Introduction
Spatial Statistics

Simon Rückinger
Institut für Soziale Pädiatrie und Jugendmedizin
Abteilung für Epidemiologie

simon.rueckinger@med.uni-muenchen.de

Aim of this session

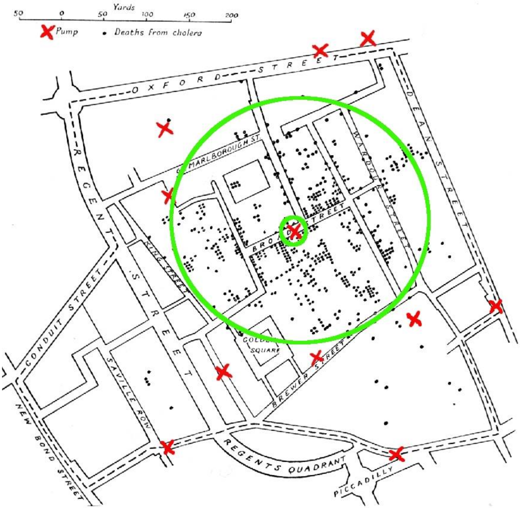
- Motivation for spatial statistics
- Learn some real-life examples
- Get first insights to the variety of modeling approaches

Cholera outbreak in London 1854 (Map by Dr. John Snow)

Motivation: A historical example

Cholera outbreak in London 1854 (Map by Dr. John Snow)

Motivation: A historical example

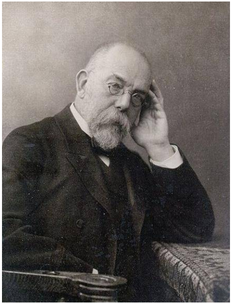


Cholera outbreak in London 1854 (Map by Dr. John Snow)

Motivation: A historical example

- Water was contamined by feces.
- One water pump was contamined with the cholera pathogen *Vibrio cholerae*.
- This explains the clustering of deaths from cholera.
- Is this a trivial example?
- No, at the beginning of the 19th century the „miasma theory of diseases“ was still well established.

John Snow (1813-1858)



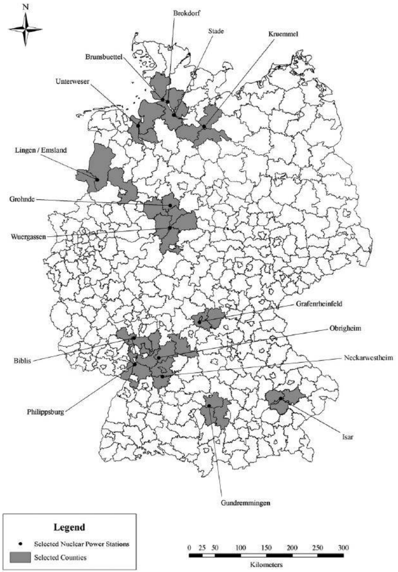
Robert Koch (1843-1910)

Talk by Robert Koch at the 10th Medical Congress in Berlin 1890:
„Koch's postulates"

Motivation for spatial statistics

- When the pathomechanism of a disease is unclear careful evaluation is crucial.
- Spatial correlation can provide important evidence.
- There are plenty of diseases where the pathomechanism is unknown or unclear.

Childhood cancer and nuclear power plants



Spix C et al. European Journal of Cancer 2008; 44: 275-284

Childhood cancer and nuclear power plants

Distance from nearest nuclear power plant (km)	Cases		Controls	
	N	%	N	%
<5	77	4.8	148	3.1
5–<10	158	9.9	464	9.8
10–<20	523	32.9	1589	33.6
20–<30	403	25.3	1181	24.9
30–<40	225	14.1	726	15.3
40–<50	137	8.6	371	7.8
>=50	69	4.3	256	5.4

Spix C et al. Case-control study on childhood cancer in the vicinity of nuclear power plants in Germany 1980-2003. European Journal of Cancer 2008; 44: 275-284

Methods applied in the paper

- Matched case-control study
- Conditional logistic regression
- Independent variable: $\frac{1}{\text{Distance from nearest power plant in km}}$
- Further covariates: None

Spix C et al. Case-control study on childhood cancer in the vicinity of nuclear power plants in Germany 1980-2003. European Journal of Cancer 2008; 44: 275-284

Subgroup	Coef	Lower 95% CL
All malignancies 1980–2003	1.18	0.46
Diagnostic groups 1980–2003		
Leukaemia	1.75	0.65
Central nervous system tumours	-1.02	-3.40
Embryonal tumours	0.52	-0.84
All malignancies except leukaemia	0.76	-0.20
First half of operation period	1.89	0.85
Second half of operation period	0.54	-0.47

Spix C et al. Case-control study on childhood cancer in the vicinity of nuclear power plants in Germany 1980-2003. European Journal of Cancer 2008; 44: 275-284

Open questions/limitations

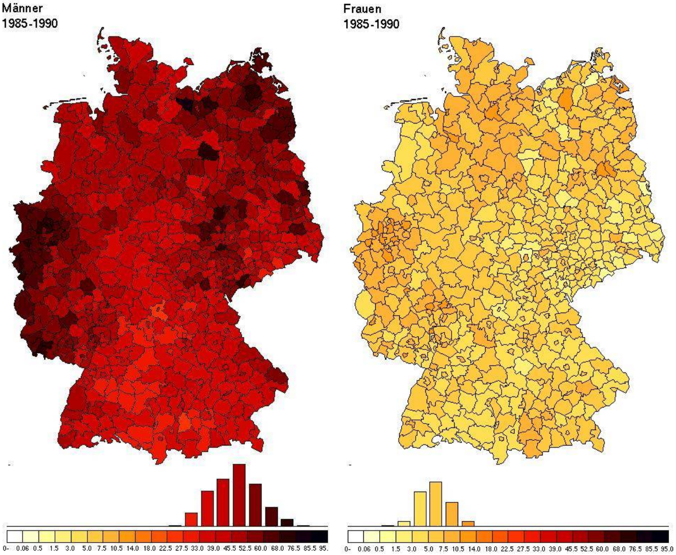
- Biologically plausible?
- Confirmed by other studies?
- Confounding?
- Attributable Risk: 0.2%

Spix C et al. Case-control study on childhood cancer in the vicinity of nuclear power plants in Germany 1980-2003. European Journal of Cancer 2008; 44: 275-284

Lung cancer mortality in Germany

- How are maps of cancer mortality generated?
- Age standardization is crucial.
- Stratification by sex seems useful.

Lung cancer mortality in Germany



http://www.dkfz.eu/de/krebsatlas/organe/162_map.html

Open questions/limitations

- How complete are cancer registries?
- Is completeness comparable between regions?
- On how many cases are the most extreme rates based?
- What are the underlying mechanisms that cause different lung cancer mortalities:
  - Different social status and lifestyle?
  - Exposure to cancer pathogens?
  - ...

Morbus Hodgkin and deprivation

Methods

- Cases from cancer registry
- Analysis on community level
- Poisson regression, outcome: cases per community
- Offset: *Log*(expected cases)
- Independent variable, e.g. Townsend deprivation score (mean on community level)
- ecological study!!

McNally RJQ et al. Geographical and ecological analyses of childhood acute leukaemias and lymphomas in north-west England. British Journal of Haematology 2003; 123, 60-65

Morbus Hodgkin and deprivation

Risk of Morbus Hodgkin by Townsend deprivation index

Quintile	RR	95% CL
1	1	–
2	5.02	(0.59–43.00)
3	3.02	(0.31–29.04)
4	4.09	(0.46–36.58)
5	13.08	(1.71–100.02)
Test for linear trend	$P = 0.001$	

McNally RJQ et al. Geographical and ecological analyses of childhood acute leukaemias and lymphomas in north-west England. British Journal of Haematology 2003; 123, 60-65

Open questions/limitations

- Biologically plausible?
- Ecological fallacy?
- Confounding?
- ...

McNally RJQ et al. Geographical and ecological analyses of childhood acute leukaemias and lymphomas in north-west England. British Journal of Haematology 2003; 123, 60-65

Parkinson Cluster

3 Parkinson „cluster“ in Canada

- 4 Parkinson cases among a TV crew of 125 people
- 4 Parkinson cases who were teching over a longer period in a mobile classroom of a college (out of 30 teachers).
- 3 Parkinson cases, among a group of 7 employees in a garment factory.

Kumar A et al. Clustering of Parkinson Disease: Shared Cause or Coincidence? Archives of Neurology 2004; 61: 1057-1060

- Calculation of the probability of Parkinson for each individual in the cluster based on the incidence (Probability of disease: p).
- Binomial probability mass function:

$$
\begin{aligned}
P(4 \text{ x Parkinson out of } 125|p) &= \binom{n}{k} p^k (1-p)^{n-k} \\
&= \binom{125}{4} \cdot p^4 \cdot (1-p)^{125-4}
\end{aligned}
$$

Kumar A et al. Clustering of Parkinson Disease: Shared Cause or Coincidence? Archives of Neurology 2004; 61: 1057-1060

Results

Results for the 3 clusters:

- $P = 7.9 \cdot 10^{-}7$
- $P = 2.6 \cdot 10^{-}7$
- $P = 1.5 \cdot 10^{-}7$

Kumar A et al. Clustering of Parkinson Disease: Shared Cause or Coincidence? Archives of Neurology 2004; 61: 1057-1060

Open questions/limitations

- This is multiple testing!
- The clusters were chosen retrospectively.
- Clustering may be expected.
- If one searches long enough one may find clusters of any disease in certain groups.

Childhood leukemia: Point processes

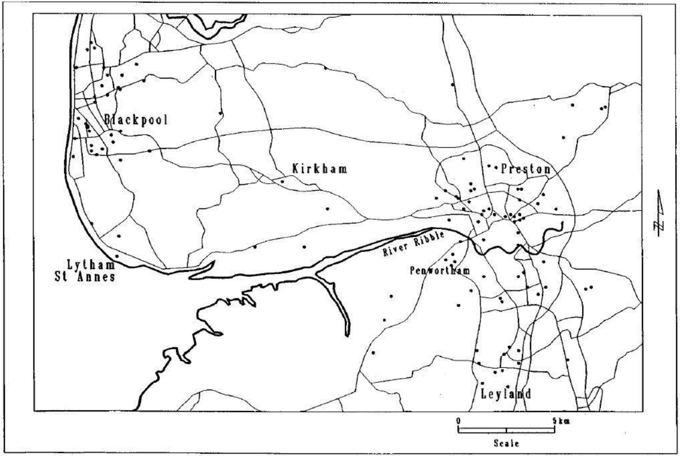


Figure 3 Locations of cases of childhood leukaemia in west-central Lancashire, 1954–92

Gatrell AC et al. Spatial point pattern analysis and its application in geographical epidemiology. Trans Inst Br Geogr 1996; 21: 256-274

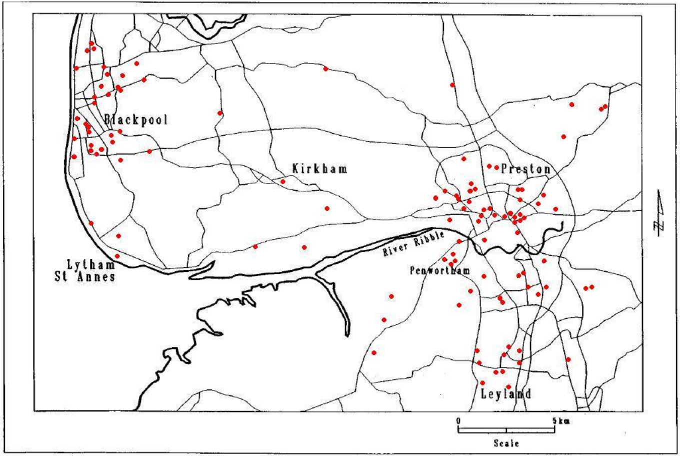


Figure 3 Locations of cases of childhood leukaemia in west-central Lancashire, 1954–92

Gatrell AC et al. Spatial point pattern analysis and its application in geographical epidemiology. Trans Inst Br Geogr 1996; 21: 256-274

Methods

- K function: The average number of events within a certain distance of a randomly chosen event divided by the average number of events per unit area.
- Calculate K function for cases.
- Calculate K function for controls.
- Difference between K functions points to clustering.

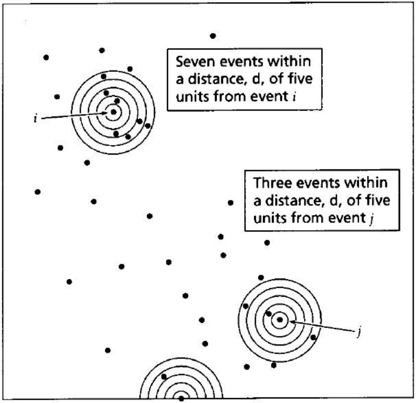


Figure 2 Estimation of a K function

Gatrell AC et al. Spatial point pattern analysis and its application in geographical epidemiology. Trans Inst Br Geogr 1996; 21: 256-274

Childhood leukemia: Point processes

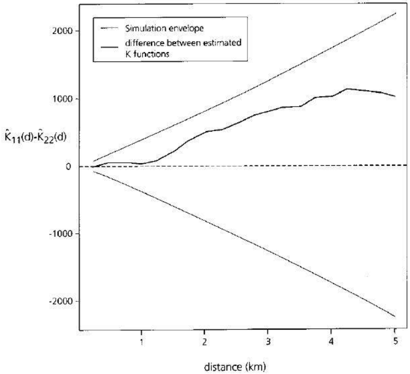


Figure 4 Difference between K functions (bold line) and simulation envelope (lighter lines) for childhood leukaemia and 'population at risk'

Gatrell AC et al. Spatial point pattern analysis and its application in geographical epidemiology. Trans Inst Br Geogr 1996; 21: 256-274

Conclusions

- Indication of clustering.
- However, no significant deviation from spatial randomness.
- Statistical Power?

Leukemia in Upstate New York

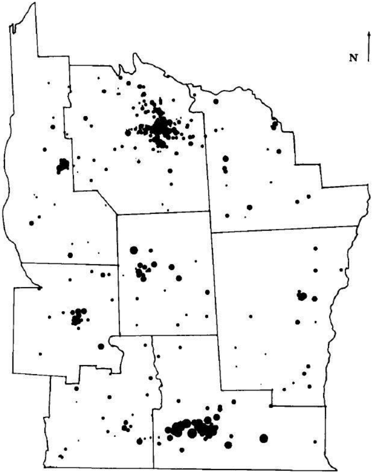


Figure 1. The 592 cases of leukaemia in Upstate New York

Kulldorff M et al. Spatial Disease Clusters: Destection and inference. Statistics in Medicine 1995; 14: 799-810

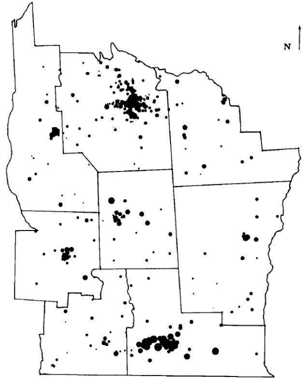


Figure 1. The 592 cases of leukaemia in Upstate New York

Leukemia cases

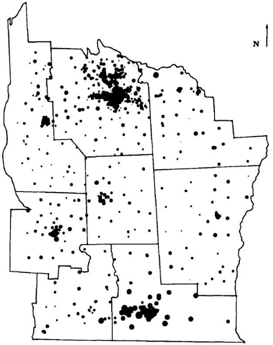


Figure 2. The population density in Upstate New York. The high density area in the north is Syracuse, and in the south Binghamton

Population

Kulldorff M et al. Spatial Disease Clusters: Destection and inference. Statistics in Medicine 1995; 14: 799-810

Methods

- Likelihood ratio test based on defined zones
- p is the probability of being a case in a zone
- q is the probabiltiy of being a case outside this zone
- H_0: $p = q$
- H_1: $p > q$

Leukemia in Upstate New York

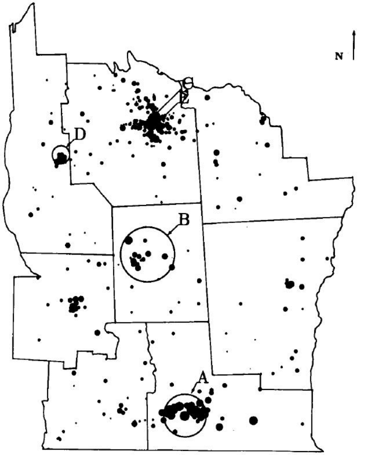


Figure 3. The most likely cluster 'A' and four other non-overlapping clusters on a map

Kulldorff M et al. Spatial Disease Clusters: Destection and inference. Statistics in Medicine 1995; 14: 799-810

Table I. The most likely cluster A and four other non-overlapping clusters. The incidence rate for the population as a whole is 0·56

Zone z	Number of cases c_z	Population n_z	Incidence rate per 1000	Relative likelihood $L(z)/L_0$	Radius in km	Rank	County
A	95·3	99608	0·96	472976	6·3	5	Broome
B	43·2	36629	1·18	21088	10·2	27	Cortland
C	55·2	56806	0·97	1911	2·9	174	Onondaga
D	26·4	23682	1·11	187	2·8	781	Cayuga
E	3·4	793	4·29	51	0	996	Onondaga

Kulldorff M et al. Spatial Disease Clusters: Destection and inference. Statistics in Medicine 1995; 14: 799-810

Lip cancer in Scotland

Tabelle: The Scottish lip cancer data.

County	Obs cases y_i	Exp cases E_i	Perc. in agric. x_i	Adjacent counties
1	9	1.4	16	5,9,11,19
2	39	8.7	16	7,10
...	...	...	...	...
56	0	1.8	10	18,24,30,33,45,55

Clayton DG et al. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 1987; 43: 671-681

Estimation of SMR?

- via maximum likelihood: $SMR_i = \frac{y_i}{E_i}$
- via Bayesian inference?
- What a priori information do we have?

Lip cancer in Scotland

Which a priori assumptions are plausible?

- General similarity of counties?
- Similarity of adjacent counties?
- Combination adjacent and general similarity?

Tabelle: Results for the area-specific relative risks from 4 different methods for the Scotish lip cancer data

Area	ML	exchangeable model	CAR prior	Convolution prior
1	6.43	4.67	4.72	4.81
2	4.48	4.20	4.47	4.44
...	...	...	...	...
56	0	0.65	0.87	0.83

Lip cancer in Scotland

(a) ML estimation

(b) exchangeable prior on log relative risk

Lessons to learn

- There are plenty diseases with unknown pathomechanism.
- Spatial correlation can provide important insights.
- There exists a broad variety of methods.
- There are also plenty of possibilities for wrong interpretations.