MUCH ADO ABOUT NULL THINGS

A replication issue in spatial stats

JEFF SAUER
SERGIO REY
TAYLOR OSHAN
LEVI JOHN WOLF

LOCAL STATISTICS

A REPLICATION FAILURE

"NULLS" IN LOCAL TESTS

The importance of null hypotheses in local statistics

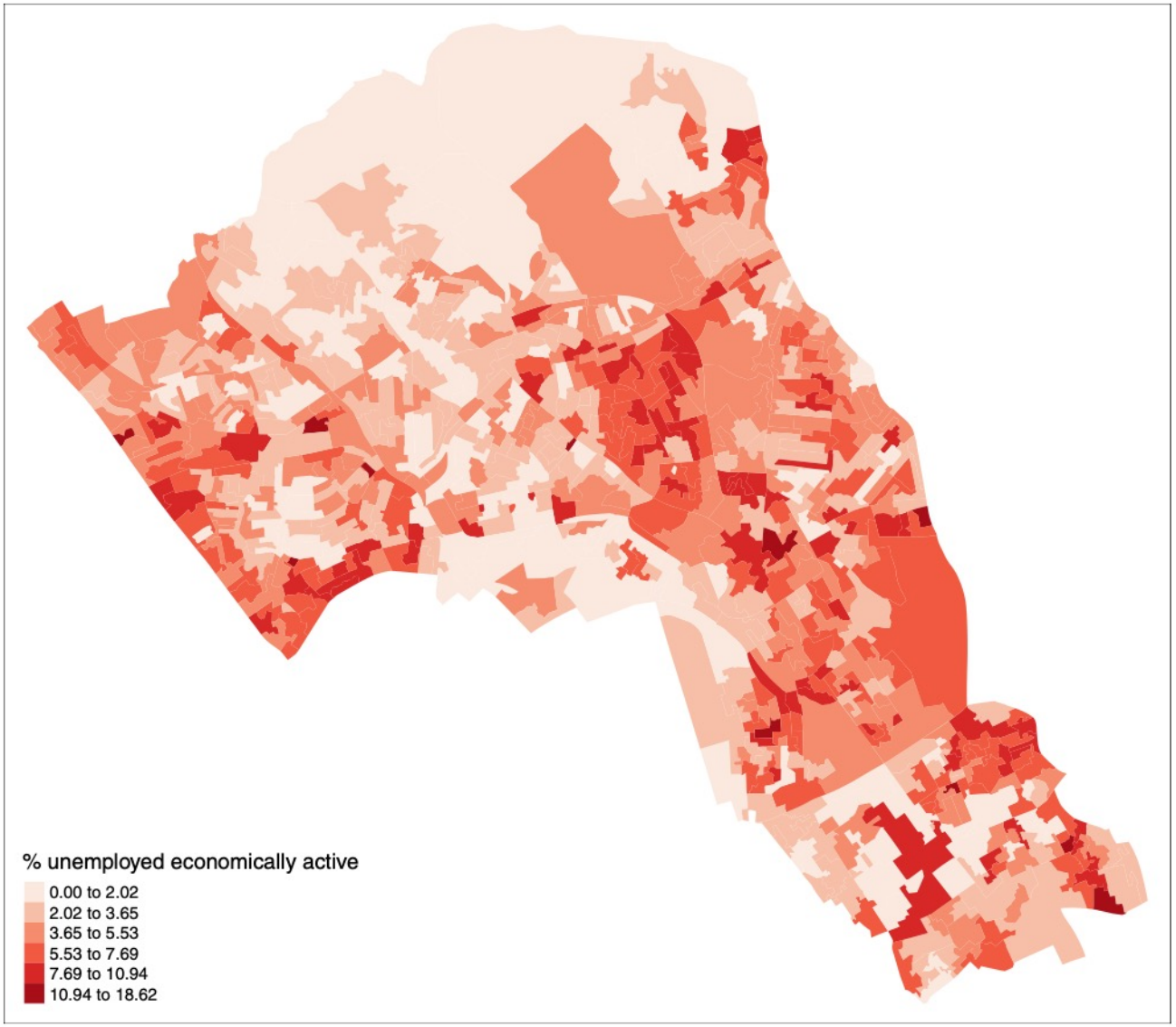
LOCAL STATISTICS

FORMALIZING INTUITION

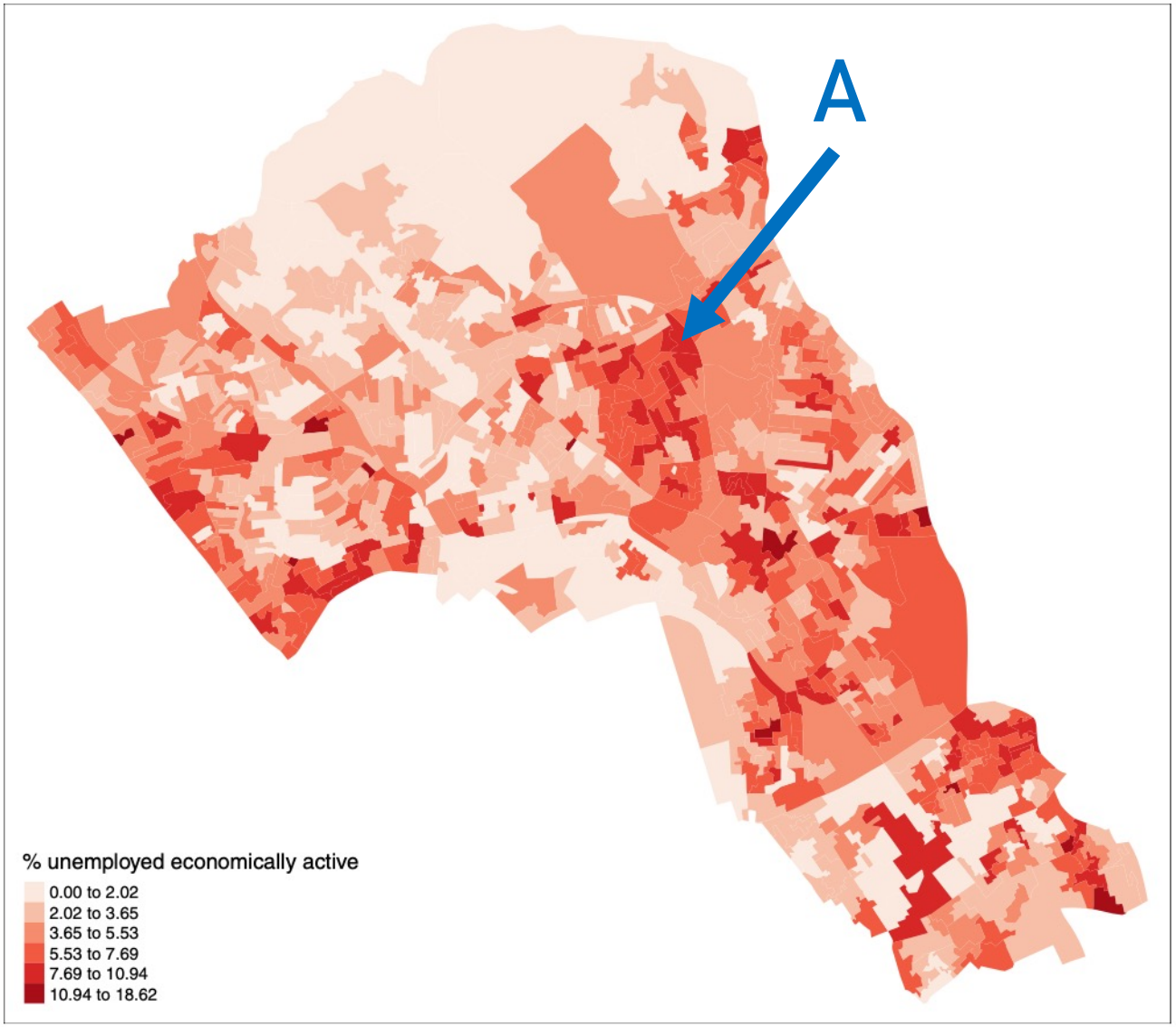
A REPLICATION FAILURE

"NULLS" IN LOCAL TESTS

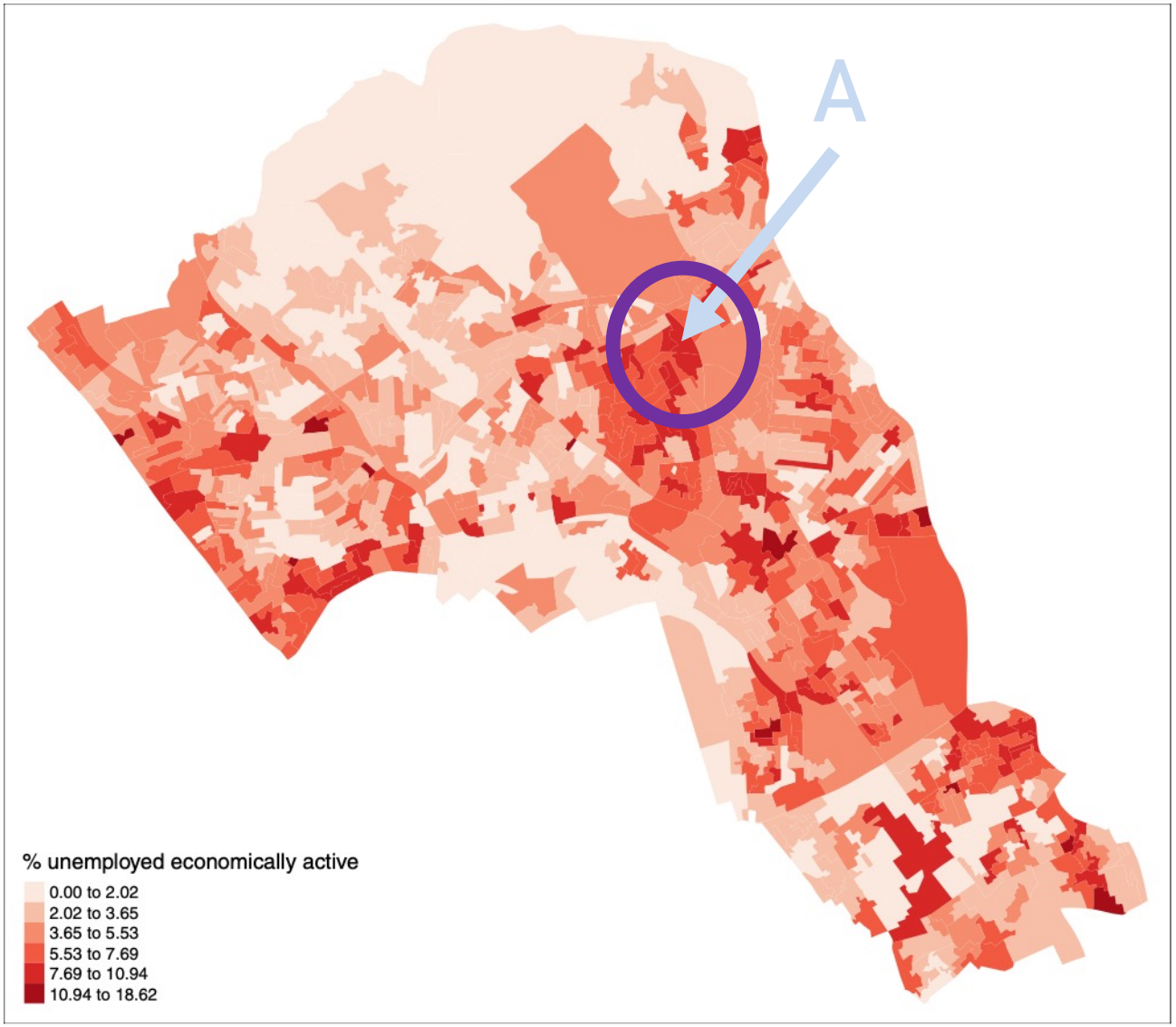
The importance of null hypotheses in local statistics



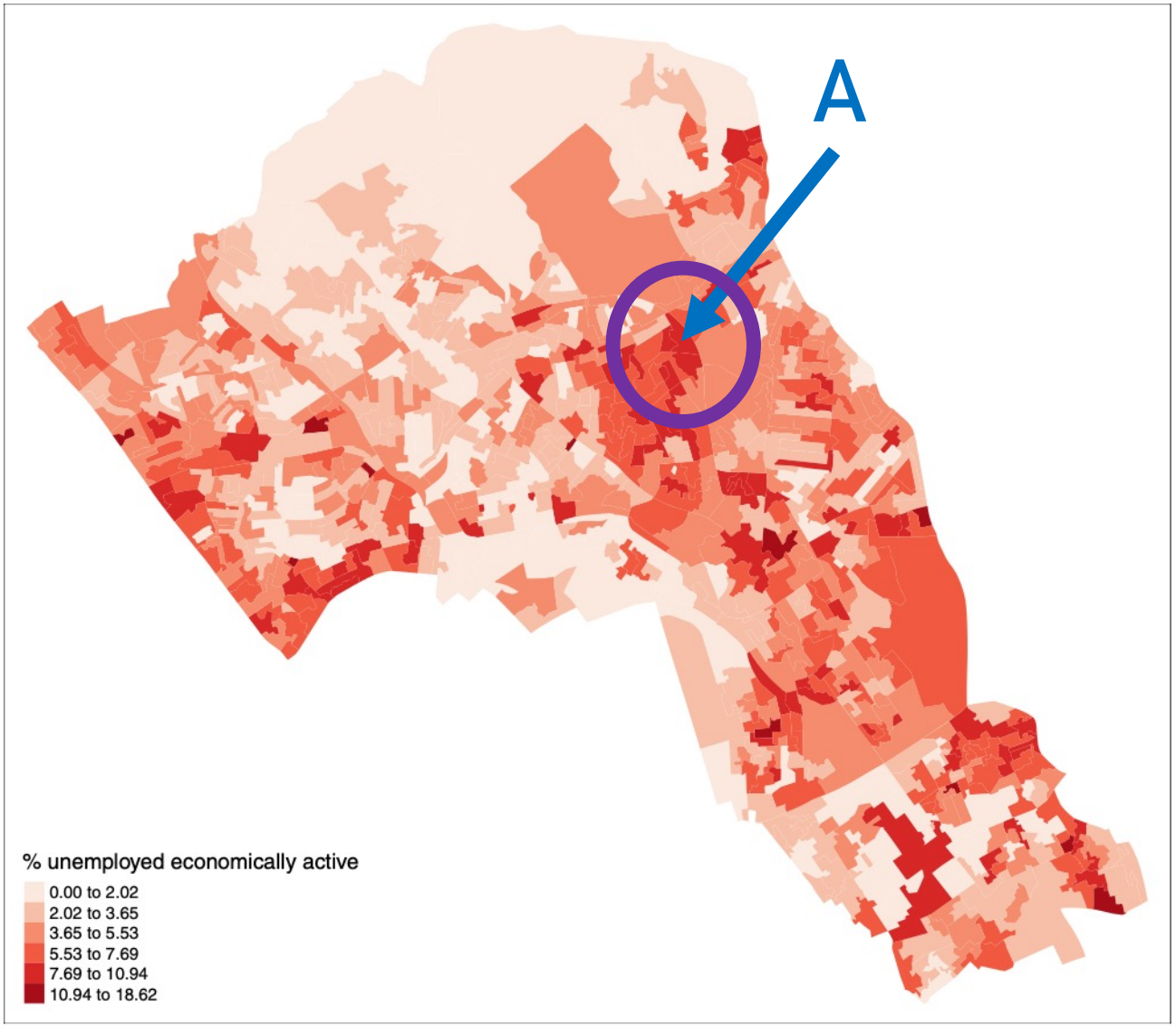
Local statistics: find outliers and/or clusters



Local statistics: find outliers and/or clusters

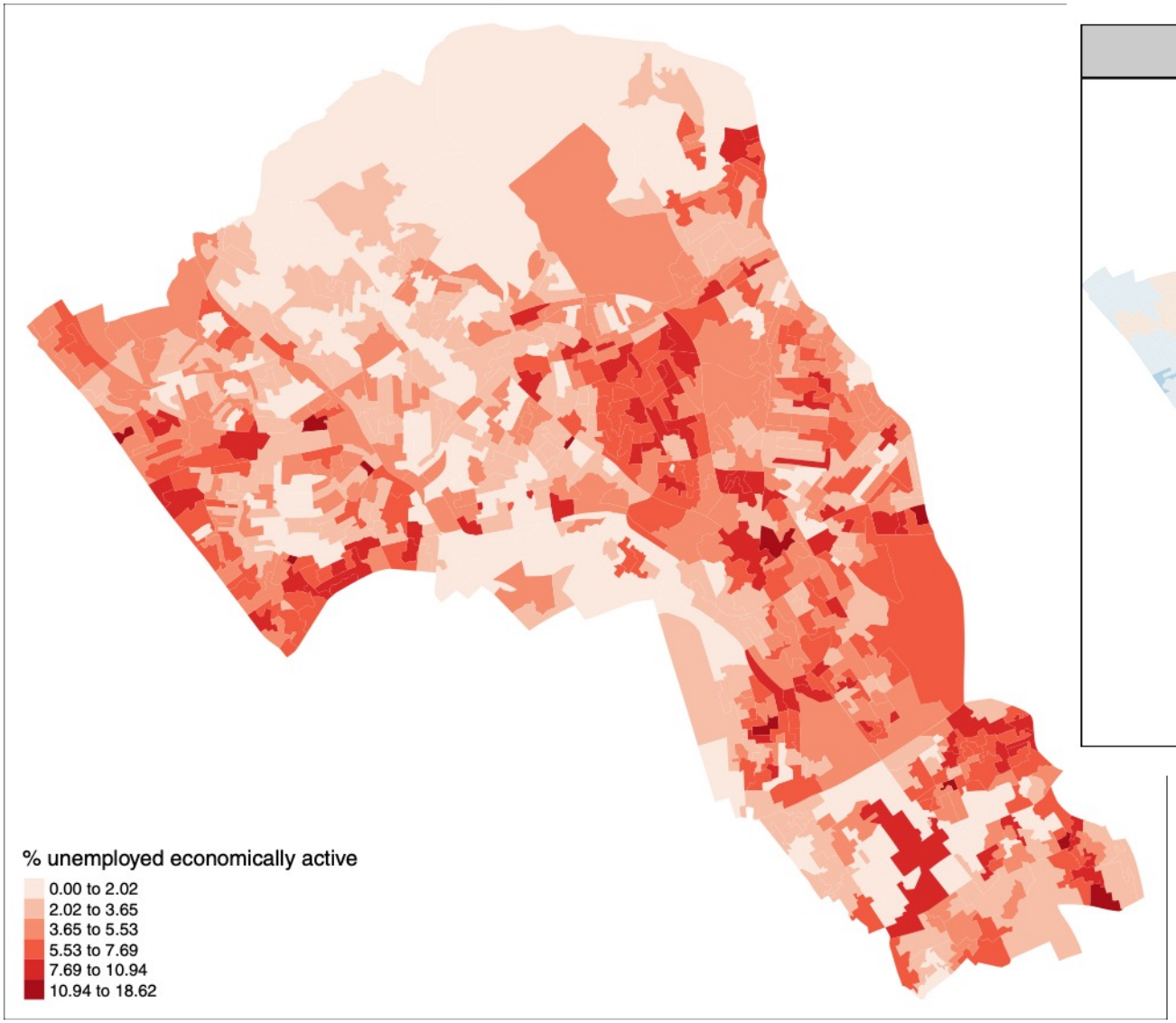


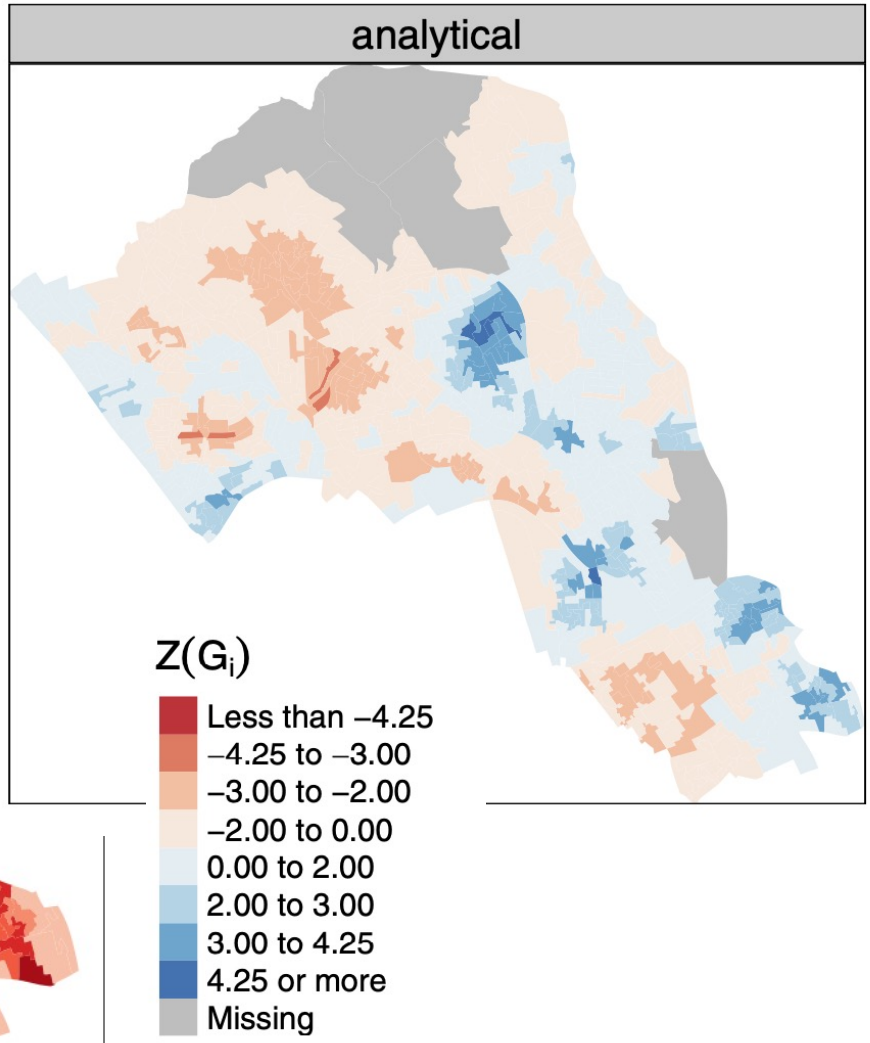
Local statistics: find outliers and/or clusters



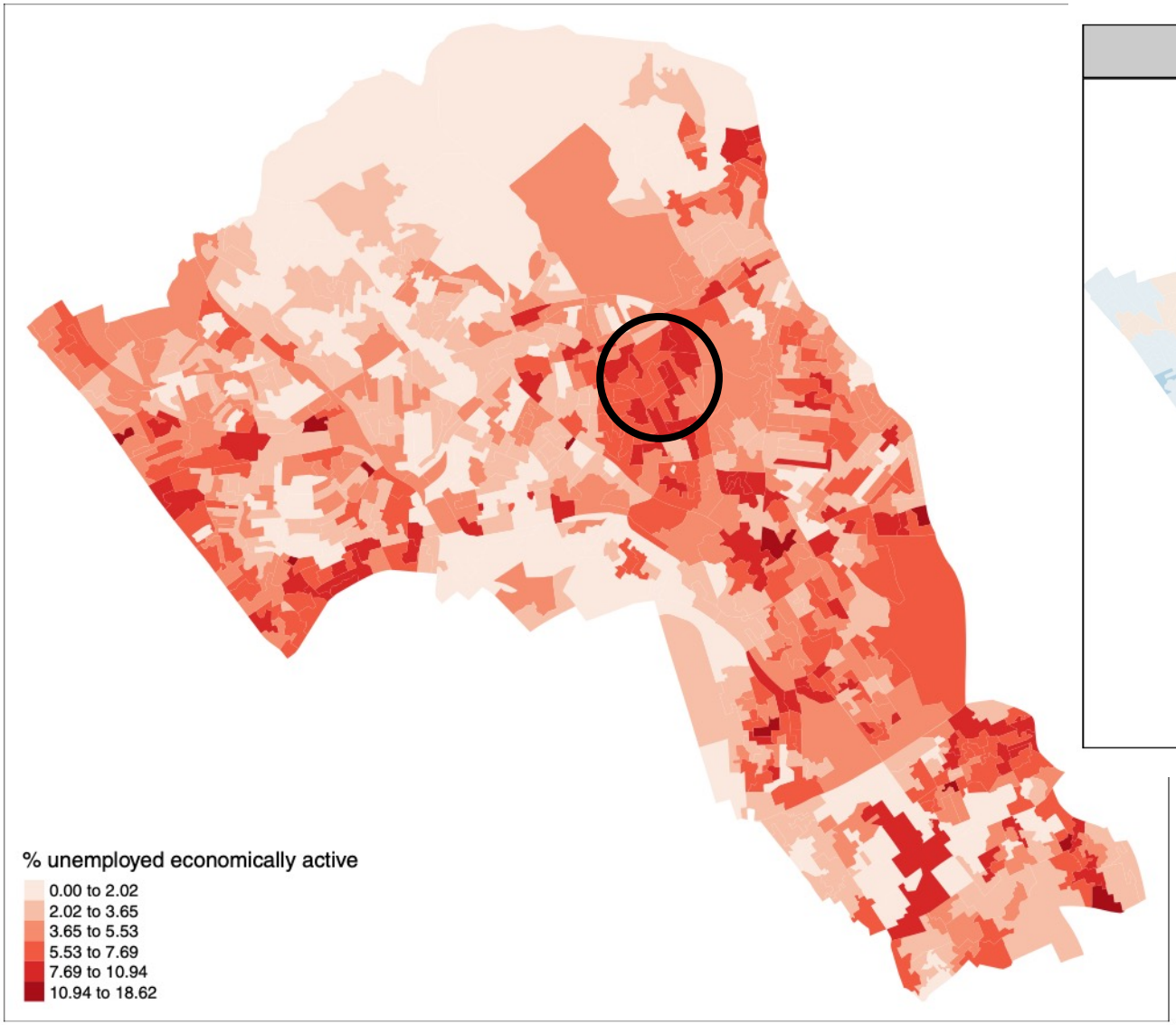
Is the distribution around this site what we'd expect?

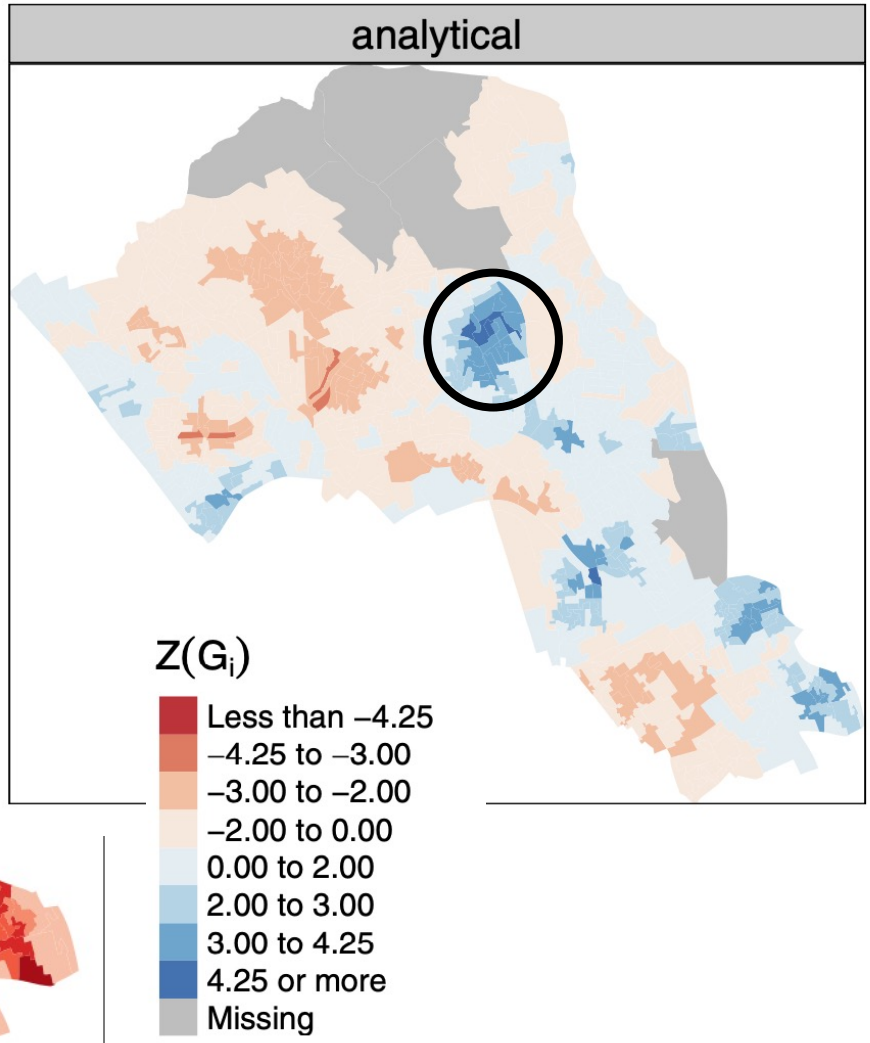
Local statistics: find outliers and/or clusters



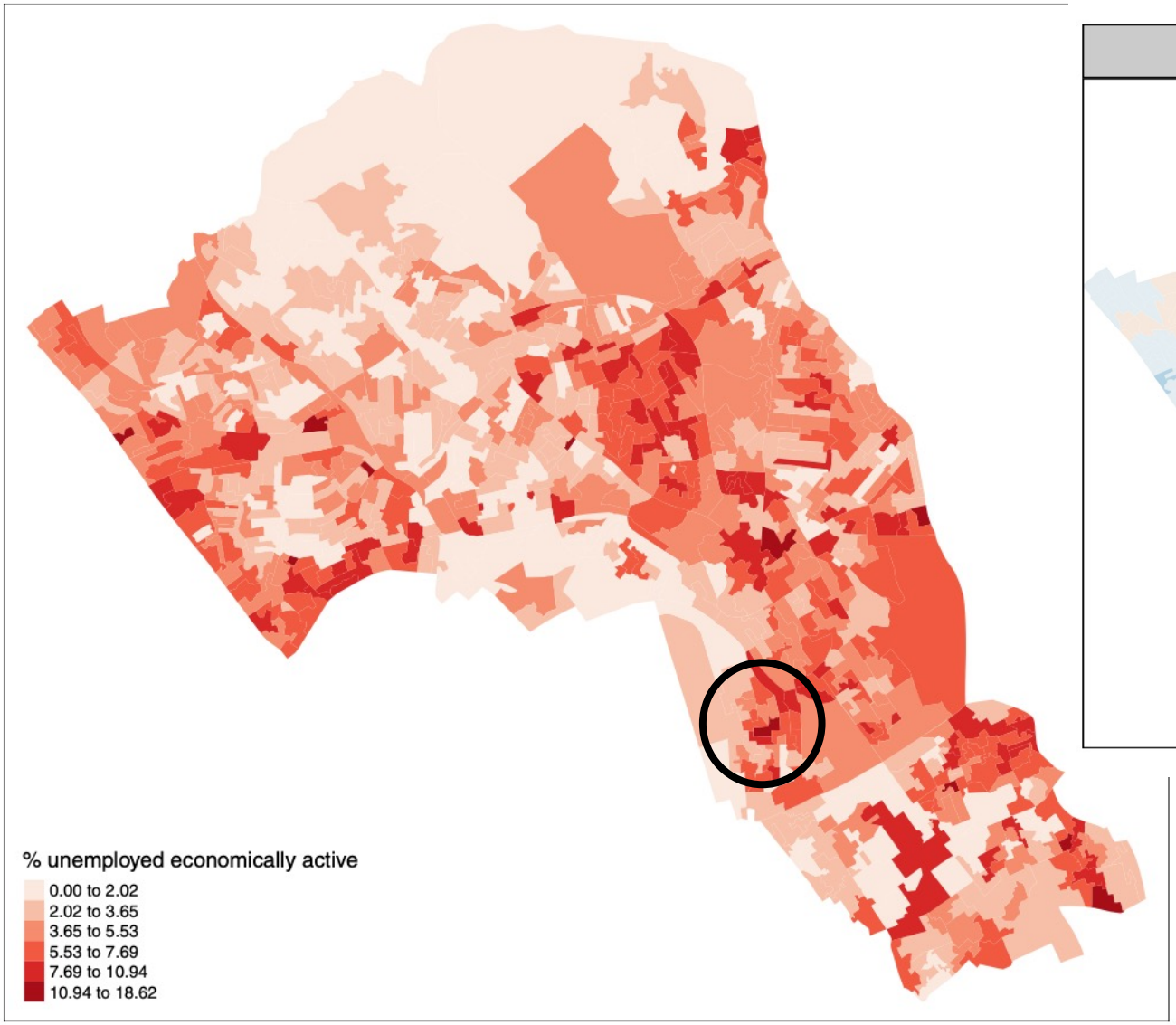


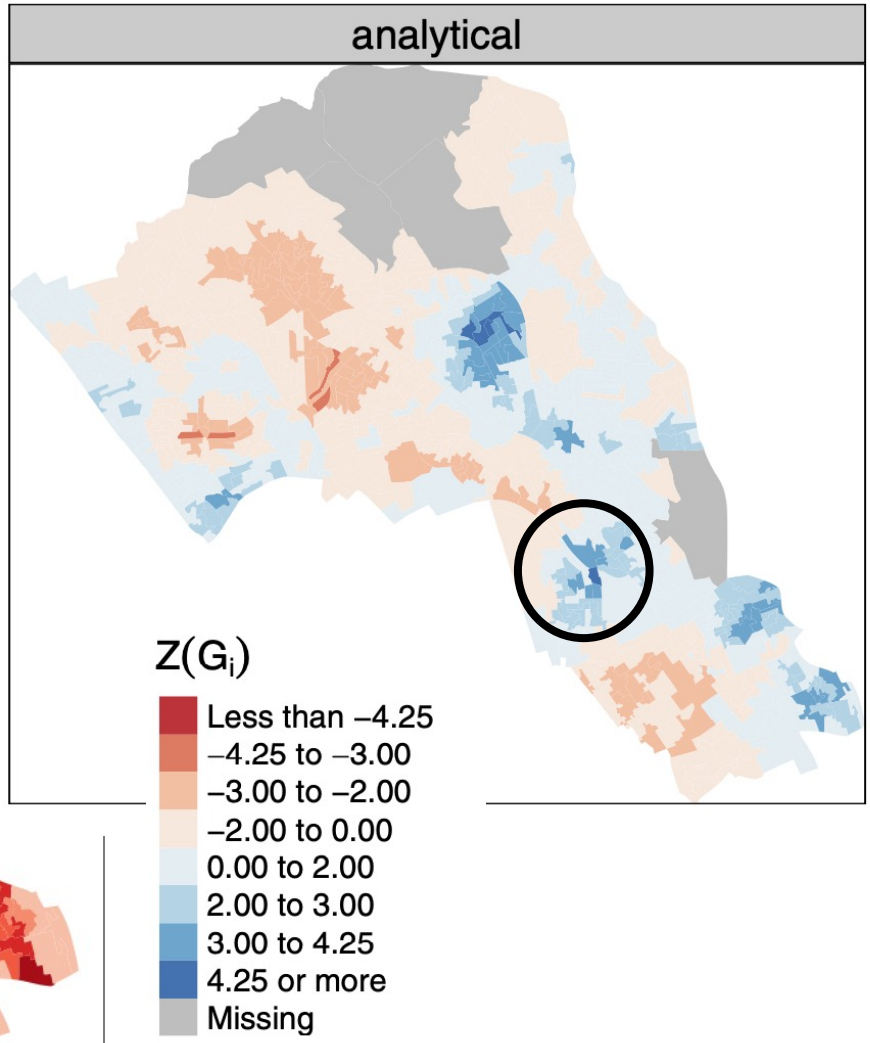
Local statistics: find outliers and/or clusters





Local statistics: find outliers and/or clusters





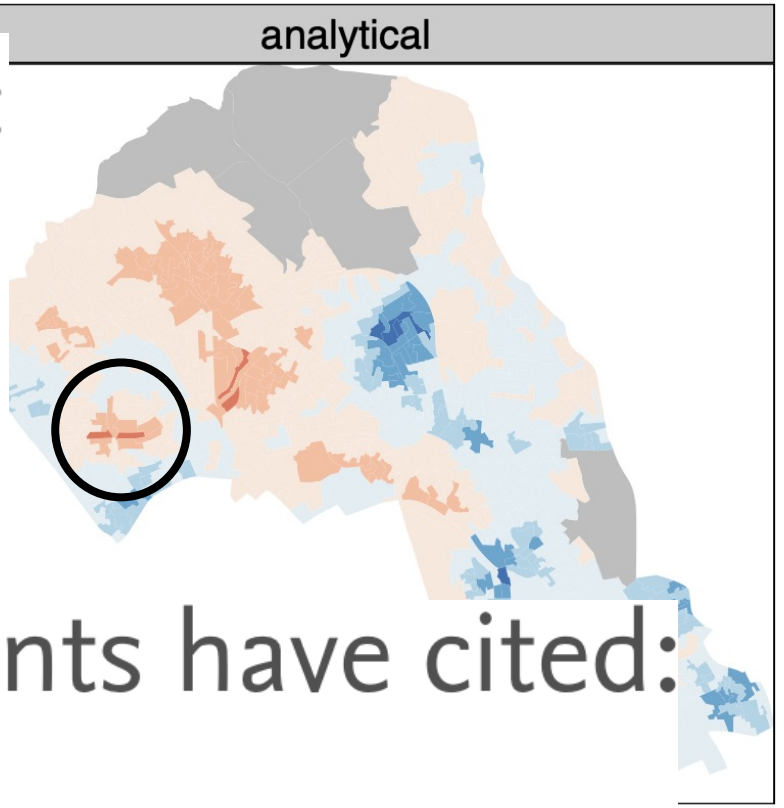
Local statistics: find outliers and/or clusters

5,679 documents have cited:

Local Indicators of Spatial Association—LISA

Anselin L.

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

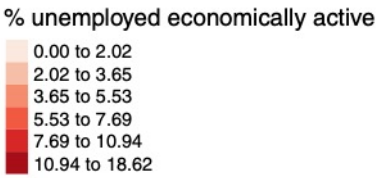


2,893 documents have cited:

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

LOCAL STATISTICS

FORMALIZING INTUITION

A REPLICATION FAILURE

"NULLS" IN LOCAL TESTS

The importance of null hypotheses in local statistics

LOCAL STATISTICS

FORMALIZING INTUITION

A REPLICATION FAILURE

FINDING "REAL" CLUSTERS

"NULLS" IN LOCAL TESTS

The importance of null hypotheses in local statistics

TEST (2018) 27:716–748
https://doi.org/10.1007/s11749-018-0599-x

ORIGINAL PAPER

Comparing implementations of global and local indicators of spatial association

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

Received: 4 April 2018 / Accepted: 14 July 2018 / Published online: 27 July 2018

doi.org/10.1007/s11749-018-0599-x

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

doi.org/10.1007/s11749-018-0599-x

ORIGINAL PAPER

Comparing implementations of global and local indicators of spatial association

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

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

doi.org/10.1007/s11749-018-0599-x

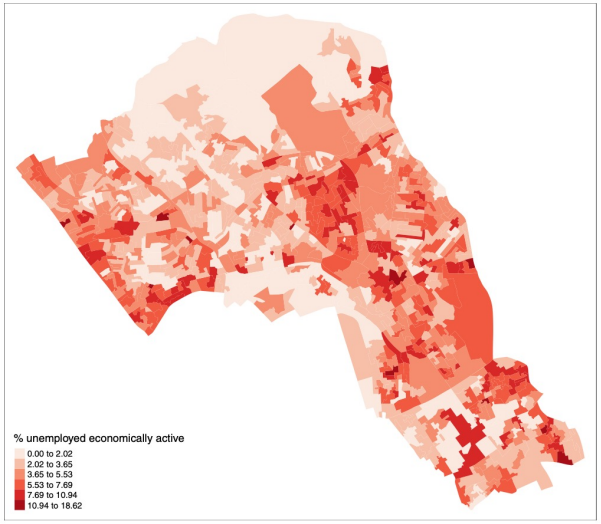
ORIGINAL PAPER

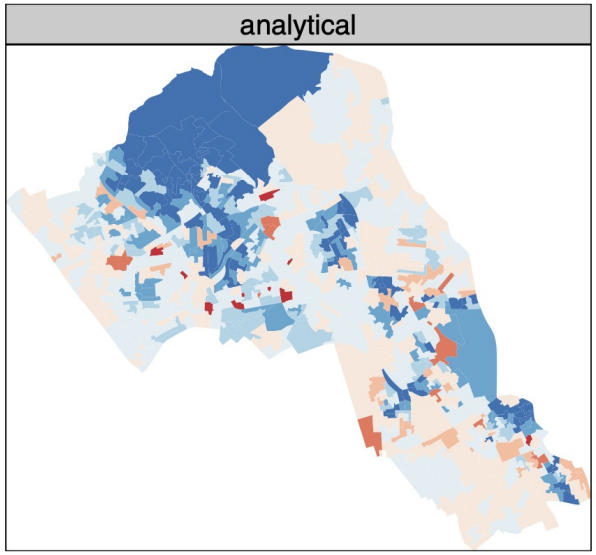
Comparing implementations of global and local indicators of spatial association

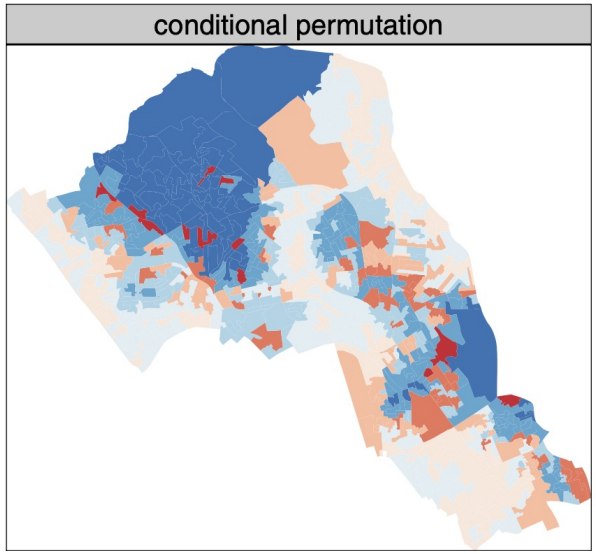
Roger S. Bivand[1] · David W. S. Wong[2]

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

doi.org/10.1007/s11749-018-0599-x

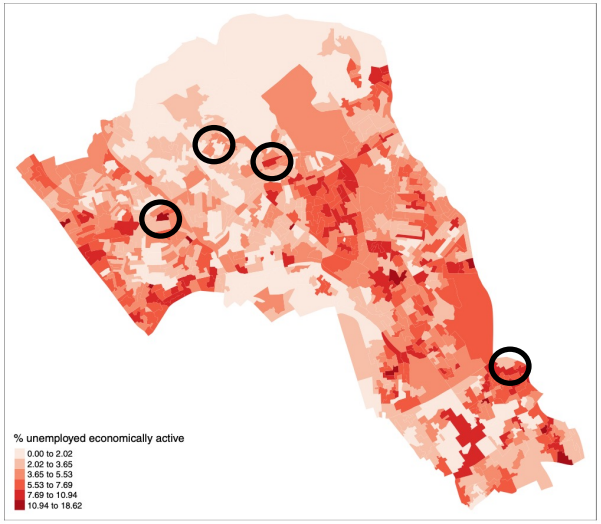


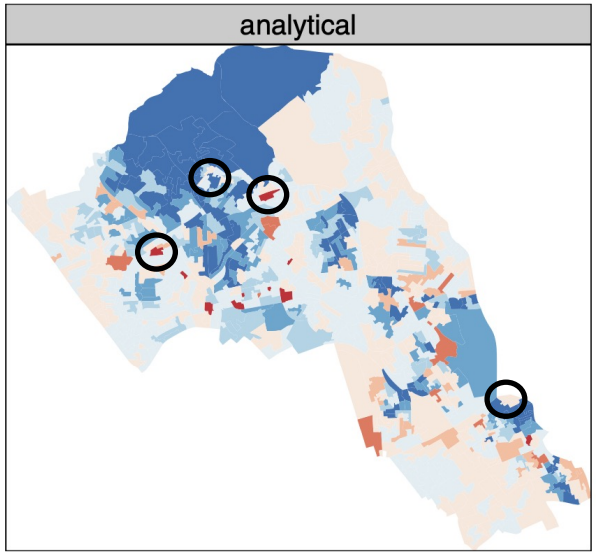


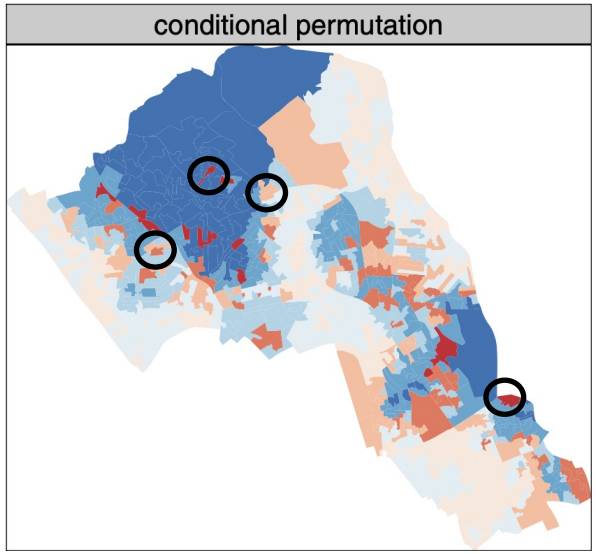


The permutation-based local z-scores were much more likely to be statistically significant […] Analytical inference yielded 267 significant OAs whereas permutation-based inference yielded 429. […] OAs were statistically significant using analytical inference but not using permutation-based inference,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

Sauer et al. (2021), p3

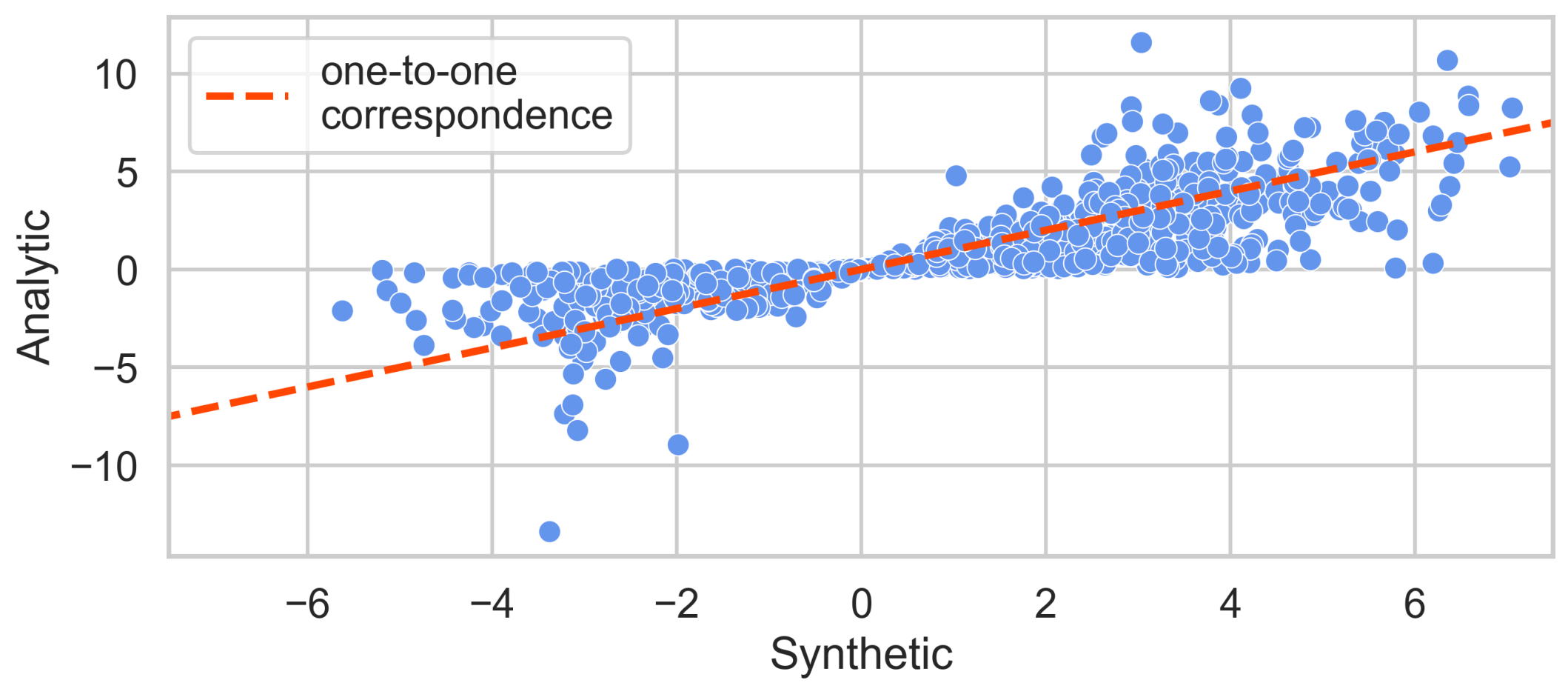






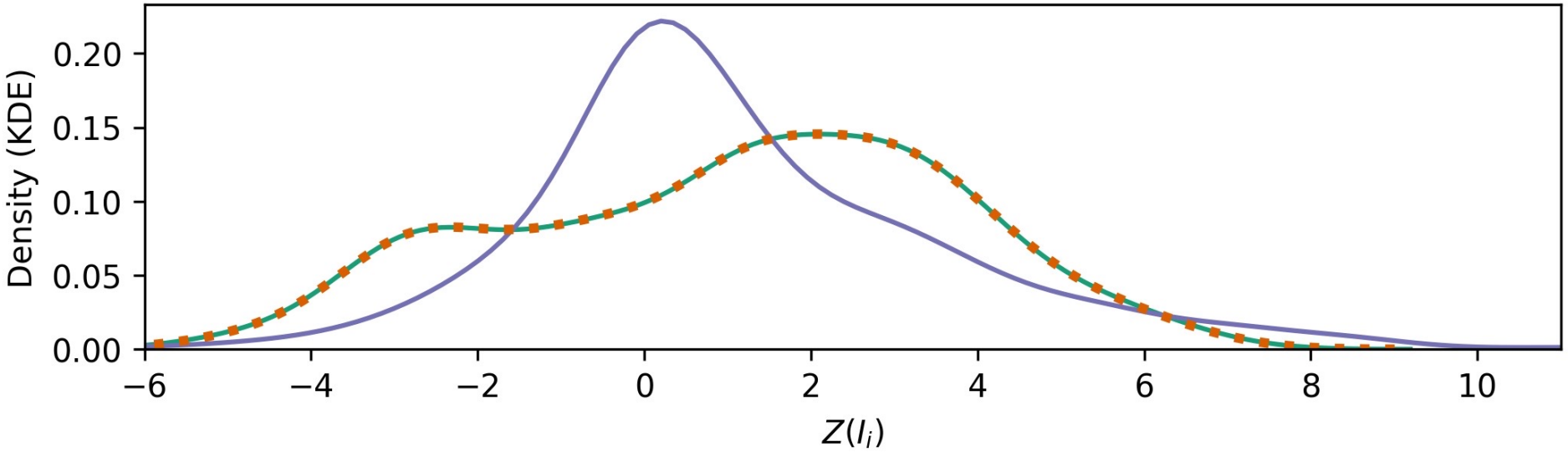
The permutation-based local z-scores were much more likely to be statistically significant […] Analytical inference yielded 267 significant OAs whereas permutation-based inference yielded 429. […] OAs were statistically significant using analytical inference but not using permutation-based inference,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

Sauer et al. (2021), p3



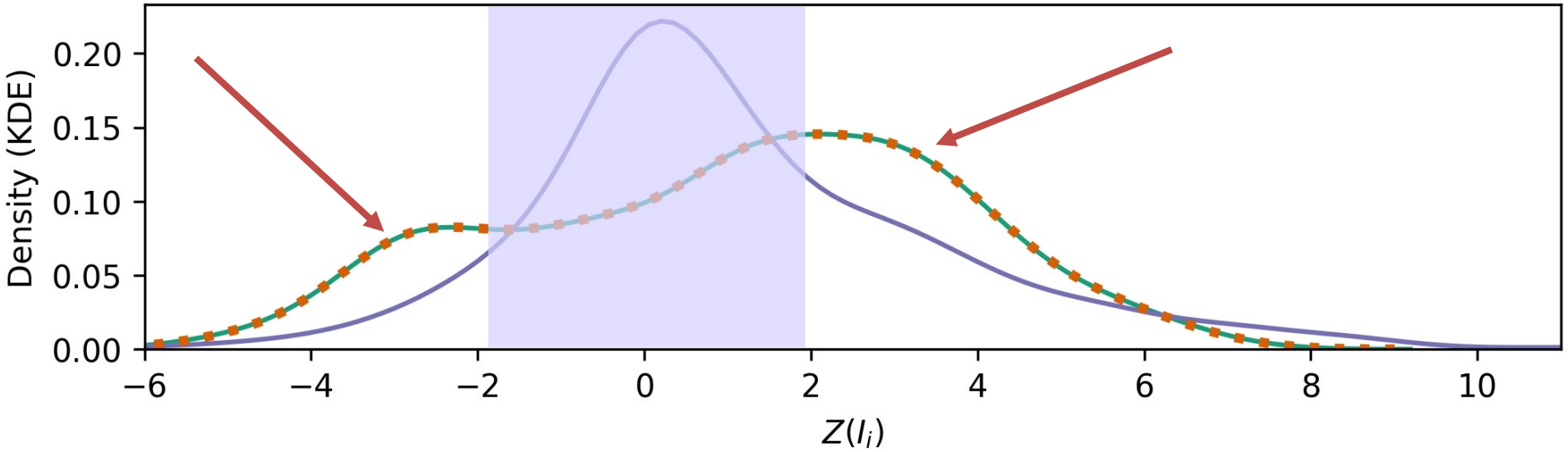
429. […] OAs were statistically significant using analytical inference but not using permutation-based inference, and vice versa.

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



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

geographical analysis

Geographical Analysis (2021) **0**, 1–17

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]

[1]Center for Geospatial Information Science, Department of Geographical Sciences, University of Maryland, College Park, MD USA, [2]Center for Geospatial Sciences, University of California Riverside, Riverside, CA USA, [3]School of Geographical Sciences, University of Bristol, Bristol, U.K.

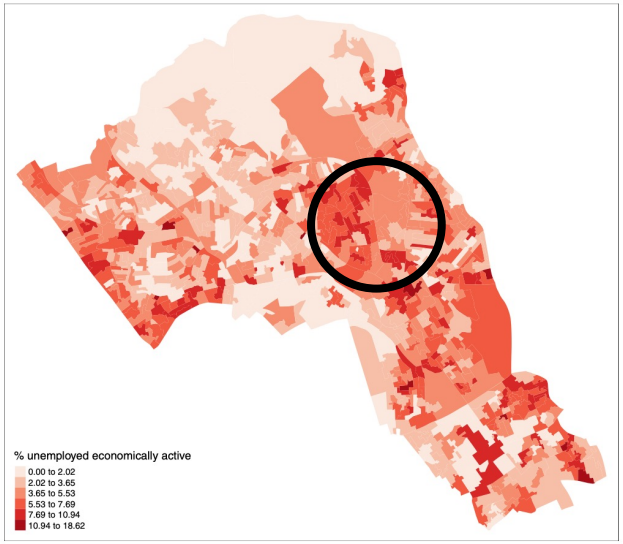
doi.org/10.1111/gean.12304

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]

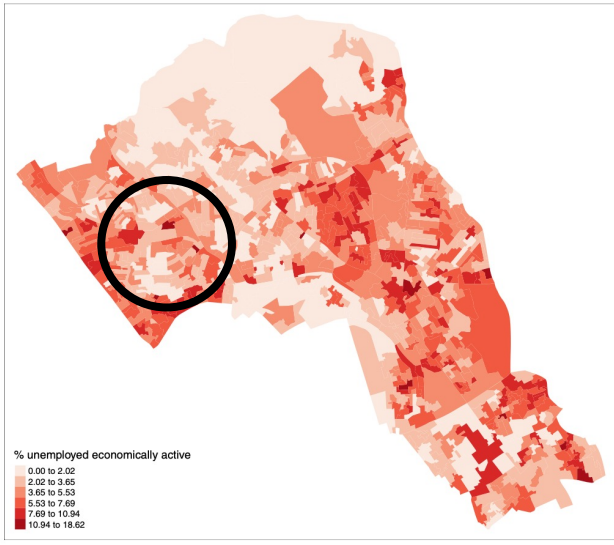
We see two different ways the divergence … could come about. Bivand & Wong (2018)'s theory is that local heteroskedasticity in the data is causing their … estimates to diverge. Our theory is that different null hypotheses drive the observed differences.

doi.org/10.1111/gean.12304



LOW
σ

HIGH
σ



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.

doi.org/10.1111/gean.12304

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.

doi.org/10.1111/gean.12304

LOCAL STATISTICS

FORMALIZING INTUITION

A REPLICATION FAILURE

FINDING "REAL" CLUSTERS

"NULLS" IN LOCAL TESTS

The importance of null hypotheses in local statistics

LOCAL STATISTICS

FORMALIZING INTUITION

A REPLICATION FAILURE

FINDING "REAL" CLUSTERS

"NULLS" IN LOCAL TESTS

ALL DEPENDS ON THEORY

The importance of null hypotheses in local statistics

ANALTYIC

SYNTHETIC

ANALTYIC

Assume normal observations distributed over a regular grid

Derive $\mathbf{E}[\Gamma]$ while assuming “all values are equally likely at any site”

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 site i constant

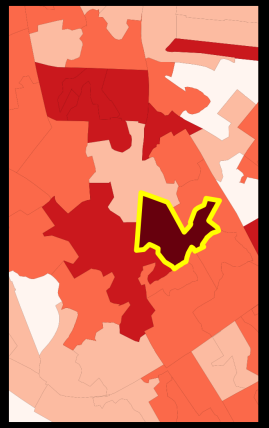
Shuffle the rest of the map

Compute a new local statistic

Repeat k - 1 times to obtain distribution of replicates, $\{\Gamma'\}$

Assess Γ against $\{\Gamma'\}$, so p is % of $\{\Gamma'\}$ at least as extreme as Γ

REAL



SYNTHETIC

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

Hold the value at site i constant

Shuffle the rest of the map

Compute a new local statistic

Repeat k - 1 times to obtain distribution of replicates, $\{\Gamma'\}$

Assess Γ against $\{\Gamma'\}$, so p is % of $\{\Gamma'\}$ at least as extreme as Γ

SYNTHETIC

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

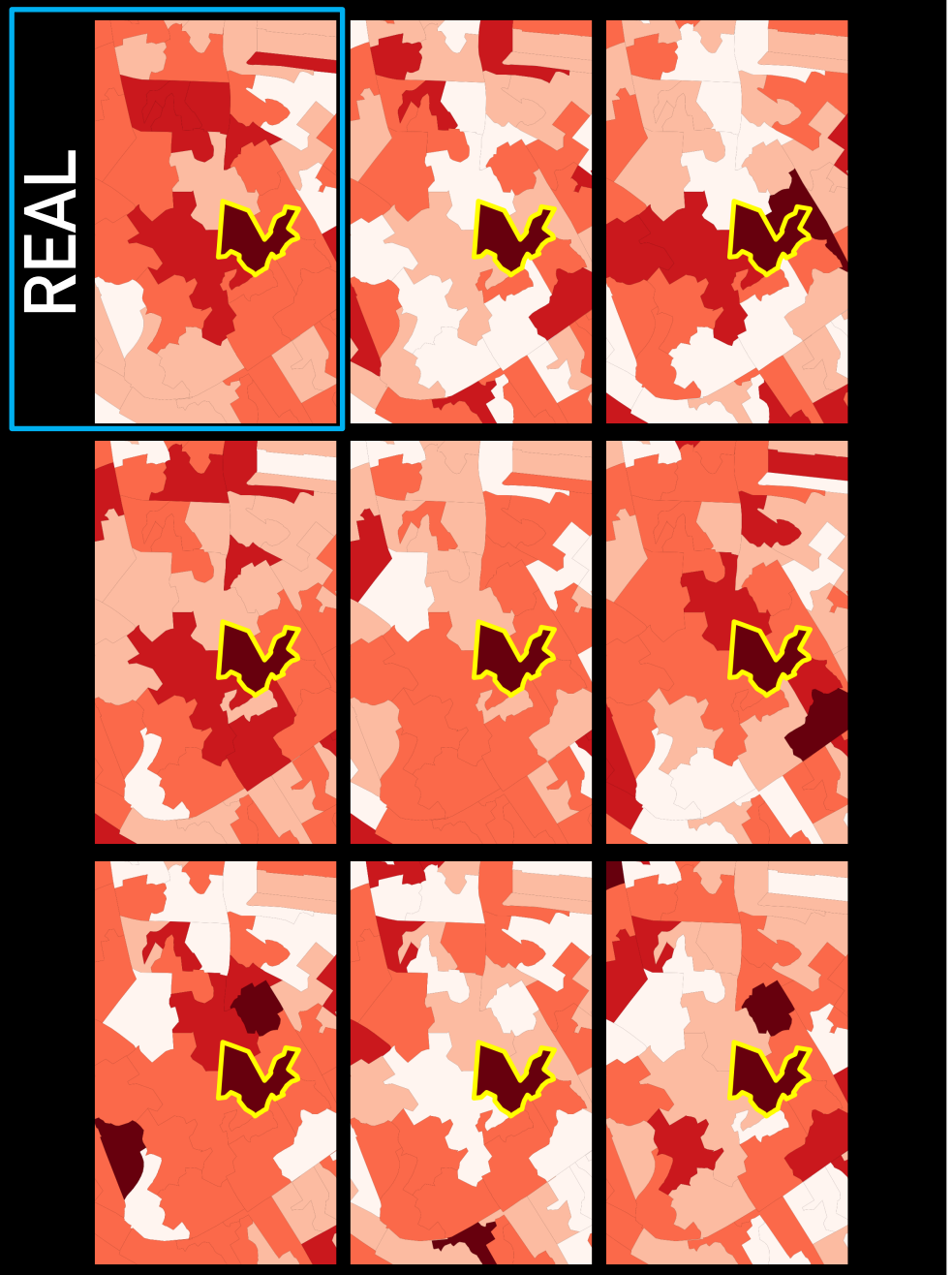
Hold the value at site i constant

Shuffle the rest of the map

Compute a new local statistic

Repeat k - 1 times to obtain distribution of replicates, $\{\Gamma'\}$

Assess Γ against $\{\Gamma'\}$, so p is % of $\{\Gamma'\}$ at least as extreme as Γ



SYNTHETIC

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

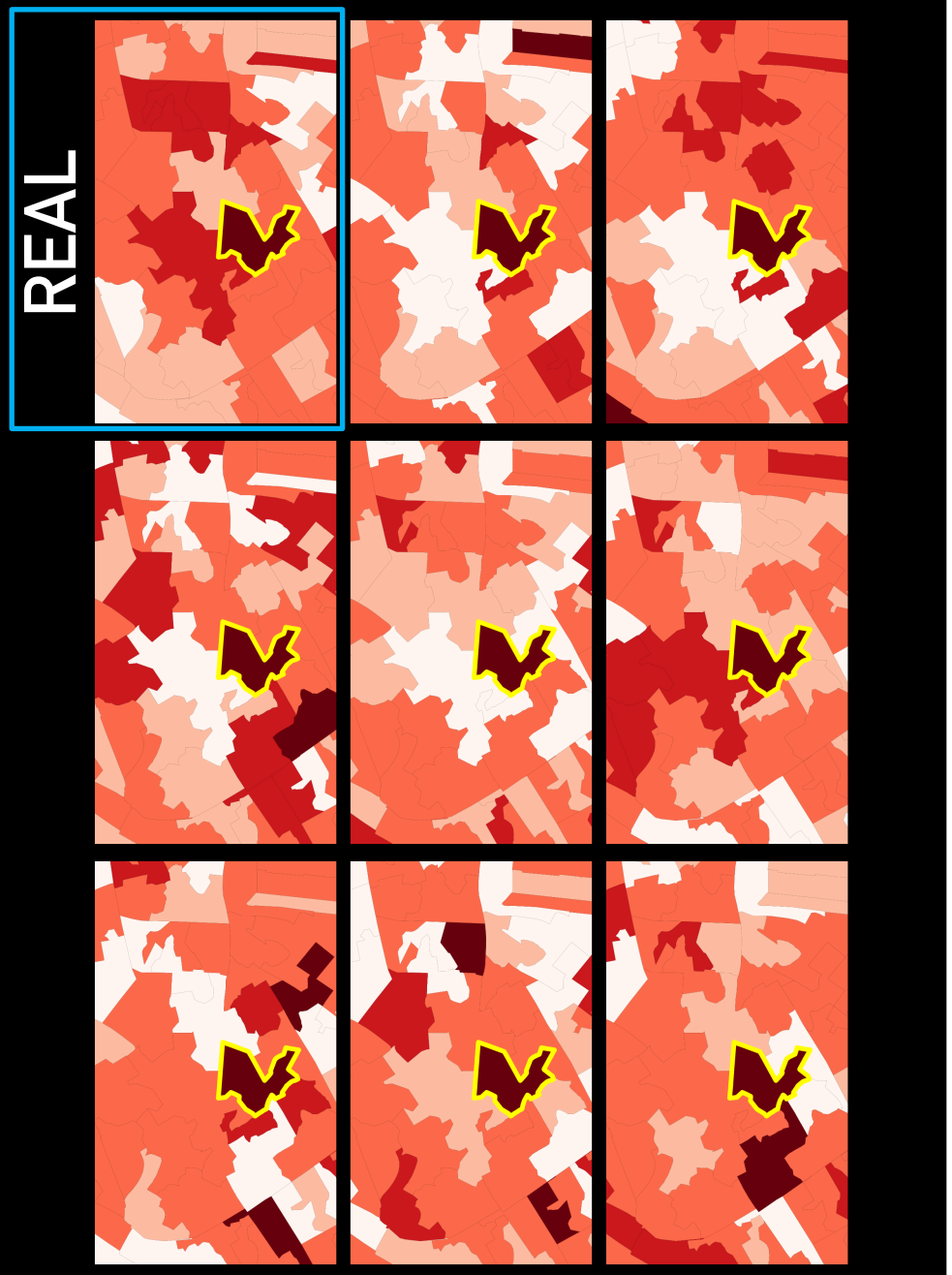
Hold the value at site i constant

Shuffle the rest of the map

Compute a new local statistic

Repeat k - 1 times to obtain distribution of replicates, $\{\Gamma'\}$

Assess Γ against $\{\Gamma'\}$, so p is % of $\{\Gamma'\}$ at least as extreme as Γ



SYNTHETIC

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

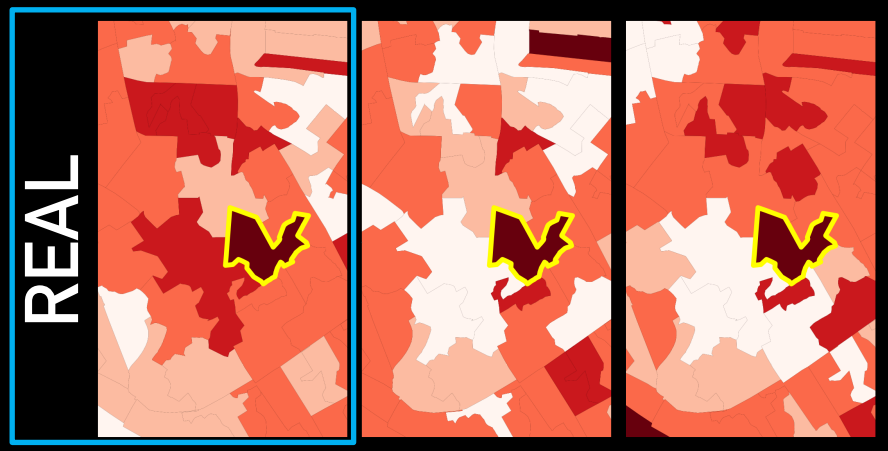
Hold the value at site i constant

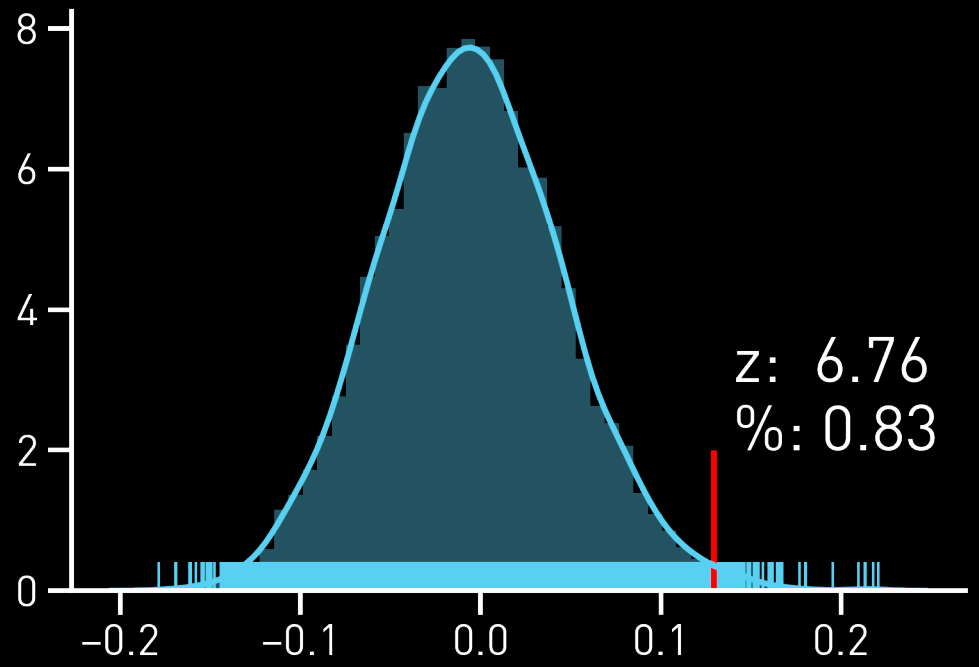
Shuffle the rest of the map

Compute a new local statistic

Repeat k - 1 times to obtain distribution of replicates, $\{\Gamma'\}$

Assess Γ against $\{\Gamma'\}$, so p is % of $\{\Gamma'\}$ at least as extreme as Γ





SYNTHETIC

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

Hold the value at site i constant

Shuffle the rest of the map

Compute a new local statistic

Repeat k - 1 times to obtain distribution of replicates, $\{\Gamma'\}$

Assess Γ against $\{\Gamma'\}$, so p is % of $\{\Gamma'\}$ at least as extreme as Γ

ANALTYIC

Assume normal observations distributed over a regular grid

Derive $\mathbf{E}[\mathbf{\Gamma}]$ while assuming "all values are equally likely at any site"

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

Construct z-score

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

Hope it generalizes!

SYNTHETIC

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

Hold the value at site i constant

Shuffle the rest of the map

Compute a new local statistic

Repeat k - 1 times to obtain distribution of replicates, $\{\mathbf{\Gamma'}\}$

Assess $\mathbf{\Gamma}$ against $\{\mathbf{\Gamma'}\}$, so p is % of $\{\mathbf{\Gamma'}\}$ at least as extreme as $\mathbf{\Gamma}$

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

?

ANALTYIC

Assume normal observations distributed over a regular grid

Derive $\mathbf{E}[\Gamma]$ while assuming "all values are equally likely at any site"

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

Construct z-score

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

Hope it generalizes!

SYNTHETIC

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

Hold the value at site i constant

Shuffle the rest of the map

Compute a new local statistic

Repeat k - 1 times to obtain distribution of replicates, $\{\Gamma'\}$

Assess Γ against $\{\Gamma'\}$, so p is % of $\{\Gamma'\}$ at least as extreme as Γ

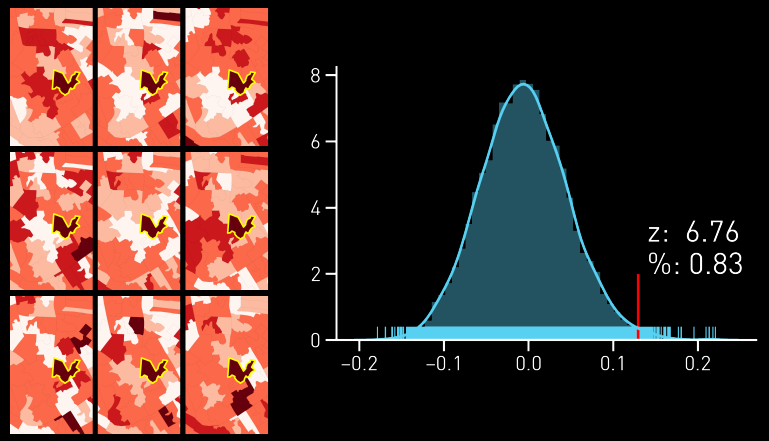
	"Total" Randomization Null *shuffle all values, so each value is equally likely at any site*	**"Conditional" Randomization Null** *for each site, the site's value is fixed and remaining sites are shuffled*
Analytical estimator *closed-form mathematical expressions for test statistics*	Originally in Anselin (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 $Var_p[I_i]$



	"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

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

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

CONDITIONAL

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

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

TOTAL

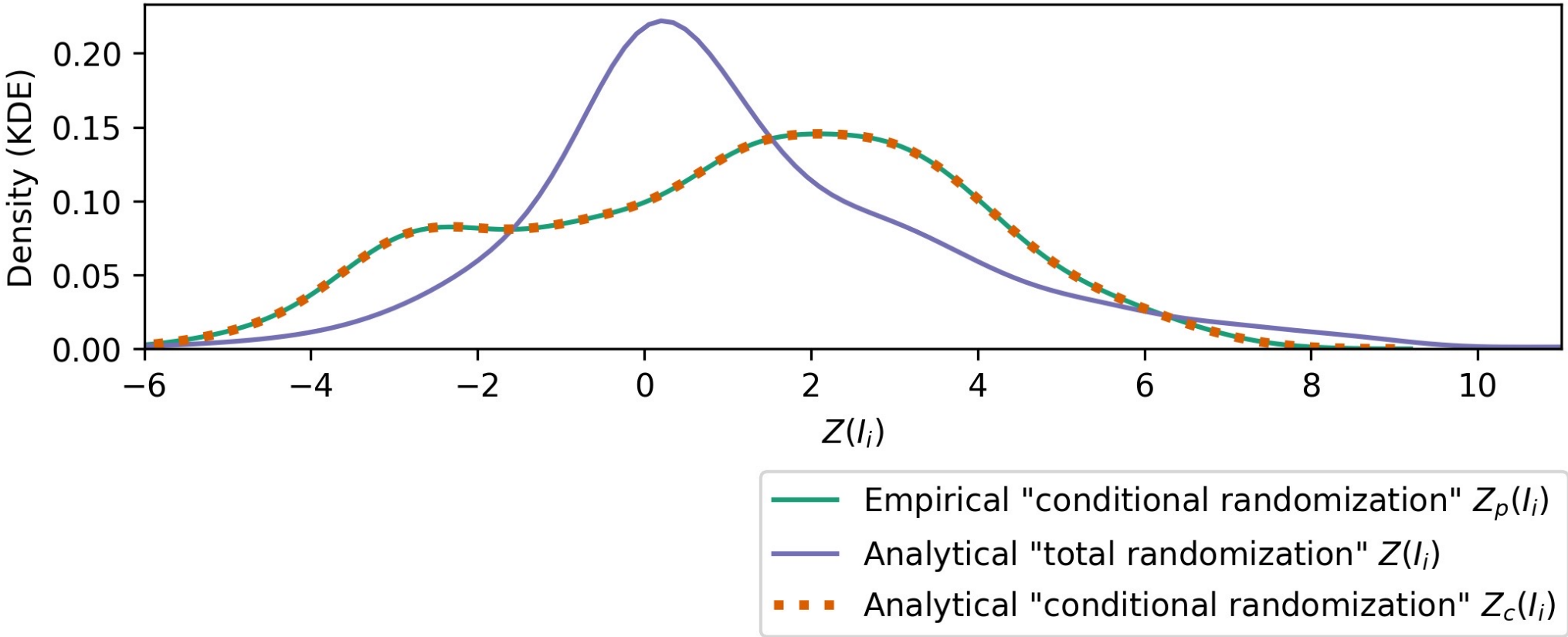
$$\mathbf{E}[I_i] = -\sum \frac{w_{ij}}{n-1}$$

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

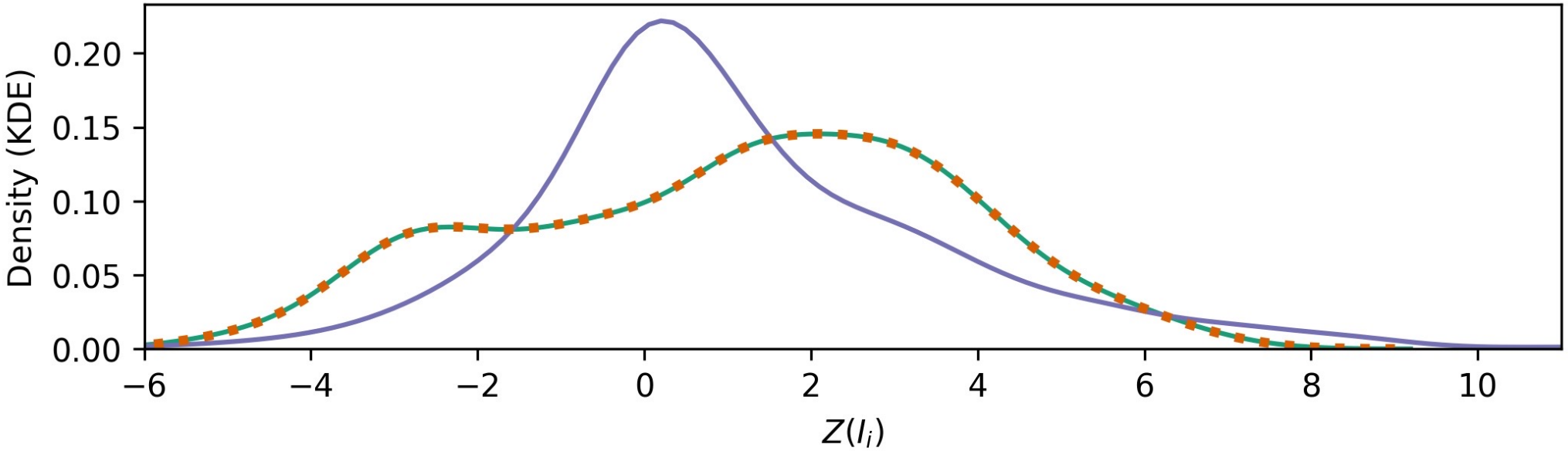
CONDITIONAL

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

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



Perfect match for Bivand & Wong (2018)'s data



	$Z(I_i)$	$Z_p(I_i)$	$Z_c(I_i)$
$Z(I_i)$	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" $Z_c(I_i)$

And practically never differ for random simulations

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.

LOCAL STATISTICS

FORMALIZING INTUITION

A REPLICATION FAILURE

FINDING "REAL" CLUSTERS

"NULLS" IN LOCAL TESTS

ALL DEPENDS ON THEORY

The importance of null hypotheses in local statistics

WHERE TO?

The importance of null hypotheses in local statistics

We do not have the right balance

a. Expecting something to replicate &
b. Reflexively arguing about its "complexity" when it doesn't

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

Michael F Goodchild & Wenwen Li
doi.org/10.1073/pnas.2015759118

"Democratizing" methods
makes our defaults "political!"

- Stats is a "science of defaults"
- These issues weren't ever solved, just supplanted by newer shinier problems

TESTING FOR LOCAL SPATIAL AUTOCORRELATION IN THE PRESENCE OF GLOBAL AUTOCORRELATION

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

MUCH ADO ABOUT NULL THINGS

A replication issue in spatial stats

JEFF SAUER
SERGIO REY
TAYLOR OSHAN
LEVI JOHN WOLF