Adrian Baddeley

Joining the dots
– inaugural lecture 7 November 2013
Joining the Dots

Statistics for spatial point patterns

Adrian Baddeley
John Snow’s map of cholera in Soho, London, 1854
Cholera in Soho 1854
Gold in Western Australia
Galaxy survey
Tree deaths in a groundwater catchment

+ tree death
• water bore
Common elements

- spatial locations of ‘events’/ ‘things’

  spatial point pattern

- additional data

  spatial covariates

- we want to investigate
  - dependence of points on covariate
  - dependence between points
What would a completely random pattern look like?
Completely random point pattern
A point process is a random mechanism that generates a random pattern of points.
Terminology: Point process

A *point process* is any random mechanism that generates a random pattern of points.
Completely random point process
Completely random point process
### Completely random point process

<table>
<thead>
<tr>
<th>1</th>
<th>1</th>
<th>2</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>3</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>
Poisson distribution (1837)

Bortkiewicz, *Das Gesetz der kleinen Zahlen* 1898
Poisson point process

The canonical model for a completely random pattern of points is the *Poisson point process*

Point locations are independent of each other; different areas of the pattern are independent of each other.
The **intensity** \( \lambda \) of a point process is the expected (mean) number of points per unit area.

Intensity could be a spatially-varying function \( \lambda(u) \) of location \( u \).
Gold in Western Australia
Problem 1: Intensity

Investigate whether the intensity $\lambda$ depends on the spatial covariate $Z$:

$$\lambda(u) = f(Z(u))$$
**Terminology: Interaction**

**Interpoint interaction** in a point process is stochastic dependence between the locations of the points.
Problem 2: Interaction

Detect and describe interaction between the points of a point process, after allowing for variation in intensity.
Gold in Western Australia
Problem 1: Intensity

Investigate whether the intensity $\lambda$ depends on the spatial covariate $Z$:

$$\lambda(u) = f(Z(u))$$
Regression

The graph illustrates a scatter plot with points plotted on the x-y axis. The x-axis ranges from 10 to 30, and the y-axis ranges from 44 to 56. The data points appear to follow a linear trend, indicating a potential correlation between the variables represented on the x and y axes.
Linear regression
The standard method for linear regression (‘least squares’) is appropriate when the errors around the line are Normally distributed.
Poisson distribution
Poisson regression
Poisson regression
Generalized Linear Models

The theory of “generalized linear models” embraces

- linear regression
- Poisson regression
- ...

This unification was achieved in 1980
Severe cyclones in Australia regression on El Niño
Scuba diving deaths in Australia
Poisson regression

3-year age groups
Pixel Logistic Regression

Proposed by statistician John Tukey 1972, developed by geologist Frits Agterberg

- Divide survey region into pixels
- Set pixel value to 1 if it contains data points, otherwise 0
- Analyse 0/1 pixel values using logistic regression
Logistic regression

\[
\log \frac{p}{1-p} = \beta_0 + \beta_1 x
\]
Gold deposits

Logistic regression analysis

Predicted probability of a gold deposit in each $2 \times 2$ km pixel

Logistic regression of $y = \text{presence/absence of gold}$ on $x = \text{distance to nearest fault}$
Claims about pixel logistic regression

- “logistic regression is a nonparametric technique” (i.e. does not make assumptions about the relation between $y$ and $x$)
- results depend on choice of pixel size
- “difficult to interpret the fitted parameters”
- small pixels $\Rightarrow$ numerical problems
Those who ignore statistics are doomed to reinvent it.
— B. Efron
Effect of pixel size

- results using different pixel sizes are **incompatible**

- there is no point process in continuous space that is consistent with logistic regression on every pixel grid

Very small pixels

For very small pixel size, pixel logistic regression is equivalent to assuming a **Poisson point process** with loglinear intensity

\[ \lambda(u) = \exp(\beta_0 + \beta_1 X(u)) \]

where \( X(u) \) is the covariate value at spatial location \( u \).


Gold deposits

Loglinear Poisson point process model

Predicted intensity of gold deposits
(number of deposits per km$^2$)
Predictions about a Poisson point process

For a Poisson point process with given parameters, it is straightforward to predict

- expected number of deposits in a given area
- probability of exactly \( n \) deposits in a given area
- probability of finding at least one deposit within \( x \) km of a given place
- etc
Until 1990 it was widely believed that mainstream statistical methods ('maximum likelihood') were "infeasible" for spatial point patterns, if interaction is present.

Instead, new methods were developed:

- Markov chain Monte Carlo
- Composite likelihood
- Moment methods
20th Century statistical methodology for point patterns

Moment methods: Ripley’s $K$ function

[Diagrams showing regular, random, and clustered patterns with corresponding plots for each type]
20th Century statistical methodology for point patterns

Critique

- Clunky
  Inflexible, slow, temperamental

- Doesn’t answer real world questions
  e.g. “How confident are you that there is gold in this region?"

- Immature
  Doesn’t provide standard statistical tools e.g. confidence interval, goodness-of-fit, leverage, influence, residuals, partial residuals
Statistical tools
Statistical methodology
Linear regression
Linear regression
Linear regression: confidence interval
Linear regression: prediction interval
Nonparametric regression
Linear regression diagnostics: residuals

The graph shows a scatter plot of residuals against the variable $x$. The residuals are the differences between the observed values and the values predicted by the regression model. The plot appears to show a random distribution of points around the zero line, indicating that the residuals are approximately normally distributed and there is no apparent pattern, which is a good sign for the model's assumptions.
Linear regression diagnostics: leverage
Linear regression diagnostics: influence
Gold deposits

Loglinear Poisson point process model

Predicted intensity of gold deposits (number of deposits per km$^2$)

Assumes intensity is a loglinear function of distance to nearest geological fault.
Validating the model

Logistic regression assumes

\[ \lambda(u) = \exp(\beta_0 + \beta_1 X(u)) \]

What if the relationship is not log-linear?

\[ \lambda(u) = \rho(X(u)) \]

How do we assess the evidence for/against a loglinear relationship?
Diagnostics for point process models

Extend Tukey’s idea to diagnostics

1. write down a diagnostic for logistic regression (rescale appropriately for pixel size)
2. take very small pixels
3. interpret as a diagnostic for point processes

Using this bridge, existing diagnostic tools from mainstream statistical science can be carried over to spatial point processes
Diagnostics for point process models

1. residuals
2. leverage
3. influence
4. partial residual
5. nonparametric smooth
1. Residuals

In linear regression, if \( \hat{y}_i \) is the fitted mean for observation \( y_i \), the residuals are

\[
  r_i = y_i - \hat{y}_i
\]

The residuals should not show a systematic pattern. If they do, this suggests that the relationship between \( x \) and \( y \) has not been correctly modelled.
Gold deposits: smoothed Pearson residuals

Diagnostics for point process models

1. residuals
2. leverage
3. influence
4. partial residual
5. nonparametric smooth
2. Leverage

In linear regression of $y$ on $x$, 

- observed response: $y_i$
- fitted response: $\hat{y}_i$
- leverage

$$h_i = \frac{d\hat{y}_i}{dy_i}$$

measures how strongly the fitted value $\hat{y}_i$ depends on the observed value $y_i$.

Large values of leverage are associated with the observations which, because of their covariate value, have a potentially strong influence on the fitted model.
Leverage for Poisson point process

For a Poisson point process with loglinear intensity

$$\lambda_{\beta}(u) = \exp(\beta^T X(u))$$

the leverage function is

$$h(u) = \lambda(u) X(u) \mathcal{I}^{-1} X(u)^T$$

Gold deposits: leverage

Leverage for fit

new
Diagnostics Poisson point process models

1. residuals
2. leverage
3. influence
4. partial residual
5. nonparametric smooth
3. Influence

In a linear model (etc), the *influence* of the \( i \)th observation is

\[
s_i = \frac{2}{p} \log \frac{L(\hat{\theta})}{L(\hat{\theta}_{(-i)})}
\]

where \( L \) is the likelihood, \( \hat{\theta} \) is the estimate of the parameter \( \theta \) using all the data, \( \hat{\theta}_{(-i)} \) is the estimate using all the data except the \( i \)th observation, and \( p \) is the number of parameters.

Large values of influence are associated with the observations which, because of their atypical response *and* high leverage, actually had a strong effect on the fitted model.
Influence for loglinear Poisson model

For the loglinear Poisson point process, the influence of data point $s_i$ is

\[ m_i = \frac{1}{p} X(s_i) \mathcal{I}_\hat{\beta}^{-1} X(s_i)^\top. \]

Gold deposits: influence

Influence for fit

new
Gold deposits: influence

Influence for fit
Gold deposits: influence

Influence for fit

Large circle at left identifies an outlier or anomaly.
Diagnostics for point process models

1. residuals
2. leverage
3. influence
4. partial residual
5. nonparametric smooth
4. Partial residuals

In linear regression

\[ y = ax + b \]

the partial residual (aka component-plus-residual) is

\[ r_i = \hat{b} x_i + y_i - \hat{y}_i \]

\( \hat{\sigma}^2 \).

A smoothed plot of \( r_i \) against \( x_i \) gives an estimate of the true relationship between \( x \) and \( y \).
Partial residuals for spatial point process model

For loglinear Poisson point process, the partial residuals are the values of $X$ at the data points $s_i$ with weights $1/\lambda(s_i)$.

Gold deposits: partial residual plot

\[ h(\text{default}) \]
Diagnostics for point process models

1. residuals
2. leverage
3. influence
4. partial residual
5. nonparametric smooth
5. Nonparametric estimate of covariate effect

Suppose that, instead of the loglinear model, the point process intensity depends on covariate $X$ through

$$\lambda(u) = \rho(X(u))$$

where the function $\rho$ is to be estimated.

Estimate $\rho$ by a kernel smoothing technique

Gold deposits: smoothed effect estimate

\[ p(\text{fault}) \]

\begin{align*}
\text{distance to nearest fault (km)}
\end{align*}
Tree deaths in Perth’s groundwater catchment
Spatially varying death rate

Tree deaths per km$^2$
Density of live trees

Compiled from 300,000 tree locations (detected from aerial imagery)

Y.M. Chang
Spatially varying death risk

Deaths per thousand trees
Covariate data

Terrain elevation

Depth to water table
Effect of depth to water table

Nonparametric estimate

\[ \rho(z) \]
Effect of terrain elevation, groundwater recharge

Partial residuals

Coming Soon . . .
Copper deposits and lineaments, Queensland
Local Poisson loglinear model
Local composite likelihood

California Redwood saplings
Local composite likelihood

Local Gibbs point process model
Local composite likelihood

Local Neyman-Scott cluster process model
Point patterns on linear networks

Road accidents in Geelong 2010–2012

G. McSwiggan
Point patterns on linear networks

Chicago street crimes

- assault
- burglary
- car theft
- damage
- robbery
- theft
- trespass
Point patterns on linear networks

Spider webs on a brick wall
Point patterns on linear networks

Dendritic spines

Kosic Lab, UCSB
Cholera in Soho 1854
Cholera in Soho 1854
Cholera in Soho 1854
Cholera in Soho 1854