Understanding Urban Pollution Through Spatial Modeling

Julia Schedler and Katherine Ensor

July 29, 2019

How can urban data be leveraged to help city and community officials manage the impact of pollution?

RICE

Analysis at Actionable Levels

The geography and layout of a particular city affects how urban data should be:

analyzed to account for geographic features such as waterways or roadways

presented at a useful level such as well-known neighborhoods or communities

Motivating Application: Hurricane Harvey Health Outcomes

Data were collected via online surveys about the impact of Hurricane Harvey, including whether the respondent experienced:

- Trouble concentrating or sleeping
- A runny nose, headache, shortness of breath, or skin rash
- Flooding in their home
- Damage as a result of the storm
- Displacement as a result of the storm

Modeling Spatial Data

Some issues to consider when modeling spatial data:

- Appropriate methods for type of spatial data (point level, lattice, point pattern)
- Accounting for Spatial Dependence / Effective Sample Size
- Choice of distance metric

This talk focuses on the analysis of lattice data.

Spatial Regression for Aggregated Data

Two popular forms of SAR (simultaneous autoregressive) models:

Spatial Errors $y = X\beta + \varepsilon; \varepsilon = \lambda W \varepsilon + u$

Spatial Lag $y = \rho W y + X\beta + v$

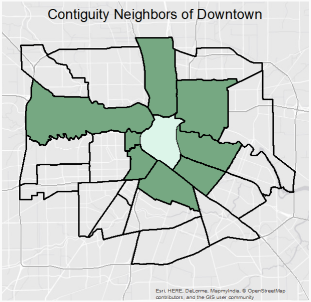
Model region i as a function of all other regions, with weight w_{ij} capturing spatial structure.

- If w_{ij} is not 0, regions i and j are "neighbors"
- Neighbors should be “close” together

Specify "closeness" by specifying neighbors of each region:

Contiguity

- First order: regions share at least one boundary point
- Second order: regions who are neighbors of first order neighbors

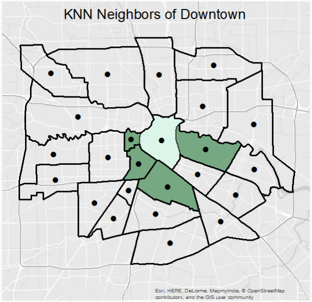


Aggregated Modeling

Specify "closeness" by specifying neighbors of each region:

K nearest neighbors

- Choose the k "closest" regions
- "closest" is based on centroid distances

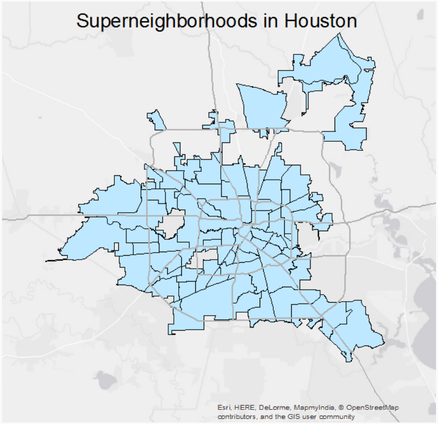


RICE

Issues with Contiguity & KNN

Contiguity

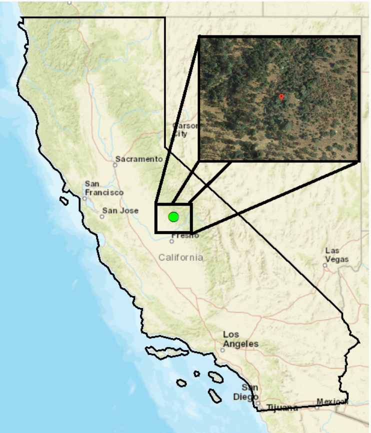
- Number of neighbors varies by size of region
- Works best when lattice is close to regular
- ignores holes or islands



Issues with Contiguity & KNN

KNN

- Centroid: single point to represent a set
  - can lie in a "remote" part of region
  - can lie *outside* the set if region is non-convex
- How to choose k?



Is there a simple way to generate a spatial weight matrix among regions that respects their geometry?

RICE

Distance between sets

Hausdorff distance measures the distances between sets as "worst case scenario" in terms of an underlying distance metric.

$$
\begin{aligned}
H(A,B) &= \max\{h(A,B), h(B,A)\} \\
&= \max\{\max_{p_a \in A} \min_{p_b \in B} d(p_a,p_b), \max_{p_b \in B} \min_{p_a \in A} d(p_a,p_b)\}
\end{aligned}
$$

- Typically used in GIS applications or image recognition
- Has not been used to generate spatial weight matrices (until now)

RICE

Issues revisited

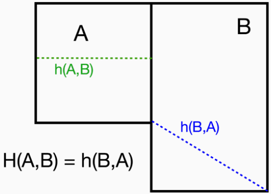
Using Hausdorff distance to define W

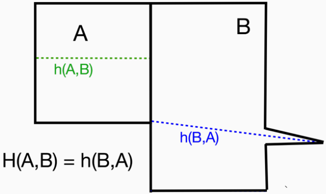
- Handles islands
- Avoids arbitrary centroid selection
- Retains flexibility of underlying distance metric

Doesn't do well with irregular geometry...

RICE

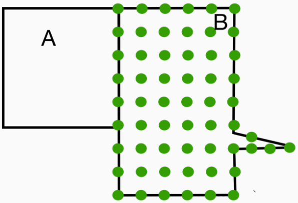
Hausdorff distance is sensitive to irregular geometry

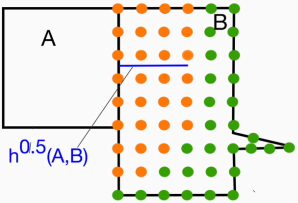




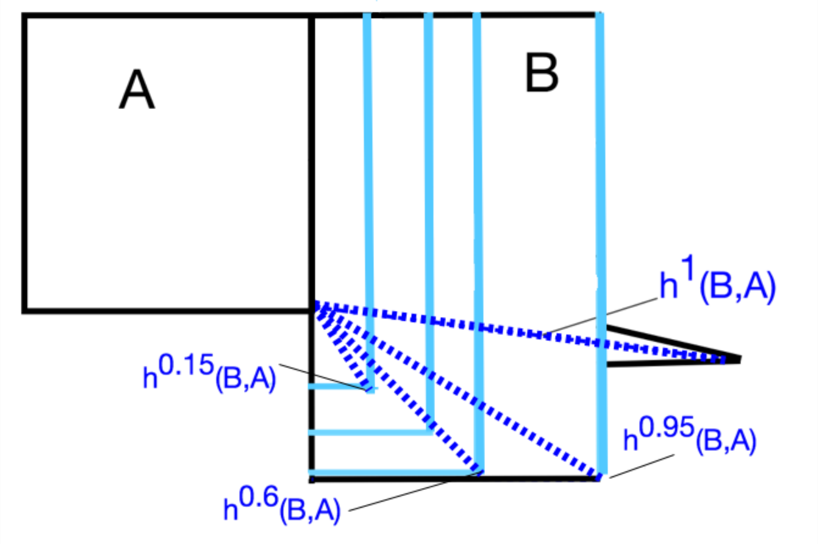
Idea: instead of using the maximum distance, use another statistic?

Extended Hausdorff Distance





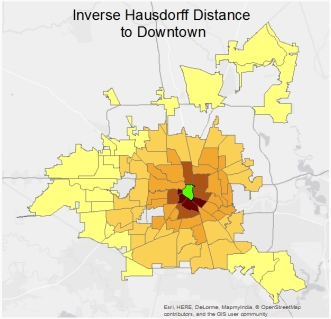
Extended Hausdorff Distance



Constructing W with Extended Hausdorff Distance

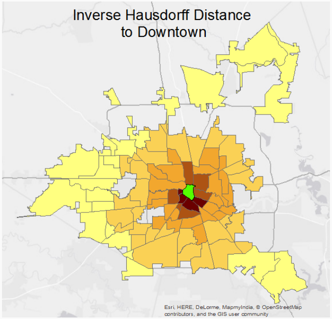
Define the entries of W as:

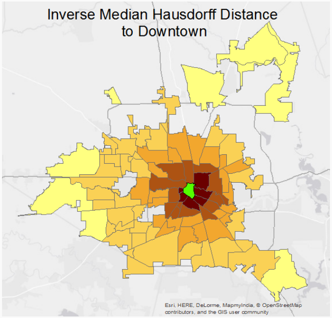
- The inverse of the Hausdorff distance; regions which are closer to region i will have larger weights in the i^{th} row of the weight matrix.
- KNN based on Hausdorff distance; the k closest regions to i will have nonzero entries in W

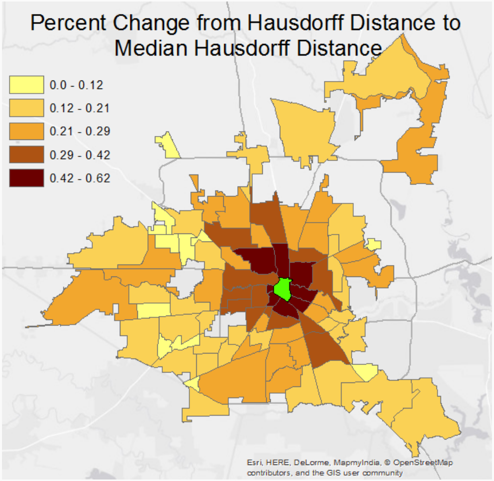


RICE

Comparing Hausdorff Distance to Median Hausdorff Distance







Considerations for using Hausdorff matrices

Computation $\binom{n}{2}$ computations where n is number of regions, but only needs to be done once per lattice/distance metric/percent area

Implementation `hausdorff` R package in development; works with existing spatial packages

Model How do various Hausdorff matrices affect model performance?

Simulations

Fitted the spatial errors and spatial lag model to data generated while varying the following:

Underlying Model Spatial Error, Spatial Lag

Weight Matrix Contiguity and KNN(k = 4) using: Centroid, Hausdorff, median Hausdorff

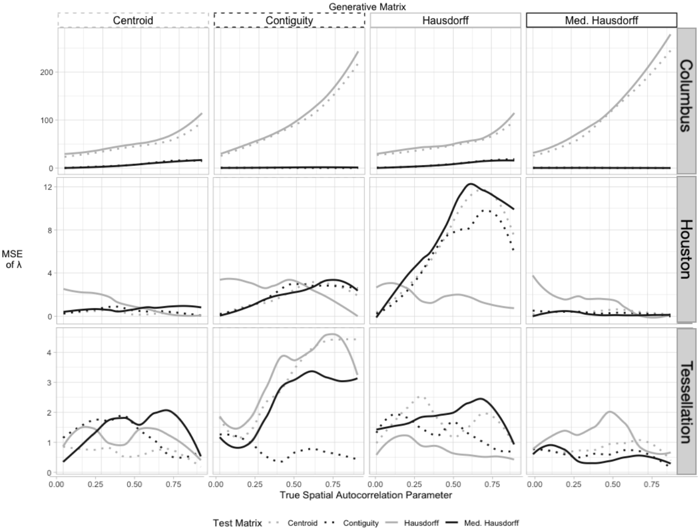
ρ/λ from 0 to 0.9 increments of 0.1

Lattice Houston Super-neighborhoods (n = 88), Columbus neighborhoods (n=49), random tessellations (n = 50)

Findings

- Results for random tessellations are not necessarily applicable to real-life lattices (tessellations are too "regular")
- The "best" (in terms of parameter estimation) weight matrix specification varies depending on the lattice

Findings



Generative Matrix
Centroid
Contiguity
Hausdorff
Med. Hausdorff
Columbus
Houston
Tessellation
MSE of λ
0
100
200
0
4
8
12
0
1
2
3
4
0.00
0.25
0.50
0.75
True Spatial Autocorrelation Parameter
Test Matrix
Centroid
Contiguity
Hausdorff
Med. Hausdorff

Hurricane Harvey Health Outcomes

HURRICANE HARVEY REGISTRY

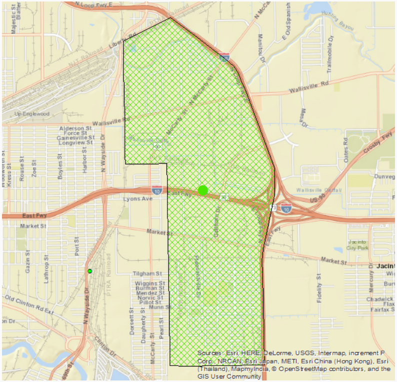
Fit a spatial regression model using various weight matrices using data collected from the HHR:

- Dependent Variable: Responses reporting trouble concentrating
- Independent Variables:
    - Estimate of probability of E. coli exposure
    - Responses reporting their home flooded
    - Responses indicating they were displaced

RICE

Another Use for (Extended) Hausdorff Distance

Use (extended) Hausdorff distance to generate spatial covariates for superneighborhoods, e.g. “distance to closest road” or “distance to bayou”



Application

It's clear that the statistical significance of the predictors in the model does not depend on which weight matrix is used. All weight matrices were able to account for the spatial dependence in the data.

	Contiguity*			Centroid			Hausdorff			Median Hausdorff		
	Estimate	Std. Err	P-value	Estimate	Std. Err	P-value	Estimate	Std. Err	P-value	Estimate	Std. Err	P-value
Intercept	1.289	0.197	<0.001	1.326	0.200	<0.001	1.341	0.200	<0.001	1.354	0.198	<0.001
Log(Dist Bayou)	0.093	0.048	0.054	0.116	0.047	0.013	0.125	0.047	0.008	0.117	0.047	0.014
Log(E. coli)	2.471	0.406	<0.001	2.424	0.408	<0.001	2.451	0.412	<0.001	2.551	0.403	<0.001
Log(Damaged)	0.560	0.119	<0.001	0.555	0.122	<0.001	0.546	0.122	<0.001	0.539	0.123	<0.001
Log(Displaced)	-0.018	0.083	0.827	-0.008	0.086	0.925	-0.001	0.085	0.993	-0.006	0.086	0.943
Lambda	0.300	0.145	0.093	0.267	0.170	0.214	0.290	0.171	0.162	0.196	0.176	0.323
Moran (residuals)	0.005	-	0.411	0.007	-	0.381	0.001	-	0.409	<0.001	-	0.419
AIC	29.076	-	-	30.358	-	-	29.942	-		30.922	-	-

* n = 64; 2 regions were islands with no neighbors.

In this case, the contiguity model seems to provide the best fit going off of AIC; the Hausdorff model is comparable and preferable in the sense that no regions were deleted.

Data Products

Extended Hausdorff weight matrices for a given lattice , underlying distance metric, and percentage can be computed once and stored on a data repository. For a given lattice, covariates based on Extended Hausdorff can be stored as well.

Future Directions

- Incorporate different distance metrics into `hausdorff` package
- Investigate selection of extended Hausdorff cutoff
- Evaluate performance of weight matrices via cross validation
    - for lattices with varying irregularity

Summary

- Exisiting methods for analysis of spatially aggregated data are not equipped to handle realities of real data
- The Hausdorff distance and Extended Hausdorff distance can handle these situations
- Computation can be lengthy, but only need to do it once
- (Extended) Hausdorff distance can accommodate any underlying distance metric

Questions?