Regional Science Policy&Practice
Regional Science Association International

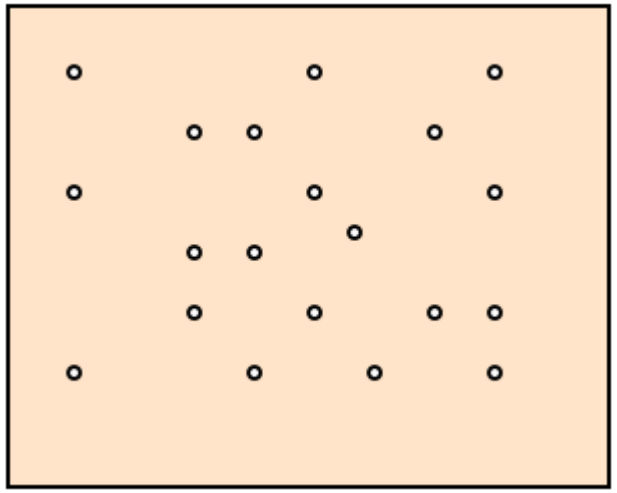
Spatial clustering based on distance

Katarzyna Kopczewska

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

Spatial clustering

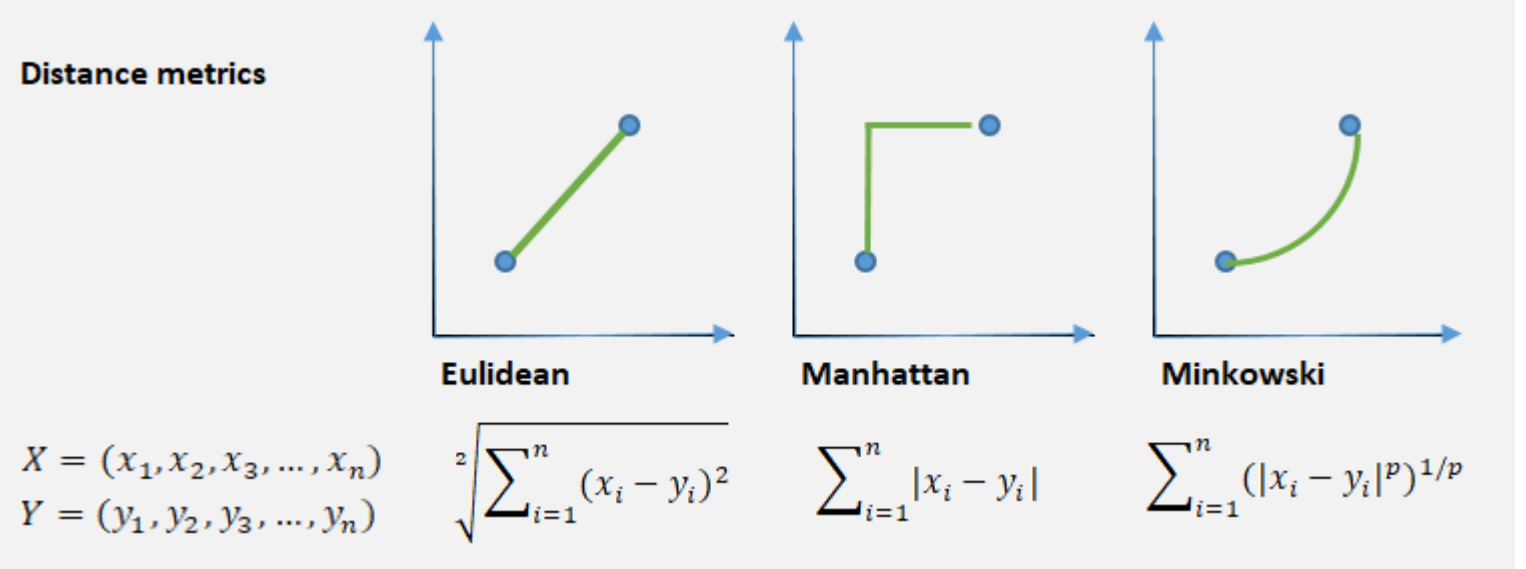
- Covered here - based on distance:
  - K-means
  - PAM & CLARA
  - Hierarchical
- Not covered here - based on density and circles:
  - DBSCAN, OPTICS, SaTScan, BNS, GAM
  - Dynamic spatio-temporal ST-DBSCAN
- Not covered here - based on network:
  - Voronoi/Dirichlet tesselation-based
- Not covered here – out of regional science:
  - spatial transcriptomics in RNA analysis



We analyse spatially located points (point-pattern), optionally with value assigned

What and how can we cluster?

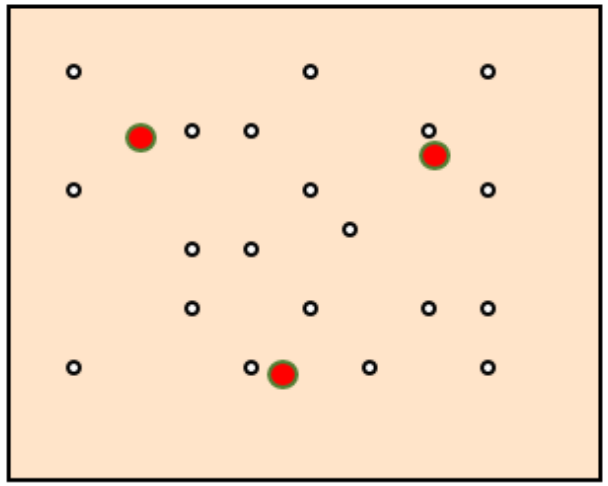
- Points geo-located in space – we cluster their coordinates (longitude and latitude, *xy*)
- Values *z* – multivariate characteristics of observations (no spatial aspect considered)
- Values *z* – multivariate characteristics of observations – to map them in their geo-location
- Geo-located points *xy* and values *z* jointly



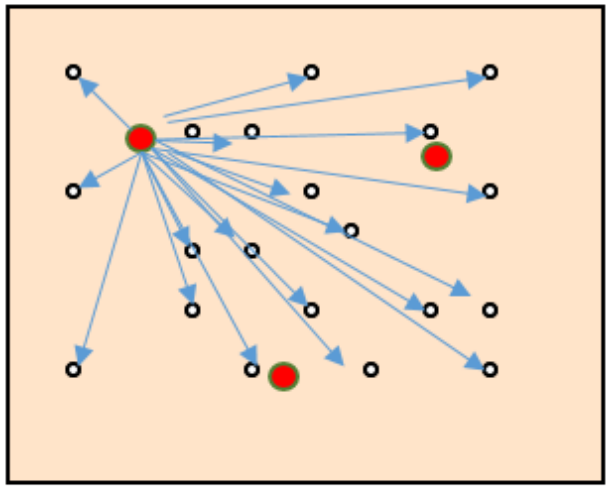
In distance-based measures the core point is DISTANCE metrics. There are many options:

- Euclidean distance – to go straight ahead
- Manhattan distance – to move around the edges of the grid
- Minkowski distance – to use curve way
- Gower distance – for qualitative data
- Mahalanobis distance – to include correlations between variables
- Hamming distance – to compare binary vectors

K-means clustering



We assume k points on the plane (any location)



We get distances between points and cores

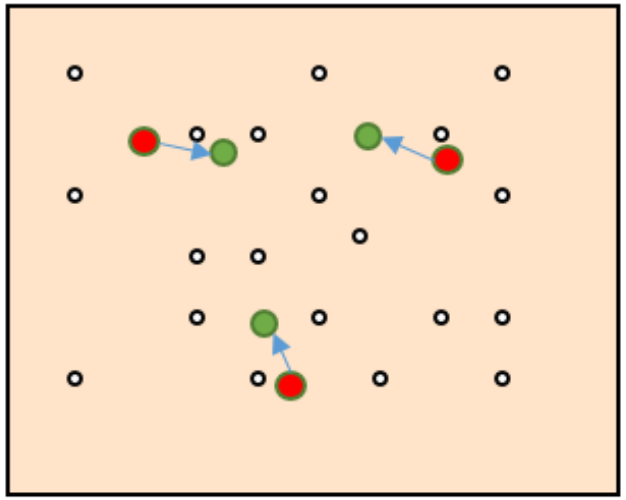
Point	dist-to-1	dist-to-2	dist-to-3	**min-dist**
1	...	...	...	...
2	...	...	...	...
3	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
n	...	...	...	...

We link points with cores by min dist **Total**

where to put ● to minimize Total distance?

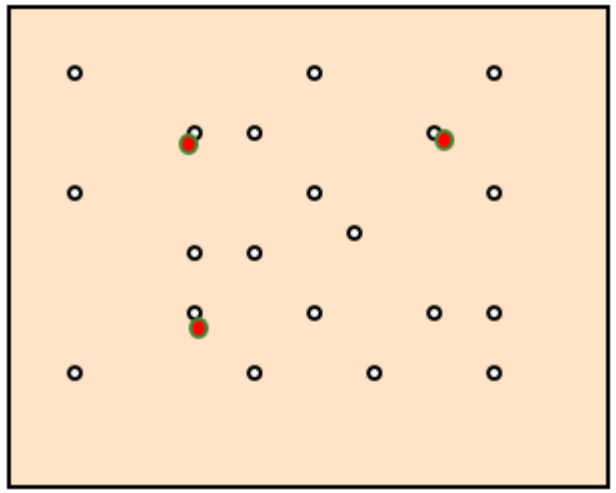
The ● are best locations.

● ● can be anywhere

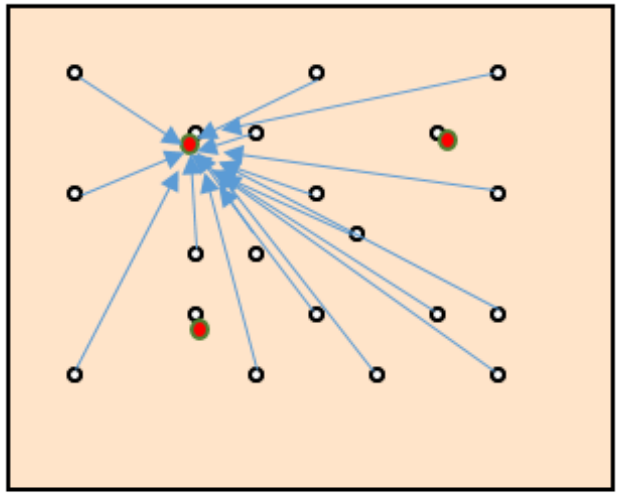


We optimise location of cores to min total dist.

Partitioning Around Medoids (PAM) & CLARA



We select k points as cores (from the sample)



We get distances between points and cores

Point	dist-to-1	dist-to-2	dist-to-3	**min-dist**
1	...	...	...	...
2	...	...	...	...
3	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
n	...	...	...	...

We link points with cores by min dist **Total**

One tries all possible combinations of cores

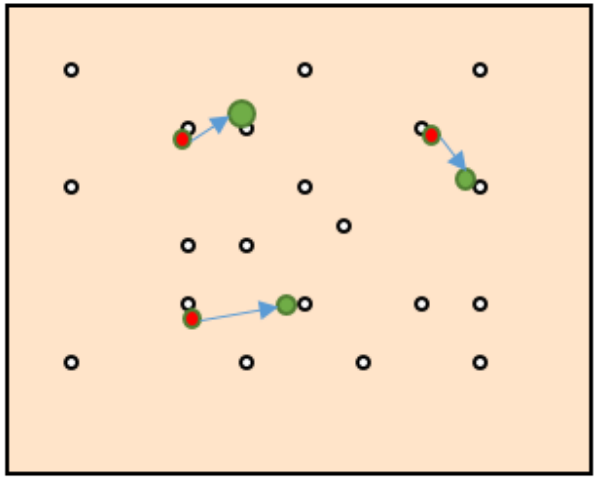
e.g. (n1, n2, n3); (n2, n4, n7); (n3, n6, n65)……

The ● are best locations.

We choose a combination

with minimum Total distance.

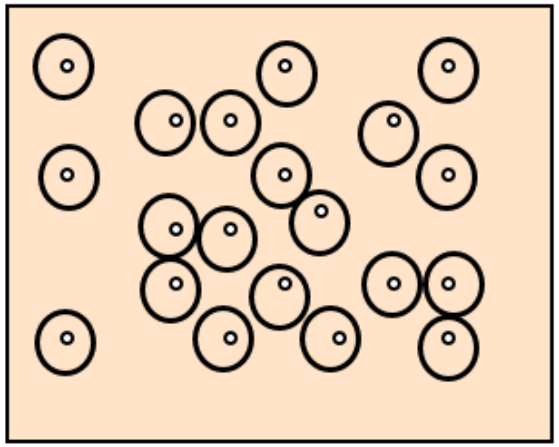
● ● are existing points



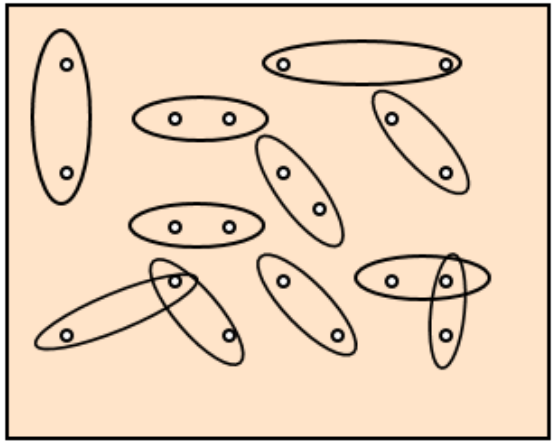
We choose best iteration (cores set) to min total dist.

CLARA is big data equivalent. It works as PAM but on subsample. The rest of points is assigned to clusters using k nearest neighbours.

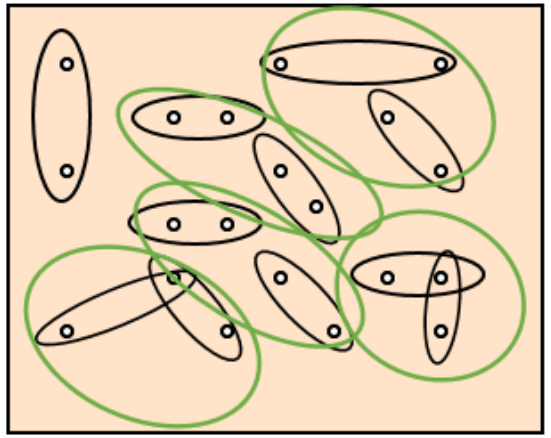
Hierarchical clustering

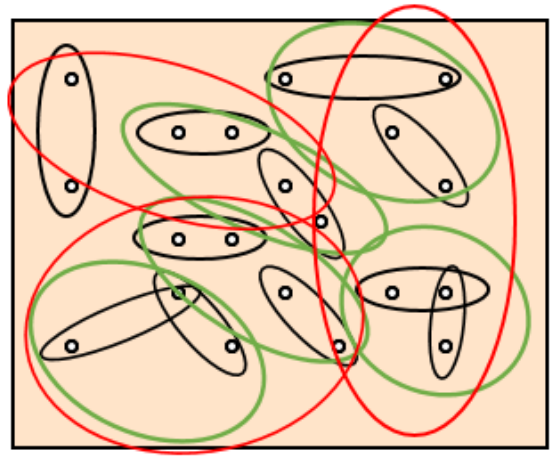


We start with singeltons - each point in own cluster.

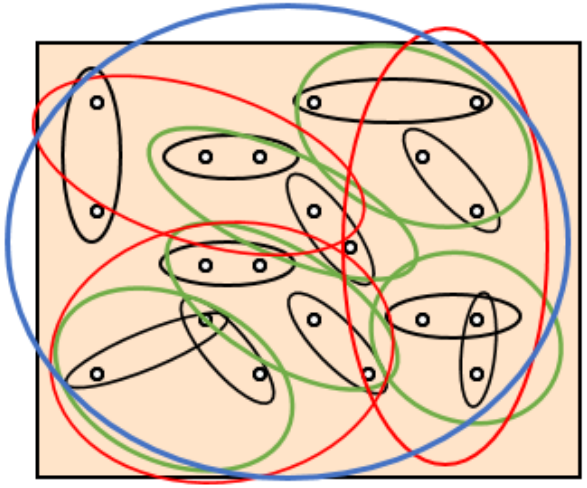


We group points with nearest neighbour.

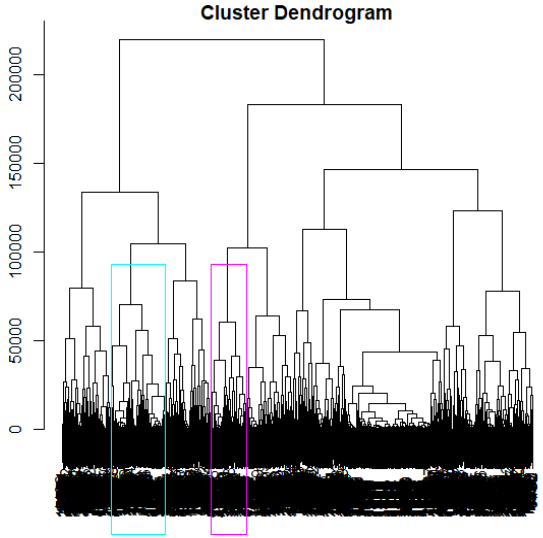




Existing points clusters are linked in bigger groups



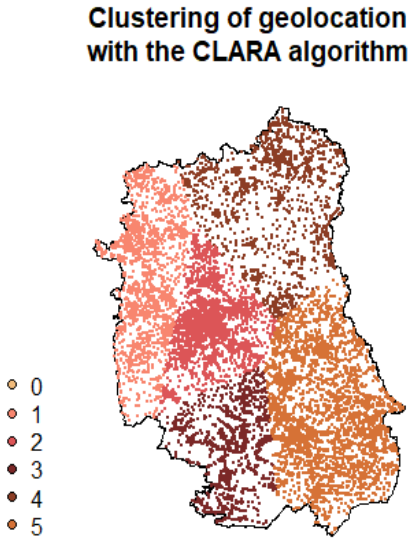
Finally, all points belong to one cluster.

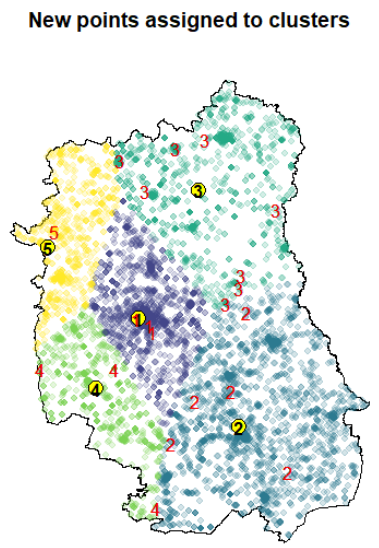


We can cut tree (dendrogram) at any heigh to decide which clustering we choose.

Clustering of geo-located points

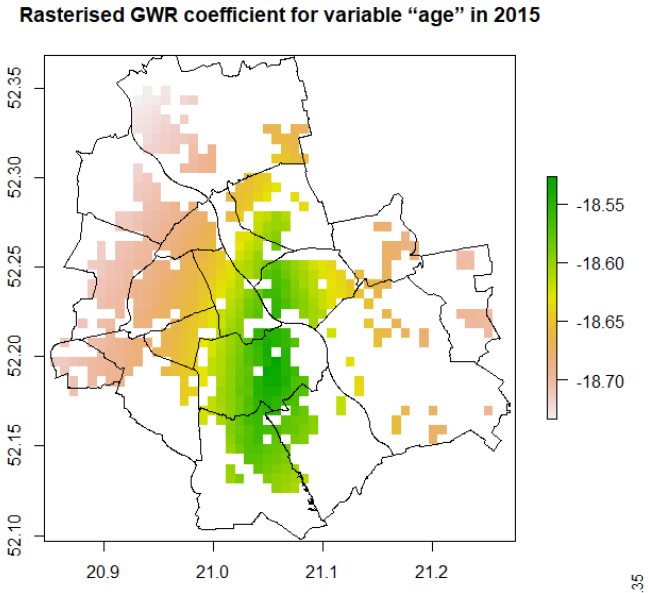
- Straighforward implementation of points clustering is to catchment areas: schools, post-offices, supermarkets, sales representatives
- With use of calibrated algorithms, one can predict cluster assignement for any new point
- In case of known *a priori* number of cores *k*, k-means optimizes the partitioning.
- Researchers complain if *k* is unknown and look for clue how to set *k*. It can be optimised by comparison different potential *k* and choosing the best division.
- Brimicombe (2007) proposed dual approach – in first step one finds the density clusters (with GAM - *Geographical Analysis Machine,* or kernel density), and in the second step, one uses them as initial points in k-means clustering. This automates the selection of *k* and speeds up the computations by setting starting centroids.





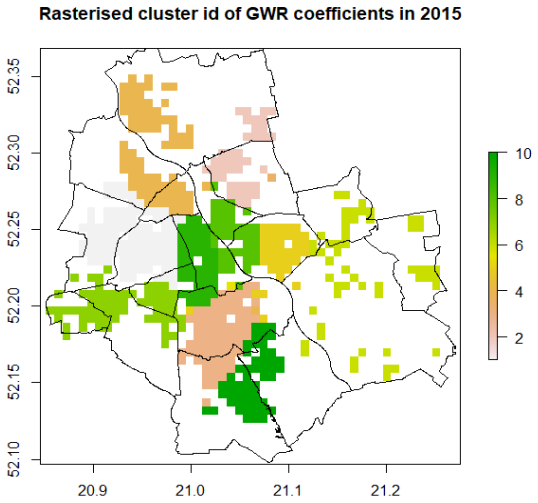
Clustering of values & mapping as clusters

- Most popular application appears in case of *Geographically Weighted Regression* (GWR) – GWR produces individual local beta coefficients for each variable and observation, what makes it difficult to summary.
- GWR coefficients in hedonic models can be clustered and mapped. Clusters are considered as submarkets.
- It is always amazing in GWR hedonic models that even if clustering of values is non-spatial, clusters are mostly continously spatial.



Values of GWR coefficients for selected variable

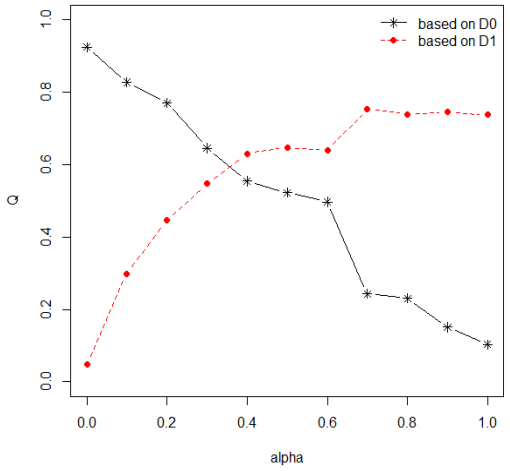
Clusters of values (GWR coefficients) for many variables



Both figures present rasterised data

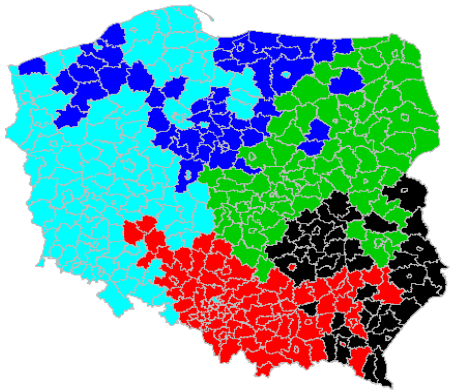
Clustering locations & values

- When clustering values in locations one may have dilema what to concentrate on.
- This dilema can be mildered by using methods that mix both clusterings.
- ClustGeo (Chavent et al., 2018) runs separate clusterings and combines them by using weighted dissimilarity (distance) matrices.
- Its extension, BootstrapClustGeo (Distefano et al.,2020) generates many potential partitionings, links spatial and non-spatial components with Hamming distance, and as ClustGeo minimizes within-cluster inertia in mixture.
- SKATER and REDCAP algorithms (Assuncao et al., 2006; Guo, 2008) build trees which are later pruned.



Cluster coherence lines for locations and values. When cross, they set mixing parameter α.

Partitioning which balances data and location.



Research studies involving clustering (1)

- **Fire distribition in Sardinia (Bajocco et al., 2015)**
  - It uses **hierarchical clustering** to group the territorial units into similarly covered areas (features are phenological metrics and spatio-temporal dynamics of the vegetated land surface (NVDI, Normalized Difference Vegetation Index from satellite photos).
  - It gave clusters defining types of territories – they were mapped with cluster ID.
  - For each cluster group they checked the frequency of fires – they assessed the natural conditions which increase and decrease fire-proneness.

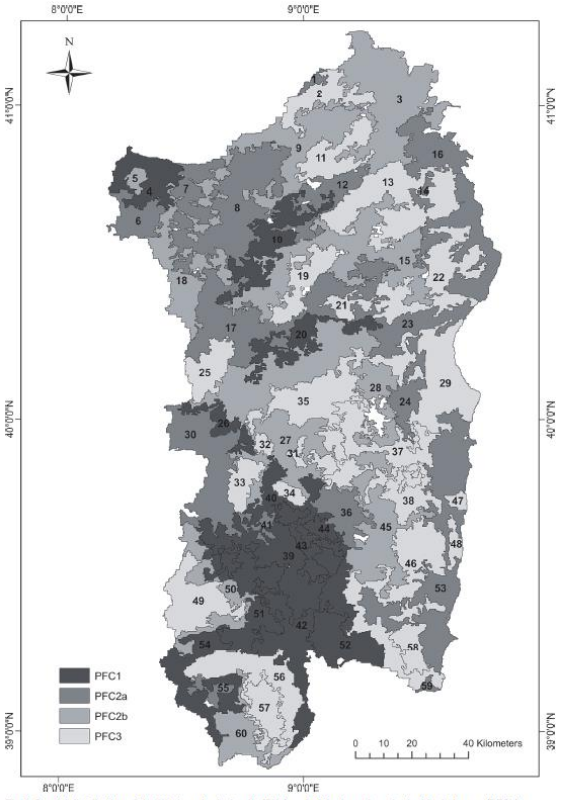


Fig 1. Spatial distribution of the 60 phenological units (PUs) and of the four phenological fuel classes (PFCs).

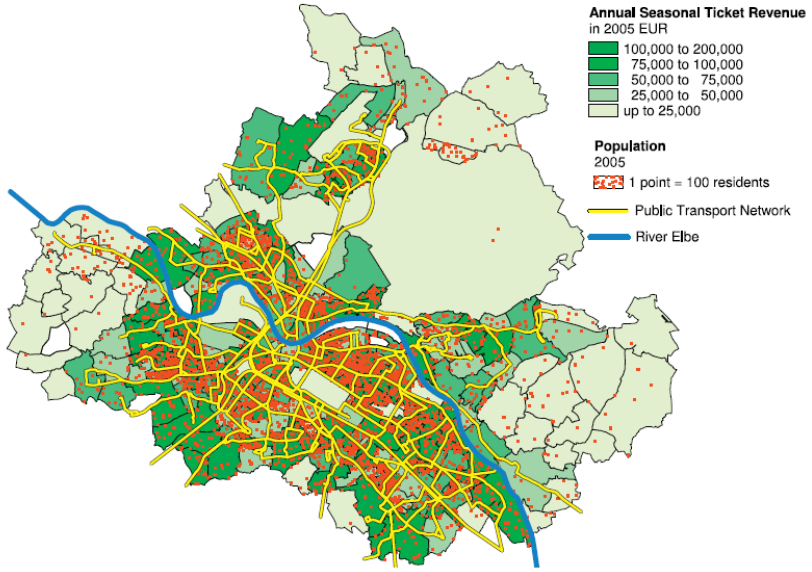
Table 4. Summary of the environmental characteristics of the phenological fuel classes (PFCs) obtained from the analysis.

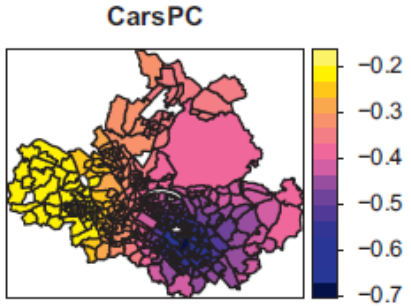
	Number of fires	Fire risk	Fuel type	Seasonal NDVI variability	Main Land Cover types	Main Climatic types
PFC1 (4210.25 km^2)	10690	High (σ = 1.98)	Fine fuel	Very high	Arable lands	Mediterranean
PFC2a (5759.63 km^2)	10341	Moderately high (σ = 1.40)	Fine fuel	High	Urban areas, Permanent crops, Heterogeneous agricultural areas and Natural grasslands and pastures	Mediterranean Transitional Temperate
PFC2b (5743.31 km^2)	4928	Moderately low (σ = 0.67)	Coarse fuel	Low	Permanent crops, Heterogeneous agricultural areas, Natural grasslands and pastures, Forest and Shrublands	Transitional Mediterranean Transitional Temperate
PFC3 (6691.31 km^2)	2785	Low (σ = 0.32)	Coarse fuel	Very low	Forest and Shrublands	Transitional Mediterranean Transitional Temperate

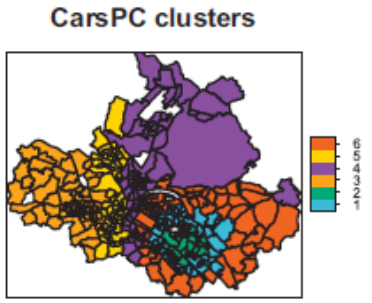
Bajocco, S., Dragoz, E., Gitas, I., Smiraglia, D., Salvati, L., & Ricotta, C. (2015). Mapping forest fuels through vegetation phenology: The role of coarse-resolution satellite time-series. PloS one, 10(3), e0119811.

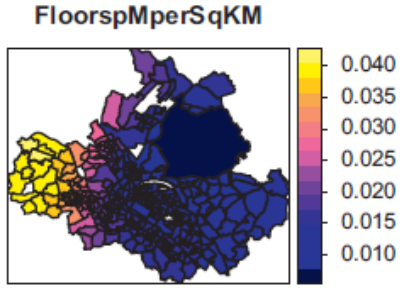
Research studies involving clustering (2)

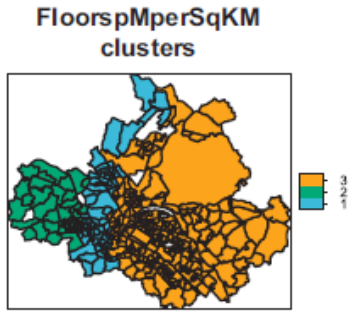
- **Spatial drift in demand for public transport tickets (Müller et al., 2013)**
  - Firstly, it estimates GWR model for demand on bus tickets, using point and district data.
  - GWR coefficients are clustered with **k-means** – individually for each variable
  - Secondly, it estimates general spatial econometric model with the same variables as in GWR (to capture spatial autocorrelation), and additionaly with dummy variables for GWR clusters (they are to capture spatial heterogenity)







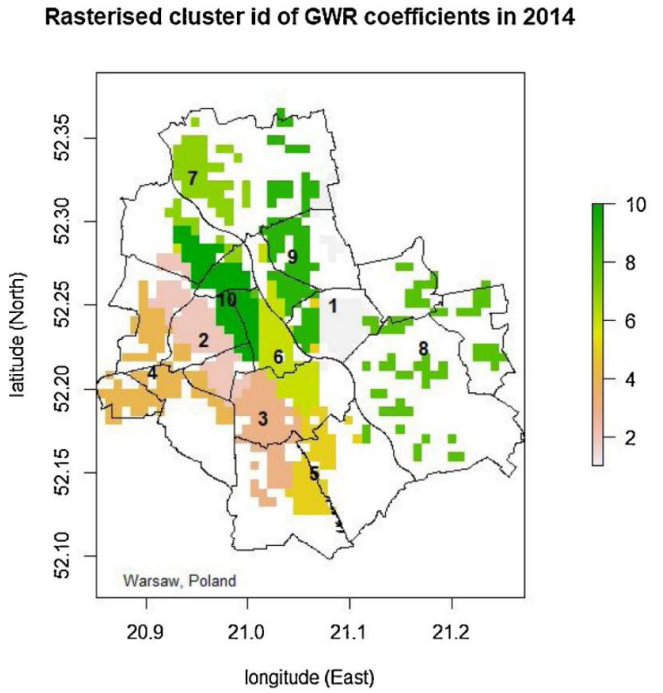


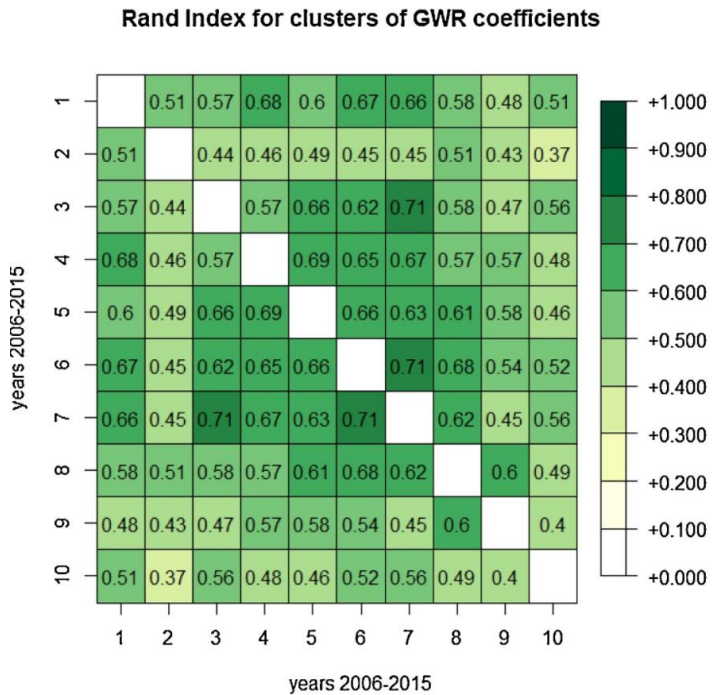


Müller, S., Wilhelm, P., & Haase, K. (2013). Spatial dependencies and spatial drift in public transport seasonal ticket revenue data. *Journal of Retailing and Consumer Services*, *20*(3), 334-348.

Research studies involving clustering (3)

- **Spatio-temporal stability of housing submarkets (Kopczewska & Ćwiakowski, 2021)**
  - It estimates annual hedonic GWR models for housing market for 10 years
  - Point coefficients were rasterised – to get common spatial reference
  - Annual GWR coefficients are clustered with **k-means** – one gets the sets of clusters for each year
  - With Rand Index / Jaccard Similarity one checks for how much the clusters overlap from period-to-period – this shows spatio-temporal stability of submarkets

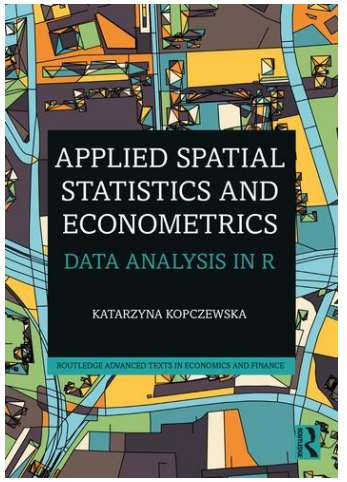




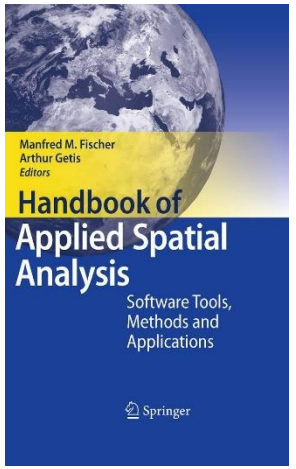
Kopczewska, K., & Ćwiakowski, P. (2021). Spatio-temporal stability of housing submarkets. Tracking spatial location of clusters of geographically weighted regression estimates of price determinants. *Land Use Policy*, *103*, 105292.

Software implementations

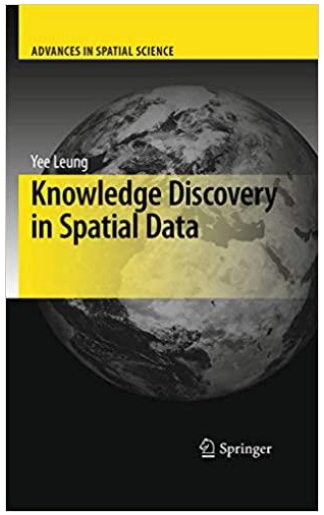
- Basic methods are available in R, in packages as: stats:: , ClusterR:: , cluster:: , fpc:: , factoextra:: , FactoMineR::
- Simultaneous clustering of values and locations (spatially constrained clustering) is in ClustGeo:: and rgeoda:: packages.
- What is more:
  - Package spatialClust:: offers Spatial Clustering using Fuzzy Geographically Weighted Clustering
  - Package SpODT:: offers spatial oblique decision tree based on the classification and regression tree
- Also, non-covered topics (network and density-based clustering) are widely available in R: in dbscan::, geoGAM::, MapGam::, SpatialCPie::, SpatialEpi::, rsatscan::, graphscan::, rflexscan::, scanstatistics:: packages.



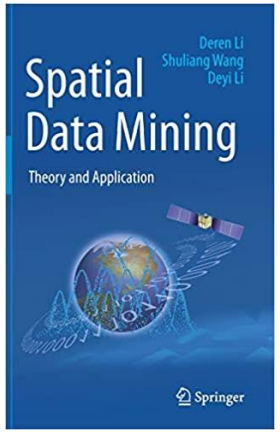
Kopczewska, K. (2021). Applied Spatial Statistics and Econometrics. Data Analysis in R, Routledge



Fischer, M. M., & Getis, A. (Eds.). (2009). Handbook of applied spatial analysis: software tools, methods and applications. Springer Science & Business Media.



Leung, Y. (2010). Knowledge discovery in spatial data (pp. 223-276). Berlin, Germany:: Springer.



Li, D., Wang, S., & Li, D. (2015). Spatial data mining. Berlin, Heidelberg:: Springer