

Point pattern analysis with Past

Øyvind Hammer, Natural History Museum, University of Oslo,
2011-01-03

Introduction

A point pattern is a distribution of points in space. We will focus on two-dimensional patterns. We will not discuss so-called *marked* point patterns, where the points have associated values such as elevation, or belong to categories. Examples of point patterns are positions of trees in a forest, earthquake epicenters, volcanic pipes, springs, submarine pockmarks, vents and seeps, fossils on a bedding plane, mineral deposits, disease outbreaks and crime occurrences. Typical questions include:

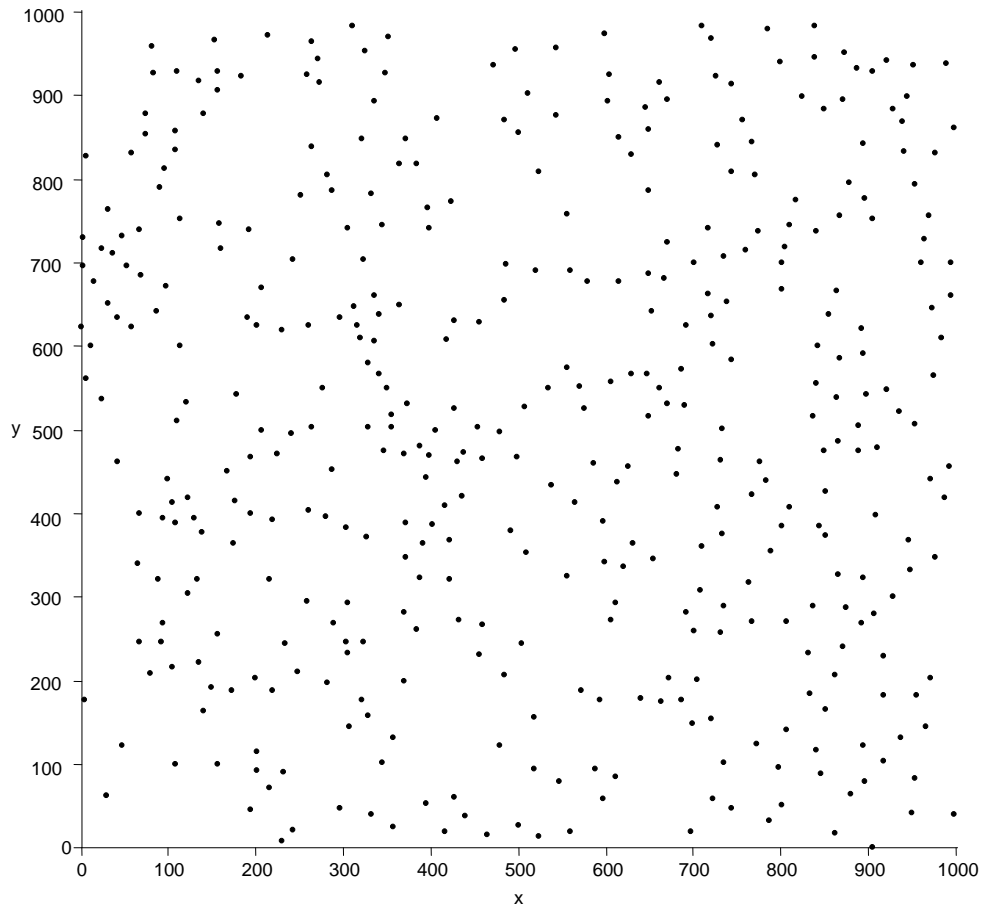
- How does the point density vary across the domain?
- Are the points completely independently distributed, or are they clustered or overdispersed?
- Do the points tend to occur along lines (e.g. earthquakes on a fault zone, springs or seeps along lineaments) ?

Typically, x-y coordinates of points are measured by clicking with a mouse on a digital image. This can be done in e.g. Photoshop or with special software such as TPSdig.

Nearest neighbour analysis

The simplest null hypothesis concerning a point pattern is that the points are distributed uniformly, randomly and independently, i.e. without regard for other points. A process generating such a pattern is called a *Poisson process*.

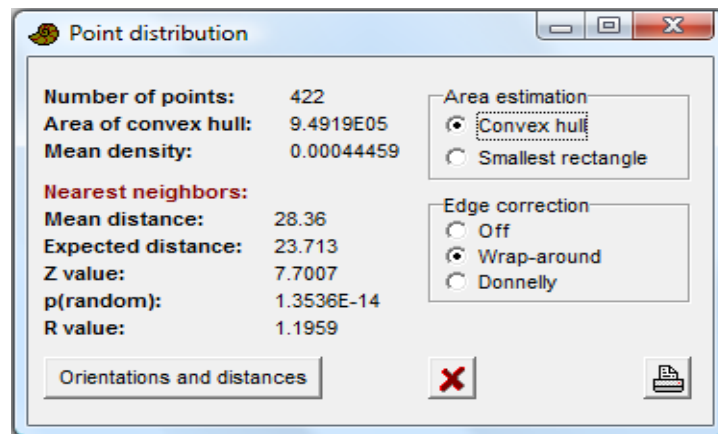
We will use a dataset containing the positions in meters (x and y) of 422 pockmarks in a 1 km² area on the Norwegian continental margin (this is highly reduced from the original data set, containing 4743 pockmarks within 12 km²). Open the file `pockmarkpos_red.dat`, select the two columns and run "XY graph" from the Plot menu.



To test the null hypothesis of a Poisson pattern, we can use the average nearest neighbour distance as our test statistic. For a Poisson pattern, the distribution of nearest neighbour distances is negative exponential, with mean

$$\mu = \frac{\sqrt{A/n}}{2}.$$

Run “Nearest neighbour” in the Geomet menu.

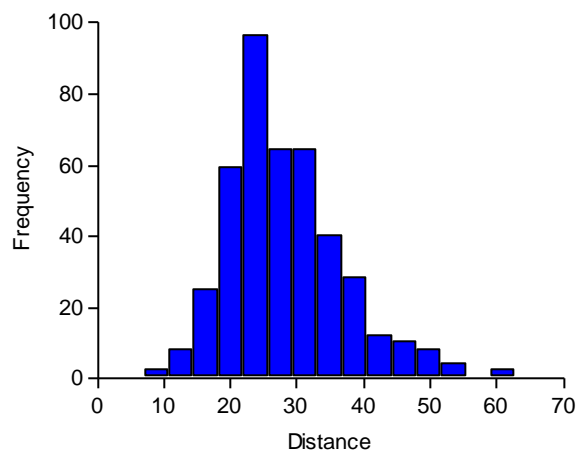


The probability that the distribution is random (Poisson process, giving an exponential nearest neighbour distribution) is presented, together with the R value:

$$R = \frac{\bar{d}}{\mu} = \frac{2\bar{d}}{\sqrt{A/n}}$$

where \bar{d} is the observed mean distance between nearest neighbours, A is the area, and n is the number of points. Clustered points give $R < 1$, Poisson patterns give $R \sim 1$, while overdispersed points give $R > 1$.

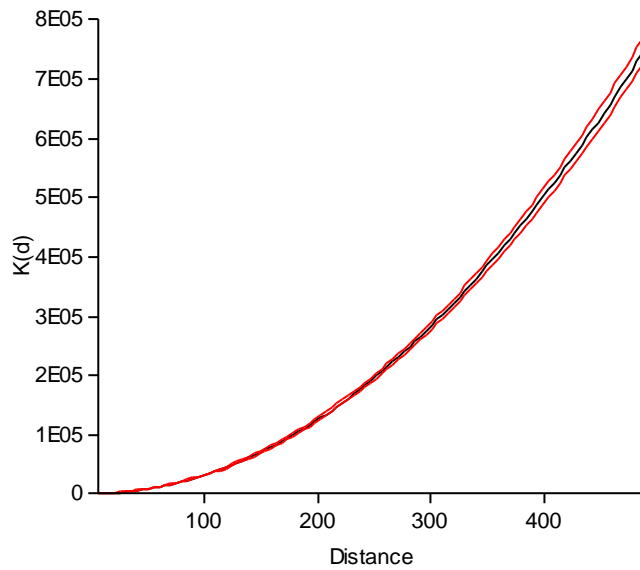
In this case, the observed mean distance is 28.36 m, while the expected distance from a Poisson pattern is 23.71 m. The difference is significant at $p < 0.001$. The points are clearly overdispersed. Clicking the “Orientations and distances” button brings up all the nearest neighbour azimuths and distances. Copy back to the spreadsheet (select “Edit labels” above the spreadsheet and paste into the topmost label row) and plot a histogram. Does it look like a negative exponential distribution?



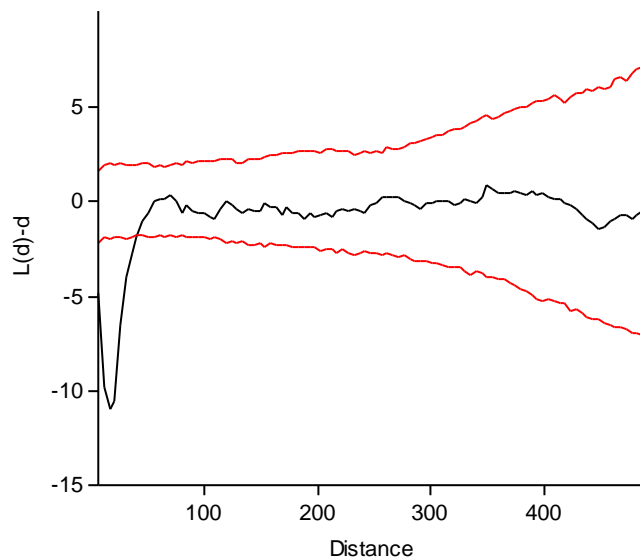
Ripley's K analysis

The problem with nearest neighbour analysis is that it does not discriminate between scales. It is quite common that points are clustered at small scales, but that these clusters are themselves overdispersed, or, conversely, that the points avoid each other locally but cluster at larger scales. In Ripley's K analysis, the points within increasing distances from one central point are counted. These counts are averaged over all central points. For a Poisson pattern, this results in a curve where the number of points increases smoothly as the square of distance. Departures from Poisson can be detected by a Monte Carlo procedure, simulating a large number of random point patterns to produce (pointwise) confidence envelopes.

With the same data set as before, run Ripley's K from the Geomet menu. The computation will take a couple of minutes. Select to plot the $K(d)$ function, and tick the “95% confidence” box.



The parabolic form is evident – the number of points increases with the square of distance. The red lines represent the 95% confidence interval from the Monte Carlo simulations. However, it is not easy to see any departures from the curve expected for a Poisson pattern. Try plotting the $L(d)$ function, which is a square-root transformation of $K(d)$ giving a straight line $L(d)=d$ under Poisson. Finally, the function $L(d)-d$ (should be zero under Poisson) gives perhaps the best picture:



The curve stays below the confidence band up to a distance of ca. 40 m, indicating local lateral inhibition. Above this scale, there seems to be little structure.

A few additional points about Ripley's K :

- The confidence band is computed *pointwise*, and does not give an exact representation of total significance. A future version of the program may attempt to correct for this, although numerical experiments indicate that the pointwise and overall confidence bands for Ripley's K do not differ enormously (Davison and Hinkley 1997, p. 417).
- It is possible to design Monte Carlo simulations using other null models than Poisson process, e.g. a "Strauss" process with lateral inhibition.
- The area may be incorrectly estimated by the smallest fitting rectangle, especially for small numbers of points. This can sometimes be spotted as giving rise to a long-range trend in $L(d)$ - d . PAST allows manual adjustment of the area to correct for this.
- If for some odd reason you are interested in fractal dimension, this can be estimated by the (asymptotic) regression slope in a log-log plot of $K(d)$.

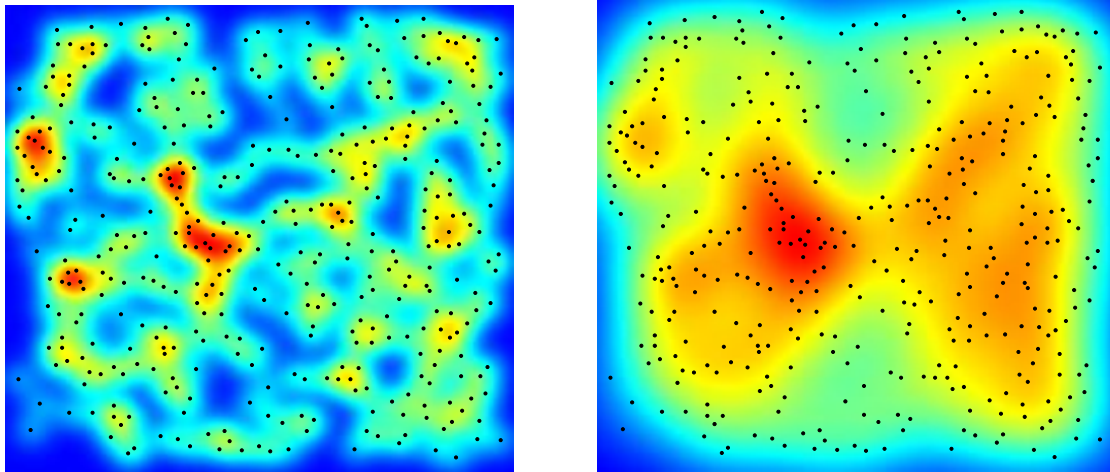
Kernel density estimation

Variations in point density across an area can often be seen by eye, but a coloured or contoured density map can be useful to visualize hot spots and density gradients. An obvious method would be to center circles of a certain, fixed size on each desired pixel in the map, and count the number of points within each circle. A smoother map with better statistical properties is obtained by downweighting distant points according to a *kernel function*. The bivariate Gaussian kernel is most commonly used:

$$f(x, y) = \frac{1}{\pi r^2} \sum_i e^{-\frac{1}{2r^2}[(x-x_i)^2 + (y-y_i)^2]}$$

The scale parameter r is critical. Although there are methods for selecting an "optimal" value, it is usually more convenient for the user to choose the scale of interest.

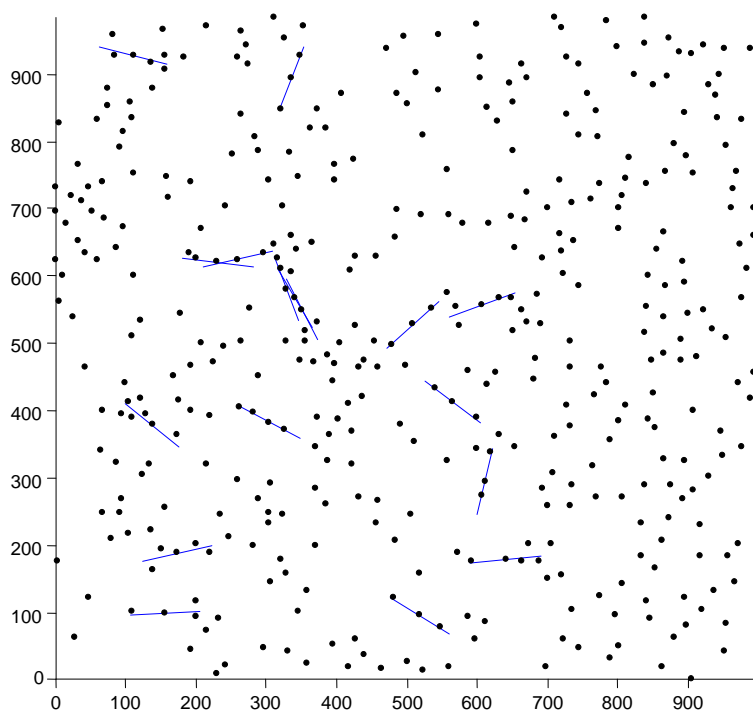
With the pockmark data set, run "Kernel density" from the Geomet menu. Experiment with different settings for grid size (number of rows and columns) and radius (scale parameter). As shown in the examples below, a small scale value ($r=30$ m) emphasizes high-frequency features in a reticulate pattern. At larger smoothing scales ($r=80$ m), a butterfly-shaped figure is produced. One of the diagonals in the pattern corresponds to a neotectonic fault. Note the apparently low density near the edges – this boundary effect is a well known artefact in kernel density estimates.



Point alignments

PAST can detect linear point alignments using the “continuous sector” method (Hammer 2009).

The *Radius* parameter sets the scale of analysis, in this case in meters (half the lineament length). *Alpha* sets the significance level for the Rayleigh test used on the procedure. Note that this is a pointwise significance, not corrected for the multiple testing of all the points. The *Dispersion filter* disables alignments with uneven distribution of points along the lineament. *View numbers* lists the alignment positions and their orientations, which can be subjected to circular statistics if required (Directions module).



For our example, with a radius of 50 m and $\alpha=0.1$, only 17 alignments are reported. This is *fewer* than expected for a random pattern. Although some pockmark strings may be present, they are not typical features in this area.

References

Davis, J.C. 1986. *Statistics and Data Analysis in Geology*. John Wiley & Sons.

Davison, A.C. and Hinkley, D.V. 1997. *Bootstrap methods and their applications*. Cambridge University Press.

Hammer, Ø. 2009. New methods for the statistical detection of point alignments. *Computers & Geosciences* 35:659-666.