science foundation ireland
fondúireacht eolaíochta éireann

StratAG
Strategic Research in Advanced Geotechnologies
Taighde Straitéiseach sna hArd-Gheoteicneolaíochtaí

Geographically Weighted Regression

Modelling spatially heterogenous processes

Martin Charlton
National Centre for Geocomputation
National University of Ireland Maynooth

Outline

- Introduction
- Spatial Data
- Geographically Weighted Regression
  - Weighting schemes
  - Calibration
  - Interpretation and inference
  - Issues
- Developments and current work

Getting wet

- It is well known that locations with increased altitudes have higher rainfall.
- In 1921 Mortyn de Carle S Salter tabulated variations in the relationship between rainfall and altitude in different parts of Britain to conclude that:

> This fact alone renders it impossible to compute any general formula for the increase of rainfall with elevation, and makes it necessary to study every record in the light of the configuration of the surrounding country.

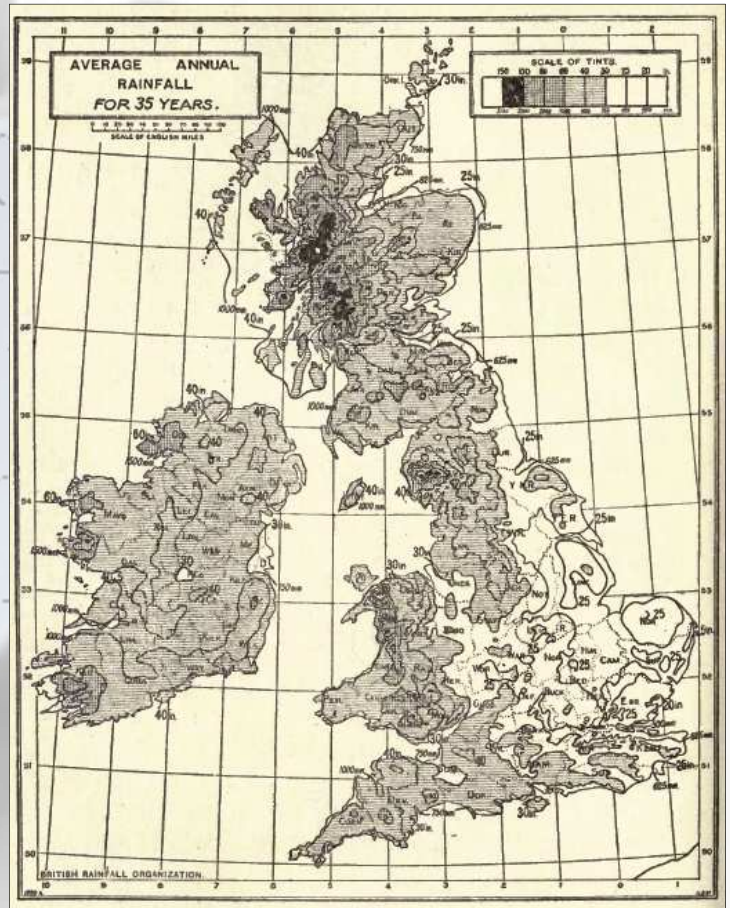
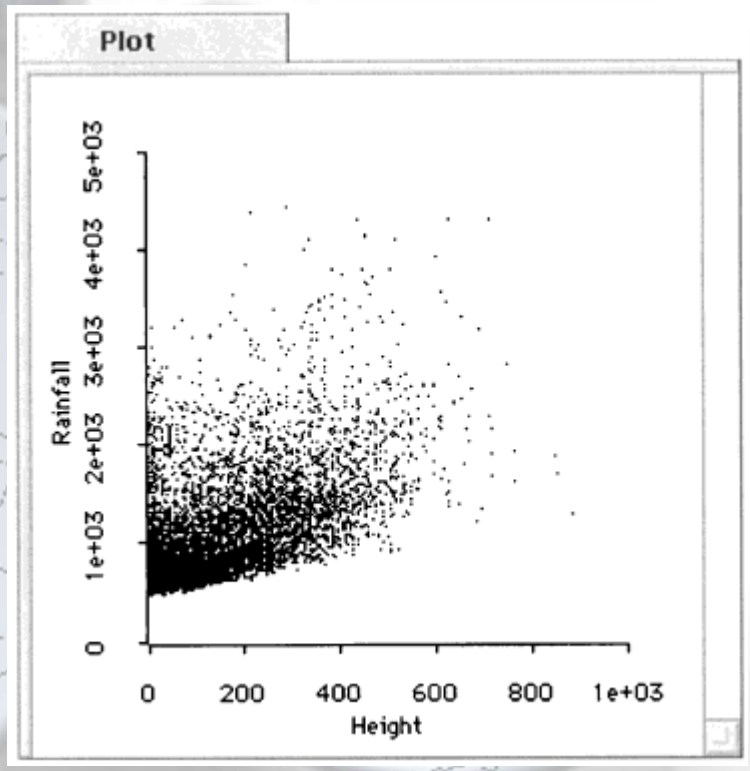


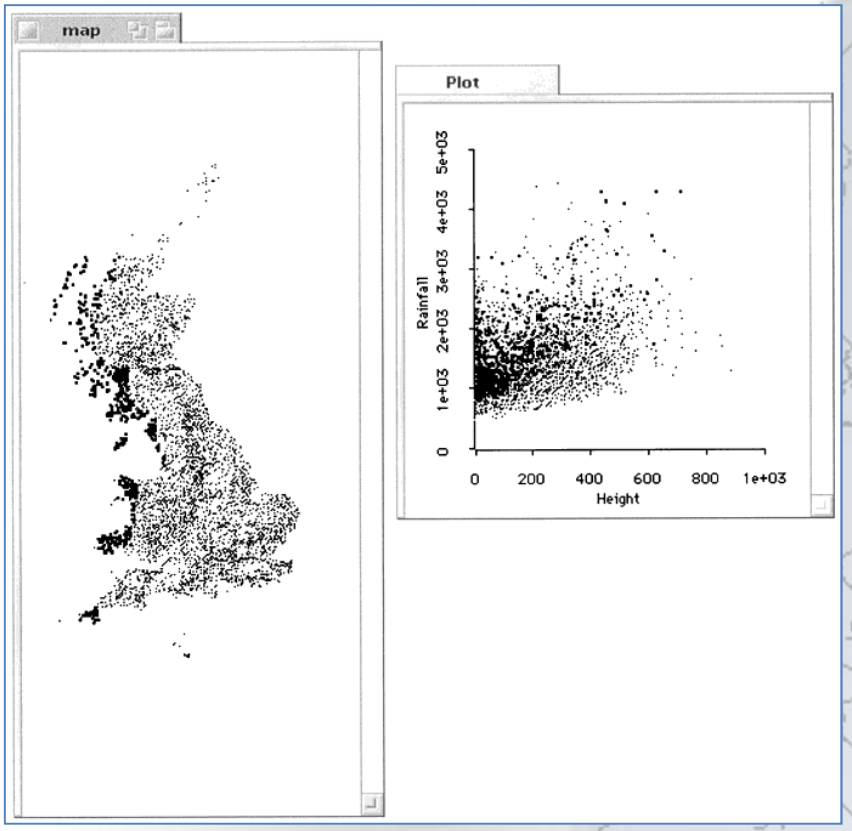
FIG. 123.—DISTRIBUTION OF AVERAGE ANNUAL RAINFALL.

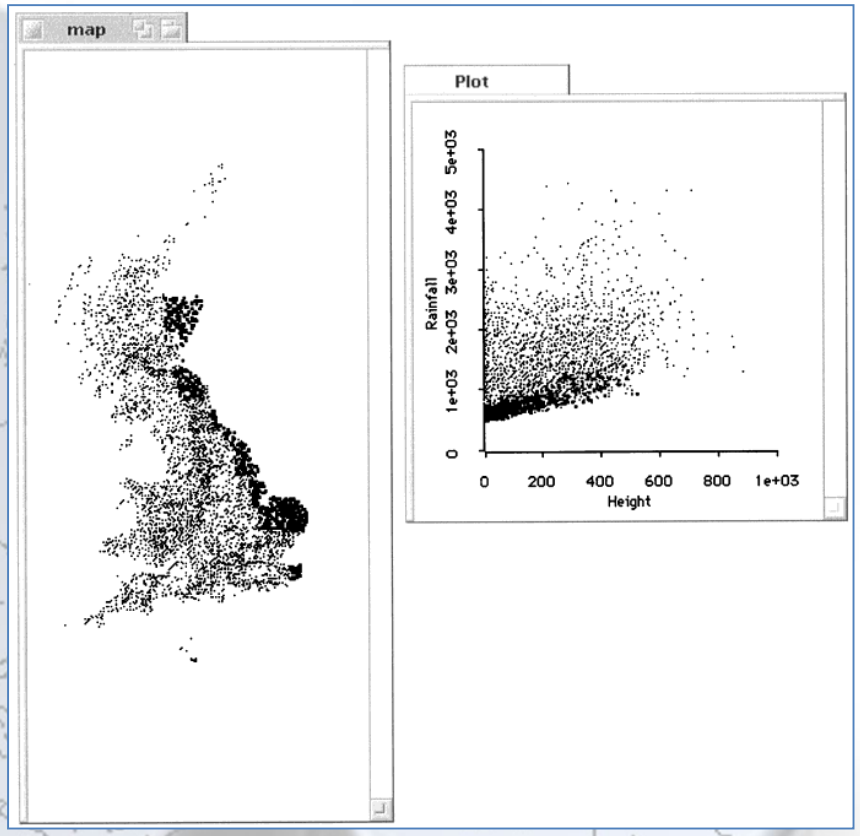
- Using more recent data, Brunsdon, McClatchey and Unwin re-examined the relationship
- The plot shows the mean annual rainfall 1961-1990 at 10925 rain gauges against the elevation of each gauge
- So we can fit a model to make some predictions:



Brunsdon et al, *Int J Climatol*, 2001

$$Rainfall_{est} = 730.83 + 2.28Height \qquad r^2 = 0.34$$





Examining spatial subsets of these data suggest that the relationship between mean annual rainfall and elevation is very different on the western and eastern coasts

Assumptions

- It's clear that in this case a single model is unlikely to satisfy many of the assumptions behind ordinary least squares regression
- The previous plots suggest that the residuals from a single model will exhibit some distinct spatial patterns – positive towards the west and north, and negative towards the east.
- So… the desirable properties of independence, normality and homoskedasticity are absent

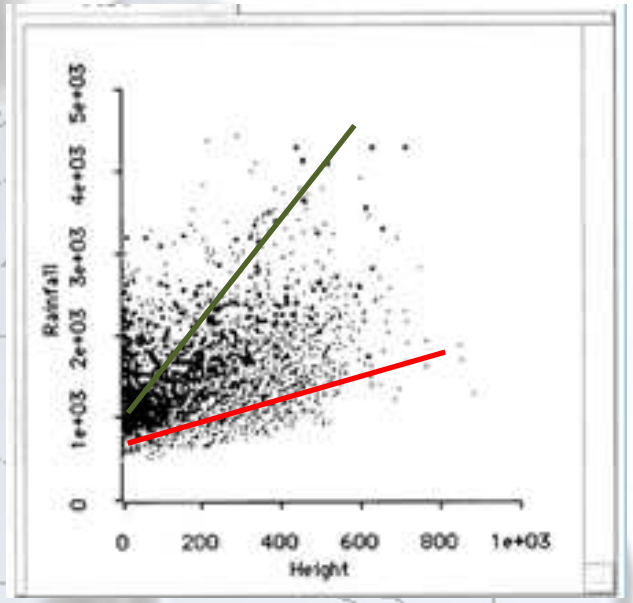
Heterogeneity

- Spatial data can be awkward
- The relationship between rainfall and elevation is not constant across the study area – it's heterogeneous

- Other data exhibit spatial heterogeneity – the relationship between the selling price of a residential property and its characteristics varies from place to place

Dealing with heterogeneity

- We could add a dummy variable (0=West, 1=East) and create an interaction term to obtain separate relationships



- But … Swansea is probably 'west', and Canterbury 'east', but what about Coventry?

Dealing with heterogeneity

- We have 10925 rain gauges
- We could compute separate regression lines for each county, or local authority districts … civil parishes … NUTS3 regions … Police forces?
- But… climate is no respecter of administrative boundaries

- And there's another problem with imposing some arbitrary underlying spatial units…

1933…

CERTAIN EFFECTS OF GROUPING UPON THE SIZE OF THE CORRELATION COEFFICIENT IN CENSUS TRACT MATERIAL[1]

By C. E. Gehlke and Katherine Biehl

Variations in the size of the correlation coefficient seem conditioned upon changes in the size of the unit used, with a smaller value of r associated with the smallest unit rather than with the largest. Various ways of grouping have considerable influence on the r, as well as has the size of the area.

Gehlke and Biehl, JASA, 1934

- Grouping? There are *n* counties in Britain: we can construct many different collections of *n* areas, built from smaller spatial units, which form complete partitions of the island

And…

- Spatial dependence is also a problem – observations with similar values tend to cluster
- Dependence is present in all directions and becomes weaker as data locations become more and more dispersed (Cressie, 1993)
- Clustering results from spatial dependence; similar values concentrate spatially (Varga, 1998)

Models of dependence

- A number of alternative models have been proposed geographers:
- Spatial lag model
  - predictions are based on the values of the x variables and the weighted values of the neighbouring y variable
- Spatial error model
  - the error terms in neighbouring spatial units are correlated
- Both require a spatial weights matrix

Spatial weights

- Common forms for spatial weights matrices include
  - 1/0 adjacent/not adjacent
  - Length of the common boundary
  - Distance between centroids of the spatial units
- First two are common for areal units

Other models

- There is a rich corpus of predictive models available under the general heading of geostatistics
- These include families of models based around the kriging approach
- The kriging variogram gives the weights in this approach

Why create models?

- Prediction – can we forecast the value of some phenomenon from the knowledge of other related phenomena?
- Inference – can we make some inference about the process which leads to spatial variation in some variable of interest
- Synthesis – a combination of variables reveals some underlying structure

Varying parameter models

- In the 1980s geographers started investigating models where the parameter estimates could vary across the study area

- Emilio Casetti – spatial expansion method
- Forster & Gorr – spatial adaptive filtering

The expansion method:

Casetti (1972), Casetti and Jones (1992)

- Suppose we have a y variable, and a single x variable, measured at locations with coordinates u and v, with a model:

$$y = \alpha + \beta x$$

- In the expansion method the coordinates are themselves functions of u and v

$$\alpha = \alpha_1 + \alpha_2 u + \alpha_3 v$$
$$\beta = \beta_1 + \beta_2 u + \beta_3 v$$

- to yield…

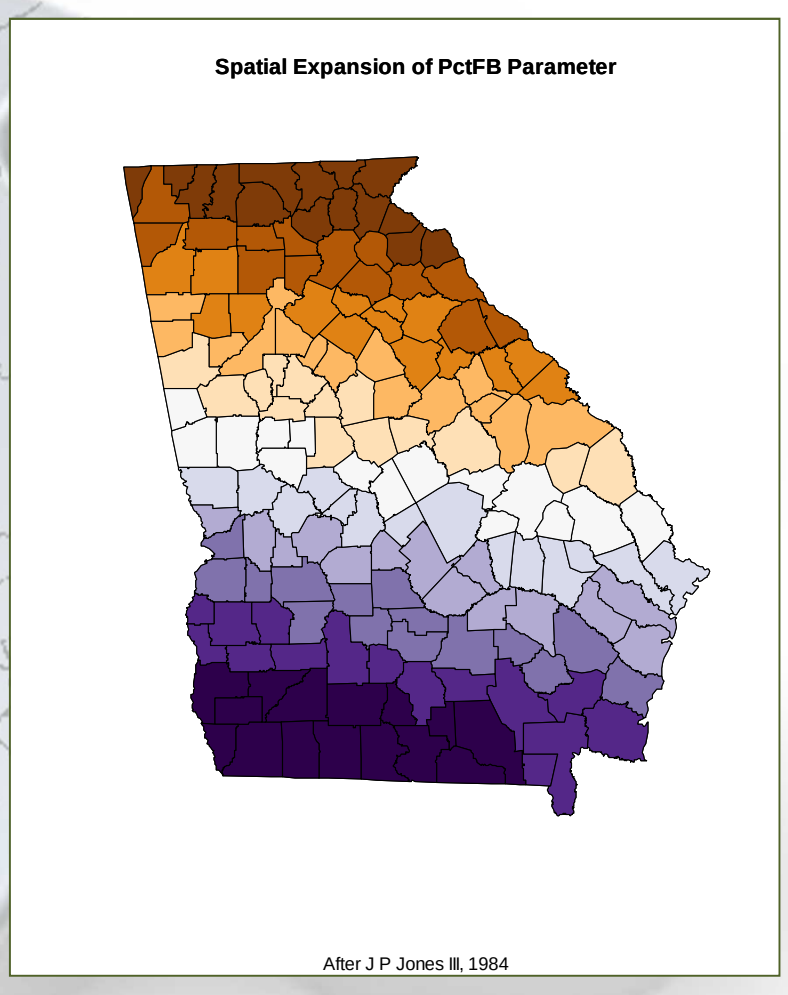
$$y = \alpha_1 + \alpha_2 u + \alpha_3 v + \beta_1 x + \beta_2 ux + \beta_3 vx$$

Example

```
lm(formula = PctBach ~ PctFB + X.PctFB + Y.PctFB)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.475e+00  4.156e-01  17.984  < 2e-16 ***
PctFB       -3.131e+01  5.539e+00  -5.653 7.37e-08 ***
X.PctFB      2.267e-06  1.978e-06   1.146    0.253
Y.PctFB      8.930e-06  1.400e-06   6.379 1.96e-09 ***
```

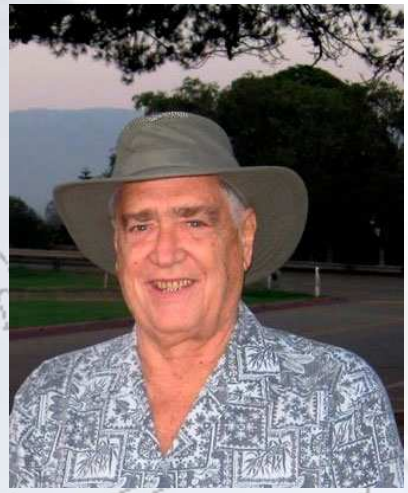
- The X term is not significant
- This is also evident from the plot of the expanded PctFB parameter – spatial variation is south -> north



Expansion method problems

- The functional relationship between the parameters and the coordinates has to be specified in advance
- We don't know what form this relationship will take – it may not necessarily be linear
- There are problems of collinearity into the model, leading to inefficient estimates of the parameters: this is magnified if we assume a more complex quadratic or cubic relationship

- One solution is not to add extra terms into the model but to take note of an assertion by the American geographer, Waldo Tobler:
- Everything is related to everything else, but near things are more related than distant things. Tobler, 1970



Global and local regression models

- We refer to a model in which the parameter estimates for every observation in the sample are identical as a **global** model.

- If the parameter estimates are allowed to vary across the study area such that every observation has its own separate set of parameter estimates we have a **local** model
- How can we estimate a local model

Ordinary Least Squares

- The OLS estimator takes the form

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

- We can weight observations with a diagonal weight matrix W

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

- The weights are in the leading diagonal of W

$$W = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \dots & \dots & w_{\dots} & \dots \\ 0 & 0 & \dots & w_n \end{bmatrix}$$

Geographical weighting

- We can estimate parameters at some location **u** with coordinates (u,v) if we make the weights dependent on the distance between each observation and **u**.

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

$$W(\mathbf{u}) = \begin{bmatrix} w(\mathbf{u})_1 & 0 & \dots & 0 \\ 0 & w(\mathbf{u})_2 & \dots & 0 \\ \dots & \dots & w(\mathbf{u})_{\dots} & \dots \\ 0 & 0 & \dots & w(\mathbf{u})_n \end{bmatrix}$$

Weighting schemes

- Suitable weighting schemes should have the property that the weight when **u** is at the location of one of the observations is 1, and decreases with increasing distance
- d_{ui} is the distance between **u** and the location of the *i*th sample, and *h* is the bandwidth

Gaussian:

$$w(\mathbf{u})_i = e^{-0.5(d_{\mathbf{u}i}/h)^2}$$

Bisquare

$$w(\mathbf{u})_i = (1-(d_{\mathbf{u}i}/h)^2)^2$$

$$w(\mathbf{u})_i = 0 \quad when \quad d_{\mathbf{u}i} > h$$

Bandwidth

- The Gaussian and bisquare schemes are referred to as **kernels**.
- The bandwidth *h* can be supplied exogenously or estimated from the data
- The larger the bandwidth the more the model parameters will approach their global values

Estimating the bandwidth

- Geographical weighting is generally more sensitive to variation in the bandwidth than the choice of kernel function
- The question arises of what an appropriate choice for the bandwidth might be
- This acts as a smoothing parameter
- A very large ($h \rightarrow \infty$) value leads to a model with no variation in the parameters – every weight is 1
- If the value is too small, the model wraps itself around the data

Bandwidth estimation

- The user can supply a a value for the bandwidth – perhaps from a previous estimation on a similar dataset
- The bandwidth can also be estimated from the data – find h which gives the best fit to the observed ys
- Minimise crossvalidation score
- Minimise Akaike Information Criterion

Kernel types and bandwidths

- Kernels can be either fixed radius or fixed local sample size
- In a fixed radius kernel the radius is specified in the same coordinate units as those for the data
- In a fixed local sample size, the radius varies so that the same number of observations is used in each local estimation – ‘adaptive’ kernel

- With GWR the complexity of the model increases as the bandwidth is decreased from ∞.
- The number of parameters in the model can be estimated from the 'hat matrix'
- $Y^* = SY$
- The 'hat matrix' converts the observed y values into the fitted y values
- A hat matrix can be derived for GWR

Hat matrix and the AIC

- The trace (sum of the elements in the leading diagonal) of the hat matrix is the number of parameters
- In the case of GWR this gives us the effective number of parameters – the smaller the bandwidth, the larger this value becomes
- The Akaike Information Criterion includes a penalty for the complexity of the model, such that in comparing models for the best value of h we are comparing like with like.

Outputs from GWR

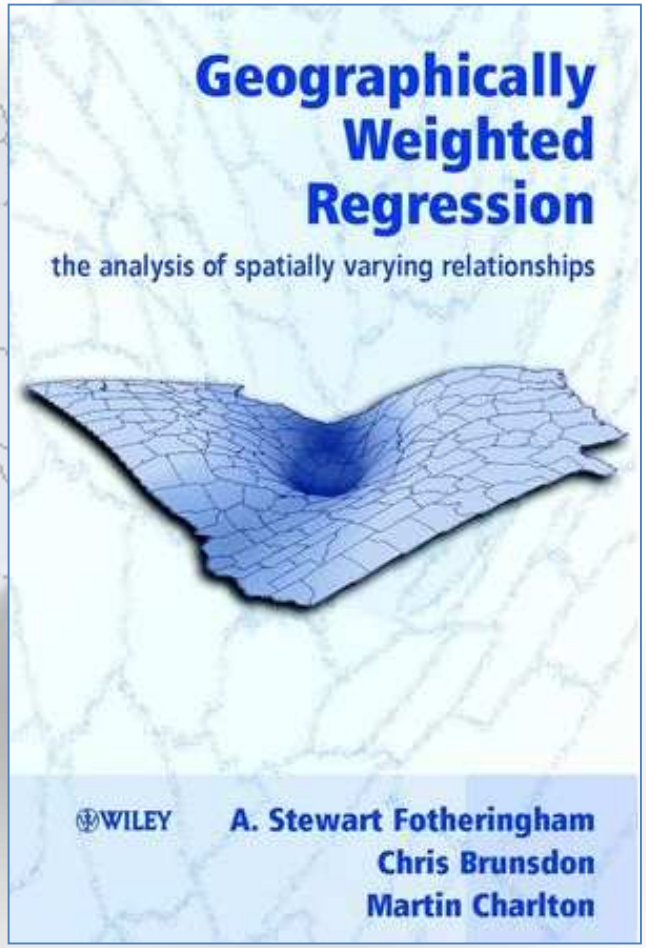
- As well as the estimates of the parameters at each location, we can estimate the locally weighted standard errors, fitted values, residuals, standardised residuals
- We can also estimate a pseudo-t value (to test the hypothesis that a local parameter estimate is zero)

Inference

- Two questions arise from the estimates.
- Is the variation in the estimates for a single parameter random?
  - Monte Carlo significance test
- Are the values of the individual parameter estimates zero?
  - Treat as local t values, but with an adjustment for multiple testing

Literature on GWR

- Although the first paper appeared in 1996, the developers authored a book which was published in 2002
- We produced some software which allowed users to run GWR models

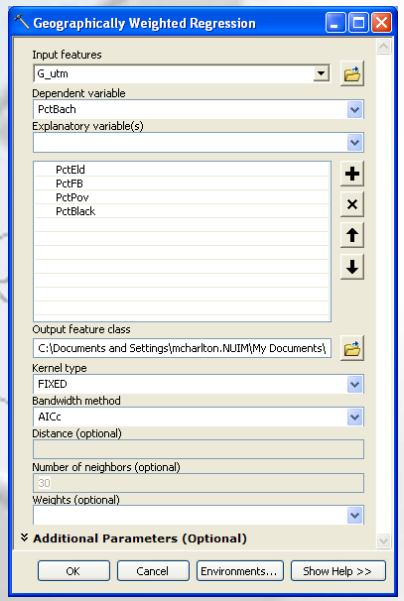


GWR: software

There are several versions of GWR:

- spgwr – R package (Bivand and Yu)
- gwrr – R package (Wheeler)
- SpaceStat – BioMedware, Ann Arbor, MI
- ArcGIS – ESRI, Redlands, CA
- GWR4 (Nakaya)

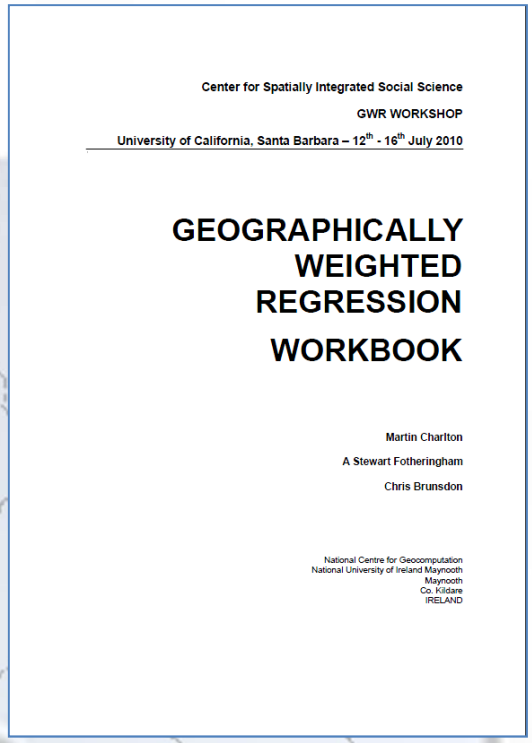
And… there's our own Windows software which we use as the basis for the workshops



Also as a core tool in ESRI's ArcGIS

GWR: outreach

- CSISS Advanced Spatial Analysis Workshop, **University of California at Santa Barbara**, July 12-16, 2010
- CSISS Advanced Spatial Analysis Workshop, **Pennsylvania State University**, June 1-6, 2008
- **Greater Manchester** Police Safety Unit, March 5-7, 2008
- **Helsinki University of Technology**, January 22-25, 2008
- GEOIDE Summer School**, St Mary's University, Halifax NS**, June 3-4 2007
- **University of Adelaide**, Australia, April 17th-20th 2007
- **Technical University of Lisbon**, Lisbon, February 21-22, 2007
- Transport Operations Research Group, **University of Newcastle upon Tyne**, November 29, 2006
- **National Centre for Geocomputation**, National University of Ireland, Maynooth, County Kildare, July 20th 2005
- Department of Geography, **University of Leeds**, Leeds, June 15th- 16th 2005
- Department of Geography, **Ritsumeikan University**, Kyoto, May 16th – 19th 2005
- Centre for Applied Spatial Analysis, **University of London**, London, 12th May 2005


Center for Spatially Integrated Social Science

GWR WORKSHOP

University of California, Santa Barbara – 12th - 16th July 2010

GEOGRAPHICALLY WEIGHTED REGRESSION

WORKBOOK

Martin Charlton

A Stewart Fotheringham

Chris Brunsdon

National Centre for Geocomputation
National University of Ireland Maynooth
Maynooth
Co. Kildare
IRELAND

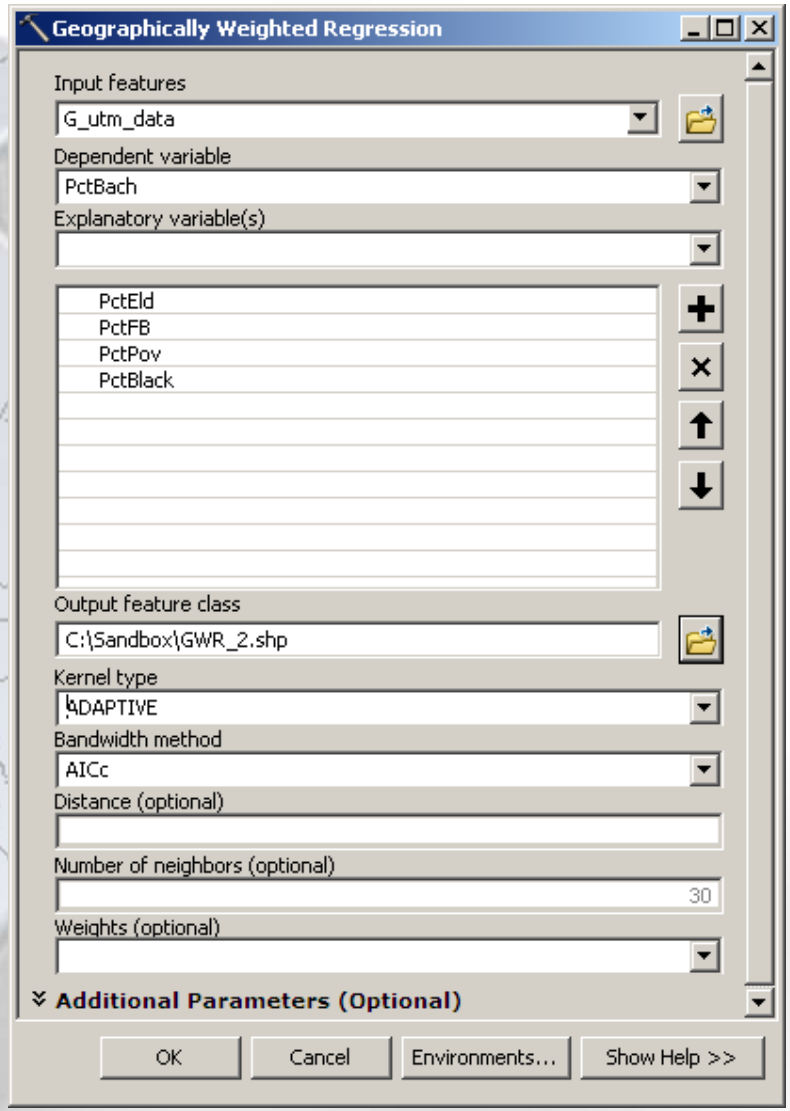
GWR and GIS

- GWR was conceived as having a close relationship with Geographic Information Systems
- GIS software
  - used in the data integration process to create suitable datasets for modelling with GWR
  - used as a diagnostic tool
    - Plot residuals from global and local models
    - Plot the estimated parameter surfaces

Using GWR

As an example we model education attainment in the counties of Georgia with four socio-economic predictors; the dependent variable is the proportion of the population education to Bachelor's degree level or higher

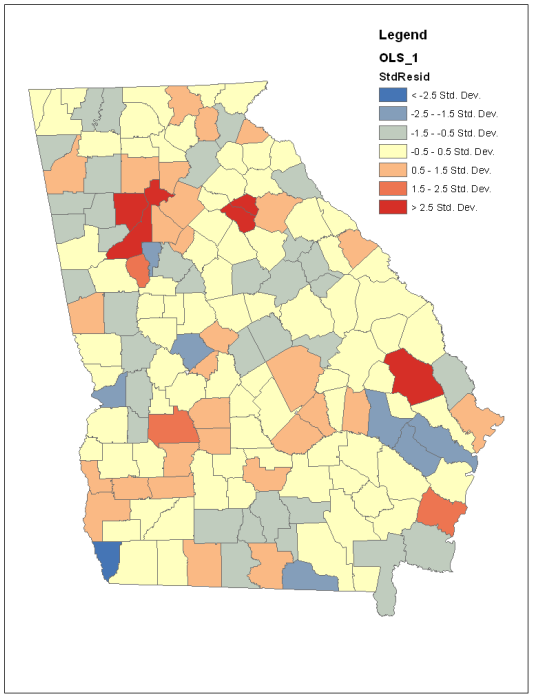
There are a number of choices available in specifying a GWR model. The ArcGIS GWR tool is one example GUI.



OLS fit

Following some initial data exploration, we fit an Ordinary Least Squares model and examine the outputs.

Is there spatial structure in the residuals? If so, then a GWR model may be an appropriate next step

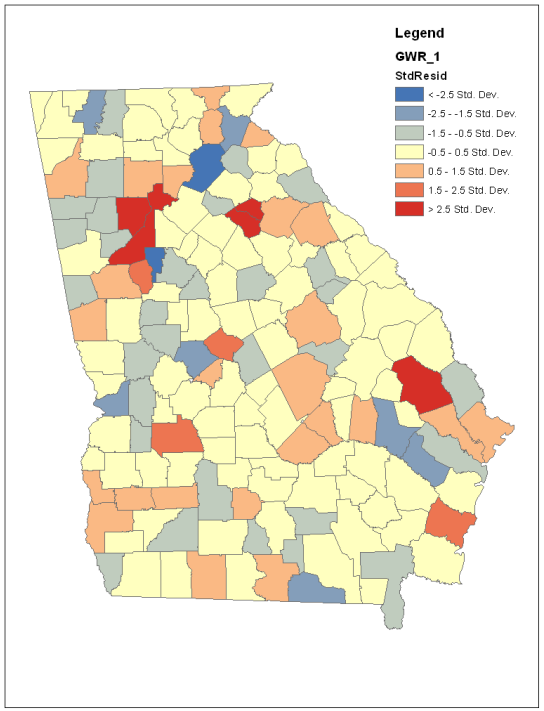


AIC is 898; coefficients have reasonable values, r^2 is 0.51, … but the residuals show some clear spatial structure: Moran's I=0.14 p=0.004

Variable	Coef	StdError	t_Stat	Prob	Robust_SE	Robust_t	Robust_Pr
Intercept	**12.67112563970**	1.63682331552	7.74129102365	**0.00000000003**	2.12548562702	5.96152026561	0.00000004073
PCTELD	**-0.10530556376**	0.13783764308	-0.76398262051	**0.44603927418**	0.15048483949	-0.69977523398	0.48511933030
PCTFB	**2.54521146415**	0.29839213403	8.52975388397	**0.00000000000**	0.59460963548	4.28047463795	0.00003617641
PCTPOV	**-0.28290660460**	0.07810457473	-3.62215152656	**0.00040528984**	0.12849100021	-2.20176202338	0.02916016987
PCTBLACK	**0.07684751375**	0.02802807341	2.74180506883	**0.00683213848**	0.03495341185	2.19856974396	0.02939123745

GWR fit

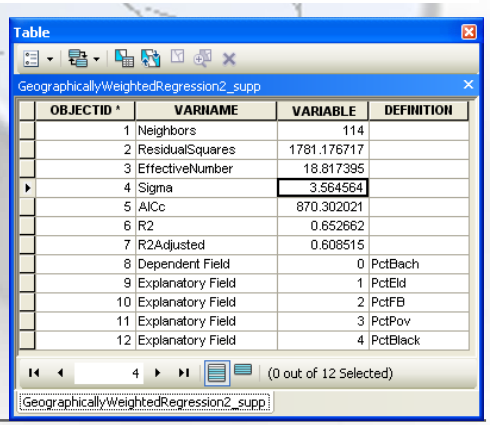
The immediate output from GWR using the ArcGIS tool is a map of the residuals – whilst the largest residuals remain, the Moran test (I=0.06 p-0.16) suggests that the evidence for the geographical pattern is less strong.



The bandwidth is 114 – using an adaptive kernel (n=159)

The fit of the model has improved (AIC reduced from 898 to 870).

There is no list of parameter estimates – these are in the attribute table of the output shapefile



OBJECTID *	VARNAME	VARIABLE	DEFINITION
1	Neighbors	114	
2	ResidualSquares	1781.176717	
3	EffectiveNumber	18.817395	
4	Sigma	3.564564	
5	AICc	870.302021	
6	R2	0.652662	
7	R2Adjusted	0.608515	
8	Dependent Field	0	PctBach
9	Explanatory Field	1	PctEld
10	Explanatory Field	2	PctFB
11	Explanatory Field	3	PctPov
12	Explanatory Field	4	PctBlack

Table

GeographicallyWeightedRegression2

OBJECTID *	Shape *	Observed PctBach	Condition Number	Local R2	Predicted	Coefficient Intercept	Coefficient #1 PctEld	Coefficient #2 PctFB	Coefficient #3 PctPov	Coefficient #4 PctBlack	Residual
1	Polygon	11.6	12.198804	0.664702	10.137337	11.043656	-0.279521	3.841697	-0.032929	0.007367	1.462663
2	Polygon	11.4	12.274694	0.664148	9.639036	10.818433	-0.242401	3.819424	-0.057734	0.014401	1.760964
3	Polygon	10.1	12.278887	0.664961	8.835155	10.704447	-0.224939	3.815855	-0.070022	0.018422	1.264845
4	Polygon	7.8	12.325626	0.664731	7.809879	10.544201	-0.193465	3.789366	-0.092207	0.024539	-0.009879
5	Polygon	5.5	12.459494	0.665076	9.714758	10.403483	-0.154509	3.747911	-0.122373	0.031986	-4.214758
6	Polygon	12	12.485815	0.665817	17.181239	10.326716	-0.135687	3.731308	-0.13692	0.035915	-5.181239
7	Polygon	8.1	12.48403	0.665694	8.685325	10.277676	-0.124236	3.71771	-0.145339	0.038305	-0.585325
8	Polygon	8.4	12.631223	0.666381	8.537637	10.238256	-0.104628	3.691405	-0.161554	0.041855	-0.137637
9	Polygon	8	12.605828	0.666719	9.515063	10.186278	-0.093769	3.680527	-0.169634	0.044415	-1.515063
10	Polygon	8.6	12.443882	0.665112	8.185605	10.524083	-0.181779	3.774624	-0.102241	0.02653	0.414395
11	Polygon	12	12.429926	0.665379	18.470835	11.107034	-0.277684	3.828734	-0.037602	0.007079	-6.470835
12	Polygon	13.6	12.431958	0.664936	11.314294	10.938121	-0.252804	3.818046	-0.053524	0.012026	2.285706
13	Polygon	11.1	12.479092	0.665054	10.522865	10.802825	-0.229822	3.805696	-0.069328	0.016646	0.577135
14	Polygon	13.1	12.496642	0.666191	10.625139	11.379309	-0.311162	3.833111	-0.017951	0.000254	2.474861
15	Polygon	9.2	12.700153	0.665609	9.993143	10.388158	-0.135195	3.71934	-0.139035	0.034924	-0.793143
16	Polygon	8.6	12.582299	0.665442	10.048617	10.746278	-0.215985	3.794529	-0.079836	0.019206	-1.448617
17	Polygon	13.7	13.094213	0.666027	11.236457	10.260104	-0.085564	3.650982	-0.177095	0.043472	2.463543
18	Polygon	5.9	12.883299	0.666364	7.205323	10.225273	-0.087109	3.660131	-0.175736	0.044352	-1.305323
19	Polygon	9.5	12.692462	0.667827	8.269976	11.653627	-0.338305	3.829174	-0.004898	-0.005238	1.230024
20	Polygon	9	12.633045	0.665087	7.77645	10.574746	-0.179782	3.762195	-0.105694	0.0261	1.22355
21	Polygon	9.1	12.793192	0.668426	7.526584	12.031292	-0.375811	3.812621	0.014799	-0.013114	1.573416
22	Polygon	15.4	12.715233	0.666089	25.474235	11.088135	-0.263942	3.813431	-0.049925	0.009308	-10.074235
23	Polygon	6.4	12.68567	0.667112	8.844834	11.373607	-0.304382	3.828406	-0.024831	0.001531	-2.444834
24	Polygon	18.4	12.895323	0.664687	14.698412	10.627397	-0.175467	3.743644	-0.11059	0.025693	3.701588
25	Polygon	9	12.99528	0.665382	10.986728	10.425575	-0.127654	3.699123	-0.145889	0.034965	-1.986728
26	Polygon	15.6	12.841647	0.665364	12.902632	10.876666	-0.226145	3.786548	-0.075147	0.016134	2.697168
27	Polygon	8	13.316406	0.671945	9.930748	12.70365	-0.430464	3.77209	0.035842	-0.023867	-1.930748
28	Polygon	9	13.010654	0.668197	10.045837	11.536771	-0.315541	3.818694	-0.021467	-0.001285	-1.045837
29	Polygon	9.7	13.085374	0.671105	11.117647	12.095656	-0.37892	3.818506	0.012669	-0.013144	-1.417647
30	Polygon	31.6	13.84051	0.663998	23.407118	10.722113	-0.159066	3.701862	-0.122437	0.023657	8.192882
31	Polygon	29.6	13.298025	0.666072	28.9282	11.150847	-0.25412	3.785082	-0.059388	0.009065	0.6718
32	Polygon	9.2	13.262973	0.66803	11.345102	11.520456	-0.307132	3.807619	-0.027803	-0.000424	-2.145102
33	Polygon	6.8	13.544101	0.664918	10.315261	10.235616	-0.065108	3.611191	-0.190487	0.045262	-3.515261
34	Polygon	12.8	13.871941	0.67577	10.132663	12.81169	-0.436992	3.77862	0.034029	-0.024432	2.667337
35	Polygon	33	13.449348	0.664404	24.505799	10.614031	-0.149255	3.702458	-0.130422	0.027996	8.494201
36	Polygon	7.6	13.598012	0.664646	10.791474	10.399866	-0.099211	3.651487	-0.165805	0.037334	-3.191474
37	Polygon	37.5	13.468356	0.672209	26.039785	12.145878	-0.377758	3.81162	0.009108	-0.01345	11.460215
38	Polygon	10.4	14.57426	0.677617	6.497664	13.654626	-0.501145	3.704227	0.053671	-0.035344	3.902336
39	Polygon	8.2	14.794555	0.671698	7.048864	14.022625	-0.519176	3.639002	0.054942	-0.038995	1.151136
40	Polygon	28.4	13.806918	0.673199	12.626465	12.267988	-0.386092	3.804645	0.011595	-0.015503	15.773535
41	Polygon	32.7	13.872868	0.663207	33.947257	11.001057	-0.208163	3.725386	-0.088744	0.014073	-1.247257
42	Polygon	9.4	13.874315	0.6695	10.414072	11.81165	-0.332622	3.797593	-0.015144	-0.006915	-1.014072
43	Polygon	7.5	14.081318	0.66436	9.477064	10.138356	-0.032538	3.565504	-0.210179	0.048796	-1.977064
44	Polygon	11	15.004711	0.675676	10.07768	12.697591	-0.417328	3.777594	0.023885	-0.023538	0.92232
45	Polygon	12	14.825278	0.66242	10.578667	10.050415	-0.006684	3.53418	-0.219067	0.04808	1.421333
46	Polygon	12	14.239848	0.663278	13.836382	10.403096	-0.085353	3.627212	-0.17089	0.035498	-1.836382
47	Polygon	18.1	14.356807	0.664473	16.599618	11.369832	-0.257622	3.744103	-0.056083	0.002502	1.500382
48	Polygon	8.8	15.343417	0.680838	8.319011	13.450556	-0.484888	3.741516	0.046592	-0.032684	0.480989
49	Polygon	5.6	15.691915	0.681014	6.192639	14.052472	-0.531934	3.685668	0.057761	-0.038355	-0.592639
50	Polygon	23.9	15.794244	0.642128	23.382791	13.787248	-0.461989	3.544341	0.019909	-0.033364	0.517209
51	Polygon	9.5	14.989915	0.666722	11.039483	11.820005	-0.309846	3.751376	-0.024461	-0.009509	-1.539483
52	Polygon	10.4	16.191606	0.657348	8.698267	14.100977	-0.50628	3.581836	0.035385	-0.036177	1.701733

1 (0 out of 159 Selected)

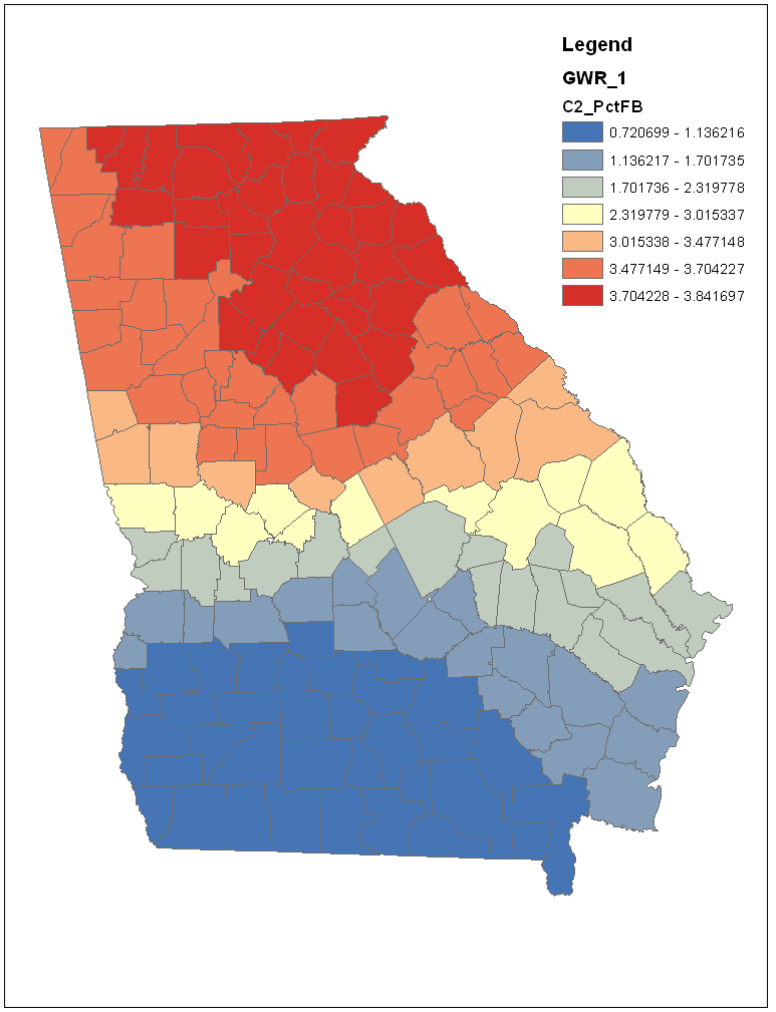
GeographicallyWeightedRegression2_supp | GeographicallyWeightedRegression2

Mappable parameters

The plot shows the variation in the parameter estimate for ‘Foreign Born’.

There’s evidence of a north-south pattern, with variation in the Foreign Born variable being more influential on variation in educational attainment in the north of the state than in the south.

The other parameter estimates and diagnostics may be mapped as well.



Estimation at non-data points

- The weighting scheme allows parameters to be estimated at locations different from those used where the data are collected
- These might be the mesh points of a regular grid
- A dataset can be split in two – a training set and a validation set – the estimation of the parameters at the validation set locations uses the training data; fitted values can be obtain if there are suitable x data at the validation set locations

Prediction at non-data points

- GWR-geostatistical hybrids have been explored by Paul Harris
  - Harris P, Fotheringham AS, Crespo R, Charlton ME (2010b) The use of geographically weighted regression for spatial prediction: an evaluation of models using simulated data sets. Mathematical Geosciences 42:657-680
  - Harris P, Brunsdon C, Fotheringham AS (2011) Links, comparisons and extensions of the geographically weighted regression model when used as a spatial predictor. Stochastic Environmental Research and Risk Assessment 25:123-138
  - Harris P, Juggins S (2011) Estimating freshwater critical load exceedance data for Great Britain using space-varying relationship models. Mathematical Geosciences 43: 265-292

Issues: model building

- In our book we propose a semi-parametric version of GWR and provide a method of estimating parameters
- Some of the variables are held as ‘global’, and the others are treated as ‘local’ with spatially varying parameters
- A question arises as to how the users decides which variables are to be global and which might be local
- We have experimented with a pseudo-stepwise method

Issues: distance metrics

- In generating the geographical weighting we need an estimate of the distance between each focus point and the rest of the data locations
- As a default we use the Euclidean distance when the coordinates of the locations are in a projected system
- If the geographical coordinates (latitude and longitude) are used, then Great Circle distances are perhaps more appropriate

Alternative distance metrics

- Atsuyuke Okabe suggested over a decade ago that we might use distances computed from a representation of the road network
- Perhaps more appropriate when modelling social processes – for example hedonic housing models
- Allows us to take into account the influence of potential barriers to movement, such as rivers
- This requires a reliable representation of the road network…

Alternative metrics

- Geographers in the USA have used the 'Manhattan' (aka 'taxicab' or 'city block') distance d=abs(dx)+(abs(dy)
- One of a generalised set of metrics known as Minkowski metrics
- $d= (abs(dx)^p+abs(dy)^p)^{1/p}$
- When p=2 we have Euclidean distances; p=1 we have Manhattan distances
- P need not be an integer, which raises the possibility of estimating a value for p optimises the goodness of fit
- Lu B, Charlton ME, Fotheringham AS, 2011, Geographically weighted regression using non-euclidean distance metrics. Spatial Statistics

Issues: collinearity

- Collinearity was raised as an issue in 2005[1]
- Collinear data: larger standard errors..
  - Potentially significant parameters cannot be identified as such
- ArcGIS tool and spgwr functions report the condition number of the geographically weighted crossproduct matrices (X'WX)
- Following suggestions of Belsey, Kuh and Welsch (Regression Diagnostics, 1980), condition numbers > 30 are held to indicate potential collinearity problems

[1]David Wheeler, Michael Tiefelsdorf, 2005, Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems* 7(2): 161-187

Collinearity

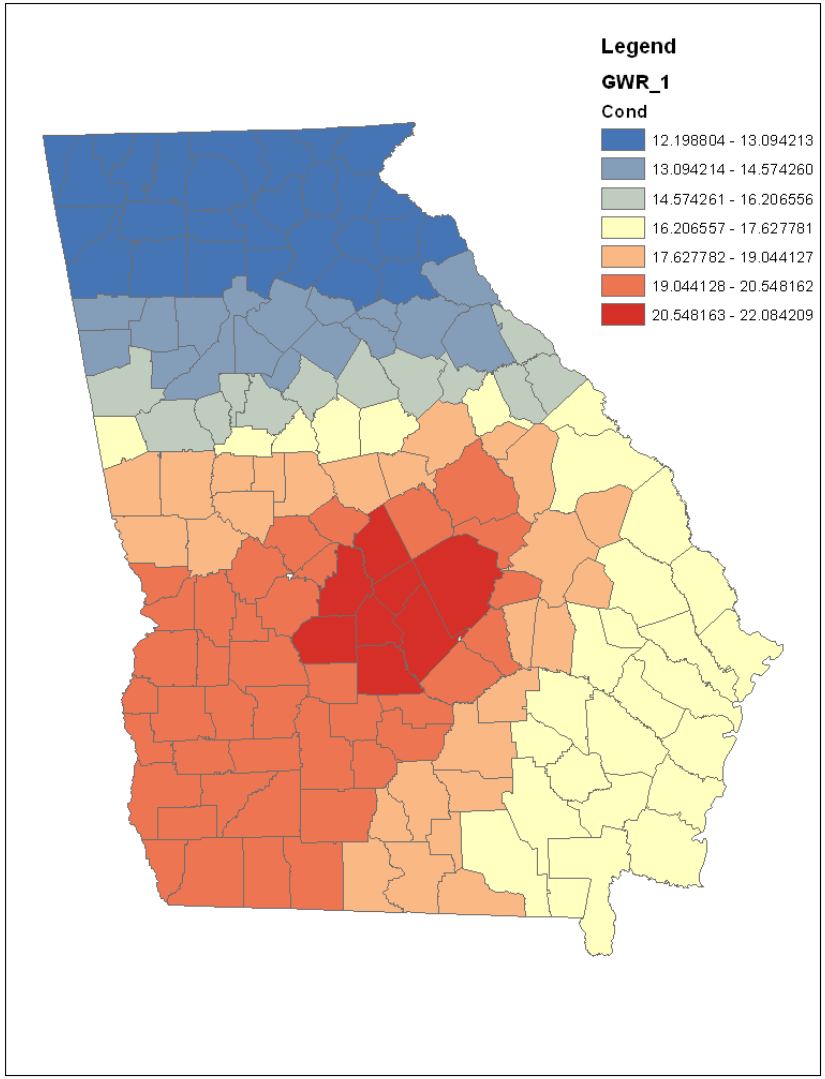
- Wheeler, in particular, has explored using geographically weighted ridge regression, and the lasso in dealing with collinear data

- Others have taken a more extreme position:
  - The predictions available for gwr() are speculative at best. Note that GWR is a notoriously unreliable technique, and simulation studies indicate that it finds pattern in coefficients even when there is none. So any tests are doubtful anyway - it should only be used for exploring the data for possible missing variables or inappropriate functional forms. (Bivand, R-sig-geo, 20-ii-2012)

Condition number

The map shows the variation in the condition number of the locally weighted X'WX matrix for each county.

Non of the values exceeds the threshold of 30, although we note that the most potentially collinear fits are in the centre of the state.



Issues: Collinearity

- One strategy is to use **ridge regression** to deal with collinear data – provides sharper estimates of the parameters
- $\beta(u) = (X'W(u)X + \lambda I)^{-1}(X'W(u)X+\lambda I)y$
- … where λ is the ridge parameter
- But what value is appropriate for λ?
- One approach is to estimate the bandwidth and λ to optimise the fit of the model
- Variation in goodness of fit appears to be more dependent on variation in bandwidth rather than variation in λ.

Issues: flexible bandwidth

- The bandwidth is a global measure
- Kernels are either fixed radius or fixed local sample size
- Apply equally to to each variable in the model
- However, spatial scale in influence of each variable may be different
- Flexible bandwidth model allows a separate bandwidth to be estimated for each variable
- Under development…
- Yang W, Fotheringham AS, Harris P, 2011, Model selection in GWR: the development of a flexible bandwidth GWR. *Geocomputation 2011*

GWPCA

- A complementary approach is to consider geographically weighted principal components
- Link to GWR through condition number – it's the ratio of the largest to the smallest eigenvalue of the cross-product matrix
  - If the data are very locally collinear, much of the local variation will be accounted for by the first component, and little by the last, so this ratio will be large

- Harris P, Brunsdon C, Charlton ME (2011a) Geographically weighted principal components analysis. International Journal of Geographical Information Science DOI:10.1080/13658816.2011.554838

GWPCA

- Output: set of eigenvectors of component loadings and associated eigenvalues for every
- Challenges for visualisation
  - Map variations in the local eigenvalues – particularly 1st and last
  - Larger values of eigenvalue 1 may indicate potential local collinearity
  - Mapping local eigenvectors… spatial variation in the contribution of variables to each component (using some form of multivariate glyph)

Other approaches

- Assuncao, 2003, Space varying coefficient models for small area data, Environmetrics
- Gelfand et al, 2003, Spatial Modeling With Spatially Varying Coefficient Processes, JASA
- Bayesian modelling framework – puts inference concerning parameters on a consistent basis
- Gelfand's models have been compared with GWR – better fit to the data, however, MCMC approach is time consuming in fitting models

Development

- The search term “geographically weighted regression” returns 67,400 hits in Google.
- We know from our own distribution of the software that interest has been shown is disciplines outside geography
- There are other GW possibilities
  - summary statistics
  - discriminant analysis
  - boxplots
  - variograms
- There are still plenty of spatial modelling challenges

Thank you