

Spatial Eigenfunction Modeling of Geo-Referenced Data in the Social Sciences

Sebastian Juhl

University of Mannheim



June 24, 2021

- many phenomena of interest to social scientists cluster in space
 - economic development, regime type, voting behavior, religious beliefs, etc.
- geographic distribution of variables provide valuable information about the underlying mechanism of interest
- spatial autocorrelation (SA) causes severe problems for common econometric methods

Adequately accounting for spatial structures is necessary to guard against false inferences and to utilize spatial information!

- so far, social scientists use a limited subset of techniques suitable to handle SA
 - exploratory spatial analysis:
 - different types of maps
 - local indicators of SA (LISA)
 - inferential models
 - parametric spatial regression models

- so far, social scientists use a limited subset of techniques suitable to handle SA
 - exploratory spatial analysis:
 - different types of maps
 - local indicators of SA (LISA)
 - inferential models
 - parametric spatial regression models
- spatial eigenfunction analysis – particularly Moran eigenvector maps (MEM) – is a simple yet powerful tool to analyze cross-sectional data structures
 - ① identification & visualization of complex (multi-scale) spatial patterns
 - ② specification, estimation, and interpretation of inferential models
 - ③ partitioning of the variation in \mathbf{y} into individual components (space, covariates, joined)
 - ④ (structure-preserving) simulation of spatially autocorrelated data

Spatial eigenfunction analysis is based on the spectral decomposition of a centered (and symmetric/ symmetrized) connectivity matrix W :

$$MWM = E\Lambda E^{-1} = E\Lambda E'$$

- demeaning projector $M = (I - \mathbf{1}\mathbf{1}'/n)$ – also ensures that all eigenvectors are orthogonal and uncorrelated
- E are all n eigenvectors
- Λ is a diagonal matrix of the eigenvalues λ

Spatial eigenfunction analysis is based on the spectral decomposition of a centered (and symmetric/ symmetrized) connectivity matrix W :

$$\mathbf{MWM} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{-1} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}'$$

- demeaning projector $\mathbf{M} = (\mathbf{I} - \mathbf{1}\mathbf{1}'/n)$ – also ensures that all eigenvectors are orthogonal and uncorrelated
- \mathbf{E} are all n eigenvectors
- $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues λ

Direct relationship to the numerator of the global Moran coefficient:

$$MC(x) = \frac{n}{\mathbf{1}'\mathbf{W}\mathbf{1}} \frac{x'\mathbf{MWM}x}{x'\mathbf{M}x}$$

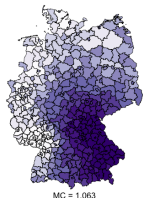
- E depict all possible distinct and mutually uncorrelated synthetic map patterns permitted by W
- corresponding eigenvalues λ indicate the level of SA

Example: 401 German NUTS-3 regions

- E depict all possible distinct and mutually uncorrelated synthetic map patterns permitted by W
- corresponding eigenvalues λ indicate the level of SA

Example: 401 German NUTS-3 regions

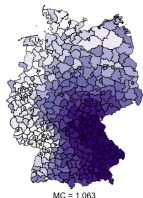
Eigenvector 1



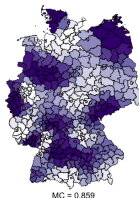
- E depict all possible distinct and mutually uncorrelated synthetic map patterns permitted by W
- corresponding eigenvalues λ indicate the level of SA

Example: 401 German NUTS-3 regions

Eigenvector 1



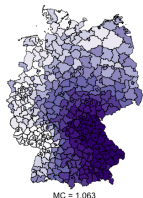
Eigenvector 25



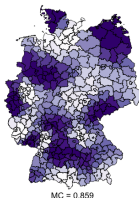
- E depict all possible distinct and mutually uncorrelated synthetic map patterns permitted by W
- corresponding eigenvalues λ indicate the level of SA

Example: 401 German NUTS-3 regions

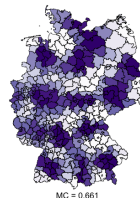
Eigenvector 1



Eigenvector 25



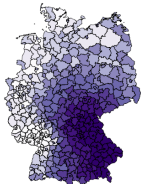
Eigenvector 50



- E depict all possible distinct and mutually uncorrelated synthetic map patterns permitted by W
- corresponding eigenvalues λ indicate the level of SA

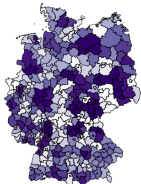
Example: 401 German NUTS-3 regions

Eigenvector 1



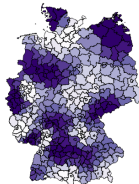
MC = 1.063

Eigenvector 75



MC = 0.514

Eigenvector 25



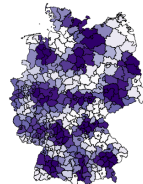
MC = 0.859

Eigenvector 170



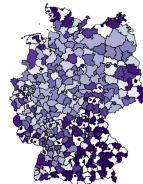
MC = 0.001

Eigenvector 50



MC = 0.661

Eigenvector 401

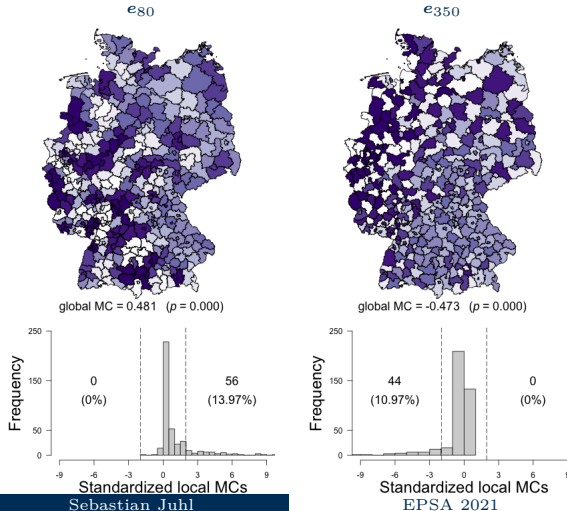


MC = -0.803

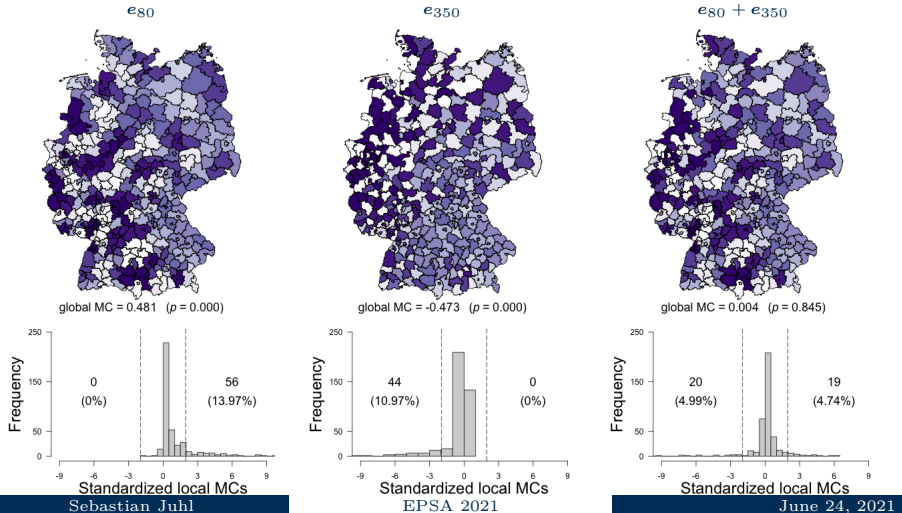
1. Identification of Complex Spatial Structures

- global and local MC statistics might fail to identify mixtures of positive and negative SA
- Example:

- global and local MC statistics might fail to identify mixtures of positive and negative SA
- Example:



- global and local MC statistics might fail to identify mixtures of positive and negative SA
- Example:



- eigenfunction analysis allows researchers to decompose the global MC into positively and negatively autocorrelated parts
 - $\text{global MC} = \text{MC}^+ + \text{MC}^-$
- decomposing the global MC helps identifying complex non-random spatial patterns

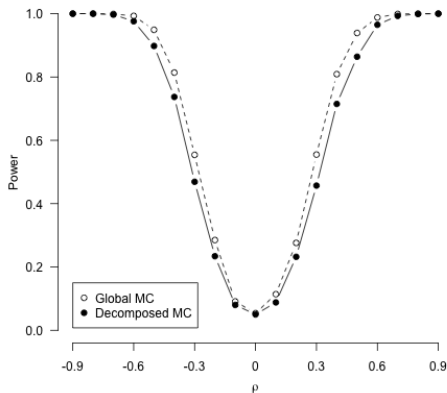
- eigenfunction analysis allows researchers to decompose the global MC into positively and negatively autocorrelated parts
 - global MC = $MC^+ + MC^-$
- decomposing the global MC helps identifying complex non-random spatial patterns

Monte Carlo setup:

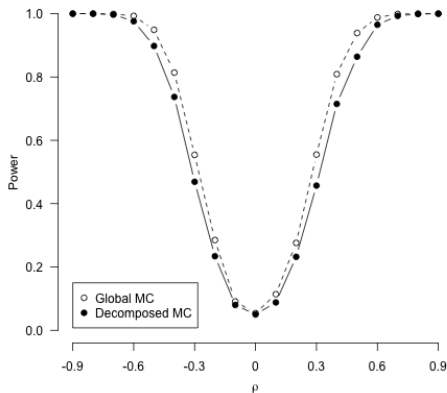
- $N = 100$ units ordered on a regular 10×10 grid
- \mathbf{W} is a symmetric contiguity matrix (rook scheme)
- ρ ranges from $-.9$ to $.9$ in steps of $.1$
- Scenario 1: simple spatial structure
 - $\mathbf{x} = (\mathbf{I} - \rho \mathbf{W}_g)^{-1} \mathbf{u}$
- Scenario 2: mixture of positive and negative SA
 - $\mathbf{x} = (\mathbf{I} - \rho \mathbf{W}_g)^{-1} \mathbf{u} + (\mathbf{I} - (-\rho) \mathbf{W}_g)^{-1} \mathbf{v}$

How does the decomposed MC based on spatial eigenfunctions performs compared to the global MC in terms of power?

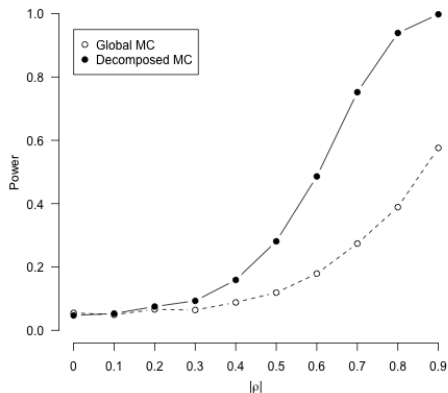
Scenario 1: Simple Spatial Structure



Scenario 1: Simple Spatial Structure



Scenario 2: Mixture of SA

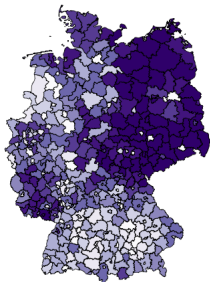


2. Model Specification, Estimation, and Interpretation

- SA causes severe problems for common econometric inferential techniques
- depending on the spatial DGP, SA can lead to
 - ❶ incorrect standard errors
 - ❷ biased and inconsistent parameter estimates
- spatial regression models address these problems but require many more or less rigid assumptions
 - knowledge of the true DGP
 - functional form assumptions
 - exact specification of SA in each regressor
 - difficult to estimate in a GLM framework
- semiparametric spatial filtering methods use Moran eigenvectors to construct a synthetic proxy variable that controls for SA

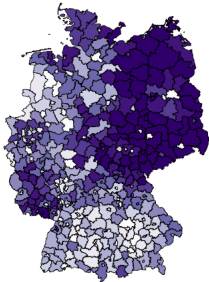
- a subset of eigenvectors can be combined to reproduce real-world map patterns
- Example: Median age in German NUTS-3 regions

Observed Median Age

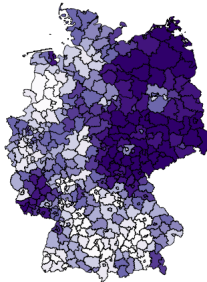


- a subset of eigenvectors can be combined to reproduce real-world map patterns
- Example: Median age in German NUTS-3 regions

Observed Median Age

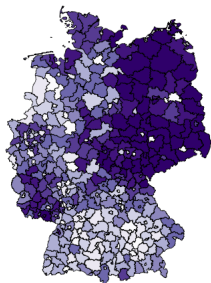


Synthetic Spatial Filter

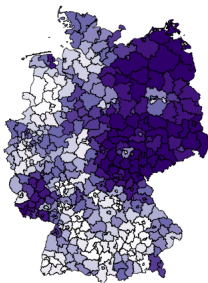


- a subset of eigenvectors can be combined to reproduce real-world map patterns
- Example: Median age in German NUTS-3 regions

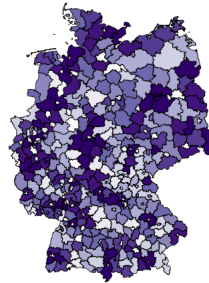
Observed Median Age



Synthetic Spatial Filter



Filtered Residuals

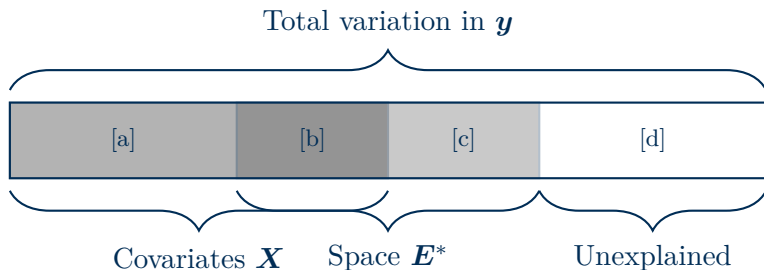


- a judiciously selected subset of eigenvectors controls for the underlying spatial pattern
- straightforward parameter estimation & interpretation

3. Variation Partitioning

How much variation in y is caused by X ?

- a common spatial structure spuriously inflates the share of variation explained by the predictors
- disentangle the individual contribution of the covariates and the spatial structure



- 1 identify a subset of eigenvectors \mathbf{E}^* that serve as spatial predictors

- 1 identify a subset of eigenvectors \mathbf{E}^* that serve as spatial predictors
- 2 regress \mathbf{y} on three sets of predictors and calculate R^2
 - 2.1 regress \mathbf{y} on \mathbf{X} (fraction $[a + b]$)
 - 2.2 regress \mathbf{y} on \mathbf{E}^* (fraction $[b + c]$)
 - 2.3 regress \mathbf{y} on \mathbf{X} and \mathbf{E}^* (fraction $[a + b + c]$)

- 1 identify a subset of eigenvectors \mathbf{E}^* that serve as spatial predictors
- 2 regress \mathbf{y} on three sets of predictors and calculate R^2
 - 2.1 regress \mathbf{y} on \mathbf{X} (fraction $[a + b]$)
 - 2.2 regress \mathbf{y} on \mathbf{E}^* (fraction $[b + c]$)
 - 2.3 regress \mathbf{y} on \mathbf{X} and \mathbf{E}^* (fraction $[a + b + c]$)
- 3 using the results from step 2, calculate individual fractions
 - 3.1 $[a] = [a + b + c] - [b + c]$
 - 3.2 $[b] = [a + b] + [b + c] - [a + b + c]$
 - 3.3 $[c] = [a + b + c] - [a + b]$
 - 3.4 $[d] = 1 - [a + b + c]$

- 1 identify a subset of eigenvectors \mathbf{E}^* that serve as spatial predictors
- 2 regress \mathbf{y} on three sets of predictors and calculate R^2
 - 2.1 regress \mathbf{y} on \mathbf{X} (fraction $[a + b]$)
 - 2.2 regress \mathbf{y} on \mathbf{E}^* (fraction $[b + c]$)
 - 2.3 regress \mathbf{y} on \mathbf{X} and \mathbf{E}^* (fraction $[a + b + c]$)
- 3 using the results from step 2, calculate individual fractions
 - 3.1 $[a] = [a + b + c] - [b + c]$
 - 3.2 $[b] = [a + b] + [b + c] - [a + b + c]$
 - 3.3 $[c] = [a + b + c] - [a + b]$
 - 3.4 $[d] = 1 - [a + b + c]$
- 4 use Moran spectral randomization to calculate R_{adj}^2

Example:

- regress GDP (\mathbf{y}) on median age (\mathbf{X})
- MC of (log) GDP: 0.347
- spatial filtering identifies 15 relevant eigenvectors (\mathbf{E}^*)

Example:

- regress GDP (\mathbf{y}) on median age (\mathbf{X})
- MC of (log) GDP: 0.347
- spatial filtering identifies 15 relevant eigenvectors (\mathbf{E}^*)

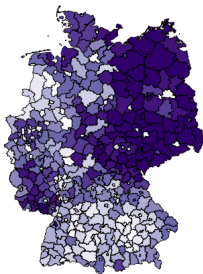
	<u>Joint Fractions</u>			<u>Individual Fractions</u>			
	$[a + b]$	$[b + c]$	$[a + b + c]$	$[a]$	$[b]$	$[c]$	$[d]$
R^2	0.286	0.246	0.523	0.277	0.009	0.237	0.477
R_{adj}^2	0.284	0.217	0.492	0.275	0.009	0.208	0.508

Note: Spatially constrained null model to calculate R_{adj}^2 based on 1,000 random permutations.

4. (Structure-Preserving) Simulation of Spatially Autocorrelated Data

- spatial multipliers $(\mathbf{I} - \rho \mathbf{W})^{-1}$ are typically used to simulate SA data
 - fixed degree of SA across simulations
 - does not preserve spatial structure
- using MEMs for simulation exercises preserves the geographic distribution

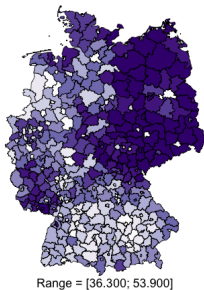
Observed Median Age



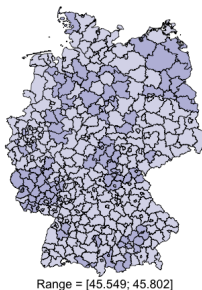
Range = [36.300; 53.900]

- spatial multipliers $(\mathbf{I} - \rho\mathbf{W})^{-1}$ are typically used to simulate SA data
 - fixed degree of SA across simulations
 - does not preserve spatial structure
- using MEMs for simulation exercises preserves the geographic distribution

Observed Median Age

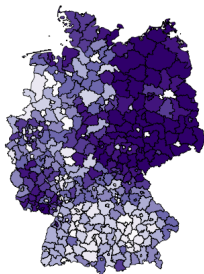


Spatial Multipliers (Mean)



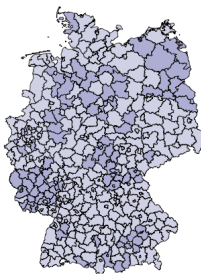
- spatial multipliers $(\mathbf{I} - \rho\mathbf{W})^{-1}$ are typically used to simulate SA data
 - fixed degree of SA across simulations
 - does not preserve spatial structure
- using MEMs for simulation exercises preserves the geographic distribution

Observed Median Age



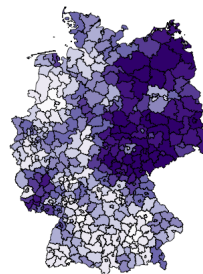
Range = [36.300; 53.900]

Spatial Multipliers (Mean)



Range = [45.549; 45.802]

MEMs (Mean)



Range = [42.206; 53.409]

- spatial eigenfunction analysis complements the statistical repertoire
- it also helps addressing methodological problems caused by SA
- MEMs allow researchers to derive additional information from geo-referenced data
- improves exploratory and inferential analysis, especially w.r.t
 - identification & visualization of complex (multi-scale) spatial patterns
 - specification, estimation, and interpretation of inferential models
 - variation partitioning
 - simulation of SA data