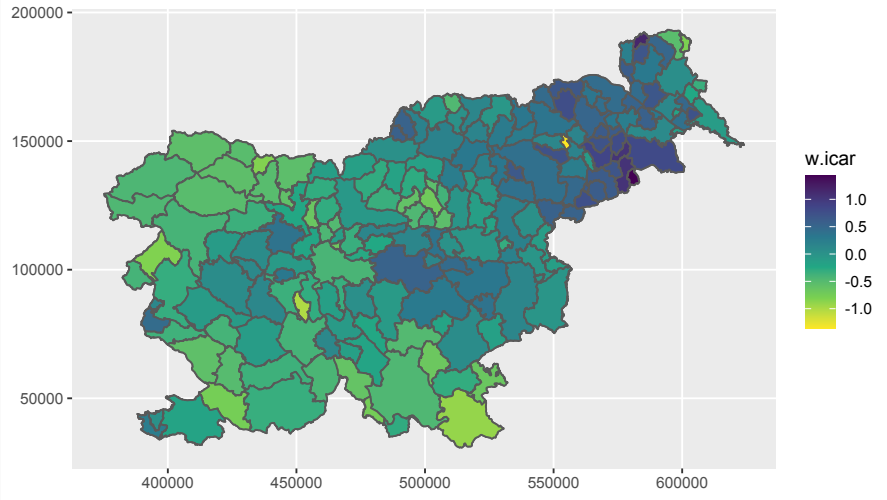
Figure: Slovenia stomach cancer data

- Observed (O_i) and expected (E_i) number of cancer counts for each of the 192 municipalities of the country
- Typically $E_i = P_i \frac{\sum_i O_i}{\sum_i P_i}$ where P_i is population of region i
- $O_i \sim Poisson(E_i \exp(\alpha + \beta SE_i + w_i))$ where $w \sim N^{-1}(0, \tau_w Q(\rho))$

Slovenia stomach cancer data



$\widehat{w}$ from DAGAR



$\widehat{w}$ from ICAR

Figure: Slovenia stomach cancer data

Table: Parameter estimates with confidence intervals and model comparison metrics

model	α	β	WAIC
Non-spatial	0.16 (0.12,0.19)	-0.14 (-0.17,-0.1)	1146.72
DAGAR	0.1 (-0.02,0.2)	-0.09 (-0.17,-0.01)	1083.03
ICAR	0.12 (0.08,0.17)	-0.04 (-0.11,0.04)	1082.33

Summary

- Areal data – visualization, measures of spatial correlation, analysis using spatial GLM
- Chloropleth maps for visualizing areal data
- Spatial proximity measures and Moran's I
- Graph-based models for areal data: ICAR, proper CAR, SAR, DAGAR
- Comparison of interpretability, accuracy, and Scalability for large areal data
- DAGAR models are positive definite and can be directly used to model or simulate any multivariate data on graphs (like imaging or social network data)
- Analysis using Stan

Review of the course

- Types of spatial data – point-referenced, areal, point-pattern
- Geospatial data exploration – surface plots, variograms
- Spatial (generalized) linear models using Gaussian Processes
- Bayesian spatial models (non-Gaussian response, missing covariate imputation, spatially-varying coefficient model)
- Large geo-spatial data: low-rank and nearest neighbor Gaussian Processes
- Geo-spatial analysis using `geoR`, `spBayes`, `spNNGP`, `BRISC`, `Stan`
- Areal data analysis — chloropleth maps, Moran's I, CAR and DAGAR models, analysis in Stan