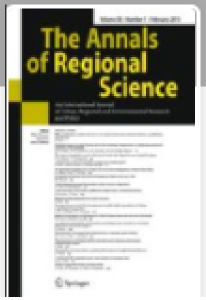
Spatial Machine Learning
collection of ideas
inspiration for new research

KATARZYNA KOPCZEWSKA

FACULTY OF ECONOMIC SCIENCES, UNIVERSITY OF WARSAW, POLAND

KKOPCZEWSKA@WNE.UW.EDU.PL



The Annals of Regional Science
An International Journal of Urban, Regional and Environmental Research and Policy

The Annals of Regional Science
https://doi.org/10.1007/s00168-021-01101-x

ORIGINAL PAPER

Check for updates

Spatial machine learning: new opportunities for regional science

Katarzyna Kopczewska[1]

Received: 30 May 2021 / Accepted: 8 December 2021

Abstract
This paper is a methodological guide to using machine learning in the spatial context. It provides an overview of the existing spatial toolbox proposed in the literature: unsupervised learning, which deals with clustering of spatial data, and supervised learning, which displaces classical spatial econometrics. It shows the potential of using this developing methodology, as well as its pitfalls. It catalogues and comments on the usage of spatial clustering methods (for locations and values, both separately and jointly) for mapping, bootstrapping, cross-validation, GWR modelling and density indicators. It provides details of spatial machine learning models, which are combined with spatial data integration, modelling, model fine-tuning and predictions to deal with spatial autocorrelation and big data. The paper delineates "already available" and "forthcoming" methods and gives inspiration for transplanting modern quantitative methods from other thematic areas to research in regional science.

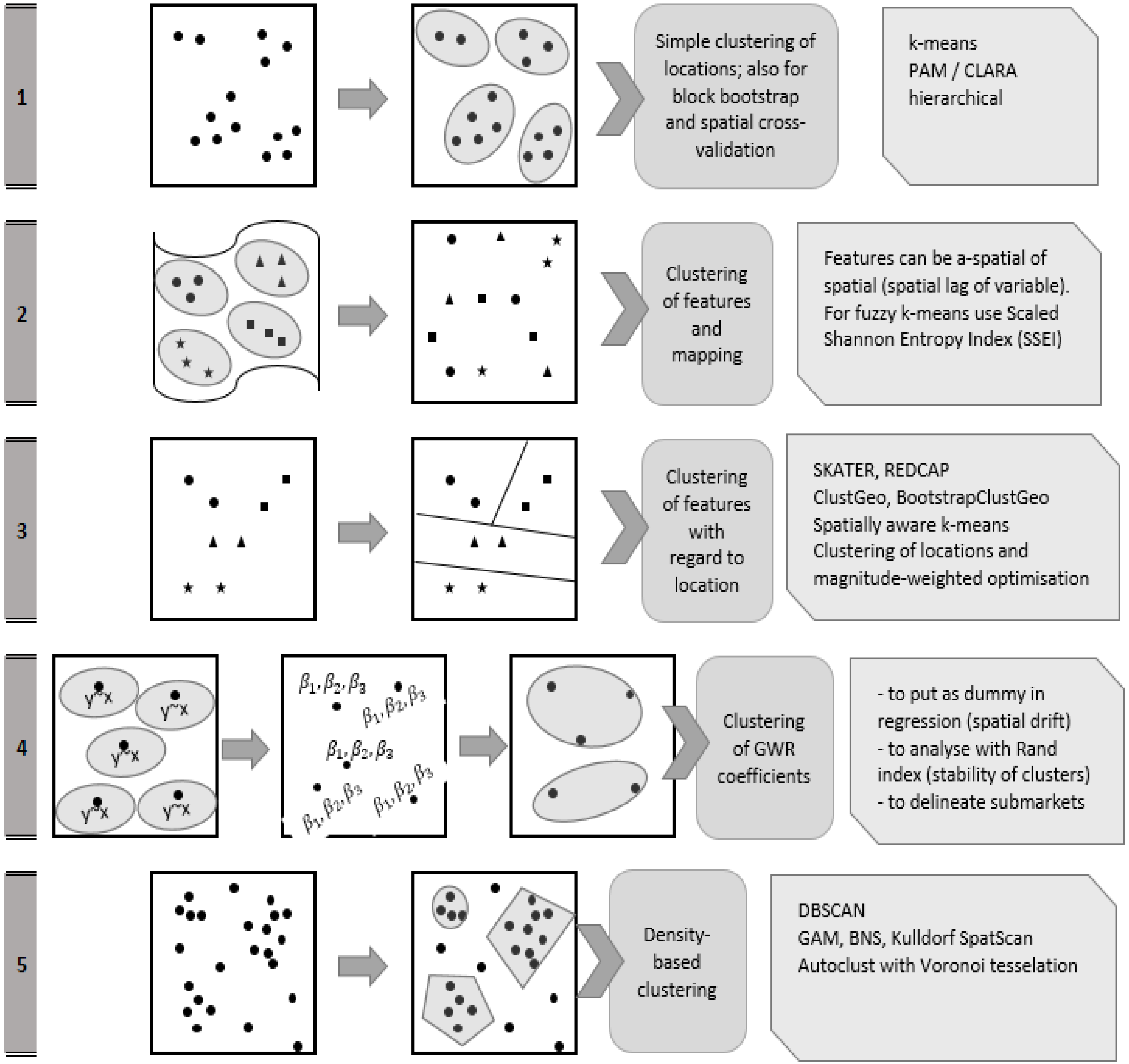
https://doi.org/10.1007/s00168-021-01101-x

JEL Classification C31 · R10 · C49

About this presentation and its motivation

- When working on geo-located point data, **spatial econometrics applications are limited due to scalability for big** (W becomes too big) data and **predictions for out-of-sample for given W**
- One can also use **GWR**, but it also has computational **problems in optimising bandwith** in big data
- As the new types of geo-located data are appearing (as sattelite data), **one needs new methods** to address those issues
- When scanning literature of last years, **there appear many concepts** (better or worse) on SPATIAL MACHINE LEARNING

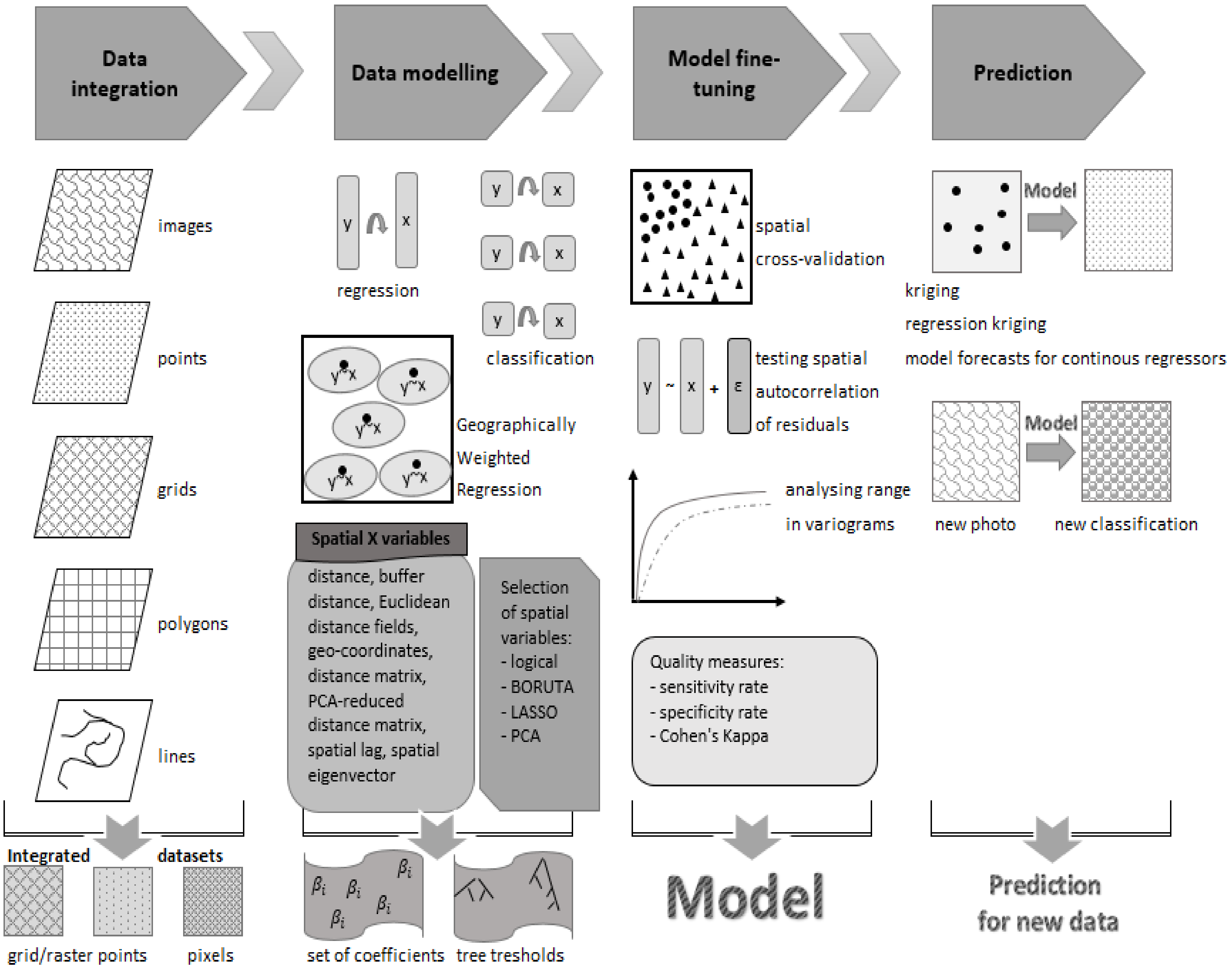
→THIS PAPER IS ITS OVERVIEW WITH CRITICAL ASSESSMENT WHAT AND HOW WAS DONE!



Spatial unsupervised learning

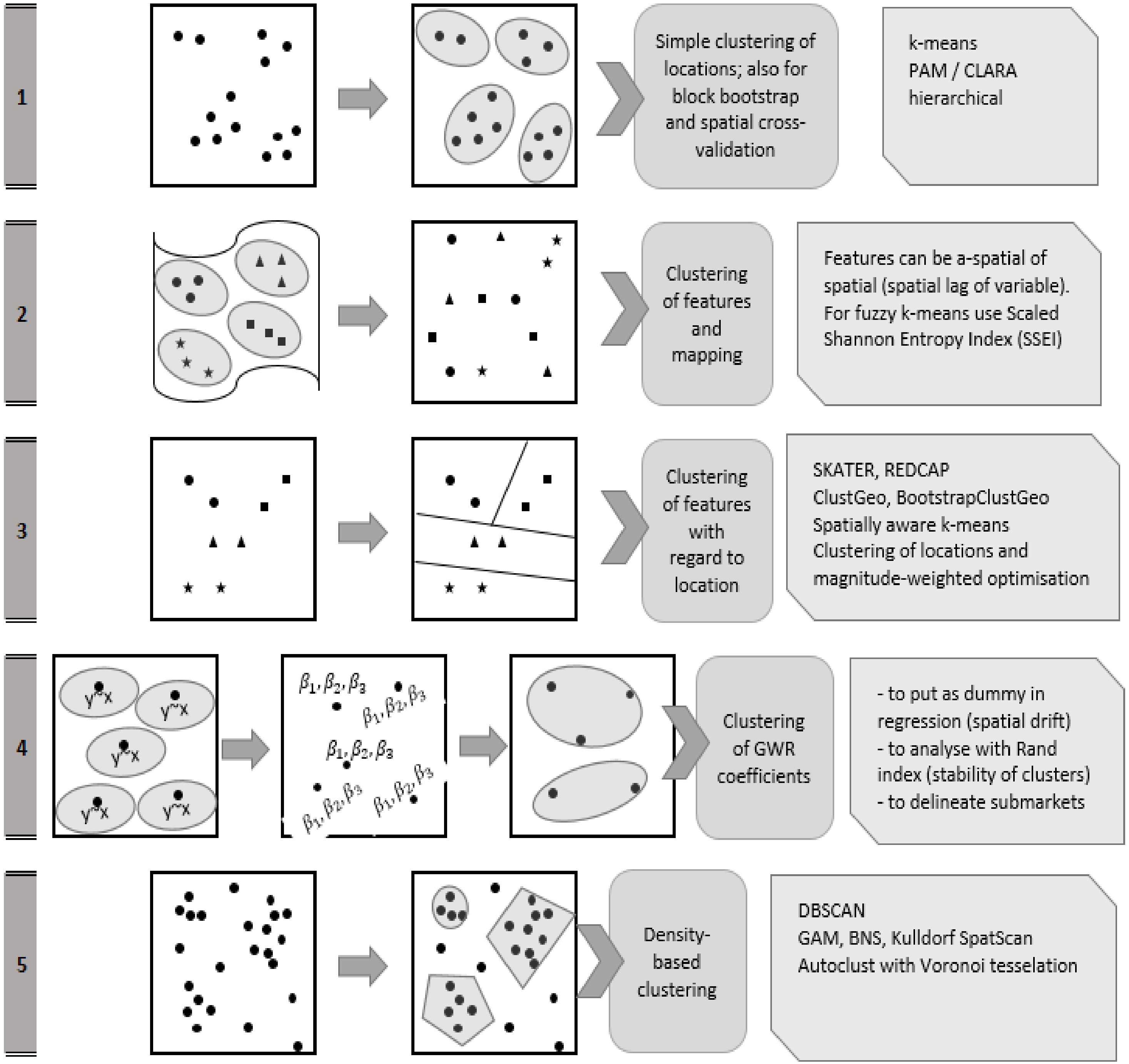
how to cluster geo-located data

Spatial supervised learning

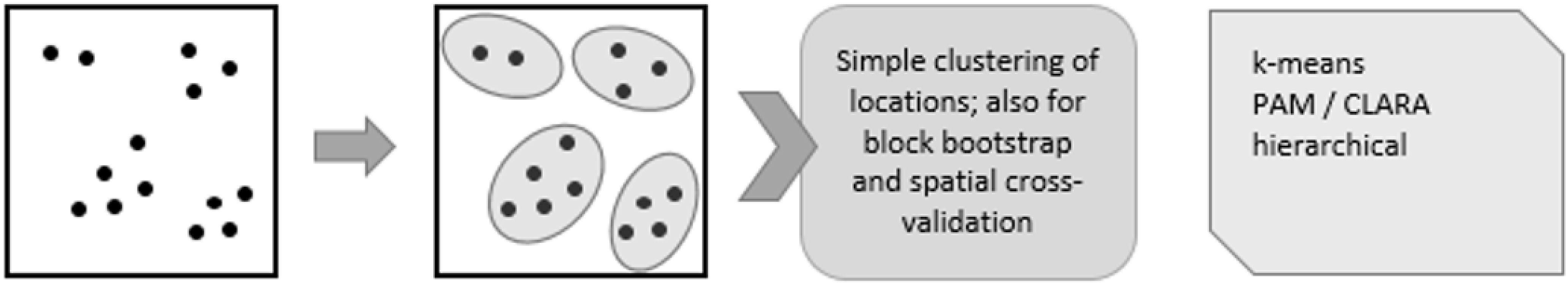


Spatial unsupervised learning

how to cluster geo-located data



1. Clustering of locations



- One uses k-means or Partitioning Around Medoids (PAM) to cluster geo-coordinates which build spatially continous groups.
- In empirical studies one can delineate catchment areas or establish territory for sales representatives
- In other applications, *k-means* helps to build **irregular non-overlapping spatial clusters** to **run spatially stratified sampling from those clusters** (e.g. Russ & Brenning, 2010; Schratz et al., 2019). This **solves the problem of inconsistency in bootstrapping** (Chernick & LaBudde, 2014; Kraamwinkel et al., 2018) and **addresses the autocorrelation in cross-validation** (discussed further).
- *K-means* irregular partitioning can also be applied in the **block bootstrap** (Hall et al., 1995; Liu & Singh, 1992). Sampling blocks of data from spatially pre-defined subsamples allows for drawing independent blocks of data but lowers the computational efficiency.

2. Clustering of features and mapping

- Bajocco et al. (2015) use **hierarchical clustering** (with dendrogram) to analyse fire distribution in Sardinia. It clusters phenological metrics and vegetated land surface of each territorial unit. It grouped the territorial units into similarly covered areas. For each cluster, one checks the fire frequency to assess the natural conditions that increase and decrease fire-proneness.

- Liu et al. (2018) run non-spatial ***k*-means** clustering to detect urban sprawl. They run a-spatial partitioning of local **spatial entropy** H calculated for a gridded population. Local spatial entropy is expressed as $H = \sum_i p_i \ln(p_i)$, where p_i is the relative population in the analysed cell and eight neighbouring grid cells and $\sum_{i=1}^{i=9} p_i = 1$. Clustering of entropy, when mapped, may delineate areas with high and low local density.

- Hengl et al. (2017) mapped soil nutrients in Africa (numer of clusters found with **hierarchical clustering** for parameterised Gaussian mixture models and BIC. They use **fuzzy k-means** and map it **Scaled Shannon Entropy Index**.

- Clusters are not always derived with a partitioning procedure – for local obesity in Switzerland Joost et al. (2019) mapped the local Getis-Ord Gi statistics for body mass index (BMI) and sugar-sweetened beverages intake frequency and concluded “**optically**” from visualisation about spatial agglomeration of high and low values of Gi.

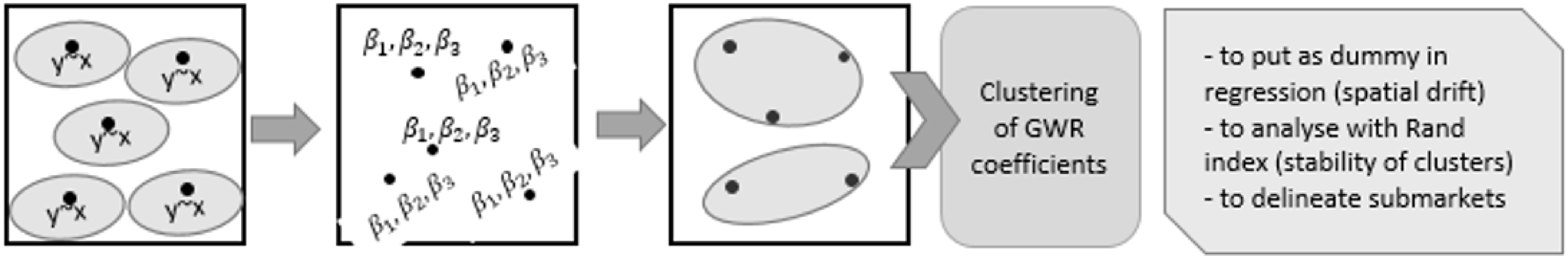
3. Clustering of features with regard to location

- **Dedicated algorithms**: SKATER (Assunção et al., 2006), REDCAP (Guo, 2008), ClustGeo (Chavent et al. 2018), Bootstrap ClustGeo (DiStefano et al., 2020)

- Combinations with *k*-means, applied to **seismic analysis** of the Aegean region (Weatherill & Burton, 2009) to capture location of earthquakes and their magnitude. ***k*-means** minimizes the total within-cluster sum of squares (difference between individual and group values) – trick is that cluster average distance was replaced by a **magnitude-weighted average distance**, which shifts the centroids of a cluster into the strongest earthquakes.

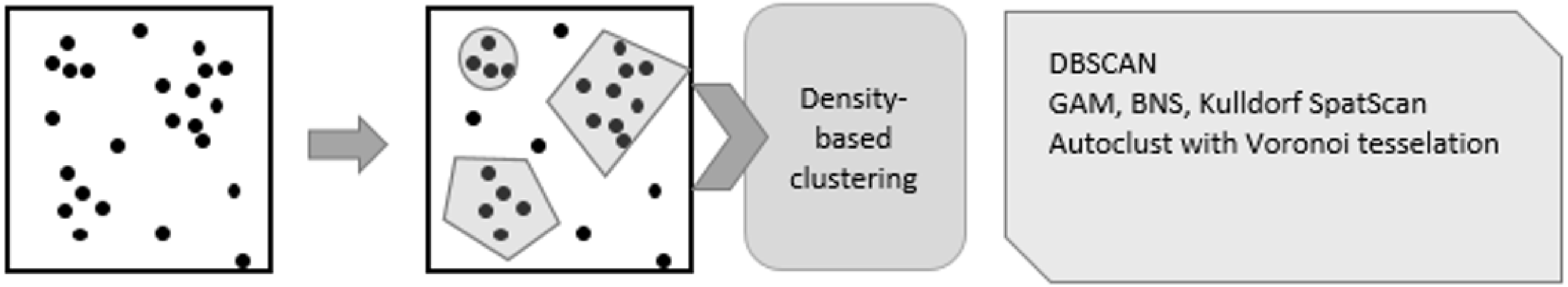
- Spatially-oriented *k*-means appears also in biostatistics. In mass spectrometry brain analysis with pixel data, spectra is like time-series. Alexandrov and Kobarg (2011) proposed a **spatially-aware *k-means* clustering**. Dissimilarity (distance) matrix between pixels is compared as a composite distance between their spectra, weighted with neighbouring spectra in radius *r*, similarly to the spatial lag concept. Even if *k-means* clustering itself has no spatial component, the distances used in clustering include neighbourhood structure.

4. Clustering of GWR coefficients



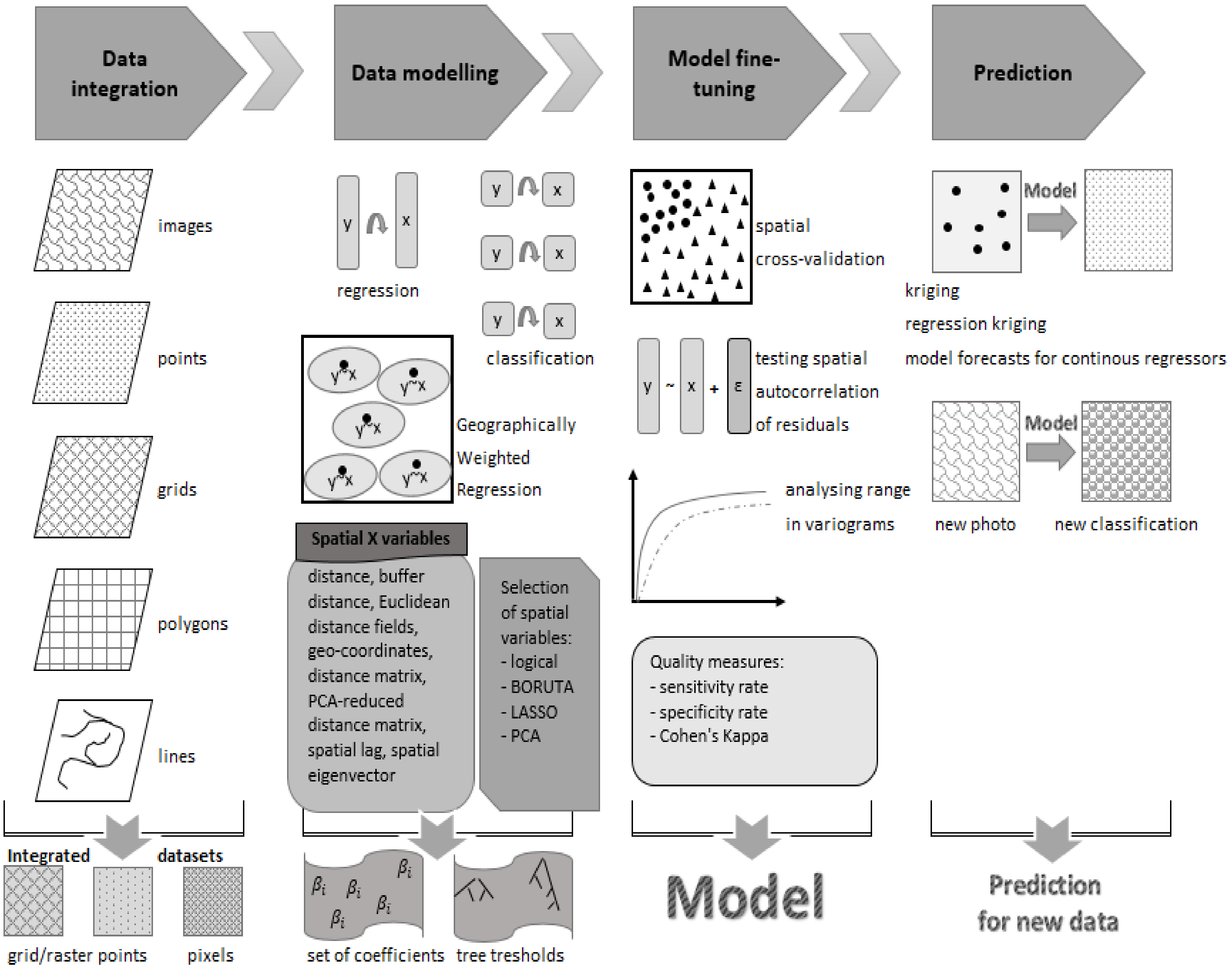
- **Clusters** of GWR individual coefficients are usually **continuous over space**,
- GWR coefficients can be used in **profiling the locations** assigned to different clusters – as obesity map (Chi et al., 2013)
- GWR coefficients can **detect spatial drift** (Müller et al., 2013) in model for public transport services - one runs two models: firstly, GWR, for which coefficiens are clustered; secondly global model which includes dummry variables on belonging to GWR clusters. It addresses heterogeneity and autocorrelation.
- GWR can **detect spatio-temporal stability** (Kopczewska & Ćwiakowski, 2021). For each period one estimates and clusters GWR coefficients. Next, they are rasterised to get median value of cluster ID in cell. Finally, one applies the Rand Index or Jaccard similarity to test the temporal similarity of clusters.
- GWR can be in **temporal version**, mostly for housing: Soltani et al. (2021) cluster with SKATER the GTWR (Geographically and Temporally Weighted Regression). Helbich et al. (2013) used mixed GWR (MGWR)

5. Density-based clustering



- The most popular is **DBSCAN** – nice application cover the retail spatial extent (Pavlis et al., 2017), tropical cyclone risk (Cai et al., 2020), in astronomy, e.g. to test the spatial distribution of Taurus stars (Joncour et al., 2018), in imaging with an airborne LIDAR technique (Wang et al., 2019), WLAN Indoor Positioning Accuracy (Wang et al., 2019) and traffic collision risk in maritime transportation (Liu et al., 2020), text data and computer codes (Mustakim et al., 2019; Reis & Costa, 2015).
- **Before DBSCAN**, there were a few other concepts of scanning statistics based on a moving circle - GAM (*Geographical Analysis Machine*), BNS (*Besag-Newell Statistic*) and Kulldorf's (1997) spatial scan statistics.
- **After DBSCAN**, there appeared a group of methods based on Voronoi / Dirichlet tessellation (Estivill-Castro & Lee, 2002; Lui et al., 2008), called Autoclust. Recently, as a rebirth, one can find proposals of its 3D implementations (Kim & Cho, 2019).

Spatial supervised learning



Supervised learning

There are two general groups of ML models:

a) **typical regressions**, which link the levels of features of variables *x* and *y*;

b) **classifiers**, which detect feature levels *x* in observed classes *y*.

In supervised learning we know the class/group, in unsupervised learning we guess it.

Typical spatial classification problem / solution is as follows:

1) from an image (e.g. pixels of a satellite photo) one extracts features of the land (e.g. vegetation index, water index, land coverage)

2) one adds geographical information (e.g. location coordinates)

3) one knows the real classification (e.g. type of crops), which is to be later forecasted with the model

4) It is to teach an algorithm to distinguish the desired image elements by linking information from the photo with the real class, where an image pixel is an individual observation.

5) next, the model can detect those elements on new photos to predict the class.

Common methods:
- Naive Bayes
- k-Nearest Neighbours
- Random Forests
- Support Vector Machines
- Artificial Neural Networks
- XGBoost
- Cubist

\+ ensemble (mixture) of them

Applications:

- in agriculture to distinguish crops, landscaping and land use (Pena & Brenning, 2015)
- geological mapping (e.g. Cracknell & Reading, 2014)
- regional socio-economic development indicators based on night-light data or land use satellite images (e.g. Cecchini et al., 2021).

Simple regression models to answer spatial questions

- Appelhans et al. (2015) explain **temperatures on Kilimanjaro** with elevation, hill slope, aspects, sly-view factor and vegetation index – they use machine learning models in a regression, and the only spatial issue is spatial interpolation with kriging.

- Liu et al. (2020) run non-spatial random forest model on socio-economic and environmental variables to **explain poverty in Yunyang, China,** using data from 348 villages. The only computational spatial component is the Moran test of residuals, which evidenced no spatial autocorrelation. The power of the study lies in merging different sources of geo-projected data: surface data for elevation, slope, land cover types and natural disasters (with resolution 30m or 1:2000); point data like access to town, market, hospital, bank, school, or industry taken from POI (Point-Of-Interest) or road density network (in scale 1:120000); and polygonal data for the labour force from a statistical office.

- Rodríguez-Pérez et al. (2020) model the **lightning-caused fire** in geo-located grid cells in Spain. They use RF, generalised additive model (GAM) and spatial models, where the fact that lightning-caused fire appeared in a given grid-cell was explained with features observed there such as vegetation type and structure, terrain, climate, and lightning characteristics.

- Gerassis et al. (2020) **map rural workers' health condition** and severe disease exposure using a ML approach. They use Bayesian Network to detect factors of good/bad health. Spatial methods appear only for interpolation of illness cases observed, which is a separate model (Point-to-Area Poisson kriging model, which deals with Spatial Count Data, unequal territories and diverse population composition). The spatial challenge was in different granulation of data: point data in the study sample and polygonal data as a basis of prediction.

Spatial cross-validation

- In spatial CV – one divides points into *k* irregular clusters (by using, e.g. *k*-means) and selects one cluster as an out-of-sample cross-validation part. Due to spatial autocorrelation between training and testing observations, simple spatial data sampling gives biased and over-optimistic predictions.
- Spatial CV increases prediction error (Liu, 2020).
- Lovelace et al. (2019) show that a spatially cross-validated model gives a lower AUROC (*Area Under the Receiver Operator Characteristic Curve*), as it is not biased with spatial autocorrelation.
- **Spatial cross-validation is becoming a standard** (e.g. Goetz et al., 2015; Meyer et al., 2019), but some studies still neglect this effect and do not address the autocorrelation problem (Park & Bae, 2015; Xu & Li, 2020).

Image recognition in spatial classification tasks

- Good example of geological mapping - **supervised lithology classification** (Cracknell & Reading, 2014).
  - As input (*X*) use the airborne geophysics and multispectral satellite data,
  - As output (*Y*) for a given territory, they use the known lithology classification, given as polygons on the image for each class.
  - They also know the *xy* coordinates of the pixels of those images.
  - They produce an algorithm which discovers the lithology classification from airborne geophysics and multispectral satellites.
  - They run three kinds of models on pixel data: i) *X*→*Y*, ii) *xy coords* → *Y*, iii) *X* & *xy coords* → *Y*.
  - This is to teach software to understand what is in the picture and give a lithology class to each pixel.

- Other example – Nicolis et al. (2020) model **dynamic statistics of earthquakes** in Chile. Their dataset of seismic events included a period of 17 years, with 86000 geo-located cases in 6575 days. For each day with an earthquake, they created a grid-based image (1°x1°) of the territory with grid-intensity estimated by an ETAS *(Epidemic-Type Aftershock Sequences*) model. Using this, they applied deep learning methods such as Long Short Term Memory (LSTM) and Convolutional Neural Networks (CNN) for spatial earthquake predictions – predicting the maximum intensity and the probability that this maximum will be in a given grid cell.

- **Images as predictors in spatial models are not always informative**.
  - Art images or face photo can predict environmental phenomena well (Fourcade et al., 2018)
  - Help 1: let’s use domain-relevant and structurally related data (Behrens et al., 2020)
  - Hepl 2: use covariates with the same or narrower range of spatial dependence of the dependent variable.

Mixtures of GWR and machine learning models

- How to take one step further from the classical analysis – **change of GWR into machine learning** solution.

- The process behind GWR lies in applying **small local regressions** on neighbouring points for each observation instead of one global estimation.
- Additionally, one decides on:
  - i) the radius and shape of the "moving geometry" (e.g. circle, ellipse)
  - ii) flexibility – fixed or adaptive kernel to address different density
  - iii) weighting scheme – uniform or distance-decaying from the core point
- Li (2019) mixed GWR with machine learning models to **improve wind speed predictions** in China by better capturing local variability. It gave a 12–16% improvement in R^2 and a decrease in RMSE.
- According to Fotheringham et al. (2017), traditional GWR should rather be substituted by **Multiscale Geographically Weighted Regression (MGWR).** In MGWR, one decides on bandwidth not only with regard to location / local density but allows for optimisation of **covariate-specific bandwidth**.

Spatial variables in machine learning models

- Latest solutions try to include **spatial components among covariates** - coordinates or distances between other points.
  - **Buffer distance** (Hengl et al., 2018) - calculated between each point of the territory and observed points, can address spatial autocorrelation better than geo-coordinates
  - **PCA-reduced distance vectors** (from distance matrix) (Ahn et al., 2020) instead of geo-coordinates – less computational requirements and reasonable good performance
  - **Euclidean distance fields** (Behrens et al., 2018) – one may inlclude seven EDF covariates: X and Y coordinates, the distances to the corners of a rectangle around the sample set and the distance to the centre location of the sample set. They address non-stationarity and spatial autocorrelation and improve predictions, as long as their range of spatial dependence is narrower than in dependent variable.

- This discussion os however open - Meyer et al. (2019) assess the inclusion of spatial covariates by quality measures such as Kappa or RMSE.
  - They claim that longitude, latitude, elevation, the Euclidean distances (also as EDF) can be unimportant or even counterproductive in spatial modelling and recommend eliminating those regressors from models.
  - They do not approve of the high fit of ML models, treating them as over-optimistic and misleading
  - They claim that in visual inspection, one observes artificial linear predictions resulting from the inclusion of longitude and latitude, and their elimination helps in making predictions real.

Perspectives (1): *efficiency in big data*

- Spatial Machine Learning enables more efficient computation in the case of big data
- Data low grantualtion is painful for classical spatial econometrics based on an *nxn* spatial weights matrix W or *nxn* distance matrix.
- Arbia et al. (2019) shows that the max. size of dataset for PC analysis is around 70,000, while already with 30,000 observations, the creation of W is challenging (Kopczewska, 2021).
- ML models, which are free of W, are automatically quicker, but addressing the autocorrelation issue is still a challenge
- New sources of data such as lightmaps of terrain (Night Earth, Europe At Night, NASA, etc.) or day photos of landscape (Google Maps, Street View etc.) bring new insights and information, and due to big-data robust analytics are useful.
- Spatial data handling (e.g., processing remote sensing image classification or spectral-spatial classification, executed with supervised learning algorithms, ensemble and deep learning) is especially helpful in big data tasks (Du et al., 2020).

Perspectives (2): *spatial patterns*

- There are different ways to address spatial heterogeneity and isotropy.
- Classical spatial econometrics was concentrated on spatial autocorrelation and mostly neglected other problems.
- Local regressions, combined with global ones, help in capturing unstable spatial patterns.
- Methods open a path for spatio-temporal modelling and studying the similarity of different layers - the dynamics connected to location can be addressed in more ways than the classical panel model.

➔ Integrating classical statistics and econometrics with machine learning provides more instruments to the modelling toolbox than a single approach.

Perspectives (3): *forecasting and new topics*

- These methods allow for better forecasting due to inherited boosting and bootstrapping in ML algorithms.
- ML results are also more flexible for spatial expanding into new points and ensemble methods give better prediction
- A shift towards spatial ML from spatial econometrics is also a change from explanation into forecasting.
- Methods drive innovations such as new indicators based on vegetation index or light indicators.
- Methods also introduce 3D solutions to spatial studies, such as social topography with 3D inequalities (Aharon-Gutman et al., 2018; Aharon-Gutman & Burg, 2019), 3D Building Information Models (Zhou et al., 2019) or urban compactness growth (Koziatek & Dragićević, 2017).
- There appear urban studies that rely on information from GoogleStreeView, by counting cars, pedestrians, bikers etc. to predict traffic (Goel et al., 2018), or counting urban disorders such as cigarette butts, trash, empty bottles, graffiti abandoned cars and houses etc. to predict neighbourhood disorder (Marco et al., 2017) or counting green vegetation index to predict safety (Li et al., 2015).

Example of application

Land Use Policy 103 (2021) 105292

Contents lists available at ScienceDirect

Land Use Policy

journal homepage: www.elsevier.com/locate/landusepol

Spatio-temporal stability of housing submarkets. Tracking spatial location of clusters of geographically weighted regression estimates of price determinants☆

Katarzyna Kopczewska *, Piotr Ćwiakowski

Faculty of Economic Sciences, University of Warsaw, Poland

ARTICLE INFO

JEL classification:
R31
C21

Keywords:
Geographically weighted regression
Spatio-temporal stability
Spatial location
Housing valuation
Submarkets
Data-driven clusters

ABSTRACT

This paper fills the gap in rich housing literature by testing the spatio-temporal stability of real estate submarkets. We start with standard Geographically Weighted Regression (GWR) estimation of the hedonic model on point data, and we cluster model coefficients to detect housing submarkets. We check spatio-temporal stability - we add novelty by comparing if clusters move over space or stay in the same place. We rasterise surface and apply the Rand Index and Jaccard Similarity to check if clusters assigned to raster cells yield stable spatial structure. This approach allows for quantitative assessments of how much determinants of price are stable over time and space. The same mechanism applied to standard errors of GWR coefficients is a good test of the spatio-temporal stability of local heteroscedasticity. A Case study of apartments' transactions in Warsaw-Poland for the 2006–2015 period, evidences relatively high spatio-temporal stability.

Goal of paper

Study on spatio-temporal stability of clusters:

- → we have geo-located point data for housing transactions
- → we estimate hedonic GWR (Geographically Weighted Regression) model for each year
- → for each year we take GWR coefficients and **cluster them**
- → we get the submarkets – the areas where the coefficients are similar (clusters)
- → we inquiry, if the submarkets are stable over time and space – are clusters changing their location (moving over space)?
    - → we divide area into raster cells and assign points to rasters
    - → in rasters, we count median / mode value of cluster ID
    - → with Rand Index / Jaccard similarity we test year-to-year cluster similarity in rasters
    - → this shows the percentage of raster cells belonging to the same cluster

This is novelty – till now this approach was never used.

dataset by years & ID, geo-located points

Year	obs ID	long (x)	lat (y)	Y	x1	x2	...	xk
Year 1	101	...	...	...	...	...	...	...
Year 1	102	...	...	...	...	...	...	...
Year 1	103	...	...	...	...	...	...	...
Year 1	104	...	...	...	...	...	...	...
Year 1	105	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
Year n	n01	...	...	...	...	...	...	...
Year n	n02	...	...	...	...	...	...	...
Year n	n03	...	...	...	...	...	...	...
Year n	n04	...	...	...	...	...	...	...
Year n	n05	...	...	...	...	...	...	...

GWR (year 1)

GWR (year n)

GWR coefficients by years & ID

β(x1)	β(x2)	...	β(xk)
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...

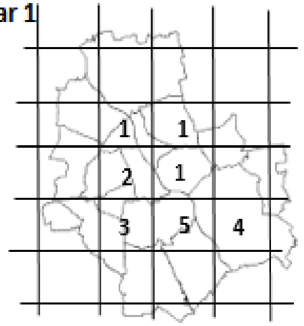
kmeans (year 1)

kmeans (year n)

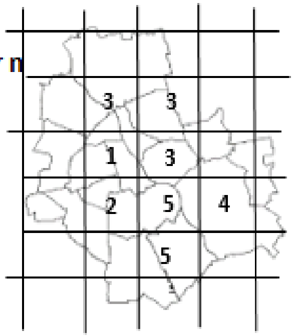
cluster ID & rasterized location

cluster ID	raster cell
1	(3,2)
1	(3,3)
2	(4,2)
4	(5,4)
5	(5,3)
...	...
3	(3,2)
3	(3,3)
4	(5,4)
5	(5,3)
5	(6,3)

rasters of dominant cluster ID by years



rasters of dominant cluster ID by years



comparison of clusters by raster cells by years

raster cell	dominant cluster ID year 1	...	...	year n
(1,1)	...	...	...	...
(1,2)	...	...	...	...
(1,3)	...	...	...	...
...	...	...	...	...
...	...	...	...	...
(3,2)	1	...	...	3
(3,3)	1	...	...	3
...	...	...	...	...
(4,2)	2	...	...	...
(4,3)	...	...	...	...
...	...	...	...	...
(5,3)	5	...	...	5
(5,4)	4	...	...	4
...	...	...	...	...

Rand Index

RI	year1	...	...	year n
year1	...	...	...	...
...	...	...	...	...
...	...	...	...	...
year n	...	...	...	...

Jaccard Index

JI	year1	...	...	year n
year1	...	...	...	...
...	...	...	...	...
...	...	...	...	...
year n	...	...	...	...

Details of hedonic model

Dataset from secondary market transactions of apartments in Warsaw from the Registry of Price and Value of Real Estate, run by the Mayor of Warsaw. For *t*=10 years between 2006 and 2015 → 65,674 observations

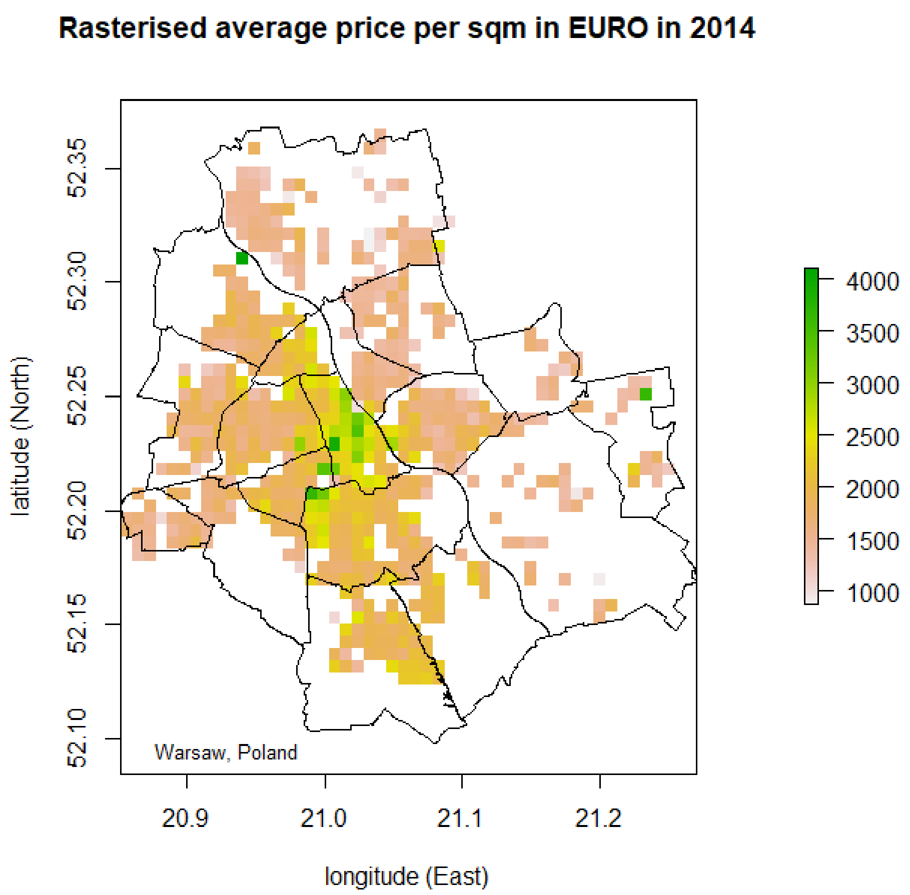
GWR model:

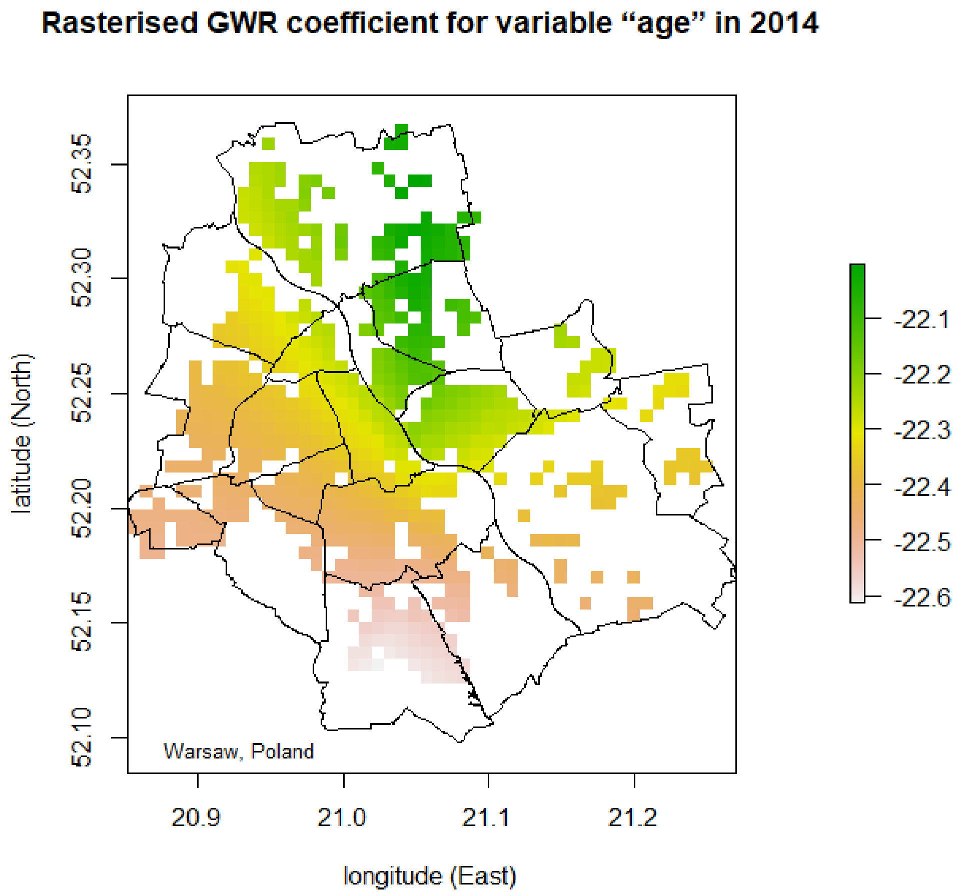
$$y_i = \beta_0(u_i, v_i) + \sum_{s=1}^{S} x_{i,s}\beta_s(u_i, v_i) + \varepsilon_i$$

where *i* is the observation number *i*=1, …,N; *s* is the feature number in the set *s*=1, …, S; y_i is the value of the dependent variable of the *i*-th observation; $x_{i,s}$ is the value of the *s*-th feature (explanatory variable) of the *i*-th observation; u_i, v_i are the geographical coordinates of the observation, $\beta_s(u_i, v_i)$ is the value of the effect of the *s*-th feature for given geographic coordinates (GWR coefficients) and ε_i is the error term.

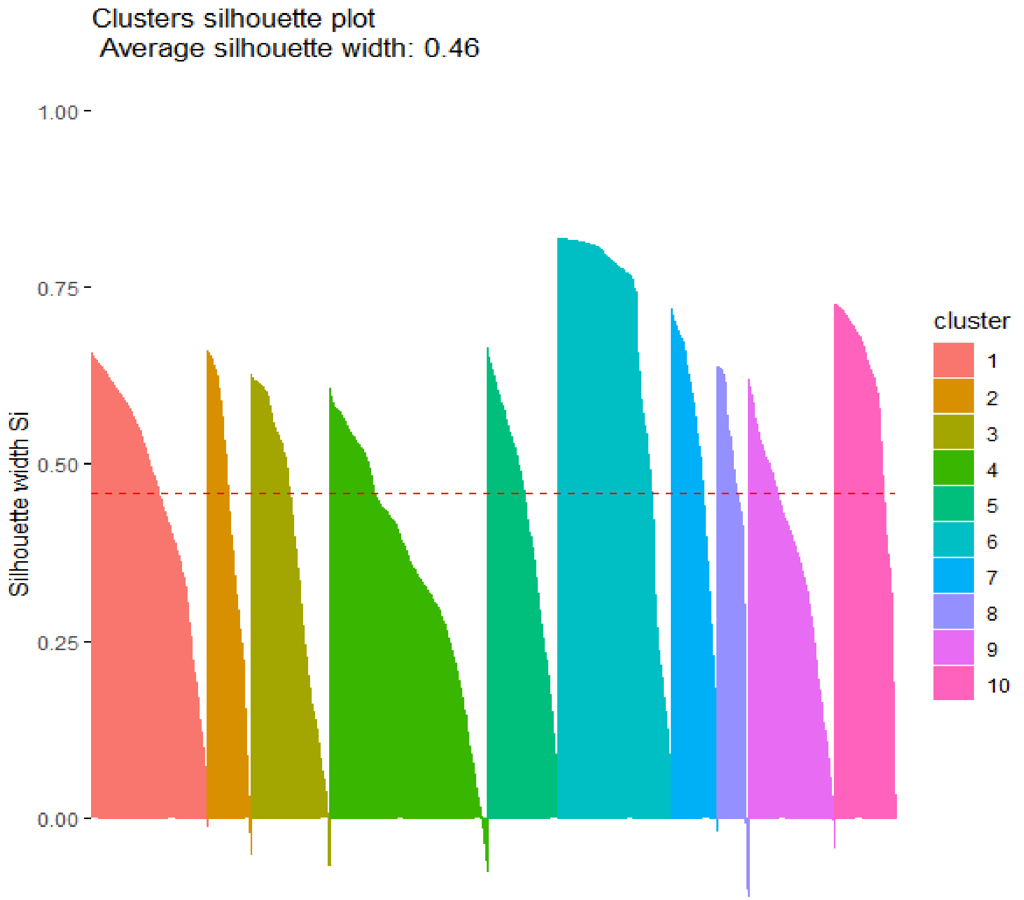
$$\begin{aligned} &price\_sqm_{it} \\ &= \beta_{0,t} + \beta_{1,t} age_{it} + \beta_{2,t} berdrooms_{it} + \beta_{3,t} floor_{it} + \beta_{4,t} floor\_0ground_{it} + \beta_{5,t} floor\_bground_{it} + \beta_{6,t} garage_{it} \\ &+ \beta_{7,t} renovation_{it} + \beta_{8,t} basement_{it} + \beta_{9,t} sq\_building_{it} + \beta_{10,t} time\_traffic_{it} + \beta_{11,t} time\_public_{it} \\ &+ \beta_{12,t} time\_driving_{it} + \beta_{13,t} dist.public_{it} + \beta_{14,t} frac\_green_{it} + \beta_{15,t} frac\_forests_{\mathrm{it}} + \beta_{16,t} count\_historic_{it} \\ &+ \beta_{17,t} frac\_historic_{it} + \beta_{18,t} \mathrm{frac\_priv}_{\mathrm{it}} + \beta_{19,t} \mathrm{rights\_owned}_{\mathrm{it}} + \varepsilon_{it} \end{aligned}$$

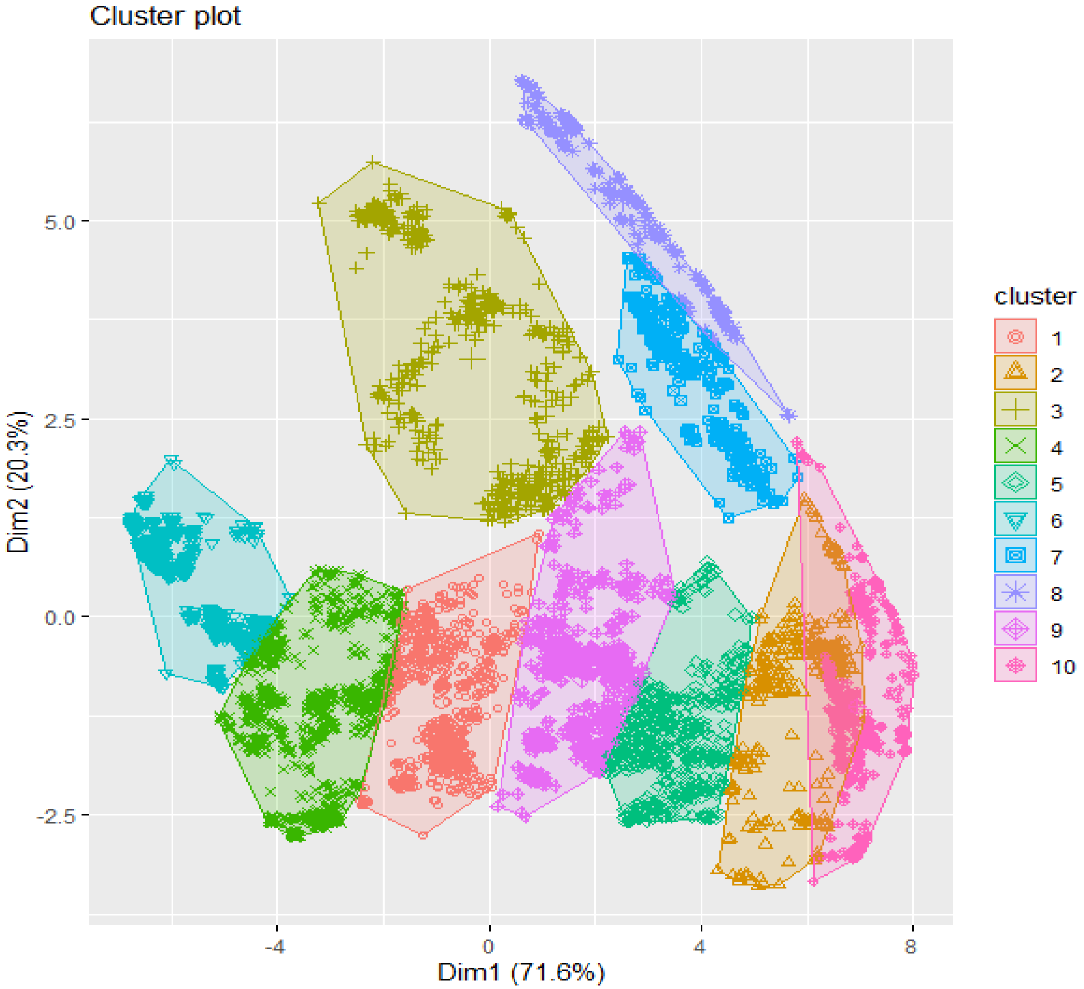
Spatial distributions of modelled pattern:
a) price per sqm in 2014,
b) GWR coefficients for variable “age” in 2014



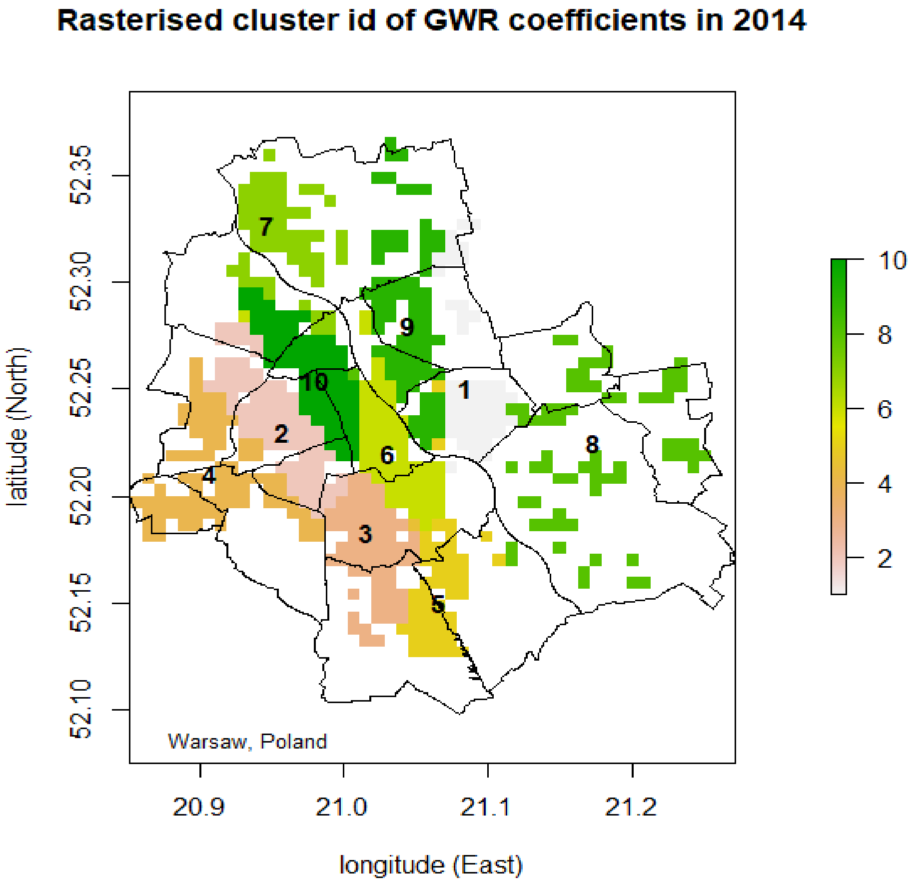


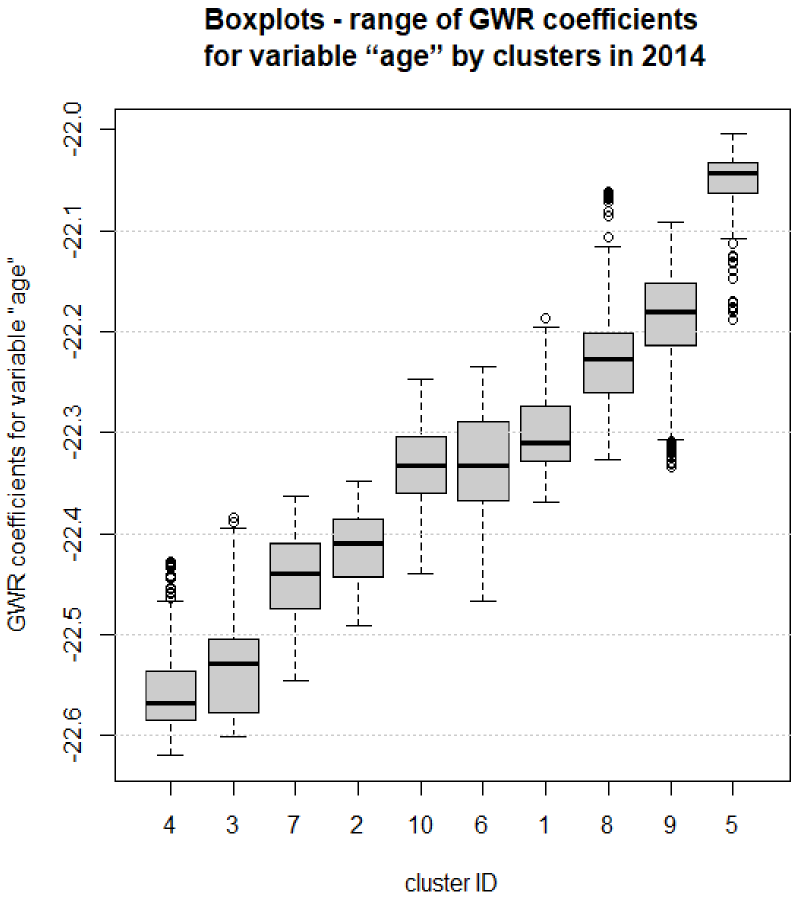
Clustered GWR coefficients:
a) Silhouette width for all clusters in 2014,
b) PCA representation of k-means clusters in 2014





Clustered GWR coefficients:
a) raster map of clusters for 2014,
b) boxplot of variable “age” by clusters in 2014





Rand Index and Jaccard Similarity

Analysis of spatio-temporal stability - by comparing values in rasters over time (as comparison of ordered vectors of cluster *id)*.

Rand and Jaccard indices compare pairs of pairs of raster cells, checking if both pairs belong to the same or different cluster in both analysed periods (t0 and t1).

Rand Index=1 means that partitions always agree (c and d are NULL) and clusterings are the same, while Rand Index=0 means that partitions migrate and do not agree for even a single pair.

Jaccard similarity is interpreted as Rand Index, but it is concentrated only on pairs that are connected, being a zoom compared to the Rand Index.

Rand Index is **R=(a+b)/(a+b+c+d)**

Jaccard similarity is **J=a/(a+c+d)**

where:

- a - in t0 the same, in t1 the same,
- b - in t0 different, in t1 different,
- c - in t0 the same, in t1 different,
- d - in t0 different, in t1 the same;

Rand: The counter is always the same (a) and always different (b), and denominator are all possible outcomes (a,b,c,d).

Jaccard: it omits a number of events which are always in different clusters (b), both in the counter and denominator.

Rand Index and Jaccard Similarity

period 1

cell A cluster 1	cell B cluster 1
cell C cluster 2	cell D cluster 2

initial pattern

period 2

cell A cluster 2	cell B cluster 2
cell C cluster 1	cell D cluster 1

relabeling of clusters
no change in pattern

period 3

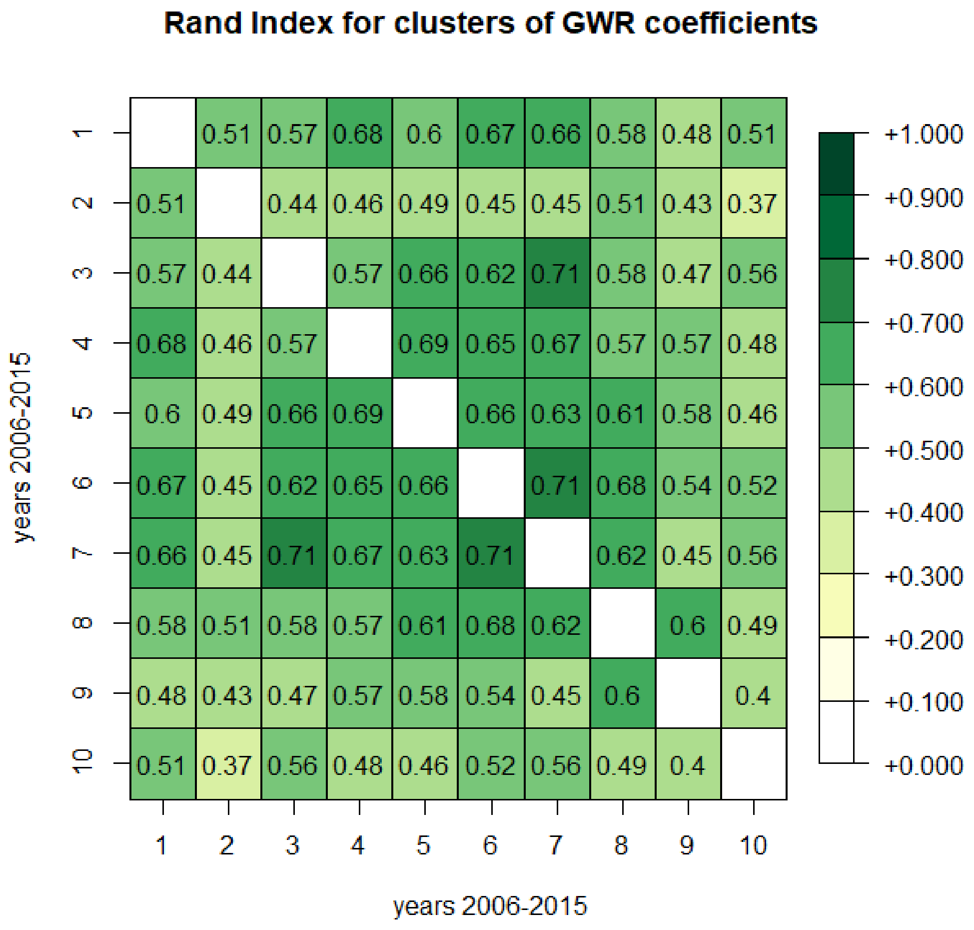
cell A cluster 1	cell B cluster 2
cell C cluster 1	cell D cluster 2

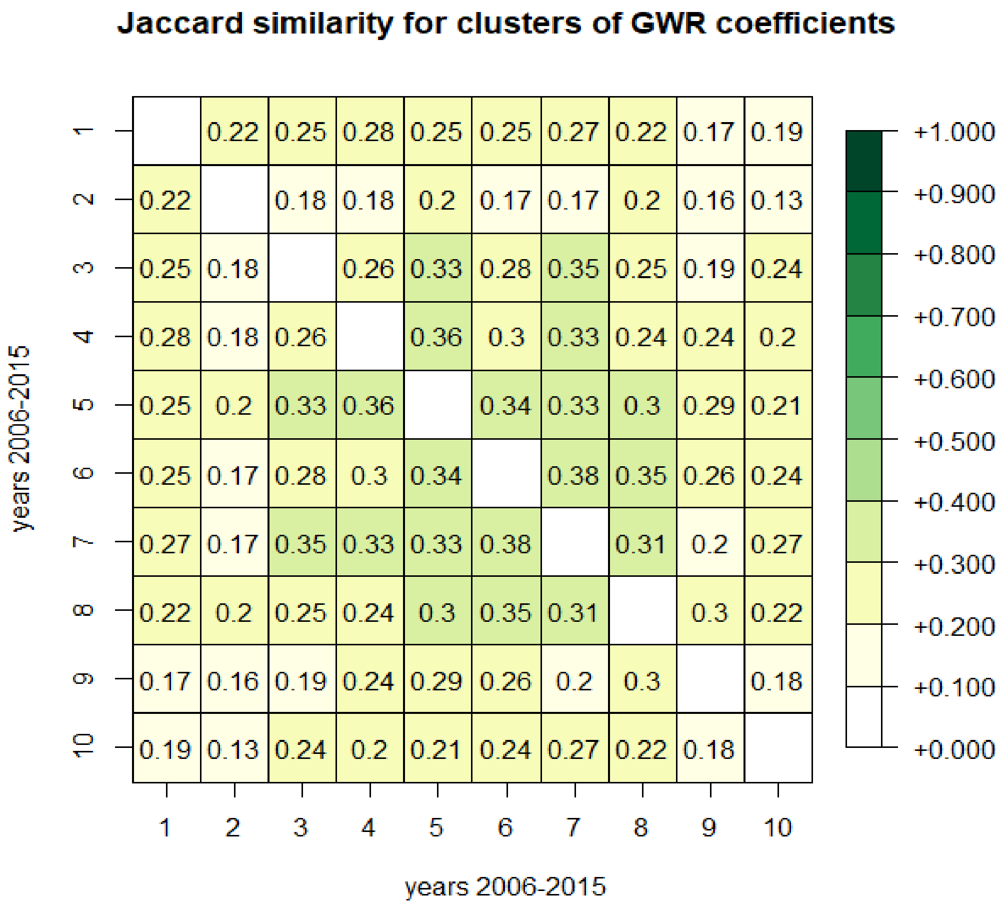
shift of clusters
change in pattern

comparison of pairs of cells in two periods	period 1 (t0) & period 2 (t1)		period 1 (t0) & period 3 (t1)	
a (in t0 the same, in t1 the same)	A:B, C:D	2	---	0
b (in t0 different, in t1 different)	A:C, B:D, A:D, B:C	4	A:D, B:C	2
c (in t0 the same, in t1 different)	---	0	A:B, C:D	2
d (in t0 different, in t1 the same)	---	0	A:C, B:D	2

Rand Index	(a+b)/(a+b+c+d)	=6/6=100%	(a+b)/(a+b+c+d)	=2/6=33%
Jaccard similarity	a/(a+c+d)	=2/2=100%	a/(a+c+d)	=0/4=0%

Rand Index & Jaccard similarity for clusters of GWR estimates in pairs year-to-year





Conclusions

It is novel methodological solution for testing spatio-temporal stability and works well

That might be used in land use policy, urban organisation etc.

It gives new insights into real estate submarkets – to understand urban trends, urban mobility, residential segregation, wealth accumulation, effects of revitalisation, equal housing opportunities

IN GENERAL, SPATIALML ALLOWS ANSWERING NEW QUESTIONS!

Thank You!