

Spatial Tools for Case Selection

Using LISA Statistics to Design Mixed-Methods Research

Imke Harbers ¹ Matthew C. Ingram ²

¹University of Amsterdam

²University at Albany, SUNY

American University
June 14, 2019

Introduction

Two Trends:

- (1) growing emphasis on **mixed-methods designs**
- (2) growing emphasis on **interdependence**, including geographic interdependence and on spatial analysis as a way to approach this interdependence

Yet, **little attention to mixed-methods research designs with spatially dependent data.**

Elsewhere, we have offered two strategies for doing this (Harbers & Ingram 2017, in *Political Analysis*)

Here, we offer two case selection strategies to integrate (a) spatial statistics with (b) qualitative analysis.

Motivation: Analytic Issues

4 perspectives on spatial dependence:

(1) Benign nuisance

- know interdependence is out there, but not substantively interested in it and assume no meaningful impact

(2) Threat to inference

- know interdependence is out there, don't have a substantive interest in it, but acknowledge that it undermines valid inferences, so account for it

(3) Substantive interest

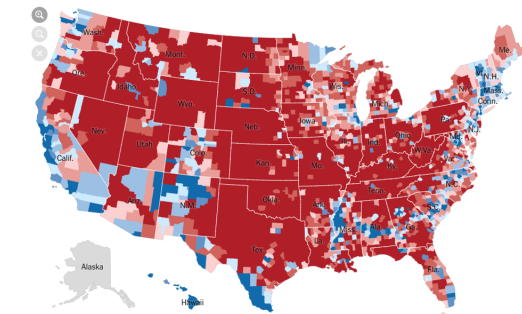
- interdependence is a key feature of phenomenon of interest and theory, and want to test effects, e.g., diffusion (theory-testing approach in "Geo-Nested Analysis")

(4) Substantive interest, but in theory-building mode
(approach in this paper)

Motivation: Spatial Dependence as Given

Interdependence inheres in social phenomena, and most social science data are likely spatial data.

Outcomes we care about are clustered in space (e.g., voting).



Motivation

Audience

- Primarily mixed-method researchers, but also quantitative researchers working with spatial data

Premises

- Geography or context as placeholder for variables yet to be uncovered
- Agnostic about reasons for spatial dependence

Running example

- County-level homicide rates in the US (Baller et al. 2001)

Table 1: Overview of Options

<i>Option</i>	<i>Motivation & Logic</i>	<i>Analytic Properties of Cases</i>	<i>Value added of Spatial Approach</i>	<i>Steps</i>	<i>Prerequisites</i>
1 <i>LISA statistics (and maps) of outcome of interest</i>	Identify clusters to visualize how an outcome of interest maps onto existing political or jurisdictional boundaries and is distributed across space <i>Guiding questions:</i> How does the spatial association (clustering) of an outcome of interest map onto political boundaries? What is the appropriate level of analysis for in-depth case studies? Are there cases that stand out from a spatial perspective, for instance because they defy regional patterns?	(1) <i>Selection based on LISA clusters:</i> a case consists of a set of connected units that form part of the same cluster For similarity clusters (high-high; low-low): Cases selected based on similar values on the dependent variable in connected units For dissimilarity clusters (low-high; high-low): Cases selected to ensure within-case variation on the dependent variable (2) <i>Selection based on LISA values and statistical significance:</i> A case consists of two or more connected units with extreme LISA values indicating strong association between neighboring units	Logic of selection aligns with mixed method designs where case selection is based on the value of the dependent variable (outcome of interest), but a “case” here consists of connected units to account for spatial dependence of the outcome and to provide leverage for understanding the origins of spatial patterns	Calculate LISA statistics, cluster identifiers, and generate Moran scatter plot and map	Geo-referenced data on outcome of interest
2 <i>LISA statistics (and maps) of residuals from baseline model</i>	Visualize spatially uneven performance of a model; identify high-high and low-low clusters to uncover omitted variables and scope conditions	<i>Selection based on LISA clusters:</i> a case consists of a set of connected units that form part of the same cluster	Logic of selection aligns with mixed method designs where case selection is based on residuals, but a “case” here consists of connected	Estimate baseline regression model; calculate LISA statistics	Geo-referenced data on outcome of interest and predictors for
	<i>Guiding questions:</i> How does the model perform across space? Which regions of the study area are well predicted? Which are poorly predicted? Are there clusters of over- and under-prediction? Do these clusters map onto political boundaries?	For similarity clusters (high-high; low-low): Cases selected based on clustering of residuals and model fit	units to account for the spatial structure of the error term; this facilitates access to case knowledge to think through what may be missing from the model	for residuals, cluster identifiers, and generate map	baseline, non-spatial model (e.g. OLS)

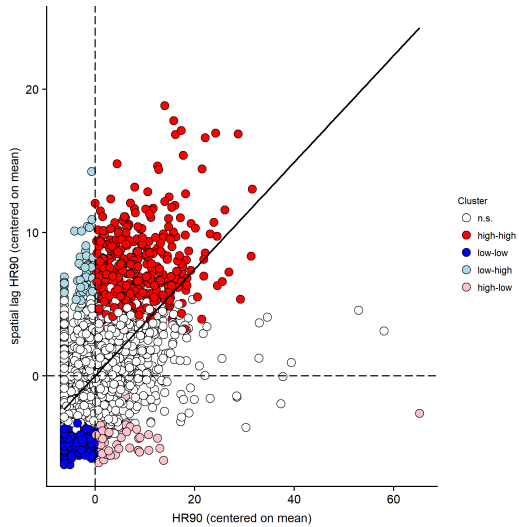
Option 1

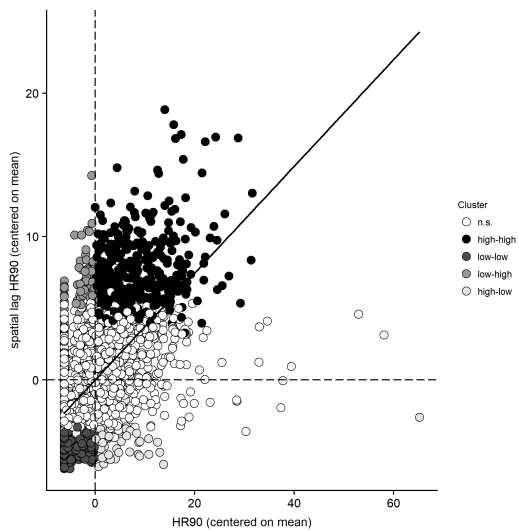
LISA statistics for outcome of interest.

Identify clusters to visualize how an outcome of interest is distributed geographically, including how it maps onto existing boundaries, e.g., administrative, political, jurisdictional.

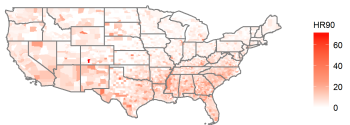
Guiding questions:

Is there clustering? How does the spatial association map onto political boundaries? What is the appropriate level of analysis for in-depth case studies? Do these patterns suggest scope conditions? Are there sites that stand out for one reason or another?

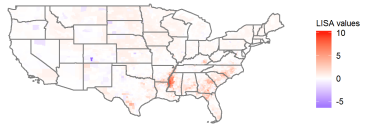




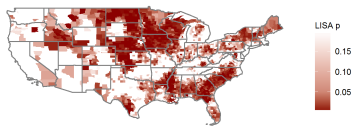
Homicide Rates (HR90)



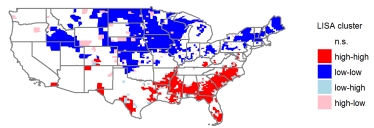
LISA Values (DV: HR90)



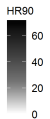
LISA p-values (DV: HR90)



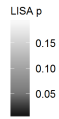
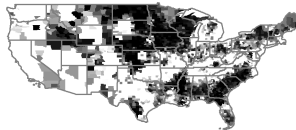
LISA Clusters (DV: HR90)



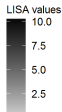
Homicide Rates (HR90)



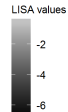
LISA p-values (DV: HR90)



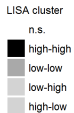
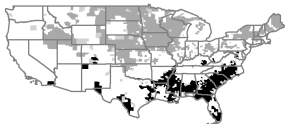
LISA Values, positive (DV: HR90)



LISA Values, negative (DV: HR90)



LISA Clusters (DV: HR90)



LISA Clusters (DV: HR90)

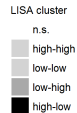


Table 2. Focal units selected based on LISA clusters**High-high**

FIPSNO	County	State	HR90	Wy	LISA	p
28055	Issaquena	Mississippi	34.92	23.06	10.397	0.002
22035	East Carroll	Louisiana	37.77	19.22	9.335	0.002
28151	Washington	Mississippi	30.42	23.12	9.286	0.002
28125	Sharkey	Mississippi	28.30	22.80	8.334	0.002
28083	Leflore	Mississippi	27.67	20.62	6.975	0.002

Low-low

FIPSNO	County	State	HR90	Wy	LISA	p
20007	Barber	Kansas	0	0.32	0.867	0.002
20109	Logan	Kansas	0	0	0.867	0.002
31101	Keith	Nebraska	0	0.09	0.867	0.002
31135	Perkins	Nebraska	0	0.08	0.867	0.002
46045	Edmunds	South Dakota	0	0.43	0.854	0.002

Low-high

FIPSNO	County	State	HR90	Wy	LISA	p
13183	Long	Georgia	0	13.08	-1.054	0.01
8007	Archuleta	Colorado	0	11.62	-1.013	0.008
48365	Panola	Texas	0	12.65	-0.947	0.018
21109	Jackson	Kentucky	0	12.72	-0.917	0.012
48271	Kinney	Texas	0	12.63	-0.904	0.016

High-low

FIPSNO	County	State	HR90	Wy	LISA	p
8053	Hinsdale	Colorado	71.38	3.58	-6.83	0.004
46063	Harding	South Dakota	19.97	0.29	-1.832	0.002
31173	Thurston	Nebraska	19.22	1.15	-1.455	0.004
17035	Cumberland	Illinois	18.74	1.84	-1.282	0.002
41023	Grant	Oregon	16.98	0.99	-1.253	0.002

Option 2

Assume have baseline, non-spatial model (e.g., OLS) and have extracted component that remains unexplained (residuals)

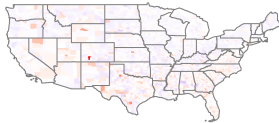
LISA statistics (and maps) of residuals.

Visualize spatially uneven performance of a model; identify high-high and low-low clusters to uncover omitted variables and scope conditions.

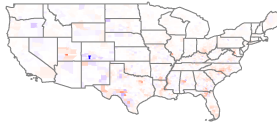
Guiding questions:

How does the model perform across space? Which regions of the study area are well predicted? Which are poorly predicted? Are there clusters of over- and under-prediction? Do these clusters map onto political boundaries?

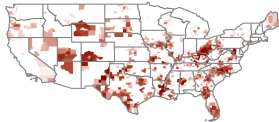
OLS Residuals)



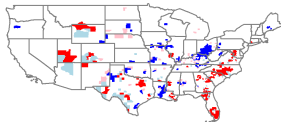
LISA Values (residuals)



LISA p-values (residuals)



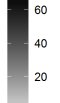
LISA Clusters (residuals)



OLS Residuals (positive)



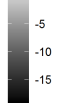
Residuals



OLS Residuals (negative)



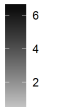
Residuals



LISA Values, positive (residuals)



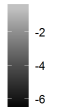
LISA values



LISA Values, negative (residuals)



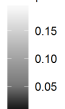
LISA values



LISA p-values (residuals)



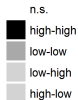
LISA p



LISA Clusters (residuals)



LISA cluster



LISA Clusters (residuals)



LISA cluster

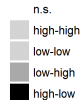


Table 3. Locations Selected based on LISA Clusters of Residuals**High-high**

FIPSNO	County	State	e	We	LISA	p
11001	Washington	D.C.	43.496	4.731	6.799	0.032
49031	Piute	Utah	27.718	3.024	4.334	0.042
24510	Baltimore City	Maryland	13.898	6.426	3.719	0.006
24033	Prince Georges	Maryland	12.395	7.178	3.650	0.004
48435	Sutton	Texas	20.895	2.700	3.501	0.032

Low-low

FIPSNO	County	State	e	We	LISA	p
1105	Perry	Alabama	-19.102	-4.249	2.840	0.004
28157	Wilkinson	Mississippi	-12.286	-5.344	2.650	0.002
1065	Hale	Alabama	-15.098	-4.185	2.462	0.002
22009	Avoyelles	Louisiana	-8.658	-4.968	1.902	0.002
22029	Concordia	Louisiana	-8.993	-5.418	1.899	0.002

Low-high

FIPSNO	County	State	e	We	LISA	p
47159	Smith	Tennessee	-0.135	3.257	-0.021	0.038
45047	Greenwood	South Carolina	-0.133	4.118	-0.022	0.016
47111	Macon	Tennessee	-0.295	3.448	-0.046	0.044
13317	Wilkes	Georgia	-0.392	4.681	-0.065	0.028
51700	Newport News	Virginia	-0.497	3.273	-0.068	0.046

High-low

FIPSNO	County	State	e	We	LISA	p
26051	Gladwin	Michigan	0.021	-2.566	-0.003	0.016
26117	Montcalm	Michigan	0.059	-2.525	-0.006	0.040
39103	Medina	Ohio	0.101	-2.617	-0.011	0.036
29217	Vernon	Missouri	0.126	-3.085	-0.015	0.028
46123	Tripp	South Dakota	0.182	-3.539	-0.027	0.006

Conclusions

Tools from spatial analysis can provide additional leverage for case selection.

- 1 Identify scope conditions
- 2 Clarify bound or unbound nature of phenomena
- 3 Examine causal mechanisms
- 4 Identify new, previously omitted variables, to generate new hypotheses and build theory

Core implication across all proposed strategies:

- need to redefine meaning of “case” as more than 1 unit of observation

Thank You

Spatial Tools for Case Selection

Using LISA Statistics to Design Mixed-Methods Research

Imke Harbers ¹ Matthew C. Ingram ²

¹University of Amsterdam

²University at Albany, SUNY

American University
June 14, 2019