# Spatial Statistics for Epidemiologists

## Part 1: Exploring Spatial Data

Leonhard Held and Birgit Schrödle

27.-29. October 2010



## Spatial autocorrelation

### Spatial autocorrelation

Spatial autocorrelation is an important general concept though rarely directly applicable in epidemiological applications!

- Correlation of a variable with itself at close values
- Generalization of autocorrelation in time
- Typically only positive spatial autocorrelation: $\rho \in [0, 1]$
- Random patterns exhibit no spatial autocorrelation: $\rho = 0$



## Moran's $I$

Compare first-order autocorrelation in time:

$$\rho = \frac{\sum_{i=2}^{n}(y_i - \bar{y})(y_{i-1} - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

### Moran's $I$

$$I = \frac{n}{w_{++}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

with weights $w_{ij}$, $w_{++} = \sum_i \sum_j w_{ij}$



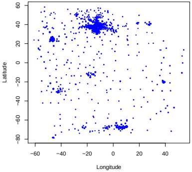
## Choice of weights

Different weights $w_{ij}$ can be chosen, for example:

- adjacency-based weights, with weight 1 if two regions $i$ and $j$ are adjacent
- k nearest neighbours weights, with weight 1 if location $j$ is one of $k$ nearest neighbours of location $i$
- distance-based weights, with weight 1 if location $j$ is within distance $d$ of location $j$

Which weights are chosen, can depend on type of the problem

## Example: Age distribution in Upstate New York




Each point represents the centroid of a census region in the counties of interest.



## Example: Age distribution in Upstate New York

- Variable of interest: Average age in 785 census regions
- Question: Is the average age spatially correlated?
- For each census region the centroid is given
- Calculate Moran's I, using adjacency-based weights obtained through triangulation

$$w_{ij} = \begin{cases} 1 & \text{if regions } i \text{ and } j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}$$

```
> tessel <- tri2nb(coords)
> weights <- nb2listw(tessel, style = "B")
> m <- moran(age, weights, length(age), Szero(weights))
> m$I
[1] 0.3894743
```



## Using k nearest neighbours

- Or use $k$ nearest neighbours
- For $k = 10$:

```
> n10 <- knearneigh(coords, k = 10)
> w10 <- nb2listw(knn2nb(n10))
> m10 <- moran(age, w10, length(age), Szero(w10))
> m10$I
[1] 0.3716742
```

- For $k = 20$:

```
> m20$I
[1] 0.3175319
```



## Testing Moran's $I$

- Significance testing of spatial correlation can be achieved by
  - assuming that the $y_i$'s are normally distributed
  - through a permutation test
- Result of a permutation test with 9999 permutations (adjacency-based weights):

```
> moranmc <- moran.mc(age, weights, nsim = 9999)
> moranmc$statistic
statistic
0.3894743
> moranmc$p.value
[1] 1e-04
```

## What is a permutation test?

- Permutation tests compare the observed data (suitably summarized in a statistic $T$) with randomly permuted data, which can be viewed as samples under the null hypothesis $H_0$.
- For example, in a statistical comparison of two groups, one may randomly permute group allocation.



## Example: Two-sample comparison

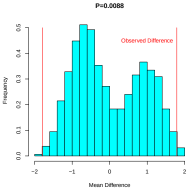
Are there differences in cholesterol levels between consumers of meat and fish?
Cholesterol levels:

| Group | Data | Mean |
|---|---|---|
| Fish | 5.42, 5.86, 6.16, 6.55, 6.8, 7.0, 7.11 | 6.4 |
| Meat | 6.51, 7.56, 7.61, 7.84, 11.50 | 8.2 |

Application of a standard $t$-Test gives a $P$ value of 0.041
Consider now all $m = 792$ permutations of group allocation under fixed sample sizes $n_1 = 7$ and $n_2 = 5$. Use difference of the means as test statistic.
Idea: The observed difference 1.79 should be unremarkable in the set of 792 possible group differences.
$\rightarrow$ Calculation of a (two-sided) $P$ value.



## $P$ value of permutation test




Computation of all permutations may be impossible for larger dataset. Solution: Draw a random sample from all populations, typically of size $m = 999$ or $m = 9999$.



## Geary's $C$

**Geary's $C$**

$$C = \frac{n-1}{2w_{++}} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- An alternative measure of spatial correlation with values between 0 and 2
- Values smaller than 1 indicate positive autocorrelation
- Test: Is $C < 1$?

## Test Geary's $C$

- Test: Is $C < 1$?
- As for Moran's I: Test under assumption of normality or permutation test
- Result of a permutation test with 9999 permutations (adjacency-based weights):

```
> gearymc <- geary.mc(age, weights, nsim = 9999)
> gearymc$statistic

statistic
0.6410651

> gearymc$p.value

[1] 1e-04
```



## Application to disease rates

- Areal disease data is typically of the form $(y_i, e_i)$ or $(y_i, n_i)$, $i = 1, \ldots, n$
- The standardized mortality rate (SMR) $y_i/e_i$ or $y_i/n_i$ is then plotted.
- However, the application of Moran's or Geary's test is invalid:
  - Even in the case of constant risk, the underlying population counts will vary, and the SMR's will not have constant variance
  - Permutations of the rates in the $n$ areas are not equally probable when the populations are different.
  - For rare diseases, the disease rates cannot be viewed as approximately continuous.



## Application to disease rates

- A better approach is to test each rate for departure from the overall rate.
- Under the null hypothesis, the associated p-values are iid uniform.
- Spatial correlation of the p-values or the associated binary test decisions (based on thresholding the p-values using a nominal error rate $\alpha$, e.g. 5%) can then be investigated.



## Example: Disease rates in Upstate New York

- Among the 780 observations, 56 are significantly greater than the overall rate (7.2%) at a 95% significance level.
- Moran's $I$ statistic is 0.029 with $p$-value 0.077 based on 9,999 permutations.

$\rightarrow$ Some evidence for heterogeneous rates

$\rightarrow$ Weak evidence for autocorrelation

## Discussion

- There a different notions of spatial autocorrelation
- Interpretation is difficult since it depends on the definition of the weights
- To investigate spatial autocorrelation of disease rates, it is useful to use binary significance test results rather than the rates themselves
- This ensures that application of a permutation test is valid



## Intensity function

- Let $N(\mathbf{A})$ be the number of events in a region $\mathbf{A}$.
- The intensity function of a spatial point process is

$$\lambda(\mathbf{s}) = \lim_{|d\mathbf{s}| \to 0} \left\{ \frac{E\left[N(d\mathbf{s})\right]}{|d\mathbf{s}|} \right\}$$

- If the process is stationary then the intensity $\lambda = \lambda(\mathbf{s})$ is the expected number of events per unit area.
- Otherwise it is to be understood locally at $\mathbf{s}$.



## Spatial kernel smoothing

- Estimate intensity function via

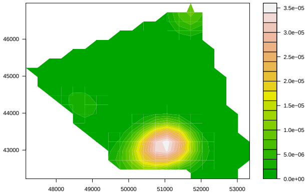
$$\hat{\lambda}(\mathbf{s}) = \frac{1}{h^2} W\left(\frac{\mathbf{s} - \mathbf{s}_i}{h}\right)$$

  where
  - $W(.)$ is a circular symmetric bivariate density function
  - $h > 0$ is a scalar which determines the amount of smoothing
- $h$ can be estimated through cross-validation



## Example: Smoothing Childhood leukaemia cases

```
> kcase <- kernel2d(loccase, polygon, h0 = 600)
> filled.contour(kcase, color = terrain.colors)
```



# Spatial clustering

**Spatial clustering**

A general tendency for cases to occur more closely together than would be compatible with random sampling from the population density

- → For point-referenced data, information on population density is required
- → For areal data, the number $n_i$ of persons at risk or the number $e_i$ of expected cases under constant risk is typically provided
- Only the spatial case-control design does not require this information!



# Tests of clustering

→ Besag & Newell (1991)

- General tests: Tests of complete spatial randomness (CSR): Interest in overall point pattern
- Focused tests: Some factor (e.g. a power plant) at a particular location has been previously hypothesized to be associated with the disease
- Tests for the detection of clusters: routine surveillance → identify "hot spots" without any preconceptions about their locations



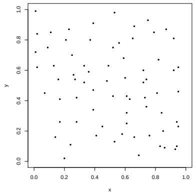
# Monte Carlo tests

- Tests for CSR are often Monte Carlo tests, since the distribution of the test statistic considered (under $H_0$) is not known
- Focus on value of a particular statistic $U$, for example smallest distance $D$ between all pairs of points
- What is the distribution of $D$ under CSR?
  - → Simulate $m$ CSR patterns, calculate distribution of $D$
- The Monte-Carlo $p$-value is defined as $k/(m+1)$ where $k$ is the rank of the actually observed value of the statistic
- Typically $m = 99, 999$ or $9999$.



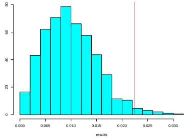
# Example: Swedish pines

The data give the locations of pine saplings in a Swedish forest.



## Example continued

Consider smallest distance between all pairs of points.
Simulate 999 point patterns under $H_0$ (no clustering)

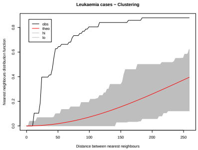



Monte-Carlo $p$-value= 0.02, some evidence for inhibition



## Testing CSR

- More advanced choices for test statistic $U$: Distribution of the nearest neighbour distance, inter-event distance, quadrat counts,...
- Example: Distribution of nearest neighbour distance of a observed point pattern (black) and under assumption of spatial randomness (red) with confidence bands (blue)






## Tests for focused clustering

- Often based on area-level data $(y_i, e_i)$, where
  - $y_i$ are the number of disease cases
  - $e_i$ are the number of expected cases

  in area $i$.
- Based on distance $d_i$ from hypothesized source $S$ of hazard
- Stone's test considers the test statistic

$$T = \max_{j=1,\dots,m} \frac{\sum_{i=1}^{j} y_i}{\sum_{i=1}^{j} e_i}$$

  where the data are ordered by distance $d_i$, so the test statistic is the maximal cumulative relative risk as distance from $S$ increases.
- An alternative is a score test, which may have higher power than Stone's test for certain alternatives



## Example: Leukemia in Upstate New York

- Hypothesis: Cases cluster around waste sites
- Example: Significant cluster around the Monarch chemicals waste site?

```
> ny <- read.table("data/newyork/ch1/CH1A.DAT")
> monarch <- c(-0.14, -67.19)
> locations <- rbind(cbind(ny[, 2], ny[, 3]), monarch)
> z <- knearneigh(locations, k = 1)
> index <- z$nn[[nrow(locations)]]
> stmontest <- stone.test(Observed ~ offset(log(Expected)), nydata,
+     model = "poisson", R = 999, region = index)
> stmontest

Stone's Test for raised incidence around locations

        Type of boots.: parametric
        Model used when sampling: Poisson
        Number of simulations: 999
        Statistic:  2.545434
        p-value :  0.016
```

## Detection of clusters: The Umhausen lung cancer hot spot

- In Umhausen, Austria, a small village in western Tirol, which has about 2600 inhabitants, an unexpectedly high rate of lung cancer has been observed (24 rather than 3.5 lung cancer cases per 1000 inhabitants in 10 years).
- This seems to be due to increased exposure to indoor radon.

  An 8700-year-old rock slide has a strong emanating power, because its rocks are heavily fractured and have a slightly increased uranium content. → 500 times higher Radon levels than those recommended by the Commission of the European Community
- However, the hot spot was not found through application of statistical techniques ...

NEJM, Volume 330:1684-1685, June 9, 1994, Number 23



## Detection of clusters

- Objective: Screening a large region for evidence of individual hot spots of disease without any preconception about their likely locations
- Two popular approaches:
  - Test by Besag and Newell (1991)
  - The scan statistic, see Kulldorff (1995)
- General problems:
  - Multiplicity
  - use of centroids for aggregated data
  - preference of circular clusters



## Example: Method of Besag and Newell

- Screening of study region for clusters including $k$ cases
- Size $k$ of clusters must be specified in advance
- No adjustment for multiple testing, but it can be tested if number of detected clusters is larger than expected

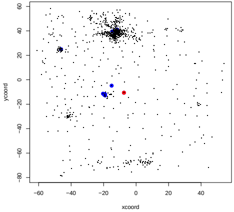
```
> bnny <- opgam(data = nydata, thegrid = nydata[, c(4, 5)], alpha = 0.05,
+     iscluster = bn.iscluster, k = 8, R = 999, model = "poisson",
+     mle = calculate.mle(nydata))
> bnny[, c(1, 2, 5)]

        x      y      pvalue
1  -46.17  25.00 0.019777091
2  -15.05  -4.82 0.040107378
3  -19.30 -12.45 0.008130742
4  -20.36 -11.41 0.045401531
5  -19.15 -11.69 0.045401531
6   -7.47 -10.59 0.001096719
7  -19.07 -12.82 0.001486029
8  -19.53 -12.46 0.008130742
9  -11.77  41.14 0.035240730
10 -14.69  39.50 0.040107378
```



## Continued: Plot of detected cluster centres

- Blue: 10 detected cluster centres
- Red: Cluster centre with smallest p-value



# Discussion

- Kernel smoothing is a technique to estimate intensity functions
- Tests of clustering can be of different form:
  - General tests: Tests of complete spatial randomness
  - Focused tests
  - Tests for the detection of clusters
- They all have limited epidemiological value as they do not quantify spatial risk and are not designed to adjust for confounders