Cluster investigations using Disease mapping methods

International workshop on Risk Factors for Childhood Leukemia Berlin May 5-7 2008

Peter Schlattmann
Institut für Biometrie und Klinische Epidemiologie
CHARITÉ CAMPUS BENJAMIN FRANKLIN

Overview

- Geographic epidemiology
- Disease Clustering
- Global clustering: Random effects models
- Focused clustering: Ecological regression

Goals of geographic epidemiology and disease mapping

- Identify high risk areas for further analytical studies, e.g stomach cancer in Bavaria
- Disease surveillance, e.g. cancer registries
- Health care system evaluation, e.g. regional distribution of avoidable death
- Cluster investigations, e.g Sellafield, Krümmel

Cluster-Definition?

"Aggregation of relatively uncommon events or diseases in space and/or time in amounts that are believed or perceived to be greater than could be expected by chance"
(Last, A dictionary of epidemiology, 1995)

Distinction

- "Global clustering" Clustering in a large geographic area
  - Methods: Small-area mapping and/or spatial statistics
- "Disease clustering" due to Point Sources
  - Methods: Ecological studies based on distance as surrogate measure of exposure
  - Potential danger. Selection bias

Example: Childhood leukaemia in the GDR 1980-1989

Several Hypotheses are under discussion for childhood leukaemia. Among these hypotheses is the concern that there is an excess risk in the vicinity of nuclear power plants or installations.

An excess is likely to cause public concern.

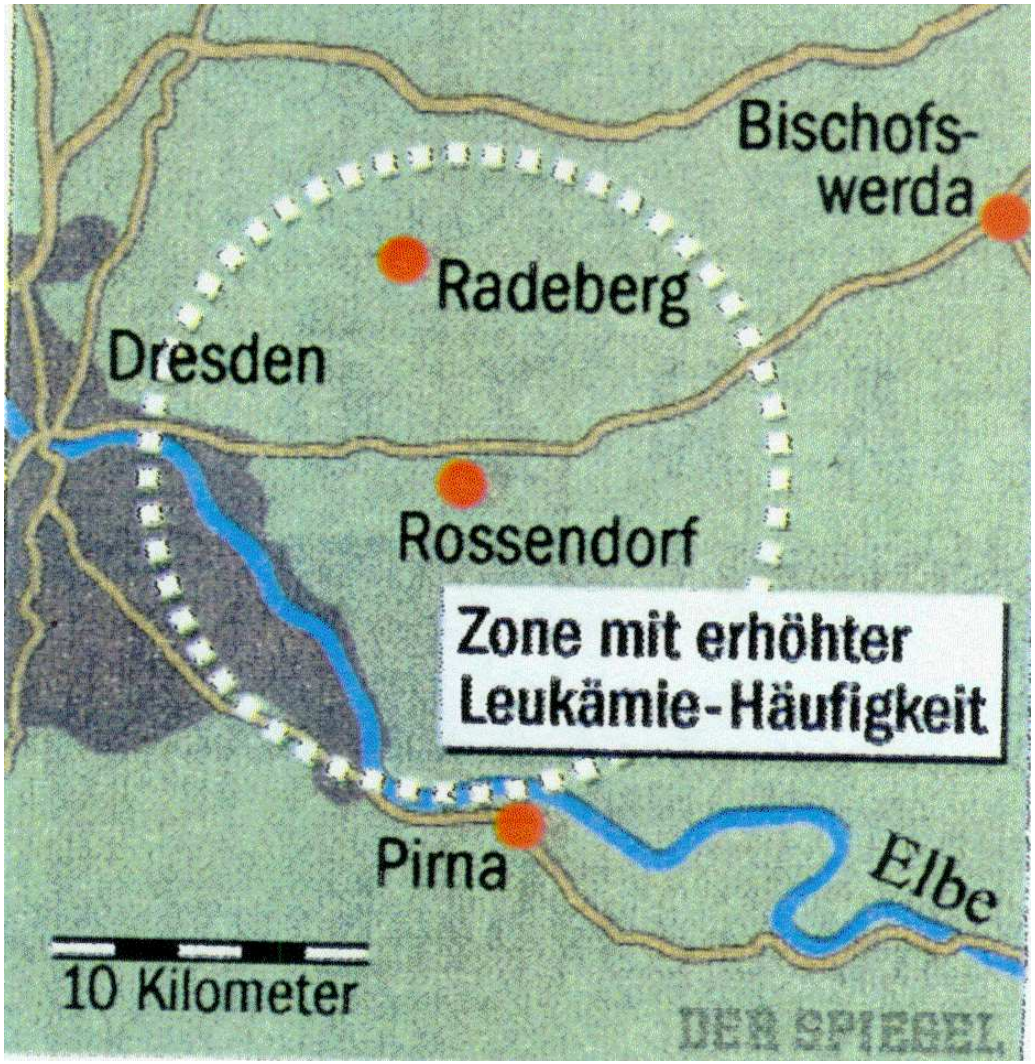

Figure 1: A putative cluster in the vicinity of Rossendorf, Der Spiegel (1996)

Investigation of this putative cluster

To avoid selection bias global clustering should be investigated before investigating a point source!
Points to consider:

- Spatial resolution
- Data are mostly only available based on administrative units, which is rarely appropriate
- A system like SAHSU as in the UK would be desirable

For the childhood leukaemia data only the resolution of "Landkreise" is available (Möhner et al, Atlas of Cancer incidence for the GDR)

Traditional methods in disease mapping

Construction of percentile maps:

- Calculate the SMR $= \frac{o_i}{E_i}$ or standardised rates for each region (o_i are the observed cases, E_i are the expected cases according to the standard)
- Classify the areas according to the percentiles of the SMR distribution
- Frequent choices: Quartiles and Quintiles

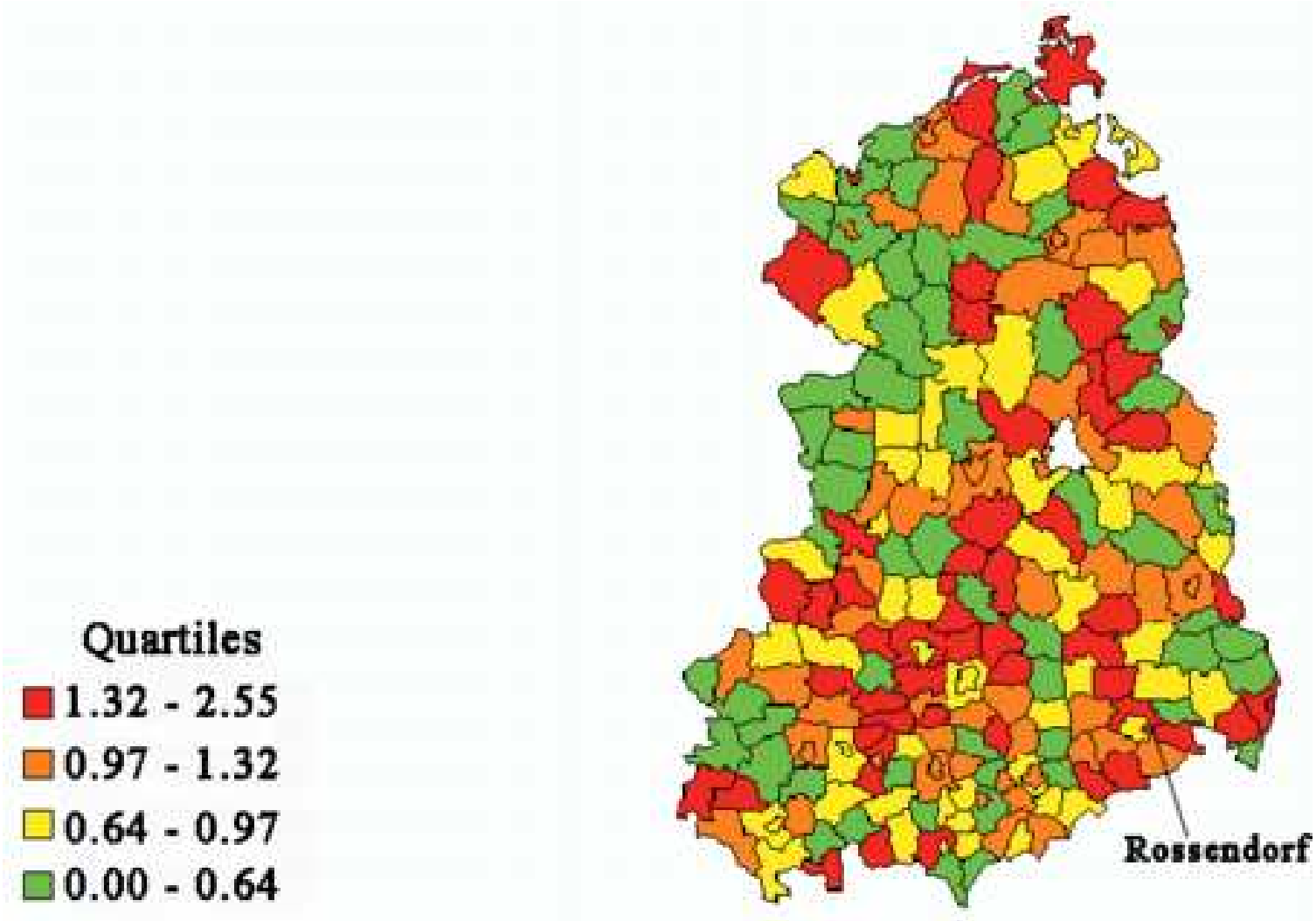

Figure 2: Percentile map: Childhood leukaemia GDR 1980-89

Problems:

- How to choose the number of categories?
- The approach is likely to reflect the heteroscedasticity of SMR's or rates due to different population size
- Example: $\text{SMR}_1 = \frac{2}{0.8} = 2.5$, $\text{SMR}_2 = \frac{20}{8} = 2.5$
- Now one case more: $\text{SMR}_1 = \frac{3}{0.8} = 3.75$, $\text{SMR}_2 = \frac{21}{8} = 2.625$

Areas with a small population tend to large relative risk (SMR) estimates!

Traditional methods II: Statistical probability maps

Since the o_i are count data frequently a Poisson distribution is assumed:

- Assume $O_i \sim Po(\theta E_i)$
- $Pr(O_i = o_i) = f(o_i, \theta, E_i) = \frac{e^{-(\theta E_i)}(\theta E_i)^{o_i}}{o_i!}$

Map construction

Calculate $P(O_i \geq o_i)$ or $P(O_i \leq o_i)$ under the null hypothesis hypothesis

- $\theta = 1$
- Or alternatively based on the Maximum-Likelihood Estimator $\hat{\theta} = \frac{\sum_{i=1}^{n} O_i}{\sum_{i=1}^{n} E_i}$, (n is the number of areas).
- Classify the individual area according to these probabilities, e.g. SMR > 1 and significant!

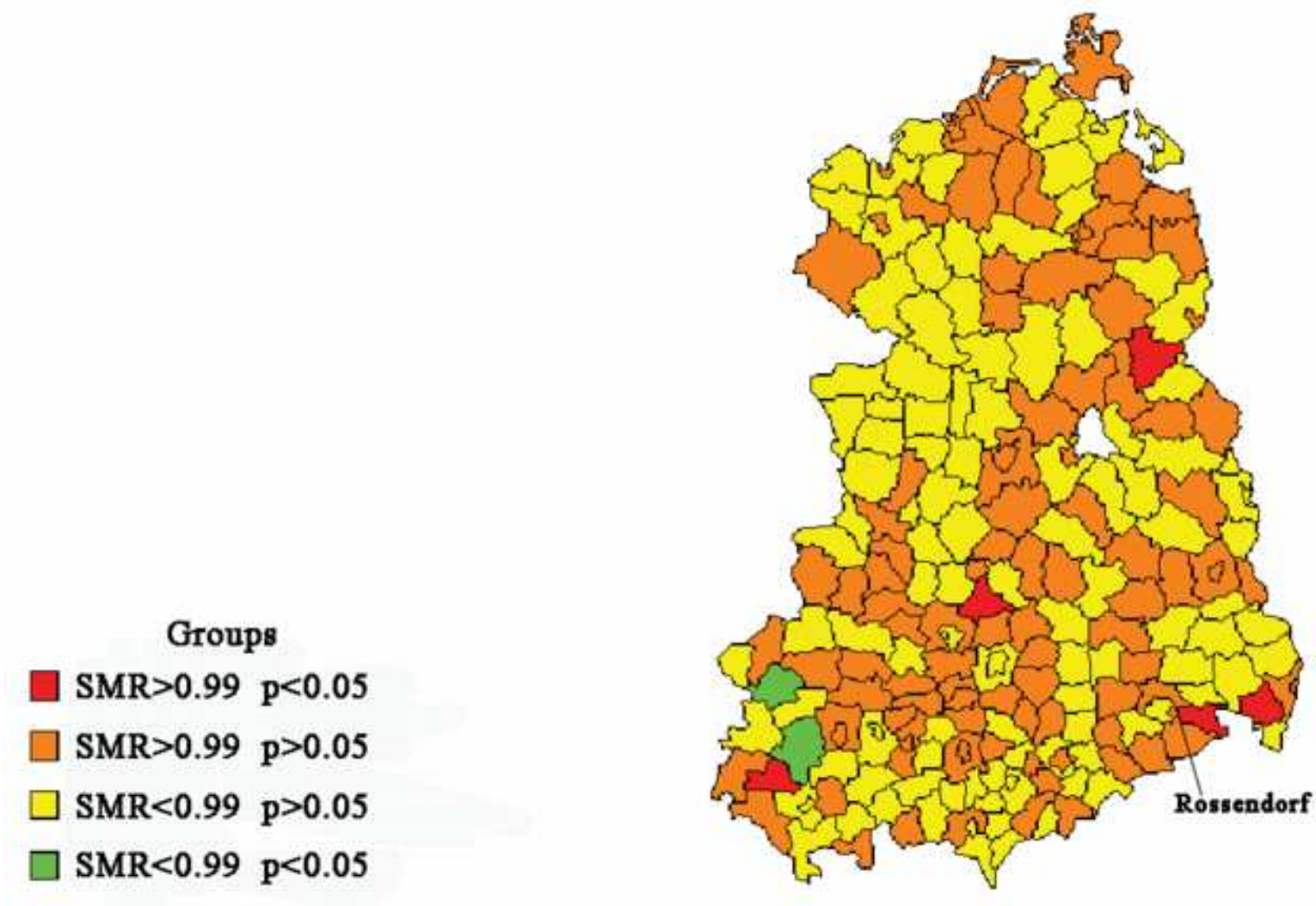

Figure 3: Statistical probability map: Childhood leukaemia GDR 1980-89

Problems

- Multiplicity: A large number of tests is performed thus the type I error is inflated
- For a significance level of 0.05 by construction 2.5% of the areas will have SMR's significantly larger than one.
- Areas with a large population tend to significant results
- The assumption of a common Poisson parameter θ may be too strong: Overdispersion occurs, this indicates heterogeneity

Is the map at random?

When assessing whether a map is at random two phenomena need to be distinguished:

- Heterogeneity: Different areas have different levels of disease risk.
  - This may due to different levels of exposure
- Autocorrelation: Neighboring areas have similar levels of disease risk. This occurs:
  - In infectious diseases or if an infectious nature of the disease is present
  - If similar patterns of exposure in neighboring areas are present

Fixed effects models in geographic epidemiology

The assumption that the observed cases O_i follow a Poisson distribution can be seen as a fixed effects model with:

$$O_i \sim \text{Poisson}(\mu_i)$$
$$\log \mu_i = \beta_0 + \log(E_i)$$

This is a Poisson regression model with an intercept only (common risk in all areas).

Testing for heterogeneity

Null hypothesis:

$$H_o : \theta_1 = \theta_2, \ldots, \theta_n \quad \text{versus}$$

$$H_1 : \exists \theta_i \neq \theta_j, i \neq j$$

$$\chi^2_{Gail} = \sum_{i=1}^{n} \frac{(o_i - \hat{\theta} E_i)^2}{\hat{\theta} E_i}$$

$$\text{Reject} : H_0 : \chi^2_{Gail} > \chi^2_{n-1,0.95}$$

Results for the childhood leukaemia data

Results were obtained with DismapWin (Schlattmnn, 1996)

```
Current map: c:\dismap\ddr87.bnd
Observed cases:OBS8089
Expected cases/ Person years:EXP8089

Gail statistic for heterogeneity:
value of test-statistic=207.627899 p-value=0.31  df=   218
```

Overdispersion

If we have a fixed effects model with $X \sim Po(\theta)$ then

- $E(X) = \theta$
- $Var(X) = \theta$
- If the empirical variance of X is larger than the theoretical variance overdispersion occurs
- This is an indication of heterogeneity
- Autocorrelation can also be a source of overdispersion

Variance decomposition for overdispersed data

If there is overdispersion present the total variance of the observed cases may be partioned in two sources of variance (Poisson variability and heterogeneity variance τ^2):

$$Var(X) = \theta + \tau^2$$

For SMR's this leads to

$$Var(O) = \theta E + \tau^2 E^2$$

$$\tau^2 = \frac{Var(O)}{E^2} - \frac{\theta}{E}$$

Random effects models in geographic epidemiology

Like in meta analysis we can develop a hierarchical model:

- First level: $O_i \sim \text{Poisson}(\theta_i E_i)$
- Second level: $\theta_i \sim P(\lambda, \tau^2)$

That is the observed cases are distributed conditional on the parameter $\theta_i E_i$ and the parameters θ_i follow a distribution with expectation λ and heterogeneity variance τ^2.

Random effects models in geographic epidemiology

If there is variability between areas, i.e. heterogeneity a random effects model may be used.

$$
\begin{aligned}
O_i &\sim \text{Poisson}(\mu_i) \\
\log \mu_i &= \beta_0 + \log(E_i) + u_i \\
u_i &\sim P(\lambda, \tau^2)
\end{aligned}
$$

This is an intercept only Poisson regression model with random effects. The distribution $P(\lambda, \tau^2)$ with expectation λ and heterogeneity variance τ^2 must be specified. A common choice is the Γ distribution with scale parameter α and shape parameter ν . (Clayton and Kaldor, 1987)

Using empirical Bayes estimators

Calculation of empirical Bayes estimators proceeds as follows:

- Idea: Use the observed data and estimate the parameters of the Gamma-distribution.
- Then use these parameters and calculate the posterior expectation of the relative risk estimate for the $i-th$ area.
- $\theta_i^{EB} = E(\theta_i | o_i, E_i, \hat{\alpha}, \hat{\nu}) = \frac{o_i + \hat{\nu}}{E_i + \hat{\alpha}}$

Properties of empirical Bayes estimators

- For "large population" areas the θ_i^{EB} will be close to the crude SMR's
- For "small population" areas the θ_i^{EB} will "shrink" to the overall mean.
- This procedure is a "borrowed strength" approach. Information is taken from the distribution of the θ_i

Results for the leukaemia data

DismapWin gives the following results:

```
Parameter estimates of Gamma-Distribution:
alpha= 1011.77472 nu= 1001.52246

Heterogeneity variance tau^2 =  0.00101
```

Please note that the heterogeneity variance is given by $\hat{\tau}^2 = \frac{\hat{\alpha}}{\hat{\nu}^2}$

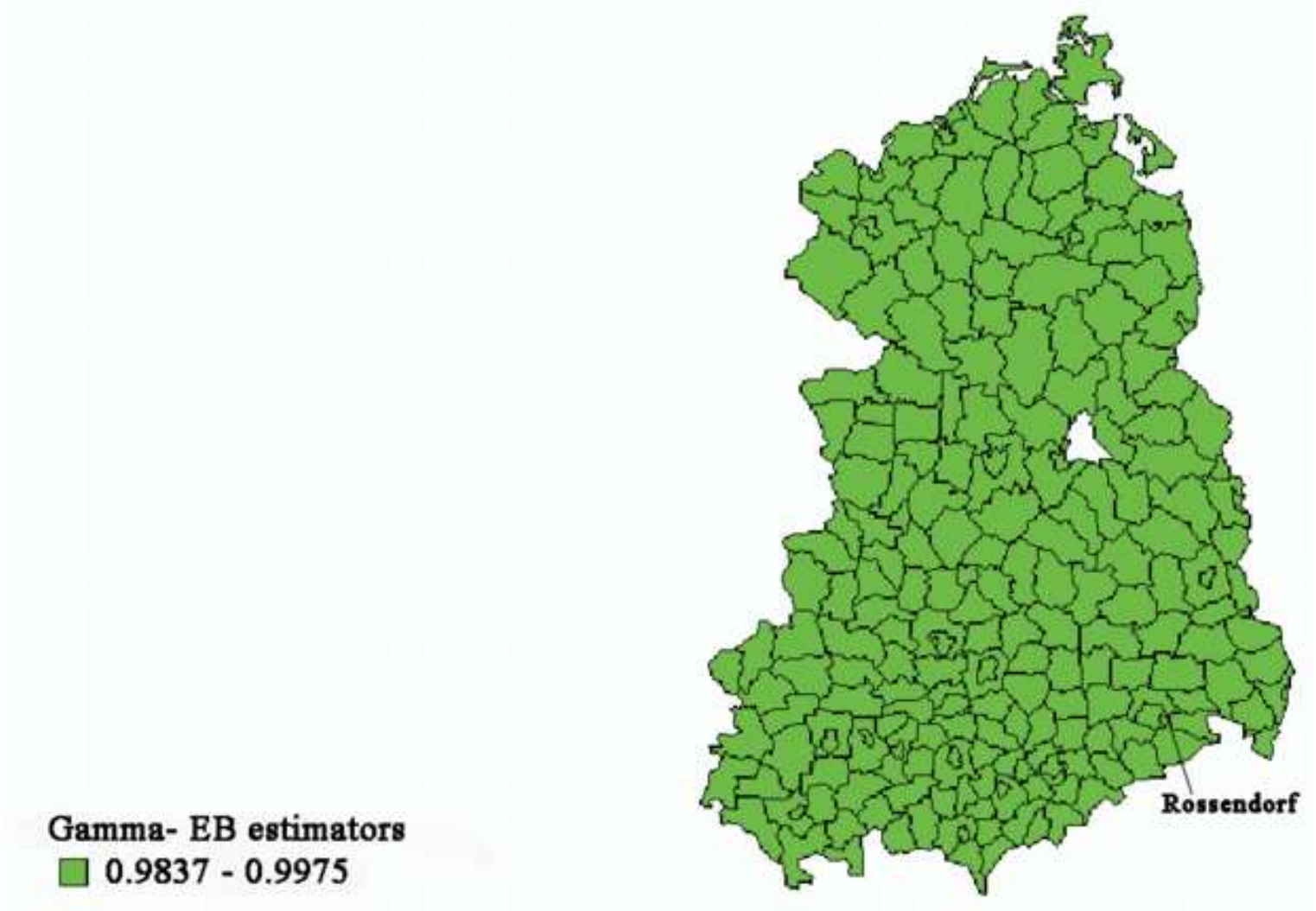

Figure 4: Map based on Gamma empirical Bayes estimates: Childhood leukaemia GDR 1980-89

A nonparametric distribution for the relative risk distribution

Here several discrete levels of risk θ_j are assumed:

$$P = \begin{bmatrix} \theta_1 & \dots & \theta_k \\ p_1 & \dots & p_k \end{bmatrix}$$

The mixture density is a weighted sum of Poisson densities for each area i:

$$f(O_i, P, E_j) = \sum_{j=1}^{k} p_j f(o_i, \theta_j, E_i), \quad with \quad \sum_{j=1}^{k} p_j = 1 \quad and \quad p_j \geq 0, \quad j = 1, \dots, k$$

Parameters estimation

Please note that the model consists of the following parameters:

- The number of components k
- The k unknown relative risks $\theta_1, \ldots, \theta_k$
- The $k-1$ unknown mixing weights $p_1, \ldots, p_{k-1}$

For finding the maximum likelihood estimates there are no closed form solutions available, iterative solutions are available.

Parameter estimation

Parameter estimation may be done as follows:

- Define a grid of possible subpopulation means
- Identify grid points with positive support
- Calculate a refined solution using a different algorithm

DismapWin: Results for the leukaemia data

```
weight  0.00000 parameter   0.00000
weight  0.00159 parameter   0.18571
weight  0.00000 parameter   0.37143
weight  0.00000 parameter   0.55714
weight  0.00000 parameter   0.74286
weight  0.70388 parameter   0.92857
weight  0.29453 parameter   1.11429
weight  0.00000 parameter   1.30000
.....
weight  0.00000 parameter   2.60000

log-likelihood at iterate=-458.07010
```

Results continued

```
Refined solution with fixed support size:

weight=  0.00817  parameter=  0.16954
weight=  0.70481  parameter=  0.99439
weight=  0.28702  parameter=  0.99439

 Heterogeneity variance tau^2 =  0.00551

log-likelihood at iterate=-457.38644
```

Results III

```
Fixed support size solution:

weight=  1.00000  parameter=  0.98962

Heterogeneity variance tau^2 =  0.00000

 log-likelihood at iterate=-457.43839
```

The heterogeneity variance for the semiparametric model

Please note that the heterogeneity variance is given by

$$\tau^2 = \sum_{j=1}^{k} p_j(\theta_j - \bar{\lambda})^2$$

$$\bar{\lambda} = \sum_{j=1}^{k} p_j \lambda_j$$

Classification of the individual area

The individual area may be classified using Bayes' theorem:

$$Pr(Z_{ij} = 1 | O_i, \hat{P}, E_i) = \frac{\hat{p_j} f(o_i, \hat{\theta}_j, E_i)}{\sum_{l=1}^{k} \hat{p_l} f(o_i, \hat{\theta}_l, E_i)}$$

The i-th area is then assigned to that subpopulation j for which it has the highest posterior probability of belonging. Z_{ij} indicates component membership.

The empirical Bayes relative risk estimate

Taking expectations gives the empirical Bayes estimate of the relative risk θ_i.

$$\hat{SMR}_i = E(\theta_i | O_i, \hat{P}, E_i) = \frac{\sum_{j=1}^{k} \hat{p}_j f(o_i, \hat{\theta}_j, E_i) \hat{\theta}_j}{\sum_{l=1}^{k} \hat{p}_l f(o_i, \hat{\theta}_l, E_i)}$$

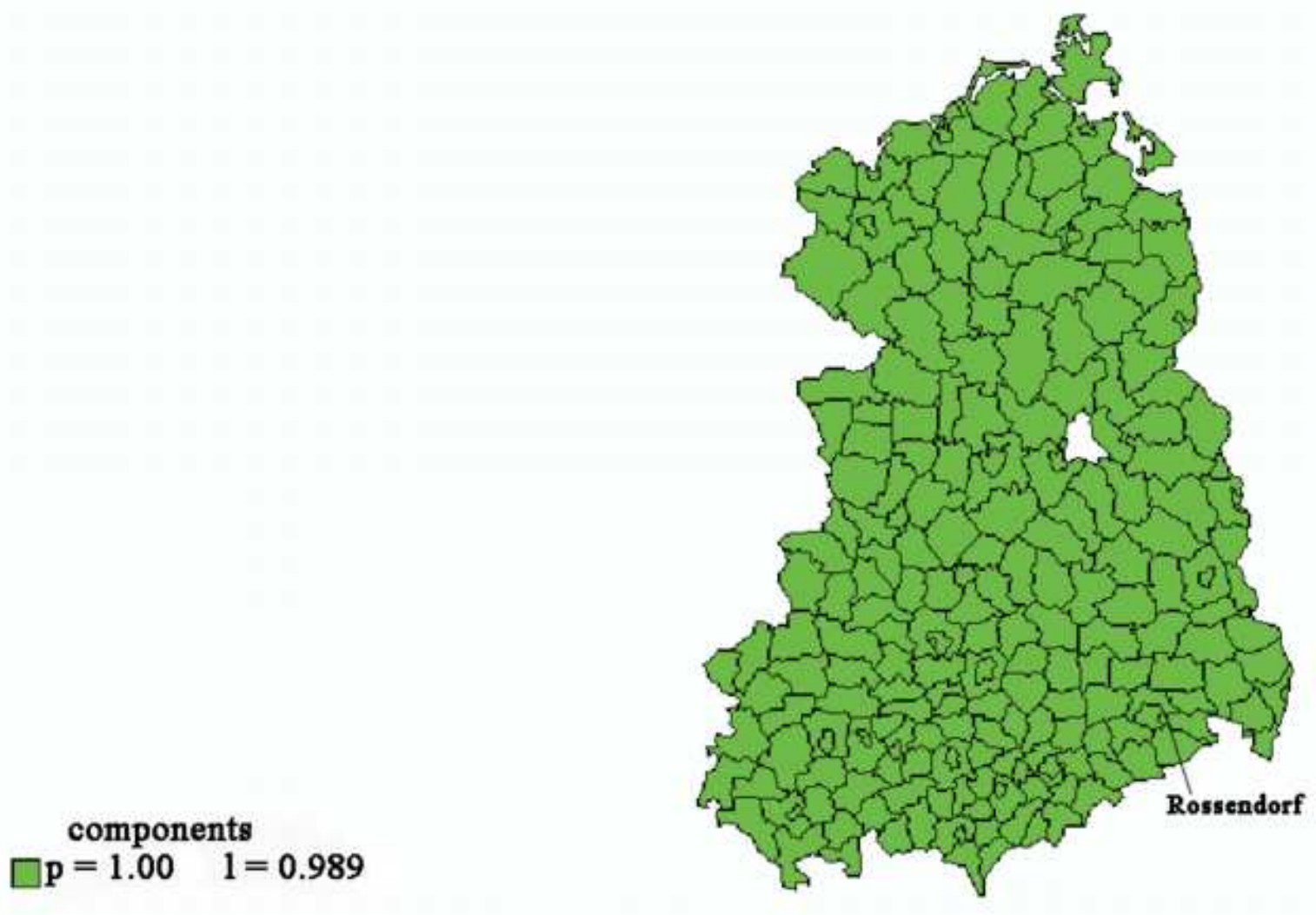

Figure 5: Map based on mixture model classification: Childhood leukaemia GDR 1980-89

Evaluation of the leukaemia data

There was public concern whether there is an excess of childhood leukaemia in the vicinity of the nuclear installation Rossendorf close to Dresden.

Table 1: Relative risk estimates for areas close to Rossendorf

Area	Cases	Expected Cases	SMR	EB	MIX-EB	$Pr(O \geq o_i)$
Dresden (City)	32	34.41	0.93	0.99	0.99	0.618
Dresden (area)	10	6.8	1.47	0.99	0.99	0.085
Sebnitz	9	3.53	2.55	0.99	0.99	0.004
Pirna	7	7.07	0.99	0.99	0.99	0.412
Bischofswerda	2	4.44	0.45	0.99	0.99	0.820

Conclusion

Based on the spatial resolution available (Landkreise) no excess risk could be identified!

Full Bayesian models

A full Bayesian model has three levels:

- First level: $O_i \sim \text{Poisson}(\theta_i E_i)$
- Second level: $\theta_i \sim P(\lambda, \tau^2)$
- Third level: The parameters λ, τ^2 have itself a distribution

Solutions can be obtained using Monte Carlo Markov Chain methods. A software implementation is given by WinBugs and GeoBugs (Best and Spiegelhalter, 1995, 2000)

Random effects models in geographic epidemiology with a spatial component

When dealing with spatial data two phenomena are sometimes distinguished

- Unstructured heterogeneity: A random effects model
- Structured heterogeneity: A random effects model with spatial dependency

The notation for the spatial random effects model

$$
\begin{aligned}
O_i &\sim \text{Poisson}(\mu_i) \\
\log \mu_i &= \beta_0 + \log(E_i) + u_i + v_i \\
u_i &\sim N(\lambda, \sigma^2) \\
v_i &\sim N(0, \tau^2 \boldsymbol{W}^{-1})
\end{aligned}
$$

This is an intercept only Poisson regression model with random effects u_i and a spatial term v_i. This type of model is mainly fitted with Bayesian methods.

Conclusions for Public Health practice

- When producing maps besides descriptive maps (Percentiles) smoothed maps based on (empirical) Bayes methods are desirable

- There are a variety of models for (empirical) Bayesian mapping. Which one should be used? Lawson et al (2000) found in a simulation study:
  - Full Bayesian model: Overall best performance
  - Gamma empirical Bayes model: Overall second best performance!!
  - Mixture models: Sometimes oversmoothing, but useful for classification

- More recommendations in the proceedings of the WHO workshop (1997) on disease mapping and risk assessment

Testing for Autocorrelation

We want to test the Null Hypothesis that all risk estimates are independently distributed against the alternative that at least one pair is correlated. For example for categorical data this gives:

- $H_0 : \forall P(X_i = x_i, X_j = x_j) = (P(X_i = x_i)P(X_j = x_j)$ vs
- $H_1 : \exists P(X_i = x_i, X_j = x_j) \neq (P(X_i = x_i)P(X_j = x_j)$

Definition of neighborhood

Two ares are said to be neighbors, if they have:

- One corner in common (Bishop's case)
- One line in common (King's case)
- One corner or one line in common (Queen's case)

Testing for autocorrelation: Moran's I

Moran's I is frequently used to test for autocorrelation in disease maps. Calculation of Moran's I requires the following steps:

- Define a nxn neighborhood matrix $\boldsymbol{W}$ with w_{ij} :
  - $w_{ij} = 1, i \neq j$ if areas R_i and R_j are adjacent
  - $w_{ij} = 0$, otherwise
- Calculate the sum $S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$
- Calculate $I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

Moran's I continued

- Calculate $E(I)$ and $Var(I)$
- Reject H_0 if $|Z_i| > Z_{1-\alpha/2}$, $Z_I = \frac{I - E[I)}{\sqrt{Var(I)}}$

Here x_i denote the SMR's or rates of the individual regions and $\bar{x}$ their mean.

Results for the leukaemia data

DismapWin gives the following output for Moran's I:

```
Moran´s I
value of test-statistic=0.020720 p-value=0.279244

Additional information:
expectation=-0.004587 variance=0.001871 z-value= 0.585090
```