The Joys of Geocoding (from a Spatial Statistician’s Perspective)

Dale L. Zimmerman
University of Iowa

October 21, 2010

Geocoding context

- Applications of spatial statistics to public health and social sciences are increasing

- Require defining a spatial location for each subject

- Historically, a person's spatial location has been defined as their place of residence

- The residential address is assigned a location reference, or *geocode*

- The geocode may be a statistical tabulation area (e.g. census tract) or point (lat-long) coordinates; we consider point geocodes only

- The process for assigning the geocode is known as *geocoding*

Geocoding methods

1. Visit each address with a GPS receiver
2. Use maps from aerial or satellite imagery together with parcel descriptions from county assessors
3. Street geocoding — Use a geographic information system (GIS) to match an address to a street name and address range (a "street segment") in a digitized street centerline file, and then interpolate the address along this segment
   - The digitized street centerline files most often used are the U.S. Census Bureau TIGER/Line files (or enhanced versions thereof)
   - Many social/public health studies utilize the services of a commercial geocoding firm

Pros and cons of street geocoding

Pros:

- cheaper
- more convenient (hence much more common)

Cons:

- less accurate
- less complete

Inaccuracy and incompleteness of street geocoding

At least 10 studies have been published measuring accuracy by distance between the street geocode and the “true” location (ascertained by GPS or imagery). These show that:

- Positional errors of several hundred meters occur regularly
- Errors tend to be larger in rural areas
- E.g., in a study involving rural addresses in 4 counties in upstate NY, 10% geocoded with errors > 1.5 km, and 5% geocoded with errors > 2.8 km

At least 6 studies of completeness of automated geocoding have been published. These show that:

- It is common for 10% to 40% of the addresses of study subjects to fail to geocode
- The problem is worst in rural areas

Potential effects

- Positional errors and incompleteness reduce power (by an unknown amount) for detecting clusters, trends, and associations with spatial covariates

- Inferences made by applying complete-data methods to the incomplete data are subject to selection bias (*geographic bias*)

Research goals

- Gain an understanding of the nature of the measurement errors and incompleteness (including geographic bias) associated with street geocoding

- Determine magnitudes of their effects on existing spatial methods of analysis

- Modify existing methods of analysis so as to properly account for them

References

Zimmerman, D.L. (2008). Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. *Biometrics* **64**:262-270.

Zimmerman, D.L. and Fang, X. (2011). Estimating spatial variation in disease risk from locations coarsened by incomplete geocoding. *Statistical Methodology*, forthcoming.

Zimmerman, D.L., Fang, X., and Mazumdar, S. (2008). Spatial clustering of the failure to geocode and its implications for the detection of disease clustering. *Statistics in Medicine* **27**:4254-4266.

Zimmerman, D.L., Fang, X., Mazumdar, S., and Rushton, G. (2007). Modeling the probability distribution of positional errors incurred by residential address geocoding. *International Journal of Health Geographics* **6**:1.

Zimmerman, D.L. and Li, J. (2010). The effects of local street network characteristics on the positional accuracy of automated geocoding for geographic health studies. *International Journal of Health Geographics* **9**:10.

Zimmerman, D.L., Li, J., and Fang, X. (2010). Spatial autocorrelation among automated geocoding errors and its effects on testing for disease clustering. *Statistics in Medicine* **29**:1025-1036.

Zimmerman, D.L. and Sun, P. (2011). Estimating spatial intensity and variation in risk from locations subject to geocoding errors. *Submitted*.

Carroll County example

- We have a complete database of 9298 addresses in Carroll County, Iowa, current as of 12-31-05

- Obtained in conjunction with a comprehensive study of rural health by UI researchers

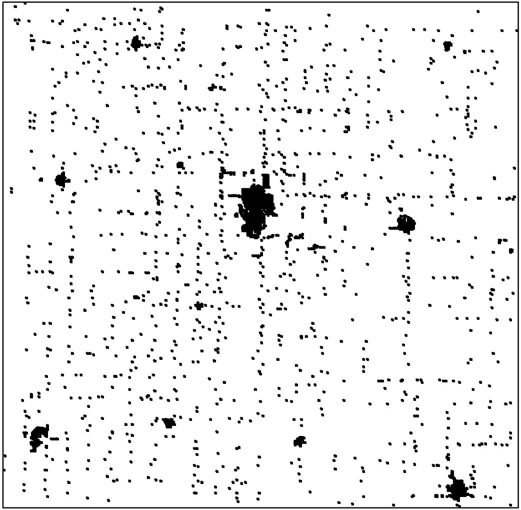
- A “true” geocode (using 2-ft resolution aerial imagery) was obtained for each address

- The addresses were street-geocoded as well

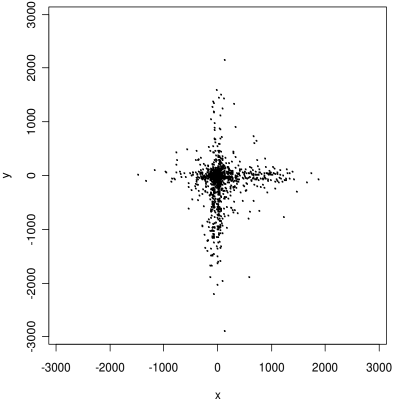
- Of the 9892 addresses, 7443 (80%; 64.3% for rural addresses, 85.4% for municipal addresses) could be street-geocoded using a 60%-matching criterion

- Zip codes were observed for all 9892 addresses

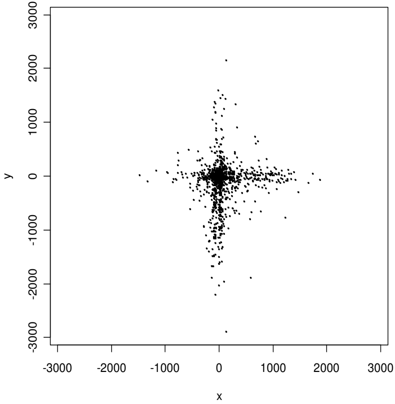
Locations of 7443 Carroll County addresses that street-geocoded:



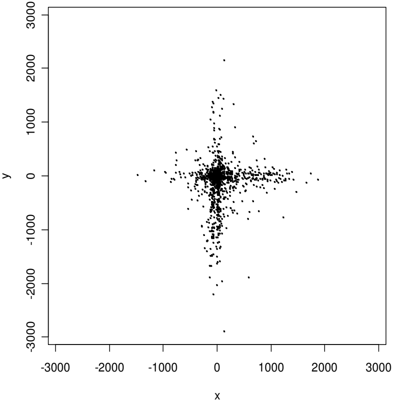
Geocoding errors for rural addresses ranged from 3 m to 2896 m (median 168 m):



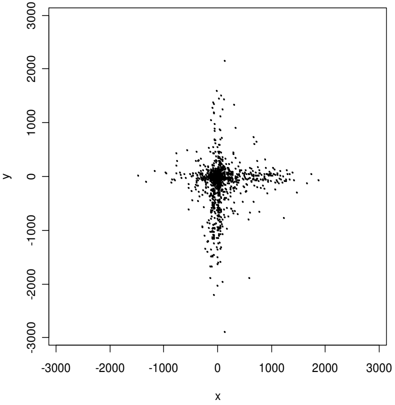
Features of this error distribution: Clustering along axes



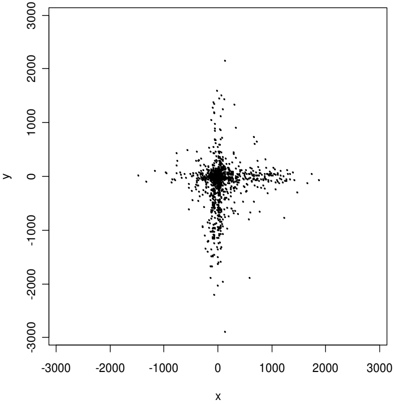
Features of this error distribution: Asymmetry wrt origin



Features of this error distribution: Outliers



Features of this error distribution: Parallel strands straddling each axis

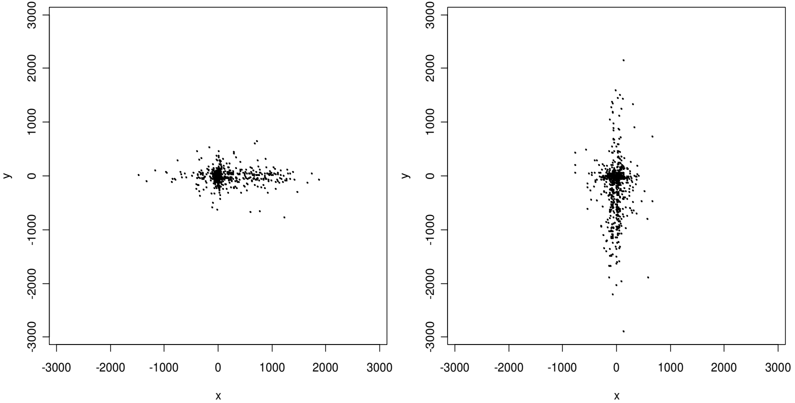


Explanations?

- Greek cross shape — rectilinearity of street network, sizable addressing or interpolation errors

  Errors tend to be aligned with the axial orientation of the corresponding street segment.

Decomposition of errors by the primary orientation (E-W or N-S) of the street on which the corresponding address lies:



Explanations? (continued):

- Asymmetry wrt origin — systematic addressing or interpolation error
- Straddling strands — offsets from street centerline
- Outliers — very large offsets or TIGER file errors/omissions

Modeling the error distribution

- No uniform or normal distribution will fit adequately
- Nor will *any* elliptical distribution
- But mixtures of normal or t distributions may fit well:

$$f(\mathbf{x};\mathbf{p},\theta) = \sum_{i=1}^{g} p_i f_i(\mathbf{x};\theta)$$

  where $p_i \geq 0$ $(i = 1,\ldots,g)$, $\sum_{i=1}^{g} p_i = 1$, and $f_i(\cdot)$ is a normal or t pdf

- Methodology and free software are available for likelihood-based estimation of $(\mathbf{p},\theta)$; see e.g. Peel and McLachlan (2000, *Statistics and Computing*)

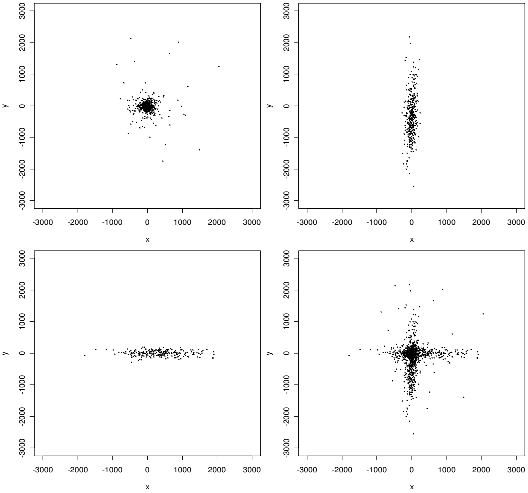
Fitted model:

- Normal and t mixtures of order 1 to 8 were fitted
- For a given order, the t mixture always fit better
- The 3- and 4-component t mixtures fit best (on the basis of BIC)
- Likelihood-based parameter estimates for the 3-component t mixture:

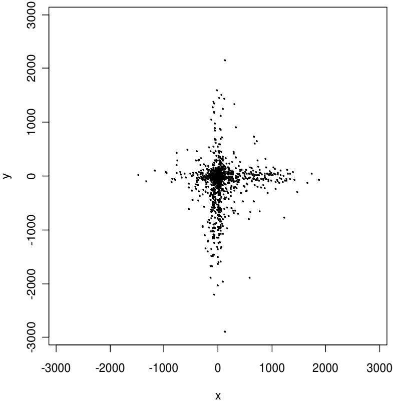
Component	p	μ_X	μ_Y	σ_X	σ_Y	ρ	df
1	0.571	-12.1	-10.7	61.6	54.1	-0.05	1.6
2	0.253	-4.7	-350.0	75.9	550.0	0.18	6.5
3	0.176	352.8	-12.6	540.3	84.9	-0.03	16.7

If errors are expressed relative to the alignment of the corresponding street (i.e. the N-S axis is rotated 90 degrees counterclockwise), a 2-component t mixture has even smaller BIC.

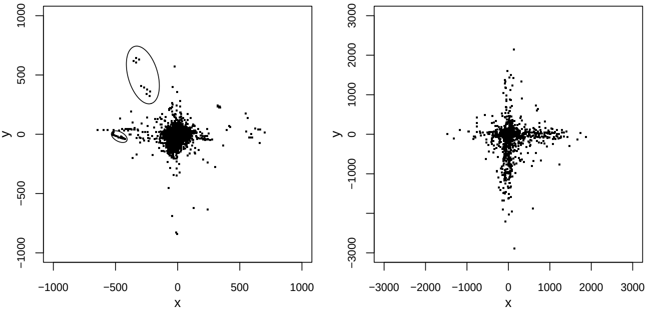
Simulated data from the fitted 3-component t mixture model



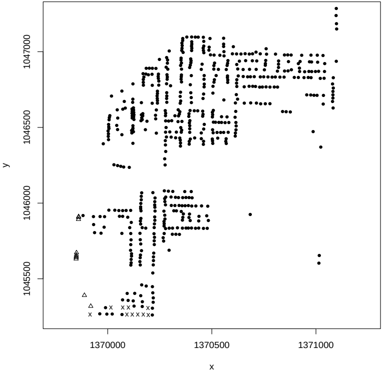
Comparison to observed error distribution:



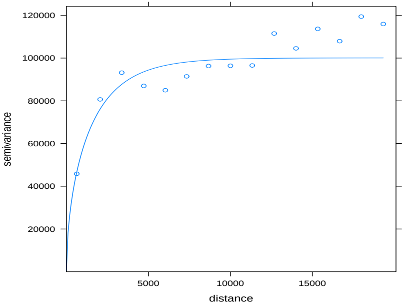
Spatial autocorrelation among geocoding errors



Locations of addresses corresponding to the larger ellipse:



Variogram of rural positional errors:



Effects of spatial autocorrelation among geocoding errors

Generally, spatial autocorrelation among the geocoding errors mitigates the loss of power attributable to them.

Demonstration:

- Superimposed upon a subset of 998 rural addresses in southern portion of Carroll County, we simulate 1000 realizations of a spatially clustered binary (cases and controls) disease process
- Realizations generated from a Gaussian random field threshold model, with varying proportions of cases (π) and correlation ranges
- Compute empirical power for Cuzick-Edwards test for disease clustering

- Repeat for geocoded addresses, and for locations obtained by independent sampling from the empirical distribution of 998 geocoding errors

		Address locations		
Range	Diseased fraction	Ground-truthed	Geocoded	Indep. error
500	0.017	0.664	0.517	0.443
	0.033	0.831	0.629	0.540
	0.067	0.987	0.918	0.874
1000	0.017	0.787	0.701	0.619
	0.033	0.944	0.842	0.771
	0.067	1.000	0.988	0.976
2000	0.017	0.895	0.813	0.780
	0.033	0.985	0.936	0.917
	0.067	1.000	0.994	0.995

Summary of features of geocoding errors

Geocoding errors have been found to be:

- non-normally distributed
- spatially autocorrelated
- systematically related to geographic covariates (e.g. rurality, length of street segment)

Proper studies of effects of geocoding errors, and the development of methods for incorporating them in analyses, should account for these features.

Measurement-error methods for intensity estimation

Consider a 2-D inhomogeneous Poisson process on study area D, with intensity function

$$\lambda(\mathbf{s}) = \lim_{|b(\mathbf{s})| \to 0} \left(\frac{E[N\{b(\mathbf{s})\}]}{|b(\mathbf{s})|} \right),$$

where $b(\mathbf{s})$ is a circular region centered at $\mathbf{s} \in D$ and $N(B)$ is the number of events that occur in a region B of area $|B|$.

Assume the intensity belongs to a parametric family $\{\lambda(\mathbf{s};\theta) : \theta \in \Theta\}$.

Let $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ represent the true locations of the n events that occur in D. Assume the geocoding is complete.

If we observe locations without error, the likelihood function is proportional to

$$L(\theta;\mathbf{s}_1,\ldots,\mathbf{s}_n) = \exp\left\{-\int_D \lambda(\mathbf{s};\theta)\,d\mathbf{s}\right\}\left\{\prod_{i=1}^{n}\lambda(\mathbf{s}_i;\theta)\right\}.$$

Suppose we don't observe $\mathbf{s}_1,\ldots,\mathbf{s}_n$ but instead observe perturbed versions $\mathbf{u}_1,\ldots,\mathbf{u}_n$. Also suppose that conditional on $\mathbf{s}_1,\ldots,\mathbf{s}_n$, the $\mathbf{u}_i$ are independent and each has pdf $g(\mathbf{u}|\mathbf{s}_i,\mathbf{x}_i,\beta,\tau)$ where $\mathbf{x}_i$ is a vector of observed covariates with associated parameters β, and τ is a vector of dispersion parameters.

Then the unconditional joint likelihood of the observed locations is proportional to

$$
\begin{aligned}
L_E(\theta, \beta, \tau; \mathbf{u}_1, \ldots \mathbf{u}_n) &= \exp\left\{ -\int_D \lambda(\mathbf{s}; \theta)\, d\mathbf{s} \right\} \\
&\quad \times \prod_{i=1}^{n} \int_D \lambda(\mathbf{s}_i; \theta) g(\mathbf{u}_i | \mathbf{s}_i, \mathbf{x}_i, \beta, \tau)\, d\mathbf{s}_i.
\end{aligned}
$$

We have found that the error disperson standard deviation needs to be of magnitude at least 5% of the dimensions of the study region in order for it to be worthwhile to account for geocoding errors.

Incompleteness of street geocoding

In addition to measurement errors, another practical reality of geocoding is incompleteness.

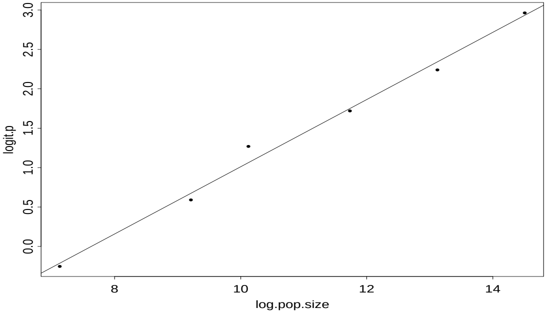
Two main features of the incompleteness:

- Relationship to local population density
- Spatial clustering of the failure to geocode
- However, we do generally have coarse geographic information (e.g. Zip codes) for the addresses that fail to geocode

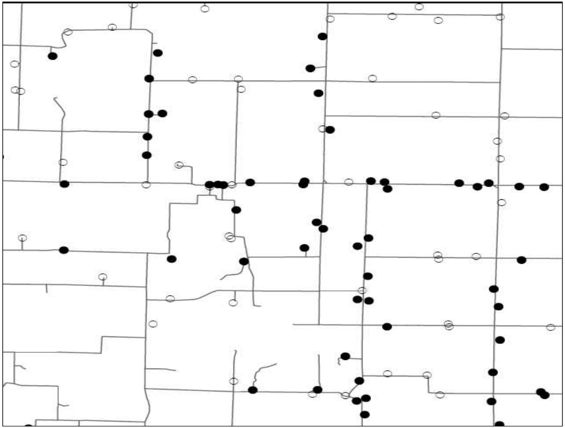
Incompleteness versus population size

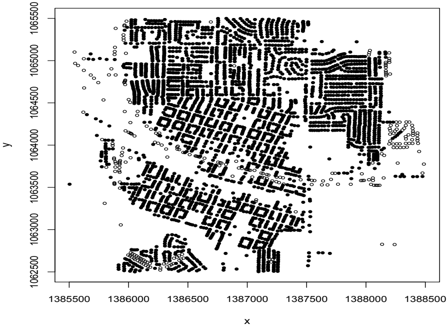
Geocoding success in the National Health Interview Survey (NHIS) (Kravets and Hadden, 2006):

Pop. size	*# counties*	*# households*	*% geocoded*
$\geq$ 1 million	300	138,281	95.1
250,000-999,999	194	48,992	90.4
50,000-249,999	106	23,379	84.8
20,000-49,999	76	17,625	78.1
2,500-19,999	110	19,805	64.4
$<$ 2,500	48	4,339	43.7


logit.p
0.0
0.5
1.0
1.5
2.0
2.5
3.0
8
10
12
14
log.pop.size

Spatial clustering of the failure to geocode




1062500
1063000
1063500
1064000
1064500
1065000
1065500
y
1385500
1386000
1386500
1387000
1387500
1388000
1388500
x

Effect of clustering of geocoding failure on detection of disease clustering

Empirical powers of size-.05 Cuzick-Edwards test for spatial clustering of disease cases (based on a rural subset of 998 addresses):

		$\pi = 0.01$		$\pi = 0.04$	
Data set	k	$r = 1000$	$r = 3333$	$r = 1000$	$r = 3333$
Complete	—	0.786	0.961	0.999	1.000
Geocoded	1	0.593	0.856	0.932	0.980
Geocoded	2	0.408	0.600	0.868	0.979
Geocoded	3	0.266	0.408	0.668	0.873

Here π is the disease case prevalence of a Gaussian random field threshold model, r is the correlation range, k is the case/control odds ratio of geocoding failure.

Nonparametric intensity estimation

- We wish to exploit the coarsened data (Zip codes) to improve nonparametric (kernel-based) estimation of the intensity.

- A generic kernel intensity estimator:

$$\tilde{\lambda}(\mathbf{s}) = \sum_{i=1}^{n} K_h(\mathbf{s} - \mathbf{s}_i) \equiv \sum_{i=1}^{n} h^{-1} K(h^{-1} \|\mathbf{s} - \mathbf{s}_i\|)$$

  where $K(\cdot)$ is a univariate symmetric kernel function and h is the bandwidth

- Used routinely for EDA of spatial point patterns (to look for peaks and troughs, associations with maps of covariates, etc.)

- $\tilde{\lambda}(\mathbf{s})$ is asymptotically unbiased if geocoding is complete

Additional notation

- For each $\mathbf{s} \in D$, define a geocoding indicator random variable

$$G(\mathbf{s}) = \begin{cases} 1, & \text{if an event at site } \mathbf{s} \text{ geocodes} \\ 0, & \text{otherwise.} \end{cases}$$

- Geocoding propensity function:

$$\phi(\mathbf{s}) = P\{G(\mathbf{s}) = 1\}.$$

  * $\phi(\mathbf{s}) \equiv 1.0 \Rightarrow$ Geocoding complete, no geographic bias
  * $\phi(\mathbf{s}) \equiv c < 1.0 \Rightarrow$ Geocoding incomplete, $\tilde{\lambda}(\mathbf{s})$ biased but not geographically so
  * $\phi(\mathbf{s}) \neq c \Rightarrow$ Geocoding incomplete, $\tilde{\lambda}(\mathbf{s})$ geographically biased

Relationship to thinned processes

- Let g_i be the observed value of $G(\mathbf{s}_i)$, and define $\mathscr{G} = \{i : g_i = 1\}$

- The events that geocode, i.e. $\{\mathbf{s}_i : i \in \mathscr{G}\}$, constitute a realization of a “thinned” point process, or more specifically an independently $\phi(\mathbf{s})$-thinned process

- Let $\lambda_T(\mathbf{s})$ denote the intensity function for the thinned process associated with the incompletely geocoded data

- Key result:
$$\lambda_T(\mathbf{s}) = \phi(\mathbf{s})\lambda(\mathbf{s})$$

Kernel intensity estimation for a thinned process

- Consider two modified kernel intensity estimators (assuming $\phi(\cdot)$ is known):

$$\bar{\lambda}(\mathbf{s}) = \{\phi(\mathbf{s})\}^{-1} \sum_{\substack{i=1 \\ i \in \mathscr{G}}}^{n} K_h(\mathbf{s} - \mathbf{s}_i)$$

  and

$$\hat{\lambda}(\mathbf{s}) = \sum_{\substack{i=1 \\ i \in \mathscr{G}}}^{n} \{\phi(\mathbf{s}_i)\}^{-1} K_h(\mathbf{s} - \mathbf{s}_i)$$

- $\hat{\lambda}(\mathbf{s})$ has better properties, so we consider it exclusively

- Although $\phi(\cdot)$ is generally unknown in practice, it can be estimated using the coarsened data, and the resulting estimate can be substituted into $\hat{\lambda}(\mathbf{s})$ to yield a coarsened-data estimator

$$\hat{\lambda}_C(\mathbf{s}) = \sum_{\substack{i=1 \\ i \in \mathscr{G}}}^{n} \{\hat{\phi}(\mathbf{s}_i)\}^{-1} K_h(\mathbf{s} - \mathbf{s}_i)$$

Carroll County example

- Incomplete-data and coarsened-data kernel intensity estimates were computed

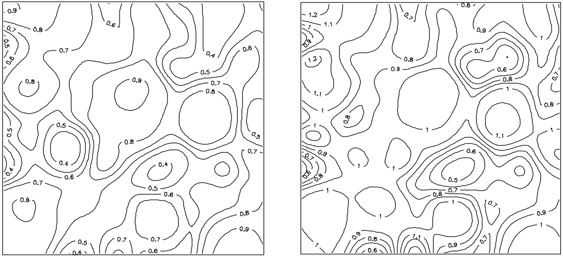
- Complete-data estimate was also computed, using geocodes obtained via aerial orthophotographs

- Complete-data estimate serves as a benchmark for comparing the other two estimates

Proportion of Carroll County addresses that geocoded, by Zip code:

Zip code	Number of addresses	Proportion geocoded
50050	3	0.667
50058	817	0.889
51401	5150	0.851
51430	359	0.738
51433	3	0.000
51436	363	0.766
51440	200	0.715
51443	829	0.779
51444	110	0.173
51449	29	0.448
51451	93	0.312
51453	17	0.529
51455	918	0.773
51459	41	0.293
51462	16	0.688
51463	298	0.591
51466	4	0.500
51467	48	0.354

Pointwise ratio of incomplete-data kernel intensity estimate (left panel) and coarsened-data kernel intensity estimate (right panel) to complete-data kernel intensity estimate (the integrated MSE of the latter is less than 50% that of the former):



Conclusions and other developments

- Incomplete geocoding can be successfully dealt with using a coarsened-data approach
- Coarsened-data kernel intensity estimator offers substantial improvements in inferences
- We've also developed a coarsened-data maximum likelihood approach for parametric intensity estimation; if the data are coarsened at random then the likelihood is proportional to

$$L_{IG}(\theta;\mathbf{X}) = \exp\left\{-\int_D \lambda(\mathbf{s};\theta)\,d\mathbf{s}\right\}\left\{\prod_{i\in\mathscr{G}}\lambda(\mathbf{s}_i;\theta)\right\}\left\{\prod_{i\notin\mathscr{G}}\int_{Z_i}\lambda(\mathbf{s};\theta)\,d\mathbf{s}\right\}$$

($X_i = \mathbf{s}_i$ if $g_i = 1$, and $X_i = Z_i$ otherwise).

- We've also developed coarsened-data methods (both nonparametric and likelihood-based) for estimation of spatial variation in log relative risk. The nonparametric estimator, e.g., is

$$\hat{\rho}_C(\mathbf{s}) = \log\left(\frac{\hat{\lambda}_{C1}(\mathbf{s})}{\hat{\lambda}_{C0}(\mathbf{s})}\right)$$