

Explaining [modelling] spatial data

Mark Stevenson

Faculty of Veterinary and Agricultural Sciences

The University of Melbourne

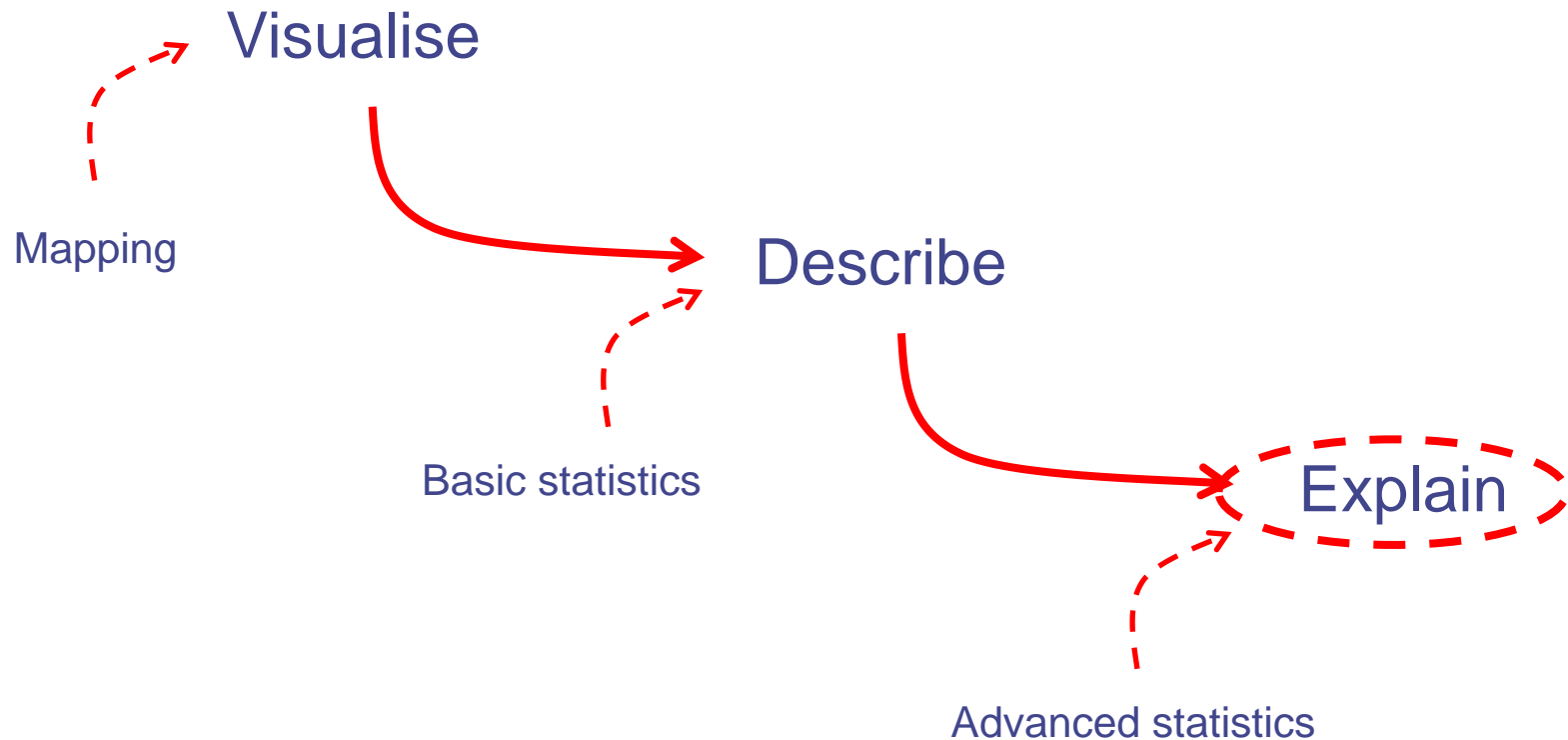
Parkville, Victoria 3010, Australia

mark.stevenson1@unimelb.edu.au



THE UNIVERSITY OF
MELBOURNE

Approach to spatial analysis



Roadmap

- Background
- Fixed-effects models
 - frequentist approaches
 - diagnostics for spatial autocorrelation
- Mixed-effects models
 - area data
 - point data
 - continuous data



Background

- What's so special about spatial 'modelling'?
 - spatial modelling allows you to estimate the strength of association between a set of explanatory variables and disease presence
 - mapping residuals from a model allows you to identify locations where disease is present that is not accounted-for by the explanatory variables you included in your model





The Australian

WAGYU update

Edition 20 - February 2002

BSE hits Japan. Hard

THE outbreak of BSE in Japan has hit beef consumption hard with 30% of consumers surveyed in Tokyo indicating they have not eaten beef since the outbreak was confirmed in Japan in September.

At retail, beef is selling at a fraction of the cost 12 months ago.

Total beef consumption for October has fallen 48% to 45,000 tons. That said, consumption of domestic product has been affected much more than imported beef, with the consumption of domestic beef falling by 77% compared to a year ago.

Australia, given its clean green image, is less affected than all other countries exporting beef to Japan. The consumer impact of BSE has

intensified as restaurants move to take beef from their menus.

Total consumption of pork has increased 12% in October as a result of substitution.

Profits for Skylark, the largest family restaurant chain in Japan, were down 7 billion yen compared to their August forecasts of 12 billion.

Compounding the affects of the BSE outbreak is the Japanese economy, which is officially in recession, shrinking .5% in the July to September quarter.

Personal consumption, a major factor in domestic demand fell 1.7%. A deteriorating business environment in Japan has led to a cut in salaries with nominal pay scales falling

1.7%. Retail prices for November published by the bank of Japan reveal a 1.4% drop in prices compared to November 2000, the largest drop in two years.

Australia's MLA has been quick to mount a major media campaign re-enforcing with Japanese consumers the "clean green" image of Australian beef. While the short term prognosis for exports to Japan is not good, the longer-term perspective is for an enhanced opportunity for Australian producers.

The wild card in the overall scenario is whether or not there are further discoveries of BSE in Japan. Each new discovery delays recovery in consumer confidence.

Bright Wagyu future, page 3



McDonalds In Japan Gives Out Free Burgers To Fight Fears Of Mad Cow



Kadohira, M., Stevenson, M., Kanayama, T., Morris, R., 2008. The epidemiology of bovine spongiform encephalopathy in Hokkaido, Japan, September 2001 to December 2006. Veterinary Record 163, 709 - 713.

PAPERS & ARTICLES

Epidemiology of bovine spongiform encephalopathy in cattle in Hokkaido, Japan, between September 2001 and December 2006

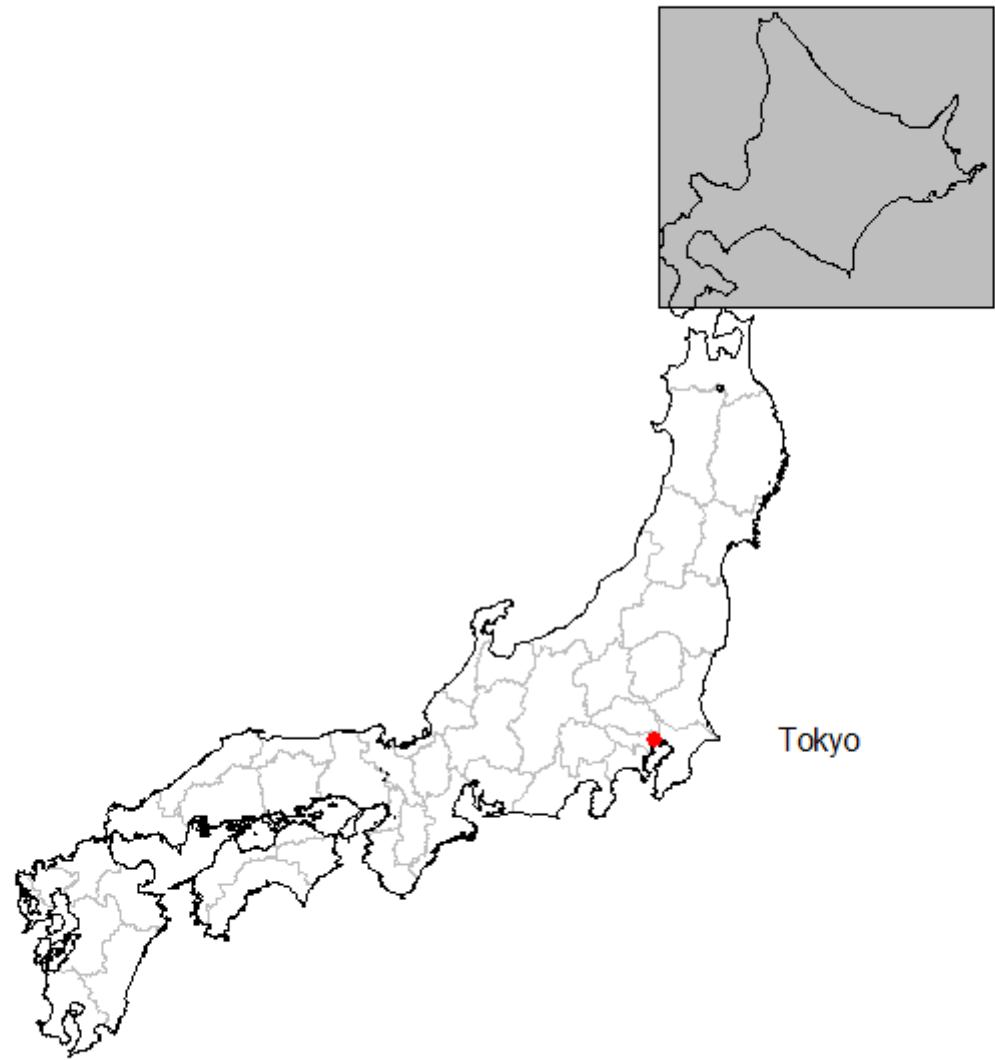
M. KADOHIRA, M. A. STEVENSON, T. KANAYAMA, R. S. MORRIS

Between October 2001 and December 2006 an estimated total of 6 million cattle in Japan were tested for BSE, with 31 returning a positive result. Exploratory mapping, the space-time scan statistic, and ordinal logistic regression have been used to describe the epidemiology of the 24 cases identified in the prefecture of Hokkaido, and to quantify the risk factors for the disease. Two birth cohort groups were affected: cattle born during a period of seven months in 1996, and cattle born between 1999 and 2001. The descriptive spatial analyses showed that eight of the 10 cases born in 1996 were born in areas with a relatively high density of dairy farms in the east of Hokkaido, but that the 14 later cases were more widely distributed throughout the prefecture, with equal numbers of cases in the east and the west. These findings provide indirect evidence of a single localised contamination of the cattle feed supply in 1996, and recycling of the infection after 1999.

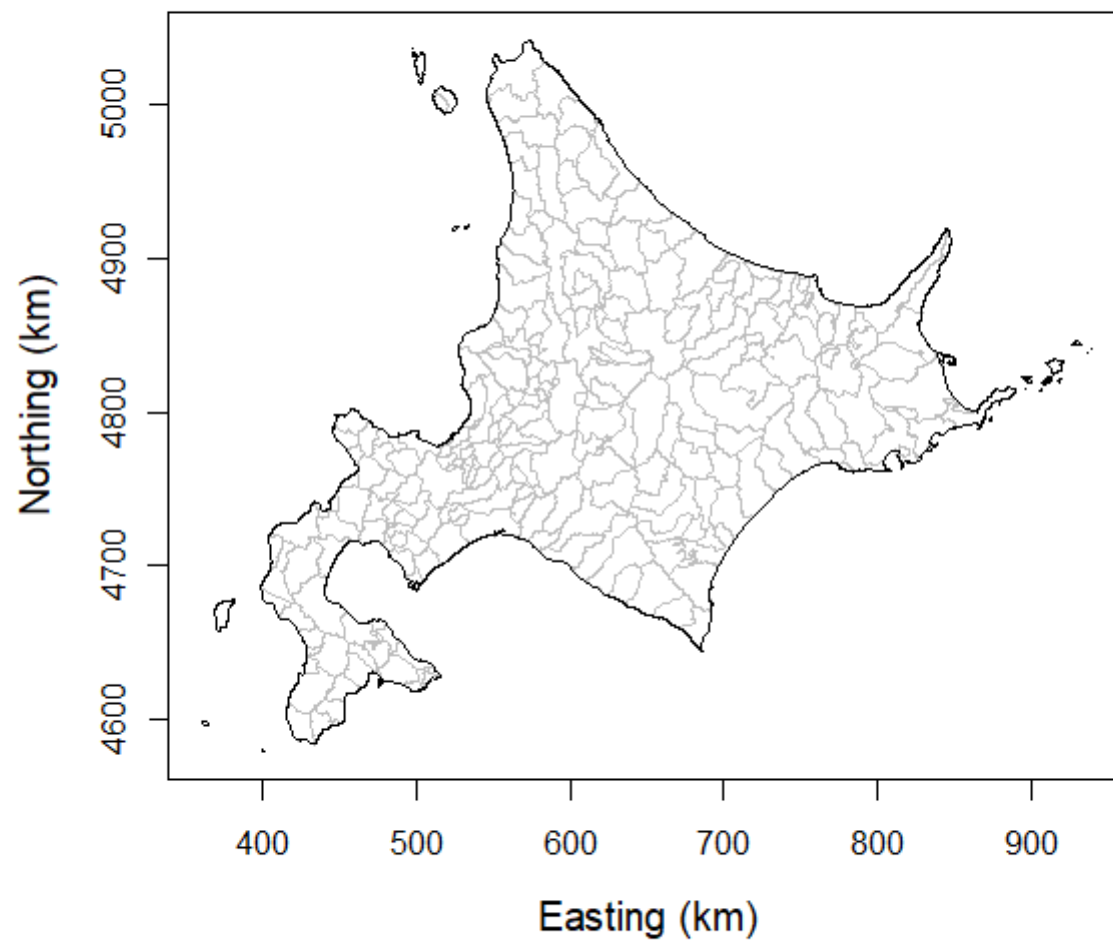
Map showing the boundaries of the 47 prefectures of Japan.



Map showing the boundaries of the 47 prefectures of Japan.



Map of Hokkaido prefecture, showing the boundaries of the 219 mainland cities.



Obihiro, Hokkaido, Japan.



University of Obihiro, Hokkaido, Japan.





Obihiro, Hokkaido, Japan.

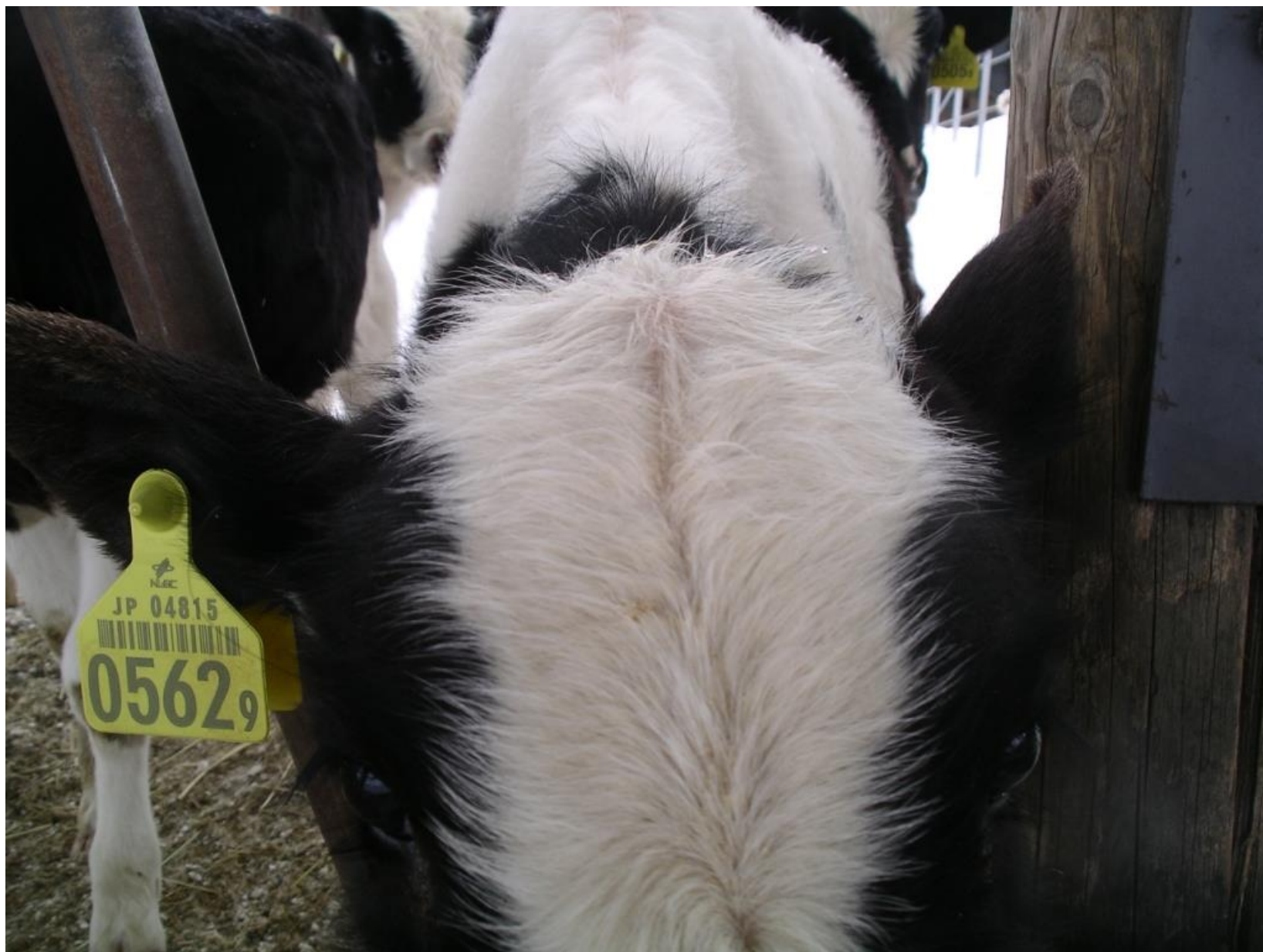


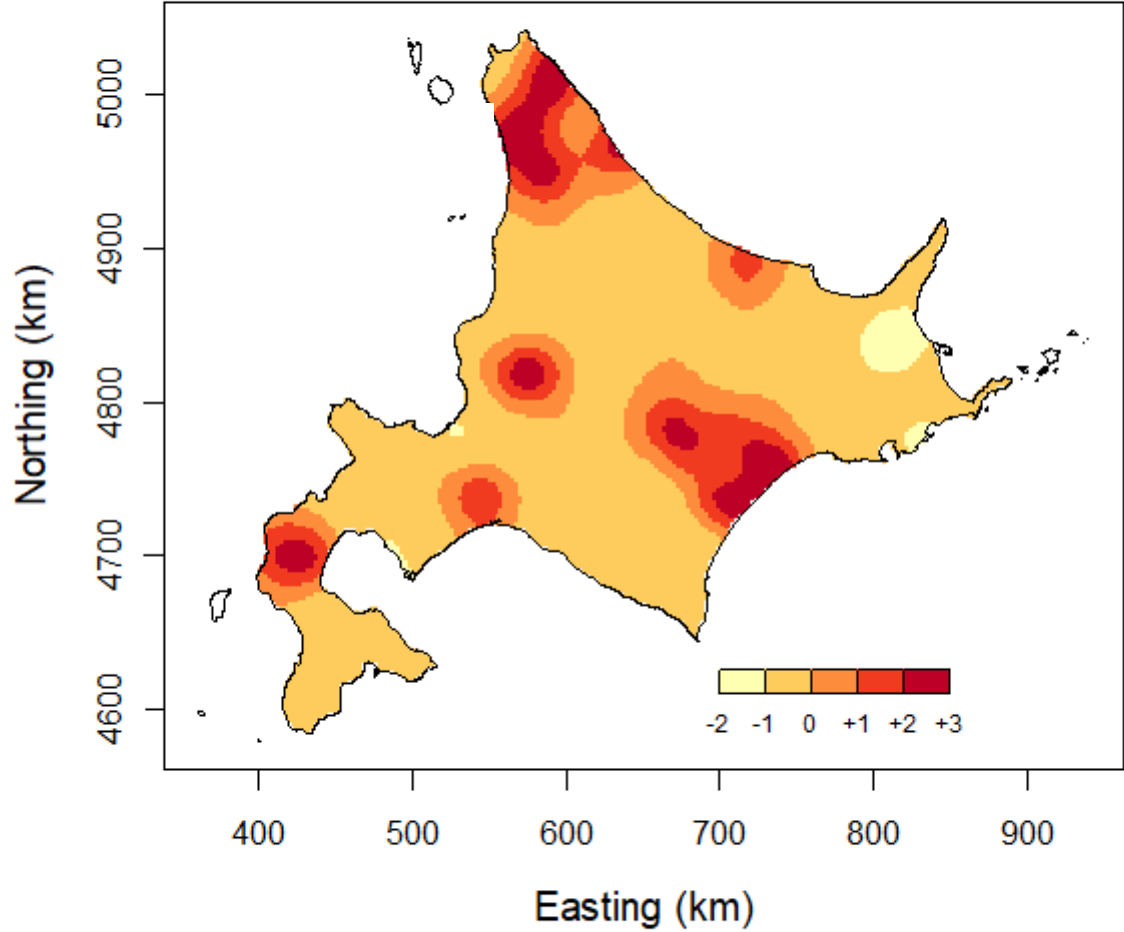
Table 1: BSE in the Japanese prefecture of Hokkaido, September 2001 to December 2006. Regression coefficients and standard errors from an ordinal logistic regression model of city level BSE risk.

Variable	Coefficient	SE	Wald z	P	OR (95% CI)
Intercepts:					
0 1	-3.4149	0.4041	-8.35	< 0.01	
1 2	-5.1255	0.6690	-7.66	< 0.01	
No. dairy farms ^a	0.4204	0.0942	4.46	< 0.01	1.52 (1.26 – 1.83) ^b

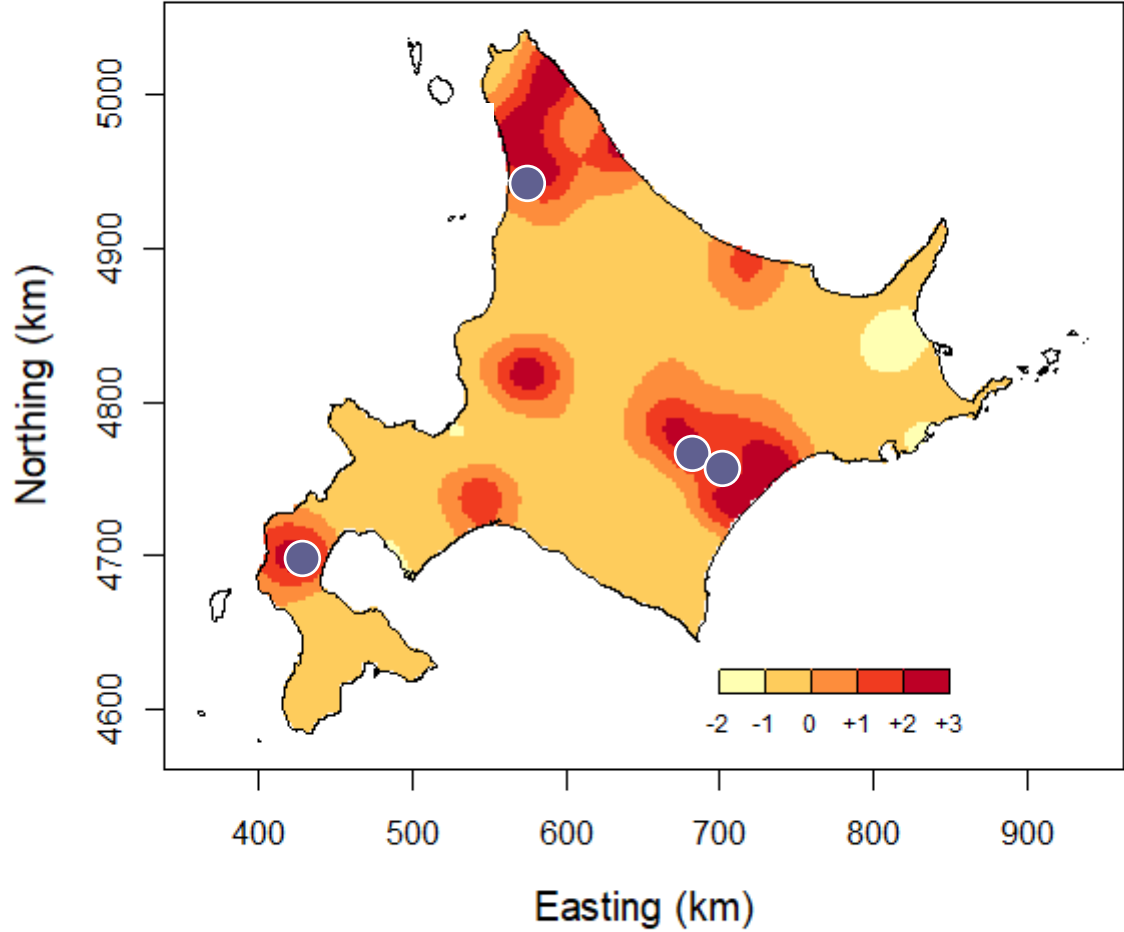
^a Number of dairy farms per city (25 farm increments).

^b Interpretation: increasing the number of dairy farms per city by 25 increased the odds of a city being BSE positive by a factor of 1.52 (95% CI 1.26 – 1.83).

BSE in the Japanese prefecture of Hokkaido, September 2001 to December 2006. Image plot showing the distribution of positive and negative sign residuals from an ordinal logistic regression model of city level BSE risk.



BSE in the Japanese prefecture of Hokkaido, September 2001 to December 2006. Image plot showing the distribution of positive and negative sign residuals from an ordinal logistic regression model of city level BSE risk. The city of birth of the four BSE cases identified up until 30 January 2009 shown as points.



Stevenson, M., Morris, R., Lawson, A., Wilesmith, J., Ryan, J., Jackson, R., 2005. Area-level risks for BSE in British cattle before and after the July 1988 meat and bone meal feed ban. *Preventive Veterinary Medicine* 69, 129 - 144.



Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Preventive Veterinary Medicine 69 (2005) 129–144

www.elsevier.com/locate/prevetmed

**PREVENTIVE
VETERINARY
MEDICINE**

Area-level risks for BSE in British cattle before and after the July 1988 meat and bone meal feed ban

M.A. Stevenson^{a,*}, R.S. Morris^a, A.B. Lawson^{b,1},
J.W. Wilesmith^c, J.B.M. Ryan^c, R. Jackson^a

^a*EpiCentre, Institute of Veterinary, Animal, and Biomedical Sciences, Massey University,
Private Bag 11-222, Palmerston North, New Zealand*

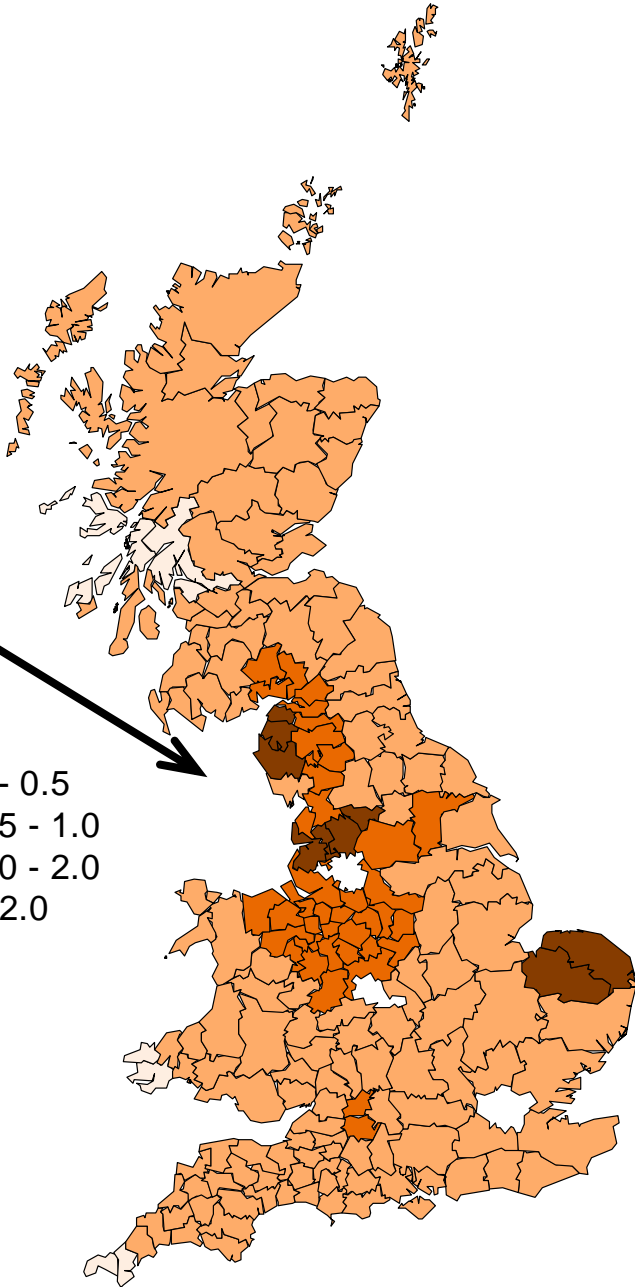
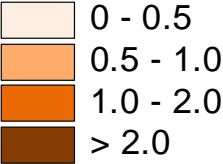
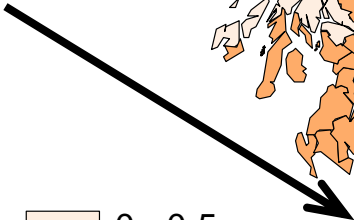
^b*Department of Mathematical Sciences, University of Aberdeen, Aberdeen AB24 3UE, UK*

^c*Epidemiology Department, Veterinary Laboratories Agency, New Haw,
Addlestone, Surrey KT15 3NB, UK*

Received 27 July 2004; received in revised form 24 January 2005; accepted 27 January 2005

Relative risk attributable to structured heterogeneity terms: post-control cohort.

Low-moderate BSE risk
not explained by
dairy:non-dairy ratio, pig:
cattle ratio, or northing



Kathmandu, Nepal January 2013.



Roadmap

- Background
- Fixed-effects models
 - frequentist approaches
 - diagnostics for spatial autocorrelation
- Mixed-effects models
 - area data
 - point data
 - continuous data



Fixed-effects models

- ‘Standard’ regression techniques we’ve learnt about
 - linear regression
 - Poisson regression
 - logistic regression
 - can be applied to spatial data (in the same way that they’re used for non-spatial applications)
 - this lecture will focus on Poisson regression

Fixed-effects models

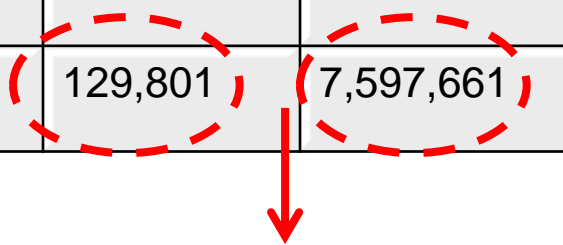
- Indirect and direct adjustment
 - allows us to compare patterns of disease in populations which have different population structures

Fixed-effects models

- Indirect adjustment
 - the expected number of disease cases is determined on the assumption that the risk [or rate] of disease in each area is the same as the national level risk of disease
 - the observed number of disease cases is compared with the expected number of disease cases to yield a standardised morbidity/mortality ratio (SMR)
 - usual to plot SMR of disease as a choropleth map

Fixed-effects models

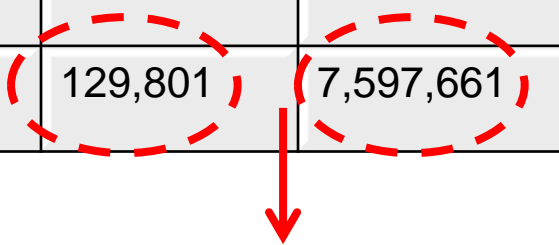
County	Observed	Population	Expected	SMR
A	691	82,220		
B	1157	34,168		
C	713	60,571		
...		
Total	129,801	7,597,661		


$$129,801 \div 7,597,661 = 0.017$$



Fixed-effects models

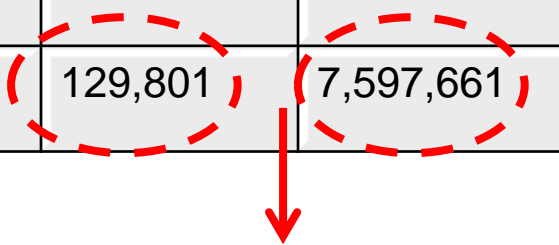
County	Observed	Population	Expected	SMR
A	691	82,220	$0.017 \times 82,220 = 1398$	
B	1157	34,168	$0.017 \times 34,168 = 581$	
C	713	60,571	$0.017 \times 60,571 = 1030$	
...	
Total	129,801	7,597,661		


$$129,801 \div 7,597,661 = 0.017$$



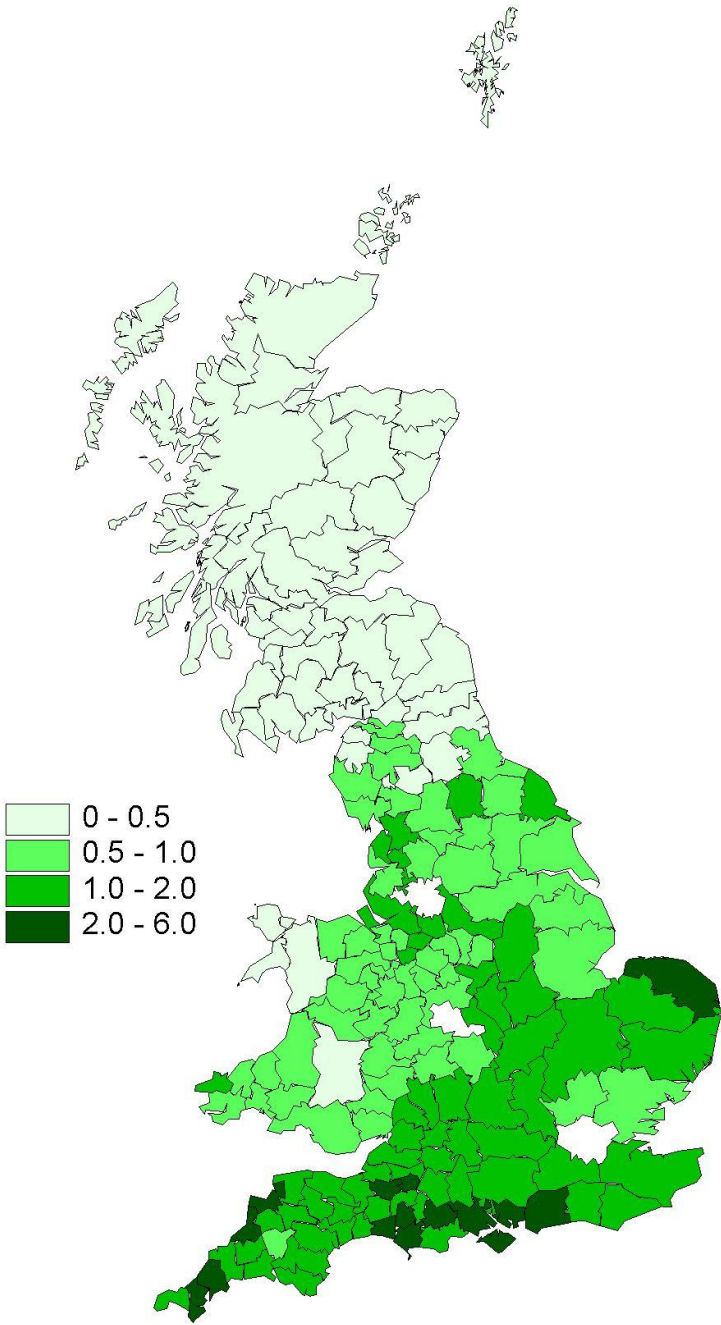
Fixed-effects models

County	Observed	Population	Expected	SMR
A	691	82,220	$0.017 \times 82,220 = 1398$	$691 \div 1398 = 0.50$
B	1157	34,168	$0.017 \times 34,168 = 581$	$1157 \div 581 = 1.99$
C	713	60,571	$0.017 \times 60,571 = 1030$	$713 \div 1030 = 0.69$
...
Total	129,801	7,597,661		

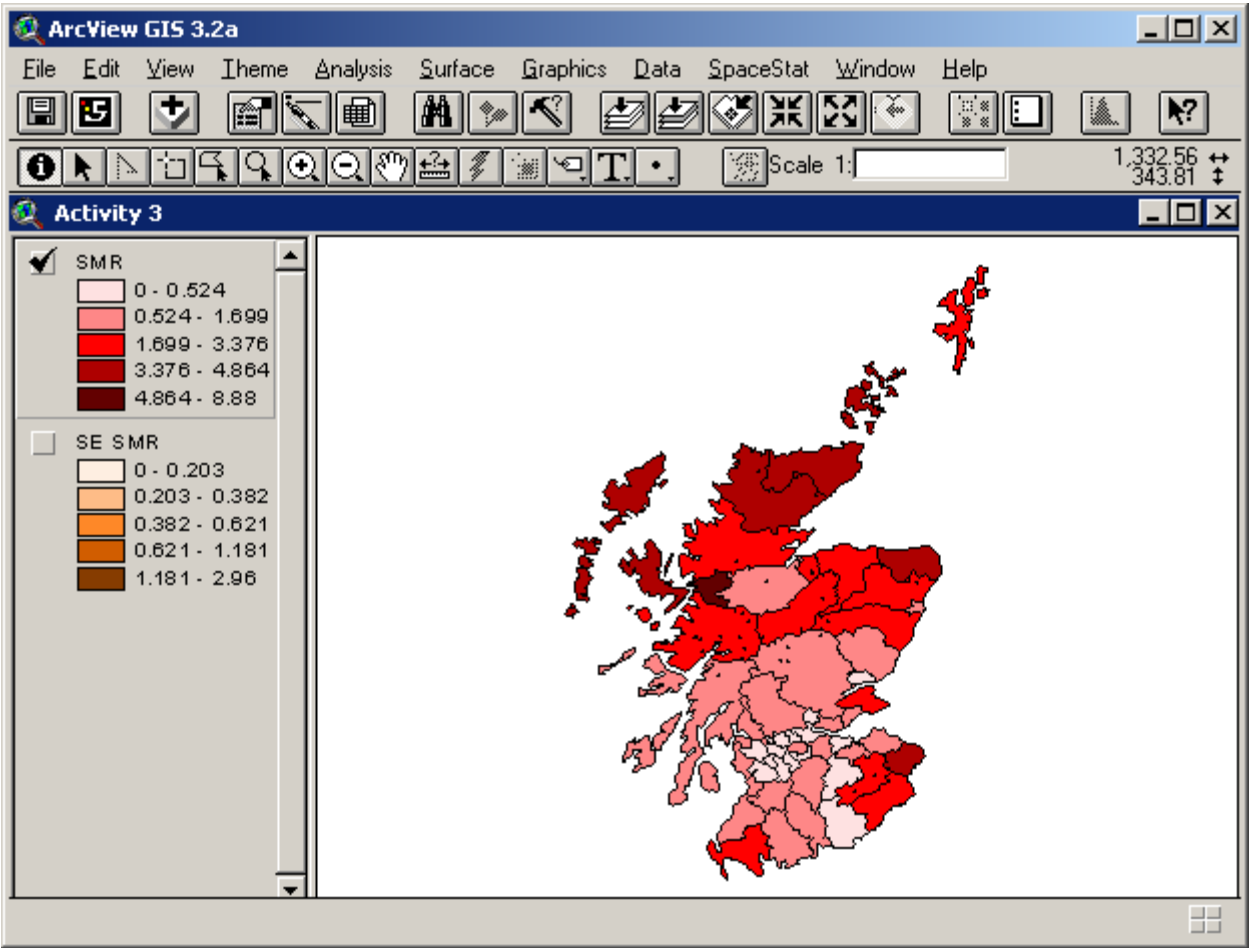

$$129,801 \div 7,597,661 = 0.017$$



SMR for BSE in British cattle born before 30 June 1988.



Choropleth maps of district-level standardised mortality ratios (SMRs) for lip cancer in Scottish districts, 1980-1985.



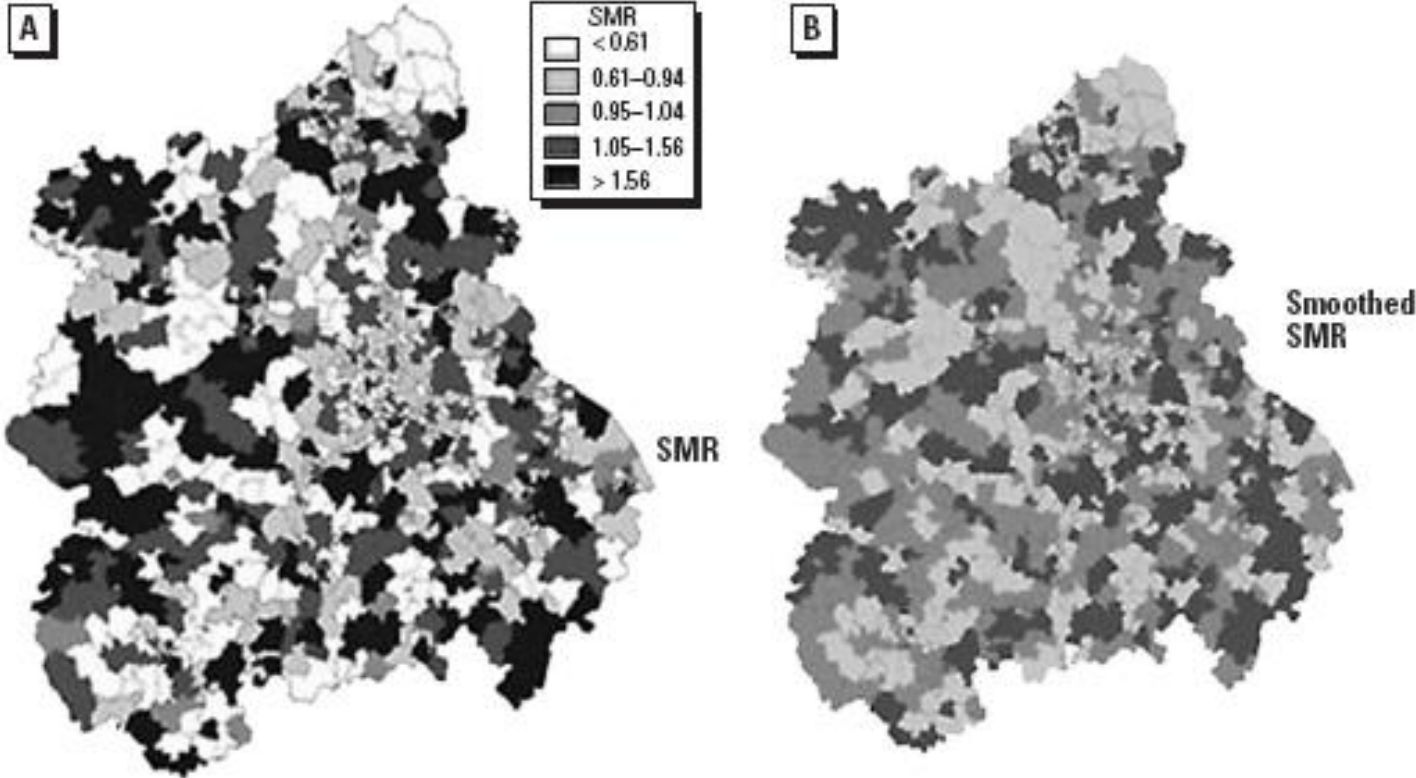


Figure 2. Adult leukemia by electoral ward in West Midlands Region, England, 1974–1986. (A). SMR; West Midlands = 1.0. (B) SMR after smoothing using empirical Bayes methods. Figure reproduced from Olsen et al. (1996), with permission of the BMJ Publishing Group.

Ferrándiz et al. (2004) Spatial analysis of the relationship between mortality from cardiovascular and cerebrovascular disease and drinking water hardness. Environmental Health Perspectives 112(9) 1037-1044.

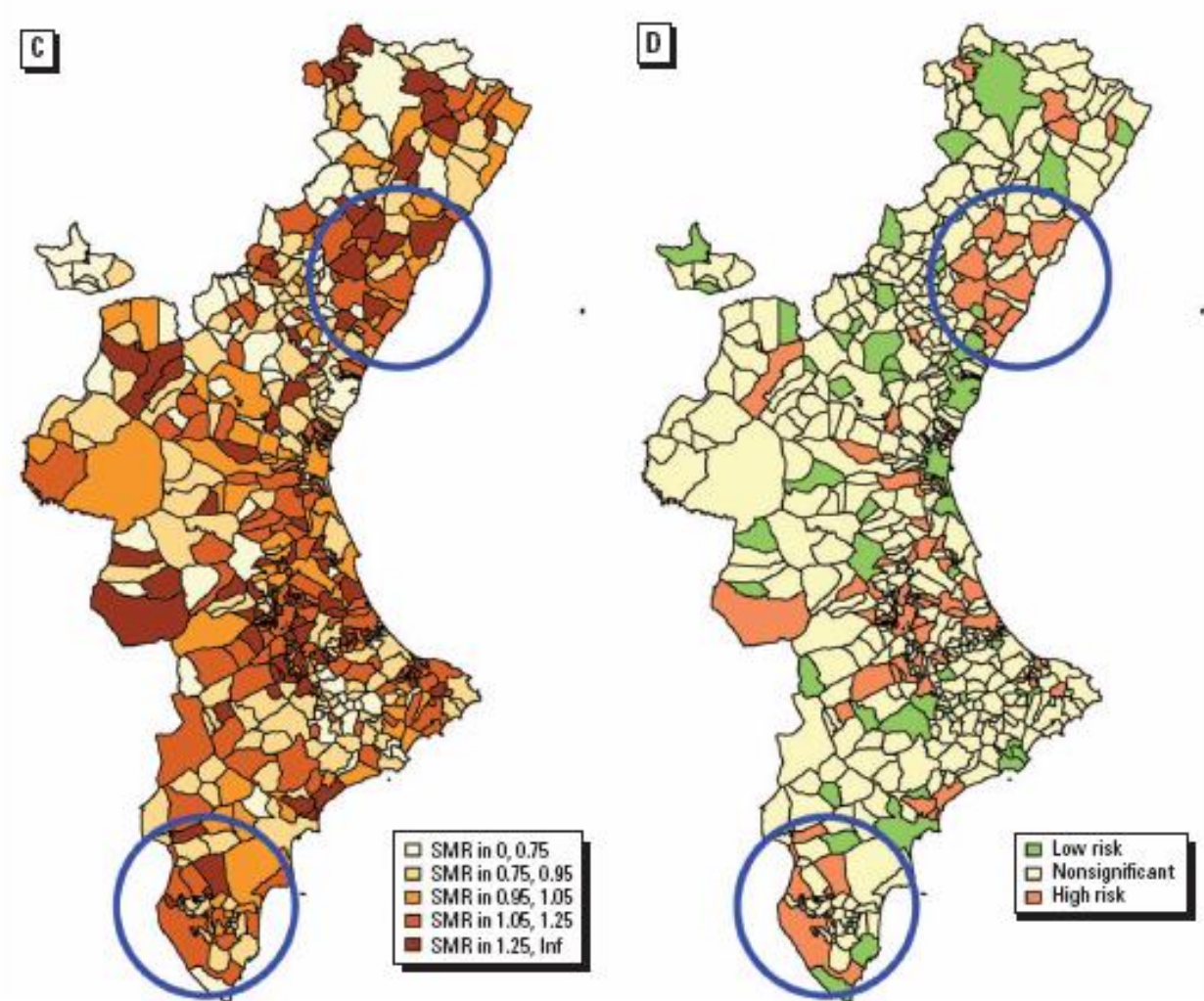


Figure 5. Disease mapping of total cerebrovascular mortality for the whole period: smoothed SMRs (A,C) and significance of 95% confidence intervals (B,D) after standardization by age, sex, and deprivation index (A,B) and further standardization by Mg (C,D). Inf, infinity. Municipalities illustrating the change of risk level when adjusting for the covariate are circled in blue.

Fixed-effects models

- So:

$$\frac{O_i}{E_i} = \text{SMR}_i$$

$$O_i = E_i \times \text{SMR}_i$$

- we say that the observed count of disease events in each area equals the expected count multiplied by an area-specific ‘modifier’ (the SMR)



Fixed-effects models

- Formally:

$$O_i = E_i \times SMR_i$$

$$\mu_i = E_i \times RR_i$$

$$\log \mu_i = \log E_i + \log RR_i$$

$$\log \mu_i = \log E_i + (\alpha + \beta_1 x_{1i} + \dots + \beta_m x_{mi}) + \varepsilon$$

the modifier is actually a risk ratio

the term $\log(E_i)$ is called an offset and represents an adjustment to deal with areas of different population size



Fixed-effects models

- Formally:

$$O_i = E_i \times SMR_i$$

$$\mu_i = E_i \times RR_i$$

$$\log \mu_i = \log E_i + \log RR_i$$

$$\log \mu_i = \log E_i + (\alpha + \beta_1 x_{1i} + \dots + \beta_m x_{mi}) + \varepsilon$$

the regression coefficients $\beta_1, \beta_2, \dots, \beta_m$ represent 'how strong' each explanatory variable is at influencing the RR

the key assumption is that this influence (i.e. effect) is constant or 'fixed' across all areas



Fixed-effects models

- Formally:

$$O_i = E_i \times SMR_i$$

$$\mu_i = E_i \times RR_i$$

$$\log \mu_i = \log E_i + \log RR_i$$

$$\log \mu_i = \log E_i + (\alpha + \beta_1 x_{1i} + \dots + \beta_m x_{mi}) + \varepsilon$$

the exponent of ε_i represents the residual relative risk in area i after adjusting for each of the covariates included in the model



Fixed-effects models

- Formally:

$$O_i = E_i \times SMR_i$$

$$\mu_i = E_i \times RR_i$$

$$\log \mu_i = \log E_i + \log RR_i$$

$$\log \mu_i = \log E_i + (\alpha + \beta_1 x_{1i} + \dots + \beta_m x_{mi}) + \varepsilon$$

```
sc.glm01 <- glm(obs ~ offset(log(exp)) + propag, family  
= poisson, data = sc.dat)
```



Fixed-effects models

Variable	Coefficient	SE	z	P	RR (95% CI)
Intercept	-0.5865	0.0701	-8.37	< 0.01	
Prop ag	0.0811	0.0060	13.46	< 0.01	1.08 (1.07 – 1.10) ^a

^a Interpretation: For unit increases in the percentage of the workforce involved in outdoor industry, the relative risk of lip cancer was increased by a factor of 1.08 (95% confidence interval 1.07 to 1.10).

Fixed-effects models

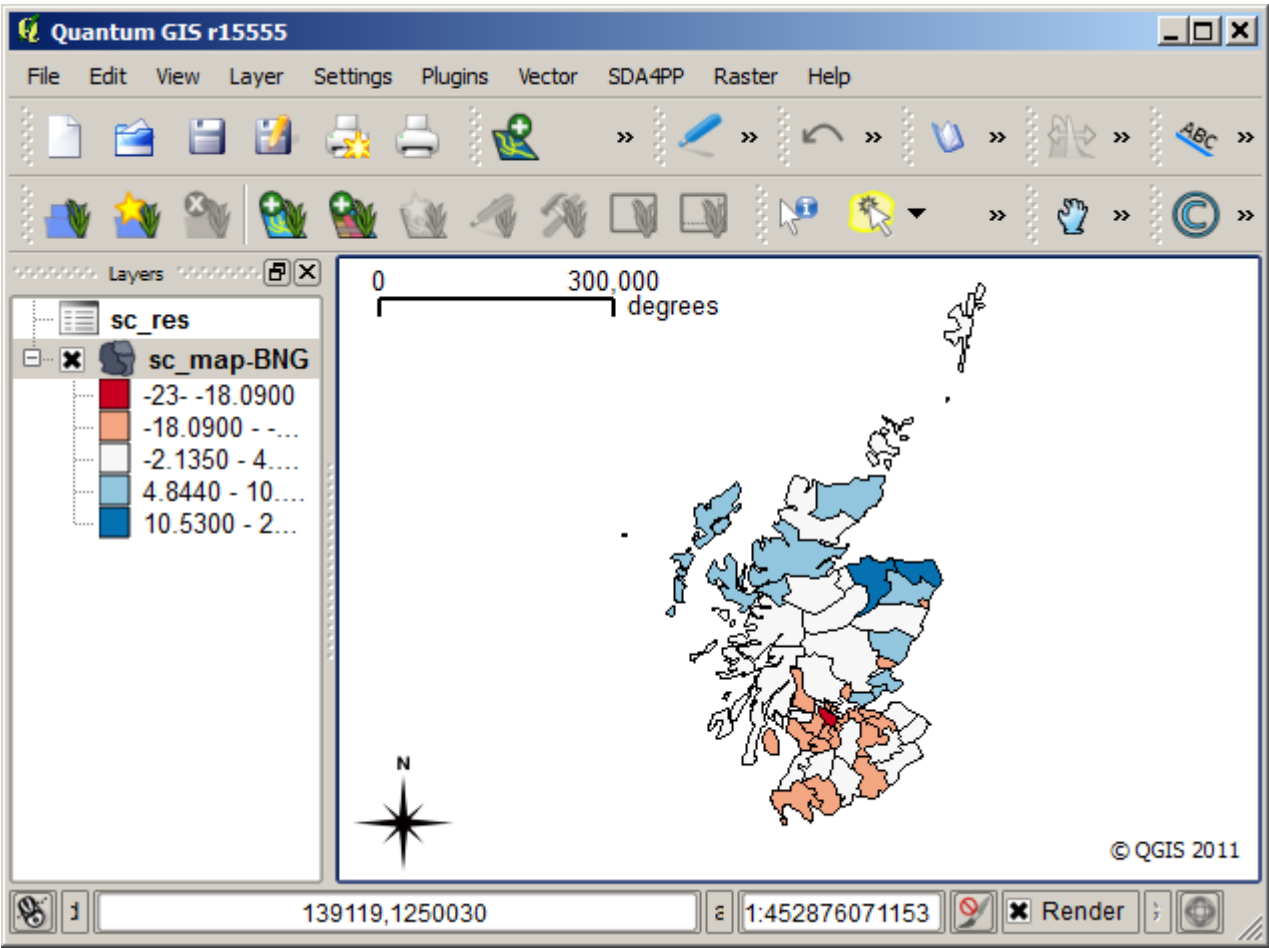
- What we've done here is account for the first order spatial properties of the data
- We now need to extend the model to account for the second order spatial properties (if they are present)
- How do we establish that second order properties are present?

Diagnostics for spatial autocorrelation

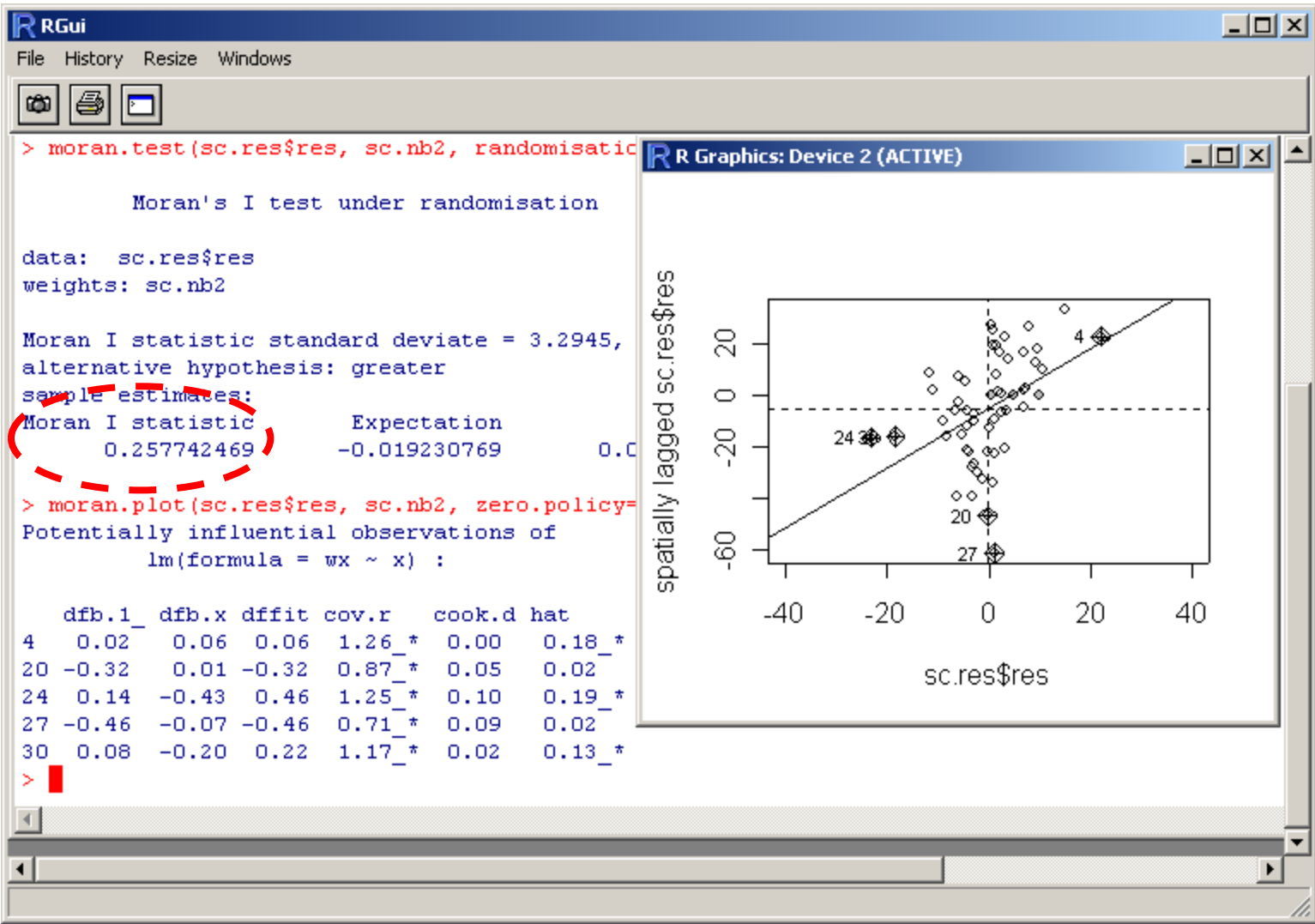
- How do we establish that second order properties are present?
 - take the residuals from the fixed-effects model and check for spatial autocorrelation in the residuals using Moran's I
 - if spatial autocorrelation is not present, then your fixed-effects model is OK
 - if spatial autocorrelation is present, you need to re-parameterise as a mixed-effects model to account for spatial autocorrelation



Choropleth maps of raw residuals from a fixed-effects model of lip cancer in Scottish districts, 1980-1985.



Moran scatterplot computed from raw residuals from a fixed-effects model of lip cancer in Scottish districts, 1980-1985.

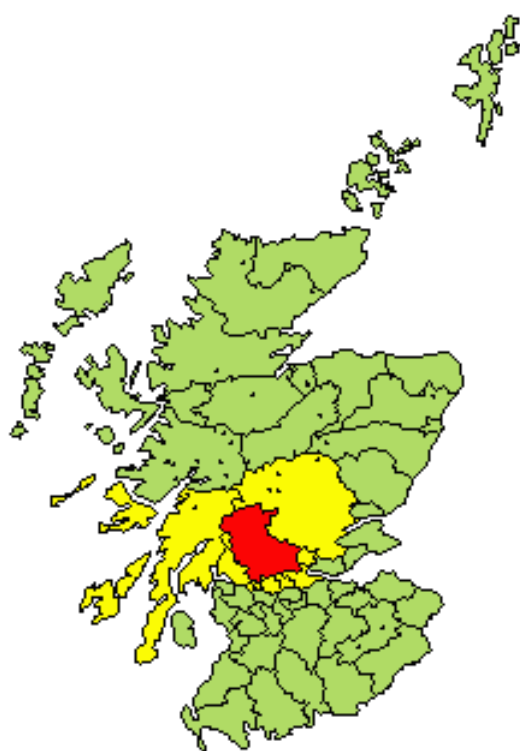


Moran's I 0.2577, P < 0.01

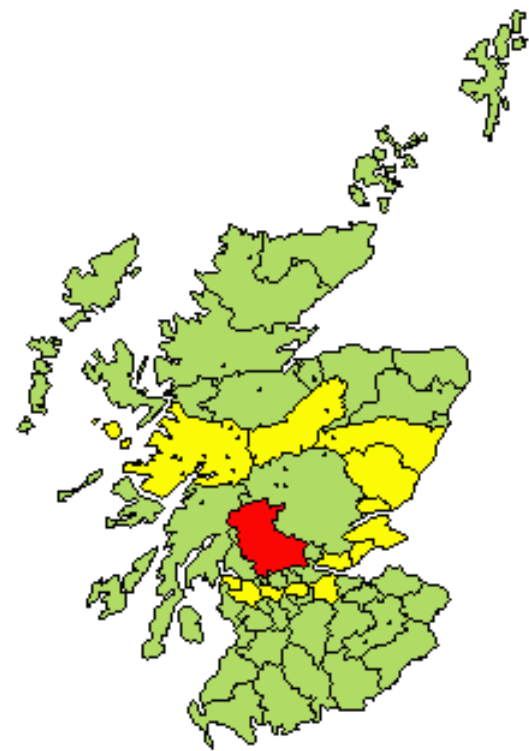
Diagnostics for spatial autocorrelation

- Other things we can do
 - plot Moran's I and its 95% confidence interval as a function different spatial proximity matrix definitions
 - this is called a correlogram

Alternative definitions for 'adjacency'. The district of interest is marked in red.

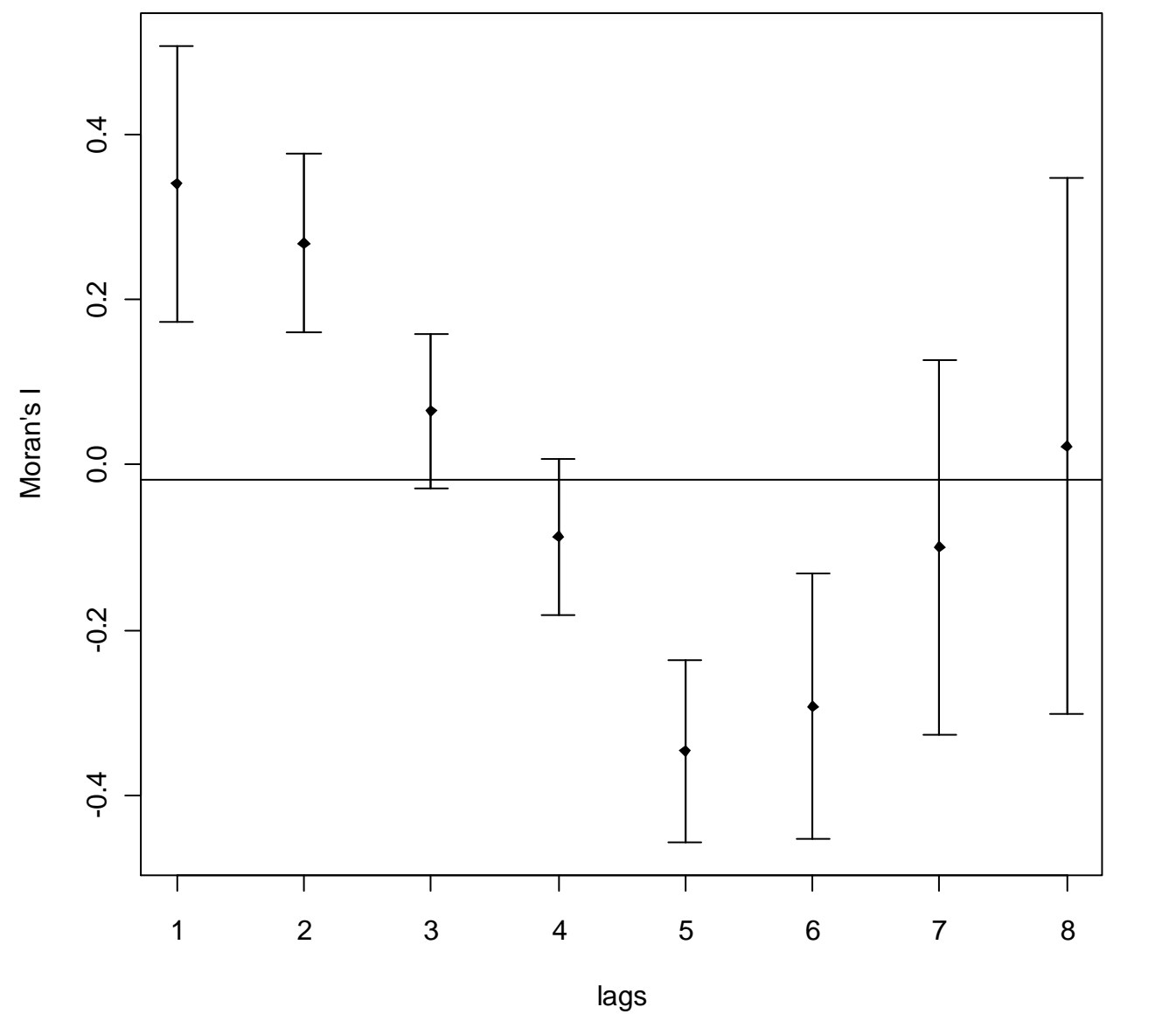


First order adjacency
Primer orden de adyacencia



Second order adjacency
Segundo orden de adyacencia

Spatial correlogram computed using the residuals from the fixed-effects model of lip cancer in Scottish districts, 1980-1985. Moran's I (and its 95% confidence interval) computed at spatial lags 1-8.



Pashupati Temple, Pashupati Nath Road, Kathmandu, Nepal January 2013.



Pashupati Temple, Pashupati Nath Road, Kathmandu, Nepal January 2013.



Roadmap

- Background
- Fixed-effects models
 - frequentist approaches
 - diagnostics for spatial autocorrelation
- Mixed-effects models
 - area data
 - point data
 - continuous data



Mixed-effects models

- What we want to achieve
 - a model that accounts for both the first order AND second order properties in our data
- How do we do it?
 - difficult (actually, very clumsy) to do using frequentist methods
 - less difficult using Bayesian methods



Mixed-effects models

- Stevenson's guide to Bayesian statistics ...

Frequentist vs Bayesian statistics

- Bayesian statistics is based on Bayes' theorem, which describes the mathematical relationship between:

the prior (or 'pre-trial') probability of an event

and

the posterior (or 'post-trial') probability of an event

given

the data that's been observed



Frequentist vs Bayesian statistics

- Suppose we have some unknown quantities θ that we want to estimate
- We collect data X

Specify a prior distribution for θ :

$p(\theta)$

Specify the distribution of X given θ :

$p(X | \theta)$

Apply Bayes' theorem to determine the distribution of θ given X :

$p(\theta | X)$



Frequentist vs Bayesian statistics

- Bayes' theorem

$$p(\theta | X) = \frac{p(\theta) p(X | \theta)}{\int p(\theta) p(X | \theta) d\theta}$$

- The prior distribution $p(\theta)$ expresses our uncertainty about θ before seeing the data
- The posterior distribution $p(\theta | X)$ expresses our uncertainty about θ after seeing the data



Frequentist vs Bayesian statistics

- In most situations these functions are difficult to evaluate analytically (due to the need to integrate high-dimensional complex distributions)
- Instead, it's straightforward to simulate realisations from the joint posterior of all parameters (using Markov chain Monte Carlo methods)
- Once we have these simulated values, posterior summaries can be obtained by simple data summaries

Frequentist vs Bayesian statistics

- So ...
 - the prior distribution $p(\theta)$ represents our uncertainty about θ before seeing the data X
 - the posterior distribution $p(\theta | X)$ represents our uncertainty about θ after seeing the data X

Frequentist vs Bayesian statistics

- The Bayesian approach

Prior The weather report said there was a 20% chance of rain today

Data I look outside and see dark clouds

Posterior Based on the weather report (my prior) and looking out the window (data gathering) I reckon there's a high probability of rain today – I'll take a raincoat to work



Mixed-effects models: area data

- Our model:

$$O_i = E_i \times SMR_i$$

$$\mu_i = E_i \times RR_i$$

$$\log \mu_i = \log E_i + \log RR_i$$

$$\log \mu_i = \log E_i + (\alpha + \beta_1 x_{1i} + \dots + \beta_m x_{mi}) + \varepsilon$$



Mixed-effects models: area data

- Our model:

$$O_i = E_i \times SMR_i$$

$$\mu_i = E_i \times RR_i$$

$$\log \mu_i = \log E_i + \log RR_i$$

$$\log \mu_i = \log E_i + (\alpha + \beta_1 x_{1i} + \dots + \beta_m x_{mi}) + \varepsilon$$

$$\log \mu_i = \log E_i + (\alpha + \beta_1 x_{1i} + \dots + \beta_m x_{mi}) + U_i + S_i + \xi_i$$

Spatially correlated random effect term. Indicates unmeasured, spatially correlated risks for disease.

Unstructured heterogeneity term to account for unmeasured non-spatial risk factors



Mixed-effects models: area data

- The spatial random effect term (S_i)
 - has a mean 0 and precision $\tau_{a.b}$
 - is normally distributed about the weighted mean of the log relative risks in the remaining areas ($i \neq j$) with variance inversely proportional to the sum of the specified spatial proximity matrix
- The non-spatial random effect term (U_i)
 - has a mean of 0 and precision $\tau_{a.h}$



Mixed-effects models: area data

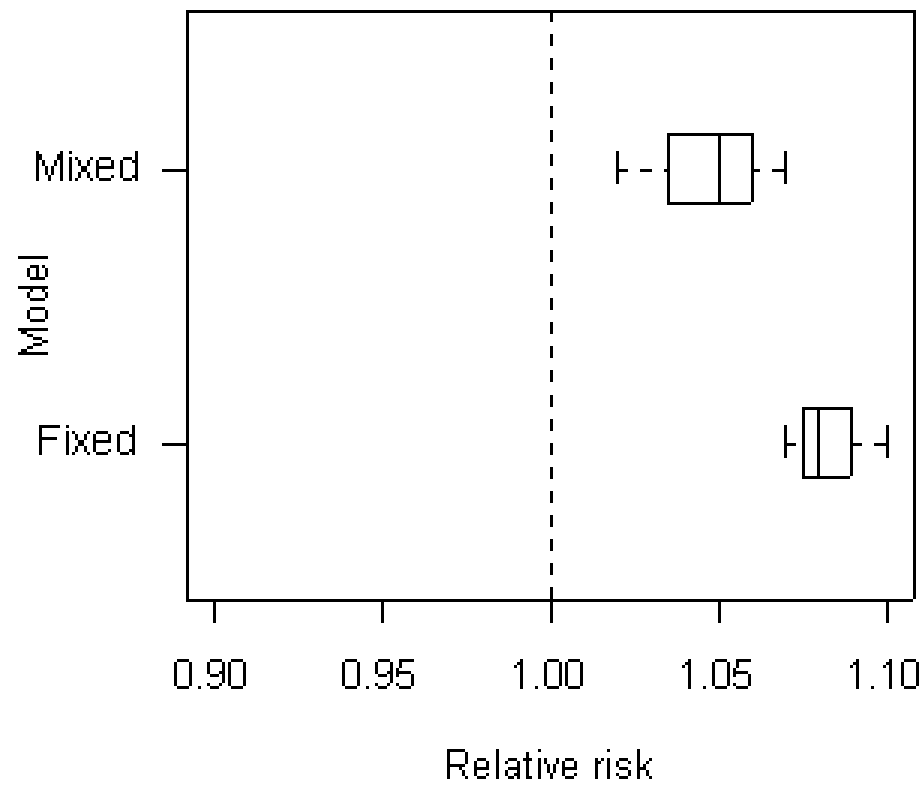
- We assign prior (actually, hyperprior) distributions to $\tau_{au.b}$ and $\tau_{au.h}$
- This controls the variability of the spatial and non-spatial random effect terms
 - when $\tau_{au.b}$ is small the variance of the spatial random effect terms will be large: a small amount of spatial smoothing is applied to the data
 - when $\tau_{au.b}$ is large the variance of the spatial random effect term will be small: a large amount of spatial smoothing is applied to the data

Variable	Post mean	SD	MC error	RR	95% CI of RR
Intercept	-0.2809	0.1215	0.0027	-	
Prop ag	0.0470	0.0128	0.0003	1.05 ^a	1.02 – 1.07
Heterogeneity ^b :					
Structured	0.6640	0.0644	0.0011	-	-
Unstructured	0.0613	0.0547	0.0025		

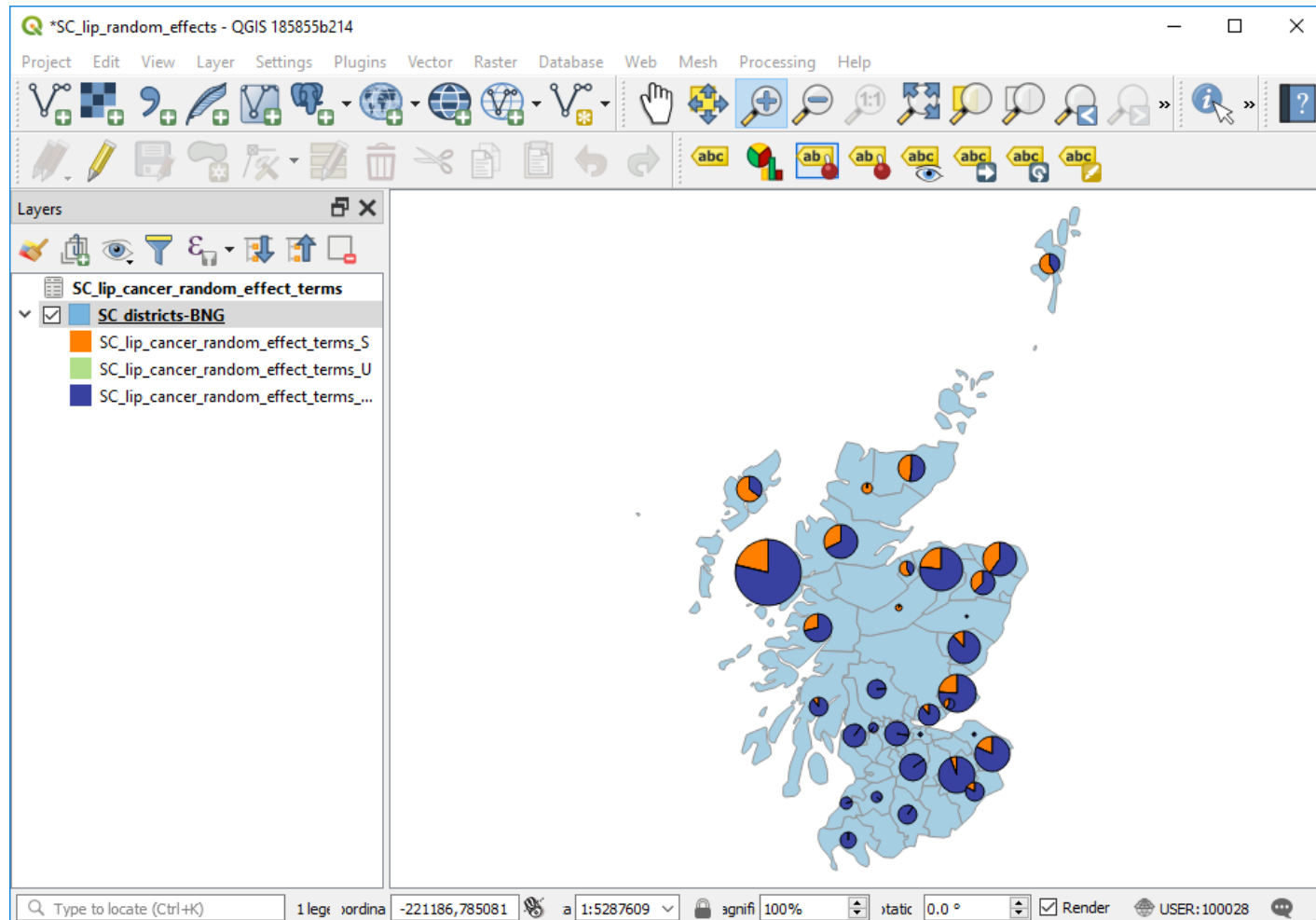
^a Interpretation: For unit increases in the percentage of the workforce involved in outdoor industry, the relative risk of lip cancer was increased by a factor of 1.05 (95% credible interval 1.02 to 1.07).

^b Empirical standard deviation of structured and unstructured heterogeneity terms.

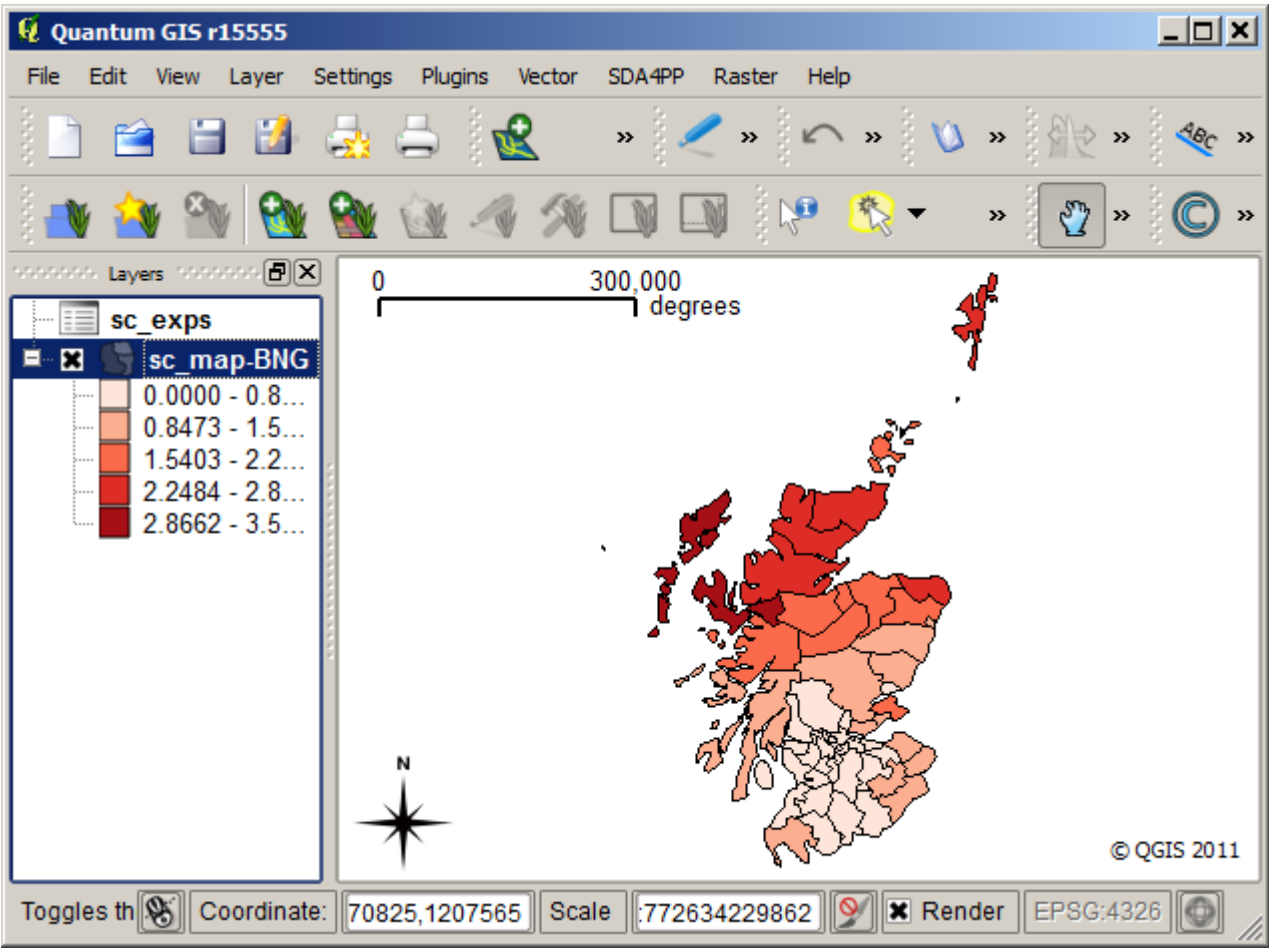
Box and whisker plots showing the 95% credible interval of the district-level relative risk of lip cancer associated with unit increases in the percentage of the workforce involved in outdoor industry.



Proportional symbol map of Scotland. For each district the magnitude of the unexplained variation in lip cancer incidence is shown by the size of each pie. The orange shading shows the proportion of unexplained variation attributable to spatially correlated effects (S_i). The light green shading shows the proportion of unexplained variation attributable to non-spatially correlated effects (U_i). The dark blue shading shows the size of the 'well behaved' residual ξ_i .



Choropleth map of exponentiated spatial random effect terms from the mixed-effects regression model of lip cancer in Scotland.



Mixed-effects models: area data

- What value did we get out of the mixed-effects model?
 - the effect of proportion of the workforce involved in outdoor industry wasn't as strong as we first thought
 - after controlling for the workforce effect, lip cancer risk increased with increasing latitude (i.e. the further north, the higher the lip cancer risk)



Kathmandu, Nepal January 2013.

