From dots to the results: spatial methods for geo-located data

Katarzyna Kopczewska

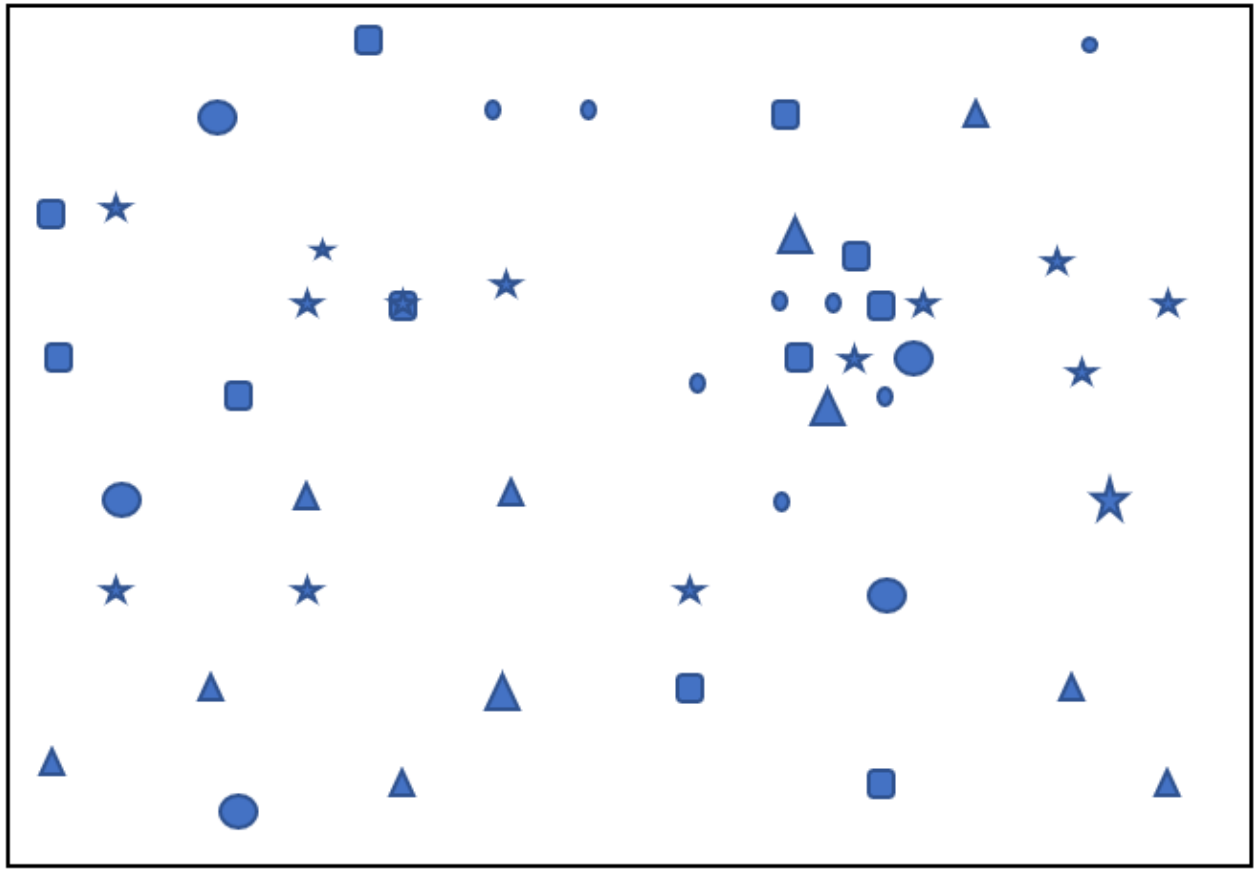
Faculty of Economic Sciences
University of Warsaw, Poland
kkopczewska@wne.uw.edu.pl

UNIVERSITY OF WARSAW | FACULTY OF ECONOMIC SCIENCES

We have dots…

/geo-referenced observations possibly with values attached/

What can we do with this information?

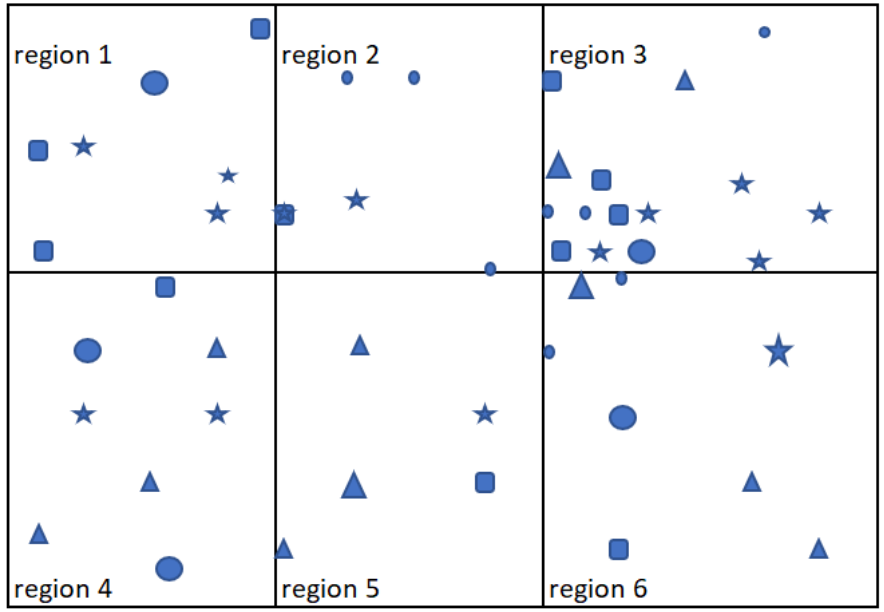


Statistics	Econometrics	Machine Learning
Spatial Statistics	Spatial Econometrics	Spatial Machine Learning

Neighbourhood, distance, relative and absolute location matter

This will be the story what to do with data that have values, location and time label.

➔ Spatial statistics



- Most „intuitive" is to aggregate dots to boxes

Great paper

Briant, A., Combes, P. P., & Lafourcade, M. (2010). Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations?. *Journal of Urban Economics*, *67*(3), 287-302.

- So, we get cross-table and we can apply all concentration /clustered-based/ measures

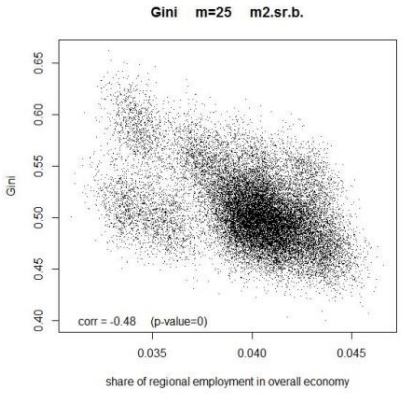
sectors	region 1	region 2	region 3	region 4	region 5	region 6	Total
▲	0	0	2	3	3	3	**11**
●	1	3	4	2	0	3	**13**
★	3	1	5	2	1	1	**13**
■	3	1	4	1	1	1	**11**
Total	**7**	**5**	**15**	**8**	**5**	**8**	**48**

Gini, Krugman dissimilarity, entropy, LQ &
Relative Specialisation, Hachman, Ogive, Refined Diversification,
Relative Diversity, Hallet, National Averages, Lillen,
Kullback-Leilbler divergence, Bruelhart-Traeger, Ellison-Glaeser,
Maurell-Sedillot, Geographic concentration and a few more

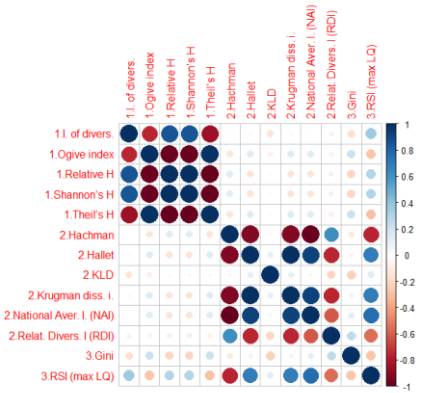
➔ Spatial statistics

- First, let's **identify** all cluster-based measures and see how to get them

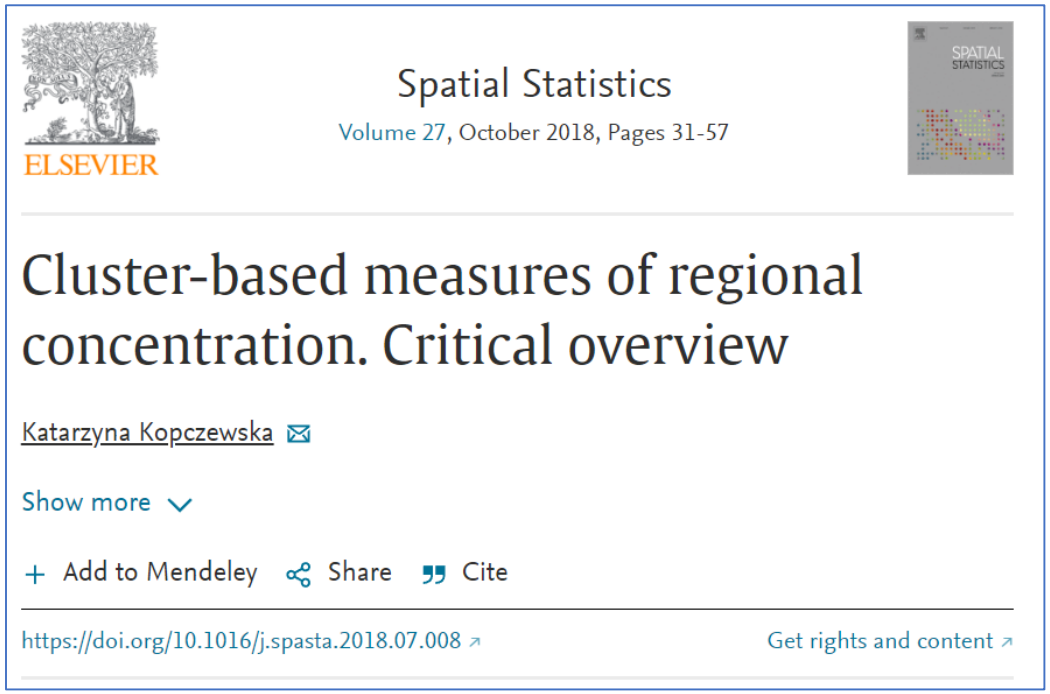
- Second, **any problems** in calculations?
  - MAUP / aggregation level
  - Can the same table bring different results?



Aggregation (x) matters for the results (y)



Measures depending on the design can bring the same or new info



https://www.sciencedirect.com/science/article/abs/pii/S2211675317302956
https://link.springer.com/book/10.1007/978-3-319-51505-2

➔ Spatial statistics

- Clustered-based measures **do not look inside** the table cells – spatial distribution can be any
- How to express **spatial agglomeration** with one number with regard to sector, size and location?

Papers in Regional Science
Regional Science Association International

FULL ARTICLE

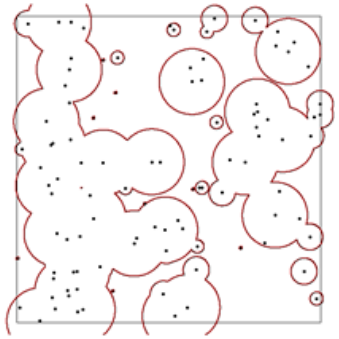
SPAG: Index of spatial agglomeration

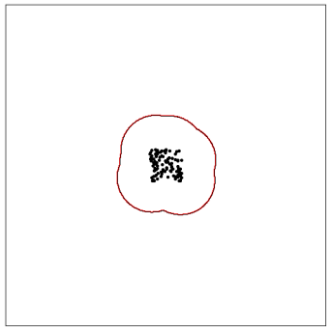
Katarzyna Kopczewska, Paweł Churski, Artur Ochojski, Adam Polko

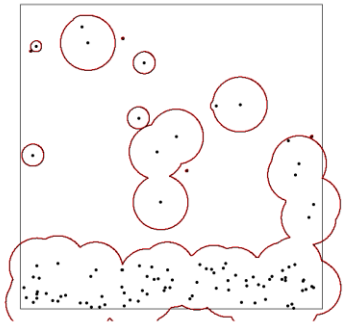
First published: 12 July 2019 | https://doi.org/10.1111/pirs.12470 | Citations: 3

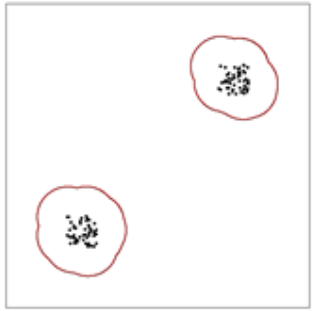
https://rsaiconnect.onlinelibrary.wiley.com/doi/abs/10.1111/pirs.12470

Random pattern	High agglomeraton	Costal & interrior locations	Two agglomerations
SPAG = 0.72	SPAG = 0.008	SPAG=0.41	SPAG=0.1



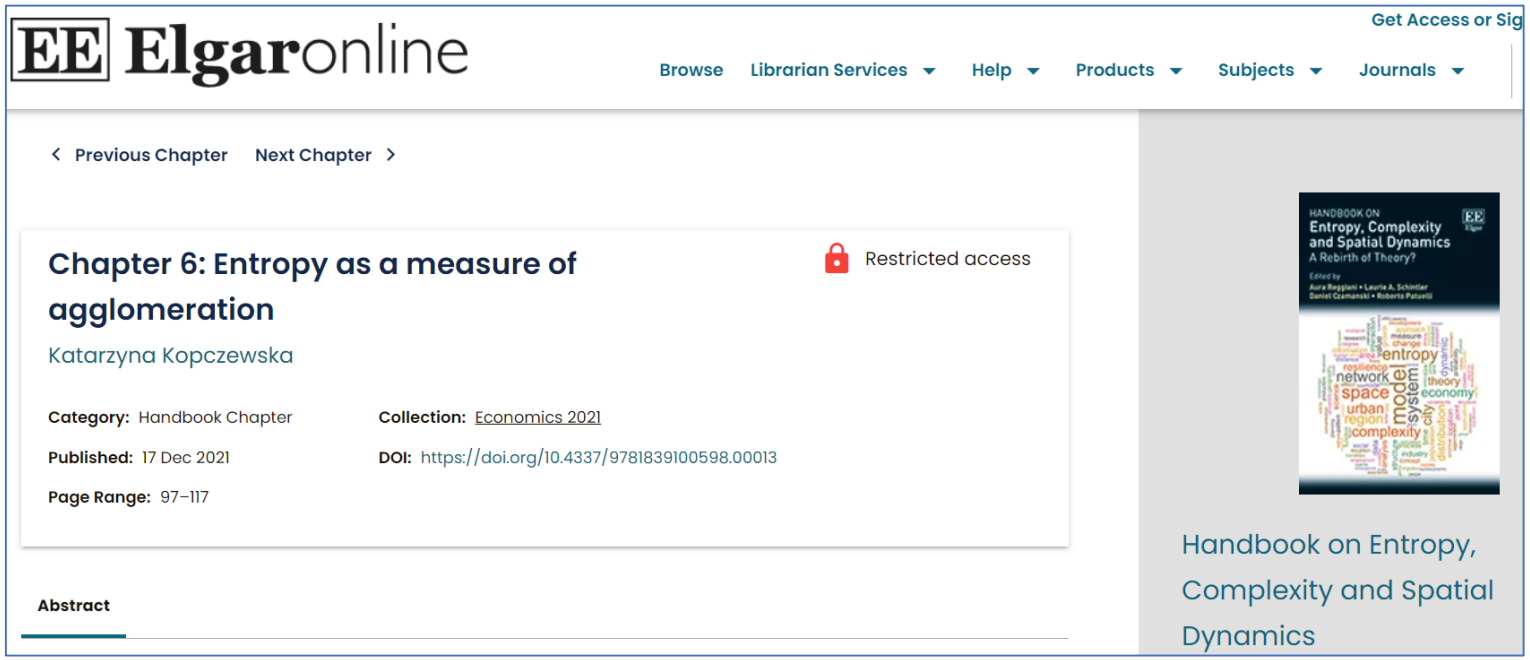


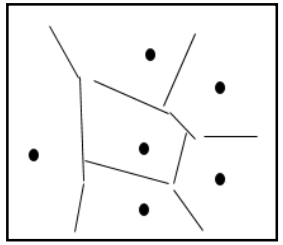




➔ Spatial statistics

- Alternative to SPAG measure of spatial agglomeration: using **entropy for tessellated** point pattern (Voronoi polygons)





+

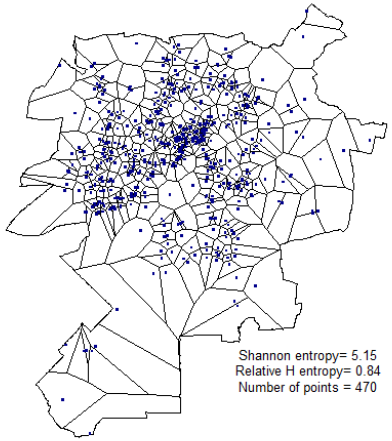
Shannon entropy: $H = -\sum_{n=1}^{N} s \cdot \log s$

Max entropy (uniform distr): $H_{max} = logn$

Relative entropy: $\mathrm{RelH} = \frac{H_{empir}}{H_{max}} = \frac{H_{empir}}{\ln n}$

=

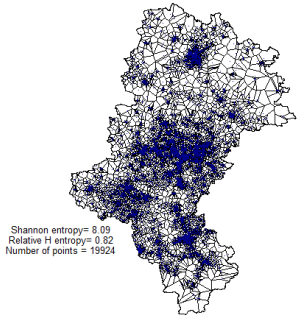
Tesselated business locations in Lublin city



Tesselated regular points

relH=1

Tesselated business locations in Silesia region



relH=0.82

https://www.elgaronline.com/display/edcoll/9781839100581/9781839100581.00013.xml
https://www.researchgate.net/publication/366020406_Entropy_as_a_measure_of_agglomeration

➔ Spatial statistics

- We know the uniform, random, agglomerated point patterns.
- Can point patterns be **natural and follow Benford's law**? Can we generate it?
- If yes, are **cities with their location and population** (3D) following Benford's law? What about Zipf law?

PLOS ONE

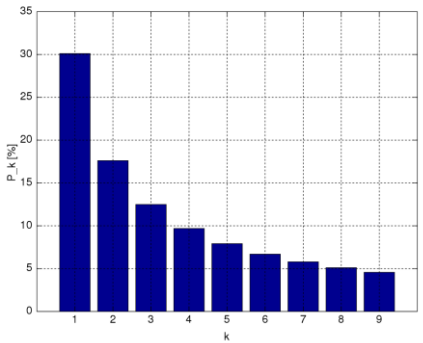
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

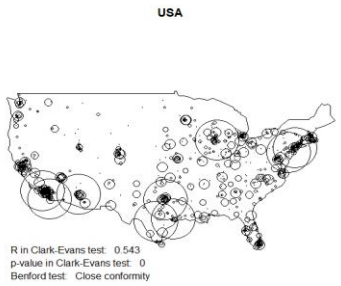
Natural spatial pattern—When mutual socio-geo distances between cities follow Benford's law

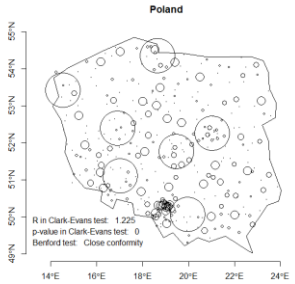
Katarzyna Kopczewska, Tomasz Kopczewski

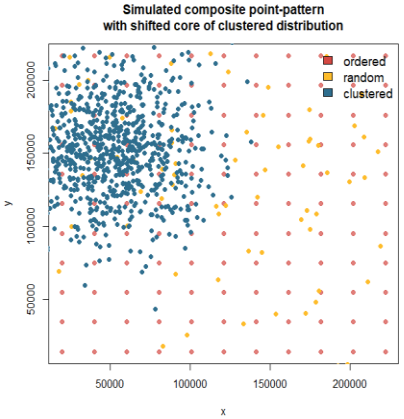
Published: October 20, 2022 • https://doi.org/10.1371/journal.pone.0276450

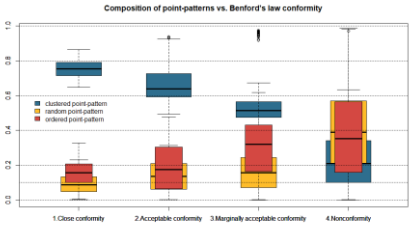


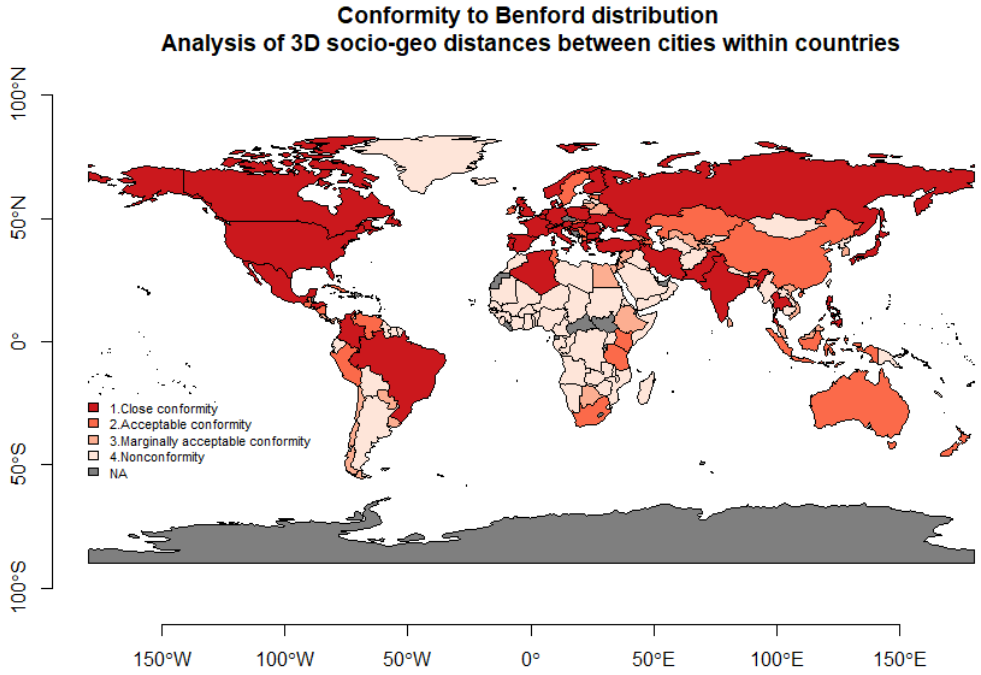
Benford's law











Criteria used: MAD Conformity - Nigrini (2012)

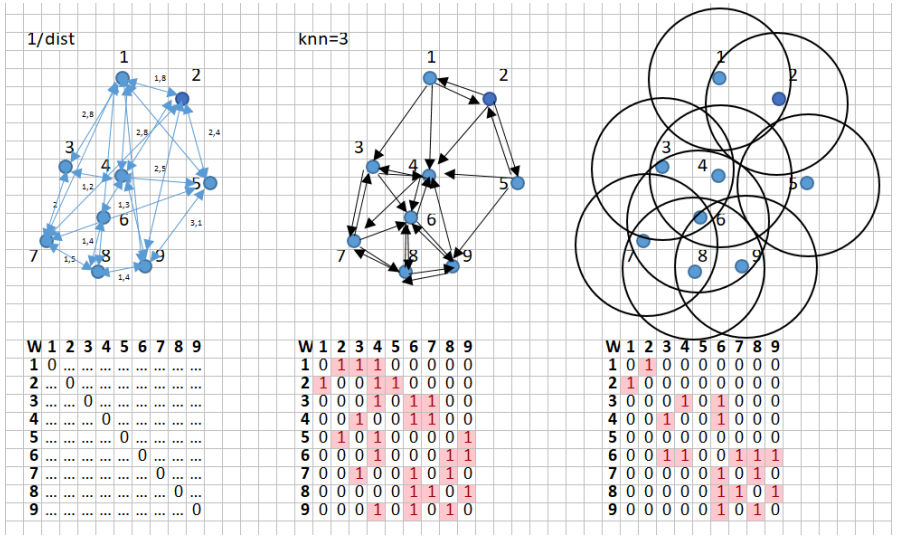
$$dist_{ij,3D} = \sqrt[2]{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

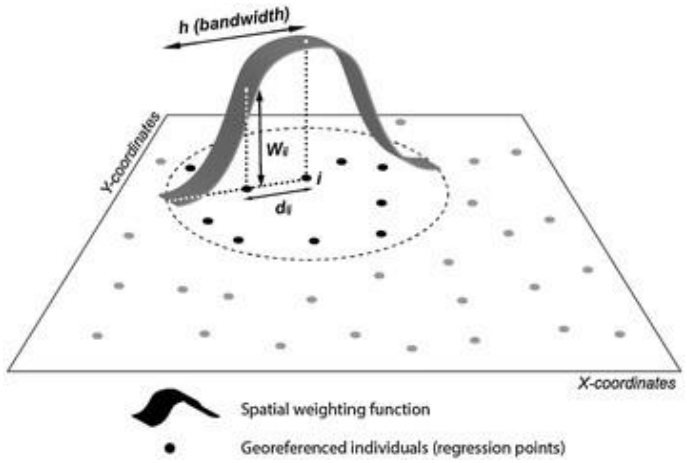
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0276450

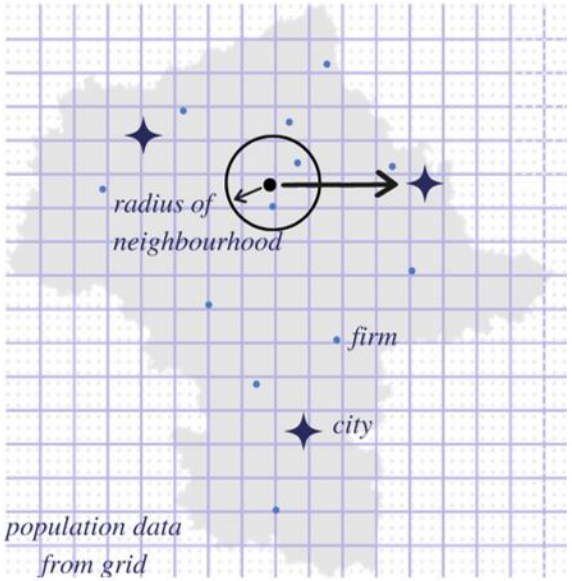
➔ ➔ Spatial econometrics

- How to incorporate info from a neighbourhood in econometric model on points:

Spatial econometrics with spatial weights matrix W	Geographically weighted regression GWR	Microgeography and agglomerations
Spatial lag for Y, X or e is the average value from the neighbourhood	Local regression on neighbours only; each observation gets its own coefficients	Making a radius around the point and counting what's inside







What can go wrong? ☺

➔ ➔ Spatial econometrics

- How to deal with big spatial data (for which W is unavailable)? → like 1 mln points (spatial econometrics works up to 100K points)
- How to predict out-of-sample for point data using spatial econometrics? → W keeps old locations, no possibility to input new locations

- We can bootstrap data and choose the best model from the candidates – estimate small model quickly
- We know well the parameters (mean is mean, sd is a function of n & sd)
- We can approximate exchange old point into new point – we use tesselation tiles to link them in pairs

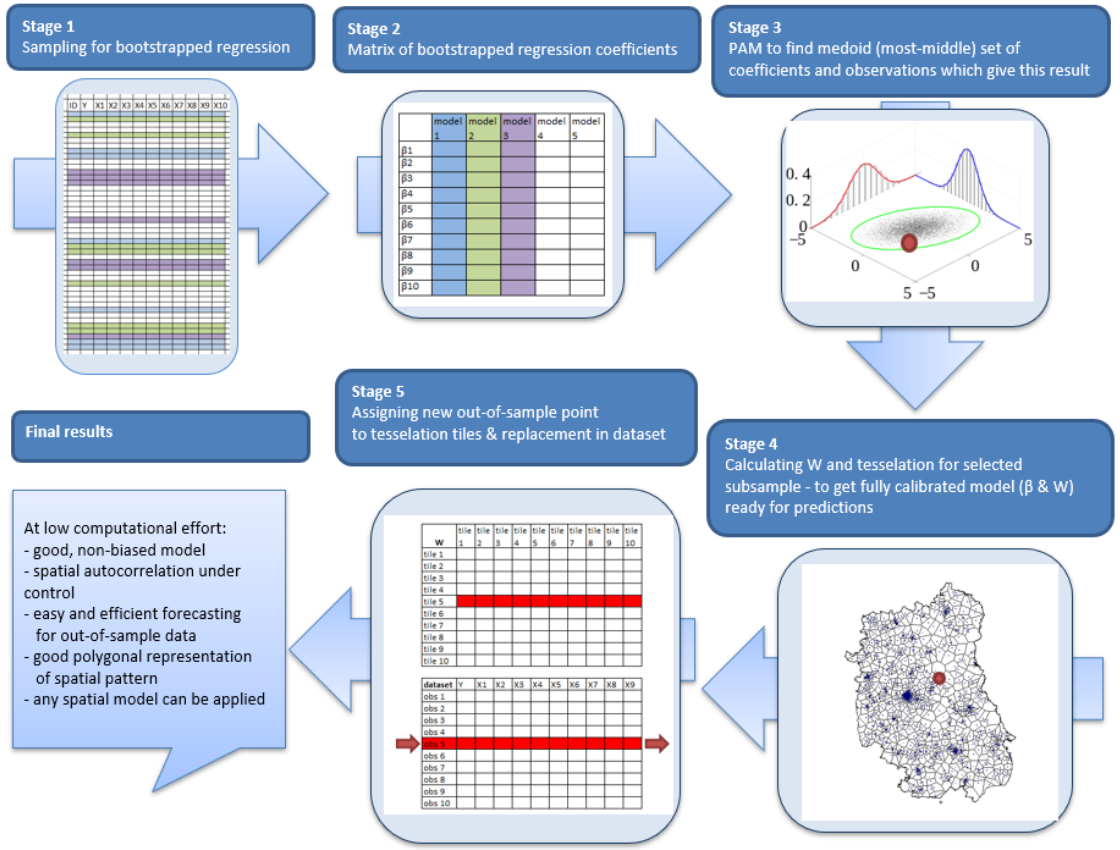
Scandinavian Journal of Statistics Theory and Applications

ORIGINAL ARTICLE | Full Access

Spatial bootstrapped microeconometrics: forecasting for out-of-sample geo-locations in big data

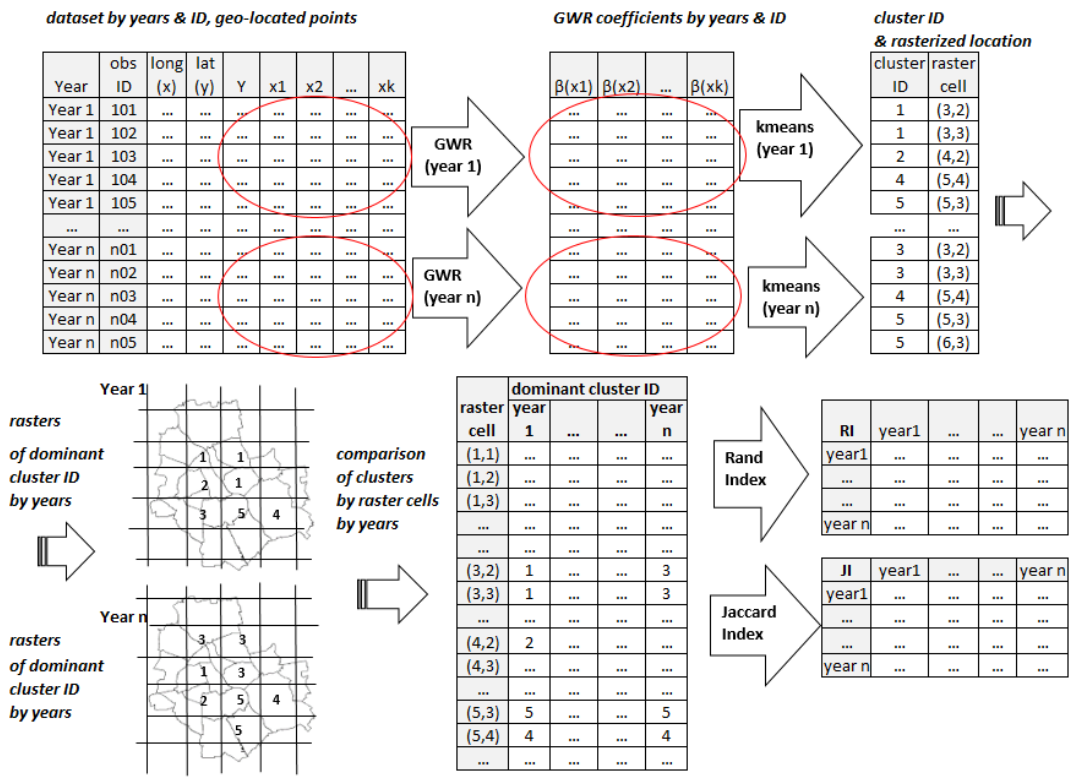
Katarzyna Kopczewska

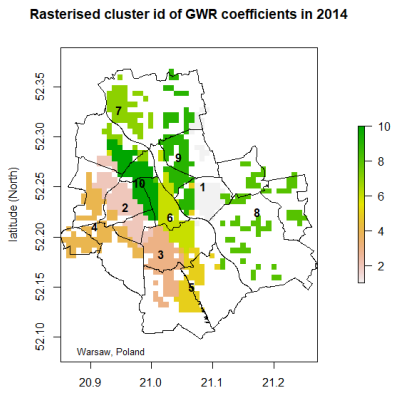
First published: 09 March 2023 | https://doi.org/10.1111/sjos.12636

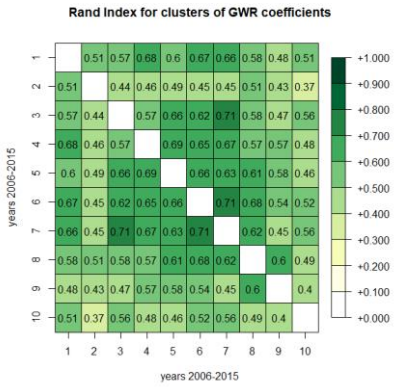


➔ ➔ Spatial econometrics

- For spatio-temporal data (e.g. housing transactions every year) we can get annual GWR hedonic models – are the valuations stable over time and space?
- We can cluster GWR coefficients to get submarkets – are they the same place every year? How long is model valid?







- **Cluster annual GWR coefficients** and **grid it** – then take cluster ID as representation for grid
- **Compare grids** if they belong to the same clusters over time – use **Rand index** (comparing pairs of pairs of points) – output is a stability index

https://www.sciencedirect.com/science/article/pii/S0264837721000156

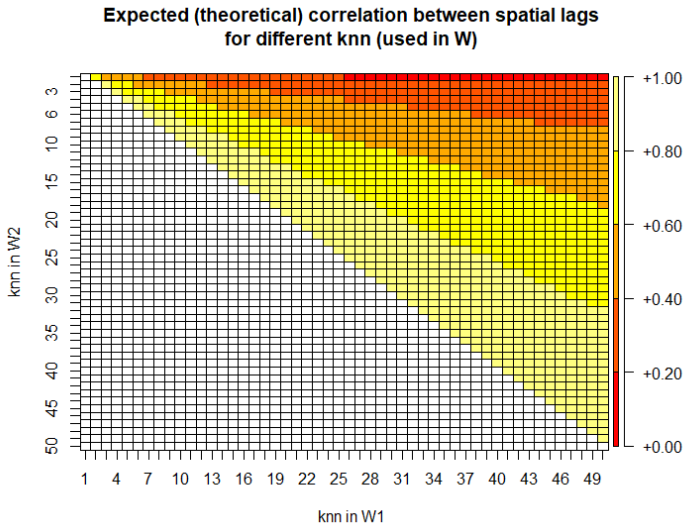
➔ ➔ Spatial econometrics

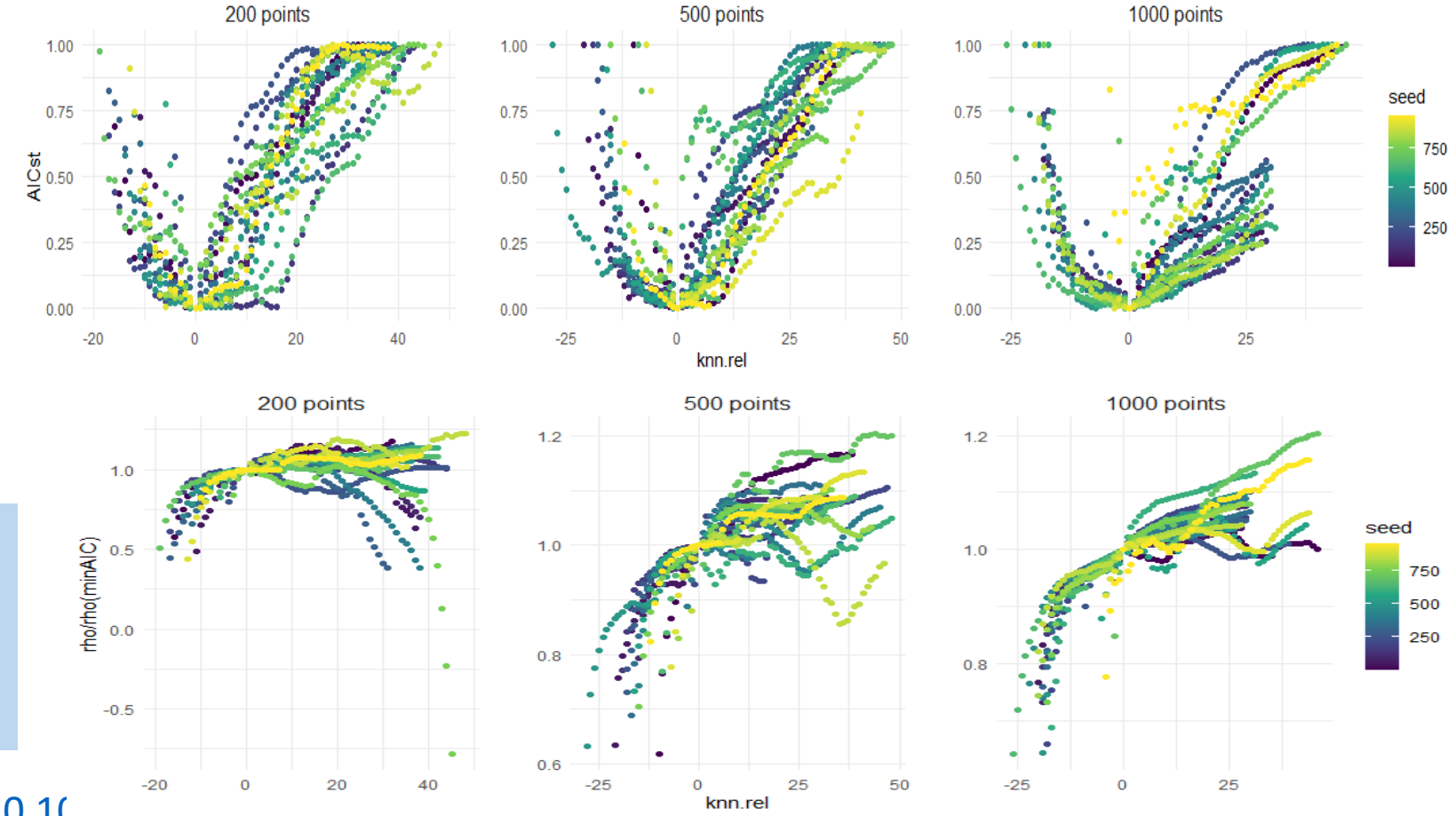
- How many k nearest neighbours (knn) should be included in W?
- Does knn matter for the result?

SPATIAL ECONOMIC ANALYSIS
https://doi.org/10.1080/17421772.2023.2176539
Routledge Taylor & Francis Group
RSA Regional Studies Association
OPEN ACCESS Check for updates

Akaike information criterion in choosing the optimal *k*-nearest neighbours of the spatial weight matrix

Maria Kubara and Katarzyna Kopczewska





- **Knn matters** – wrongly set generates significant **bias of coefficients**
- **Use AIC** – check candidate knn and choose the model with lowest AIC

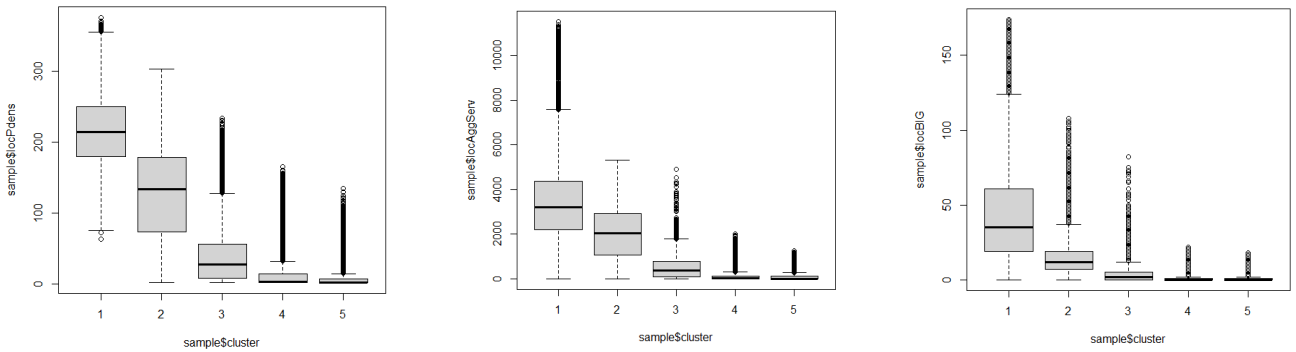
https://www.tandfonline.com/doi/full/10.1

➔ ➔ Spatial econometrics

- Does local density of points matters for their nature and relations they generate?
- How to deal with spatial heterogeneity of these data?

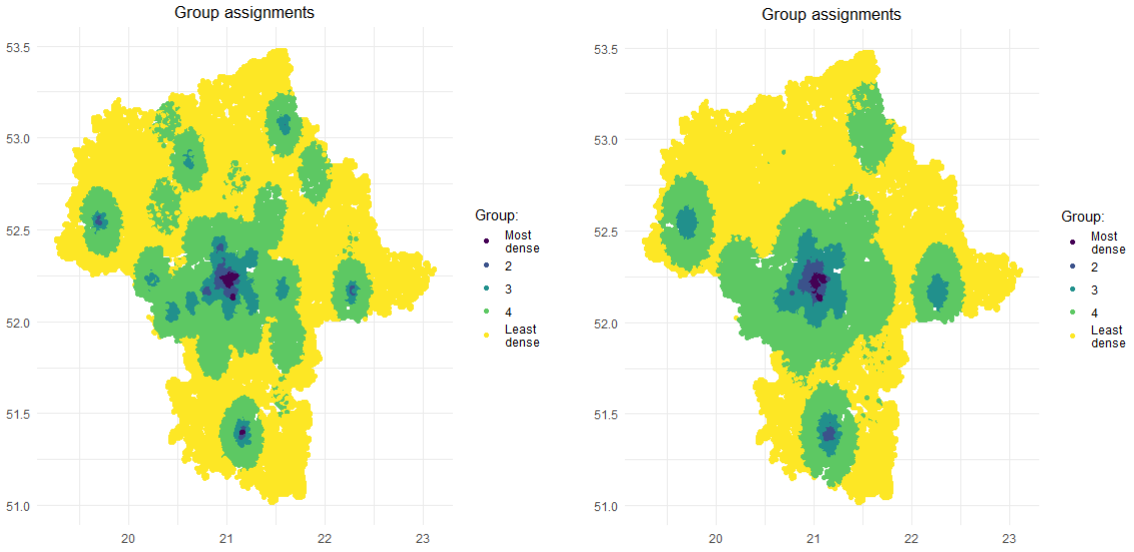
Spatial switching regimes regression: density-based cluster approach to deal with spatial heterogeneity

Katarzyna Kopczewska, Ewa Dobrowolska, Anil Bera – will be ready soon



Close neighbourhood of firms differs among density groups
➔ Economic relations differ too - in direction and strength!

- Divide data into DBSCAN density clusters (ML algorithm)
- Estimate models in subgroups and compare them
- Information from neighbourhood calculated in radius from point (e.g number of firms from given sector, number of inhabitants)



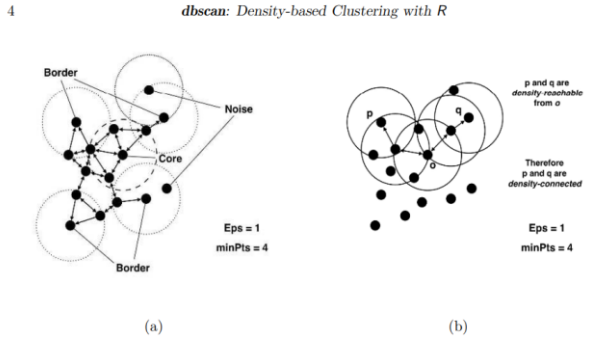


Figure 1: Concepts used the DBSCAN family of algorithms. (a) shows examples for the three point classes, core, border, and noise points, (b) illustrates the concept of density-reachability and density-connectivity.

DBSCAN checks if numer of points within the radius exceeds a threshold – if yes, it is high density cluster

➔ ➔ ➔ Spatial machine learning

- Novel solution with huge potential
- Challenge: how to include spatial component?
- What is done and what not?
- New perspectives for regional and urban studies

Home › The Annals of Regional Science › Article

Original Paper | Open Access | Published: 24 December 2021

Spatial machine learning: new opportunities for regional science

Katarzyna Kopczewska

The Annals of Regional Science **68**, 713–755 (2022) | Cite this article

10k Accesses | **18** Citations | **32** Altmetric | Metrics

Why use ML?

- Why not? ☺
- Dealing with non-linearity, partial impact
- Finding new patterns, also spatio-temporal
- Higher computational efficiency
- New types of data possible to include (photos, pixels, etc)
- Better possibilities to forecast

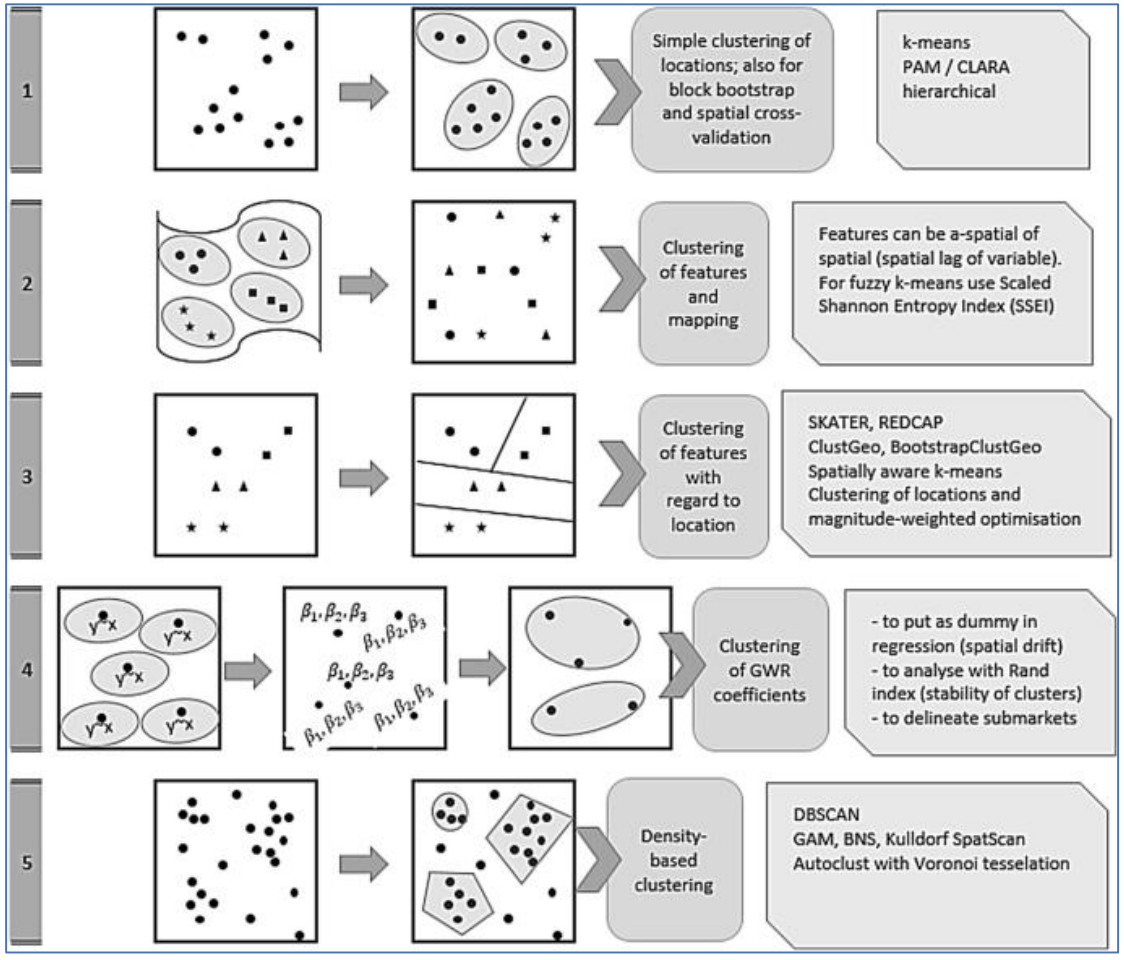
Unsupervised learning

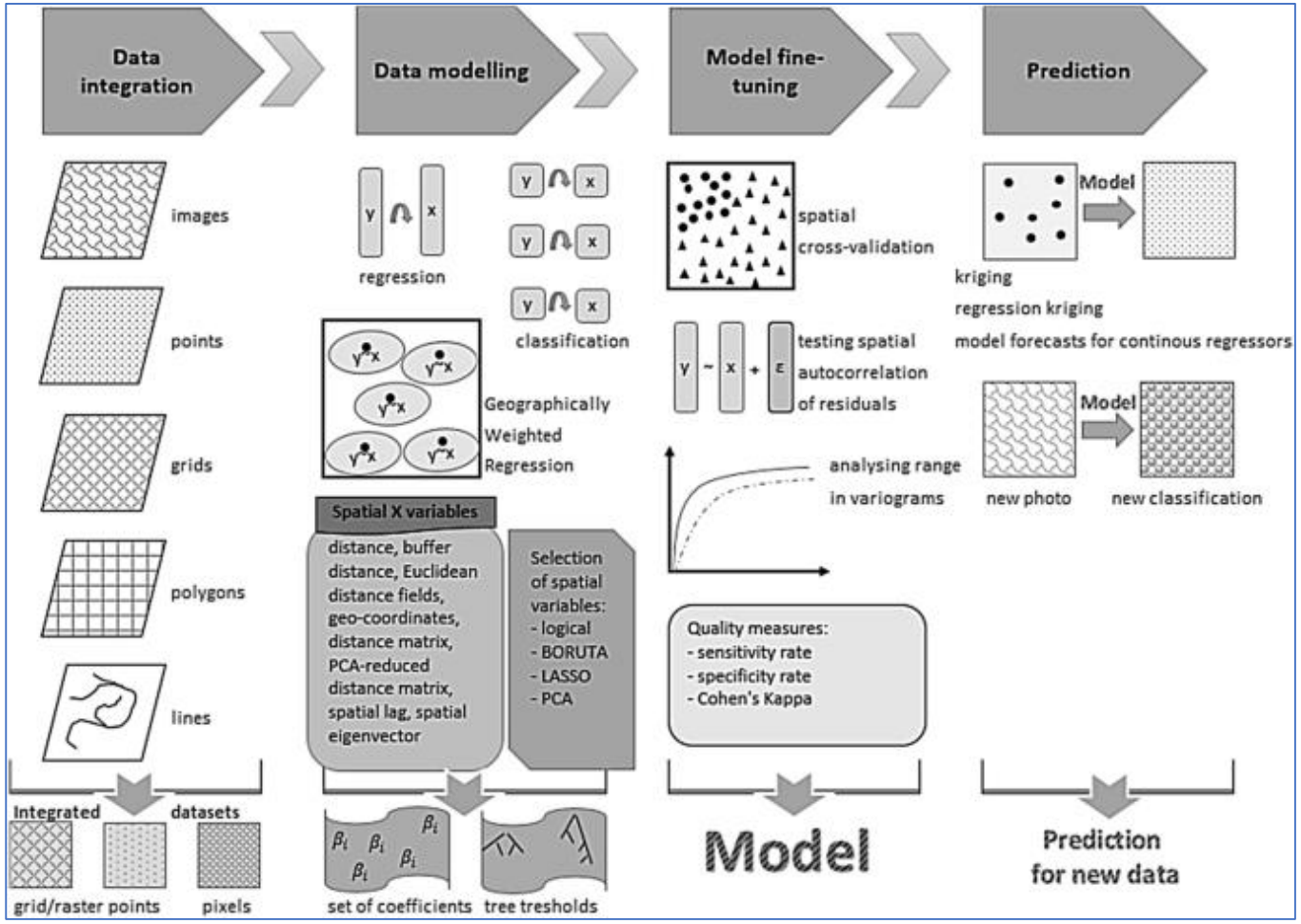
Clustering:
- locations
- features and mapping
- features in space
- GWR coefficients
- density-based

Supervised learning

Regression models:
- Simple regressions
- Spatial cross-validation
- Image recognition in spatial classification
- Mixtures of GWR and ML
- Spatial variables

Unsupervised and supervised spatial machine learning





➔ ➔ ➔ Spatial Machine Learning

Perspectives:

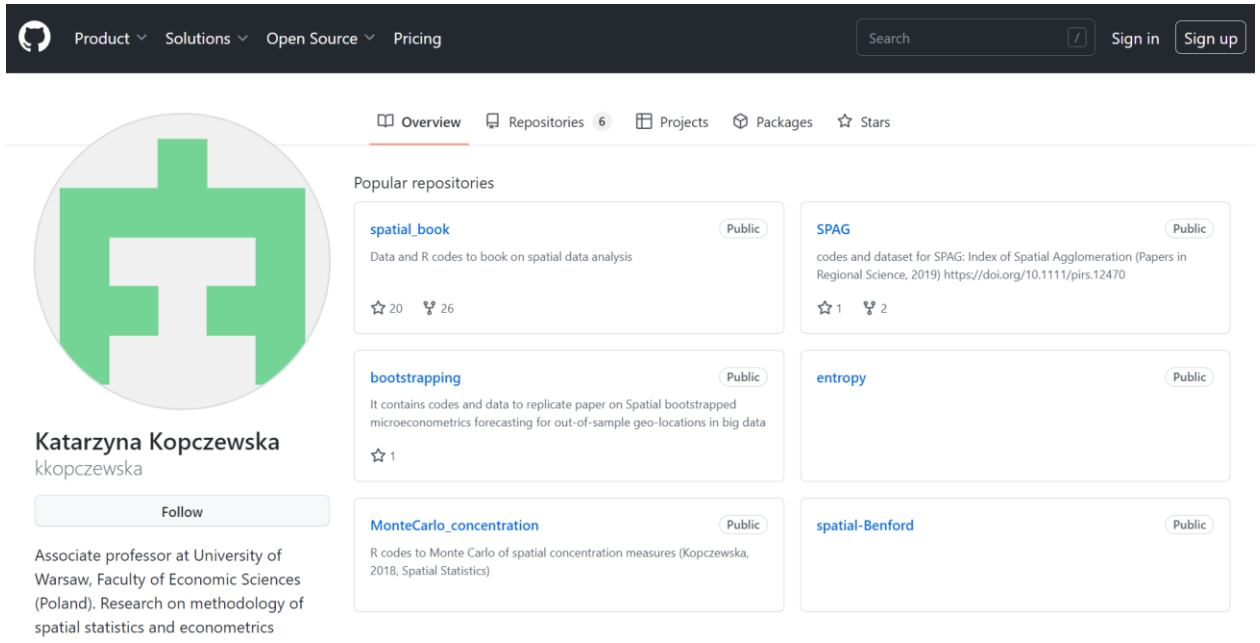
- Big data analytics (to help spatial econometrics)
- Using new sources of data – nightlights, day photo of landscape, spectral data, vegetation indicators
- Dealing with spatial heterogeneity and isotropy
- Spatio-temporal modeling with many layers (also of different granulation)
- Better forecasting (new approach - „econometric" models not only to explain)

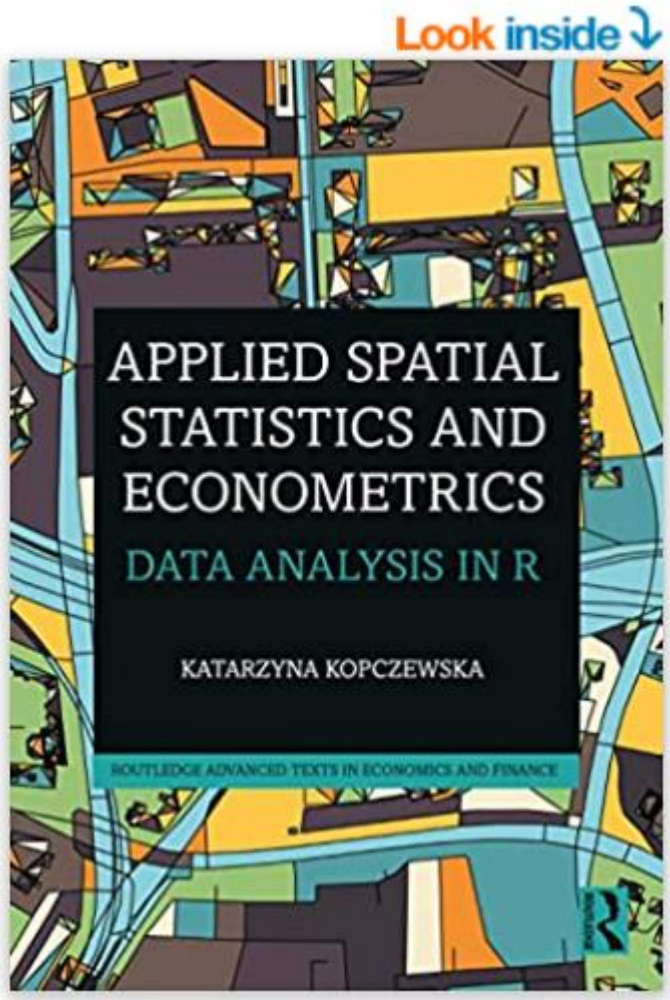
What will come soon?

- Spatial 3D solutions
- Like in life sciences: standards of reporting, computational reproducibility, workflow managers (ready-to-use environments)
- Communication with those who believe in p-value ☺

➔ ➔ ➔ ➔ R implementations

- Toolbox to run the analysis is finally the core of our research
- R still wins with Python – especially since C++ implementations are available
- Making codes available at Github makes the science really open





Applied Spatial Statistics and Econometrics: Data Analysis in R (Routledge Advanced Texts in Economics and Finance) 1st Edition

https://www.routledge.com/Applied-Spatial-Statistics-and-Econometrics-Data-Analysis-in-R/Kopczewska/p/book/9780367470760
https://github.com/kkopczewska

Thank you!

Katarzyna Kopczewska

Faculty of Economic Sciences
University of Warsaw, Poland

kkopczewska@wne.uw.edu.pl

@KathyKopczewska
@SpatialWarsaw