Understanding hotspots using Geographically Weighted Regression

Spencer Chainey

Director of Geographic Information Science

Follow me on Twitter @SpencerChainey

UCL Jill Dando Institute of Crime Science

Overview

- Identifying hotspots
- Regression – the basics
- Applying a GWR model
  - OLS model; diagnostics
  - GWR model
  - Examples:
    - Hotspots of serious violent crime
    - Hotspots of residential burglary

- Presentation: focus more on technique rather than causes
- Research parallel to RTM
- Do not expect you to be experts in GWR at the end …

Identifying hotspots

- Use hotspot analysis for operational targeting, strategic analysis, problem solving, investigative analysis, performance analysis
- Techniques: spatial ellipses, thematic mapping of geographic units, KDE, Gi*
  - Identify concentration of crime

Chainey,S.P., Tompson,L., Uhlig,S. (2008). ***The utility of hotspot mapping for predicting spatial patterns of crime***. Security Journal

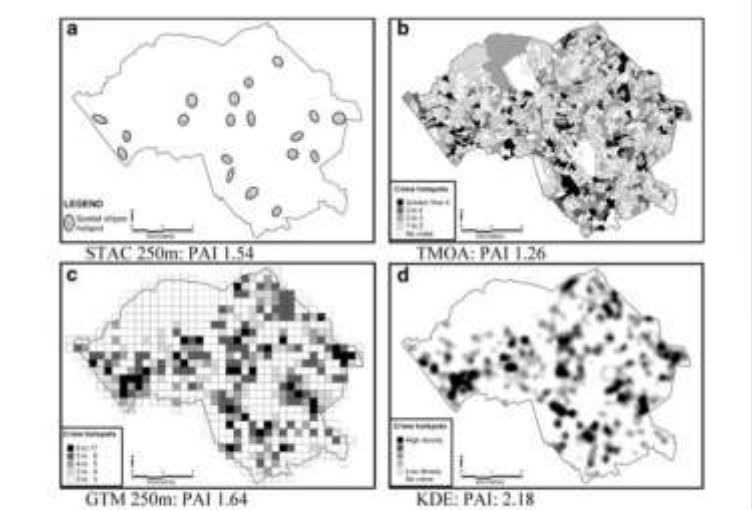
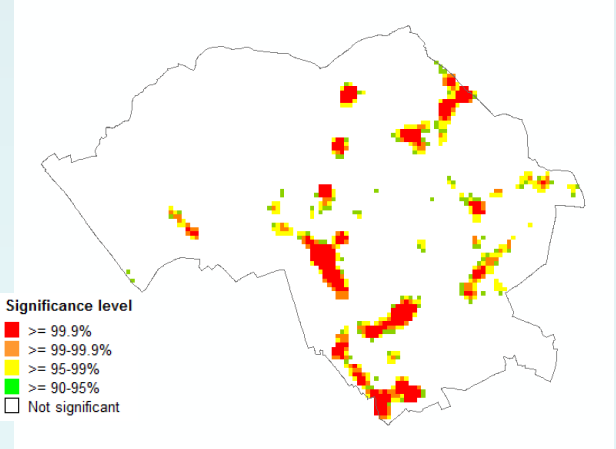
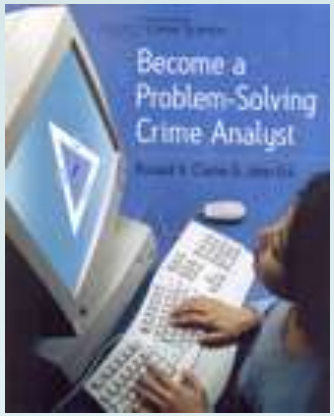


Figure 4. Hotspot maps generated from 3 months of residential burglary input data (measurement date of the 1 January 2003) using (a) STAC, (b) thematic mapping of output areas, (c) grid thematic mapping and (d) KDE. Each map is shown with its PAI value, based on 1 month of measurement data.



Identifying hotspots

- Hotspot mapping: start of a process
  - further analysis seeks to understand what is causing it
- Relies on skill of analyst
  - knowledge of appropriate theory
  - Use of accurate data
  - Analysis: fusion of data and theory from which the 'cause' of the problem can be interpreted
- Tendency to use other descriptive analytical techniques (offender network analysis, temporal analysis, MO etc)
- Need to make more use of inferential techniques to provide statistical support



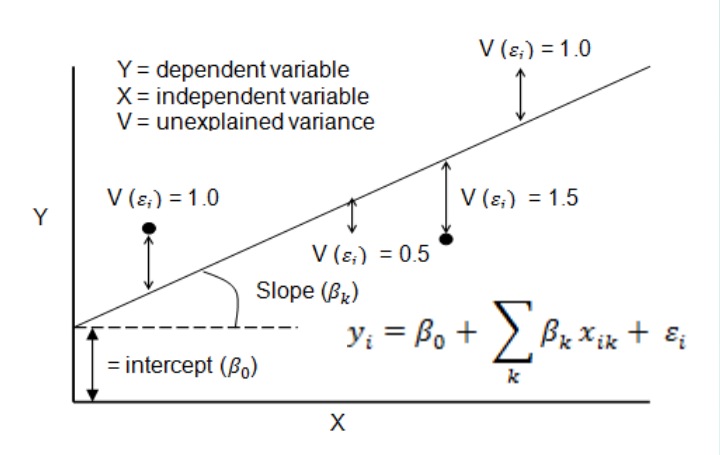
Regression analysis – the basics

- Test relationship between one phenomenon (e.g. a crime problem) and factors that are considered to cause the phenomenon
  - Statistically determines if factors help explain the phenomenon
  - The strength in the relationship
  - Identifies if other key factors are missing

Regression analysis – the basics

- Ordinary Least Squares regression (OLS)

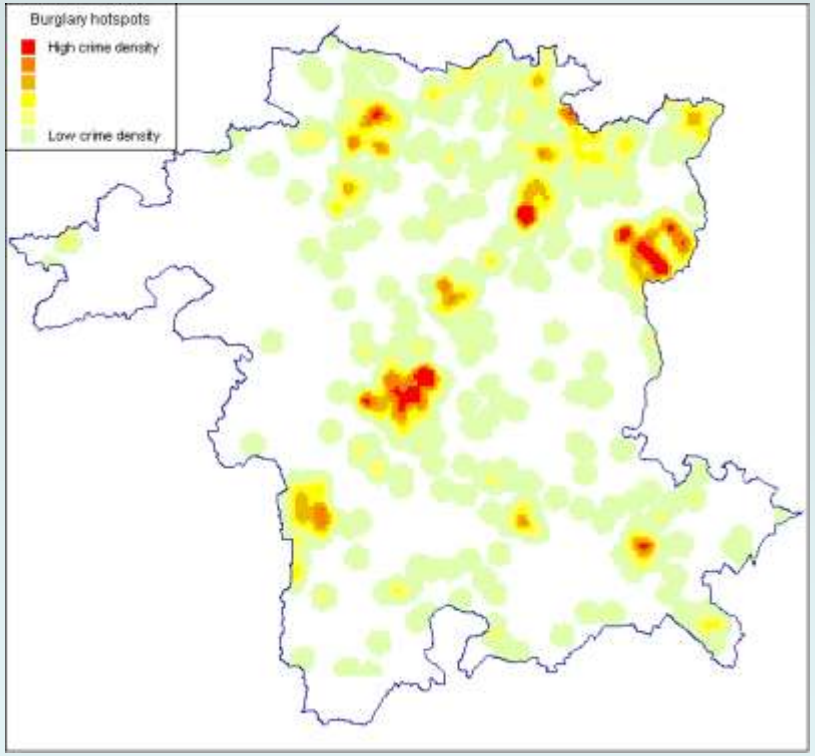
  Scatterplot: phenomenon that is being studied and variable that is being used to explain it
  - Aim is to fit straight line on chart that minimises location of points from this line
  - If all data fits on regression line, we have a perfect relationship!
  - Likely to have certain amount of 'unexplained variance'



- Call the phenomenon under study: dependent variable (y)
- Causal factors: explanatory variables (x) (or independent variables)

Regression analysis – the basics

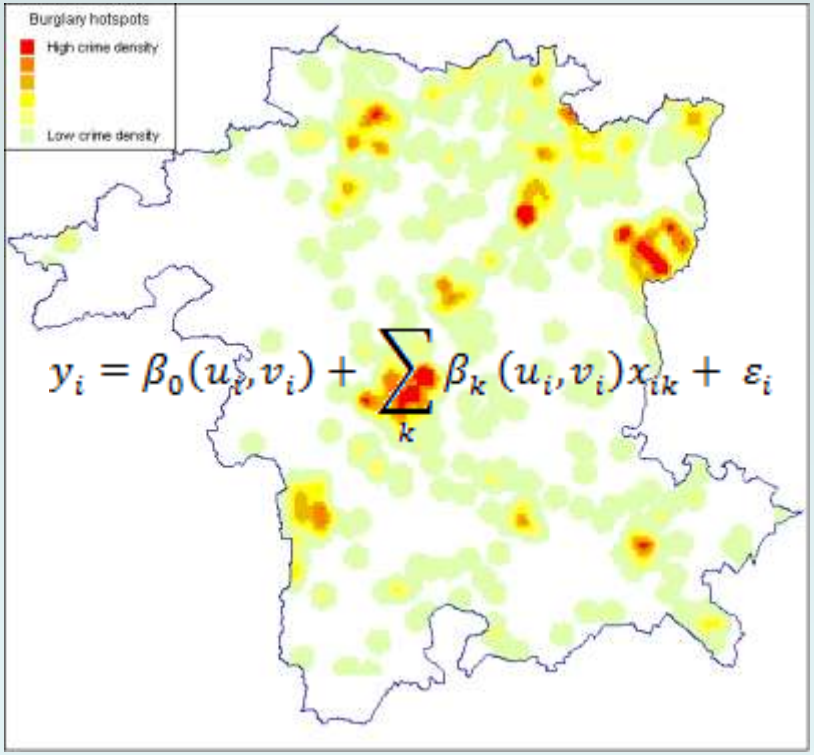
- OLS assumes relationships are consistent geographically
  - Reasons that explain phenomenon in one location, are exactly the same as those in another
  - e.g. if hotspots of burglary are present in two locations, assuming that reasons for these hotspots are exactly the same, influenced equally by the factors causing burglary.



- Reality: unlikely to be the case
  - Rarely do variables exhibit spatial stationarity i.e. the variable is not equally constant in all directions

Geographically weighted regression
(Fotheringham, Brunsdon and Charlton, 2002)

- GWR recognises that relationships between variables are likely to vary across space
  - Generates regression equation for explaining this relationship for every area on a map
- Can do in ArcGIS 9.3 or higher
  - Or use GWR software or R code from National University of Ireland



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

Geographically weighted regression

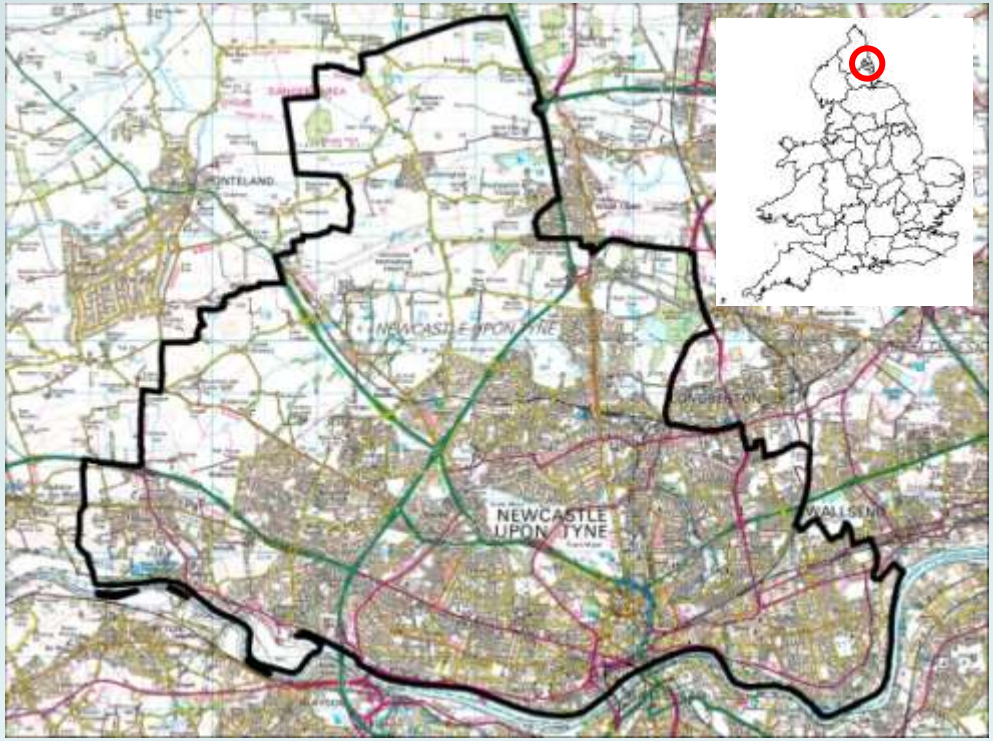
- To produce a good GWR model:
  - OLS model: to see if we have a relationship worth exploring
  - Statistical diagnostic tests: model performance, significance and bias
  - Then proceed to GWR
- Two approaches to explanatory variable selection:
  - hypothesis testing: select variable based on it providing a plausible reason to explain why hotspot is present
  - exploratory approach: apply a range of disparate variables, then attempt to explain the relationship
    - Reality, choice of variables naturally be selective in terms of data availability, data quality, and curiosity

      i.e. rather than selecting number of variables at random, we begin this by determining variables that would be interesting to test

Geographically weighted regression

Examples:

- Newcastle-upon-Tyne, England (Pop: 190,000)
- Serious violent crime (hypothesis testing; grid cells)
- Residential burglary (exploratory; Census areas - Lower Super Output Areas)

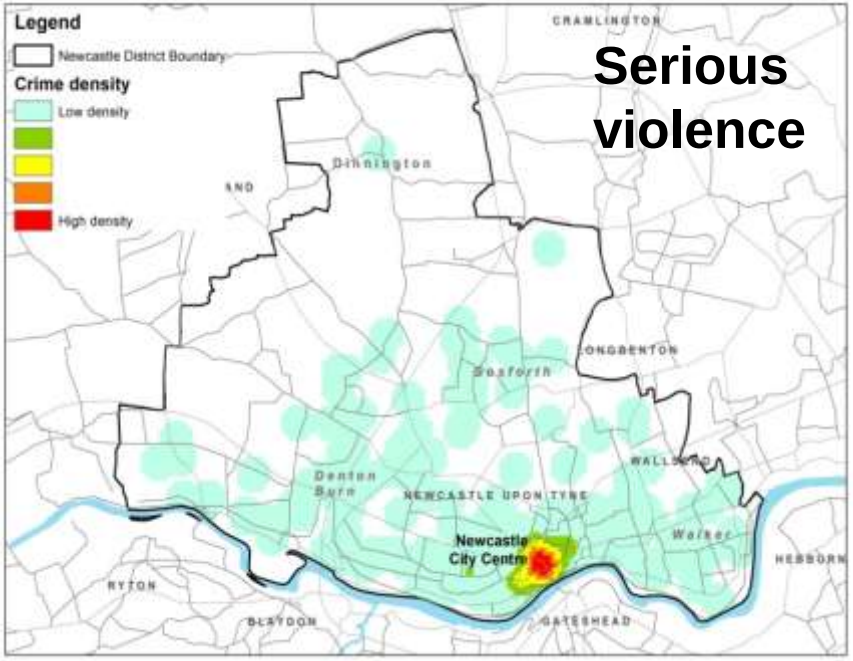


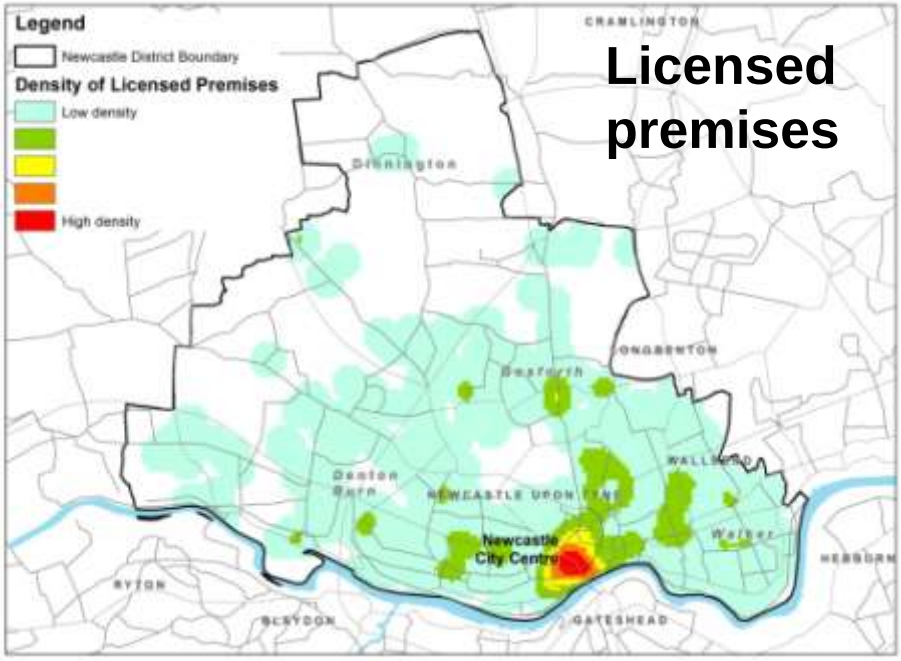
Steps (all in ArcGIS):

- Hotspots (and scatterplots)
- OLS and diagnostics:
  - Model performance, significance and bias
- GWR

Understanding hotspots of serious violence
(Serious bodily harm, wounding)

- Hypothesis: hotspots of serious violence are related to locations where there is a clustering of pubs, bars and night clubs (licensed premises)
- Would usually consider more than one variable, but for ease of explanation using one in first example





Understanding hotspots of serious violence

Model performance: is there a statistical relationship?

Explanatory variables:	Adjusted R^2	Coefficient for explanatory variable(s)	Koenker statistic	Robust probability
Licensed premises	0.87	0.44	Statistically significant	Statistically significant

Interpretation (from OLS and other diagnostics):

- Adjusted R^2: very strong positive relationship i.e. licensed premises explain 87% of variation in serious violence
- Coefficient: for every one unit change in licensed premises, serious violence increases by 0.44. Means little though if the relationship is not statistically significant
- Koenker statistic: if statistically significant then we must use robust probability rather than standard probability results
- Robust probability (of T test scores): significant
- **Our explanatory variable, licensed premises, is statistically valuable to the model (using OLS regression)**

Understanding hotspots of serious violence

Model significance: is there a statistical relationship?

- Joint F statistic or Joint Wald statistic: 0.01, significant to 99%

  (if Koenker is statistically significant we use Joint Wald statistic)
- Koenker statistic: significant (heteroscedasticity and/or nonstationarity in model)
  - Do the explanatory variables have a consistent relationship with the dependent variable
  - Stationarity: spatial processes represented by explanatory variables behave the same everywhere
  - Heteroscedasticity: variation in relationship between dependent variable and each explanatory variable does not change with changes in the magnitude of explanatory variable

    i.e. relationship between serious violence and licensed premises was just as accurate in areas where there were small number of licensed premises than in areas where there were high numbers of licensed premises

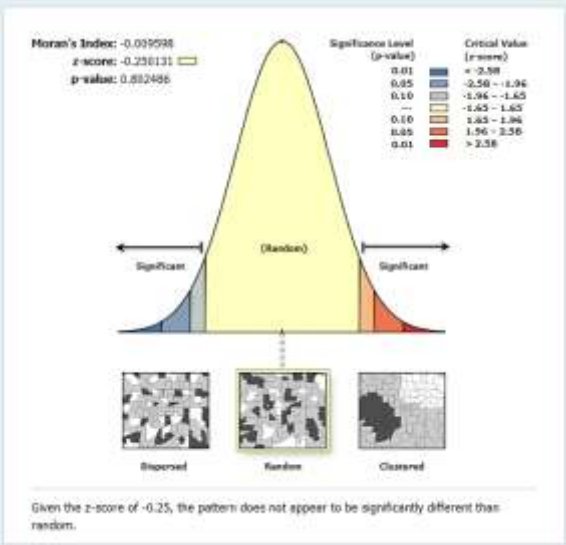
Interpretation

- Relationship between serious violence and licensed premises varies across space, and is therefore a good candidate for GWR

Understanding hotspots of serious violence

Model bias: is there a statistical relationship?

- Jarque-Bera statistic: not significant
  - Are residuals normally distributed? (difference between observed value of dependent variable and estimated value)
- Moran's I of residuals: not clustered
  - If clustered then suggests key explanatory variable(s) missing from our model



Interpretation

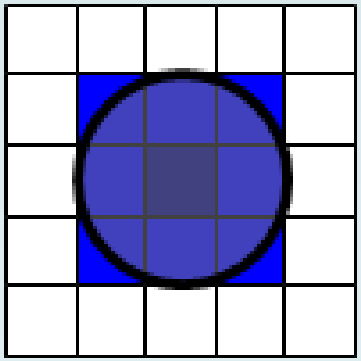
- Jarque-Bera: model is not biased
- Moran's I: residuals are not clustered
- ✓ Passed tests required prior to producing GWR model

- AICc (compare model performance the lower the better): 1899
- Reminder: all this can be done in ArcGIS/GWR

Understanding hotspots of serious violence

GWR – bandwidth and model selection

- Need to specify bandwidth (size of neighbourhood over which relationships between the dependent variable and the explanatory variables will be explored)
  - Bandwidth size optimisation methods: Cross Validation and AICc (aim is to minimise AICc value)
  - User specified bandwidth size
  - Fixed or adaptive



OLS AICc	GWR AICc (bandwidth = 1297)	GWR CV (bandwidth = 1297m)	GWR User Defined, fixed (bandwidth = 285m)	GWR Adaptive AICc	GWR Adaptive CV
1899	1710	1710	7044	1668	1668

Interpretation

- Adaptive AICc/CV was deemed to be best model
- GWR models better than OLS (except for user defined 285m bandwidth)

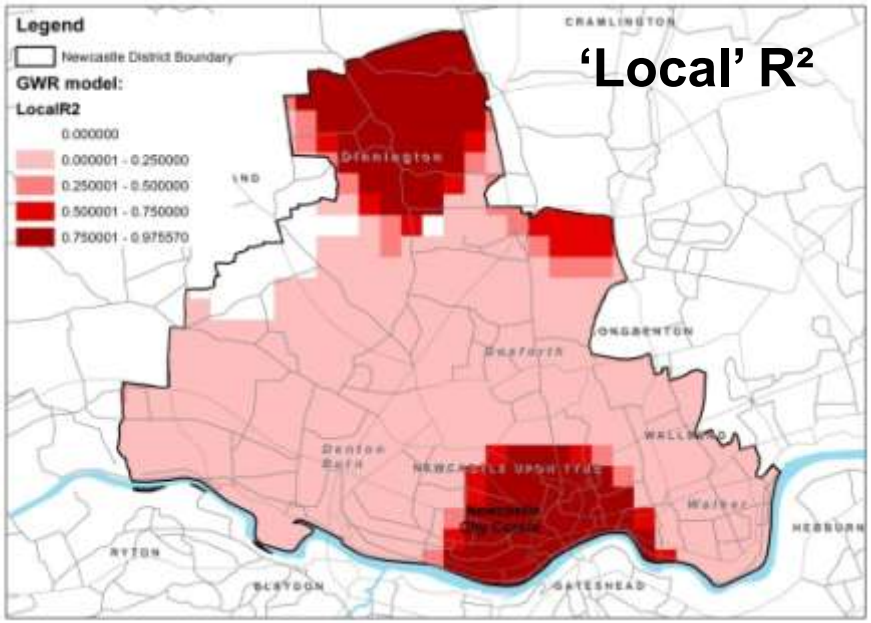
  i.e. model that recognises spatial variations in relationship between serious violence and licensed premises is a better model

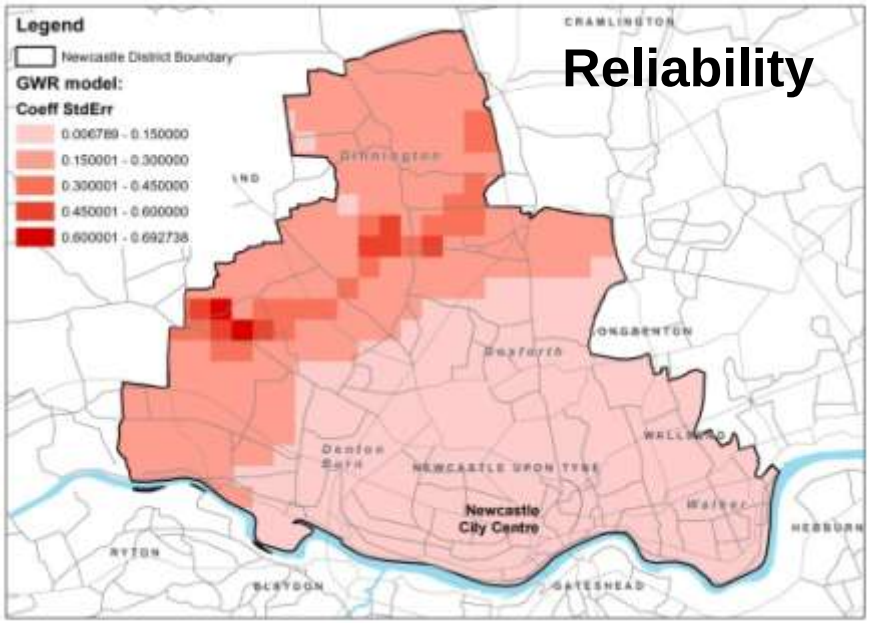
Understanding hotspots of serious violence

GWR – mapping the results

Model (adaptive AICc)

- Adjusted R^2 of 0.92 (recall OLS model: adjusted R^2 of 0.87)
  i.e. GWR model stronger at capturing relationship between licensed premises and serious violence
- Relationship between location of licensed premises and serious violence varies across Newcastle: strongest in and around the city centre (and reliable) and to the north of the district (but less reliable)

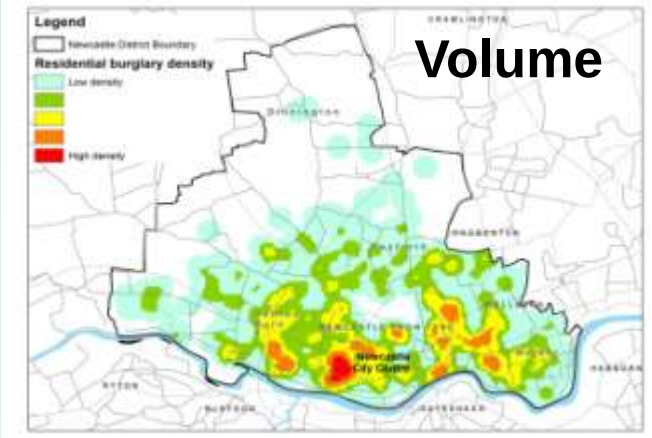


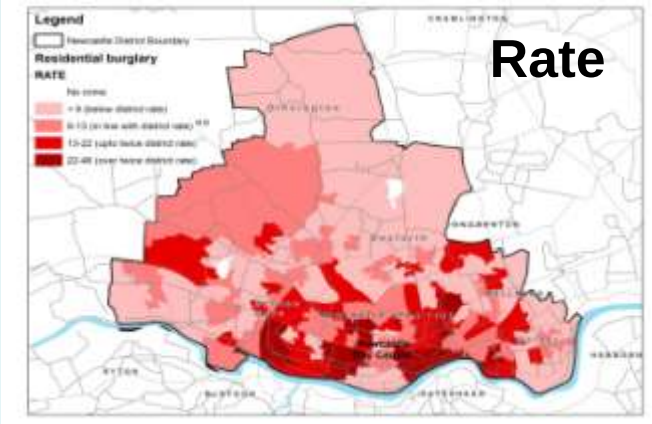


Understanding hotspots of residential burglary

Exploratory approach

- N = 1304; rate (Newcastle) = 11.1 per 1000 household per annum





Dependent variable: burglary rate

Explanatory variables:

- Population with a low level of educational attainment (Census 2001: Less than Key stage 2)
- Asian population (Indian, Pakistani and Bangladeshi)
- Black population (Census 2001)
- Chinese population (Census 2001)
- Student population (aged 18 or over) (Census 2001)
- Offenders of serious acquisitive crime (burglary, vehicle crime, street robbery, street theft) (Northumbria Police)
- Income deprivation (Index of Deprivation 2010)
- Education, skills and training deprivation (Index of Deprivation 2010)
- Living environment deprivation (Index of Deprivation 2010)

Understanding hotspots of residential burglary

Model performance (nine explanatory variables):

- Adjusted R^2: 0.41
- Koenker statistic: not significant therefore use probability from T test
- Probability: Asian pop, student pop, offenders of SAC, income deprivation significant
- VIF (>7.5): identifies variable redundancy

Explanatory variables:	Coefficient for explanatory variables	Probability	VIF
Low level of educational attainment	0.011	0.9	5.3
Asian population	0.158	0.0002*	1.3
Black population	-0.944	0.131	1.7
Chinese population	0.21	0.296	1.8
Student population	0.22	0.0006*	2.7
Offenders of serious acquisitive crime	0.876	0.0000*	2.0
Income deprivation	-20.1	0.027*	7.6
Education, skills and training deprivation	0.071	0.223	10.2
Living environment deprivation	0.044	0.492	1.3

Interpretation

- *Education, skills and training deprivation* (VIF > 7.5) likely to be 'telling the same story' as *low level of educational attainment*
- Only statistically significant variables have role in our model: re-run model with these
- *Income deprivation* VIF > 7.5, but removing other variables may remove redundancy

Understanding hotspots of residential burglary

Model performance (four explanatory variables):

- Adjusted R^2: 0.42
- Koenker statistic: significant
- Probability: all significant
- VIF (>7.5): none
- AICc: 1133 (initial model 1138)

Explanatory variables:	Coefficient for explanatory variable(s)	Robust probability	VIF
Asian population	0.141	0.003*	1.0
Student population	0.22	0.0000*	1.1
Offenders of serious acquisitive crime	0.92	0.0000*	1.8
Income deprivation	-9.65	0.034*	2.0

Interpretation

- Model is better than initial model
- But only tells less than half of the 'story'

  i.e. 42% of explanation for residential burglary

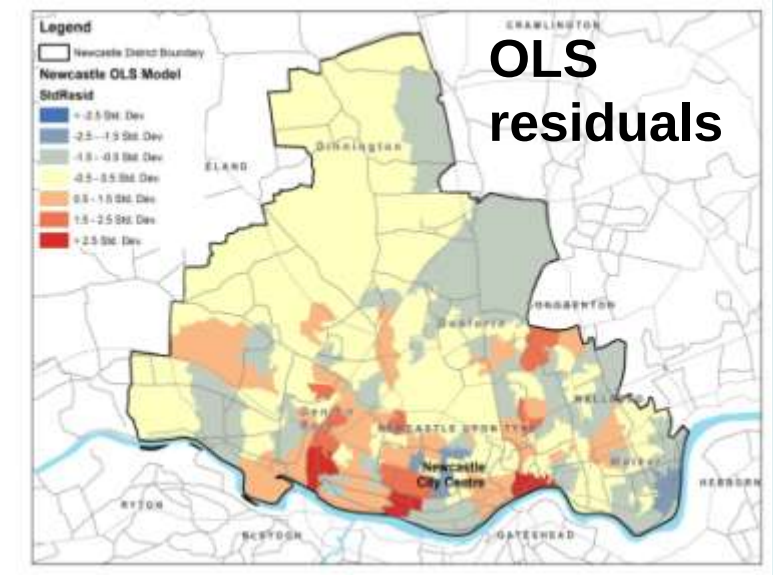
Understanding hotspots of residential burglary

Model significance and bias:

- Joint F statistic or Joint Wald statistic: 0.03, significant to 95%

  (if Koenker is statistically significant we use Joint Wald statistic)
- Koenker statistic: significant

  (heteroscedasticity and/or nonstationarity in model)
- Jarque-Bera statistic: significant
- Moran's I: clustered residuals



Interpretation

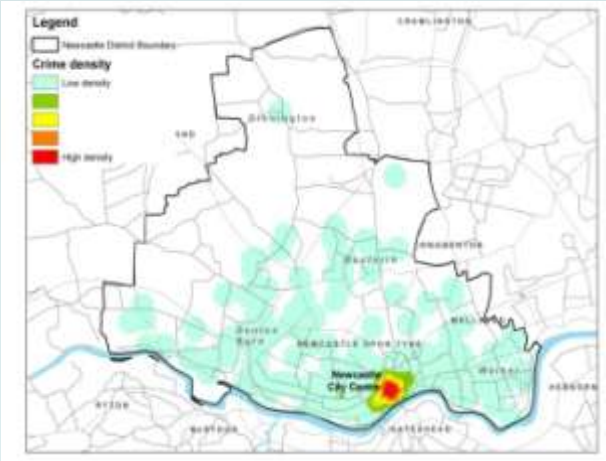
- Model is statistically significant, but it is bias
- One or more key variables are missing
- We can not proceed to using GWR ... ☹
- Need to identify additional variables; use map of residuals to help consider variables that are missing ...

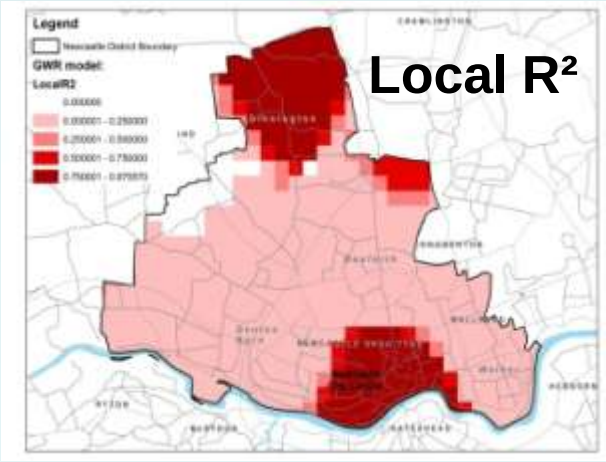
GWR: exploratory vs hypothesis testing

- Exploratory approach
  - Selecting variables and simply inputting them into a model in hope they will provide answers to explain a phenomenon is unrealistic!
- Hypothesis testing approach
  - enabling subjective reasoning for determining which variables to include in a model
  - If OLS model is biased and/or residuals are clustered, this requires the researcher to consider what variables are missing from the model, and test their inclusion
    - Map/scatterplot distribution of new variables to see if appear to be related to dependent variable

Conclusions (1)

- GWR provides means of exploring spatially varying relationships
  - Identify hotspots, use GWR to help understand why hotspots are present
  - Also illustrates that explanations for crime problems are often complex, rather than being based on a single explanation for every area
  - Does require sound theoretical reasoning to ensure model reliability
- Current research (parallel to RTM):
  - Determining causal factors
  - Spatially varying weightings
  - ‘Cause map’: risky places

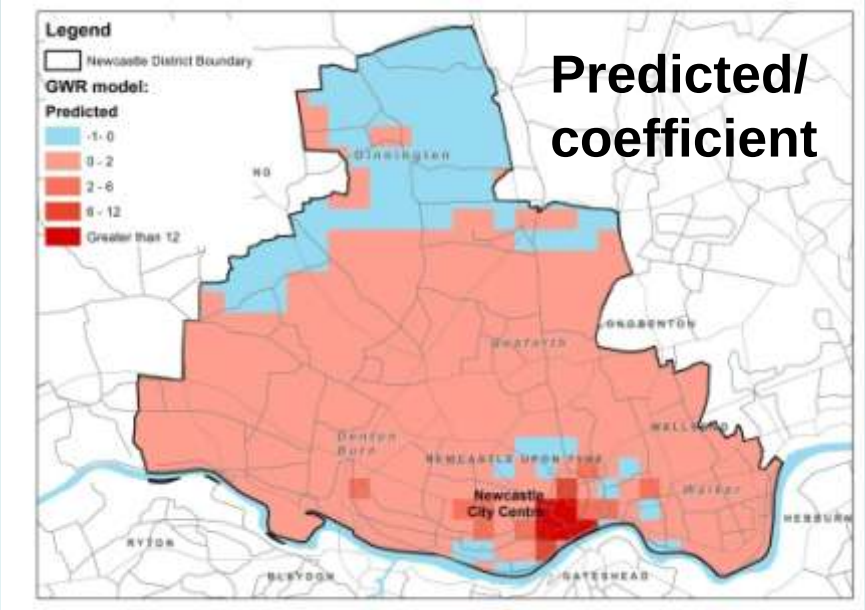




Conclusions (2)

- Crime prediction
  - Coefficients for each explanatory variable provide indication of how dependent variable would change for each unit change in the explanatory variable, for each area

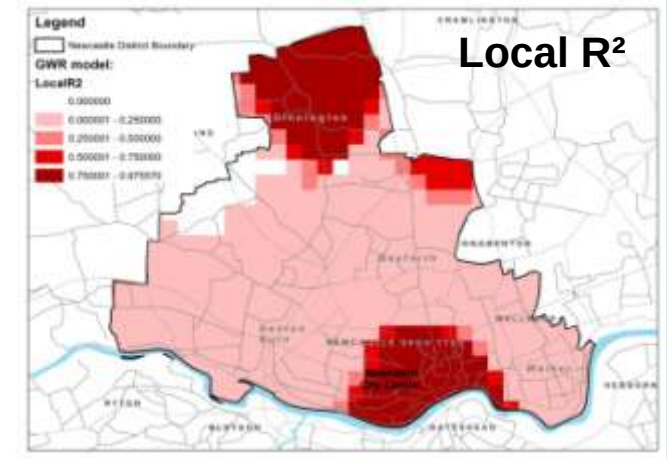
    E.g. two new licensed premises open in Newcastle. One is in area where coefficient between licensed premises and serious violence is higher – in this area greatest risk of increase in serious violence



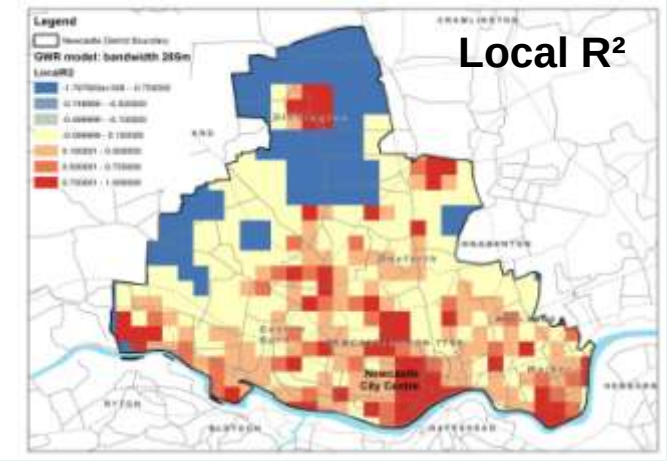
Conclusions (3)

- Bandwidth size ...
  - In the effort to minimise AICc, the bandwidth (fixed or adaptive) tends to be quite large (averaged approximately 1250m)
  - Spatially varying relationships may not be as 'local' as we'd like
  - Need for more research that explores issues when bandwidth is smaller than advised 'optimal' size
    - And influence of size of the geographic unit on bandwidth size and GWR results
- Future research: temporally sensitive GWR
  - Spatio-temporal varying relationships

Adaptive AICc (AICc = 1688)



User defined 285m (AICc = 7044)



Summary

- Does GWR replace good interpretative analysis?
  - No, complements it
  - Offers statistical robustness to claims for explaining why crime happens at certain locations
- Need to apply sound theory, logic and reasoning on why crime is likely to concentrate in order to sensibly direct GWR analysis
- When explanatory variables do not provide the whole answer
  - further hypothesise on other explanatory reasons
  - build these into our model and measure performance improvement
- Once established a reliable model, potential to use it to explore how change in one explanatory variable would impact upon future crime levels
  - providing a means of helping to identify crime prevention opportunities

Thankyou

Some useful resources

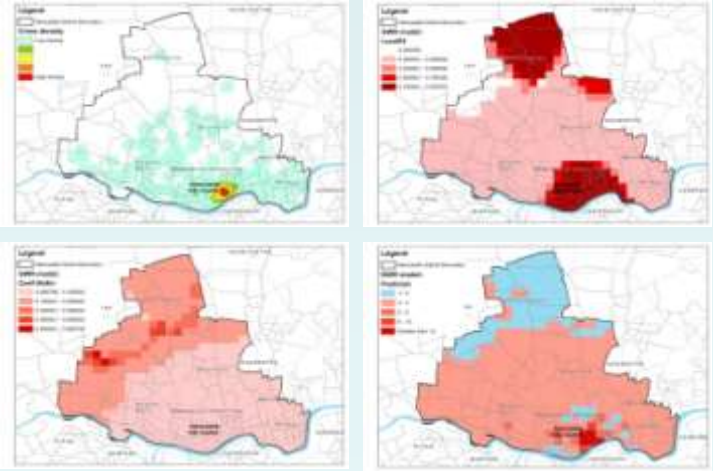
- COMING SOON: Chainey (2012). Understanding hotspots using GWR. In M. Leitner (ed), *Crime Mapping and Modelling*
- Fotheringham, Brunsdon and Charlton (2002). *Geographically Weighted Regression*
- ESRI ArcGIS Resource Centre and Help
  - Some excellent resources by Lauren Rosenshein and Lauren Scott
- Slides on website by Friday 28th October

Spencer Chainey

The Jill Dando Institute of Security and Crime Science

University College London

E: s.chainey@ucl.ac.uk

W: www.ucl.ac.uk/scs

Follow me on Twitter @SpencerChainey