Applied Geoscience Data Analysis using Matlab

12 Nov 2013

(Lecture 21)

Spatial Statistics

Focus on three techniques:

- Scanline Analysis (test for clustering)
- Box Counting Methods (test for randomness or uniformness)
- Nearest Neighbor (test for clustering & randomness)

Types of spatial distributions:

<table>
<thead>
<tr>
<th>Uniform</th>
<th>Random</th>
<th>(clustered) Clumped</th>
</tr>
</thead>
</table>

Scanline Methods – Count occurrence along transects

Coefficient of variation (Cv) = std of spacing divided by the mean along the scanline

\[ C_v = \frac{\sigma}{\mu} \]

- \( C_v = 0 \) (Uniform)
- \( C_v = 1 \) (Random)
- \( C_v > 1 \) (Clumped (or clustered))

Variance < Mean  Variance = Mean  Variance > Mean
Box Counting Methods (uniform)

Consider some spatial point process... e.g. lightning strikes

Let's test the null hypothesis that the distribution of lightning strikes in an area is uniform.

10 x 10 km² area containing 100 strikes.
Divided into 25 boxes.

If the distribution is uniform, how many strikes are expected in each of the 25 sub areas?

10 x 10 km² area containing 100 historic strikes.
Divided into 25 sub-boxes = 4 strikes per box.

Nexp = 4 * ones(25,1);
Nexp = Nexp(:,);

How many are observed in each of the 25 sub boxes?

figure; hist3(data,[5 5]), view(30,70)
Nobs = hist3(data,[5 5]);
Nobs = Nobs(:,); % Produces a 25 x 1 vector of the number in each cell

Number of Strikes in Box

Boxes 1-25

Red = observed
Blue = expected if the distribution is uniform
Chi-Squared (an old friend)

\[
\text{chi2value} = \text{sum}((\text{Nobs} - \text{Nexp})^2 / \text{Nexp}) = 14
\]

Test at the 95% level .......

\[
\text{chi2\_crit} = \text{chi2inv}(0.95, 25-1-1) = 35.17
\]

Can we **not** reject the null hypothesis that the strikes are uniformly distributed?

---

Box Counting Methods (Poisson or Random)

Let's test the null hypothesis that the distribution of lightning strikes in an area is **Poisson (Random)**.

10 x 10 km² area containing 100 strikes.

Divided into 49 boxes.

---

Review of Poisson Model (lecture 7)

Recall from univariable stats...

\[
P(x) = \frac{(t\lambda)^x e^{-t\lambda}}{x!}, \quad x = 0,1,2,3,\cdots n\quad \text{where}
\]

\(P(x)\) is the probability of getting \(x\) number of events during a time window of duration \(t\). The event rate, \(\lambda\), is the number of events per unit time.

---

Review of Poisson Model (lecture 7)

**Old Example:** The rate of flooding on a river is \(\lambda = 0.044\) floods/year (~2.2 floods every 50 years)

What is the probability of having NO floods \((x=0)\) is the next 50 years?

\[
P(0) = e^{-0.044*50} = 0.11.
\]

The probability of having at least 1 flood is then: 1-P(0)=0.89.
### Review of Poisson Model (lecture 7)

**Matlab:**

```
>> x=0.10;
>> P=poisspdf(x,0.044*50);
>> plot(x,P)
```

- `poisspdf` is used to compute the probability of a given number of events in a fixed interval.

### Test if spatial distribution is random

- `hist3` is used to create a 3D histogram of the data.
- `counts` is used to extract the counts from the histogram.

```
counts = hist3(data,[7 7]);
```

- `bar` is used to create a bar chart of the counts.

```
x = 0 : 1 : 5; % number of strikes in a box
Nobser = hist(counts,x); % number areas with x strikes
figure; bar(x,Nobser)
```
What is the expected spatial Poisson distribution

\[ P(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, 3, \ldots n \]

\[ \text{Nexp}(x) = 49 \times P(x) \]

\text{Nexp} is the expected number of boxes with \( x \) number of strikes in it.

\( \lambda = 100 / 49 \)

\( x = 0 : 5; \) number of strikes in a box

\( \text{Nexp} = 49 \times \text{exp}(-100/49) \times (100/49)^x / \text{factorial}(x) \)

OR

\( \text{Nexp} = 49 \times \text{poisspdf}(x, 100/49) \)

Test with chi-squared

\( \gg \text{chi2} = \text{sum}((\text{Nobser} - \text{Nexp})^2 / \text{Nexp}) \)

\( \text{chi2} = 1.4 \)

\( \text{chi2inv}(0.95, 6 - 1 - 1) \)

\( \text{ans} = \)

\( 9.49 \)

Can we cannot reject the null hypothesis that the strikes are randomly distributed?

Histogram showing number of boxes with \( x \) strikes

There are a total of 6 bins here

Nearest Neighbor Approach to Evaluating Spatial Randomness and Clustering

- More common approach
- Avoids making subjective choices about binning
Nearest Neighbor (NN) Approach

\[ r = \text{distance between each point and its nearest neighbor} \]
\[ n = \text{number of points} \]

Calculate the average NN distance

\[ R_{\text{avg}} = \frac{\sum r}{n} \]

Test statistic \( c \)

\[ c = \frac{R_{\text{avg}} - R_{\text{exp}}}{\sigma_{R_{\text{exp}}}} \quad \text{Clark & Evans, 1954} \]

Where \( \sigma_{R_{\text{exp}}} \) is the standard error of the means of the NN distances in a randomly distributed population

For 95\% confidence, if \( |c| > 1.96 \) then we can reject the null hypothesis that the data are randomly distributed

Calculate the expected average NN distance for Poisson Distribution

\[ R_{\text{exp}} = \frac{1}{2\sqrt{\rho}} \quad \rho = \text{density of points} = \text{number/area} \]

Then calculate the ratio of the two

\[ R = \frac{R_{\text{avg}}}{R_{\text{exp}}} \]

- \( R > 1 \) (uniform) indicates objects are further apart than you might expect for a random model
- \( R \approx 1 \) (Poisson) indicates random distribution
- \( R < 1 \) (clustered) indicates objects are closer together than you might expect for a random model

Calculating nearest neighbor distance in Matlab:

\[ k \text{nnsearch finds K nearest neighbors} \]
\[ [\_ , D] = \text{knnsearch}(X,Y,'distance','dist_method','k',' #near_neighb) \]
returns a vector \( D \) containing the distances between each row of \( Y \) and its \( k \) closest points in \( X \). If we use \( X, X \), then we can calculate the distance from every point, to every other.

\[ X(1:3,:) \% \text{positions of data in space} \]
\[ \text{ans} = \]
\[ 2.1896 \quad 5.7165 \]
\[ 0.4704 \quad 8.0241 \]
\[ 6.7864 \quad 0.3305 \]

\[ [\_ , \text{dist}] = \text{knnsearch}(X,X,'distance','euclidean','k',\text{length}(X)); \]
\[ \text{figure; imagesc(dist); colorbar} \]
Calculate average NN distance
Ravg=mean(min(dist(:,2:end),[],2)); % skip the first column, since it’s zero; find minimum across each row, take the average = 0.5171

Calculate expected average NN distance
n=100; area=100; % 100 object in 100 square units area
den=length(data)/area; % density
Rexp=1/(2*sqrt(den)); % R expected = 0.5
R = \frac{R_{avg}}{R_{exp}} = 1.0342
✓ R approximately = 1 (Poisson) indicates random distribution

\[ \sigma_{R_{exp}} = \frac{0.2613}{\sqrt{n \rho}} \]
\[ c = \frac{R_{avg} - R_{exp}}{\sigma_{R_{exp}}} \]
S=0.2613/sqrt(n*den); % define the standard error =0.0261
c=(Ravg-Rexp)/S; % calculate test criteria =0.6547

For 95% confidence, if |c| > 1.96 then we can reject the null hypothesis that the data are randomly distributed

So... in this case we cannot reject the null hypothesis.

Note that you can call other distance functions. So if the data in X are [lat, lon] you can use Matlab’s distance function.

[idx, dist] = knnsearch(X,X,'distance',@distance,'k',length(X));

The key paper on nearest neighbor methods

Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations
Authors: Philip J. Clark and Francis C. Evans
Reviewed work(s)
Published by: Ecological Society of America
Stable URL: http://www.jstor.org/stable/1931034