Geographically Weighted Regression

Mark Green

DASC507 – Advanced Biostatistics II

Analysis Methods for Complex Data Structures

Outline

- Linear regression – please not another recap
- Spatial non-stationarity – why it matters
- Geographically Weighted Regression (GWR)
  - Spatial kernel
  - Bandwidth
- Criticisms
- Case study example

Linear regression – please not another recap

Outcome of interest

Coefficient applied to X

$$y = \alpha + \beta X + \mathrm{e}$$

Constant

Explanatory variables

Error term

Linear regression – please not another recap

- Beta coefficients for explanatory variables give an ‘overall’ or ‘average’ association – helpful for describing an association
- This assumes that the influence of a variable on an outcome is consistent across varying contexts
- Parameters may not always be ‘stationary’, but may be non-stationary when capturing heterogenous patterns
- Time-series is a common form of non-stationarity, as is spatial patterns

Spatial non-stationarity – why it matters

Conceptual reasons

- A large body of literature shows that neighbourhoods matter for health (e.g., air quality, access to services)
- Neighbourhoods or places imprint local effects (e.g., resilience, community support, local resources, place-unique issues)
- Context vs composition matters – but hard to separate out

Spatial non-stationarity – why it matters

Methodological reasons

- Global models cannot incorporate local differences between places easily
- Local processes/effects/modifiers hard to measure
- Non-stationarity may simply just be omitted variables (but still need to account for possibility)
- Multi-level models might help but don't acknowledge space explicitly

Geographically Weighted Regression (GWR)

- Extension of linear (OLS) regression model
- Allows for spatially varying coefficients
- Demonstrates how associations between explanatory variables to an outcome variable may differ by location both in strength of association and direction of relationship
- Can be used for individual- and area-level data (must have spatial location for both)
- A data-driven approach for estimating parameters

Geographically Weighted Regression (GWR)

GWR works through:

- For each observation (individual or area)
  - Select surrounding/neighbouring observations based on a search window
  - Run a regression model on just these observations (local regression model)
- Repeat process for each data point (i.e., n regression models are run)
- Compare to global regression model of all data points
- Plot regression coefficients
- Smile

Geographically Weighted Regression (GWR)

Our modified OLS equation becomes:

$$y_i(u) = \alpha_i(u) + \beta_i(u)X_i + e_i$$

Where *i* is observation and *u* is 'conditioned on place' (vector of co-ordinates). To estimate β, we use this magic:

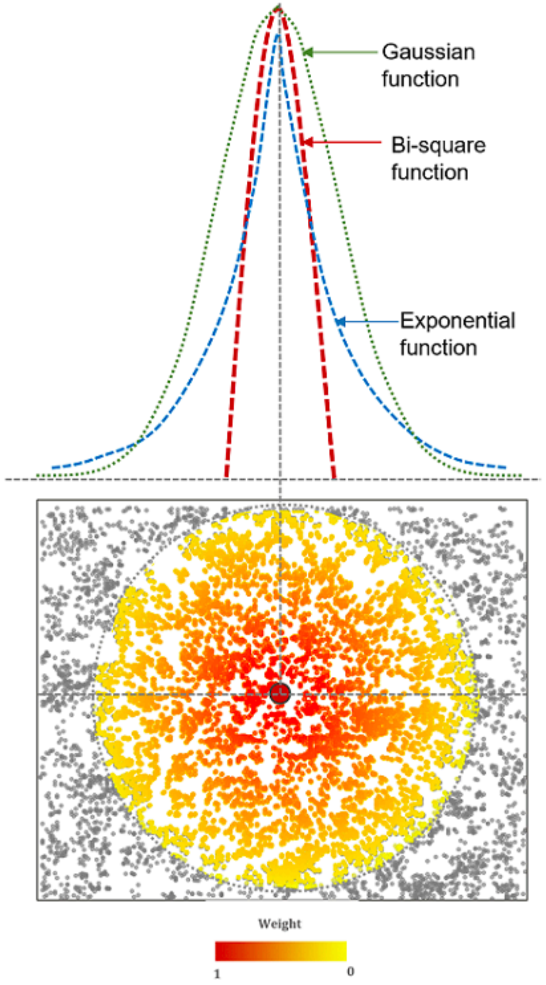
$$\hat{\beta}(u) = (X^T W(u) X)^{-1} X^T W(u) y$$

Where *W* is our spatial weights again. This is similar to a weighted least squares global model (just conditioned on location)

Geographically Weighted Regression (GWR)

To fit local regression models, we need to define the spatial kernel

- Defines how to weight observations
- Data located close to the data point being estimated are given greater importance -> remember Tobler's first law of Geography?
- Need to define how weighting changes with distance
- Weights sum to 1



Geographically Weighted Regression (GWR)

There are loads of kernels out there, although the choice doesn't always make a big difference in reality

- Bisquare $w_i(u) = (1 - z^2)^2$
- Epanechnikov $w_i(u) = 1 - z^2$

For both $z = (x_i - x_0)/h$ and *z* = 0 for x_i - x_0 > *h* (i.e., cut point)

Where x_i is the observation being evaluated for inclusion, x_0 is the location being modelled, *h* is bandwidth. Tl;dr distance between observations divided by max distance to select observations from.

Geographically Weighted Regression (GWR)

Continuous / Gaussian approaches exist. For example

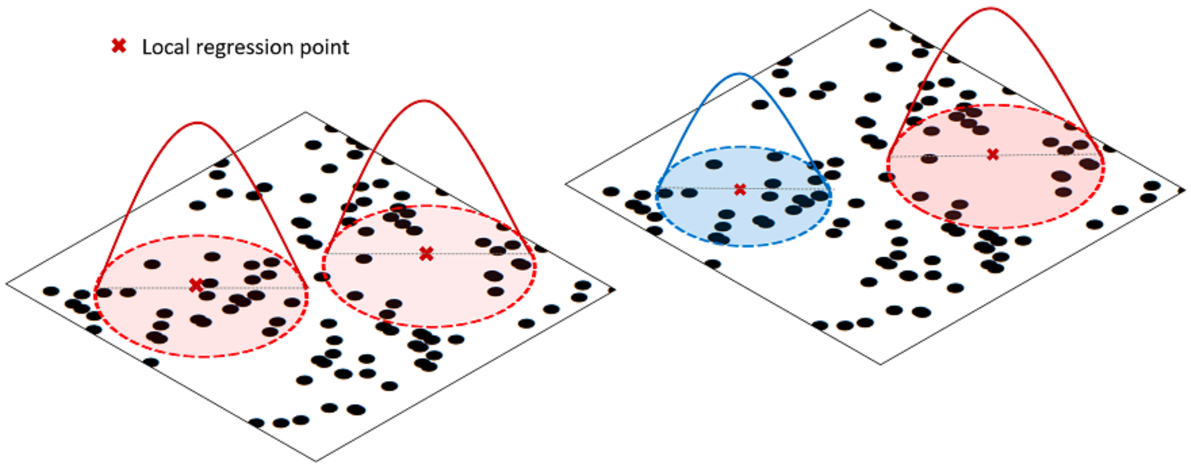
- Infinite range $w_i(u) = \exp(z^2/2)$

These approaches may not have a cut-point and therefore all data points are used in model estimation (with far away points contributing very little).

Geographically Weighted Regression (GWR)

We must also select a bandwidth

- Defines extent of observations to be considered for the local regression
- Bandwidth defines the area to be covered
  - Fixed: same distance used for each regression
  - Adaptive: varying bandwidths are used for each regression, depending on local parameters



Fixed kernel (constant distance across the study area)

Adaptive kernel (constant number of nearest neighbors across the study area)

Geographically Weighted Regression (GWR)

Choice of bandwidth is more difficult to get right and can have a large impact on your results. Important trade off between

- Bias
- Variance

Bandwidths need to be large enough to capture enough data to generate reliable estimates (precision depends on *n*), but not too large that estimates don’t reflect spatial patterns.

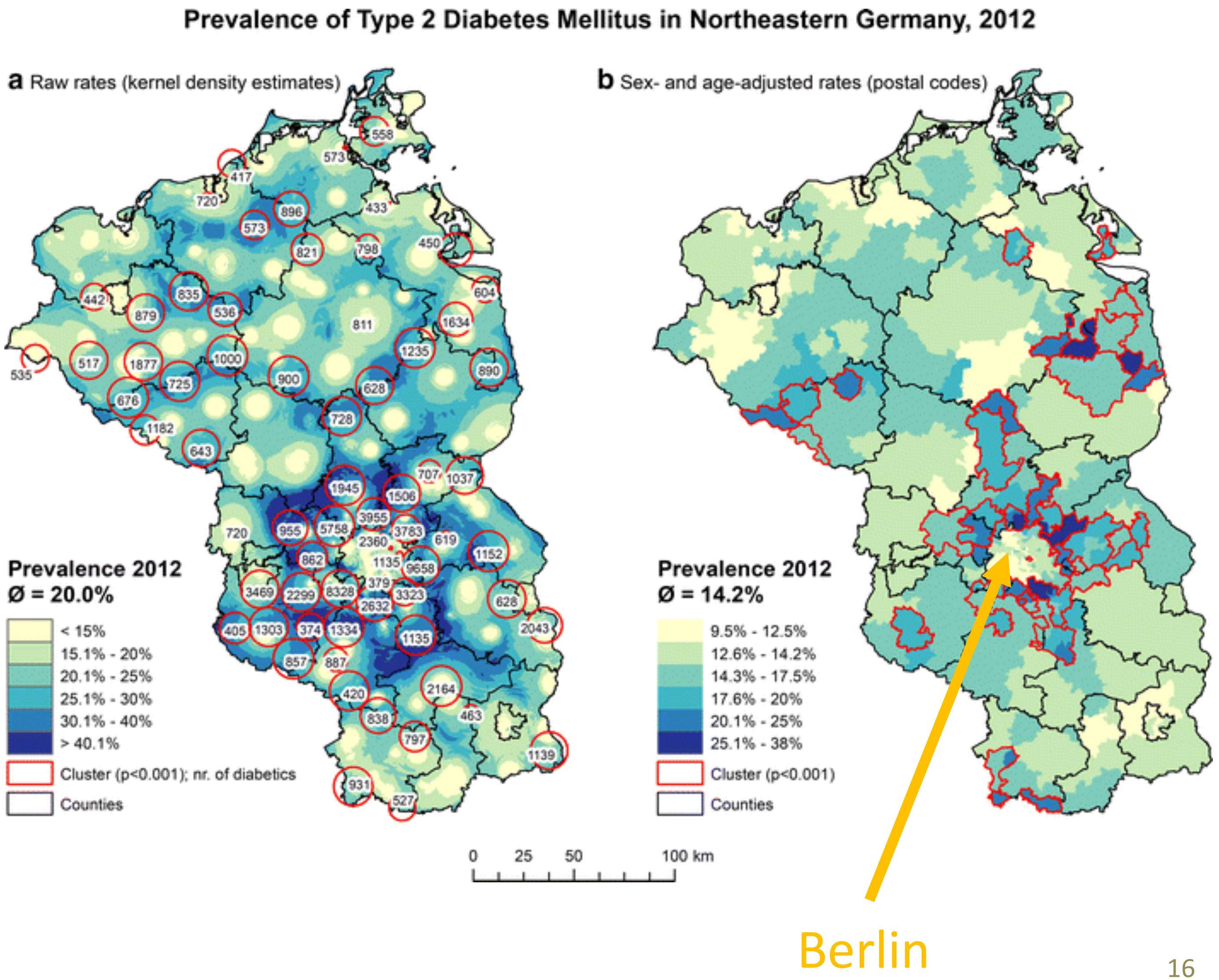
Can be ‘optimized’ using rule-based approaches or cross-validation

Criticisms

- Local multicollinearity can be problematic for local coefficients
- Simulation studies suggest GWR cannot always detected underlying spatial patterns in data or find patterns when there are none
- Difficulty in validating results/models (best as exploratory method)
- Model results can be unstable and not always replicable with small samples sizes (try re-running the model in the practical again and again)
- Bigger sample sizes help with robustness, but at the expense of computational time
- Multivariate models can be difficulty to get enough data points with enough variability in characteristics at smaller bandwidths

Case Study

Kahul et al. 2016. Do the risk factors for type 2 diabetes mellitus vary by location? A spatial analysis of health insurance claims in Northeastern Germany using kernel density estimation and geographically weighted regression. *International Journal of Health Geographics* **15**: 38.



Case Study

Variable	Coefficient	VIF
Intercept	2.259540***	
Persons aged 65–79 (%)	0.027251***	1.656689
Persons aged 80 and older (%)	0.010704**	1.650654
Unemployed persons aged 55–65 (%)	0.013354***	2.593295
Employed persons (%)	−0.006181**	1.602619
Mean income tax	0.000780**	2.272369
Non-married couples (%)	0.014524*	1.45273
Adjusted R2	0.44	
AICc	−313	
Global Moran’s I of residuals	I = 0.264 (p < 0.001)	

Global OLS Regression model

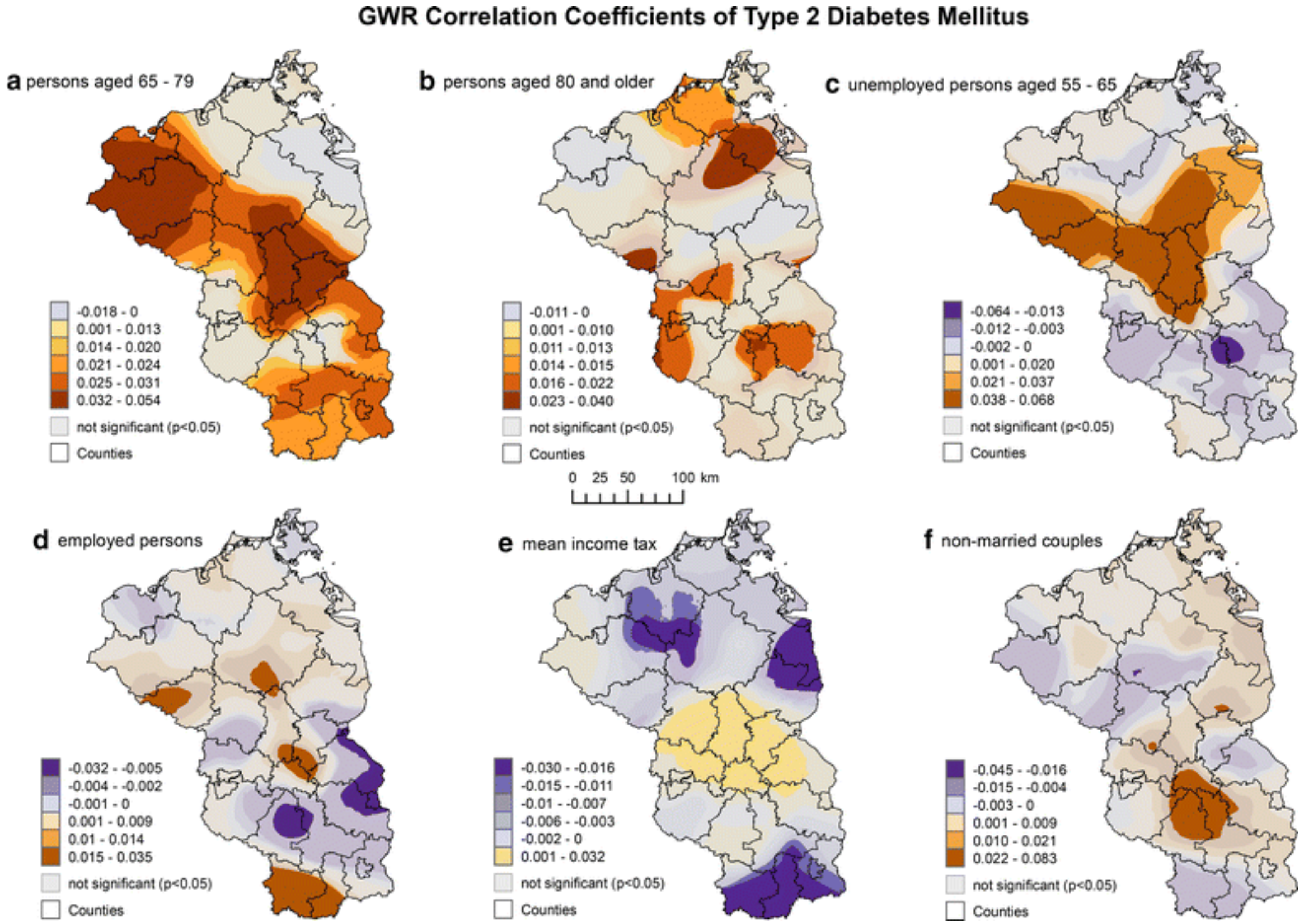
Case Study

Comparing the effects of optimisation, kernel and bandwidth choices, as well as if model residuals remain spatially clustered

Model	AICc	Adjusted R^2	Moran’s I of residuals
Adaptive, Gaussian, AICc	−347	0.51	$p < 0.001$
Adaptive, Gaussian, AIC	−347	0.51	$p < 0.001$
Adaptive, Gaussian, BIC	−315	0.44	$p < 0.001$
Adaptive, Gaussian, CV	−347	0.51	$p < 0.001$
Fixed, Gaussian, AICc	−385	0.62	$p < 0.05$
Fixed, Gaussian, AIC	−265	0.66	$p > 0.05$
Fixed, Gaussian, BIC	−316	0.44	$p < 0.001$
Fixed, Gaussian, CV	−370	0.64	$p > 0.05$
Adaptive, bi-square, AICc	−394	0.63	$p < 0.001$
Adaptive, bi-square, AIC	−374	0.66	$p > 0.05$
Adaptive, bi-square, BIC	−320	0.45	$p < 0.001$
Fixed, bi-square, AICc	−385	0.62	$p < 0.01$
Fixed, bi-square, AIC	40	0.68	$p > 0.05$
Fixed, bi-square, BIC	−316	0.44	$p < 0.001$

Case Study

Mapping coefficients



Further reading

- Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). "Geographically weighted regression: a method for exploring spatial nonstationarity". *Geographical analysis* **28**(4): 281-298.
- Comber, A., et al. (2020). The GWR route map: a guide to the informed application of Geographically Weighted Regression. *arXiv* https://arxiv.org/abs/2004.06070.
- Nakaya, T., Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2005). "Geographically weighted Poisson regression for disease association mapping". *Statistics in medicine* **24**(17): 2695-2717.
- Páez, A., Farber, S., & Wheeler, D. (2011). "A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships". *Environment and Planning A* **43**(12): 2992-3010.