Understanding the modifiable areal unit problem

Robin Flowerdew

School of Geography and Geosciences, University of St Andrews

March 2009

Acknowledgements

- Mick Green (Lancaster) and David Steel (Wollongong), statisticians
- ESRC / JISC for purchase of Census data (Crown copyright) for UK academic community
- David Martin (Southampton) for use of AZTool
- University of Canterbury for Erskine Visiting Fellowship
- Clive Sabel, Jamie Pearce and David Manley for help with data manipulation

- What is the Modifiable Areal Unit Problem (MAUP)?
- Investigating neighbourhood boundaries and health
- Why does it happen?
  - Local and regional effects
  - Spatial autocorrelation
  - Local processes
- Identifying processes – scales and areal zones
- Deriving data to reflect the processes
- Where does this leave us?

What is the MAUP?

The same basic data yield different results when aggregated in different ways

- First identified by Gehlke and Biehl (1934)
- Affects many types of analysis, including correlation and regression
- Applies where data are aggregated to areal units which could take many forms e.g. postcode sectors, local government units, store catchment areas, grid squares
- Work by (among others) Openshaw (1984), Fotheringham & Wong (1991), Tranmer & Steel (2001), on how and why the MAUP exists, and what can be done about it.

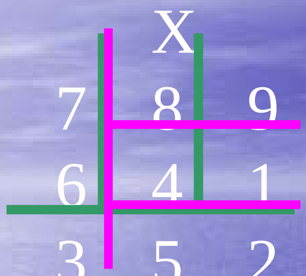
The MAUP – scale and zonation issues

- Two aspects – the **scale effect**, showing major analytical differences depending on the size of units used (generally correlations more pronounced for bigger units)
- - the **zonation effect** (Openshaw calls it the aggregation effect), showing major differences depending on how the study area is divided up, even at the same scale

Zonation effect

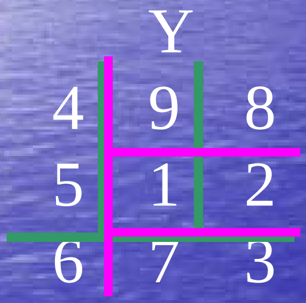
- Simple example shows how different zonal systems give very different results from same data without major variations in zone sizes or elongated shapes

Variables X and Y are defined for a 3x3 grid



X

7	8	9
6	4	1
3	5	2



Y

4	9	8
5	1	2
6	7	3

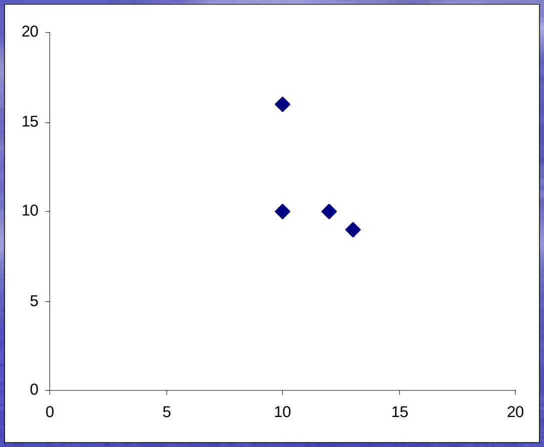
$r(X,Y) = .700$

$r_1(X,Y) = -.642$

$r_2(X,Y) = .913$

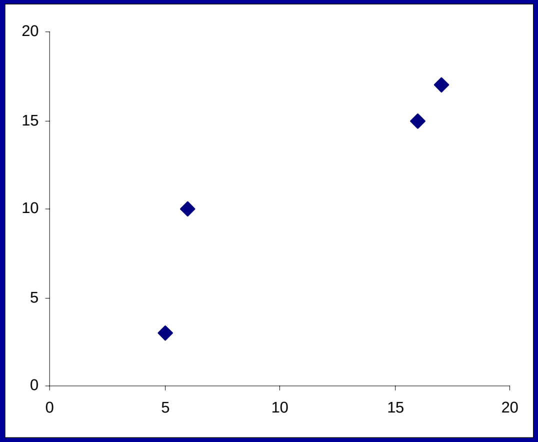
Zonation 1:

X	Y
13	9
12	10
10	10
10	16



Zonation 2:

X	Y
16	15
17	17
5	3
7	10



The MAUP in practice

- Taylor and Openshaw (1979) found that correlations in Iowa between Republican voting and percentage of old people could vary from

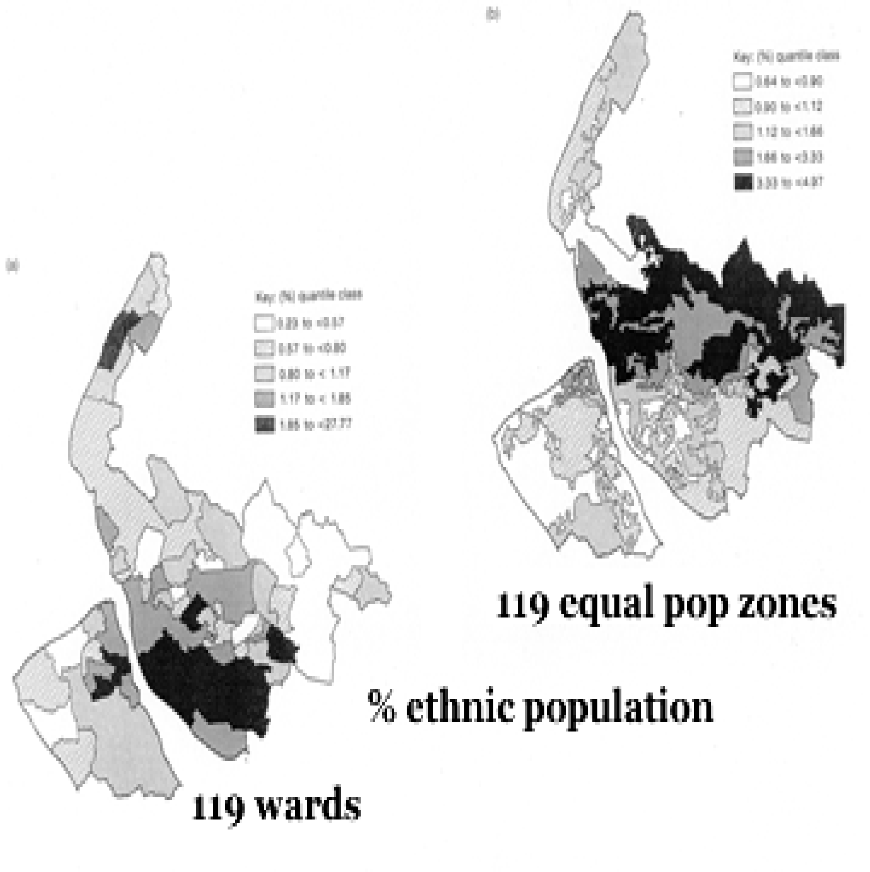
-.97 to +.99 depending on how counties were aggregated.

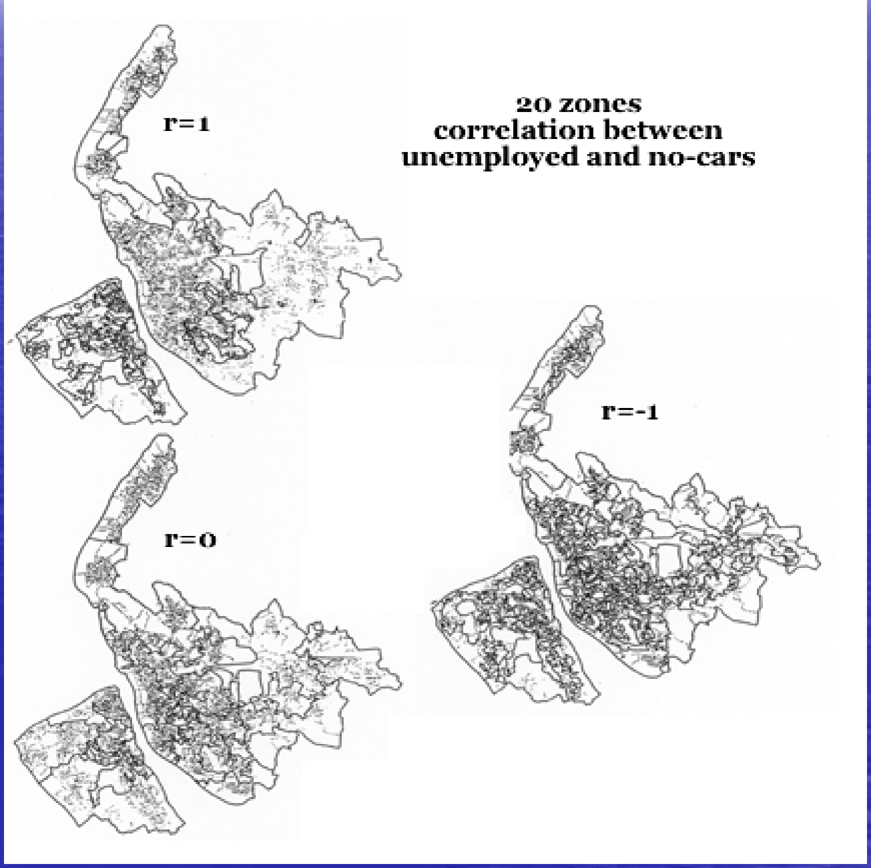
Openshaw and Rao (1995) achieved correlations between unemployment and ‘no car households’ in Merseyside from -1.00 to +1.00

But shapes are convoluted and sizes variable

Much less variation for more realistic zonal schemes

Shape and the MAUP





All shapes and sizes

But shapes are convoluted and sizes variable

Much less variation for more realistic zonal schemes

Manley (2006) took pairs of census variables and correlated them at ward and ED level – statistically significant differences in almost all cases

Example: the neighbourhood effect in health geography

- Often suggested health may be affected by contextual effects – health in the neighbourhood may affect individuals' health
- But how big is a neighbourhood? Does it matter where we draw the boundaries? Recent research (Flowerdew, Sabel and Manley 2008) looks at this and finds, for the case investigated, that the MAUP is not too worrying

Wards as neighbourhoods

- Ward-level and ED-level figures calculated for % limiting long-term illness *(pcllti)*
- ED-level *pcllti* then modelled as a function of other % variables
- Then ward-level *pcllti* added as a contextual effect
- Models weighted by population

Results – ward system

- *Ward-level*

  Pcllti = .956 + .363 *pcpens* + .032 *pcnonw* + .334 *pcmunem* $R^2 = .870$

- *ED-level (without neighbourhood effect)*

  Pcllti = 1.136 + .385 *pcpens* + .058 *pcnonw* +.262 *pcmunem* $R^2 = .733$

- *ED-level (with neighbourhood effect)*

  Pcllti = -.465 + .351 *pcpens* + .037 *pcnonw*

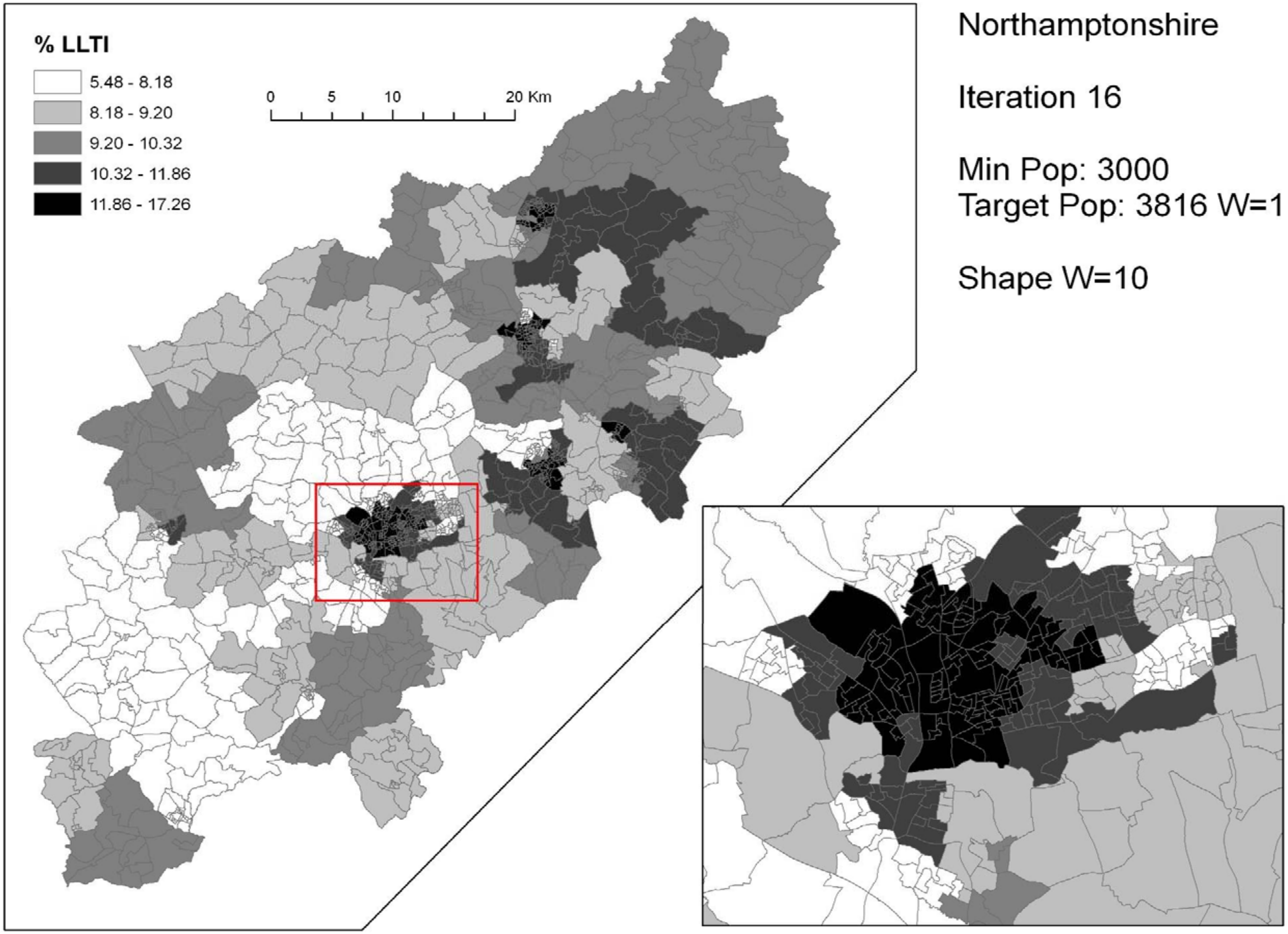
- + .228 *pcmunem* + .225 *pclltiwd* $R^2 = .747$

Designing new zonal systems

- So far, assumed ‘neighbourhood’ = ‘ward’
- But what if the boundaries were different?
- Zone design software will generate sets of pseudo-wards, based on several criteria:
- Pop. threshold
- Pop. target (e.g. average ward pop.)
- Shape (perimeter squared / area)
- Homogeneity

Northamptonshire example

- Study area needed
  - Has to be reasonably sized
  - No coastline
  - Range of settlement sizes
  - Knowledge of places

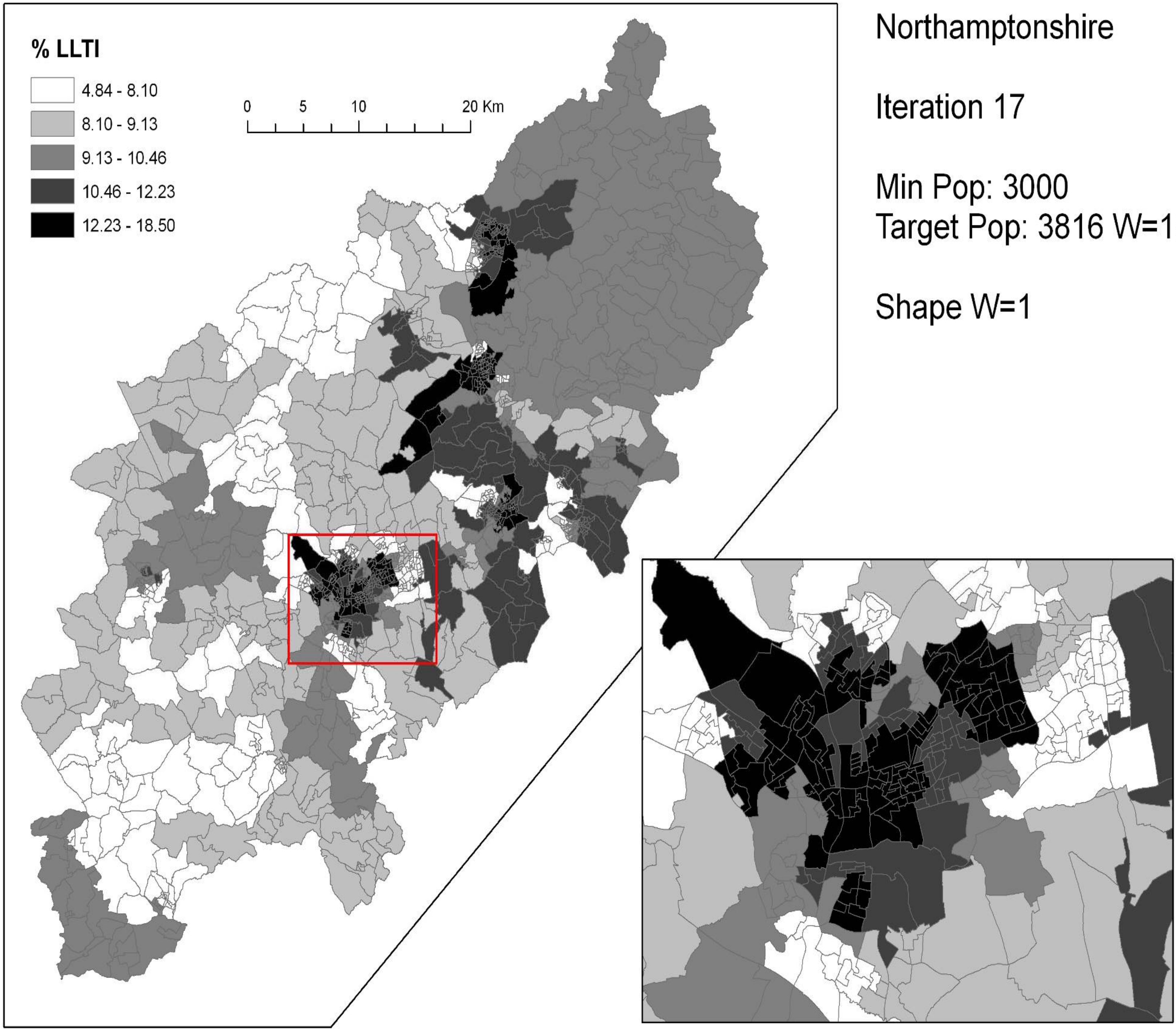


Northamptonshire

Iteration 16

Min Pop: 3000
Target Pop: 3816 W=1

Shape W=10



Northamptonshire

Iteration 17

Min Pop: 3000
Target Pop: 3816 W=10

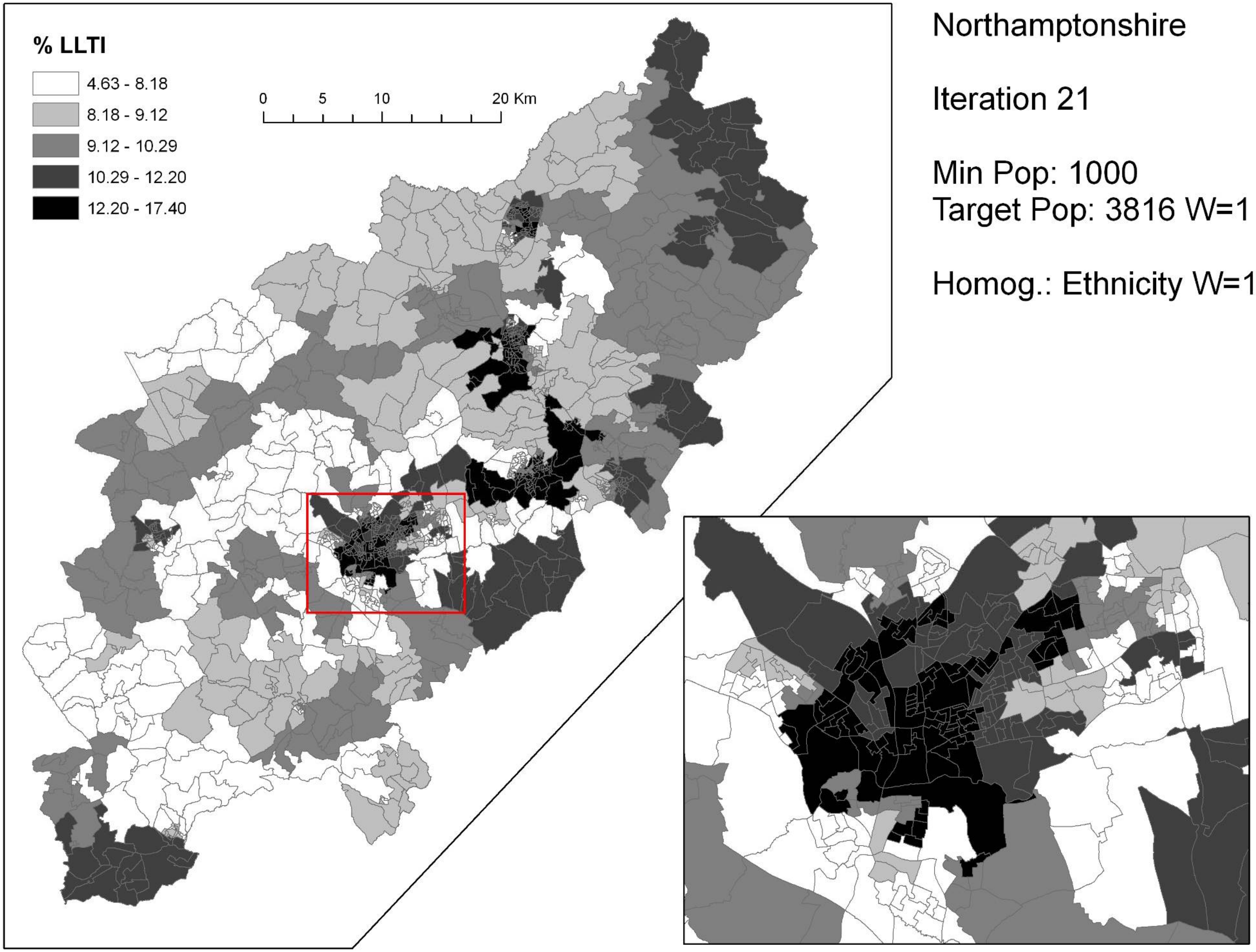
Shape W=1

Northamptonshire

Iteration 21

Min Pop: 1000
Target Pop: 3816 W=1

Homog.: Ethnicity W=1



The zonal systems

- Differences in overall patterns, based on:

a) the effect of constraints

b) random variation (process starts from randomly selected ‘seeds’)

c) often, whether zones of different types happen to be grouped together

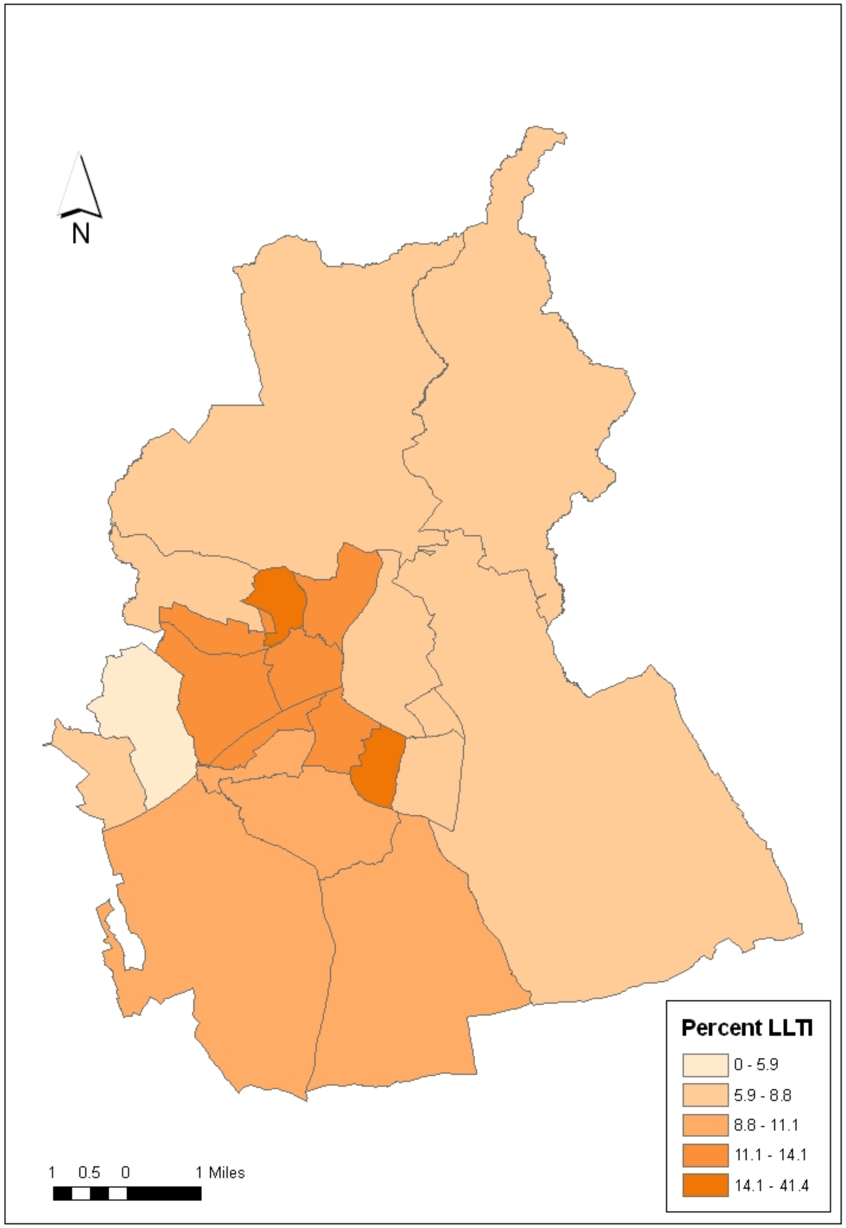
Regression results

Run	R^2	Pseudo-ward	Standard	Sig.
		coefficient	error	
3	.789	.168	.030	yes
4	.789	.168	.031	yes
5	.787	.136	.031	yes
6	.790	.179	.031	yes
8	.793	.201	.028	yes
11	.787	.120	.031	yes
14	.788	.140	.029	yes
15	.788	.145	.032	yes
16	.786	.117	.032	yes
17	.788	.157	.032	yes
21	.789	.162	.031	yes

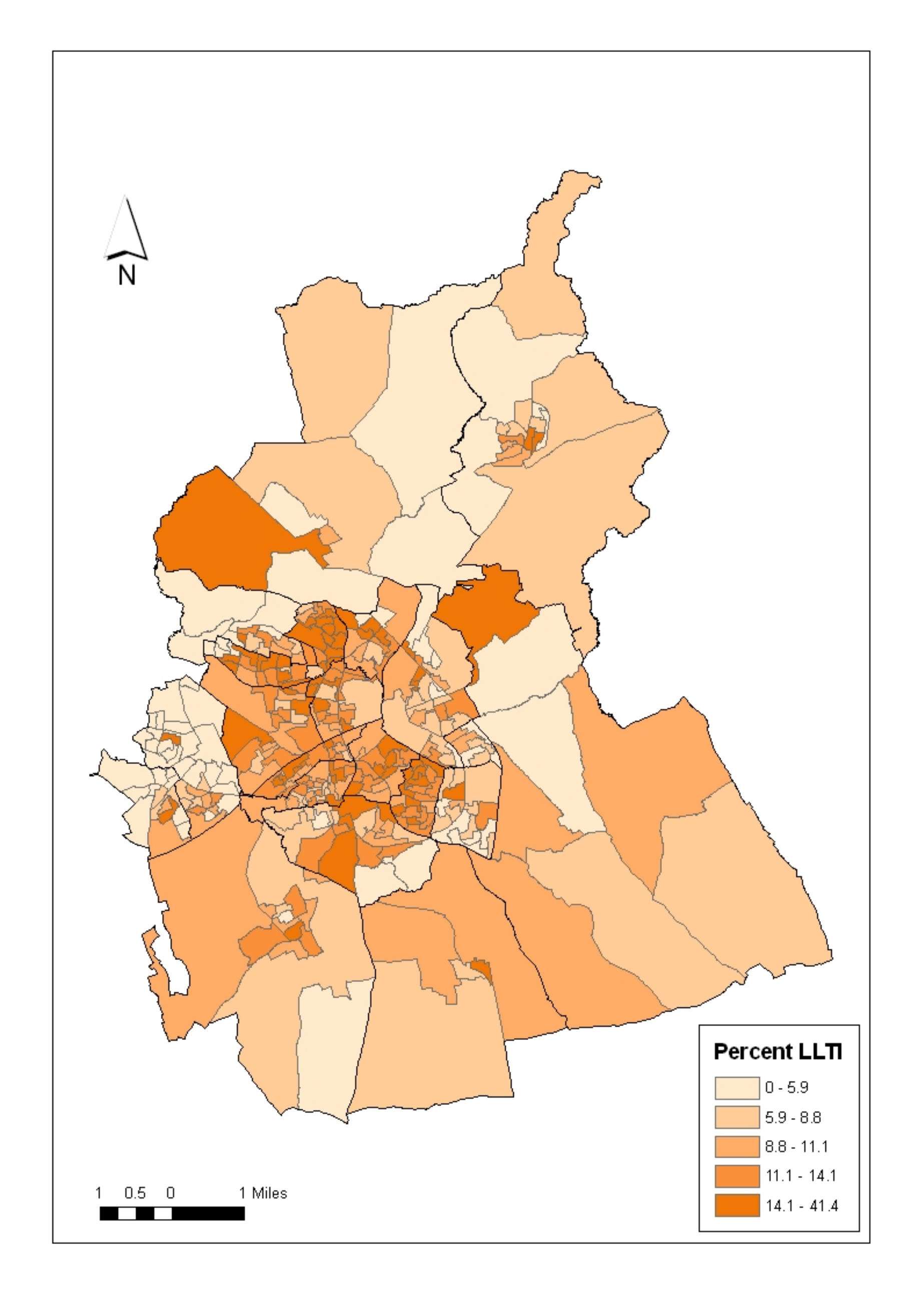
Further analysis - Thamesdown

- District of 170,000 around Swindon
- 21 wards, urban centre, rural periphery

- Thamesdown wards, 1991
- Limiting long-term illness



- Thamesdown EDs, 1991 Limiting long-term illness



Pseudo-ward systems

- Pseudo-ward systems can be generated with different constraints, of which population equality, shape and homogeneity are the main factors
- All systems had 19-22 zones of reasonably equal populations and sensible shapes

Variations in correlation structure

- The zonation effect suggests that there could be important changes in correlation coefficients for different zonal systems

X	Y	highest r	lowest r
Pcllti	pcwhite	.334	-.045
Pcllti	pckids	-.296	-.378
Pcllti	pcpens	.883	.828
Pcllti	pcunem	.863	.810
Pcllti	pcoo	.538	.329
Pcllti	pcla	.826	.701
Pcunem	pcla	.817	.686

Regression of illness on other variables

- At ED level,
- Pcllti = 9.391 - .075 pcwhite + .374 pcpens + .363 pcunem + .179 pcla
- (R^2 = .847)
- Adding neighbourhood variable, gives
- Pcllti = 9.225 - .078 pcwhite + .368 pcpens + .298 pcunem + .179 pcla + .067 nbdpw (s.e. .049)   (R^2 = .847)
- i.e. neighbourhood variable is not significant

The neighbourhood effect in other zonal systems

- To date, R ^{2} values have ranged from .849 to .995
- In most cases, neighbourhood effect is positive and small but statistically significant
- Biggest impact was to raise R^2 from .980 to .995

Conclusions about neighbourhood boundaries

- Possible to generate pseudo-ward zonal systems, with different constraints
- When relevant, analysts should try different zonal systems to test robustness of results
- In Northamptonshire, the neighbourhood effect varies from .117 to .201 – however it’s always +ve and significant – it does not matter ***in this case*** how you draw the boundaries
- In Thamesdown, it varies from .084 to .731 – not always significant – perhaps it does matter here!

Why does it happen?

- Size of correlation coefficient *r* is greatly influenced by highest and lowest values:
- if high X associated with high Y (and low X with low Y), you get high positive *r*
- If high X associated with low Y (and low X with high Y) – high negative *r*

- If aggregation groups zones with high X values (and / or high Y values) this effect is intensified
- If it groups high X zones with low X zones (or with average X zones), the effect is diluted

MAUP and spatial autocorrelation

- If the high values of X or Y are close together (positive spatial autocorrelation), grouping is likely to intensify the correlation of X and Y
- If high values are scattered around the study area (low spatial autocorrelation), grouping will diminish their effect
- MAUP effect results from interplay of these
- Effect of grouping may be largely a chance factor – or may come from a desire for homogeneity from people drawing the boundaries

Local and regional effects

- Processes with spatial aspects may generate the MAUP
- E.g. the labour market – unemployment in zone i depends on job vacancies not just in i but in surrounding zones too
- The housing market likewise

i.e. local and regional effects

Problem: identifying the region! Work assumed local = ED and regional = ward, but why should labour market respect ward boundaries?

Local processes

- Often our understanding of the distribution of X or Y reflects local knowledge – we know a housing estate is located here, a good school is located there, an ethnic enclave is over there
- Such information can in part explain mapped patterns and correlations, and can suggest the configuration of zones that best reflects geographical reality
- i.e. descriptive empirical studies can inform modelling

Identifying processes – scales and areal units

- Need to recognise that geographical processes occur at particular scales, and studying them at the wrong scale may not be very helpful
- Further, data are not always available at the right scale – usually impossible to disaggregate data below its scale of supply
- Even if data are available for very small areas, it may not be clear how to aggregate them up

Finding data to reflect the processes

- MAUP has a zonation effect too – even if data are at the right scale, they may not be configured in a way that reflects the processes going on
- Should we design a set of zones specifically to fit an empirical problem? (perhaps by finding a set of zones to maximise correlation)
- Would this be ‘cheating’?

Optimal scales may vary spatially

- Note also that appropriate scales may be different for different places – social segregation for example may affect whole wards in big cities but a few EDs in towns
- Appropriate scales may be different even within the same study area

Where does this leave us? - 1

- Scales of process are not always the same as scales for which we have data – likewise for configurations
- We need to be critical of the data for these reasons as well as many others
- We need to think about the processes being studied and the scales of data needed

Where does this leave us? - 2

- Most statistical work using modifiable areal unit data is deficient – because zonal system doesn’t fit processes - and probably underestimates the strength of relationships
- Statistical results are not independent of the zonal systems the data come from
- Probably worth analysing data at several different scales, noting the differences and using them to help identify processes
- Good zone design software becoming available