Aalto University

GEOGRAPHICALY WEIGHTED MODELS AND REGRESSION (GWR)

Spatial Data Mining

24.1.2019 Jaakko Madetoja

(slides also by James Culley and Olga Špatenková)

Today

- Lecture: Geographically weighted models and regression
  - Introduction
  - Global models: Theory and examples
  - Local models: Theory and examples
    - Geographically weighted regression

- On Monday: Computer class exercise at 10:15 in Maari-E in Maarintalo

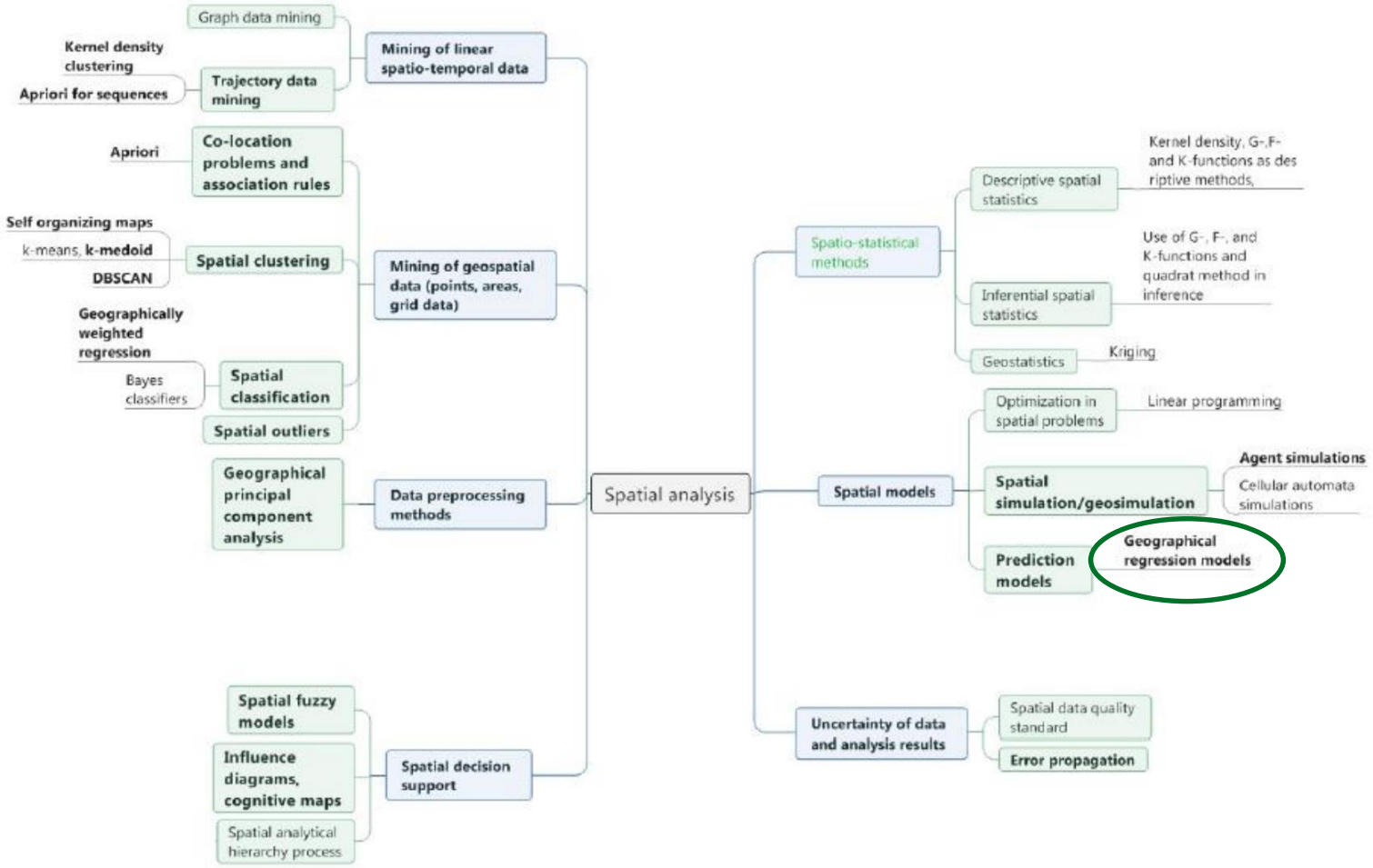
Learning goals

- After this lecture, you are able to
  - Discuss problems with global statistical models
  - Explain how geographically weighted models are created
  - List and explain the required steps in geographically weighted regression

Geographically Weighted Regression (GWR)

- Most of this lecture is based on materials found in the book (not available in libraries)
  - A. S. Fotheringham, C. Brunsdon, M. Charlton: Geographically weighted regression – the analysis of spatially varying relationships, Wiley 2002
- See also Chapter 7 in Discovering Spatio-Temporal Relationships: A Case Study of Risk Modelling of Domestic Fires by Olga Špatenková, 2009, http://lib.tkk.fi/Diss/2009/isbn9789522482334/ or Chapter 3.1 (theory), 6.1 and 6.2 (example) in Error propagation in geographically weighted regression by Jaakko Madetoja, 2018, https://aaltodoc.aalto.fi/handle/123456789/29575

The big picture



Problem

- Imagine reading a book on climate of United States
- If book contained info on data averaged across whole country
  - E.g
    - Mean averaged sunshine
    - Mean average temperature
    - Mean average rainfall
- How could we tell anything about the climate in any particular location of the United States
- Global rather than Local *observations*

Problem

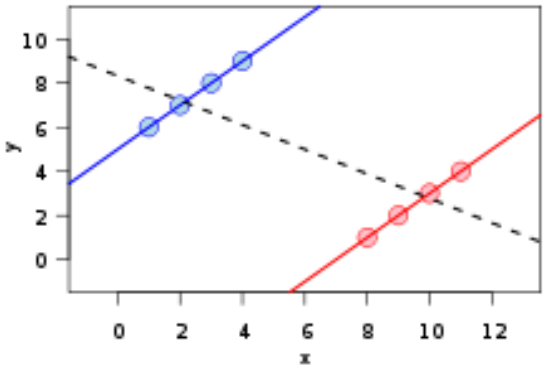
- Now imagine a model were we are trying to explain house prices by a list of variables such as
    - Size of house
    - Number of bedrooms
    - Age of house
- In a standard regression (I will explain later) the resulting parameter estimates are average relationships across study area
- E.g in Rural areas age of house may have different impact on house prices than in Urban Areas
- Global rather than Local *statistics*

Spatial heterogeneity

- This is caused by the processes generating the sampled data we are studying exhibiting non-stationarity.
- This basically means that the processes generating the observed attributes might vary over space rather than being constant as is assumed in the use of most traditional types of statistical analysis.

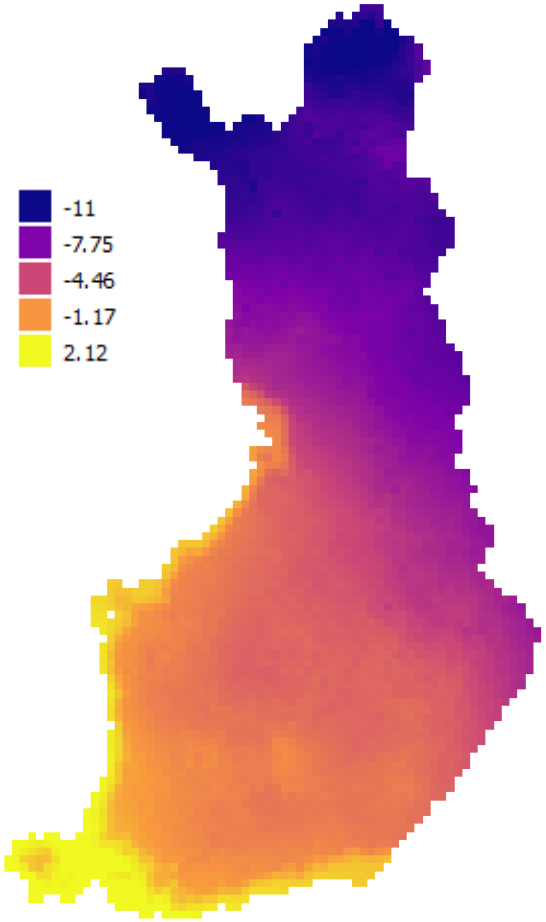
- The global models is usually an average of local models

- Simpson's paradox



Global model example: Mean

- **Mean:**
  - The data: temperature
  - Global value: -4.6



Global model example: Regression

- If our usual formula for a linear regression equation is.
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$
  - y = dependent variable – i.e. one we are trying to explain
  - x = explanatory variable
  - β = parameter indicating the influence of x on y
  - ε = error term
- Goal
  - To estimate the values of the parameters
  - To provide some diagnostics on reliability of parameters
- Important note on the symbology: y and x are variables; not coordinates!

Regression coefficients

- The most important part of the results of regression: Estimates for the coefficients β_1, β_2, …
- How to explain them: Looking at the formula, we can see that if an independent variable x_1 increases by one, the dependent variable y increases by β_1 assuming that all other independent variables remain the same.
- Note that this is slightly different from correlation

Regression

- If our usual formula for a linear regression equation is.
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$

- This equation implies that for each relationship between y and an x variable, a single parameter is estimated which is assumed to be constant across the study region.
- Consequently if there is spatial nonstationarity the resulting single parameter estimate would then represent and average of the different processes operating over space.
- This is the problem of spatial heterogeneity.

Example

Figure 2.1 *London boroughs*

Example

y: house price
x: attributes of a house

Figure 2.2 *Location of sampled house prices*

Example

Table 2.2 *Global regression parameter estimates*

Variable	Parameter estimate	Standard error	T value
Intercept	58 900	2 524	23.3
FLRAREA	697	14	49.3
BLDPWW1	−2 340	601	−3.9
BLDPOSTW	−2 786	903	−3.1
BLD60S	−5 177	1 043	−5.0
BLD70S	−2 421	1 174	−2.1
BLD80S	6 315	920	6.9
TYPDETCH	−4 215	4 038	−1.0
TYPTRRD	3 465	2 120	1.6
TYPBNGLW	23 437	6 156	3.8
TYPFLAT	3 239	2 057	1.6
GARAGE	5 956	564	10.6
CENTHEAT	7 777	627	12.4
BATH2	22 297	1 166	19.1
PROF	72	24	3.0
UNEMPLOY	−211	38	−5.5
FLRDETCH	205	27	7.5
FLRFLAT	−123	22	−5.6
FLRBNGLW	−87	65	−1.4
FLRTRRD	−119	19	−6.2
$\log_e$(DISTCL)	−18 137	604	−30.1

Note: $R^2 = 0.60$

Goodness of fit

- How well the model fits the data; the higher the value the better the model
- R^2 value
  - Percentage of variance in the y variable accounted for by the variance in the model
  - In this case 60% of the variations in house prices are explained by the model
  - This lead 40% unexplained

Example

Table 2.2 *Global regression parameter estimates*

Variable	Parameter estimate	Standard error	T value
Intercept	58 900	2 524	23.3
FLRAREA	697	14	49.3
BLDPWW1	−2 340	601	−3.9
BLDPOSTW	−2 786	903	−3.1
BLD60S	−5 177	1 043	−5.0
BLD70S	−2 421	1 174	−2.1
BLD80S	6 315	920	6.9
TYPDETCH	−4 215	4 038	−1.0
TYPTRRD	3 465	2 120	1.6
TYPBNGLW	23 437	6 156	3.8
TYPFLAT	3 239	2 057	1.6
GARAGE	5 956	564	10.6
CENTHEAT	7 777	627	12.4
BATH2	22 297	1 166	19.1
PROF	72	24	3.0
UNEMPLOY	−211	38	−5.5
FLRDETCH	205	27	7.5
FLRFLAT	−123	22	−5.6
FLRBNGLW	−87	65	−1.4
FLRTRRD	−119	19	−6.2
$\log_e$(DISTCL)	−18 137	604	−30.1

Note: $R^2 = 0.60$

T-values

- Problem of zero valued parameters
- Standard error - a measure of uncertainty of a parameter estimate
- Dividing each parameter estimate by its standard error – T-value

Statistical significance test

- Null hypothesis: the value of a parameter is zero
- Calculate of t-value
- Compare with 1.96
- If T > 1.96 or T < -1.96, reject the null hypothesis and conclude that the parameter value is significantly different from zero

- Also usually p-values – much easier to interpret as they tell you % significance
  - E.g 0.01 means the estimate is significant at 1%

Residuals and spatial autocorrelation

- Residuals: Difference between the observed and predicted values
  - Positive – underestimation
  - Negative – overestimation

- Residuals are not randomly distributed, but they are spatially autocorrelated

- Assumption in regression: Residuals should be randomly distributed; if they are not, there's a problem with the model. Locals models can be a solution to the problem.

Example

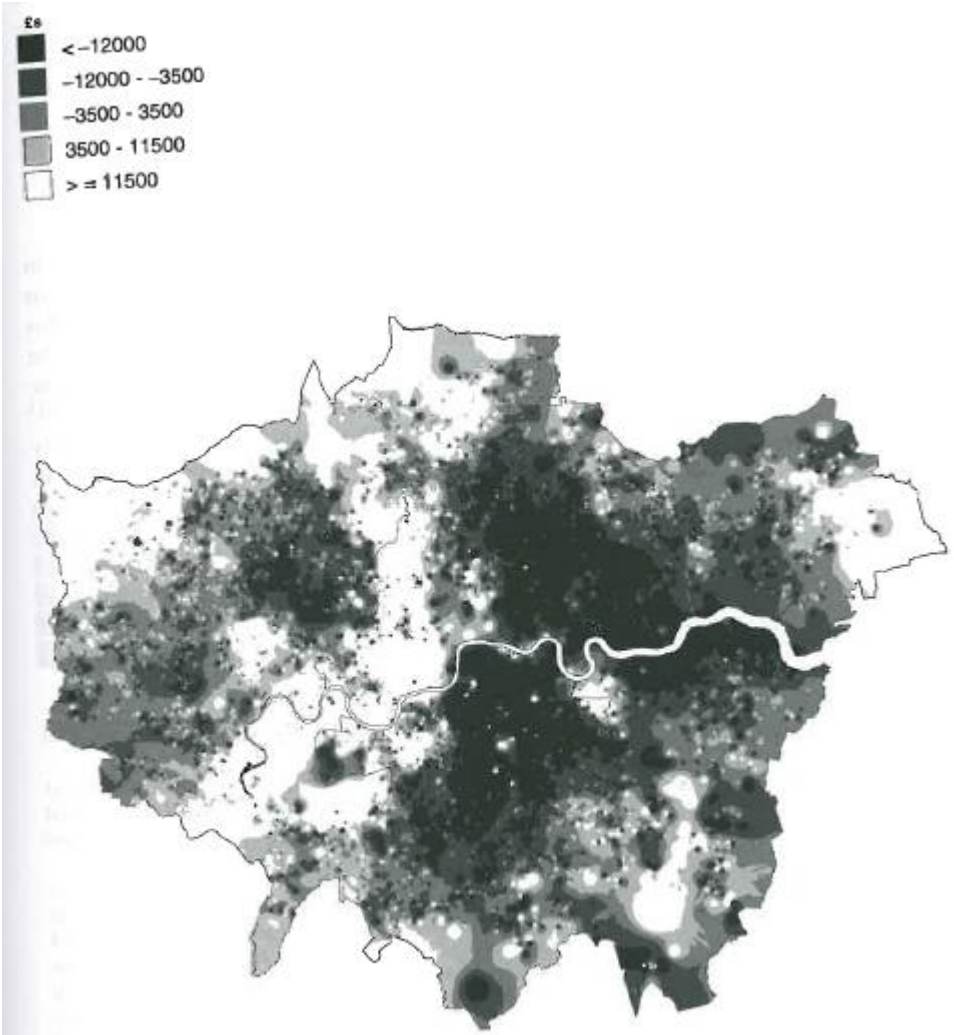


Figure 2.4 *Global model residual surface*

Local models

- So how can we account for these non-stationary relationships?
- We want to look at methods that will reveal individual parameters for locations across the study area to see if they vary significantly
- Simplest way is to run a regression for many subareas

Example: Mean

- **Mean for every region**

Example: Regression

Figure 2.1 *London boroughs*

Results

Table 2.5 *Price/m² (£) of flats and terraced housing in each London borough from separate calibrations of the global hedonic model*

Borough	Price/m² (£) Flat	Terraced	Ratio Terrace/Flat	R^2
Barking	310	609	1.96	0.70
Barnet	528	579	1.10	0.75
Bexley	106	80	0.75	0.86
Brent	263	310	1.18	0.73
Bromley	399	427	1.07	0.83
Camden	897	179	0.20	0.69
City	***	***	***	***
Croydon	329	216	0.66	0.83
Ealing	464	350	0.75	0.63
Enfield	326	615	1.89	0.85
Greenwich	629	611	0.98	0.53
Hackney	432	612	1.42	0.71
Hammersmith	524	1 272	2.43	0.82
Haringey	543	623	1.15	0.73
Harrow	233	444	1.91	0.47
Havering	104	555	5.34	0.67
Hillingdon	265	270	1.02	0.71
Hounslow	513	733	1.43	0.65
Islington	595	889	1.49	0.80
Kensington	1 574	2 019	1.28	0.75
Kingston	141	605	4.29	0.81
Lambeth	350	606	1.73	0.72
Lewisham	268	513	1.91	0.76
Merton	517	554	1.07	0.64
Newham	267	249	0.93	0.56
Redbridge	420	518	1.23	0.77
Richmond	866	713	0.82	0.75
Southwark	667	498	0.75	0.72
Sutton	311	572	1.84	0.82
Tower Hamlets	628	381	0.61	0.79
Waltham Forest	257	320	1.25	0.80
Wandsworth	563	780	1.39	0.68
Westminster	626	1 672	2.67	0.64

*** Insufficient data to calibrate model

Results

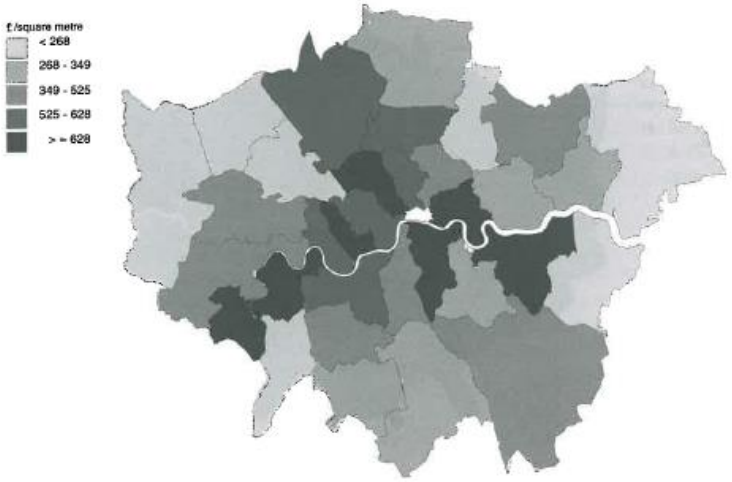


Figure 2.5 Price/m² (£) of flats estimated from separate regressions for each London borough

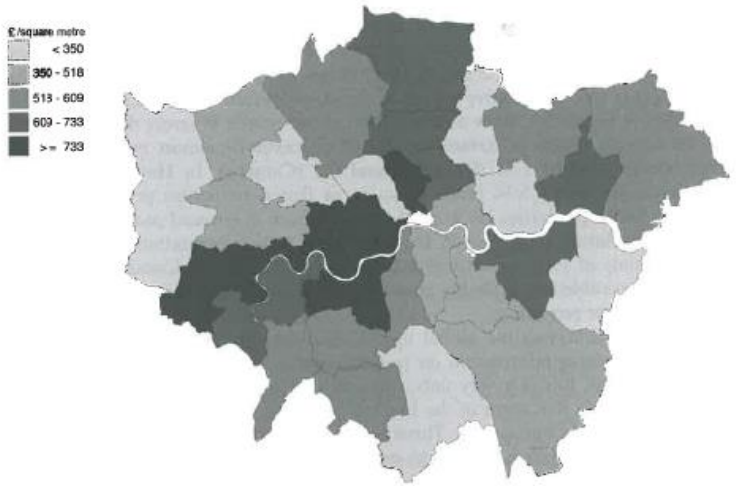


Figure 2.6 Price/m² (£) of terraced properties estimated from separate regressions for each London borough

What is wrong with this approach?

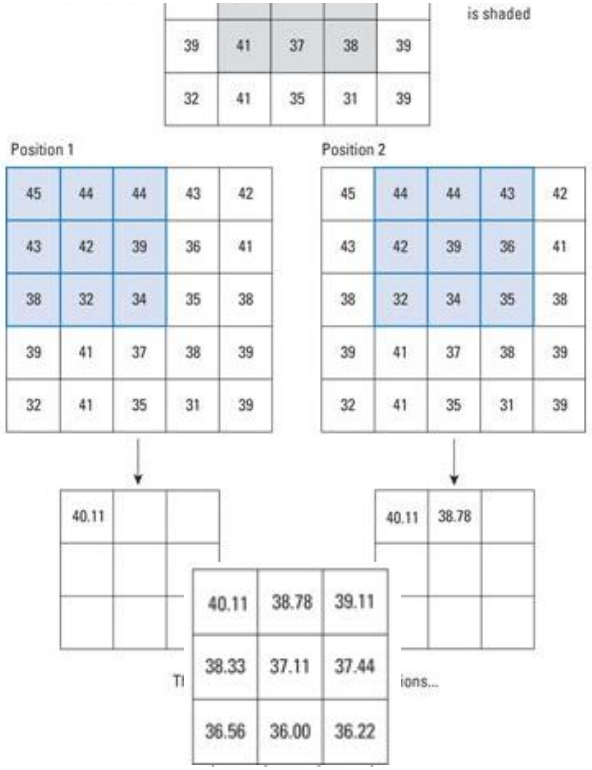
Problems with subareas approach

- The division is artificial, not based on the phenomenon.
- MAUP (Modifiable Areal Unit Problem): Different division yields different results
- Objects on different sides of a border belong to different models

Moving window approach

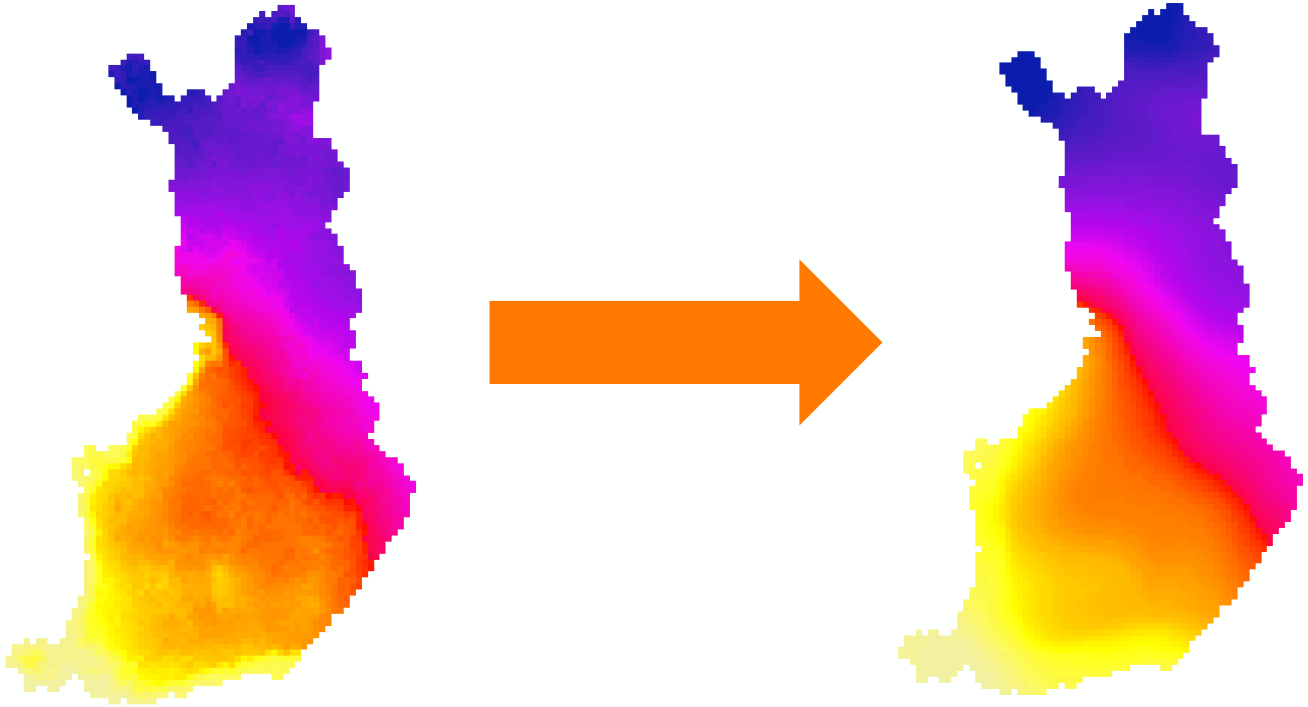
- A moving window would overcome the problem that the area boundaries are not necessarily the boundaries of the underlying spatial process
- Methods works by
  - Creating a grid of points
  - Creating an area(square or circular) around each grid point
  - Create a separate model for each point on just the data that falls inside the area of selection
  - Map the results

Moving window mean



Moving window mean

- Mean in the area of 10 x 10 pixels (100 km x 100 km)



Moving window regression

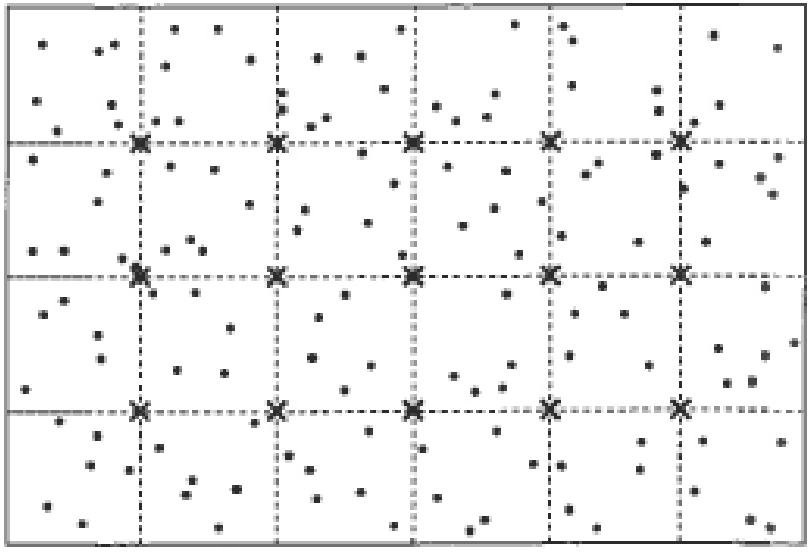


Figure 2.8 *An example of moving window regression*

Example

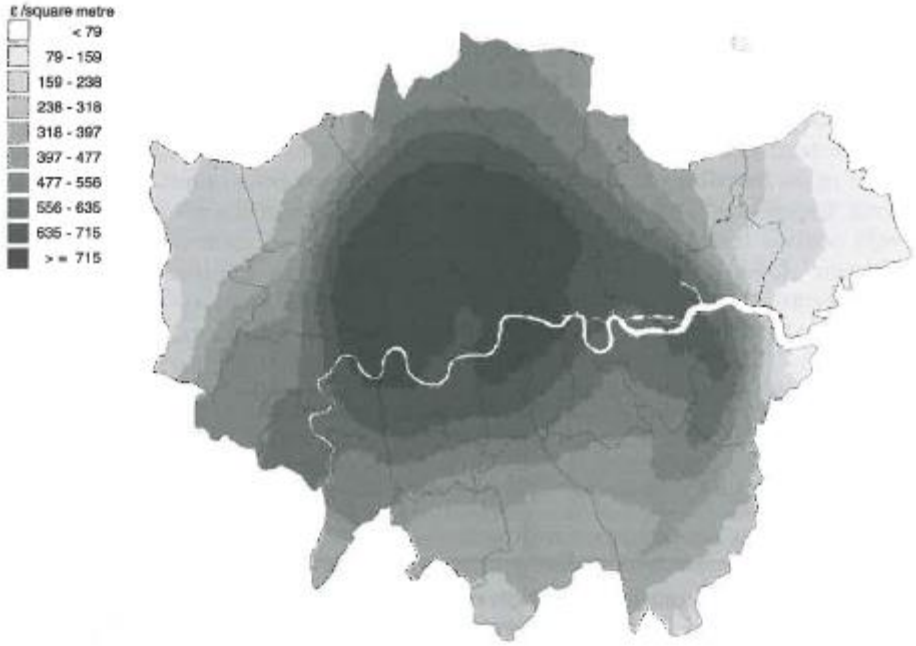


Figure 2.9 *Value of flatted properties (£/m²) from a moving window regression*

What is wrong with this approach?

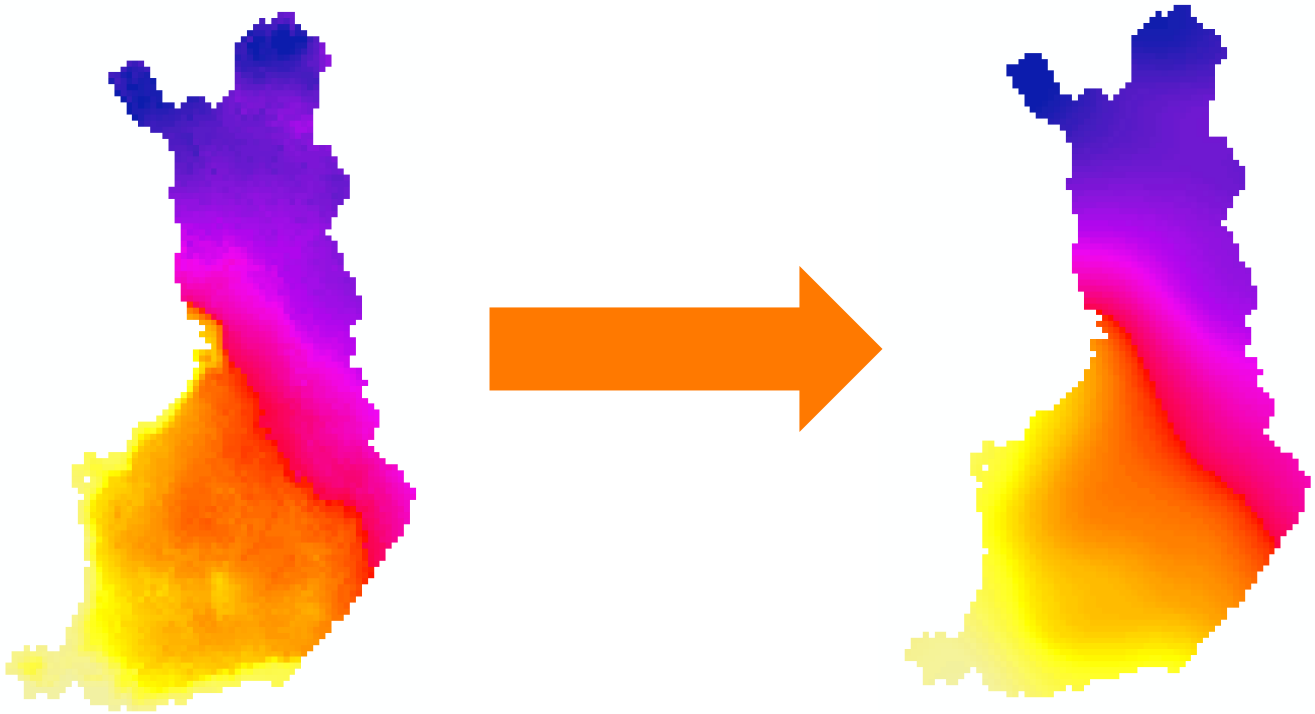
Problems with moving window approach

- The division is not artificial anymore, but the problem with the border remains

Geographically weighted models

- Removes the problem with the border: utilize gradual weights for points
- Geographically weighted model: Create a statistical model for each point so that nearby points are given a bigger weight than faraway points.
- Can be applied to every (?) statistical method that has a weighted version
  - Ready tool for at least the following: Descriptive statistics (mean, median, variance, covariance, correlation), different regression models (linear, Poisson, logistic, heteroscedastic, robust), discriminant analysis, principal component analysis (PCA; next lecture), parallel coordinate plot (PCP)
- Software: R, GWR4, ArcGIS, QGIS

Geographically weighted mean



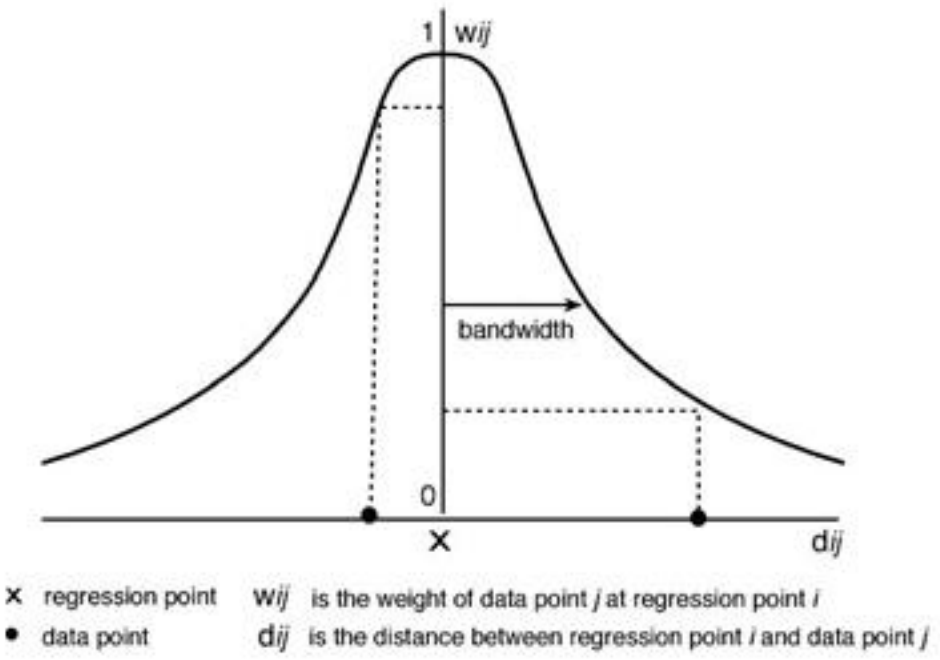
Geographic Weighted Regression (GWR)

- The geographic weighted version of the regression equation is.
  - $y_i = \beta_{0i} + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \cdots + \beta_{ni}x_{ni} + \varepsilon_i$
- Where *i* refers to the location at which data on y and x are measured and at which local estimates of the parameters are obtained.
- The regression equation is now influenced by a spatial weights matrix where observations nearer to *i* are given more weight than observations further away.

GWR

- $y_i = \beta_{0i} + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \cdots + \beta_{ni}x_{ni} + \varepsilon_i$
- Parameters β_{ni} are not constant
- Parameters are function of location
- Weights – closer observations influence the regression more than distant ones
- Weighting can be done using a Gaussian or similar decreasing function
    - For example: $w_{ij} = \left[1 - (d_{ij}/b)^2\right]^2 \; if \; d_{ij} < b$

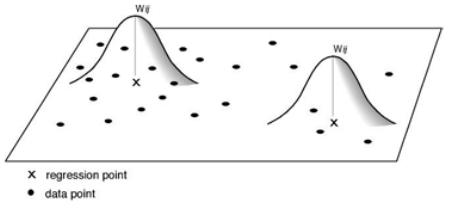
Spatial Weighting Function



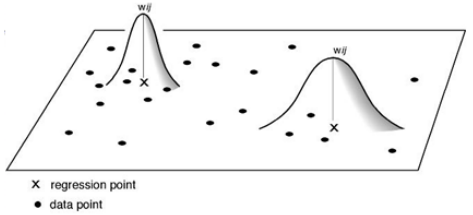
From Fotheringham et al., 2002

Weighting Schemes

- Fixed



- Spatially adaptive



From Fotheringham et al., 2002

Choice of weighting function

- Results of GWR are relatively insensitive to the choice of weighting function
- Results are sensitive to the bandwidth or the number of the nearest neighbours
  - Too small – large variance in the estimates
  - Too large – large bias
- Bandwidth size is optimized using either
  - Crossvalidation score or
    - $CV = \sum_i [y_i - y_{\neq i}^* (h)]^2$
  - Akaike Information Criterion
    - AIC_c = Deviance + 2k [n/(n-k-1)]

Example

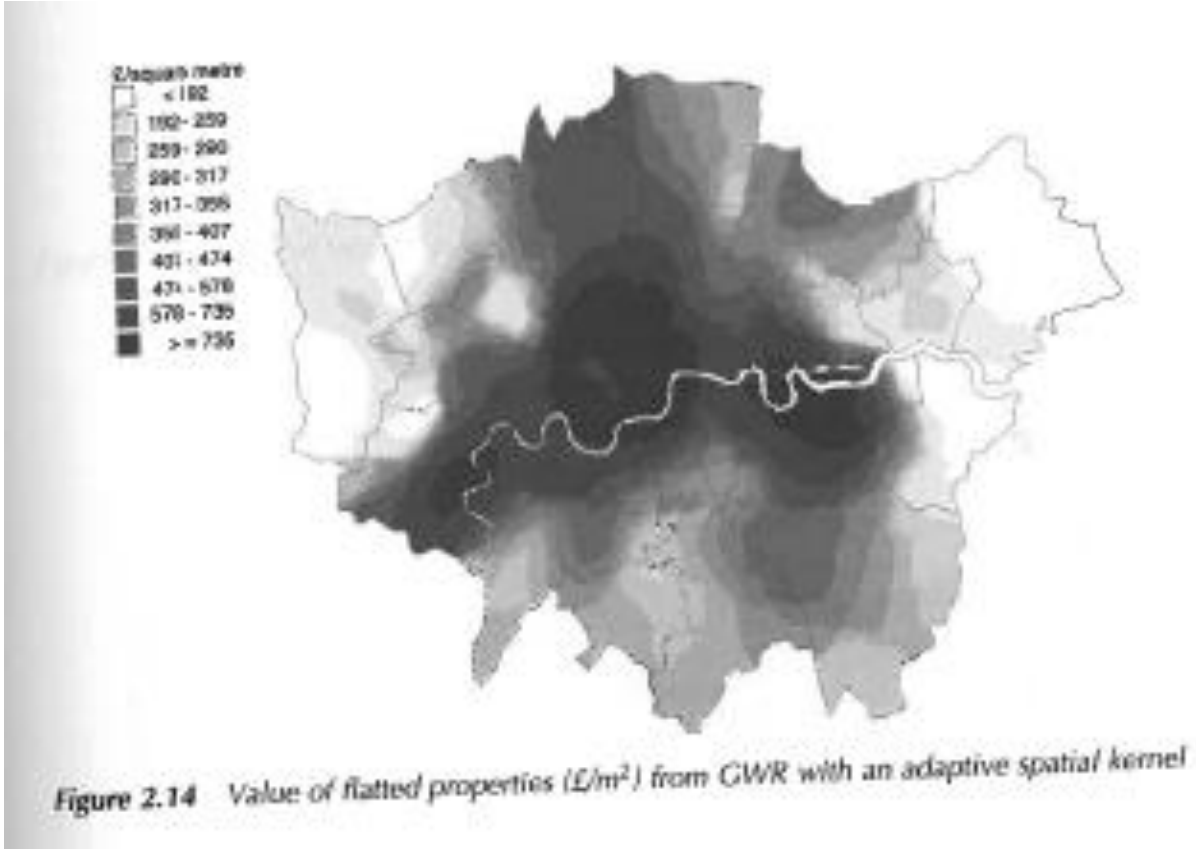


Figure 2.14 *Value of flatted properties (£/m²) from GWR with an adaptive spatial kernel*

Example

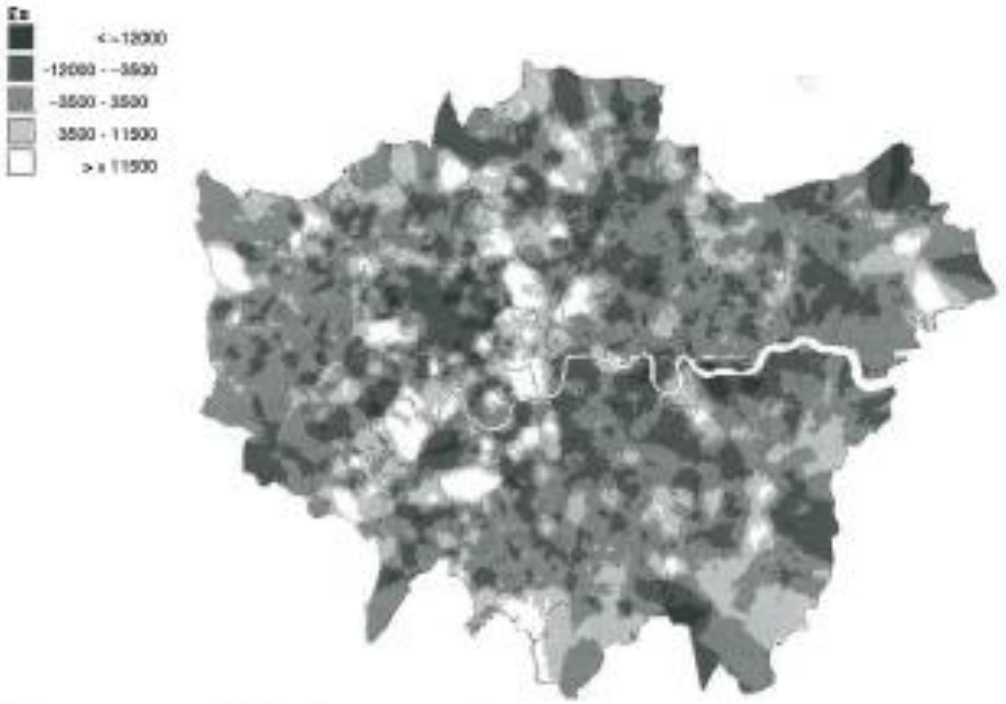


Figure 2.19 *Residuals from the GWR model*

Conclusion: General flow of GWR analysis

(0. Data preprocessing and descriptive analysis)

1. Run global regression model; for example Ordinary Least Squares (OLS) regression
2. Analyze the results; coefficients, R-squared, residuals
   - Run Moran's Index for the residuals; if they are autocorrelated, GWR might be a solution
3. Run the same GWR model
4. Analyze the results; local coefficients, local R-squared, model R-squared, residuals
   - Run Moran's Index for the residuals; did the level of autocorrelation decrease?
5. Compare OLS and GWR results: which model is better, are the results conflicting?

Concluding Remarks

- Information on spatial non-stationarity in relationships

- Residuals from GWR are generally lower compared to global models
- Residuals from GWR are generally less spatially autocorrelated

- ‘Spatial Microscope’ – observing relationships on various scales, i.e. using different bandwidth sizes

Example: Building fires in Helsinki (slides from Olga Špatenková)

- Incident data from Fire and Rescue authorities in Helsinki (2005-2007)
  - geocoded by addresses and coordinates
  - attributes (incident type, response time, etc.)
- Census data from Statistics Finland (2006)
  - aggregated to 250 x 250 m grid
  - plenty of socio-economic figures

Data preparation

- Kernel density for incidents data
- Cells with no data excluded from analysis
- Map overlay according to census grids

Data selection for the model

- Dependent variable
  - building fire density (kernel density)
- Independent variables
  - building age
  - building type (?)
  - population density (count/cell)
  - workplace density (count/cell)
  - households with kids (ratio)
  - households with adults only (ratio)
  - households with pensioners (ratio)
  - average income (€)
  - education
  - unemployment (ratio)

Global Model

AIC	11855.9		
Adjusted r^2	0.43		
Coeff. of determination	0.43		
Parameter	Estimate	St Err	T
Intercept	0.5e-1	9.2e-1	0.1
Building age	-0.6e-4	4.5e-4	-0.1
Building type	**5.7e-1**	**1.6e-1**	**3.7**
Pop. dens.	**82.8e-4**	**3.6e-4**	**23.2**
Workplace dens.	**35.1e-4**	**2.6e-4**	**13.7**
Children	**-3.3e+0**	**1.0e+0**	**-3.2**
Adults	1.8e+0	1.0e+0	1.8
Pens	-3.0e-1	9.7e-1	-0.3
Incomes	2.9e-6	3.7e-6	0.8
Education	-0.2e+0	1.1e+0	-0.1
Unemployment	**4.9e+0**	**1.9e+0**	**2.6**

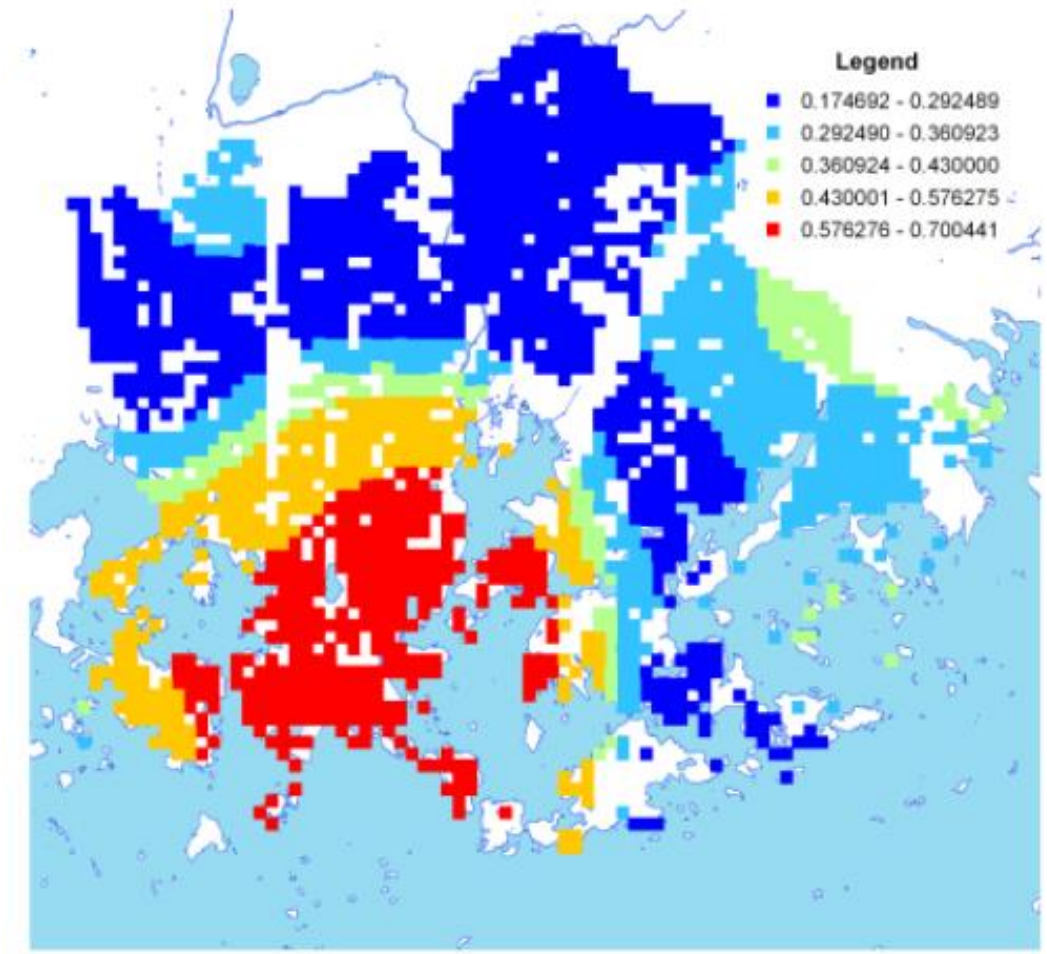
- Moran's index for residuals missing

GWR Model

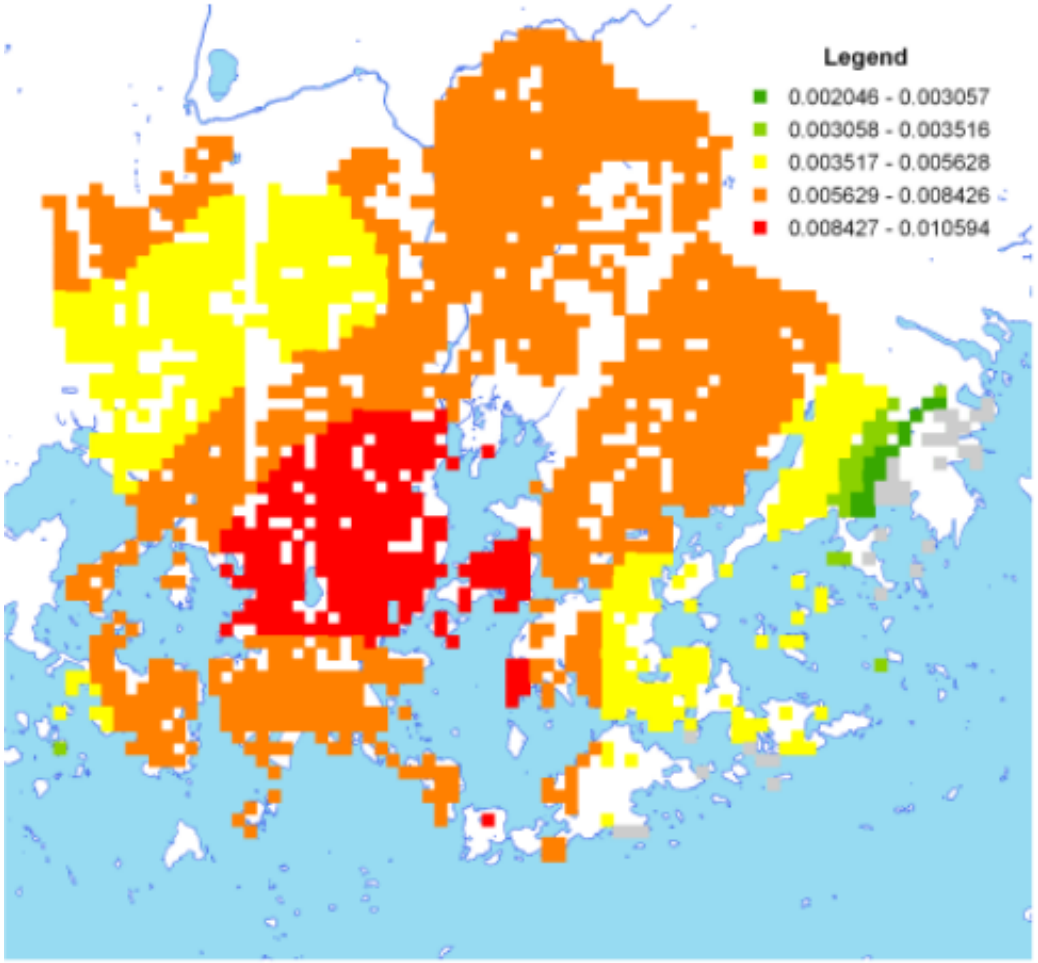
Nr. of observations	2121
Bandwidth (in data units)	1909.7
Effective nr. of parameters	113.6
AIC	11656.0
Adjusted r^2	0.51
Coeff. of determination	0.53

- Moran's index for residuals missing

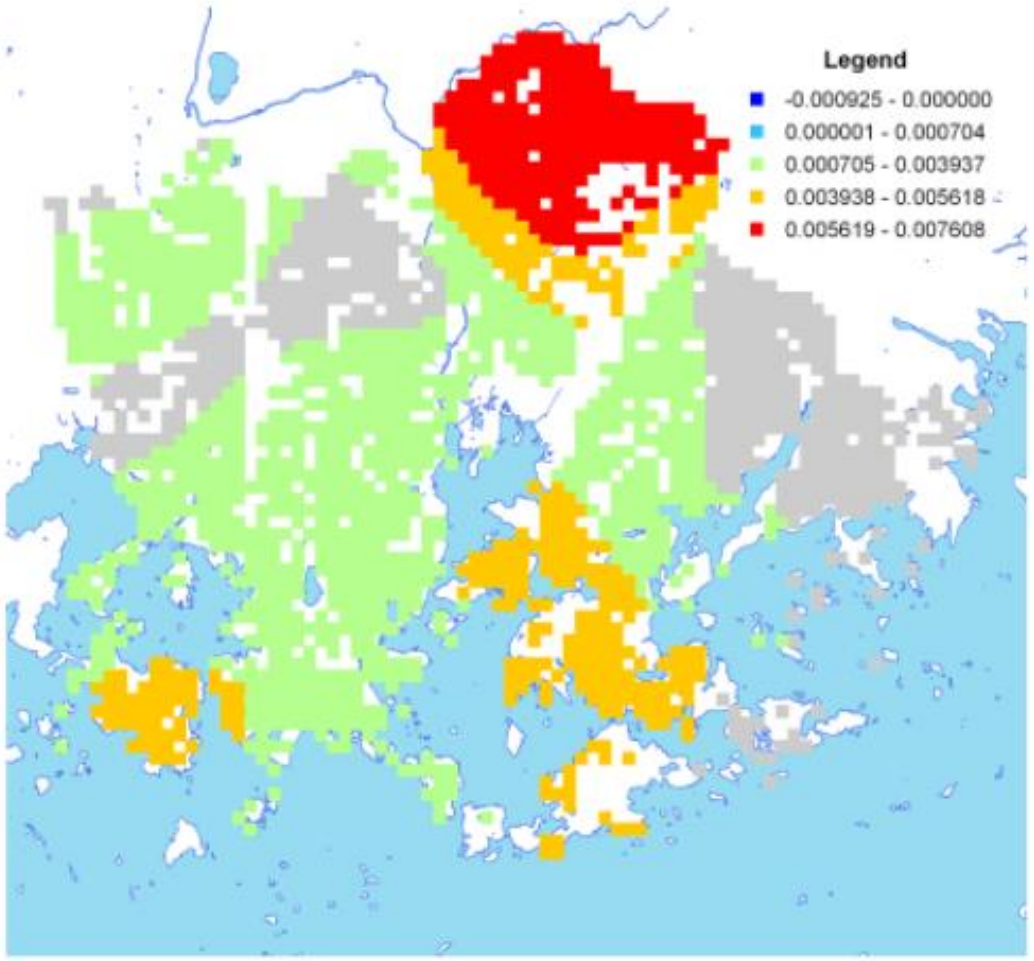
Local R-squared



Local coefficients for population density

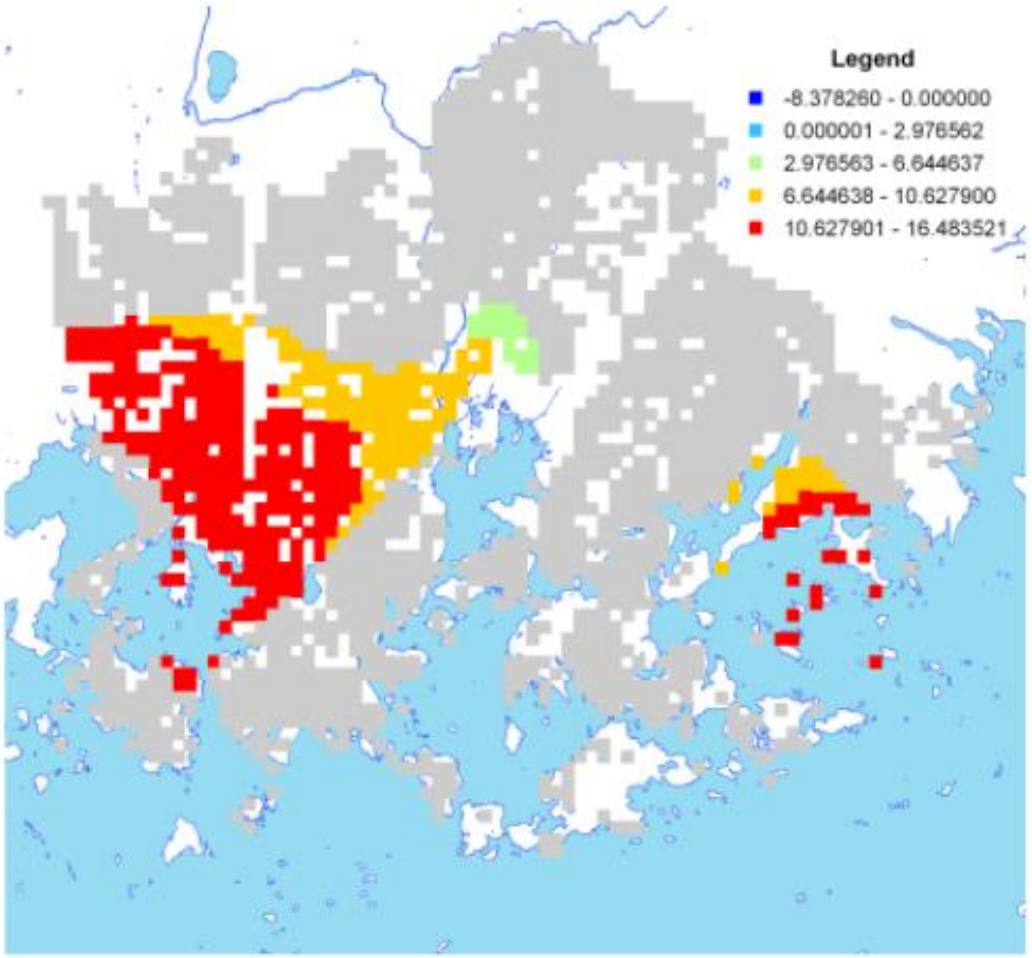


Local coefficients for workplace density



Grey= non-significant areas

Local coefficients for unemployment



Note: This studies the relationship between unemployment and building fires assuming that the income (and other attributes) stays the same!

Conclusions

- Quantification of the relations between studied variables; coefficients!
- Spatial variations in the relations!
- Necessary interpretation of parameter values and t-values at the same time

Regression as a prediction model

- The method presented this far shows how to use GWR in its main functionality: **explaining relationships** between variables (explanatory analysis) using the coefficient parameter estimates

- You can also use GWR in **prediction** (statistical inference): First, create a regression model with a data set.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

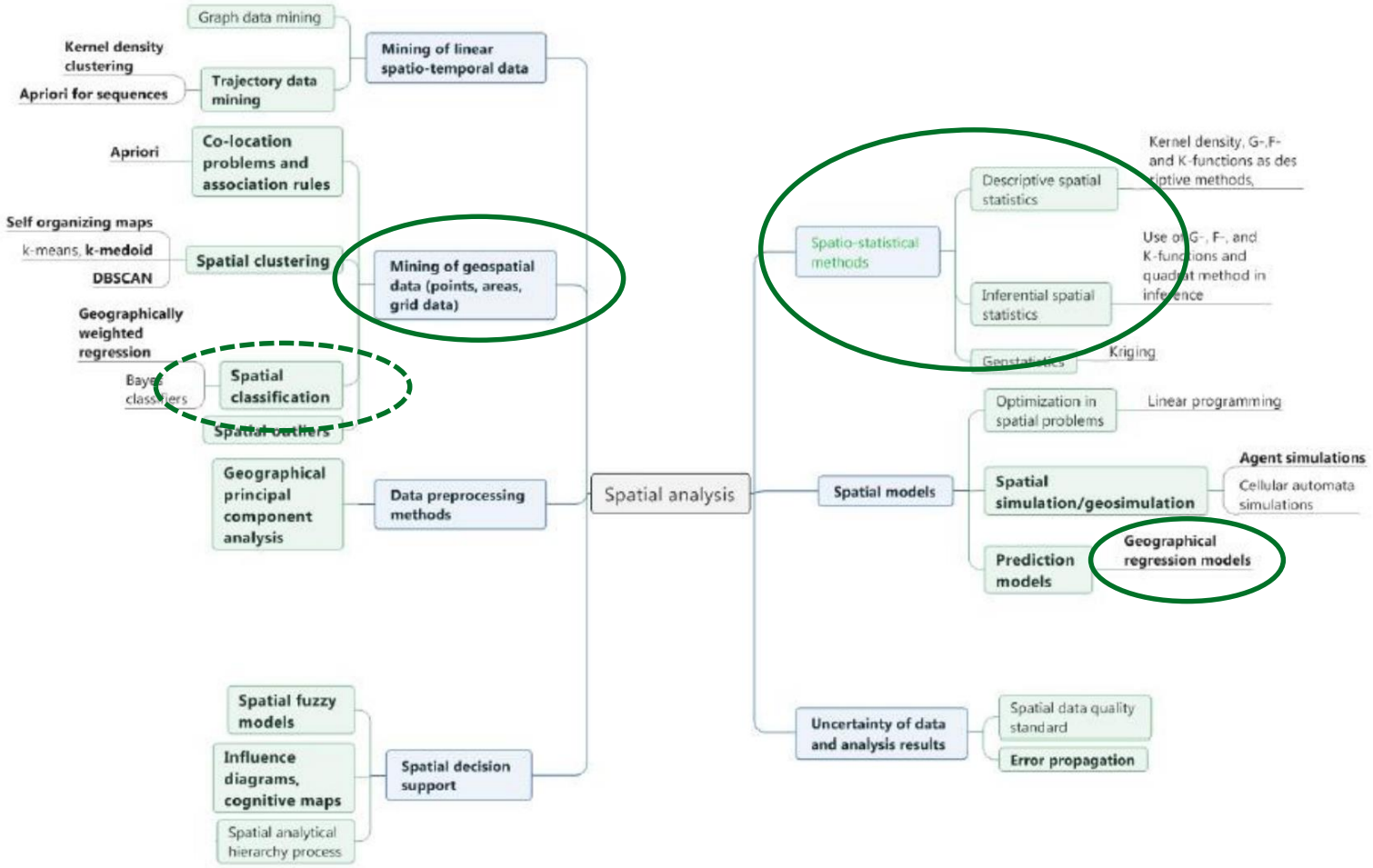
Then, you can calculate an unknown y using known x's

PROS x CONS

- Well established statistical method
- Multivariate analysis
- Suitable for spatial data
- Good connection of the results to a map
- Visualization of the results
- Spatial incorporation; a true spatial model
- Can be used for describing the relationships as well as prediction

- Pre-processing can be time consuming
- Tedious temporal analysis
- Ongoing discussion on exact statistical viability

The big picture



Thank you!