

Richard Harris, Min hua Jen, David Kilham,  
Chris Brunson, Claire Jarvis, Edward Thomas

# Developing Grid enabled spatial regression models

Second International Conference on e-Social Science,  
June 30, 2006

ESRC Grant Number: RES-149-25-1041



# Outline

- Crash course in localized, spatial statistics
  - Exemplified by GWR (Geographically Weighted Regression)
- Consider the model fitting algorithm
  - Sequential repeat testing → 'embarrassingly parallel', concurrent testing
- Linking GWR to the National Grid Service
- Acknowledge an intellectual legacy
  - Professor Stan Openshaw (and colleagues)
    - e.g. *Google* Geographical Analysis Machine (GAM)

## The paradox

- The use of linear regression modelling is widespread in geographical and other quantitative social sciences
- But, it makes little sense to look for geographies of the relationship between  $Y$  and  $X$  using a method assuming **spatial independence of the regression residuals**
  - Geography Vs non-geography

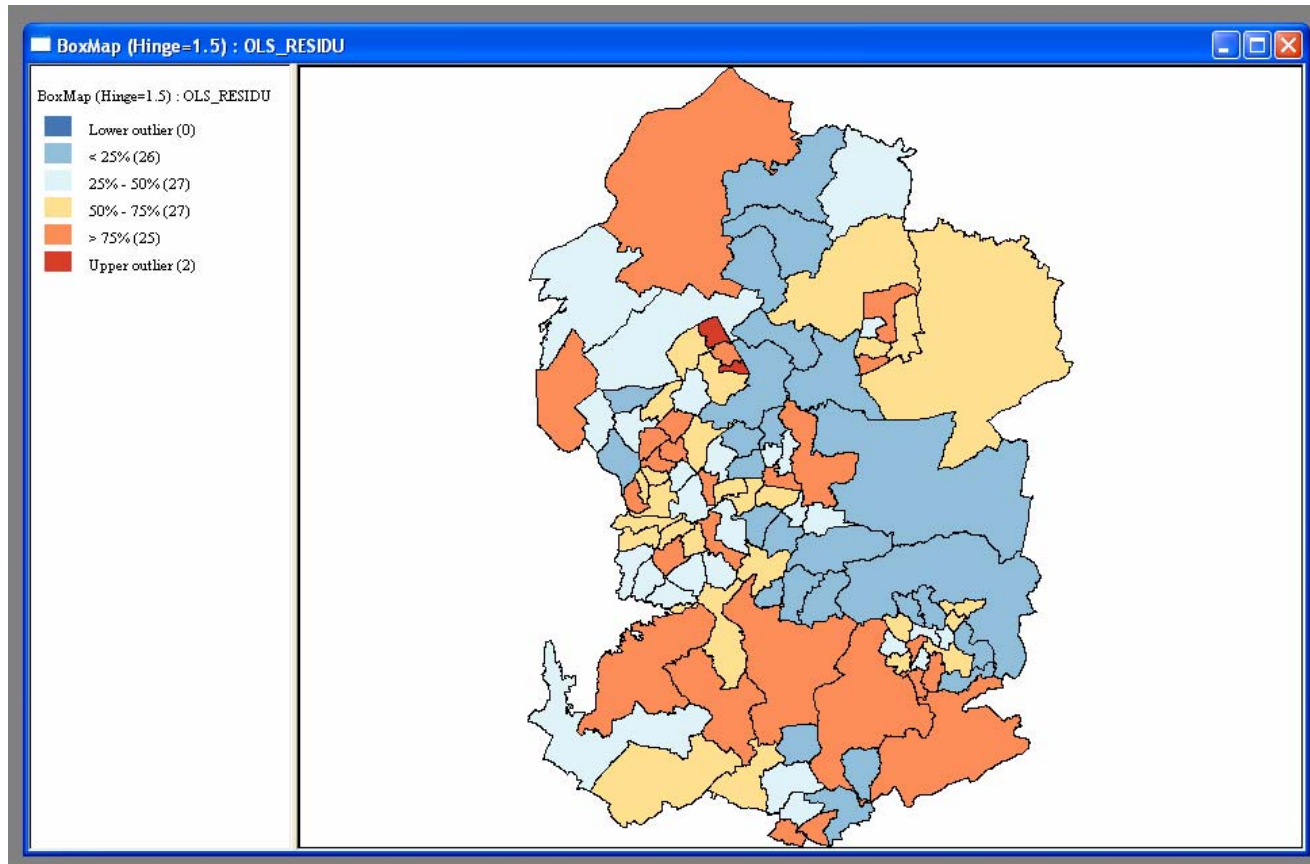
## Fitting a global (non-spatial) model

- Consider the following model of income in Bristol, Bath, North Somerset and South Gloucestershire
  - $y = \beta_0 + \sum_k \beta_k x_k + \epsilon$

	Variable definition	Source	$\beta_k$	t value
y	Estimate of net household weekly income (£) (2001/2)	Office for National Statistics (ONS)		
$x_{k=0}$	Intercept		343	18.82
$x_{k=1}$	Proportion of households with two cars or vans	ONS: 2001 Census	447	17.25
$x_{k=2}$	Proportion of people aged 16-74 in employment working in elementary occupations	ONS: 2001 Census	-539	-7.65
$x_{k=3}$	Proportion of all people born in Scotland	ONS: 2001 Census	2629	5.33
$x_{k=4}$	Proportion of all people of 'other' ethnic group	ONS: 2001 Census	-2779	-3.73

R-sq (adj.) 90.7%

# Mapping the residuals



## But what next?

- If patterns of spatial autocorrelation are found
  - Important local variations in relationships between the regression variables have been 'averaged away' by the global fit.
  - Model will be biased (standard errors are too low)
  - Offers little statistical explanation about the cause of local variation.
  - If used for prediction / interpolation, the model risks generating erroneous values.

# Geographically weighted regression (GWR)

- GWR treats the regression model

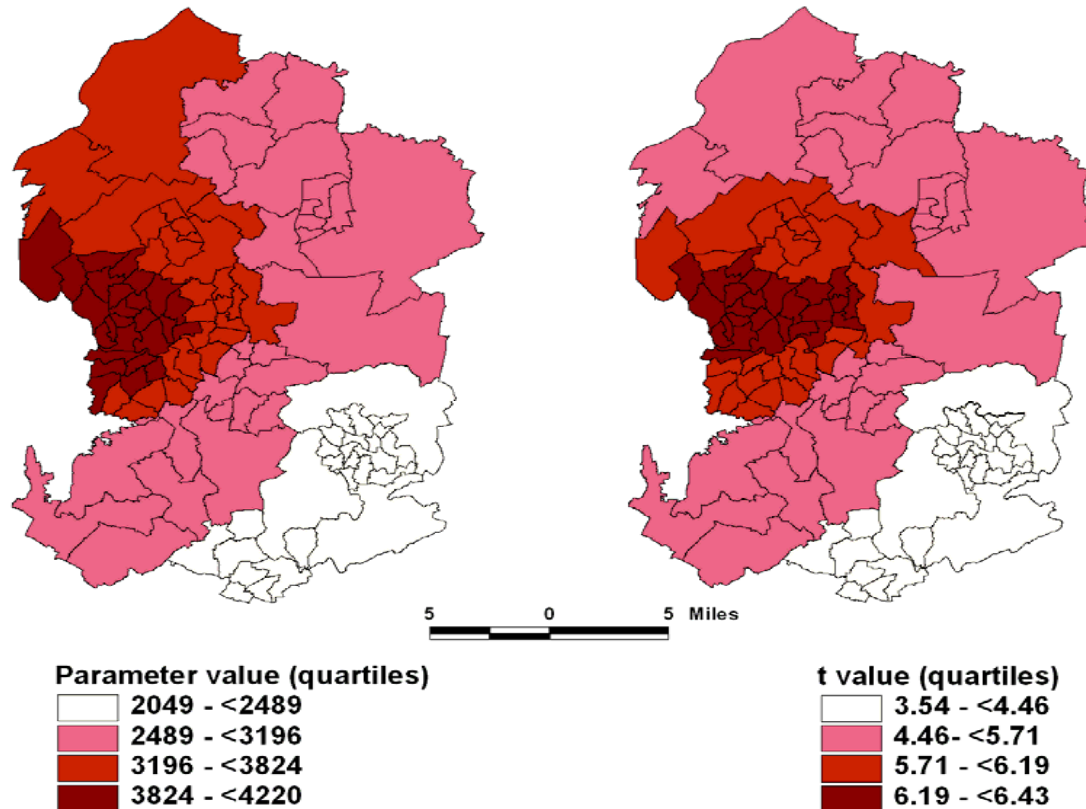
$$y = \beta_0 + \sum_k \beta_k x_k + \epsilon$$

- as a special case of

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \epsilon_i$$

- where  $\beta_k(u_i, v_i)$  is a realisation of the continuous function  $\beta_k(u, v)$  at point  $i$

# GWR fitted Scottish parameter and t values across the study region



## How does it work?

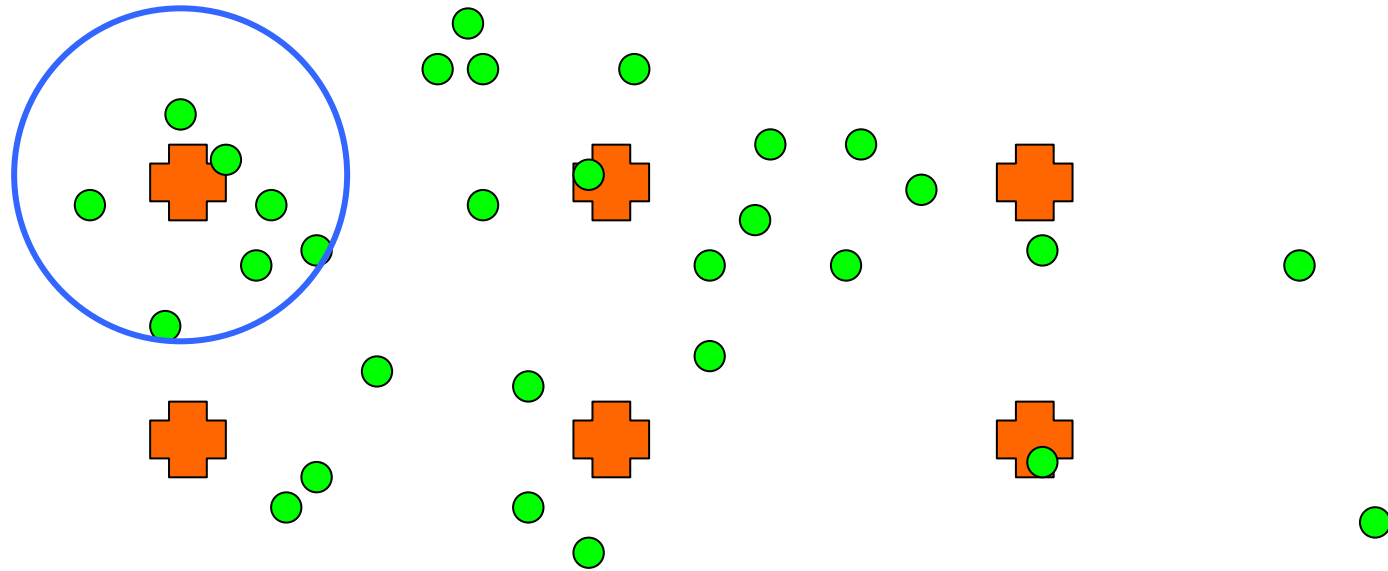
- GWR is a local estimation procedure

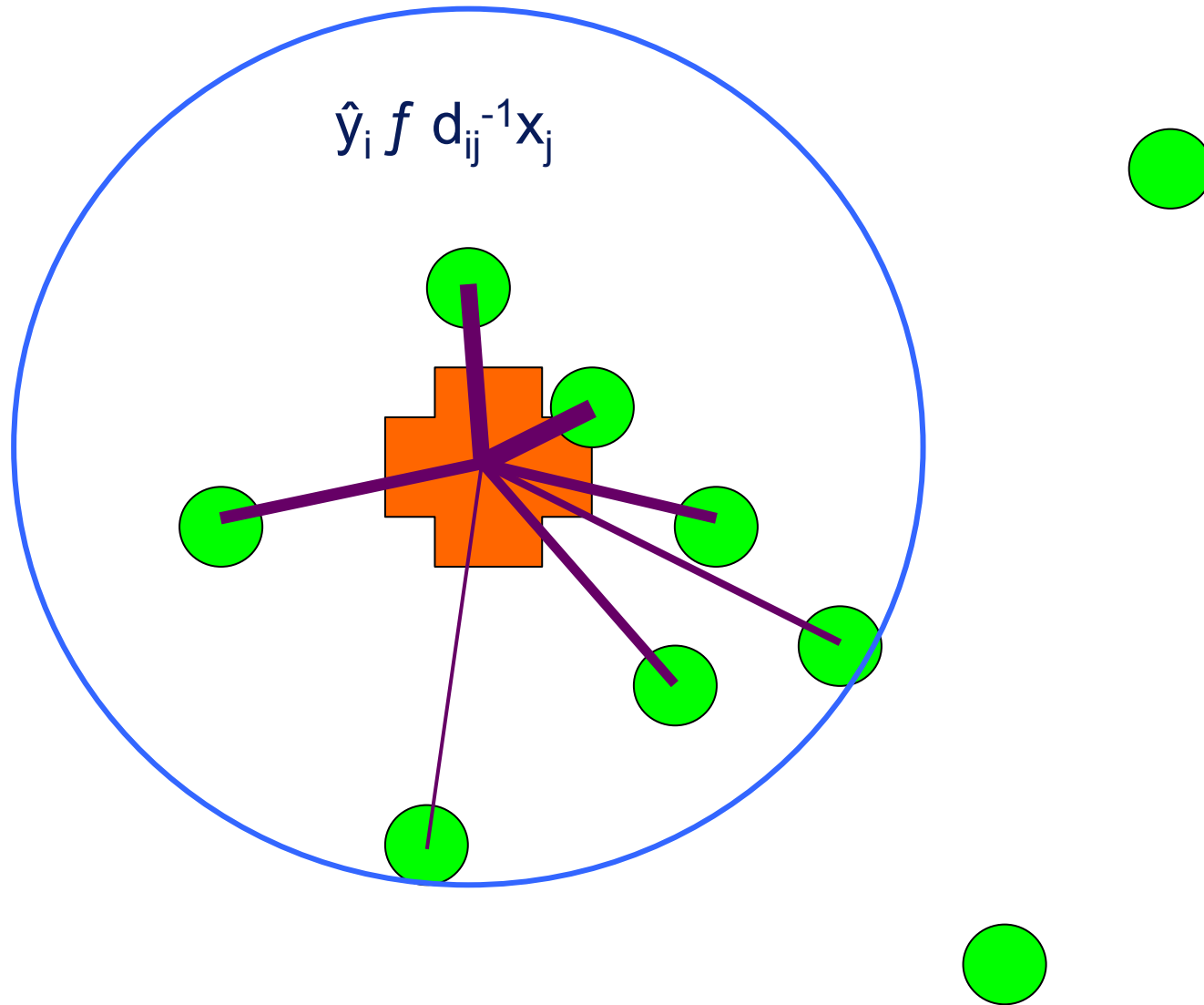
$\hat{y}_i$  is estimated by fitting a kernel around  $i$

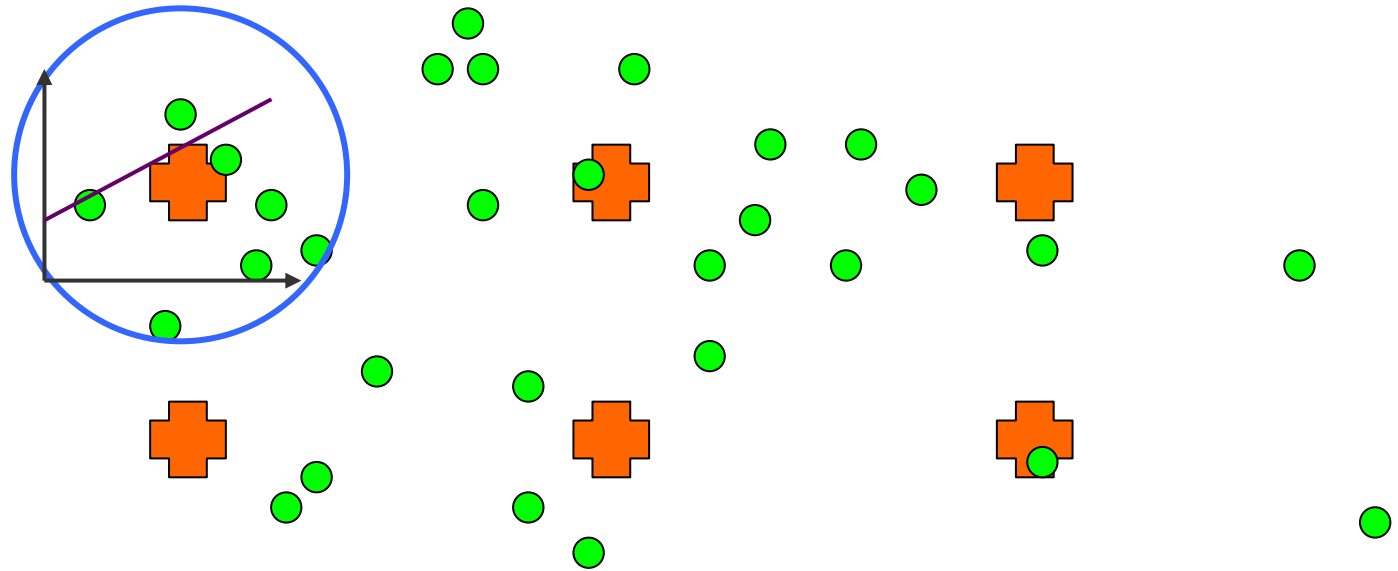
- of size  $(r \propto n_i)$  or  $(n \propto r_i)$ 
  - Non adaptive or adaptive kernel
- Within the kernel, contribution of  $j_{\text{th}}$  point to  $\hat{y}_i$ 

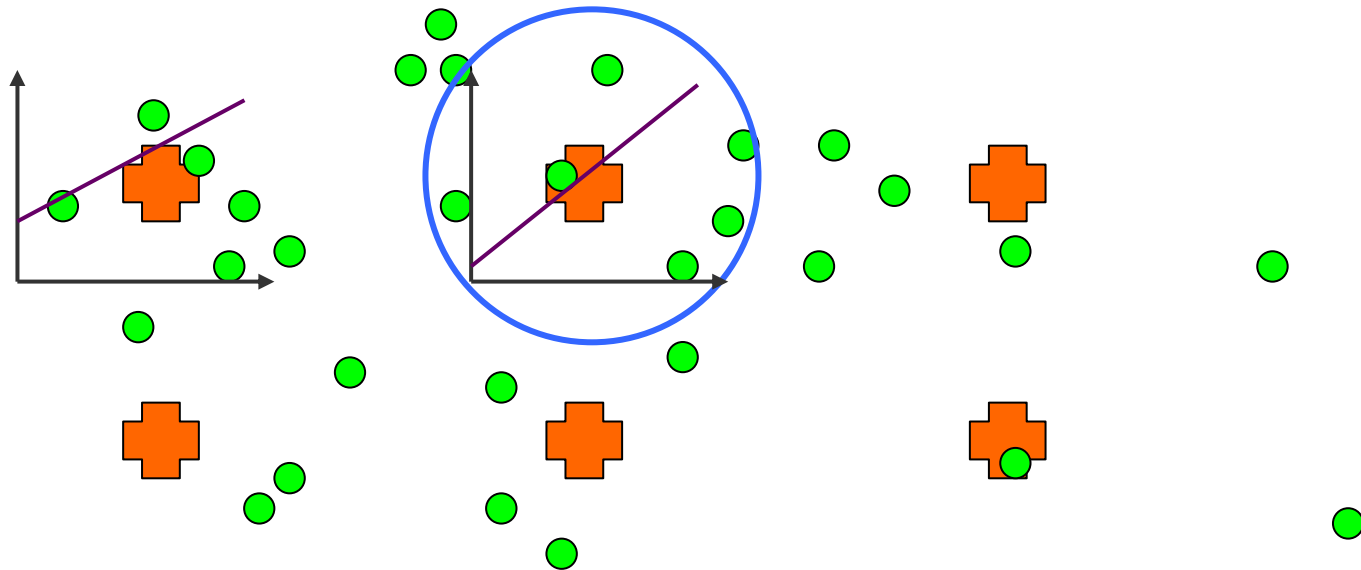
$$w_j \propto d_{ij}^{-1}$$
  - Inverse distance weighting
  - Because  $\beta_k(u,v)$  is assumed to be a continuous

# Simple GWR estimation process

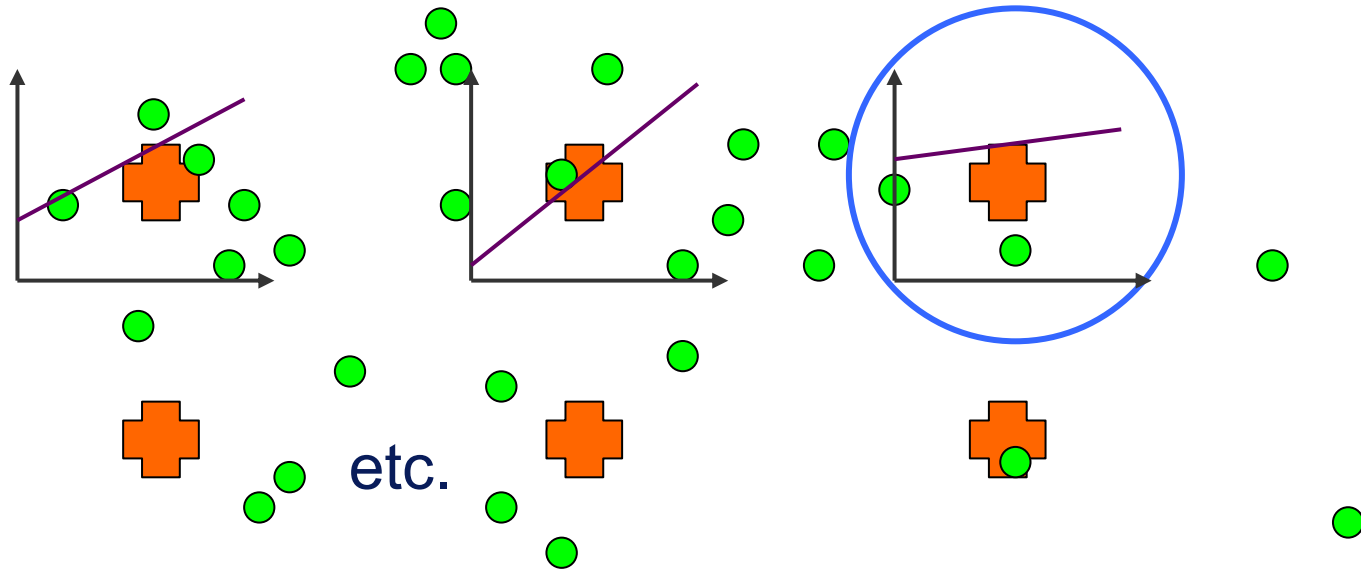




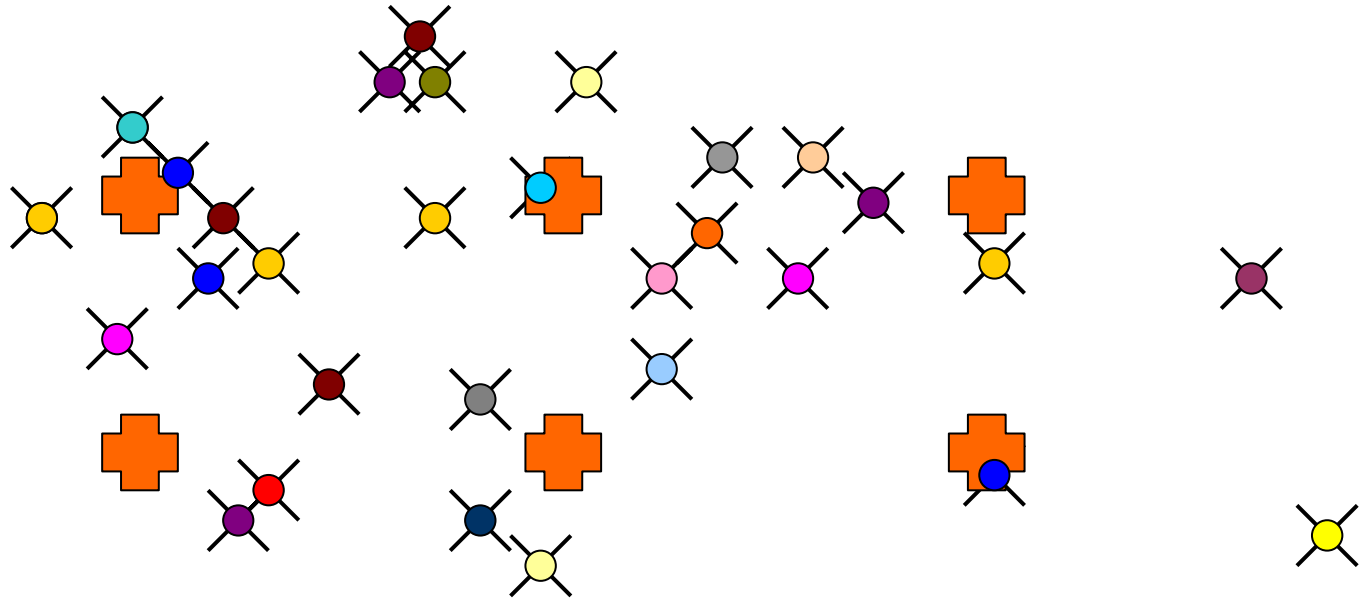




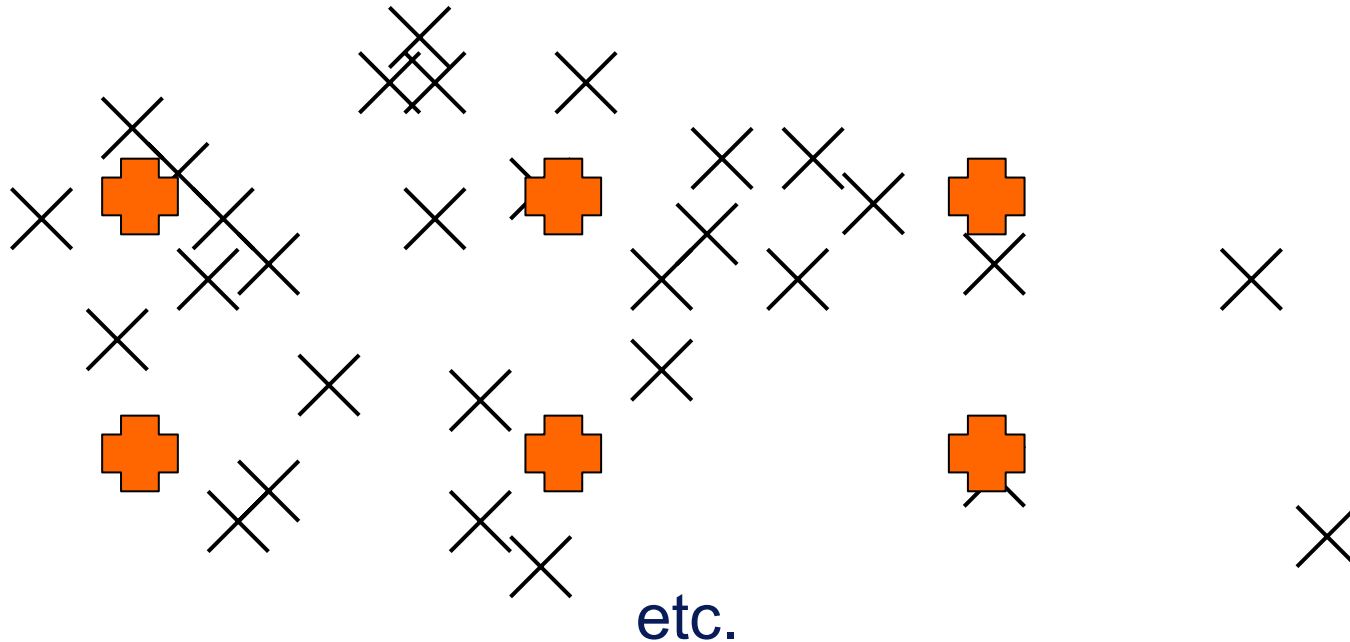
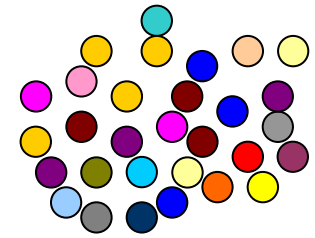
- Can then examine  $\beta_k(u_i, v_i)$  for all  $i$  and  $k$  and look for significant variation in the model parameters across the study region



# GWR simulation process (for model diagnostics)



- Geography of the data collection points and the overall distribution of the data attributes are fixed.
- Given the data and given the study region is there any statistical evidence for a non-random, non-stationary process?



# Computationally repetitious

- Create spatial index
- Calibrate search window
- Fit model
  - Estimate model in 'search window' 1
  - Estimate model in 'search window' 2
  - ...
  - Estimate model in 'search window' l
- Re-distribute data → replicate 1
  - ...
  - Estimate model in 'search window' l
- Re-distribute data → replicate 2
  - ...
- ...
- Re-distribute data → replicate m
  - ...
- Pool results and calculate diagnostics

## 'Embarrassingly parallel'

- Algorithmically, GWR can be reduced to two functions
  - one to calculate the weights
  - the other to fit the regression
- $d$  is one of  $n$  subsets of the core dataset  $D$ . Weights are calculated using function 1 concurrently applied to each of the  $n$  subsets. The data, now including weights, are passed to function 2 (a weighted regression fit), concurrently applied to each of the  $n$  subsets to give an outcome,  $z$ . The results are then pooled for a final assessment.

$$d \subset D$$

$$D \cup \{d_1, d_2, \dots, d_n\}$$

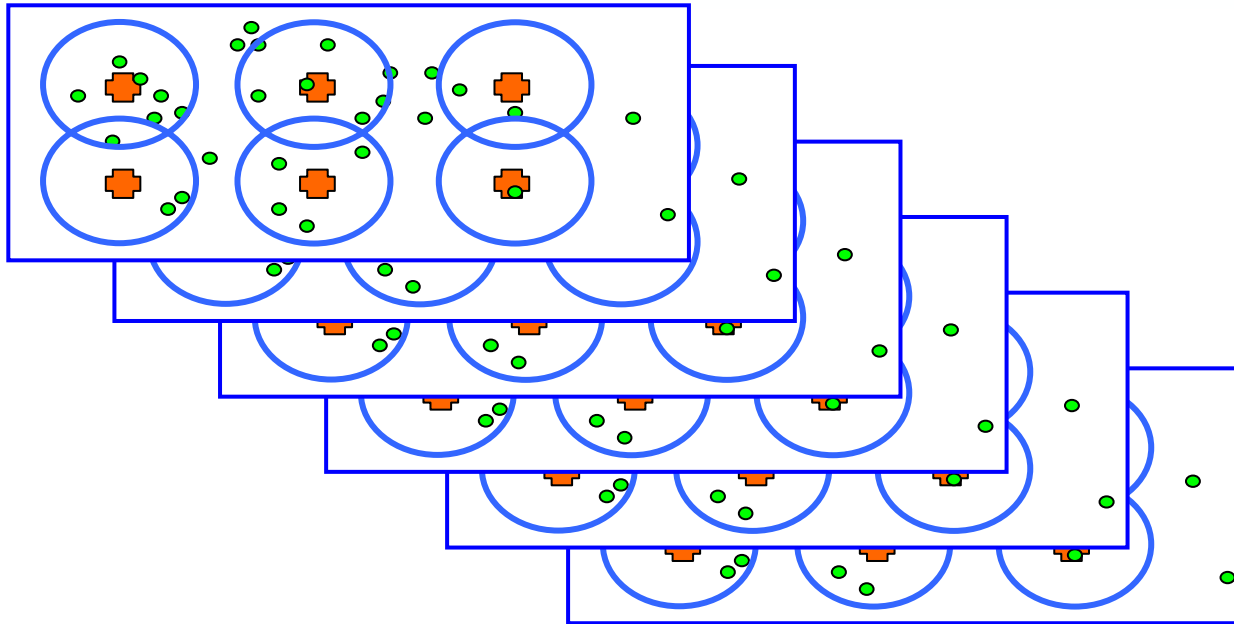
$$w_d \leftarrow f_1(d)$$

$$W \cup \{w_{d1}, w_{d2}, \dots, w_{dn}\}$$

$$z_d \leftarrow f_2(d, w_d)$$

$$Z \cup \{z_{d1}, z_{d2}, \dots, z_{dn}\}$$

## e.g. GWR estimation & simulation process



- $l \times (m + 1)$  instances of a regression fitting function
  - $l$  is the number of estimation points
  - $m$  is the number of replications
- The function and the core dataset never change: only the subset of the data being used in each case does

# Linking GWR to the computational Grid

- LOCAL
  - There are four versions of GWR
    - GWR with Visual Basic GUI
    - GWR as Fortran 77 code
    - GWR as R functions (research team's code)
    - GWR as spgwr R package
- REMOTE
  - School of Geographical Sciences, Bristol Condor pool
  - University of Bristol Condor pool
  - National Grid Service (NGS)
- Some options
  - Link R code to local cluster using *Snow* package
    - provides a high-level interface for using a workstation cluster for parallel computations in R
      - <http://www.sfu.ca/~sblay/R/snow.html>
  - Use Fortran code, interfacing with NGS
  - Use GROWL as middleware between R and NGS

Option	+	-
Snow on local cluster	Useful for Bristol as a quantitative ESRC research method node	Access for other users?
Fortran code with NGS	Easiest?	Rather <i>ad hoc</i> and 'standalone'. Ease of further development?
R with GROWL and NGS	Collaborative, interoperable, potential for further development through R	

## Summary

- GWR, like many other methods of **localized spatial analysis**, is characterised by **multiple repeat testing** as the **data are divided into geographical regions** and also **randomly redistributed** many times to simulate confidence intervals.
- The computational grid offers possibility to speed-up the process by **running GWR's sequences of calibration, analysis and non parametric simulation in parallel**.
- By exploring the suitability of GWR to the Grid environment and using it to develop a GWR-based index of deprivation in England and Wales, the project will consider the greater potential of e-social science for **data-rich spatial analysis**.