SPACE -TIME MIXTURE MODELING FOR DISEASE MAPPING

Dankmar Böhning[1]
Peter Schlattmann and Heinz Holling

Department of Epidemiology
Institute of Social Medicine and Medical Psychology, Free University Berlin

Invited Presentation at the 98 IBC in Cape Town

Session on

Space-Time Modeling of Epidemiological Data

Organized by Andrew Lawson

OVERVIEW

- Introduction

- Disease Mapping in Space Using Mixtures

- Disease Mapping in Space *and Time*

FREQUENT OBJECTIVE IN GEOGRAPHIC EPIDEMIOLOGY

to present that part of the spatial variation of a disease occurrence distribution, which cannot be explained by the different distribution of **known factors** in the various regions nor is due to **random** variation

HOPE: hints to **unknown** risk factors!

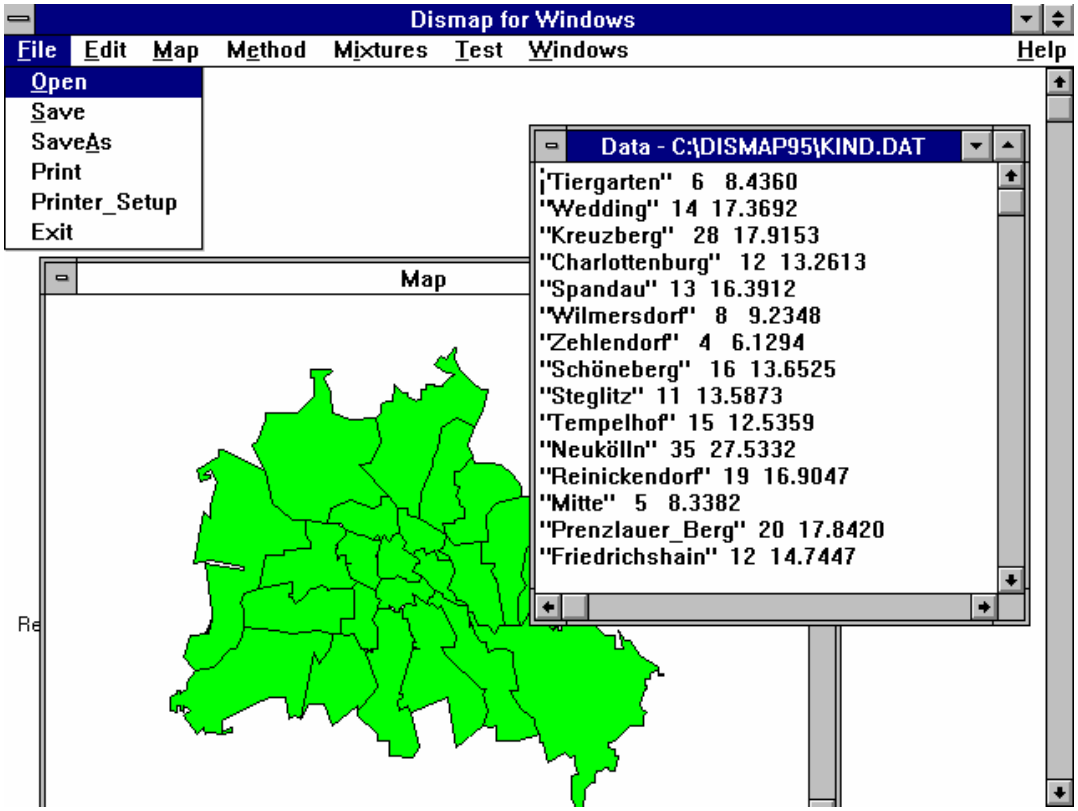
INTRODUCTION

frequently used OCCURRENCE MEASURE in epidemiology and public health institutions

$SMR_i = O_i/E_i$ in the i-th region

O_i observed death (mortality) or disease (morbidity) cases

E_i expected cases computed from external reference population



TWO „CLASSIC" BIOMETRIC METHODS

a) classification based on a certain percentile of the empirical SMR-distribution

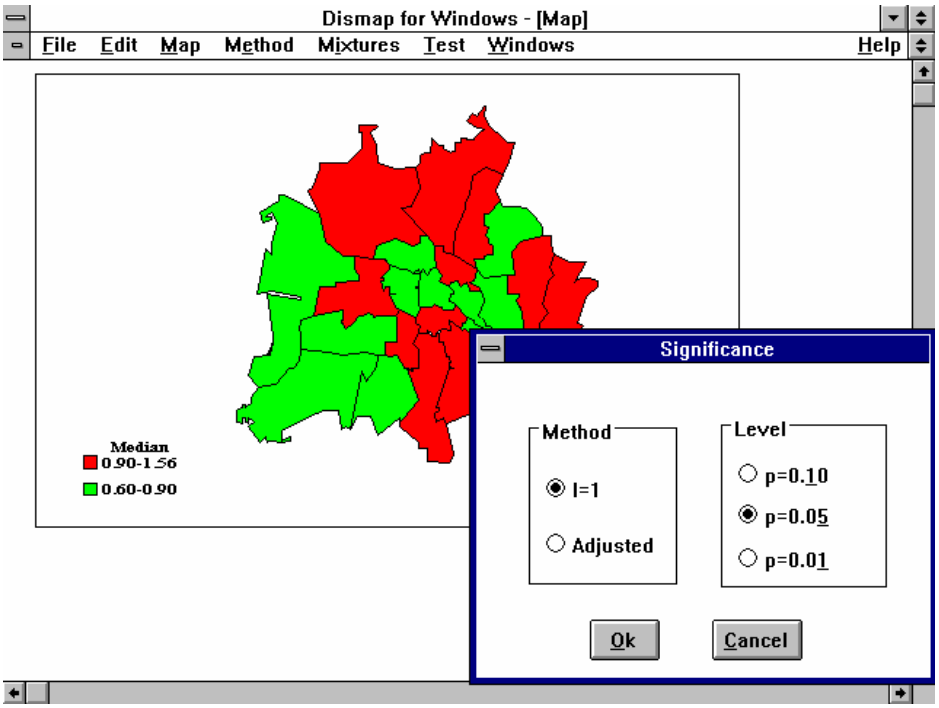
b) Poissondistribution $Po(o_i,\lambda E_i)$

$$= \exp(\lambda E_i)(\lambda E_i)^{o_i} / o_i!$$

classification based on the P-value under the Poissondistribution

$$P(O_i \geq o_i) = Po(o_i,\lambda E_i) + Po(o_i+1,\lambda E_i) + \ldots$$

INFANT MORTALITY IN BERLIN 1991



Disadvantage of conventional methods:

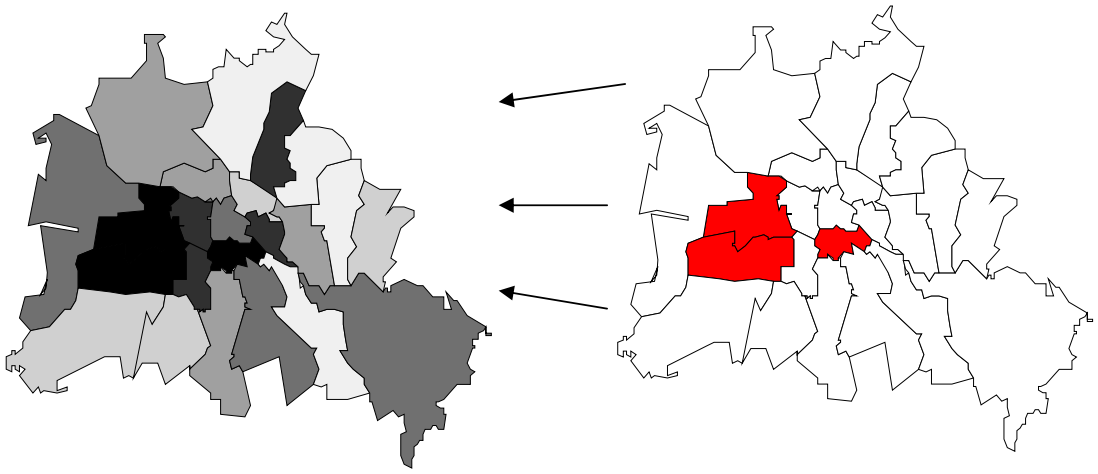
represent random variation on the map

Disease Mapping in Space using Mixtures

observed map of risk structure

true, but unobserved map of risk structure (*latent* or *hidden* map)

23 city areas of Berlin



n areas given

k (unknown) components: $\lambda_1,...,\lambda_k$

$O_1,O_2,...,O_n$
$E_1,E_2,...,E_n$ } observed data
$\mathbf{Z}_1,\mathbf{Z}_2,...,\mathbf{Z}_n$ unobserved data

$\mathbf{Z_i} = (Z_{i1}, ...,Z_{ik})$ with $Z_{ij} = 1$ meaning: area i is from component with risk λ_j

MODEL FOR: $SMR_i = O_i/E_i$ is

$$O_i \sim Po(\lambda_j E_i)$$

conditionally O_i is from area with mortality rate λ_j

let p_j probability of being from component with $\lambda=\lambda_j$

then, *unconditionally*

$$O_i \sim p_1 Po\,(\lambda_1 E_i) + p_2 Po\,(\lambda_2 E_i) + ...+ p_k Po\,(\lambda_k E_i)$$

is a *nonparametric mixture of Poisson* distributions

INTERPRETATION

k= 1: *homogeneous* risk structure
k= 2: two risk groups
k= 3: three risk groups } *heterogeneous* risk structure
...

maximum likelihood estimation of the parameters $p_1,\lambda_1,...,p_k,\lambda_k$ inclusively number of components k,

$$P = \begin{pmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_k \\ p_1 & p_2 & \dots & p_k \end{pmatrix}$$

leads to the *nonparametric maximum likelihood estimate* of a mixing distribution
(Laird 1978, Simar 1974, Lindsay 1983, ... Lesperance and Kalbfleisch 1992, Aitkin 1996, ...)

theoretical well-investigated (Lindsay 1995)
algorithmically possible (Böhning 1995 *JSPI*)

Map CONSTRUCTION

each area i is classified into component (color) j such that the posterior distribution

$$f(\lambda_j/o_i,E_i,\hat{P}) = Po(o_i, \hat{\lambda}_j E_i)\,\hat{p}_j / \Sigma_l\, Po(o_i, \hat{\lambda}_l E_i)\,\hat{p}_l$$

is maximized !

TWO APPLICATIONS

Health Region: 219 counties of the former German Democratic Republic (*The 5 New States of Germany*)

1) Incidence on Lung Cancer (ICD 162) for Women
   - 1980 –1989

   Estimate of Mixing Distribution (NPMLE):

$$\begin{pmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_k \\ p_1 & p_2 & \dots & p_k \end{pmatrix} = \begin{pmatrix} 1.33 & 0.98 & 0.78 & 0.56 \\ 0.12 & 0.27 & 0.51 & 0.10 \end{pmatrix},$$

$$\hat{k} = 4$$

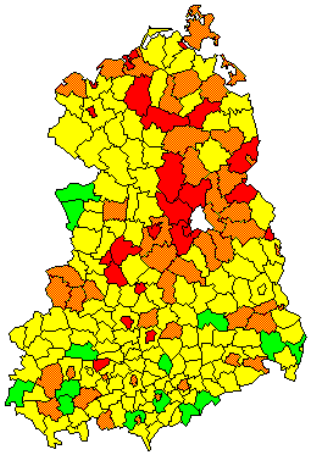
2) Incidence on Mamma-Carcinoma (ICD 174)
   - 1980 – 1989

   Estimate of Mixing Distribution (NPMLE):

$$\begin{pmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_k \\ p_1 & p_2 & \dots & p_k \end{pmatrix} = \begin{pmatrix} 1.21 & 1.15 & 1.04 & 0.92 & 0.73 \\ 0.02 & 0.06 & 0.20 & 0.68 & 0.04 \end{pmatrix},$$

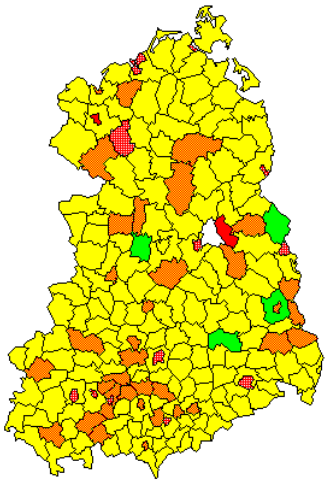
$$\hat{k} = 5$$

ICD 162 in women GDR 1980-89



risk structure
p=0.12 rr=1.33
p=0.27 rr=0.98
p=0.51 rr=0.78
p=0.10 rr=0.56

ICD 174 in women GDR 1980-89



risk structure
p=0.02 rr=1.21
p=0.06 rr=1.15
p=0.20 rr=1.04
p=0.68 rr=0.92
p=0.04 rr=0.73

Disease Mapping in Space *and Time*

n areas given for T periods

$O_1^{(1)},O_2^{(1)},...,O_n^{(1)}$; $O_1^{(2)},O_2^{(2)},...,O_n^{(2)}$; ....
$O_1^{(T)},O_2^{(T)},...,O_n^{(T)}$
$E_1^{(1)},E_2^{(1)},...,E_n^{(1)}$; $E_1^{(2)},E_2^{(2)},...,E_n^{(2)}$; ....
$E_1^{(T)},E_2^{(T)},...,E_n^{(T)}$;
$\mathbf{Z}_1^{(1)},\mathbf{Z}_2^{(1)},...,\mathbf{Z}_n^{(1)}$; $\mathbf{Z}_1^{(2)},\mathbf{Z}_2^{(2)},...,\mathbf{Z}_n^{(2)}$; ....
$\mathbf{Z}_1^{(T)},\mathbf{Z}_2^{(T)},...,\mathbf{Z}_n^{(T)}$;

IN THE TWO APPLICATIONS

Incidence on Lung Cancer (ICD 162) for Women
- 1980 – 1989
- 1970 – 1979
- 1960 – 1969

Incidence on Mamma-Carcinoma (ICD 174)
- 1980 – 1989
- 1970 – 1979
- 1960 – 1969

TWO ANALYSIS OPTIONS

a)

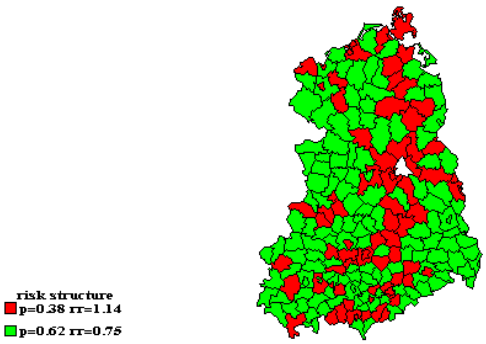
$\mathbf{Z_i^{(t)}} = (Z_{i1}^{(t)}, ...,Z_{ik}^{(t)})$ with $Z_{ij}^{(t)} = 1$ meaning:

for *each* period: area i is from component with risk $\lambda_j^{(t)}$, j=1,...,k , where k might depend on t

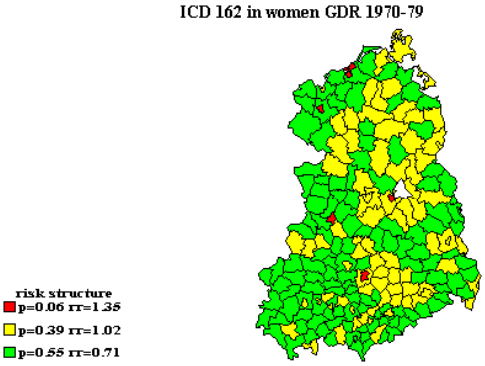
Result: T mixture models, for each time period one:

$$O_i^{(t)} \sim p_1^{(t)}\, Po\,(\lambda_1^{(t)}\, E_i^{(t)}) + p_2^{(t)}\, Po\,(\lambda_2^{(t)}\, E_i^{(t)}) + ...+ p_k^{(t)}\, Po\,(\lambda_k^{(t)}\, E_i^{(t)})$$

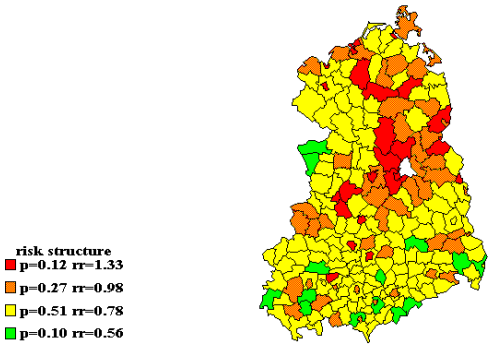
ICD 162 in women GDR 1960-69



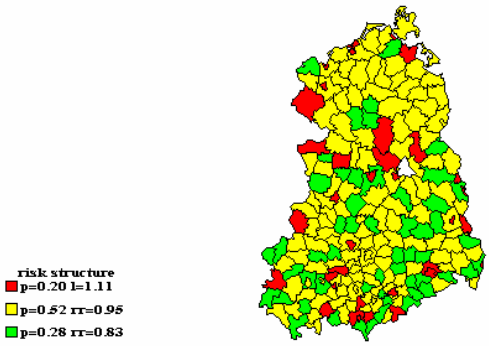
ICD 162 in women GDR 1970-79



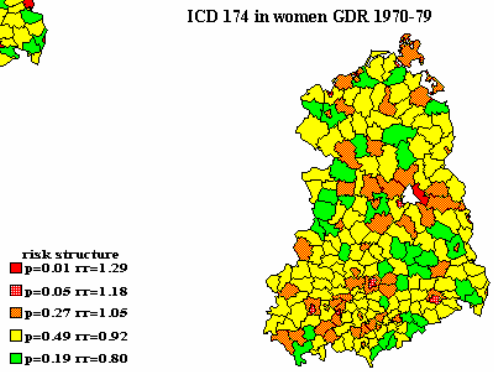
ICD 162 in women GDR 1980-89



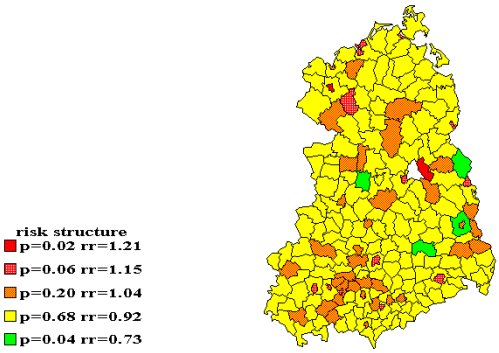
ICD 174 in women GDR 1960-69



ICD 174 in women GDR 1970-79



ICD 174 in women GDR 1980-89



b)
$\mathbf{Z_i^{(t)}} = (Z_{i1}^{(t)}, ..., Z_{ik}^{(t)})$ with $Z_{ij}^{(t)} = 1$ meaning:

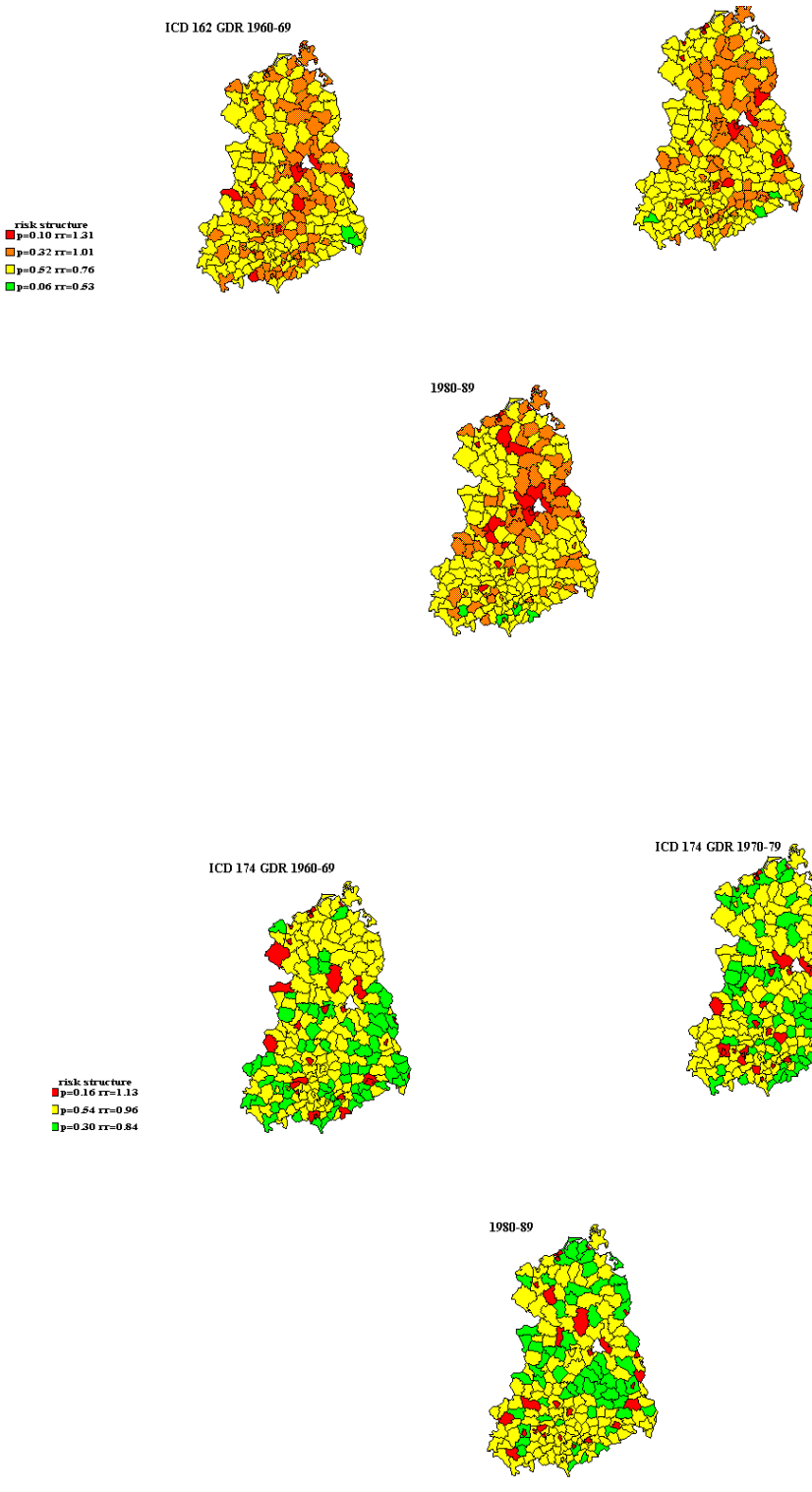
for *all* periods : area i is from component with risk λ_j, j=1,...,k , where λ_j and k *does not* depend on t

Result: one mixture model

$$O_i^{(t)} \sim p_1 Po(\lambda_1 E_i^{(t)}) + p_2 Po(\lambda_2 E_i^{(t)}) + ... + p_k Po(\lambda_k E_i^{(t)})$$

Note: in *both* cases T maps are drawn, however in
a) there are $k^{(1)} + ... + k^{(T)}$ colors
b) there are k colors

INTERPRETATION:
option b) is attractive since it is looking for joint *space-time* components (clusters)


ICD 162 GDR 1960-69
1980-89
risk structure
p=0.10 rr=1.31
p=0.32 rr=1.01
p=0.52 rr=0.76
p=0.06 rr=0.53
ICD 174 GDR 1960-69
ICD 174 GDR 1970-79
1980-89
risk structure
p=0.16 rr=1.13
p=0.54 rr=0.96
p=0.30 rr=0.84

Option c)

MORE COMPLEX MIXED MODELING

$$O_i^{(t)} \sim p_1 \, Po\,(\lambda_{i1}^{(t)} E_i^{(t)}) + p_2 \, Po\,(\lambda_{i2}^{(t)} E_i^{(t)}) + \ldots + p_k Po\,(\lambda_{ik} E_i^{(t)})$$

however,

$$\log(\lambda_{ij}^{(t)} E_i^{(t)}) = \log(E_i^{(t)}) + \log(\lambda_{ij}^{(t)}) = \log(E_i^{(t)}) + \alpha_j + \beta_j t + \textit{further covariates}$$

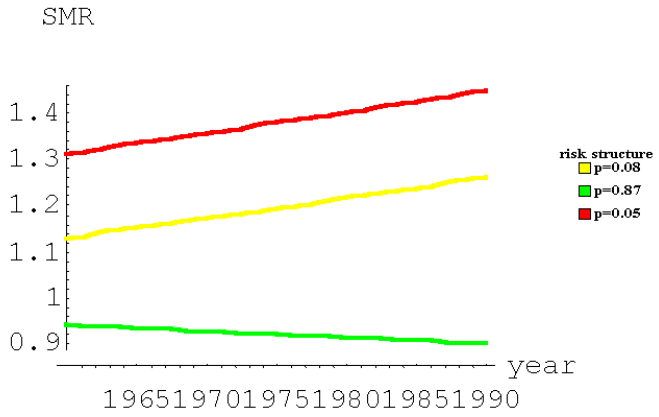
leading to

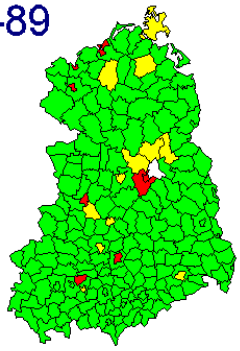
$$O_i^{(t)} \sim \sum_{j=1}^{k} p_j \, Po\,\{E_i^{(t)} \exp(\alpha_j + \beta_j t)\}$$

NOTE: in this case the map consists of the k log-linear models (each model one color)

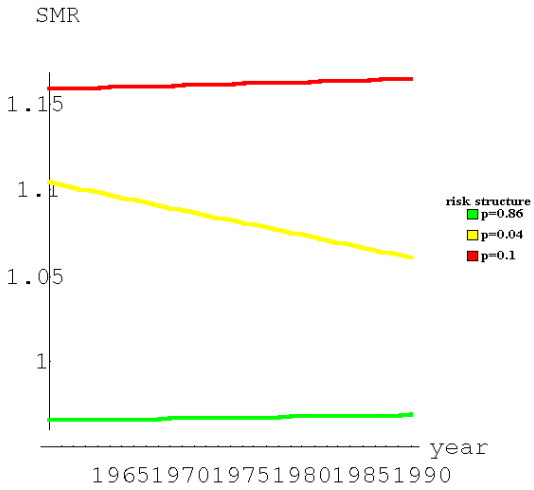
DISADVANTAGE: increased complexity: mixing over intercept or effect parameter? or both?

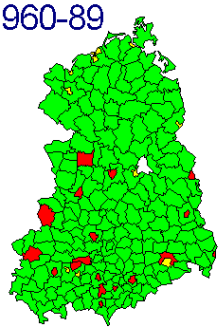
ICD 162 GDR 1960-89





ICD 174 GDR 1960-89





In Conclusion

⇒ investigated possibilities of consistently estimating heterogeneity via nonparametric mixture models

⇒ availability of these procedures in C.A.MAN and DISMAP

SOME HINTS TO RECENT REFERENCES

Lawson, Böhning, Biggeri, Lesaffre, Viel, and Bertolini (Eds.), 1998. *Disease Mapping and Risk Assessment for Public Health*. Wiley & Sons.

Böhning, 1998. *Computer Assisted Analysis of Mixtures and Applications: Disease Mapping, Meta-Analysis and others*. Chapman & Hall