Spatial Modeling and Prediction of County-level Employment-growth Data

N. Ganesh, USDA/NASS

Motivation

- We observe $\{y_i : i \in S\}$, where $S \subset U$ with $|S| = m$ and $|U| = M$. S is the set of sampled counties and U is the set of all counties. The objective is to predict the unobserved y_i's.
- Consider the following linear model for the y_i's:

$$y_i = \mathbf{x}_i'\beta + v_i, \ i = 1, \cdots, m$$

  where $v_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.
- In order to obtain better predictors (compared to the predictors obtained from the above model), we propose a spatially correlated variance-covariance matrix for the v_i's.

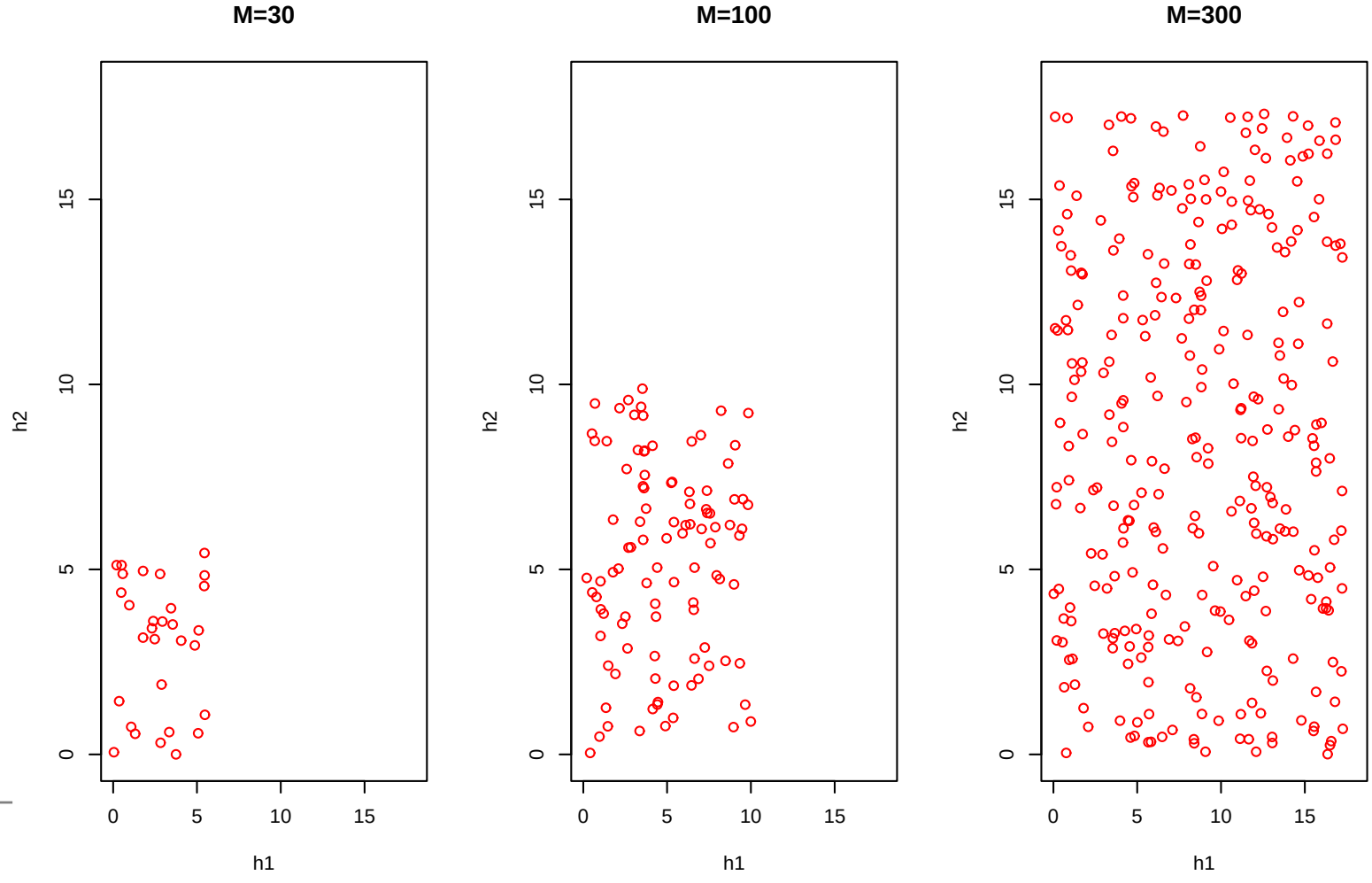
BLUP & EBLUP

For a general linear mixed model $y = X\beta + Zv + e$, let's say we are interested in predicting $t = \ell'\beta + a'v$.

- The best linear unbiased predictor (BLUP) $\hat{t}$ of t refers to the predictor among all linear (in y) unbiased predictors that has smallest mean squared error.
- The BLUP depends on unknown variance components; the EBLUP refers to the BLUP with estimated variance components.

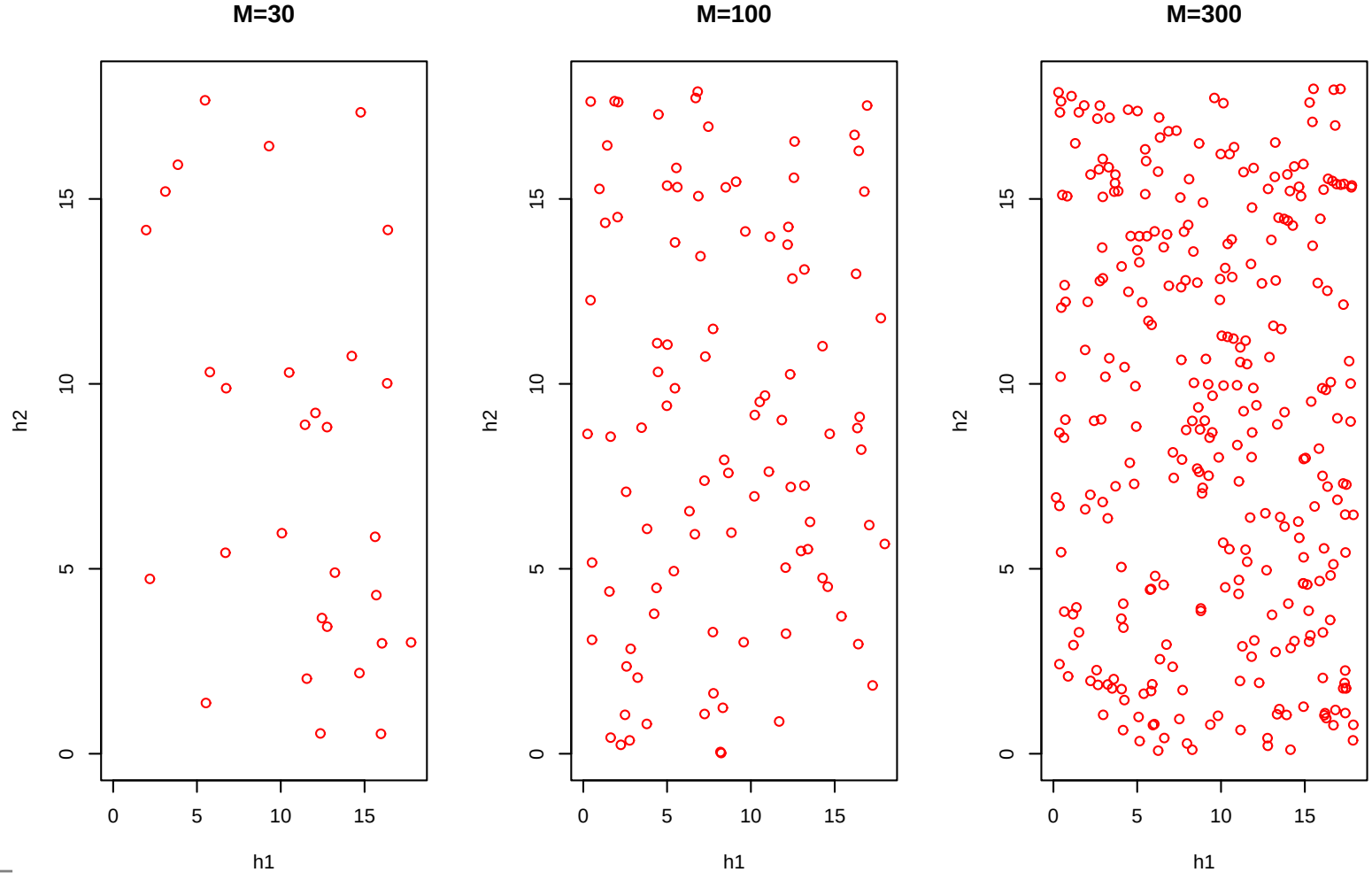
Increasing domain asymptotics

Increasing domain asymptotics: More and more observations are sampled over an increasing domain. An example of a point pattern for increasing domain asymptotics:



Infill asymptotics

Infill asymptotics: Observations are increasingly sampled over a bounded domain. An example of a point pattern for infill asymptotics:



Asymptotics for spatial data

One of the most popular covariance models for spatial data, the exponential covariance model with nugget effect is given by

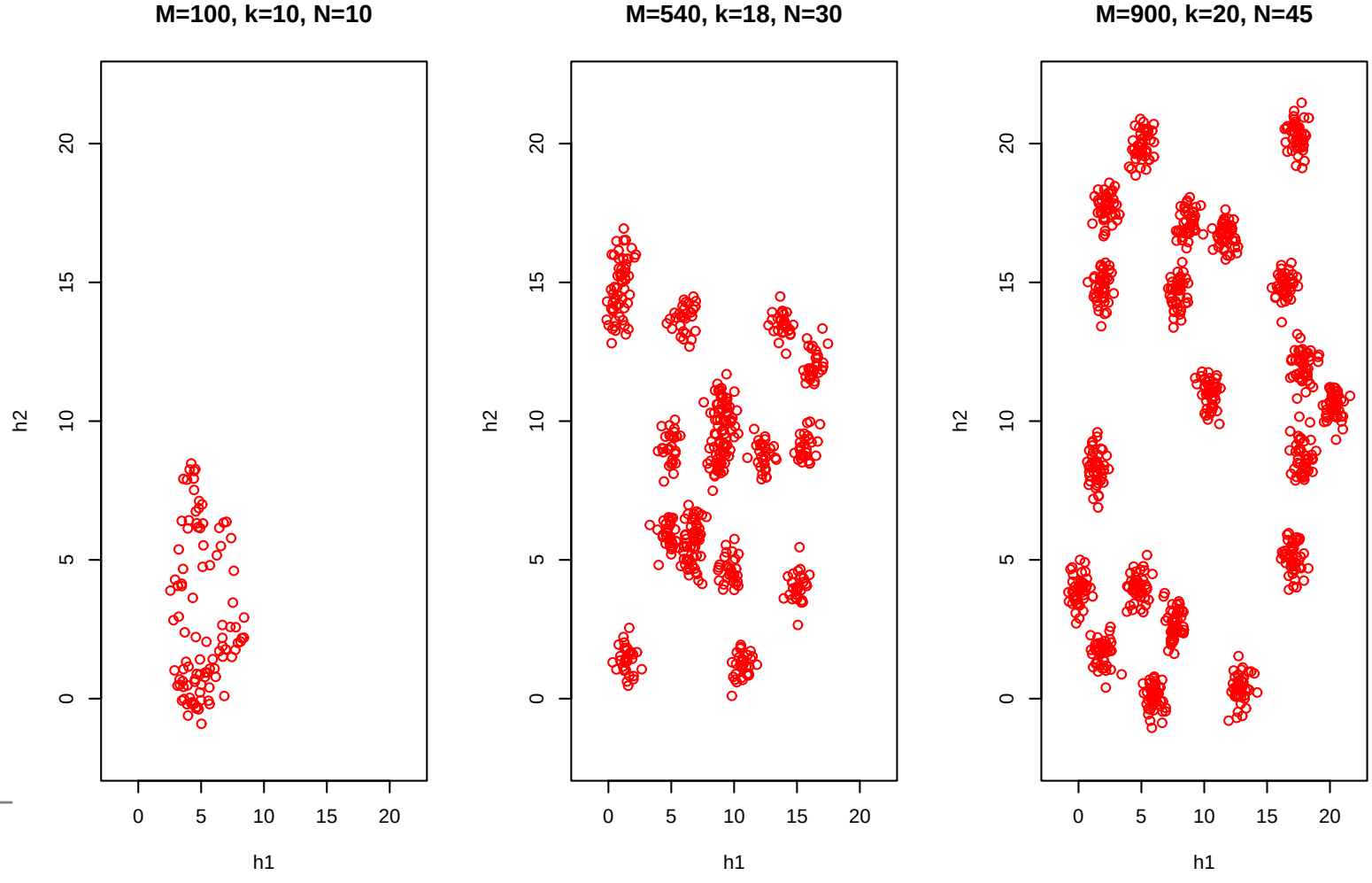
$$C(\mathbf{h}_i, \mathbf{h}_j) = \begin{cases} \sigma^2 + \delta & \text{if} \quad i = j, \\ \delta \exp(-\lambda \|\mathbf{h}_i - \mathbf{h}_j\|) & \text{if} \quad i \neq j, \end{cases}$$

where $\mathbf{h}_i$,$\mathbf{h}_j$ are the spatial locations of the observations, $\delta \geq 0$, $\lambda \geq 0$, $\sigma^2 > 0$.

- Under *increasing domain asymptotics* the MLE for $\boldsymbol{\eta} = (\delta, \lambda, \sigma^2)$ is $\sqrt{m}$-consistent (Mardia & Marshall).
- Under *infill asymptotics* and when the spatial locations h_i are on a lattice on $[0, 1]$, the MLE for σ^2 is $\sqrt{m}$-consistent. However, δ and λ can not be simultaneously consistently estimated, but the MLE for $\delta\lambda$ is $m^{\frac{1}{4}}$-consistent (Chen et al). *No asymptotic results exist for any other spatial pattern.*

Proposed asymptotic framework

The proposed asymptotic framework involves well separated blocks with number of blocks and number of observations in each block increasing with M. An example of a point pattern for proposed asymptotic framework:



Scaling

- Let $\mathbf{z}_i^\star$ be a vector of spatial locations and certain categorical and continuous variables which measure similarity.
- The $\mathbf{z}_i^\star$ are thought to be in an increasing domain.
- The scaled values $\mathbf{z}_i = K\mathbf{z}_i^\star$ are bounded.
- $\|\mathbf{z}_i - \mathbf{z}_j\| = K\|\mathbf{z}_i^\star - \mathbf{z}_j^\star\|$, where $K = M^{-p}, 0 < p < 1/d$, $d = \dim(\mathbf{z}_i^\star)$.

Covariance model for the v_i's

The proposed covariance model for the v_i's is

$$\Sigma_{\mathsf{U}} = \sigma^2 I_M + \delta A_{\mathsf{U}}$$

where the $(i,j)^{th}$ entry of A_{U} is given by

$$A_{ij} = \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|),$$

where $\delta \geq 0$, $\lambda \geq 0$, $\sigma^2 > 0$.

Assumption (C)

- The set of counties U can be partitioned into k ($= k(M)$ increasing to ∞ with M) blocks $C_1, \ldots, C_k$, with block sizes $N_1, \ldots, N_k$ such that $\sum_{l=1}^{k} N_l = M$.
- From each block C_l, n_l of the N_l counties are sampled such that $\sum_{l=1}^{k} n_l = m$. The n_l's are assumed to be non-random.
- The asymptotic framework that is considered is $k \to \infty$ and for each l, $N_l \to \infty$, $n_l \to \infty$ such that $0 < \lim_{n_l, N_l \to \infty} \frac{n_l}{N_l} < \infty$.

Definition of a "block"

For $l = 1, \ldots, k$,

$$\limsup_{M \to \infty} M^p \sup_{i,j \in C_l} \|\mathbf{z}_i - \mathbf{z}_j\| < \infty,$$

and for all $l_1 \neq l_2$,

$$\liminf_{M \to \infty} \frac{M^p}{\log M} \inf_{i \in C_{l_1}, j \in C_{l_2}} \|\mathbf{z}_i - \mathbf{z}_j\| = \infty.$$

Moreover, it is assumed that for $l = 1, \ldots, k$, $\exists\ c_l > 0$ such that

$$\lim_{N_l \to \infty} \frac{1}{N_l^2} \sum_{i,j \in C_l} I_{[M^p \|\mathbf{z}_i - \mathbf{z}_j\| \geq c_l]} = \epsilon_l > 0$$

Prediction under a simplified model

Consider the following simplified model for $\mathrm{cov}(v_i, v_j)$:

$$cov(v_i, v_j) = \begin{cases} \delta & \text{if } \ i \neq j, i, j, \in C_l \text{ for some } l, \\ \sigma^2 + \delta & \text{if } \ i = j, \\ 0 & \text{otherwise.} \end{cases}$$

For an unobserved y, when $(\beta, \delta, \sigma^2)$ are known, we compare the best linear unbiased predictors $\hat{y}^*$ and $\hat{y}$ obtained under the independence model and the above model. The relative risk $R(\hat{y}^*, \hat{y})$ (which is calculated assuming the above model is the true model) is given by

$$\begin{aligned} R(\hat{y}^*, \hat{y}) &= \frac{\mathrm{MSE}(\hat{y}^*)}{\mathrm{MSE}(\hat{y})} \\ &= \frac{(\sigma^2 + \delta)(1 + n_l\delta/\sigma^2)}{(\sigma^2 + \delta)(1 + n_l\delta/\sigma^2) - n_l\delta^2/\sigma^2} \\ &\to 1 + \frac{\delta}{\sigma^2} \end{aligned}$$

Some comments on relative risk

Table 1: Relative risk calculation.

δ	σ^2	$n_l = 1$	$n_l = 2$	$n_l = 5$	$n_l = 10$	$n_l = 20$	limit
0.4	0.6	1.19	1.30	1.44	1.53	1.59	1.67
0.5	0.5	1.33	1.50	1.71	1.83	1.91	2.00
0.6	0.4	1.56	1.82	2.12	2.29	2.38	2.50

From the above calculation, in order to achieve large relative risk (≥ 1.5) for the EBLUP under the proposed model compared to the EBLUP under the independence model, we require:

- At least $5 - 10$ "neighboring" counties.
- Block radius be small and δ/σ^2 be large.

Parameter estimation

- The parameter τ^2 is defined as $\tau^2 = \delta + \sigma^2$. An estimator $(\widehat{\beta}, \hat{\tau}^2)$ for (β, τ^2) is given by

$$\begin{aligned} \widehat{\beta} &= (X'X)^{-1}X'y \\ \hat{\tau}^2 &= \frac{1}{m}\sum_{i=1}^{m}(y_i - \mathbf{x}_i'\widehat{\beta})^2. \end{aligned}$$

- An estimator $(\hat{\delta}, \hat{\lambda})$ for (δ, λ) is given by minimizing:

$$h(\delta, \lambda; \mathbf{y}) = \sum_{l=1}^{k} \sum_{\substack{i,j \in C_l \\ i \neq j}} \left(\hat{\epsilon}_i \hat{\epsilon}_j - \delta \exp(-\lambda M^p \|\mathbf{z}_i - \mathbf{z}_j\|) \right)^2$$

where $\hat{\epsilon}_i = y_i - \mathbf{x}_i'\widehat{\beta}$.

Theorem 1

- Suppose $\tau_o^2 > 0$ and m, $M \to \infty$ such that $0 < \lim_{m,M\to\infty}(m/M) < \infty$, then under certain regularity conditions, $(\widehat{\beta}, \hat{\tau}^2)$ is consistent for (β_o, τ_o^2). Moreover,

$$\begin{pmatrix} [\mathrm{var}(\widehat{\beta})]^{-\frac{1}{2}} & \mathbf{0} \\ \mathbf{0}' & [\mathrm{var}(\hat{\tau}^2)]^{-\frac{1}{2}} \end{pmatrix} \begin{pmatrix} \widehat{\beta} - \beta_{\mathrm{o}} \\ \hat{\tau}^2 - \tau_o^2 \end{pmatrix} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{I}).$$

- Under some further regularity conditions, for $i = 1 \ldots, q$, the asymptotic variance of $\widehat{\beta}_i$ and the asymptotic variance of $\hat{\tau}^2$ are $O(\sum_{l=1}^{k} n_l^2/m^2)$.
- Furthermore, when all the n_l's grow at the same rate, the asymptotic variances of $\widehat{\beta}_i$'s and $\hat{\tau}^2$ are $O(1/k)$.

Theorem 2

- Suppose $\delta_o > 0$, $\lambda_o > 0$ and Assumption (C) is true. Then under certain regularity conditions $(\hat{\delta}, \hat{\lambda})$ is consistent for (δ_o, λ_o). Moreover,

$$\frac{\sum_{l=1}^{k} n_l^2}{(\sum_{l=1}^{k} n_l^4)^{\frac{1}{2}}} K_o^{-\frac{1}{2}} L_o \begin{pmatrix} \hat{\delta} - \delta_o \\ \hat{\lambda} - \lambda_o \end{pmatrix} \xrightarrow{d} \mathrm{N}(\mathbf{0}_2, \mathbf{I}_2)$$

  where K_o and L_o are 2×2 matrices with bounded entries that depend on $\delta_o, \lambda_o, \sigma_o^2$.

- When all the n_l's grow at the same rate, the asymptotic variances of $\hat{\delta}$ and $\hat{\lambda}$ are $O(1/k)$.

MLE for $\kappa = (\beta, \delta, \lambda, \sigma^2)$

- For the model and asymptotic framework that has been considered, it is conjectured that the MLE $\widehat{\kappa}_{\text{ML}}$ of κ is consistent and

$$\Big(\mathcal{I}(\kappa)\Big)^{\frac{1}{2}}(\widehat{\kappa} - \kappa_o) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{I})$$

  where $\mathcal{I}(\kappa)$ is the information matrix.

- For general patterns for $\mathbf{z}_i$, proving the above result is quite difficult. The technical difficulties have to do with not being able to write the inverse of the var$(\mathbf{y})$ in closed form.

- Asymptotic results for the MLE under infill asymptotics were derived only for spatial patterns for which the inverse of the var$(\mathbf{y})$ could be written in closed form (Chen et al., Low & Lam and Ying).

Simulation summary: LSE & MLE

- Let $\rho = \delta/(\delta + \sigma^2)$, i.e. ρ is the largest possible correlation between the residuals. LSE refers to the estimators given in Theorems 1-2
- The relative efficiency of LSE and MLE depends on ρ. Large values of ρ correspond to small values of relative efficiency of LSE and MLE.
- When ρ and k are small, estimating λ is problematic. However, the frequency of such cases decreases as k increases.
- In certain cases, when estimating λ, the relative efficiency of $\hat{\lambda}$ and $\hat{\lambda}_{\text{ML}}$ is greater than 1. This is due to k being too small. In such cases whenever k is increased, the relative efficiency decreases to a number smaller than 1.

Simulation summary: LSE & MLE

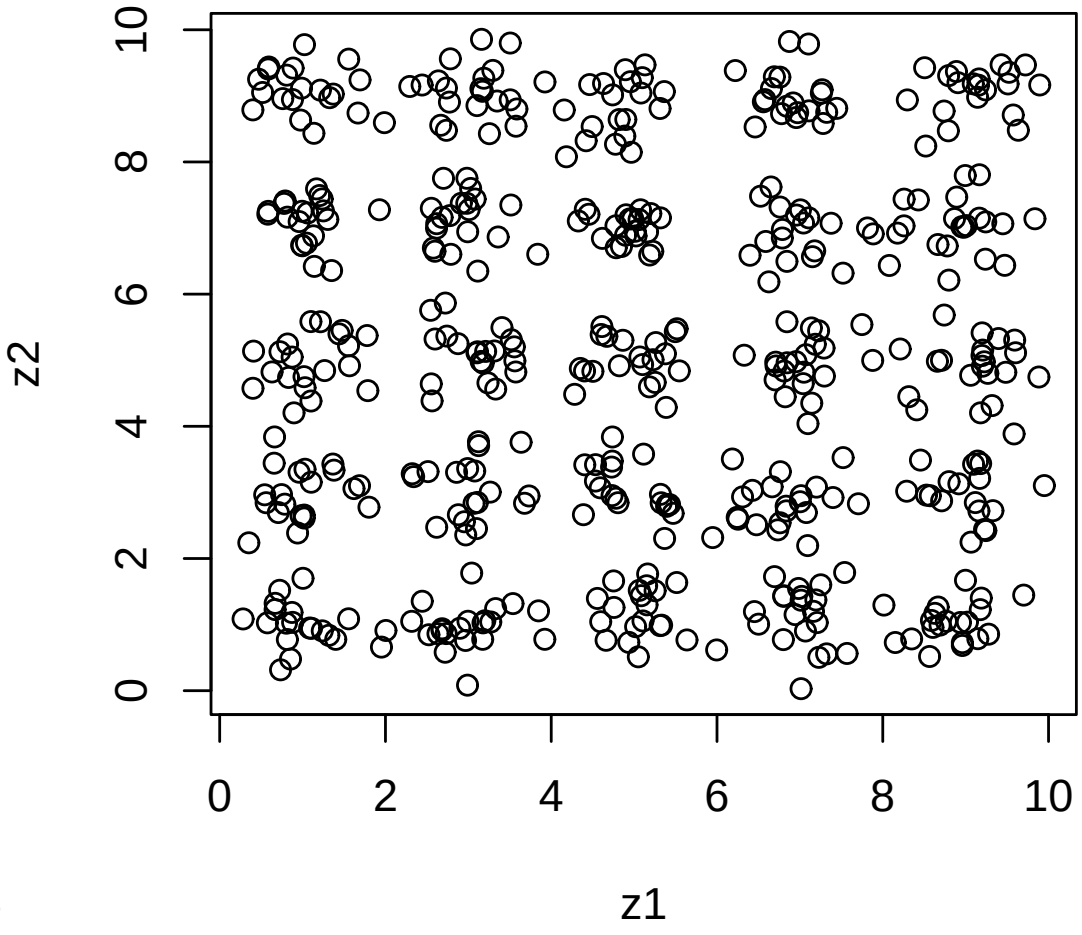
- For various parameter settings, the relative efficiency of $\hat{\tau}^2$ and $\hat{\tau}^2_{\text{ML}}$ is approximately 1.
- The relative risk of the EBLUP obtained under the proposed model and independence model does not depend on the method of estimation of the parameters.

Table 2: $k = 20$, $n = 20$, $m = 400$ (500 runs for LSE, 100 runs for MLE).

		LSE			MLE		
par.	tr. val.	mean	e.s.e.	s.e.	mean	e.s.e.	s.e.
β_1	1.00	0.998	0.184	0.173	0.994	0.133	0.134
β_2	1.00	1.046	1.148	1.050	1.015	0.730	0.696
δ	0.60	0.619	0.192	0.181	0.600	0.120	0.116
λ	0.54	0.538	0.207	0.214	0.599	0.184	0.159
σ^2	0.40	0.365	0.146	-	0.392	0.082	0.074
τ^2	1.00	0.984	0.106	0.103	0.992	0.097	0.096

Correlated blocks

Point pattern for which median between block correlation (for neighboring blocks) is $1/8$ median within block correlation.



Correlated blocks

Table 3: Summary of LSE when blocks are correlated, $k = 25$, $n = 20$, $m = 500$. (500 simulation runs)

		LSE			
par.	tr. val.	mean	med.	e.s.e.	s.e.
δ	0.3	0.305	0.294	0.105	0.112
λ	0.3	0.319	0.308	0.178	0.175

Table 4: Summary of LSE when blocks are correlated, $k = 25$, $n = 20$, $m = 500$. (500 simulation runs)

		LSE			
par.	tr. val.	mean	med.	e.s.e.	s.e.
δ	0.6	0.618	0.603	0.175	0.171
λ	0.3	0.327	0.318	0.129	0.111

Some comments on estimation

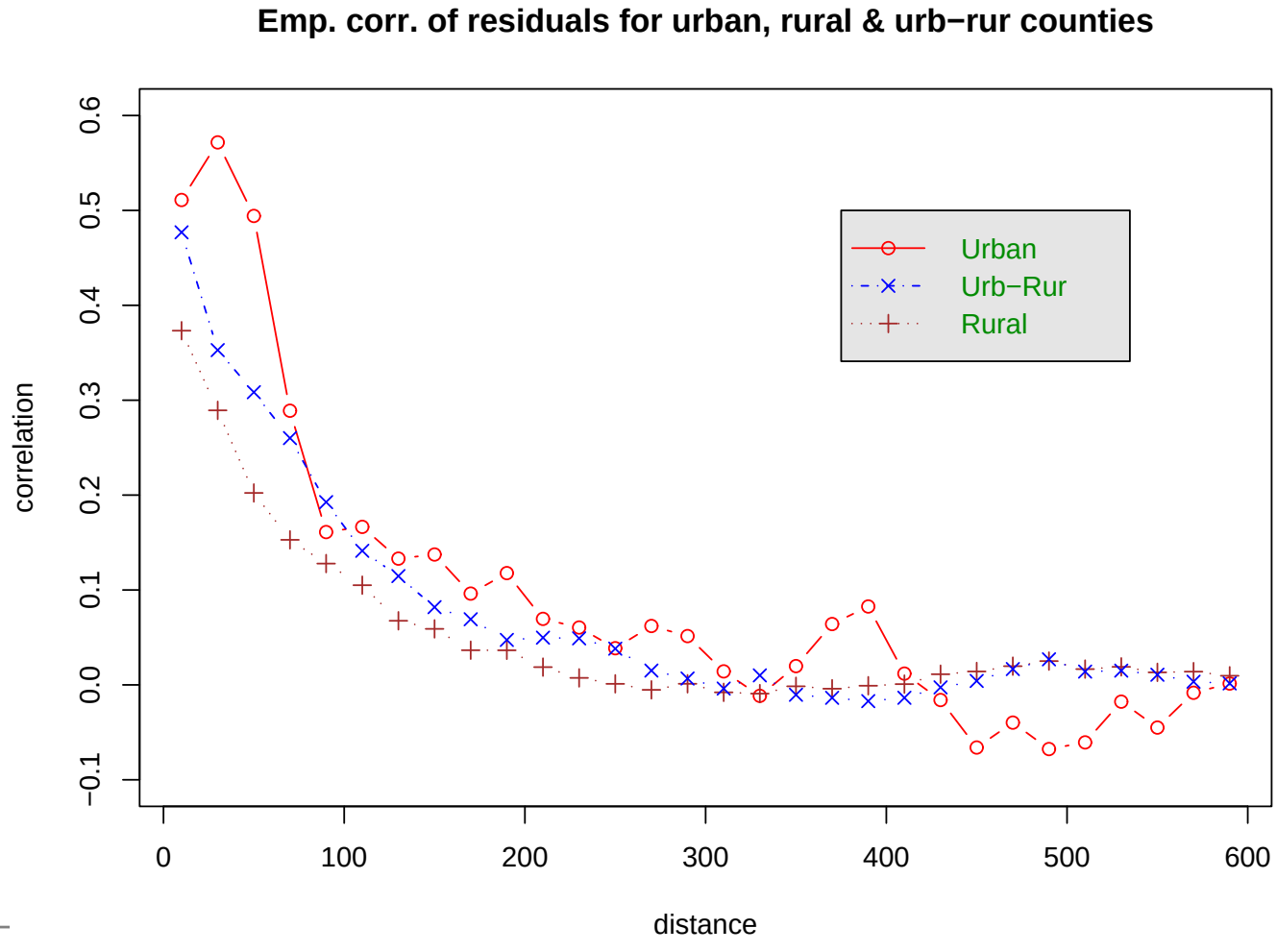
- For good estimation of the parameters we require
  - A large number of blocks.
  - The block radius be not too small.
  - At least 10-15 counties per block.
- Let $L(\beta, \delta, \lambda, \sigma^2; y)$ be the log likelihood under the proposed model. Simulations suggest $(\tilde{\delta}, \tilde{\lambda})$ obtained by maximizing $L(\widehat{\beta}, \delta, \lambda, \hat{\tau}^2 - \delta; y)$ is more efficient than $(\hat{\delta}, \hat{\lambda})$ given in Theorem 2.
- Moreover, the above method is substantially faster to implement than the MLE, approximately $1/3$ the running time of the MLE when $k = 40$, $n = 20$. Also, the LSE takes about $1/25$ the running time of MLE.

Data analysis

- The data set that was analyzed was U.S. county level employment growth rates (source: Wheeler). The data set includes 14 covariates. The response variable was county-level civilian employment growth-rate between 1980-1990.
- The objective was to compare the EBLUP obtained under the proposed model and the EBLUP obtained under the independence model.
- Among the 3106 counties, 4 counties with missing covariates were deleted.
- Stepwise AIC criterion was used to select the best set of covariates.
- Counties with a population of at least 500000 were self-represented (81 counties). Simple random sample was used to sample the remaining 719 of the 3021 counties and I pretend that only the sampled counties are observed.

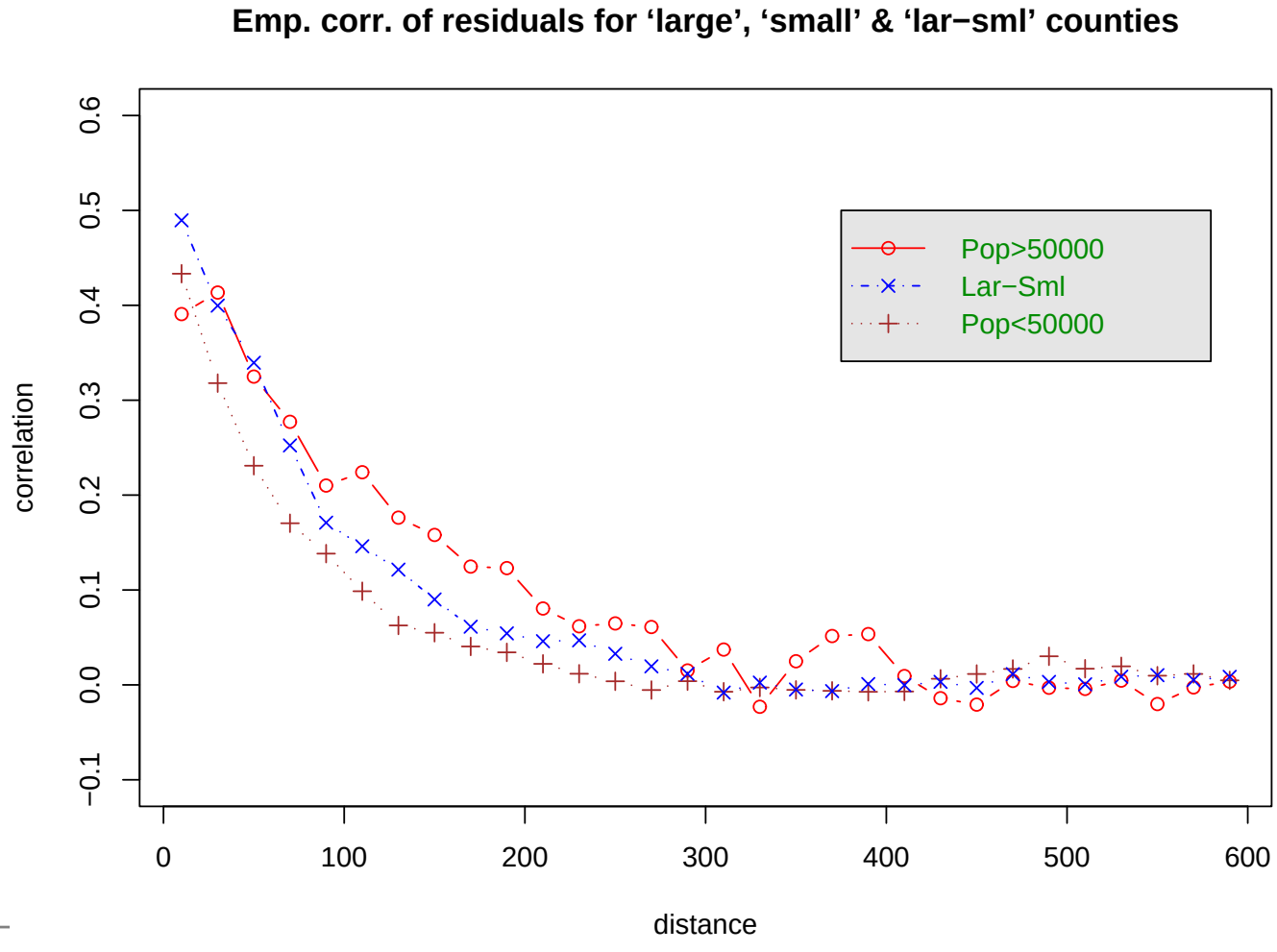
Plot of the empirical correlation

For all counties, by urban/rural county type, a plot of the empirical correlation of the residuals from the independence model:



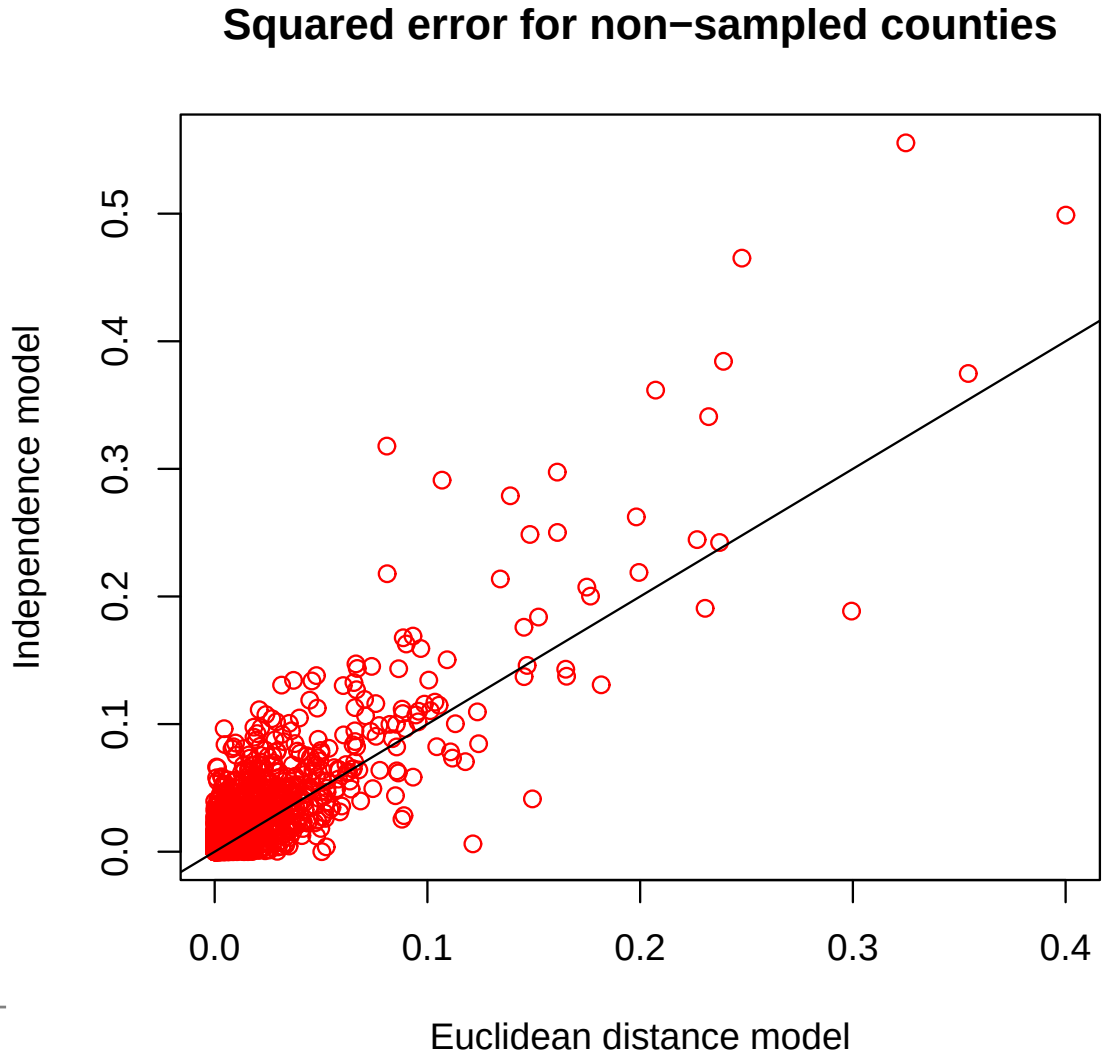
Plot of the empirical correlation

For all counties, by county population size, a plot of the empirical correlation of the residuals from the independence model:

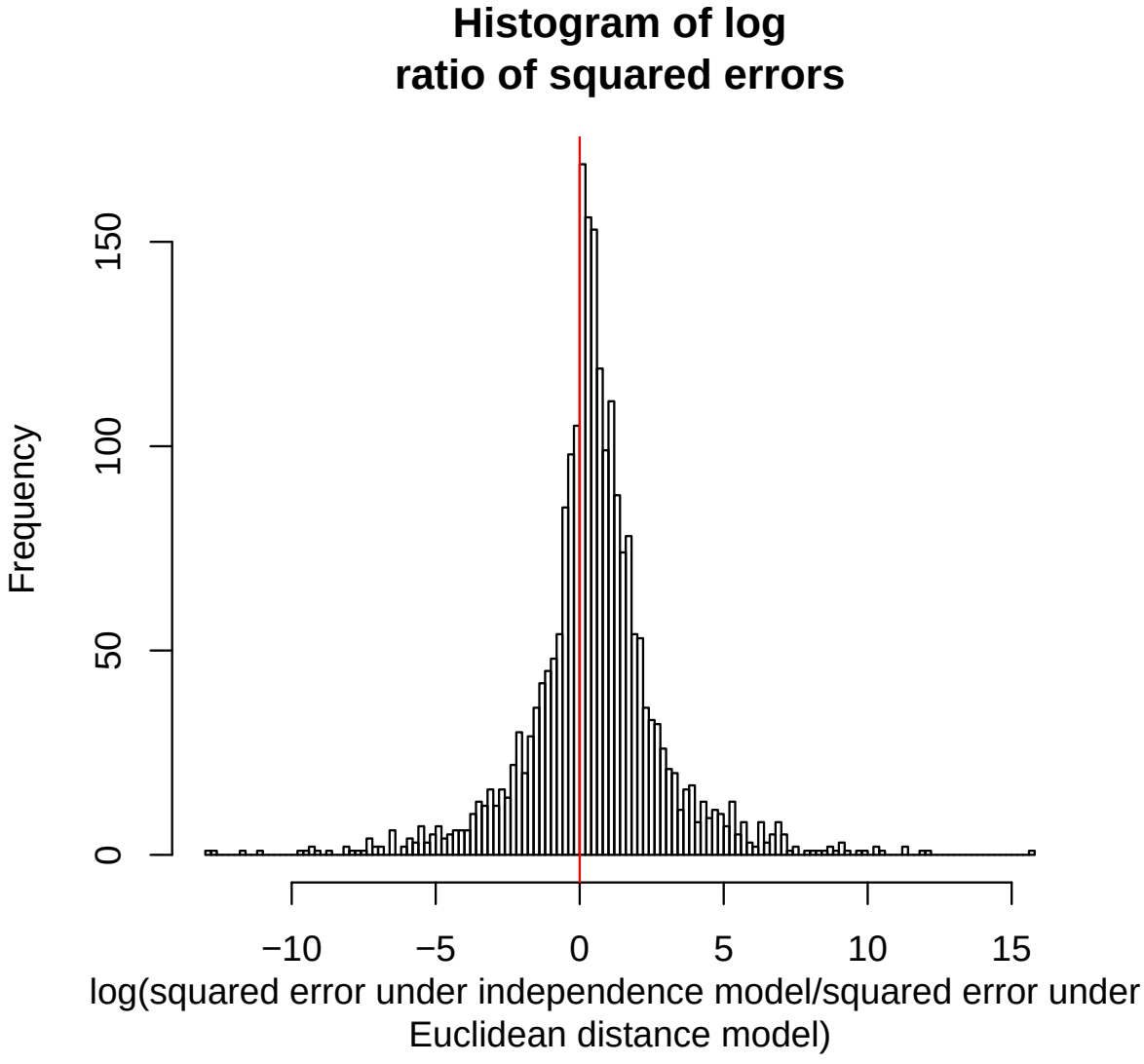


Plot of squared error

For non-sampled counties, a plot of the squared error using independence model and “Euclidean distance” model.

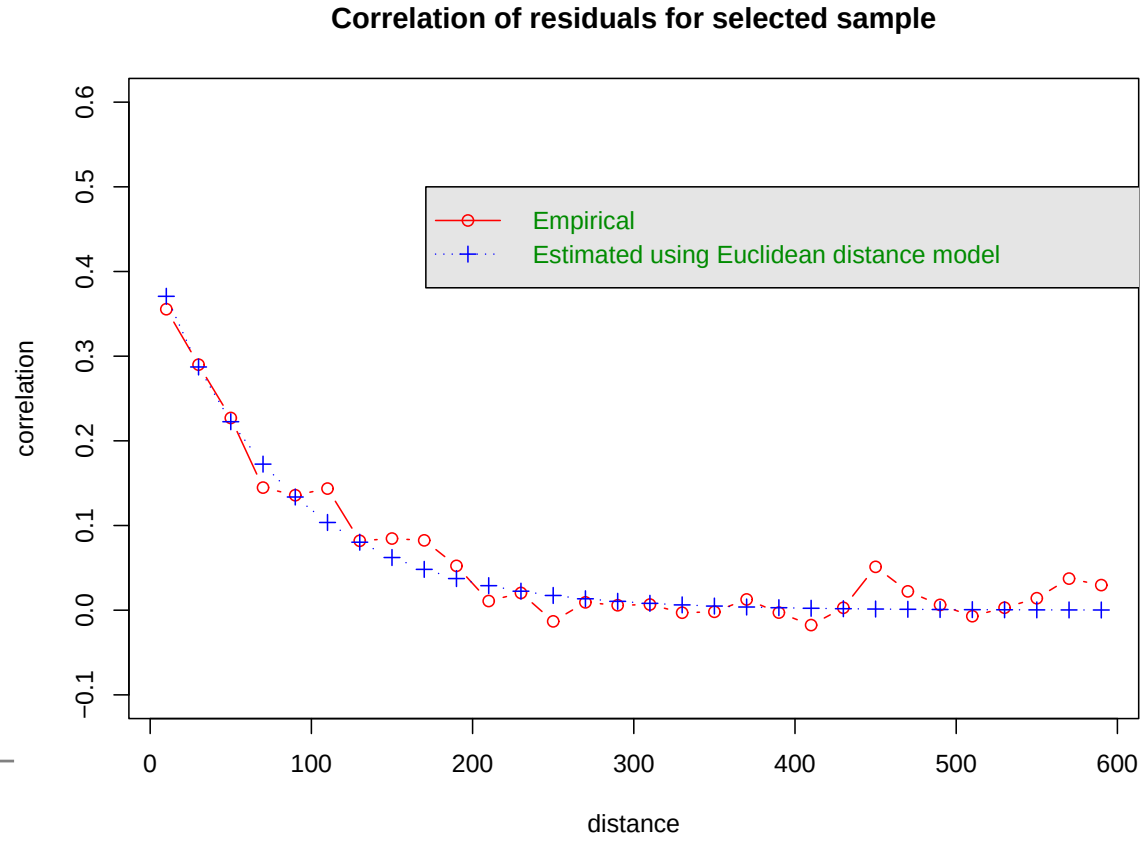


Histogram of log ratio of squared error



Plot of the empirical & estimated corr.

Using the sampled counties, a plot of the empirical correlation of the residuals from the independence model and the estimated correlation using the “Euclidean distance” model:



Comments

- The objective is to try to consider non-spatial covariates in addition to spatial locations in the residual covariance matrix. We would like to define a new distance metric which takes non-spatial covariates into account.
- Since neighboring urban counties are more correlated than neighboring urban-rural or rural counties, we added a small penalty term to the distances of these counties.

Ratio of average squared error

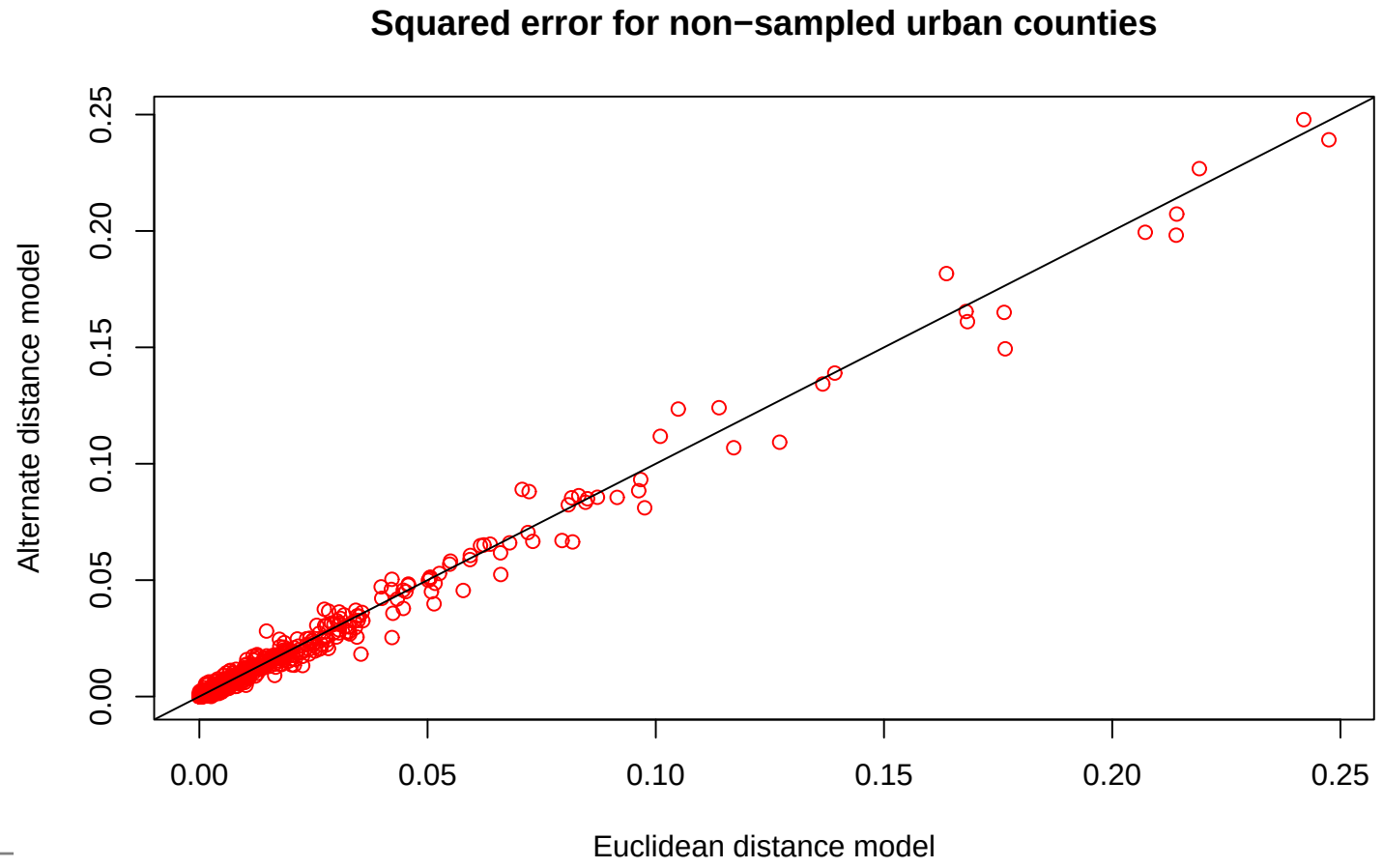
Let “Ind” be the average squared error when the independence model is used. Similarly “Euc” and “Alt” denote the average squared error when the Euclidean distance model and alternate distance model are used.

Table 5: Ratio of average squared error

	Ind/Euc	Ind/Alt
all unobserved counties	1.496	1.487
unobserved urban counties	1.575	1.612
unobs. urb. counties with $\geq$ 3 urb. ngbrs. within 40 mi.	1.516	1.597

Plot of squared error

For non-sampled urban counties, a plot of the squared error using “alternate distance” model and “Euclidean distance” model.



Frequency table of number of neighbors

Table 6: Number of sampled counties within 40 miles of a non-sampled county.

# of nghbr counties	0	1	2	3	4	5	6	7	8	9
frequency	506	639	523	328	155	83	38	16	9	5

Table 7: Number of sampled urban counties within 40 miles of a non-sampled urban county.

# of nghbr counties	0	1	2	3	4	5	6	7	8	9
frequency	190	148	111	53	34	20	2	1	6	4

Concluding remarks

- Possibly use a previous survey to construct the distance matrix for the alternate distance model.
- Consider other non-spatial covariates such as county unemployment rate, county land area and county population size to penalize dissimilar counties.

References

- Chen, H., Simpson, D.G. & Ying, Z. (2000), Infill asymptotics for a stochastic process model with measurement error, *Statistica Sinica*, **10**, 141-156.
- Loh, W. & Lam, T. (2000), Estimating structured correlation matrices in Gaussian random field models, *Annals of Statistics*, **28**, 880-904.
- Mardia, K.V. & Marshall, R.J. (1984), Maximum likelihood estimation of models for residual covariance in spatial regression, *Biometrika*, **71**, 135-146.
- Wheeler, C.H. (2003), U.S. Counties 1998. http://qed.econ.queensu.ca/jae/ 2003-v18.1/wheeler.
- Ying, Z. (1993), Maximum likelihood estimation of parameters under a spatial sampling scheme, *Annals of Statistics*, **21**, 1567-1590.