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

Sebastian Juhl

University of Mannheim

June 24, 2021

Motivation

- 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!

This Project

- 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

This Project

- 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 $\boldsymbol{y}$ into individual components (space, covariates, joined)
  4. (structure-preserving) simulation of spatially autocorrelated data

SA & Spatial Eigenfunctions

Spatial eigenfunction alanysis is based on the spectral decomposition of a centered (and symmetric/ symmetrized) connectivity matrix $\boldsymbol{W}$:

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

- demeaning projector $\boldsymbol{M} = (\boldsymbol{I} - \mathbf{11}'/n)$ – also ensures that all eigenvectors are orthogonal and uncorrelated
- $\boldsymbol{E}$ are all n eigenvectors
- $\boldsymbol{\Lambda}$ is a diagonal matrix of the eigenvalues $\boldsymbol{\lambda}$

SA & Spatial Eigenfunctions

Spatial eigenfunction alanysis is based on the spectral decomposition of a centered (and symmetric/ symmetrized) connectivity matrix $\boldsymbol{W}$:

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

- demeaning projector $\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{11}'/n)$ – also ensures that all eigenvectors are orthogonal and uncorrelated
- $\boldsymbol{E}$ are all n eigenvectors
- $\boldsymbol{\Lambda}$ is a diagonal matrix of the eigenvalues $\boldsymbol{\lambda}$

Direct relationship to the numerator of the global Moran coefficient:

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

Eigenvectors as Synthetic Map Patterns

- $\boldsymbol{E}$ depict <u>all</u> possible distinct and mutually uncorrelated synthetic map patterns permitted by $\boldsymbol{W}$
- corresponding eigenvalues $\boldsymbol{\lambda}$ indicate the level of SA

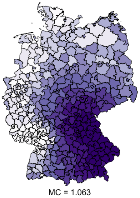
<u>Example:</u> 401 German NUTS-3 regions

Eigenvectors as Synthetic Map Patterns

- $\boldsymbol{E}$ depict <u>all</u> possible distinct and mutually uncorrelated synthetic map patterns permitted by $\boldsymbol{W}$
- corresponding eigenvalues $\boldsymbol{\lambda}$ indicate the level of SA

Example: 401 German NUTS-3 regions

Eigenvector 1

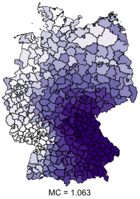


Eigenvectors as Synthetic Map Patterns

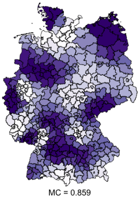
- $\boldsymbol{E}$ depict all possible distinct and mutually uncorrelated synthetic map patterns permitted by $\boldsymbol{W}$
- corresponding eigenvalues $\boldsymbol{\lambda}$ indicate the level of SA

Example: 401 German NUTS-3 regions

Eigenvector 1



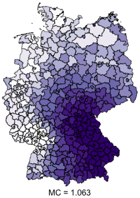
Eigenvector 25



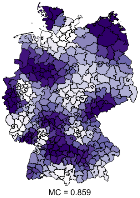
- $\boldsymbol{E}$ depict all possible distinct and mutually uncorrelated synthetic map patterns permitted by $\boldsymbol{W}$
- corresponding eigenvalues $\boldsymbol{\lambda}$ indicate the level of SA

Example: 401 German NUTS-3 regions

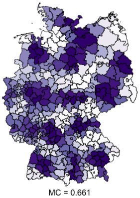
Eigenvector 1



Eigenvector 25



Eigenvector 50

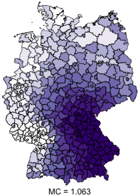


Eigenvectors as Synthetic Map Patterns

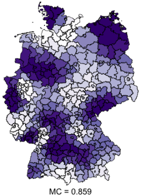
- ***E*** depict <u>all</u> possible distinct and mutually uncorrelated synthetic map patterns permitted by ***W***
- corresponding eigenvalues $\boldsymbol{\lambda}$ indicate the level of SA

<u>Example:</u> 401 German NUTS-3 regions

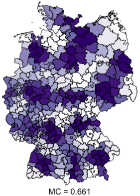
Eigenvector 1



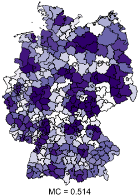
Eigenvector 25



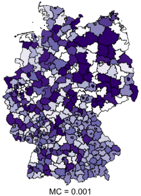
Eigenvector 50



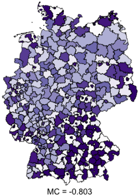
Eigenvector 75



Eigenvector 170



Eigenvector 401



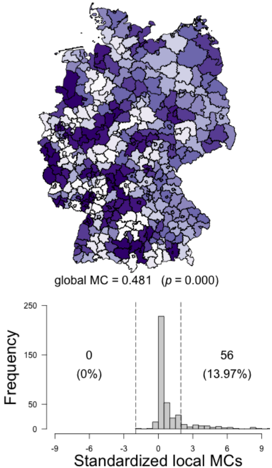
1. Identification of Complex Spatial Structures

Identification of Complex SA Structures

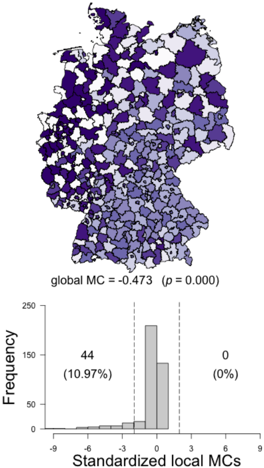
- global and local MC statistics might fail to identify mixtures of positive and negative SA
- Example:

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

e_{80}



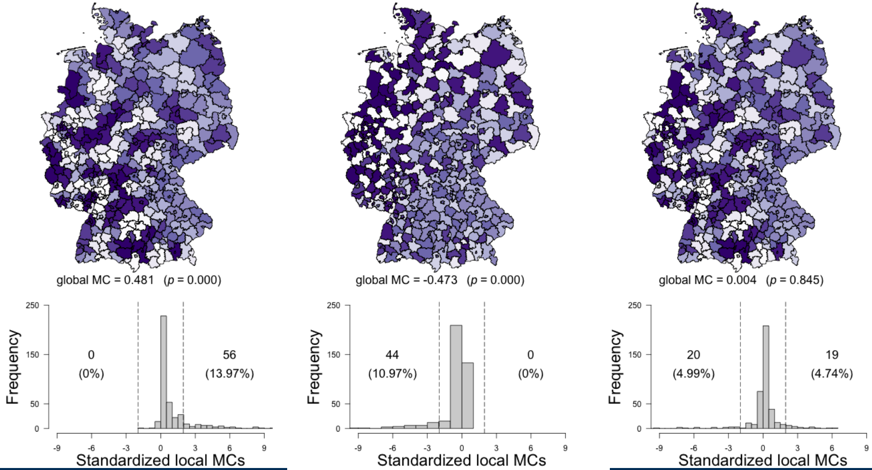
e_{350}



Identification of Complex SA Structures

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

e_{80} | e_{350} | $e_{80} + e_{350}$



Identification of Complex SA Structures

- 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

Identification of Complex SA Structures

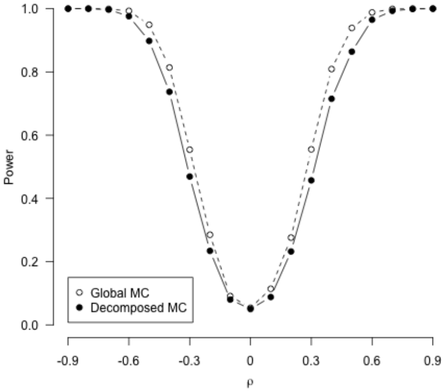
- eigenfunction analysis allows researchers to decompose the global MC into positively and negatively autocorrelated parts
  - global MC = $\text{MC}^{+} + \text{MC}^{-}$
- decomposing the global MC helps identifying complex non-random spatial patterns

Monte Carlo setup:

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

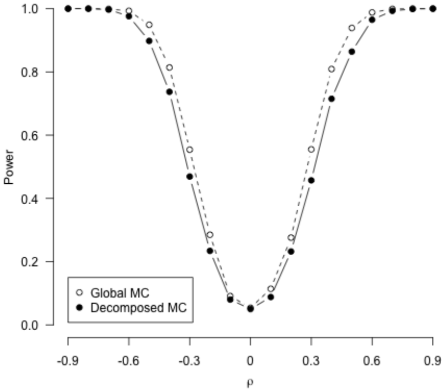
How does the decomposed MC based on spatial eigenfunctions performs compared to the global MC in terms of power?

Scenario 1: Simple Spatial Structure

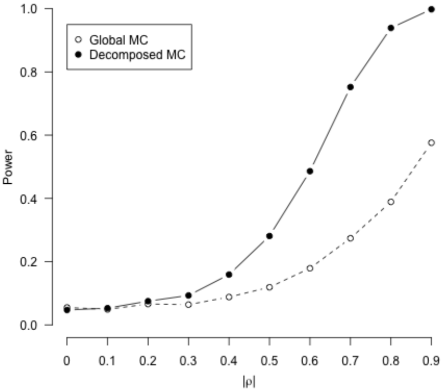


Identification of Complex SA Structures

Scenario 1: Simple Spatial Structure



Scenario 2: Mixture of SA



2. Model Specification, Estimation, and Interpretation

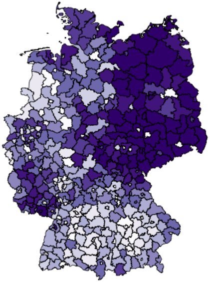
Specification, Estimation, and Interpretation

- SA causes severe problems for common econometric inferential techniques
- depending on the spatial DGP, SA can lead to
  1. incorrect standard errors
  2. 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

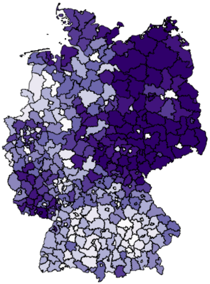
- a subset of eigenvectors can be combined to reproduce real-world map patterns
- Example: Median age in German NUTS-3 regions (2017)

Observed Median Age

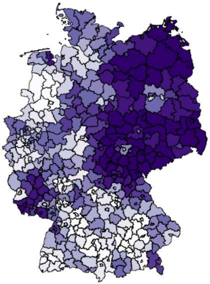


- a subset of eigenvectors can be combined to reproduce real-world map patterns
- Example: Median age in German NUTS-3 regions (2017)

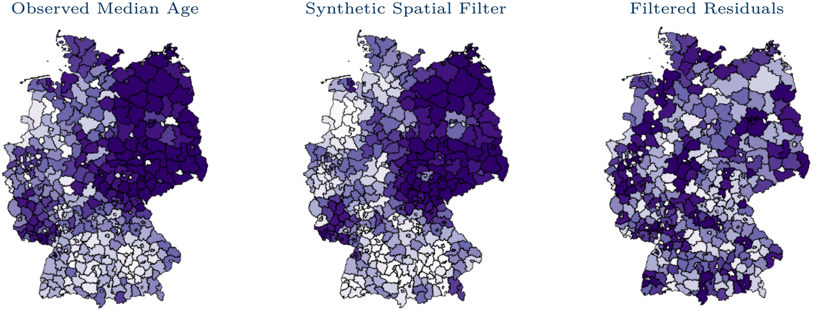
Observed Median Age



Synthetic Spatial Filter



- a subset of eigenvectors can be combined to reproduce real-world map patterns
- Example: Median age in German NUTS-3 regions (2017)



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

3. Variation Partitioning

Variation Partitioning

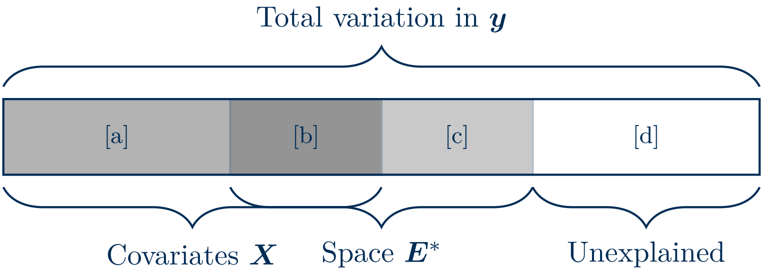
How much variation in $\boldsymbol{y}$ is caused by $\boldsymbol{X}$?

- 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

Variation Partitioning

How much variation in $\boldsymbol{y}$ is caused by $\boldsymbol{X}$?

- 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



Variation Partitioning

UNIVERSITY OF MANNHEIM

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

Variation Partitioning

UNIVERSITY OF MANNHEIM

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

2 regress $\boldsymbol{y}$ on three sets of predictors and calculate R^2

  2.1 regress $\boldsymbol{y}$ on $\boldsymbol{X}$ (fraction $[a+b]$)

  2.2 regress $\boldsymbol{y}$ on $\boldsymbol{E}^*$ (fraction $[b+c]$)

  2.3 regress $\boldsymbol{y}$ on $\boldsymbol{X}$ and $\boldsymbol{E}^*$ (fraction $[a+b+c]$)

Variation Partitioning

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

Variation Partitioning

Example:

- regress GDP ($\boldsymbol{y}$) on median age ($\boldsymbol{X}$)
- MC of (log) GDP: 0.347 ($p = 0.000$)
- spatial filtering identifies 15 relevant eigenvectors ($\boldsymbol{E}^*$)

Variation Partitioning

UNIVERSITY OF MANNHEIM

Example:

- regress GDP ($\boldsymbol{y}$) on median age ($\boldsymbol{X}$)
- MC of (log) GDP: 0.347 ($p = 0.000$)
- spatial filtering identifies 15 relevant eigenvectors ($\boldsymbol{E}^*$)

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

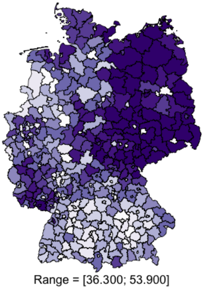
Note: Spatially constrained null model to calculate R^2_{adj} based on $1,000$ random permutations.

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

Simulating Spatial Data

- spatial multipliers $(\boldsymbol{I} - \rho \boldsymbol{W})^{-1}$ 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

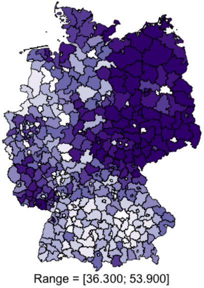
Observed Median Age



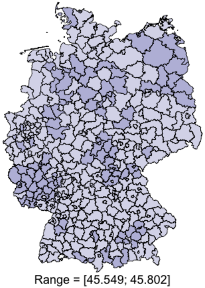
Simulating Spatial Data

- spatial multipliers $(\boldsymbol{I} - \rho \boldsymbol{W})^{-1}$ 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



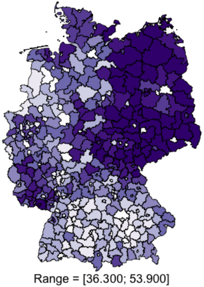
Spatial Multipliers (Means)



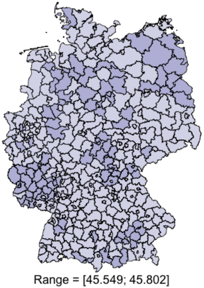
Simulating Spatial Data

- spatial multipliers $(\boldsymbol{I} - \rho\boldsymbol{W})^{-1}$ 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

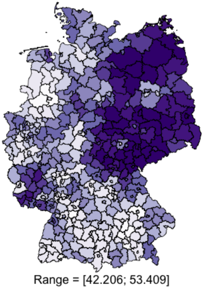
Observed Median Age



Spatial Multipliers (Means)



MEMs (Means)



Conclusion

- spatial eigenfunction analysis complements the statistical repertoire
  - it 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

Additional Resources

`spfilteR` package:

CRAN: https://CRAN.R-project.org/package=spfilteR

GitHub: https://github.com/sjuhl/spfilteR

Feedback & suggestions are highly appreciated!

`sebastian.juhl@gess.uni-mannheim.de`

www.sebastianjuhl.com