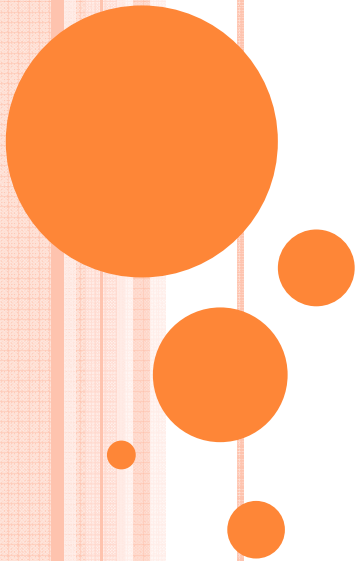
Harris, R. J., Brunsdon, C., & Grose, C. (2008). Geographical statistics and the grid: grid-enabled modelling of local and global geographies.

Peer reviewed version



Geographical Statistics & The Grid

Rich Harris, Chris Brunson and Daniel Grose

(Universities of Bristol, Leicester & Lancaster)

Outline

- About Geographically Weighted Regression (GWR)
  - Chris Brunsdon, Department of Geography, University of Leicester
  - cb179@le.ac.uk
- An example to illustrate a problem of using GWR with large datasets
- ‘The solution’
  - Rich Harris, School of Geographical Sciences, University of Bristol
  - rich.harris@bris.ac.uk

Local Vs Global Statistics

- Global
  - similarities across space
  - single-valued statistics
  - non-mappable
  - GIS “unfriendly”
  - search for regularities
  - aspatial
- Local
  - differences across space
  - multi-valued statistics
  - mappable
  - GIS “friendly”
  - search for exceptions
  - spatial

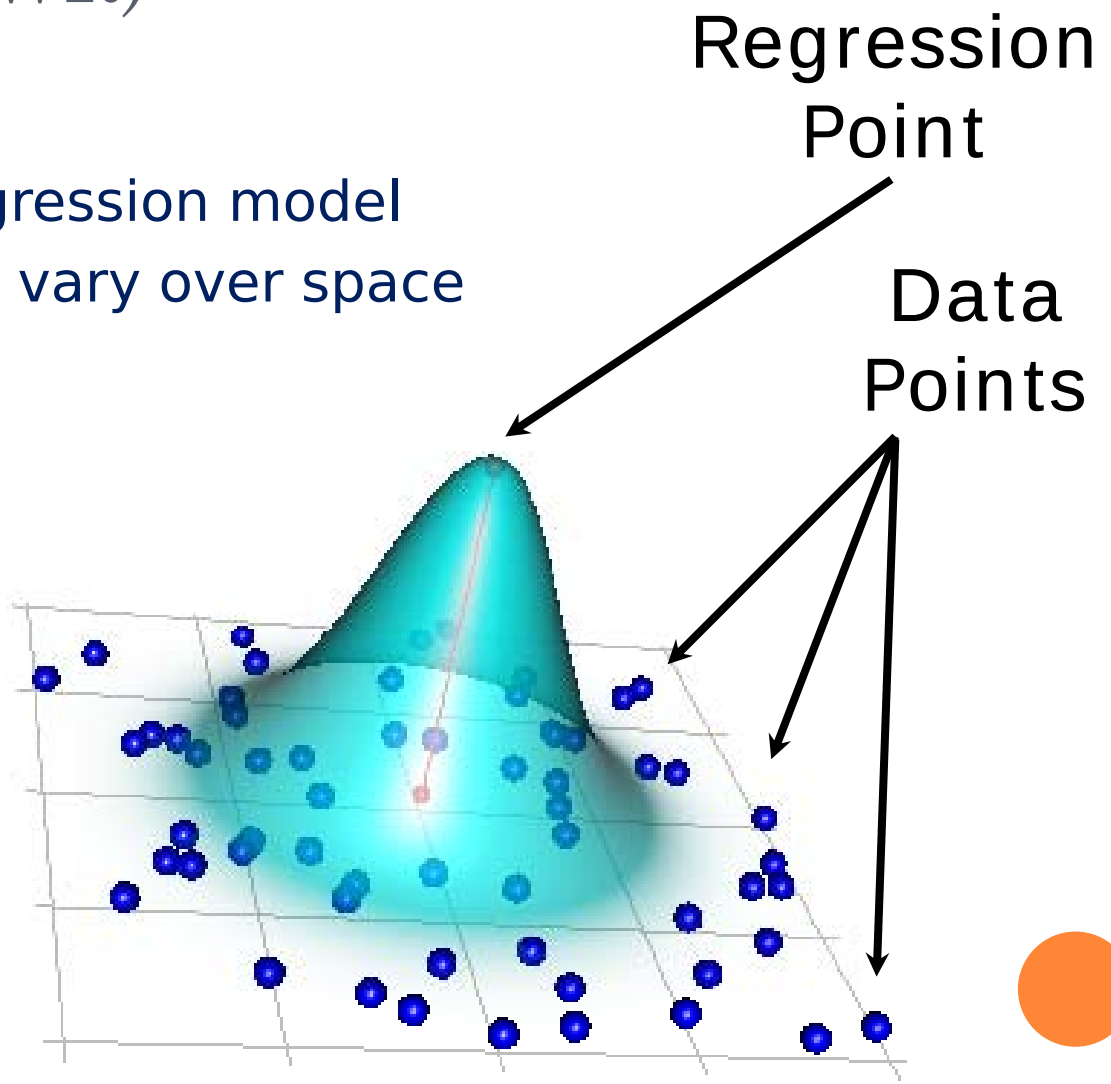
- Local statistics are spatial disaggregations of global statistics

Why might relationships vary spatially?

- Sampling variation
- Relationships intrinsically different across space e.g. differences in attitudes, preferences or different administrative, political or other contextual effects produce different responses to the same stimuli.
- Model misspecification - suppose a global statement can ultimately be made but models not properly specified to allow us to make it. Local models good indicator of how model is misspecified.

Geographically Weighted Regression (GWR)

- What is it?
  - Extension of regression model
  - Allows model to vary over space
- How it works...



In GWR we can also…

- estimate local standard errors
- calculate local leverage measures
- perform tests to assess the significance of the spatial variation in the local parameter estimates
- perform tests to determine if the local model performs better than the global one

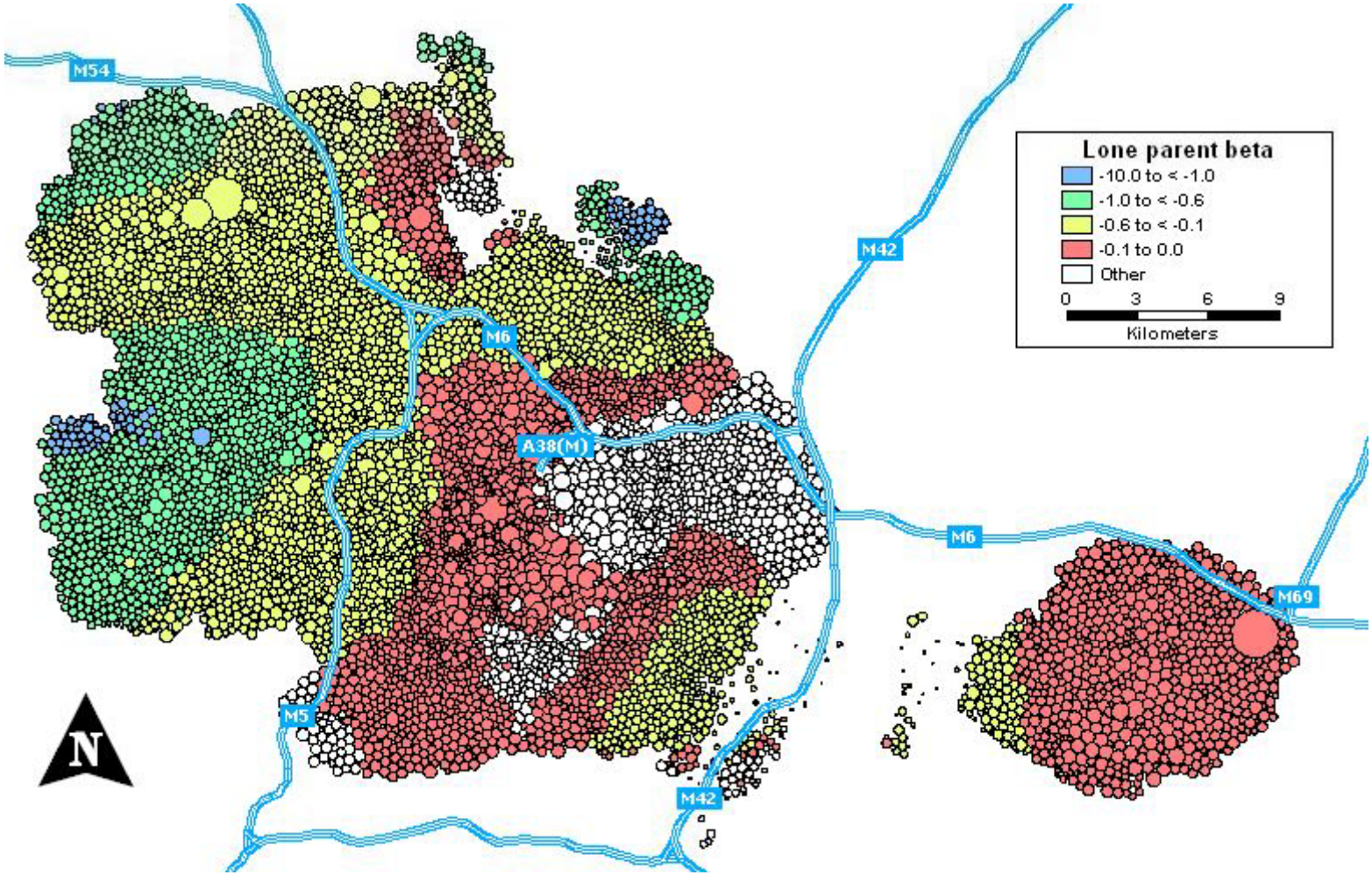
But

- Computationally very demanding
- Need to fit weighted regression models in several places
- Sometimes not viable on a single computer
- How do we address this problem?

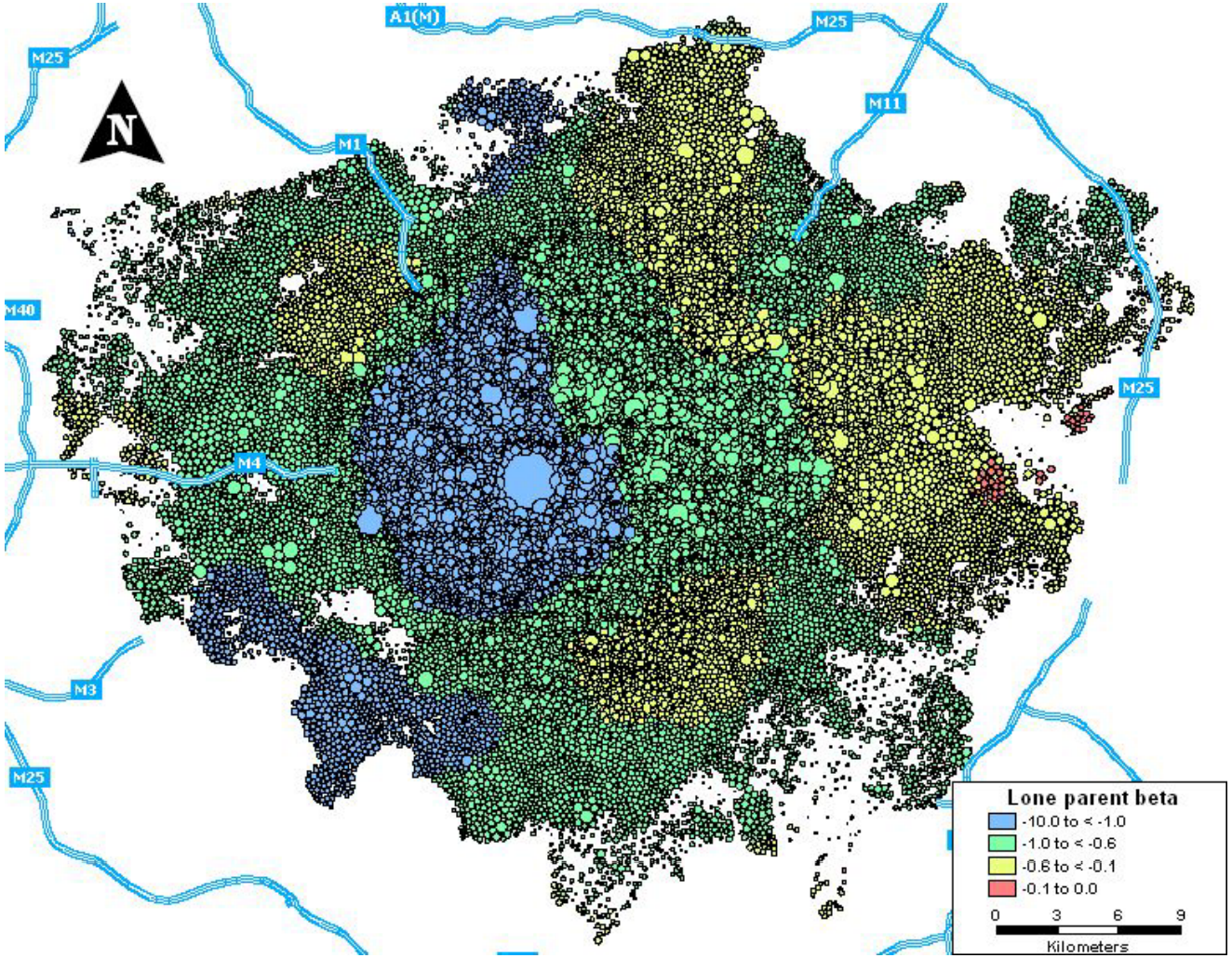
Example

- Y: Proportion of households without a car
- X_1: Proportion of persons of working age unemployed
- X_2: Proportion of households in public housing
- X_3: Proportion of households that are lone parent households
- X_4: Proportion of persons 16 or above that are single
- X_5: Proportion of persons that are “white British”
- n = 165,665

Spatial variation in the lone parent coefficient (West Midlands)



Spatial variation in the lone parent coefficient (London)



Scaling problems

- There are n regression models
- But actually there are many more:
  - n× g
  - g is the number of iterations to optimize the bandwidth, b
- And you also need to calculate the distance matrix, D: (n^2)

Takes a long time!

- If n = 100,000, on a single processor
  - Would take about half a day to calculate D
  - Would take about a fortnight to find b
- But GWR is intended for exploratory analysis!
- The main bottleneck is the calibration of b
  - Because the regression calculations are $O(n^3)$, the distance calculations are $O(n^2)$

Fortunately

- The regression models are fitted entirely independent of each other. The results are pooled and compared *at the end*.
- The process is sequential but it can also be embarrassingly parallel
  - For GWR, the (distance-weighted regression) function stays the same, only the data are changing.
  - Each spatial subset is handled separately from the next.
- True of many methods of spatially localized analysis
- Suitable for a computational grid
  - The UK's National Grid Service (NGS)

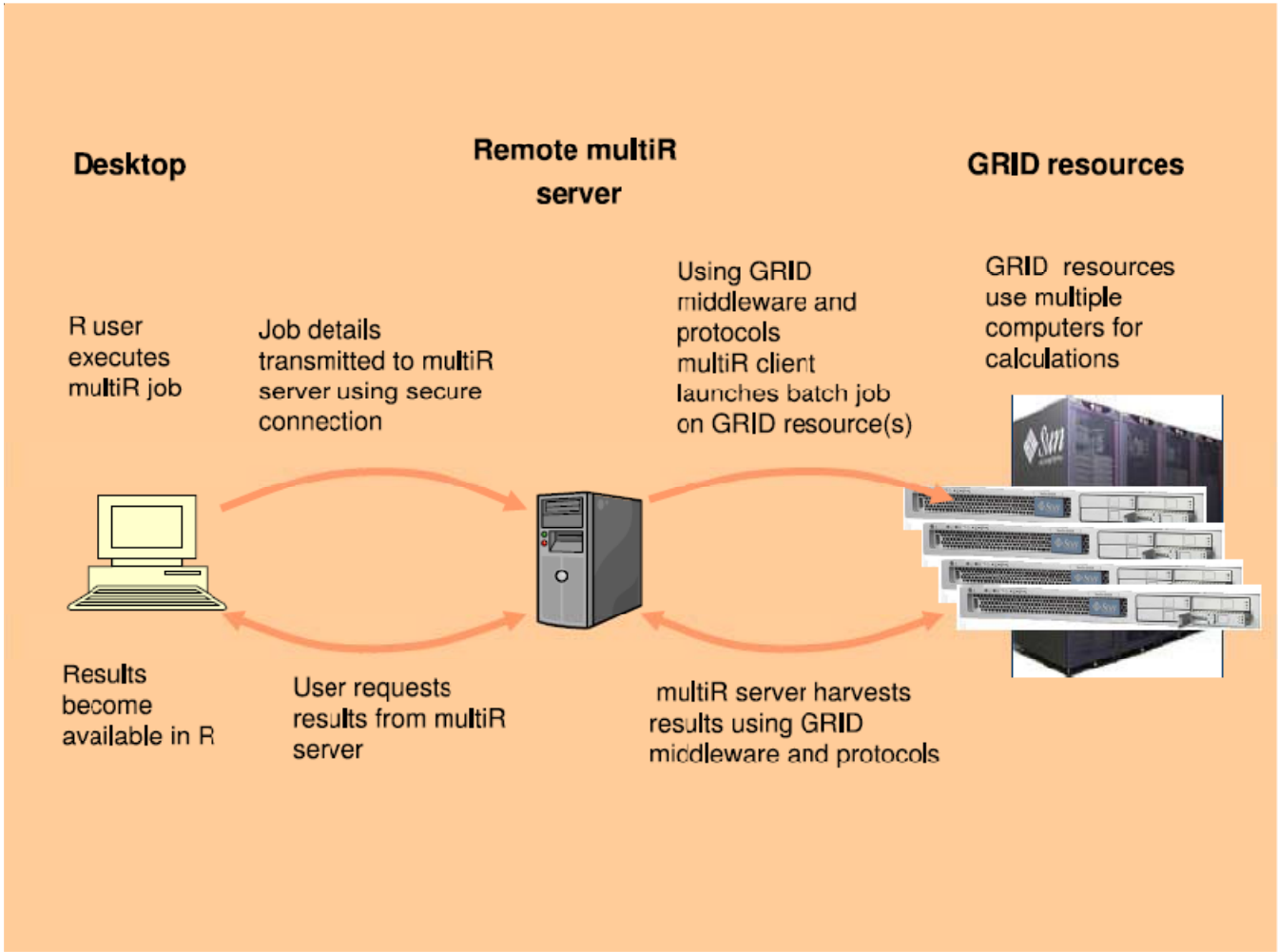
multiR and Parallel R

- R is a free software environment for statistical computing and graphics
  - http://www.r-project.org/
- There is an implementation of GWR in R
  - The spgwr package
- In R, there is a method to invoke a function a number of times with varying argument values
  - sapply
- The idea is to invoke the function on different processors running on the NGS.
- multiR (thanks to Daniel Grose) is both an ‘add in’ to R and a server (currently at Lancaster) which provides middleware between desktop R and the NGS.

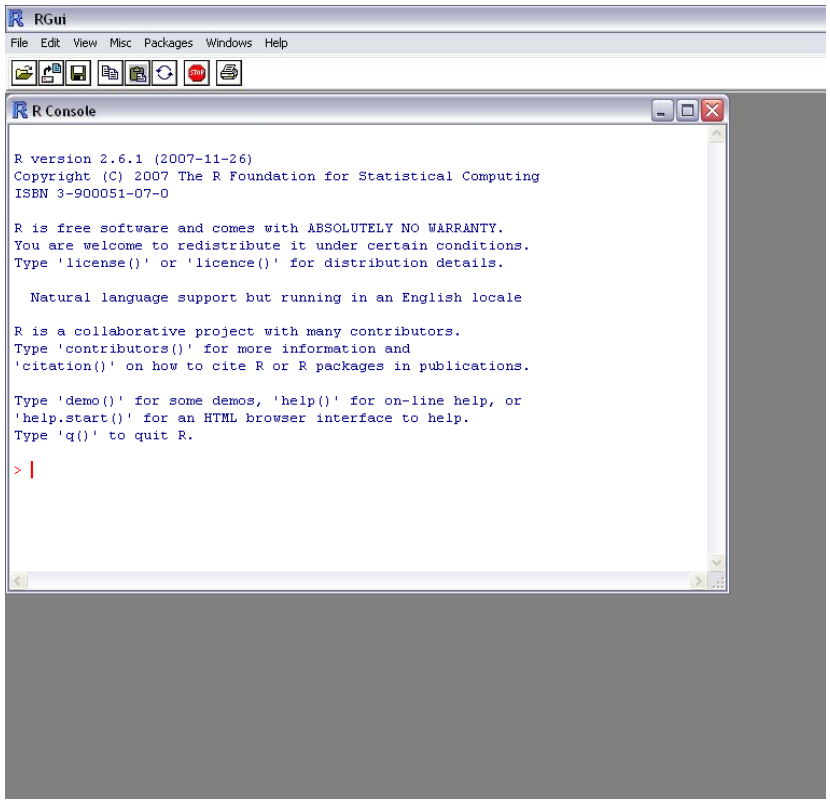
Example

- Y: Proportion of households without a car
- X_1: Proportion of persons of working age unemployed
- X_2: Proportion of households in public housing
- X_3: Proportion of households that are lone parent households
- X_4: Proportion of persons 16 or above that are single
- X_5: Proportion of persons that are “white British”
- n = 165,665

The three tier client/server architecture employed by multiR



This is R (in Windows)



- Available free from
  - http://cran.r-project.org/
- We have a tutorial
  - www.esrcsocietytoday.ac.uk
  - type ‘Grid Enabled Spatial Regression Models’ into the Search

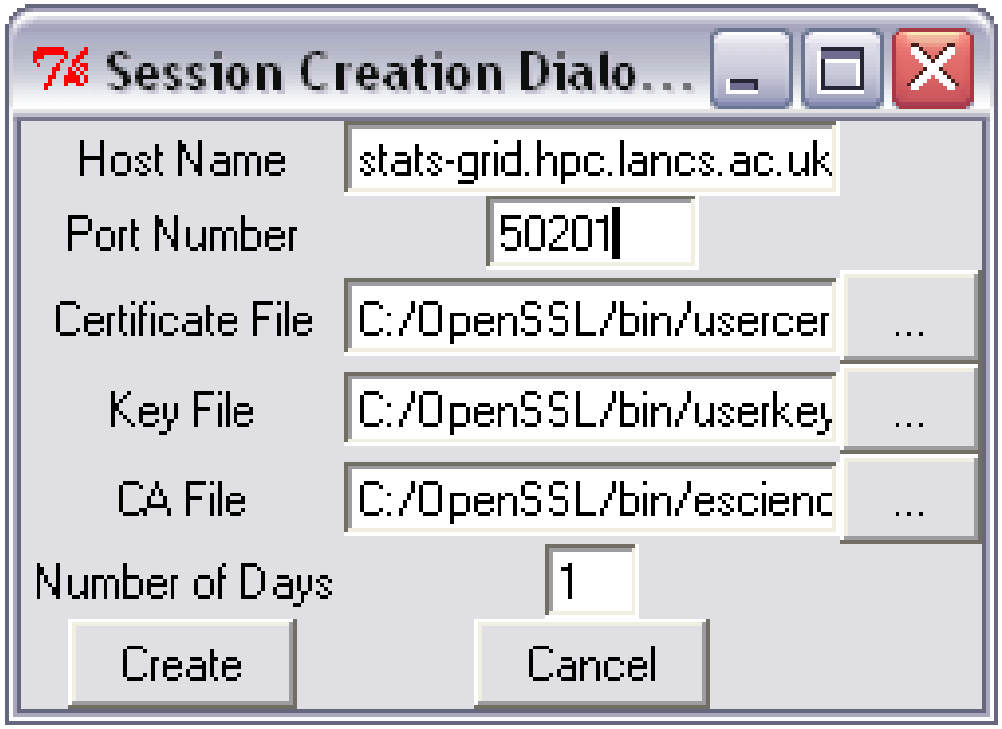
Fitting a GWR model in R

```
> library(spgwr)
> load("carsmsoa.RData")
> names(car.msoa)
[1] "Name"     "Borough"  "ProfMan"  "Renting"
  "HHNoCar"  "Easting"  "Northing"
> coords = cbind(car.msoa$Easting,
  car.msoa$Northing)
> bandwidth = gwr.sel(HHNoCar ~ Renting, data =
  car.msoa, coords)
Bandwidth: 25571.63 CV score: 4.278767
Bandwidth: 41334.45 CV score: 4.373939
…
Bandwidth: 1467.076 CV score: 2.922873
> gwr.model1 = gwr(HHNoCar ~ Renting, data =
  car.msoa, coords, bandwidth)
```

Fitting a Grid run GWR model

```
> library(spgwr.dist)
> load("D:\\Data\\GWR-
  Workshop\\Data\\Exercises\\carsmsoa.RData")
> names(car.msoa)
[1] "Name"     "Borough"  "ProfMan"  "Renting"
  "HHNoCar"  "Easting"  "Northing"
> coords = cbind(car.msoa$Easting,
  car.msoa$Northing)
> session = multiR.session.dlg()
> bandwidth = gwr.sel.dist(session, HHNoCar ~
  Renting, data = car.msoa, coords,
  max.processors = 50)
> gwr.model2 = gwr.dist(session, HHNoCar ~
  Renting, data = car.msoa, coords, bandwidth,
  max.processors = 50)
```

'The Session'



- It is loading and dealing with the various security certificates that are needed to use the NGS

Obtaining the certificates

- This is required for all NGS services
- User certificates:
  - Apply at https://ca.grid-support.ac.uk/
    - The certificate is issued for a web browser: the one you apply from needs to be the same as the one which will receive it.
  - Then need to:
    - Export the certificate from the browser
    - use OpenSSL toolkit to convert into two files (the certificate and its **key file**) and to generate a **proxy certificate**
    - See www.grid-support.ac.uk/content/view/67/184/
  - You also need to apply for an NGS account
    - http://www.grid-support.ac.uk/content/view/221/171/
- CA certificate:
  - Download from http://www.grid-support.ac.uk/content/view/182/184/

Conclusions and Caveats

- There is no point in using Grid GWR for small data sets (e.g. n < 1000)
- And you should not expect instant results with large datasets (it still takes a second or so for each regression fit and you don’t have that many processors)
- You cannot presently disconnect from R when running a Grid GWR but that should follow (and return later to collect the results)
- The software are still being tested

Potential

- multiR is more generic than GWR
- Imagine

```
function1 = function(its_parameters) {
  what it does
}
```

- Then

```
some_results = multiR(session, function1,
  list=(the_data))
```

- In other words, the function is running in parallel on the NGS
- Applications include spatial statistics, geostatistics, ‘hot spot analysis’, simulation, etc.

Acknowledgments

- The ESRC (RES-149-25-1041)
- The National Centre for e-Social Science (NCeSS)
- Particular thanks to the Centre for e-Science, Lancaster University