ADVANCES IN SPATIAL DATA SCIENCE

Causality and reproducibility

LEVI JOHN WOLF

levi.john.wolf@bristol.ac.uk

LEVI JOHN WOLF (he/him)

Spatializing partisan gerrymandering forensics
PhD Arizona State University, 2017
University of Bristol & Alan Turing Institute (2017-2023)

CONSULTING & ENGAGEMENT

Nextdoor.com CARTO MondialRelay
Boundary Commission for England BCC Living Rent Commission

METHODS

Space-time exploratory data analysis
Spatial optimization
Probabilistic programming
Spatial machine learning
Bayesian spatial econometrics
Local modelling

SUBJECTS

Districting, voting, and elections
Housing and renting in cities
Economic inequality
Segregation, gentrification, and urban change
Species distribution modelling
Urban planning & governance

A little bit about me

THE CITY AND CAUSALITY

MUCH ADO ABOUT NULL THINGS

Rethinking causality in city science & spatial analysis

THE CITY AND CAUSALITY

what's replication got to do with it?

MUCH ADO ABOUT NULL THINGS

Rethinking causality in city science & spatial analysis

Environment and Planning B: Urban Analytics and City Science

Impact Factor: **3.5**
5-Year Impact Factor: **3.9**

Editorial board

Managing Editor

Michael Batty	University College London, UK

Editors

Seraphim Alvanides	Northumbria University, UK
Daniel Arribas-Bel	University of Liverpool, UK
Andrew Crooks	State University of New York at Buffalo
Linda See	International Institute for Applied Systems Analysis (IIASA), Austria
Levi Wolf	University of Bristol, UK

Cities, laws, and history

Environment and Planning B: Urban Analytics and City Science

Impact Factor: **3.5**
5-Year Impact Factor: **3.9**

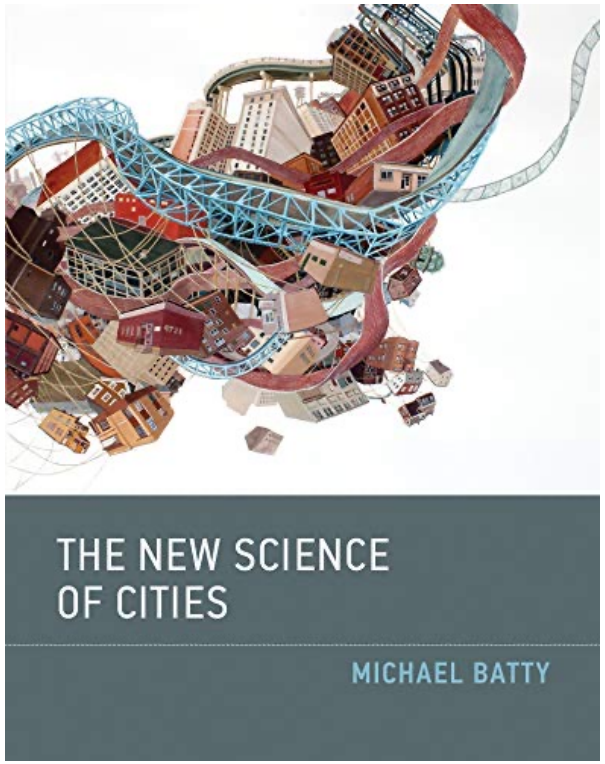
Editorial board

Managing Editor

Michael Batty	University College London, UK

Editors

Seraphim Alvanides	Northumbria University, UK
Daniel Arribas-Bel	University of Liverpool, UK
Andrew Crooks	State University of New York at Buffalo
Linda See	International Institute for Applied Systems Analysis (IIASA), Austria
Levi Wolf	University of Bristol, UK



Cities, laws, and history

Environment and Planning B: Urban Analytics and City Science

Impact Factor: **3.5**
5-Year Impact Factor: **3.9**

Editorial board

Managing Editor

Michael Batty	University College London, UK

Editors

Seraphim Alvanides	Northumbria University, UK
Daniel Arribas-Bel	University of Liverpool, UK
Andrew Crooks	State University of New York at Buffalo
Linda See	International Institute for Applied Systems Analysis (IIASA), Austria
Levi Wolf	University of Bristol, UK

Cities, laws, and history

Article

B Urban Analytics and City Science

A roundtable discussion: Defining urban data science

EPB: Urban Analytics and City Science
2019, Vol. 46(9) 1756–1768

DOI: 10.1177/2399808319882826
journals.sagepub.com/home/epb
SAGE

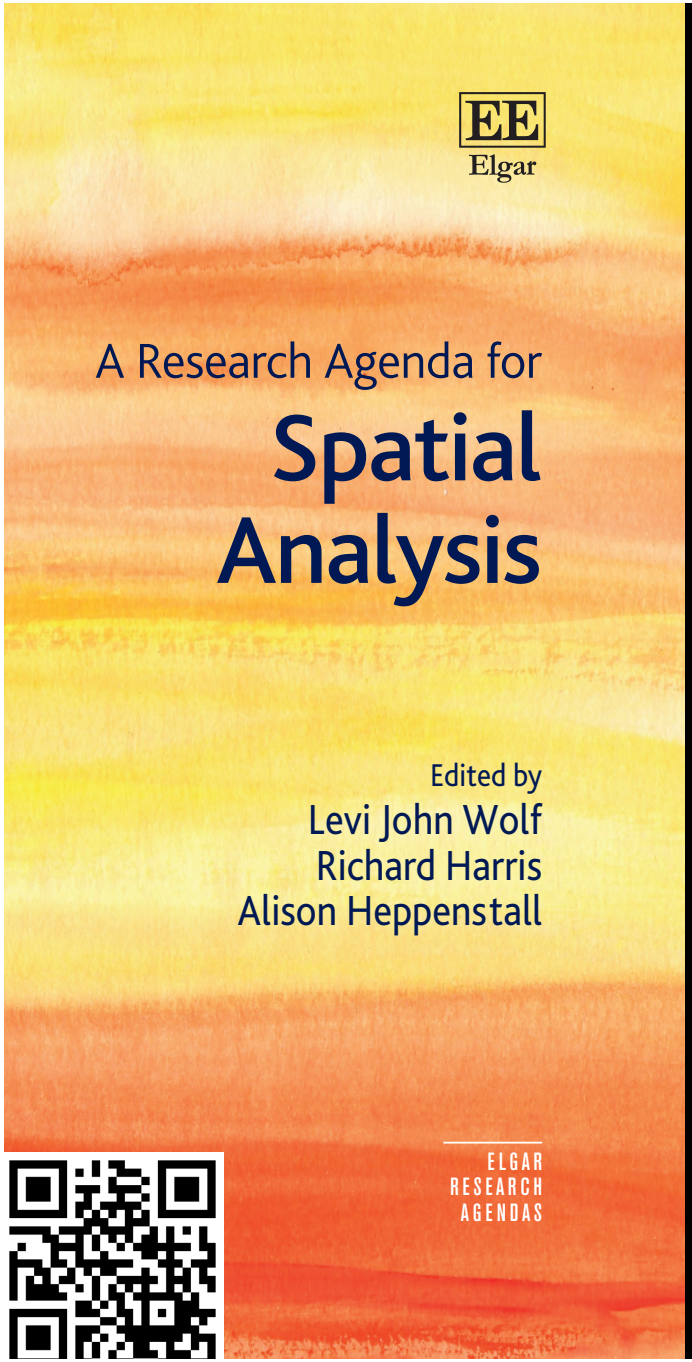
Organizers

Wei Kang, Taylor Oshan, Levi John Wolf

Participants

Geoff Boeing, Vanessa Frias-Martinez, Song Gao, Ate Poorthuis, Wenfei Xu

Urban Data Science is sometimes the dilettante practice of doing things with software and data about living in cities [...] to be a science, urban data science both has to build from theory and give back to theory.



A challenge for spatial analysis, as for many areas of applied social science, data science and statistics, is answering the question of causality ... what caused something to happen or arise where and when it did?
(Wolf, Harris, Heppenstall)

generative modelling done right can both contribute to a new form of causal inference and to the larger program of social sciences: the simultaneous search for generalised explanations of social phenomena and recognition of the uniqueness of historical events.
(Cottineau)

there are causal processes operating at a level (or levels) beyond the individual, but if the level included in an analysis does not correspond with the level(s) of the process then it is possible it will be mis-estimated
(Petrovic, Manley, & Van Ham)

Editorial

B Urban Analytics and City Science

Beyond open science: Data, code, and causality

Levi John Wolf

EPB: Urban Analytics and City Science
2023, Vol. 50(9) 2333–2336

DOI: 10.1177/23998083231210180
journals.sagepub.com/home/epb
Sage

Indeed, city science is relatively unusual in the social sciences, in that one can still often find work seeking 'laws'

A COMPUTER MOVIE SIMULATING URBAN GROWTH IN THE DETROIT REGION

W. R. TOBLER

University of Michigan

The LAWS *of* MIGRATION.
By E. G. RAVENSTEIN, ESQ., F.R.G.S.

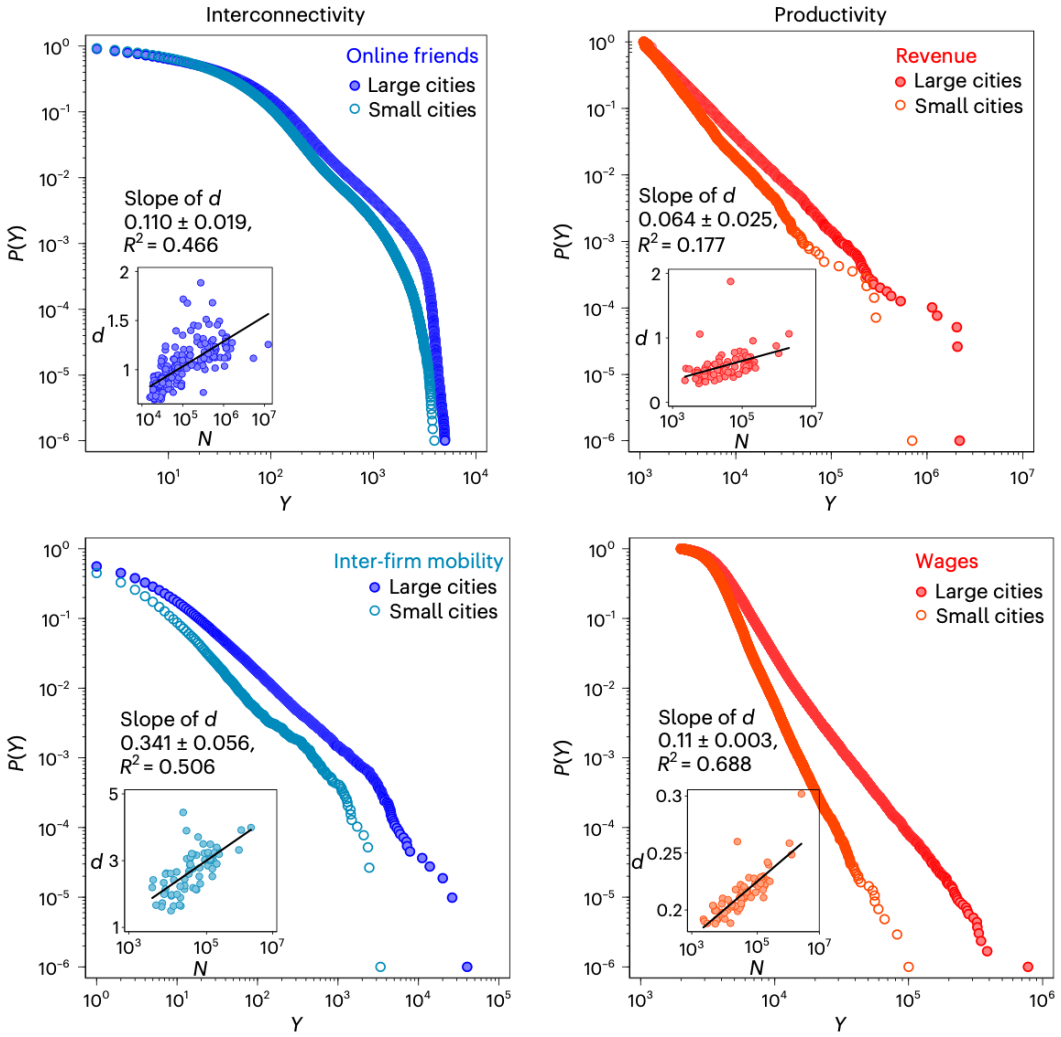
HUMAN BEHAVIOR

AND

THE PRINCIPLE OF LEAST EFFORT

An Introduction to Human Ecology

GEORGE KINGSLEY ZIPF, Ph.D.



Arvidsson et al. (2023)

laws

Replication across space and time must be weak in the social and environmental sciences

Michael F Goodchild & Wenwen Li

Proceedings of the National Academy of Sciences, 2021

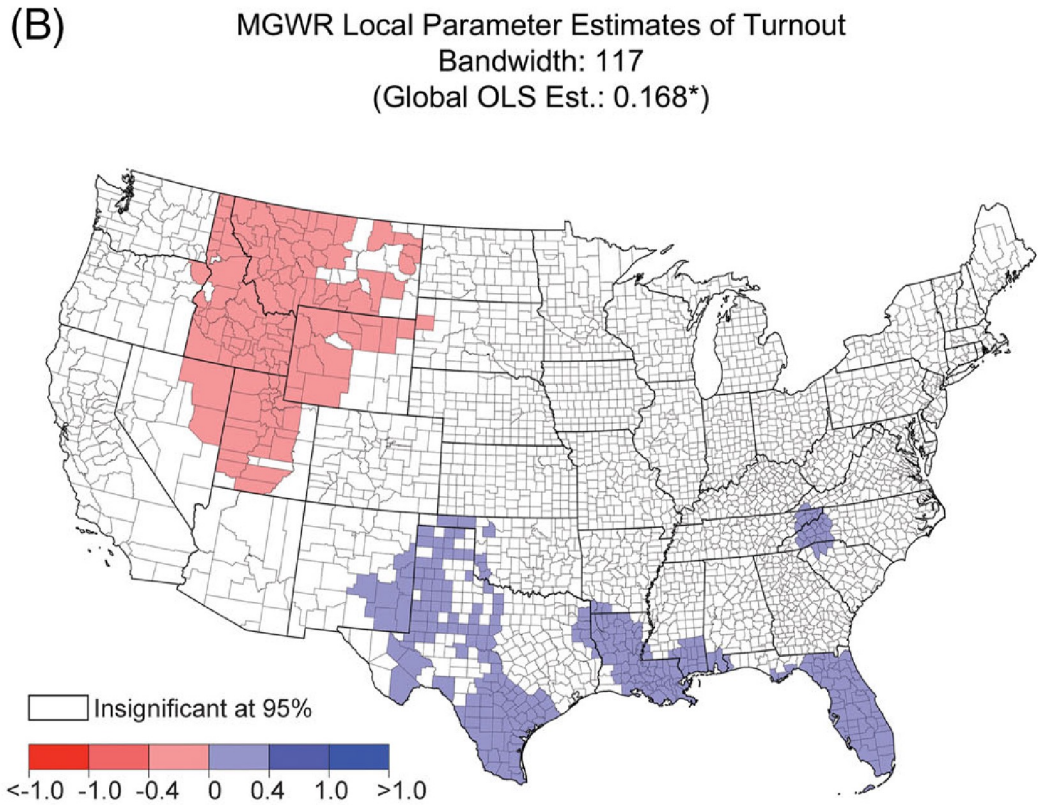
doi.org/10.1073/pnas.2015759118

GEOGRAPHIC INFORMATION SCIENCE II: Mesogeography: social physics, GIScience, and the quest for geographic knowledge

Harvey J. Miller

Progress in Human Geography, 2018

doi.org/10.1177/0309132517712154



Scale, Context, and Heterogeneity: A Spatial Analytical Perspective on the 2016 U.S. Presidential Election

A. Stewart Fotheringham,* Ziqi Li,† and Levi John Wolf‡

Contextual 'laws'?

Geography Compass

Rethinking Causality in Quantitative Human Geography

Mirah Zhang Levi John Wolf

Rather than accepting Goodchild & Li's "weak replicability," we must rigorously work together to understand where and why replicability fails. Spatial analysts must get specific about what "scale" or "context" does

preprint

Editorial

B Urban Analytics and City Science

Beyond open science: Data, code, and causality

EPB: Urban Analytics and City Science
2023, Vol. 50(9) 2333–2336

DOI: 10.1177/23998083231210180
journals.sagepub.com/home/epb
Sage

Levi John Wolf

Many contemporary definitions of causality, such as Lewis's "counterfactual theory of causation," will result in very different replication practices

nature human behaviour

Article

https://doi.org/10.1038/s41562-022-01509-1

Urban scaling laws arise from within-city inequalities

Received: 12 August 2021

Accepted: 6 December 2022

Martin Arvidsson [1], Niclas Lovsjö [1] & Marc Keuschnigg [1,2]

The causal processes underlying [inequality within cities] cities [is] an indispensable element of urban scaling

Law-as-cause	*fails if inequality does not drive scaling swh.*
"Weak Replicability" (Goodchild & Li 2021)	*still holds if inequality drives scaling differently (or not at all) in different places/times*
"Strong contextuality" (Zhang & Wolf, 2024)	*holds iff we can specify where/when inequality does not drive scaling*

Editorial

B Urban Analytics and City Science

Beyond open science: Data, code, and causality

EPB: Urban Analytics and City Science
2023, Vol. 50(9) 2333–2336

DOI: 10.1177/23998083231210180
journals.sagepub.com/home/epb
Sage

Levi John Wolf

To really achieve the open science aim of "replication," we need to have theories that can be replicated.

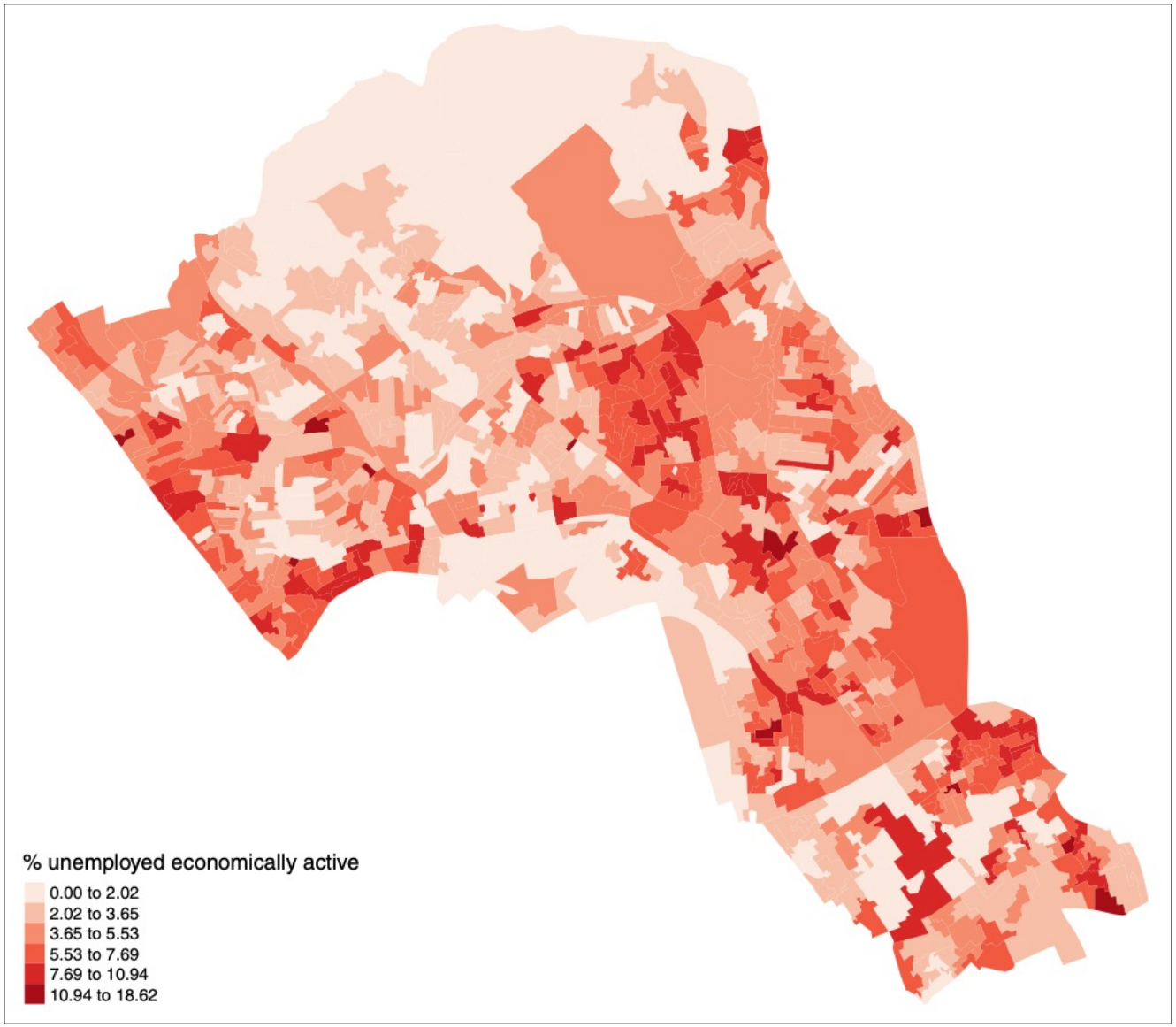
THE CITY AND CAUSALITY

what's replication got to do with it?

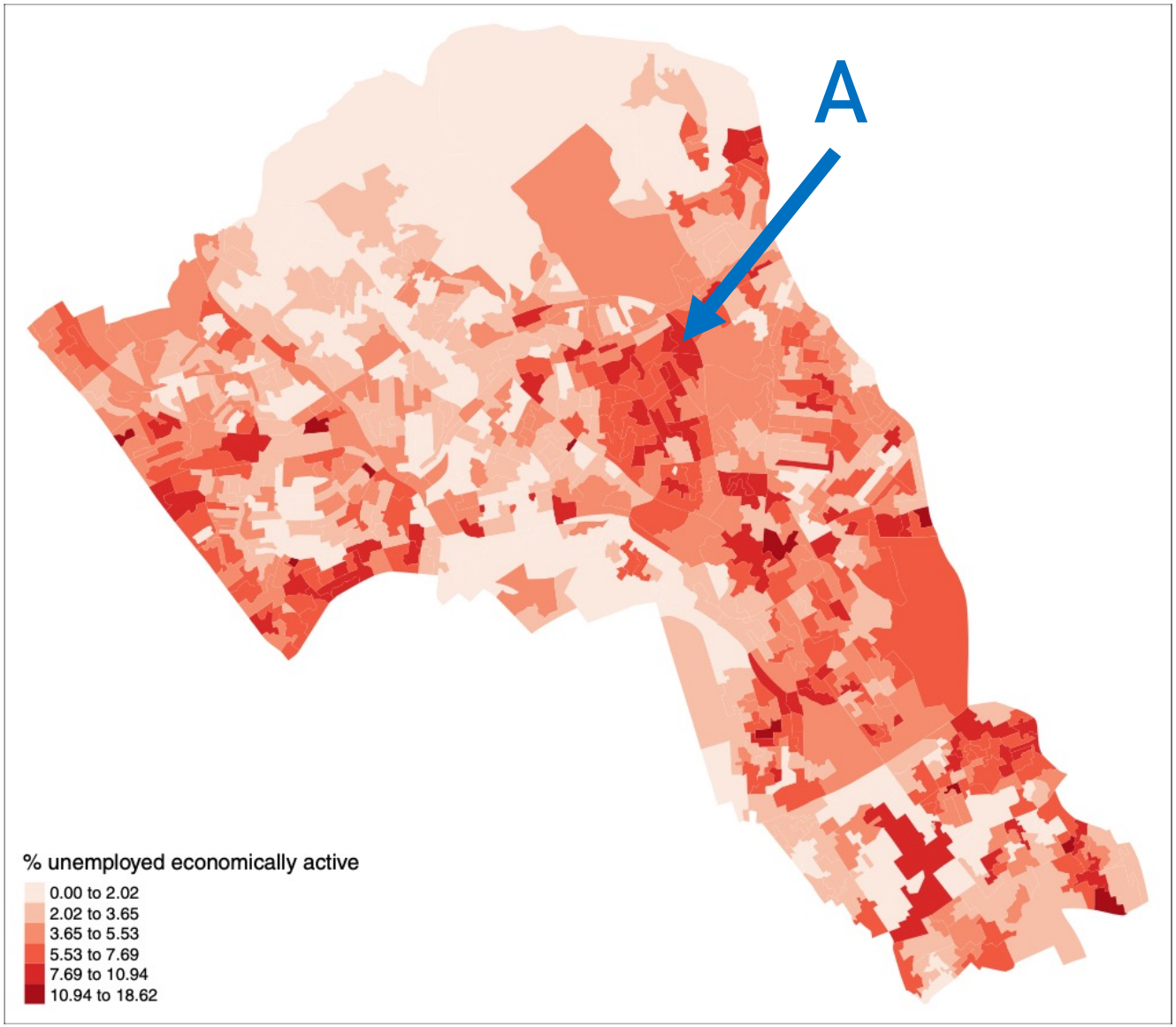
MUCH ADO ABOUT NULL THINGS

a recent replication challenge

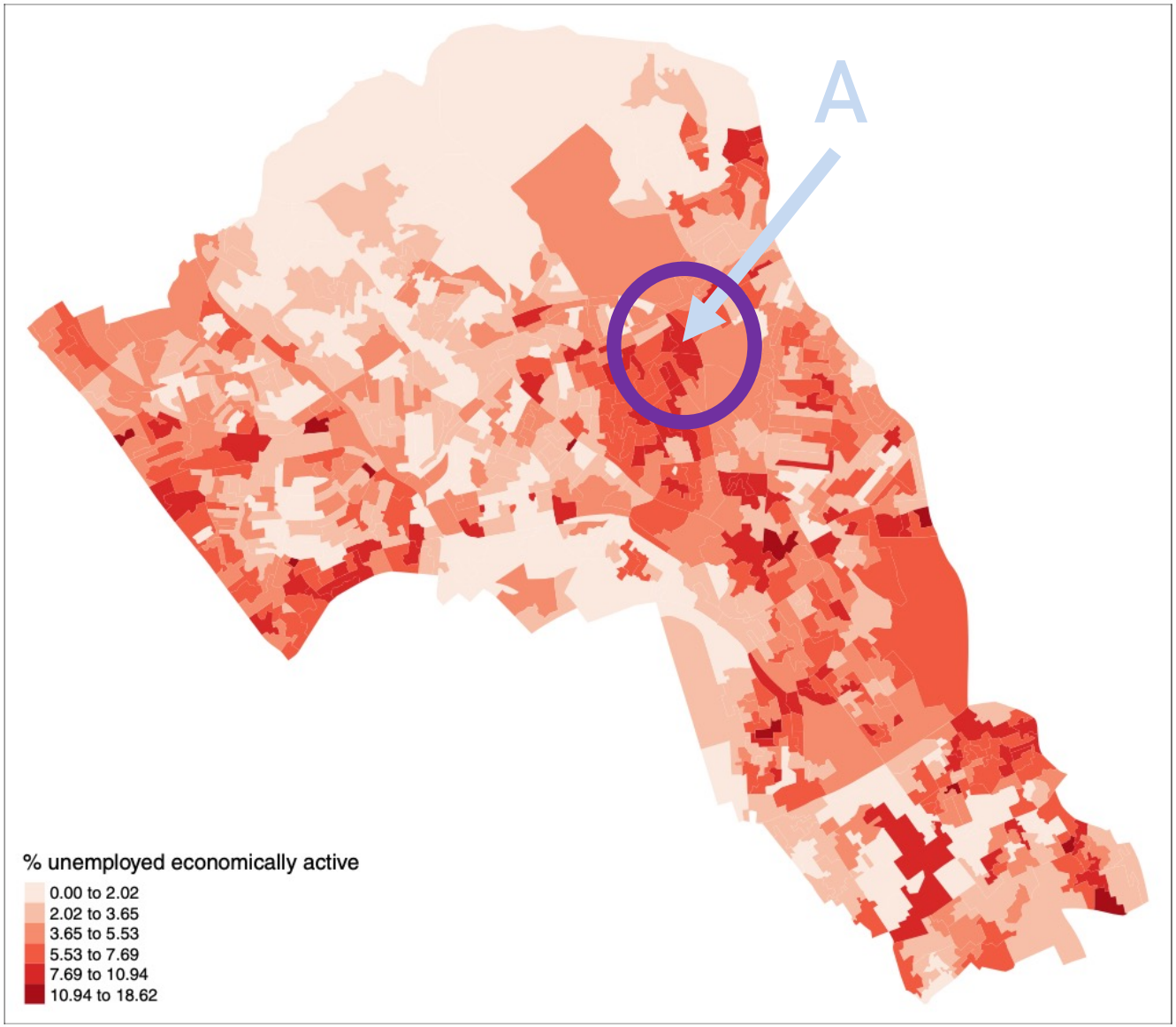
Rethinking causality in city science & spatial analysis



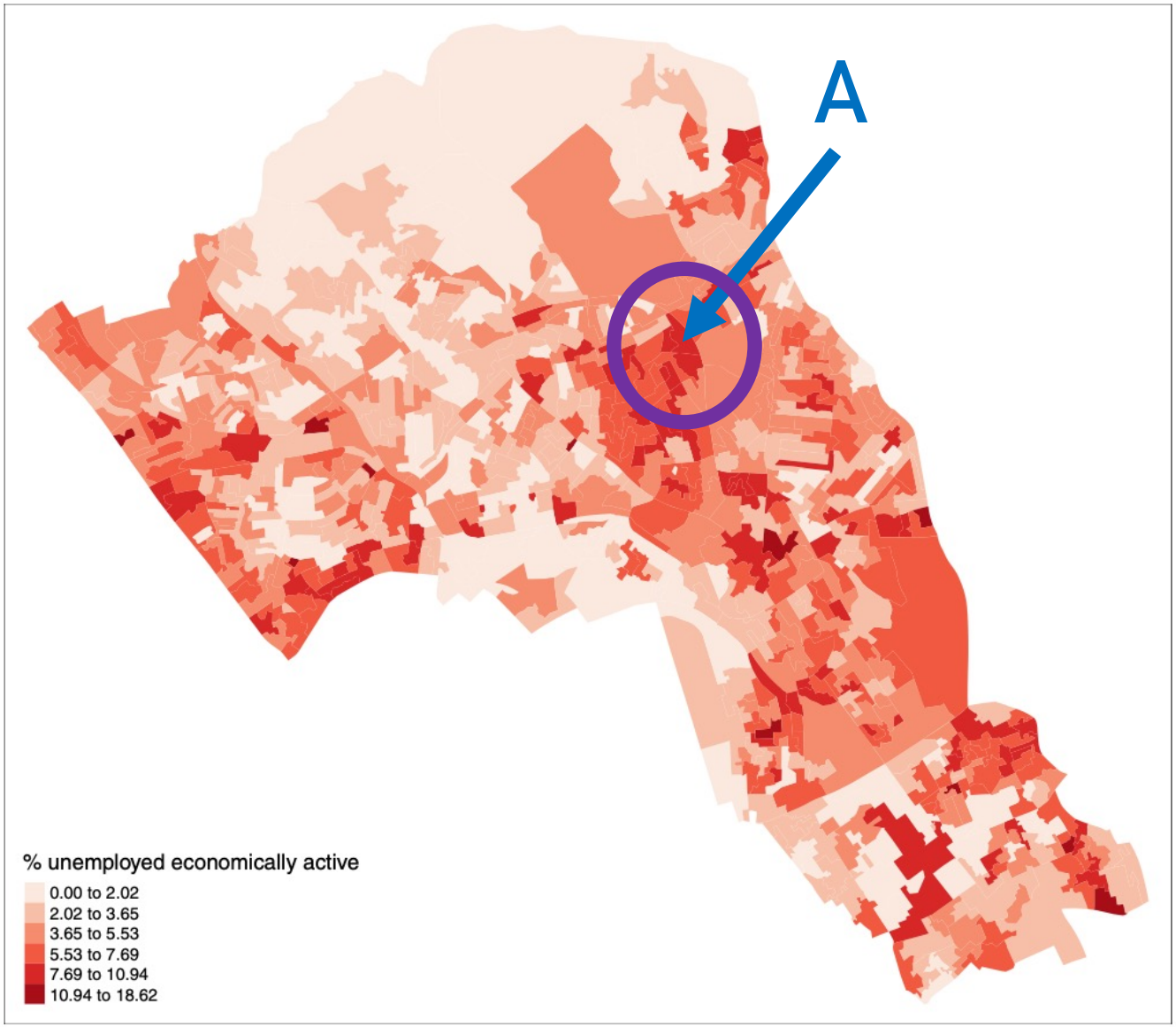
Local statistics: find outliers and/or clusters



Local statistics: find outliers and/or clusters

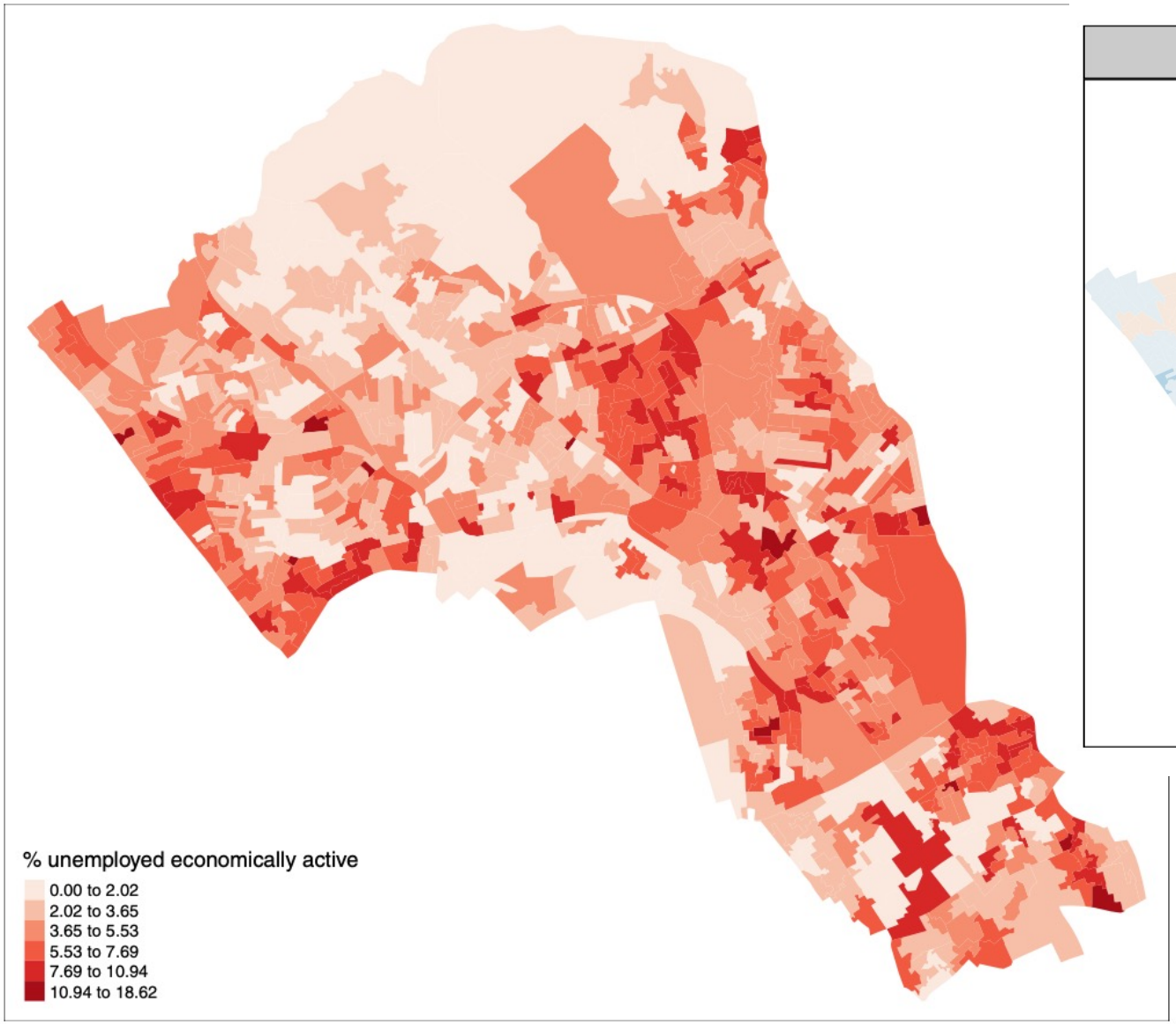


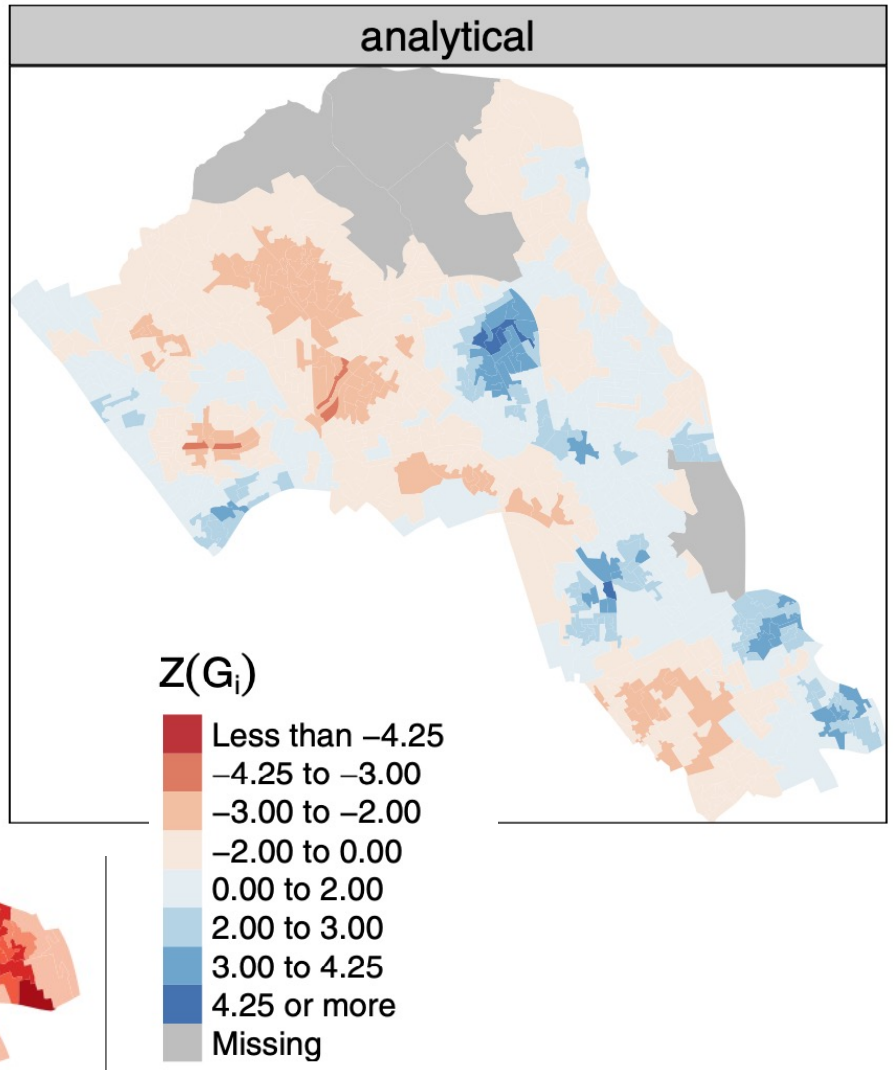
Local statistics: find outliers and/or clusters



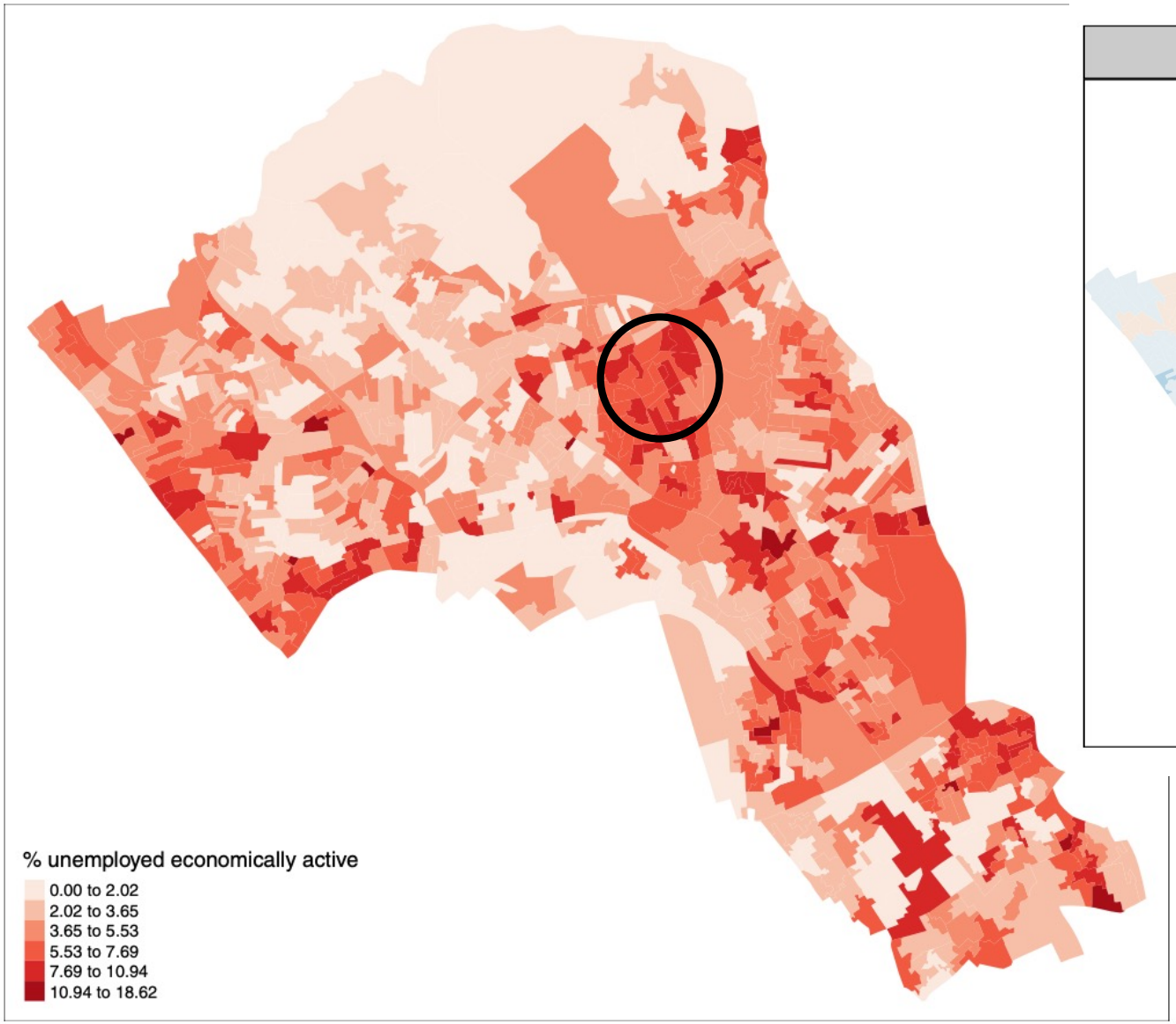
Is this site unusual for the values in its surrounding area?

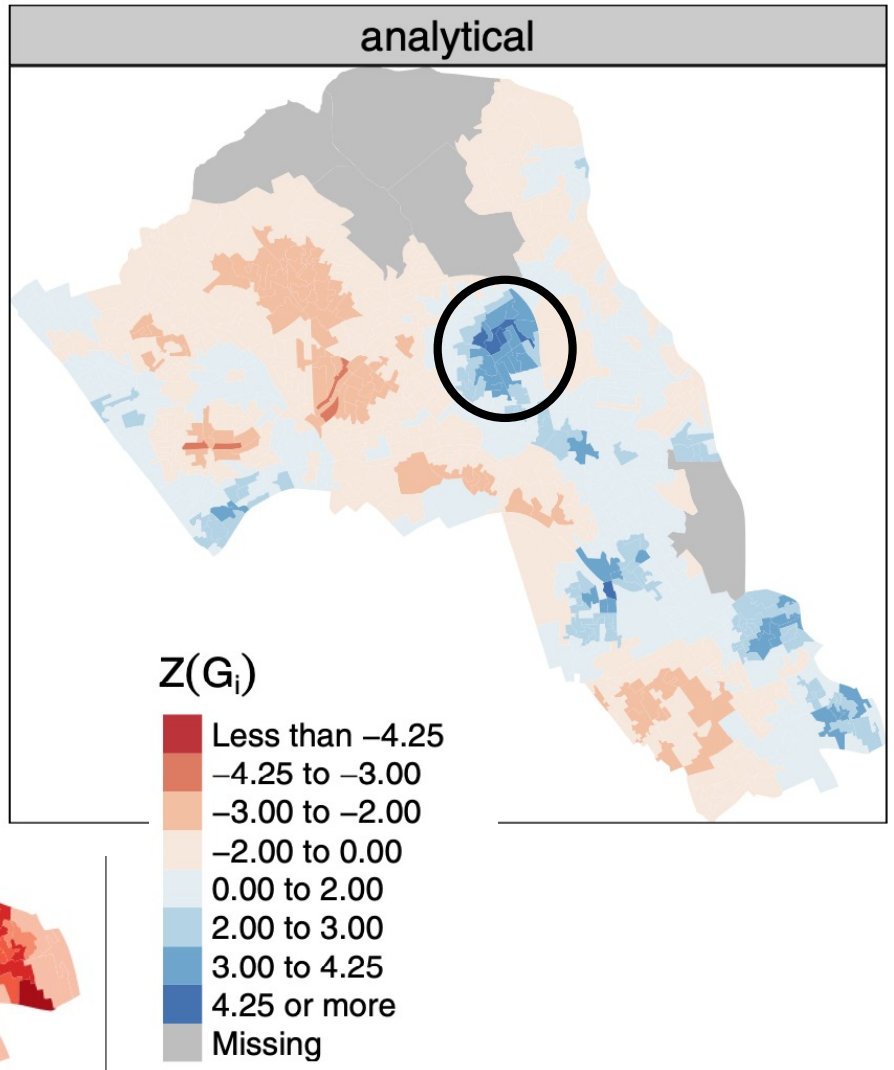
Local statistics: find outliers and/or clusters





Local statistics: find outliers and/or clusters





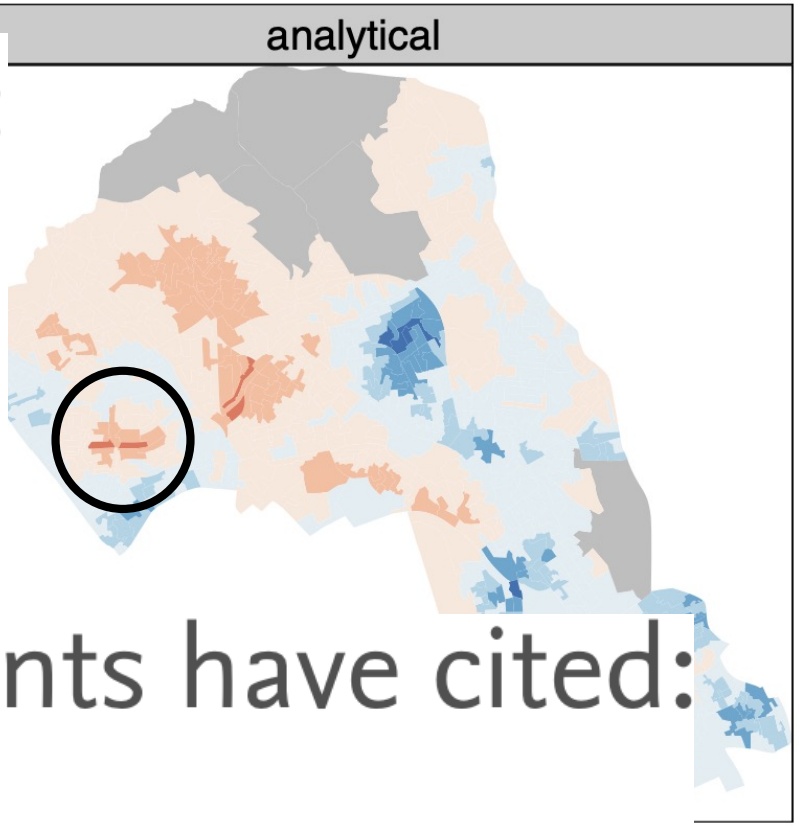
Local statistics: find outliers and/or clusters

5,679 documents have cited:

Local Indicators of Spatial Association—LISA

Anselin L.

(1995) Geographical Analysis, 27 (2) , pp. 93-115.

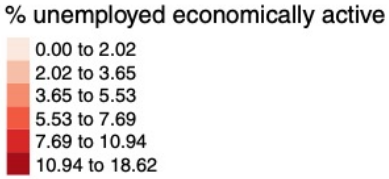


2,893 documents have cited:

The Analysis of Spatial Association by Use of Distance Statistics

Getis A., Ord J.K.

(1992) Geographical Analysis, 24 (3) , pp. 189-206.



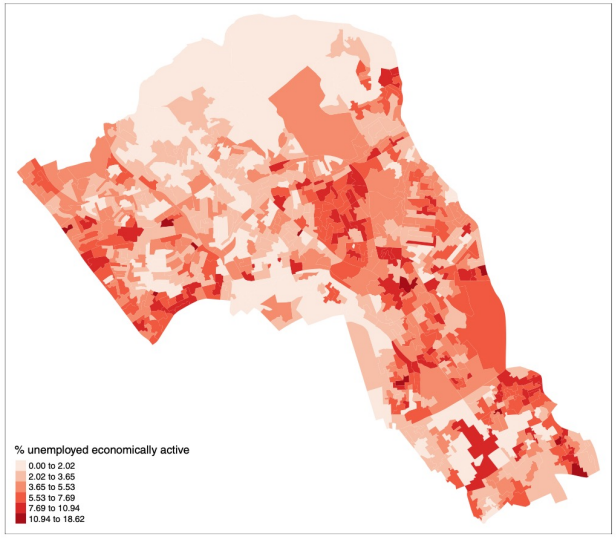
Local statistics: find outliers and/or clusters

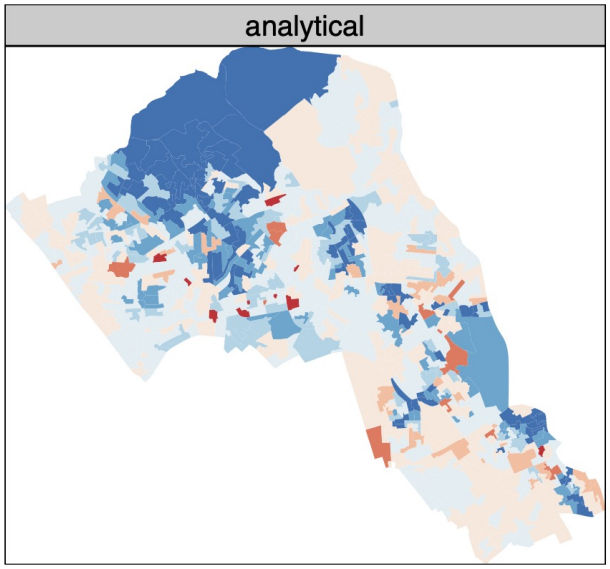
ORIGINAL PAPER

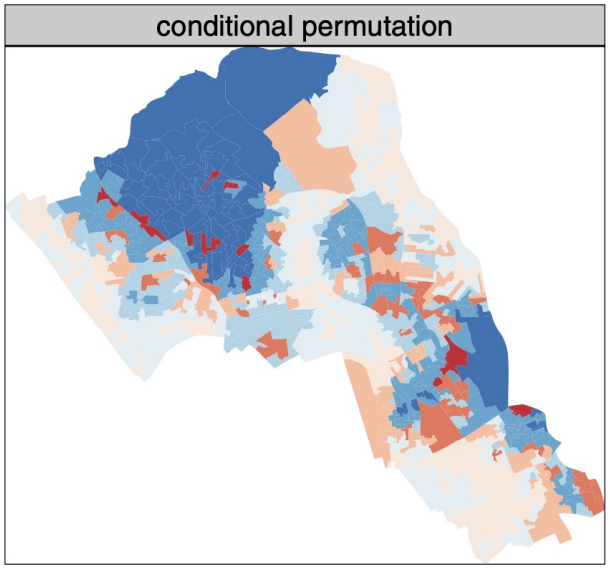
Comparing implementations of global and local indicators of spatial association

Roger S. Bivand[1] · David W. S. Wong[2]

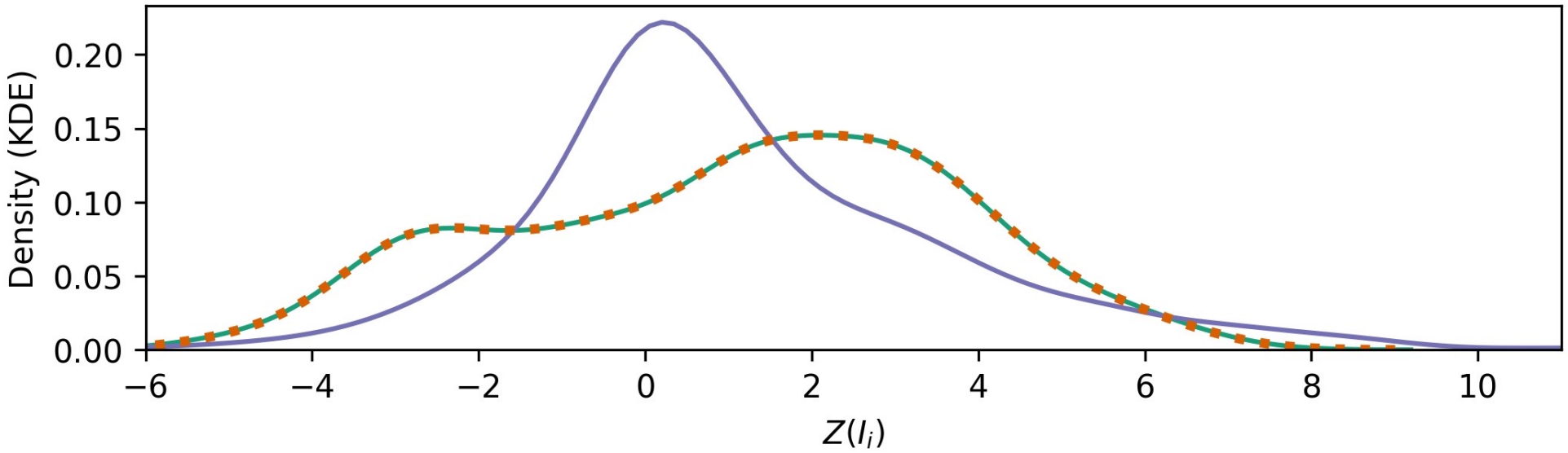
User choices for local measures both of software and of inferential method over and above the handling of multiple comparisons will have consequences for conclusions





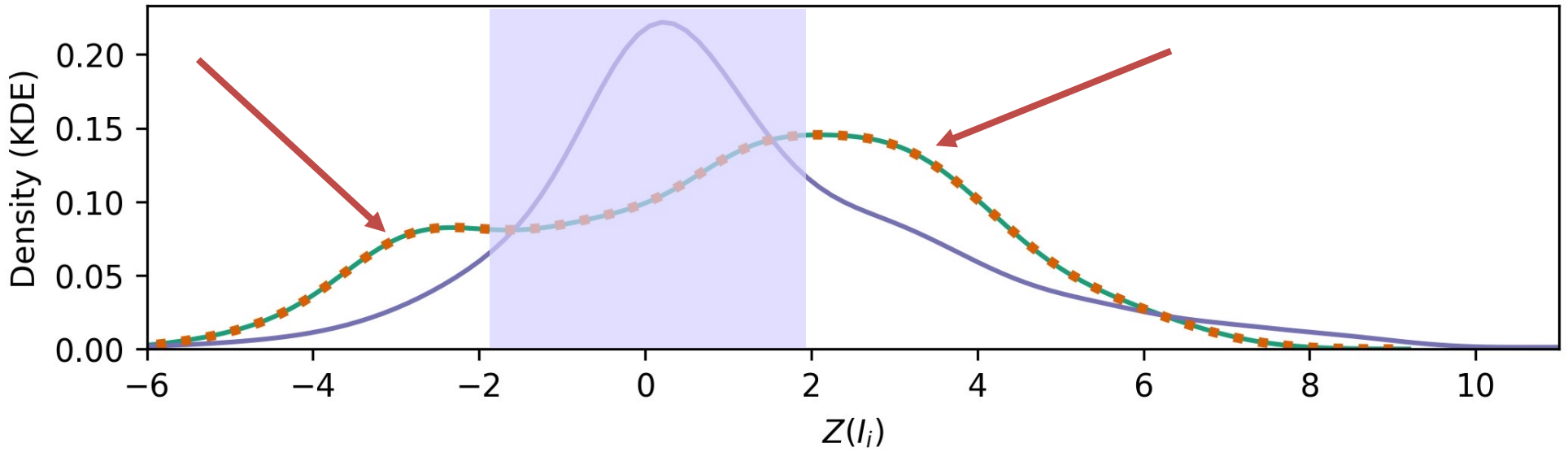


[T]he spatial patterns of Z values generated by conditional permutation for local measures differ considerably from those calculated using analytical methods.



The permutation-based local z-scores were much more likely to be statistically significant […] Analytical inference yielded 267 significant OAs whereas permutation-based inference yielded 429. […] OAs were statistically significant using analytical inference but not using permutation-based inference, and vice versa.

Sauer et al. (2021), p3



The permutation-based local z-scores were much more likely to be statistically significant […] Analytical inference yielded 267 significant OAs whereas permutation-based inference yielded 429. […] OAs were statistically significant using analytical inference but not using permutation-based inference, and vice versa.

Sauer et al. (2021), p3

Texts in Statistical Science

Geographic Data Science with Python

Sergio Rey
Dani Arribas-Bel
Levi John Wolf

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

The R Series

Spatial Data Science

With Applications in R

Edzer Pebesma
Roger Bivand

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

CrossMark

rs

differ

The Importance of Null Hypotheses: Understanding Differences in Local Moran's I_i under Heteroskedasticity

Jeffery Sauer[1], Taylor Oshan[1], Sergio Rey[2], Levi John Wolf[3]

We see two different ways the divergence … could come about. Bivand & Wong (2018)'s theory is that local heteroskedasticity in the data is causing their … estimates to diverge. Our theory is that different null hypotheses drive the observed differences.

ANALTYIC

SYNTHETIC

ANALTYIC

Assume normal observations distributed over a regular grid

Derive the expectation $\mathbf{E}[\Gamma]$

Use to define $\mathbf{Var}[\Gamma]$

Construct z-score

Assess against a standard normal distribution: $|z_i| > 1.96$

Hope it generalizes!

SYNTHETIC

ANALTYIC

$$\mathbf{E}[I_i] = -\sum_{j \neq i} \frac{w_{ij}}{n-1}$$

$$\mathbf{Var}[I_i] = \frac{w_{i(2)}(n-b_2)}{n-1} + \frac{2w_{i(kh)}2b_2-n}{(n-1)(n-2)} + \left[\sum_{i \neq j} \frac{w_{ij}}{n-1}\right]^2$$

SYNTHETIC

ANALTYIC

$$\mathbf{E}[I_i] = -\sum_{j \neq i} \frac{w_{ij}}{n-1}$$

$$\mathbf{Var}[I_i] = \frac{w_{i(2)}(n-b_2)}{n-1} + \frac{2w_{i(kh)}2b_2-n}{(n-1)(n-2)} + \left[\sum_{i \neq j} \frac{w_{ij}}{n-1}\right]^2$$

SYNTHETIC

Assume the data that you *do have* is like data you *could have*

Hold the value at the site you're testing constant

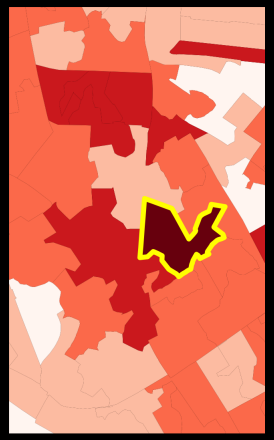
Shuffle the *rest* of the map

Compute a new local statistic

Repeat many times to obtain distribution of replicates,

Describe your site *relative to* the replicates

REAL



SYNTHETIC

Assume the data that you *do have* is like data you *could have*

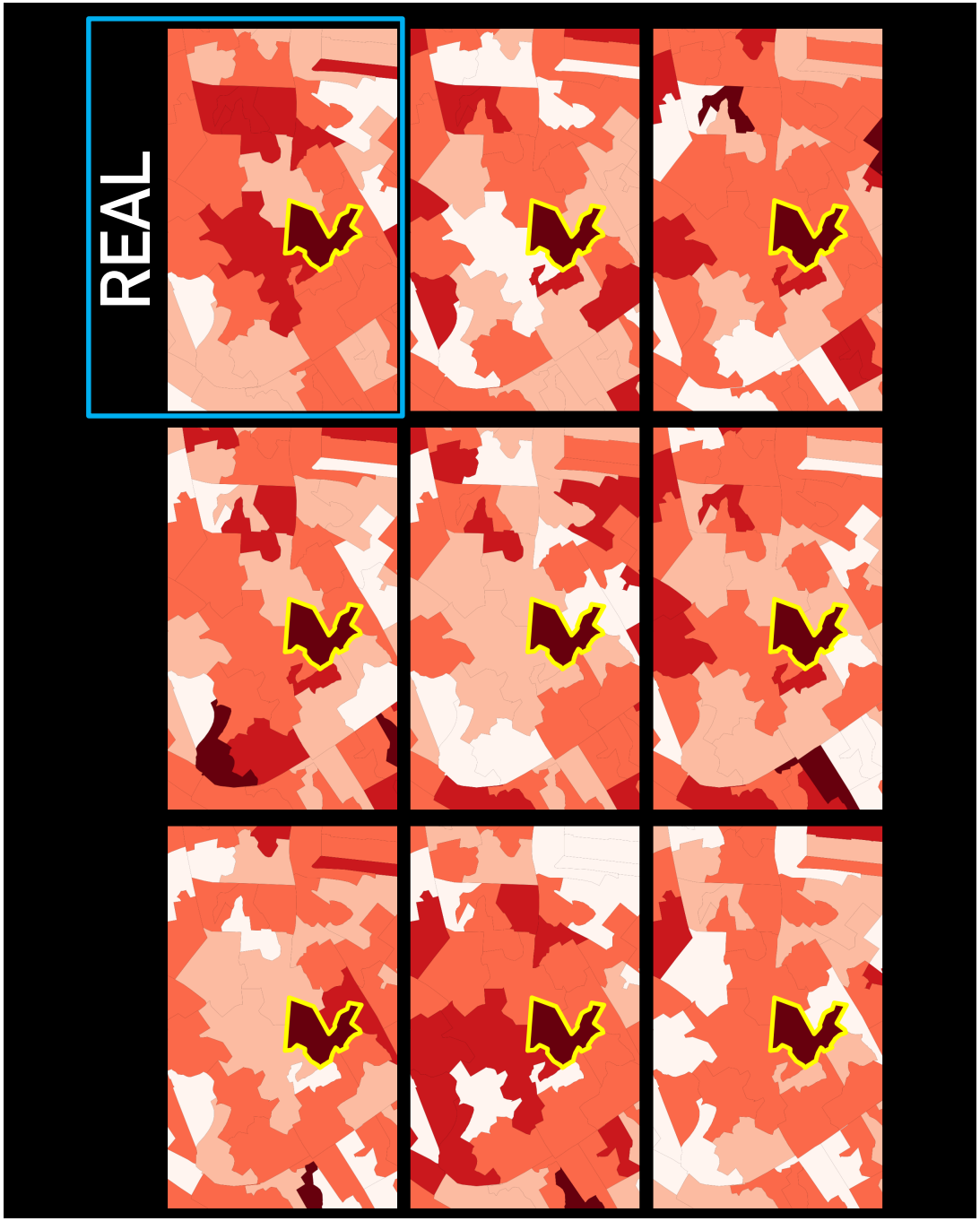
Hold the value at the site you're testing constant

Shuffle the *rest* of the map

Compute a new local statistic

Repeat many times to obtain distribution of replicates,

Describe your site *relative to* the replicates



SYNTHETIC

Assume the data that you *do have* is like data you *could have*

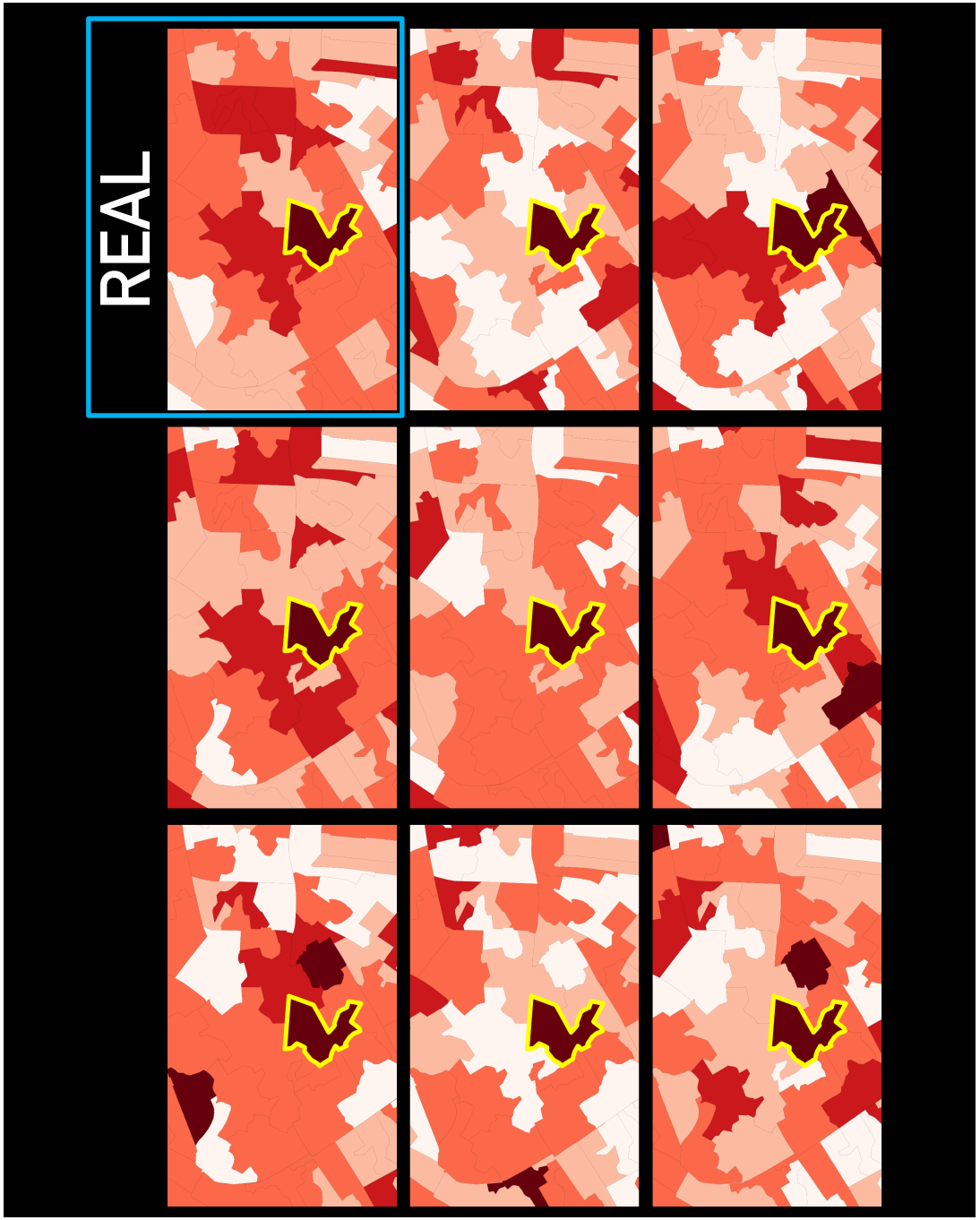
Hold the value at the site you're testing constant

Shuffle the *rest* of the map

Compute a new local statistic

Repeat many times to obtain distribution of replicates,

Describe your site *relative to* the replicates



SYNTHETIC

Assume the data that you *do have* is like data you *could have*

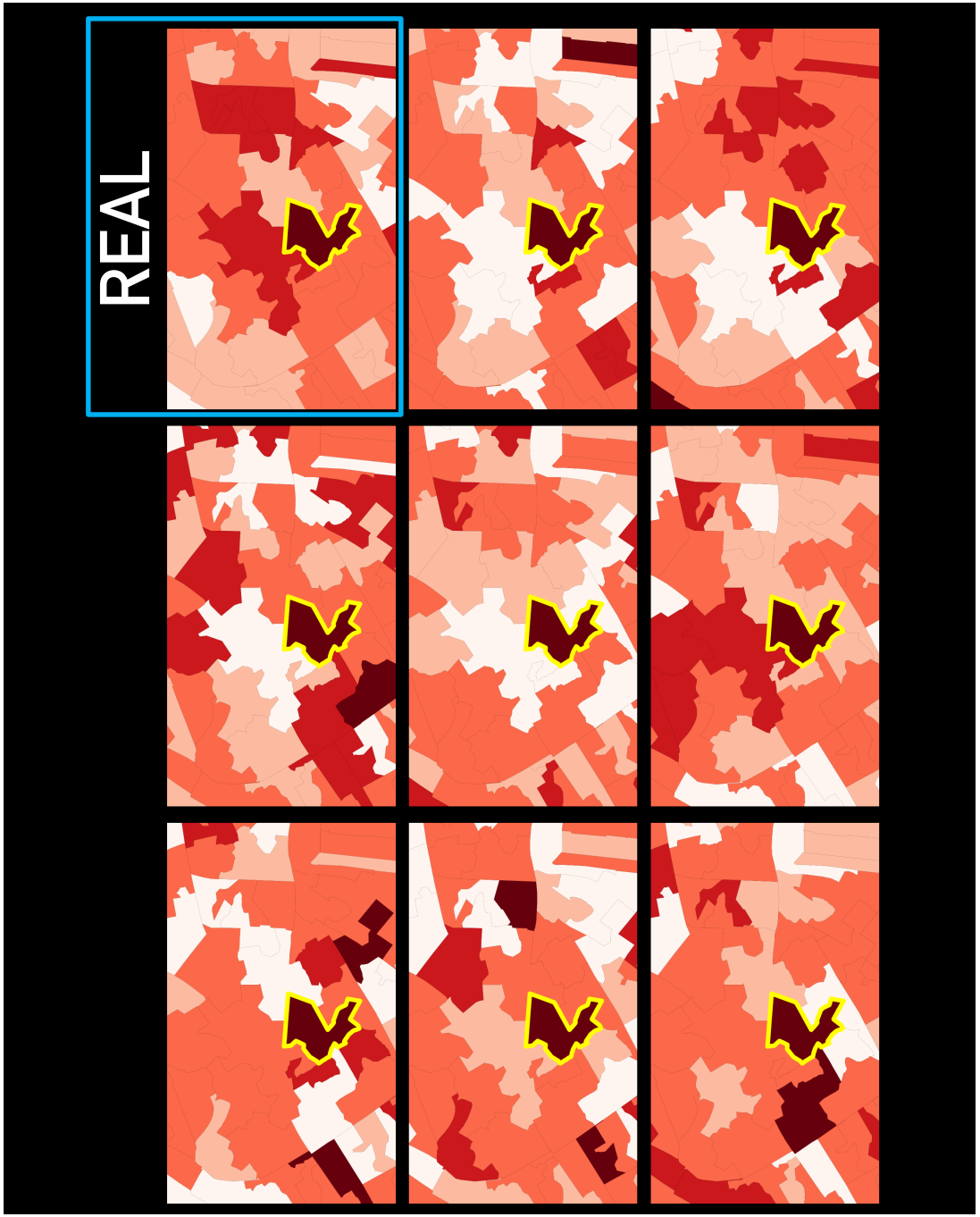
Hold the value at the site you're testing constant

Shuffle the *rest* of the map

Compute a new local statistic

Repeat many times to obtain distribution of replicates,

Describe your site *relative to* the replicates



SYNTHETIC

Assume the data that you *do have* is like data you *could have*

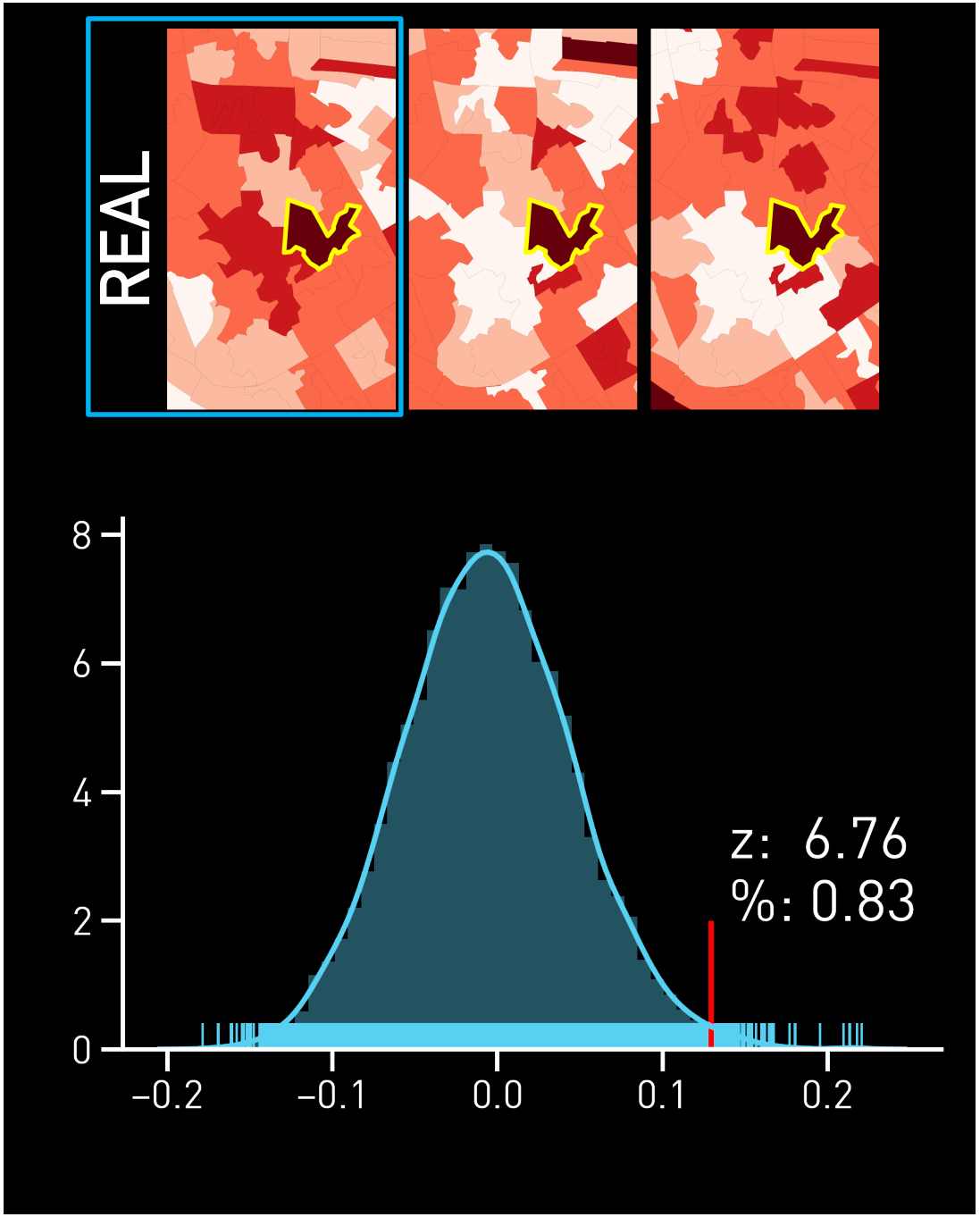
Hold the value at the site you're testing constant

Shuffle the *rest* of the map

Compute a new local statistic

Repeat many times to obtain distribution of replicates,

Describe your site *relative to* the replicates



SYNTHETIC

Assume the data that you *do have* is like data you *could have*

Hold the value at the site you're testing constant

Shuffle the *rest* of the map

Compute a new local statistic

Repeat many times to obtain distribution of replicates,

Describe your site *relative to* the replicates

ANALTYIC

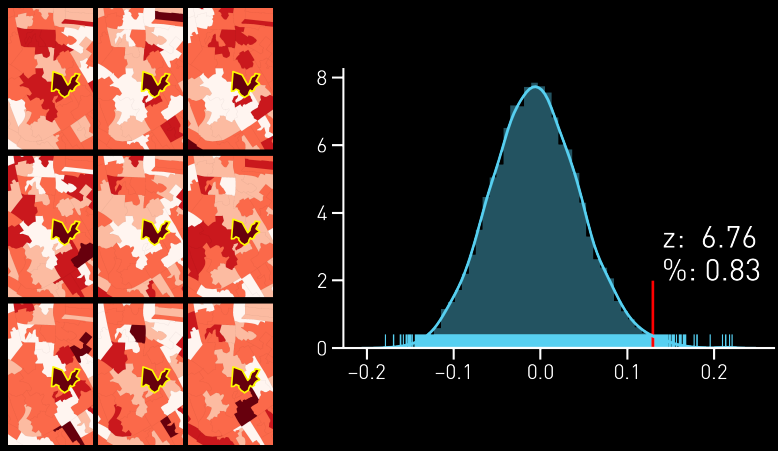
$$\mathbf{E}[I_i] = -\sum \frac{w_{ij}}{n-1}$$

$$\mathbf{Var}[I_i] = \frac{w_{i(2)}(n-b_2)}{n-1} + \frac{2w_{i(kh)}2b_2-n}{(n-1)(n-2)} + \left[\sum_{i \neq j} \frac{w_{ij}}{n-1}\right]^2$$

SYNTHETIC

?

	"Total" Randomization Null *shuffle all values, so each value is equally likely at any site*	**"Conditional" Randomization Null** *for each site, the site's value is fixed and remaining sites are shuffled*
Analytical estimator *closed-form mathematical expressions for test statistics*	Originally in Anselin (1995). Implemented in **pysal** and **spdep**. Used in Bivand & Wong (2018). Denoted here as $E[I_i]$ and $Var[I_i]$	Originally in Sokal (1998). Not implemented in **pysal** and **spdep** or considered by Bivand & Wong (2018). Denoted here as $E_c[I_i]$ and $Var_c[I_i]$
Permutation estimator *test statistics computed from set of simulated maps*	Not considered or implemented.	Originally in Anselin (1995). Implemented in **pysal**. Used in Bivand and Wong (2018). Denoted here as $E_p[I_i]$ and $Var_p[I_i]$



	"Total" Randomization Null *shuffle all values, so each value is equally likely at any site*	**"Conditional" Randomization Null** *for each site, the site's value is fixed and remaining sites are shuffled*
Analytical estimator *closed-form mathematical expressions for test statistics*	Originally in Anselin (1995). Implemented in **pysal** and **spdep**. Used in Bivand & Wong (2018). Denoted here as $E[I_i]$ and $Var[I_i]$	Originally in Sokal (1998). Not implemented in **pysal** and **spdep** or considered by Bivand & Wong (2018). Denoted here as $E_c[I_i]$ and $Var_c[I_i]$
Permutation estimator *test statistics computed from set of simulated maps*	Not considered or implemented.	Originally in Anselin (1995). Implemented in **pysal**. Used in Bivand and Wong (2018). Denoted here as $E_p[I_i]$ and $Var_p[I_i]$

$$\mathbf{E}[I_i] = -\sum_{j \neq i} \frac{w_{ij}}{n-1}$$

$$\mathbf{Var}[I_i] = \frac{w_{i(2)}(n-b_2)}{n-1} + \frac{2w_{i(kh)}2b_2-n}{(n-1)(n-2)} + \left[\sum_{i \neq j} \frac{w_{ij}}{n-1}\right]^2$$

	"Total" Randomization Null *shuffle all values, so each value is equally likely at any site*	**"Conditional" Randomization Null** *for each site, the site's value is fixed and remaining sites are shuffled*
Analytical estimator *closed-form mathematical expressions for test statistics*	Originally in Anselin (1995). Implemented in **pysal** and **spdep**. Used in Bivand & Wong (2018). Denoted here as $E[I_i]$ and $Var[I_i]$	Originally in Sokal (1998). Not implemented in **pysal** and **spdep** or considered by Bivand & Wong (2018). Denoted here as $E_c[I_i]$ and $Var_c[I_i]$
Permutation estimator *test statistics computed from set of simulated maps*	Not considered or implemented.	Originally in Anselin (1995). Implemented in **pysal**. Used in Bivand and Wong (2018). Denoted here as $E_p[I_i]$ and $Var_p[I_i]$

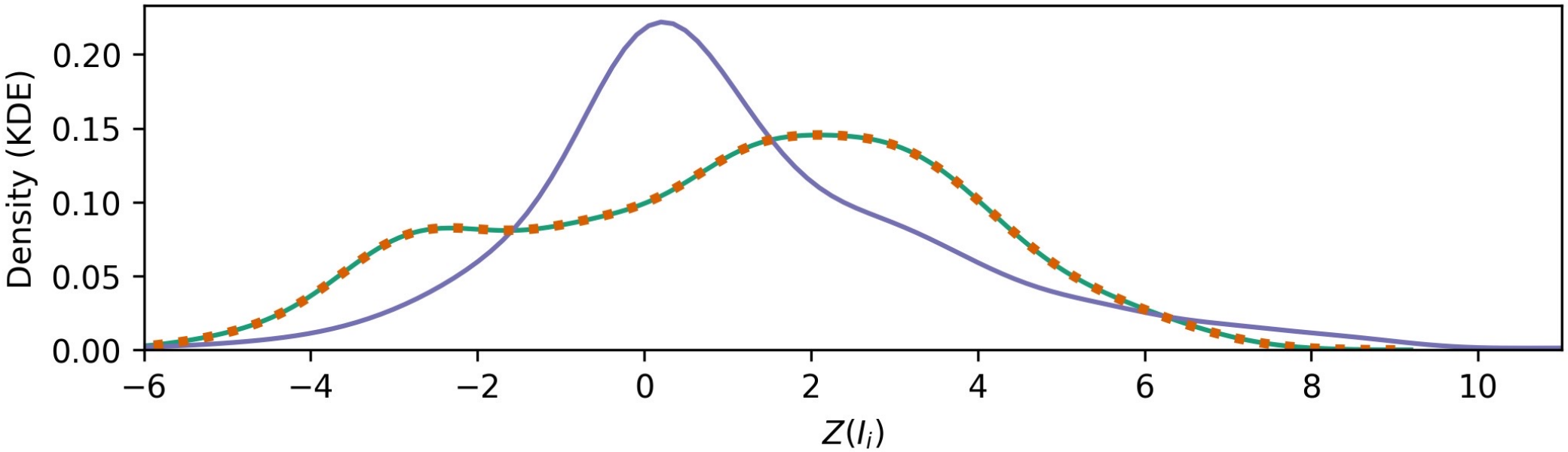
$$\mathbf{E}_c[I_i] = -z_i^2 \sum_{i \neq j} \frac{w_{ij}}{n-1}$$

$$\mathbf{Var}_c[I_i] = \left[\frac{z_i}{m_2}\right]^2 \left[\frac{n}{n-2}\right] \left[w_{i(2)} - \frac{\left(\sum_{i \neq j} w_{ij}\right)^2}{n-1}\right] \left[m_2 - \frac{z_i^2}{n-1}\right]$$

	"Total" Randomization Null *shuffle all values, so each value is equally likely at any site*	**"Conditional" Randomization Null** *for each site, the site's value is fixed and remaining sites are shuffled*
Analytical estimator *closed-form mathematical expressions for test statistics*	Originally in Anselin (1995). Implemented in **pysal** and **spdep**. Used in Bivand & Wong (2018). Denoted here as $E[I_i]$ and $Var[I_i]$	Originally in Sokal (1998). Not implemented in **pysal** and **spdep** or considered by Bivand & Wong (2018). Denoted here as $E_c[I_i]$ and $Var_c[I_i]$
Permutation estimator *test statistics computed from set of simulated maps*	Not considered or implemented.	Originally in Anselin (1995). Implemented in **pysal**. Used in Bivand and Wong (2018). Denoted here as $E_p[I_i]$ and $Var_p[I_i]$

$$\mathbf{E}[I_i] = -\sum_{j \neq i} \frac{w_{ij}}{n-1} \quad \mathbf{E}_c[I_i] = -z_i^2 \sum_{i \neq j} \frac{w_{ij}}{n-1}$$

Null likelihood for the stat depends on the observed valueat the site!



	$Z(I_i)$	$Z_p(I_i)$	$Z_c(I_i)$
$Z(I_i)$	1.00	0.85	0.85
$Z_p(I_i)$	0.85	1.00	1.00
$Z_c(I_i)$	0.85	1.00	1.00

Empirical "conditional randomization" $Z_p(I_i)$

Analytical "total randomization" $Z(I_i)$

Analytical "conditional randomization" $Z_c(I_i)$

Same null, same findings over different estimators

The Importance of Null Hypotheses: Understanding Differences in Local Moran's I_i under Heteroskedasticity

Jeffery Sauer[1], Taylor Oshan[1], Sergio Rey[2], Levi John Wolf[3]

A vast body of work has expanded upon Anselin (1995). But, we [still] tend to distinguish the "total" and "conditional" null by computation, not conceptualization. Our null hypotheses should be more explicitly acknowledged, however, because they are themselves design decisions.

We do not have the right balance

a. Expecting something to replicate &
b. Post hoc justifying with spatial or scale "complexity" when it doesn't

Replication across space and time must be weak in the social and environmental sciences
Michael F Goodchild & Wenwen Li
Proceedings of the National Academy of Sciences, 2021
doi.org/10.1073/pnas.2015759118

GEOGRAPHIC INFORMATION SCIENCE II: Mesogeography: social physics, GIScience, and the quest for geographic knowledge
Harvey J. Miller
Progress in Human Geography, 2018
doi.org/10.1177/0309132517712154

"Democratizing" methods
makes our defaults "political!"

- Some questions were never answered, but stats is a science of defaults
- Empirical research now scales well, but theoretical understanding is contextual!

TESTING FOR LOCAL SPATIAL AUTOCORRELATION IN THE PRESENCE OF GLOBAL AUTOCORRELATION

J. Keith Ord & Arthur Getis
doi.org/10.1111/0022-4146.00224

accepted,
preprint

Confounded Local Inference: Extending Local Moran Stats to Handle Confounding

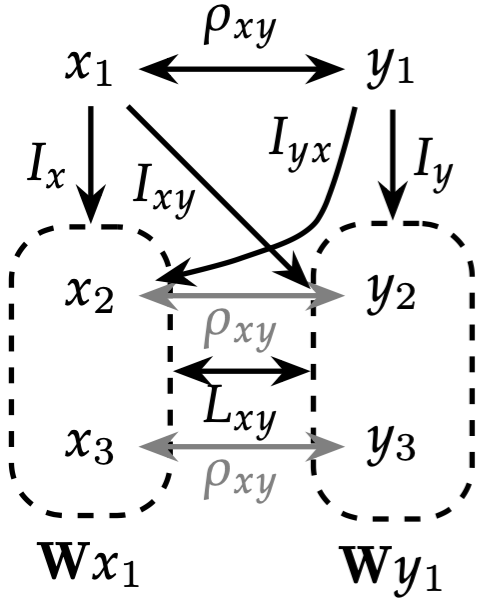
Levi John Wolf

Connecting theory to (local) statistics

accepted,
preprint

Confounded Local Inference: Extending Local Moran Stats to Handle Confounding

Levi John Wolf



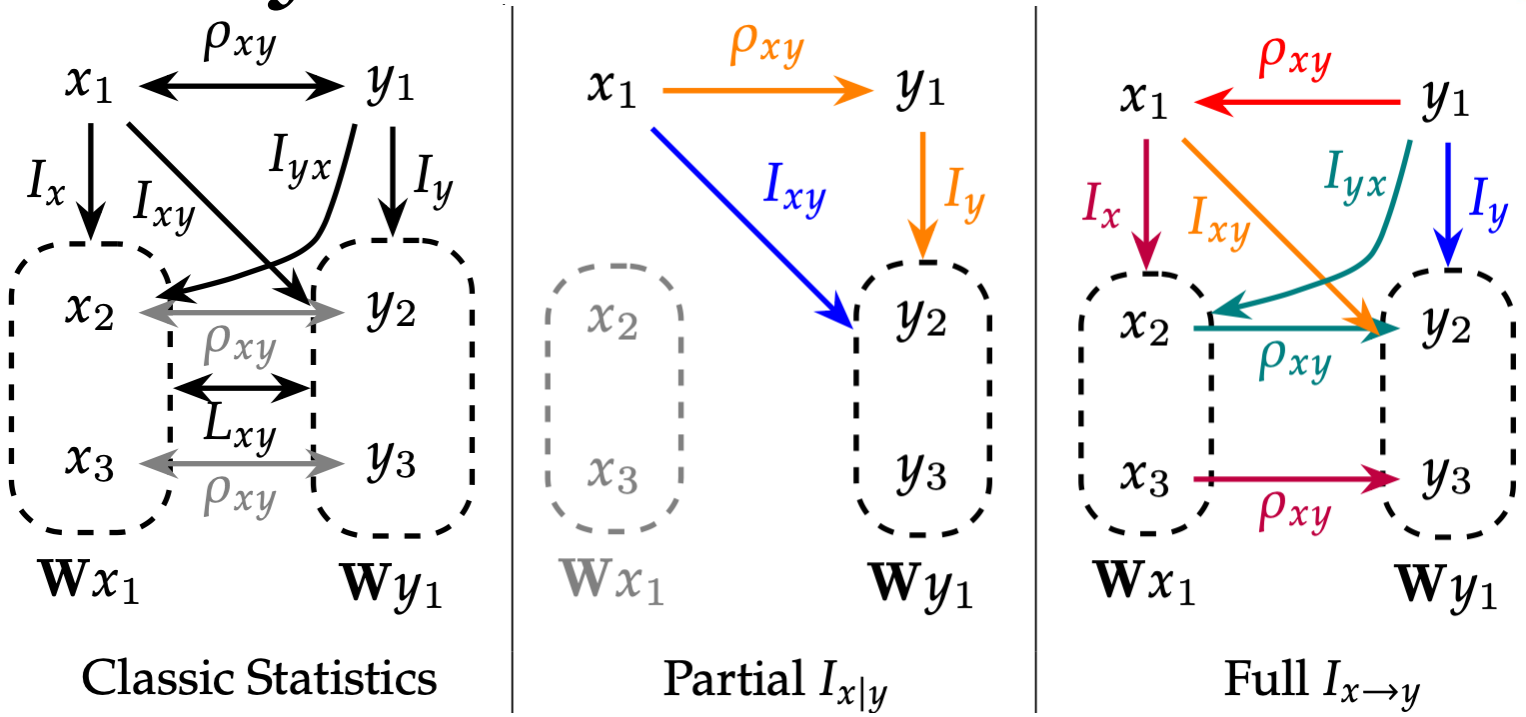
Classic Statistics

Connecting theory to (local) statistics

accepted,
preprint

Confounded Local Inference: Extending Local Moran Stats to Handle Confounding

Levi John Wolf



Connecting theory to (local) statistics

accepted,
preprint

Confounded Local Inference: Extending Local Moran Stats to Handle Confounding

Levi John Wolf

Name		Statistic
Unviariate	$\mathbf{I}_y$	$y \circ \mathbf{W}y$
Bivariate	$\mathbf{I}_{xy}$	$x \circ \mathbf{W}y$
Partial	$\mathbf{I}_{y\|x}$	$\mathbf{I}_y - \rho_{xy}\mathbf{I}_{xy}$
Auxiliary	$\mathbf{I}_{x\to y}$	$\mathbf{I}_{y\|x} - \rho_{xy}\mathbf{I}_{yx} - \rho_{xy}^2\mathbf{I}_x$

Connecting theory to (local) statistics

accepted, preprint

Confounded Local Inference: Extending Local Moran Stats to Handle Confounding

Levi John Wolf

Name		Statistic	
Unviariate	$\mathbf{I}_y$	$y \circ \mathbf{W}y$	*Are house prices high here?*
Bivariate	$\mathbf{I}_{xy}$	$x \circ \mathbf{W}y$	*Are big houses expensive here?*
Partial	$\mathbf{I}_{y\|x}$	$\mathbf{I}_y - \rho_{xy}\mathbf{I}_{xy}$	*Is this house's price unusual for this area, given its size?*
Auxiliary	$\mathbf{I}_{x\rightarrow y}$	$\mathbf{I}_{y\|x} - \rho_{xy}\mathbf{I}_{yx} - \rho^2_{xy}\mathbf{I}_x$	*Is the "extra" you pay for this house given its size unusual for this area?*

Connecting theory to (local) statistics

THE CITY AND CAUSALITY

replication is needed
for city science to work

MUCH ADO ABOUT NULL THINGS

and replication is
harder than it looks!

Rethinking causality in city science & spatial analysis

ADVANCES IN SPATIAL DATA SCIENCE

Causality and reproducibility

LEVI JOHN WOLF

levi.john.wolf@bristol.ac.uk