Mari Myllymäki

January 23, 2012

**Data sets**

You may use these data sets for your studies or you can have your own data set.

Some references where the data has been considered are given here for the data sets and references may also be found from the specified source, for example.

- Spanish towns.

  The example origins from Glass and Tobler (1971) and has been further considered by Ripley (1977, 1988) and Stoyan et al. (1995). The data consist of the positions of 69 towns in a square window of side length 40 miles within the Spanish Plateau south-east of Madrid. The locations of towns relative to each others provide information for researchers in demography.

  Some references:
  Glass, L. and Tobler, W.R. (1971). Uniform distribution of objects in a homogeneous field: Cities on a plain. Nature 233 (5314), 67-68.

  Source:
  http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470014911.html,
  http://users.jyu.fi/ penttine/ppstatistics/

- Tharandter Wald.

  These data observed in a 56 m $\times$ 38 m rectangle come from a Norway spruce forest in Saxony (Germany), see Illian et al. (2008). The data consist of 134 spruce trees with diameter at breast height varying between 16 and 37 cm. The stand was originally planted and was later thinned by a forester with the aim of obtaining equal-sized trees.

  - How are the trees distributed?
  - Are the marks independently distributed?

  Some references:
  Illian et al. (2008). Statistical Analysis and Modelling of Spatial Point Patterns. Wiley, Chichester.

  Grabarnik, P., Myllymaki, M. and Stoyan, D. (2011). Correct testing of mark hypothesis for marked point patterns. To appear in Ecological Modelling.

  Source:
  R/spatstat, data(spruces)

- Redwood data.

The data represent the locations of 62 seedlings and saplings of California redwood trees in a square sampling region. They originate from Strauss (1975); the present data are a subset extracted by Ripley (1977) in a subregion that has been rescaled to a unit square.

- How are the trees distributed?
- How does the intensity vary over the region?
- What is the scale of clustering of trees?

Source:
R/spatstat, data(redwood)

- Redwood full data.

  These data represent the locations of 195 seedlings and saplings of California redwood trees in a square sampling region. They were described and analysed by Strauss (1975). This is the "*full*" dataset; most writers have analysed a subset extracted by Ripley (1977) which is available as 'redwood'.

  Strauss (1975) divided the sampling region into two subregions I and II demarcated by a diagonal line across the region. The spatial pattern appears to be slightly regular in region I and strongly clustered in region II.

  - What is the distribution of trees in the subregion I? In the subregion II? How are they different?
  (- Is the boundary correctly defined?)

  Source:
  R/spatstat, data(redwoodfull)

- Waterstriders.

  Cross-sectional data on positions of waterstriders (at last larval stage) at a specific time point on a water surface. Waterstriders are arthropods living on the surface of ponds. They are able to communicate by sending and receiving signals using the water surface as a medium. In addition, ecologists know from experiments that the later larval stages and male adults form their own territories and tend to prevent other males from entering these. Does the data support this? What kind of dependence is there between the waterstriders at last larval stage?

  Source:
  http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470014911.html, see pages 8, 138 in the book, http://users.jyu.fi/ penttine/ppstatistics/

- Frost shake of oaks.

  The data consist of 392 oak trees observed in a 100 m $\times$ 100 m square. For analysis purposes the sound oaks are marked by 1, whereas the oaks suffering from frost shake have mark 2.

  Frost shake is a split in a tree trunk, which is caused by interaction of frost and sun and lowers wood quality. Since not much is known on the processes inducing frost shake, it is biologically interesting to test whether frost shake occurred randomly within the stand.

  - Are the marks (sounds/frost shake) independently distributed?

Some references:
Goreaud, F., Pelissier, R., 1999. On explicit formulas of edge effect correction for Ripley's K-function. J. Veg. Science 10, 433-438.

Illian et al. (2008). Statistical Analysis and Modelling of Spatial Point Patterns. Wiley, Chichester.

Grabarnik, P., Myllymaki, M. and Stoyan, D. (2011). Correct testing of mark hypothesis for marked point patterns. To appear in Ecological Modelling.

Source:
http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470014911.html,
http://users.jyu.fi/ penttine/ppstatistics/

- Amacrine cells.

  Austin Hughes' data: a point pattern of displaced amacrine cells in the retina of a rabbit. 152 "on" cells and 142 "off" cells in a rectangular sampling frame.

  We are interested in studying the spatial dependency between the positions of on and off cells because we hope this will tell us something about how the two cell types emerge during development: do the two cell types emerge from a single undifferentiated population, or do they develop independently of each other? (Diggle et al. Modelling the bivariate spatial distribution of Amacrine cells)

  Is there a pronounced inhibitory effect? Can two on cells, or two off cells, be located arbitrarily close together? What about an on and an off cell?

  Some references:
  Diggle, P. J. (1986). Displaced amacrine cells in the retina of a rabbit: analysis of a bivariate spatial point pattern. J. Neurosci. Meth. 18, 115–125.

  Source:
  R/spatstat, data(amacrine) P. Diggle's homepage:
  http://www.lancs.ac.uk/staff/diggle/pointpatternbook/Datasets/
  amacrines.explain_explain.txt

- Ants.

  These data give the spatial locations of nests of two species of ants, Messor wasmanni and Cataglyphis bicolor, recorded by Professor R.D. Harkness at a site in northern Greece, and described in Harkness & Isham (1983). The full dataset (supplied here) has an irregular polygonal boundary, while most analyses have been confined to two rectangular subsets of the pattern (also supplied here).

  The harvester ant M. wasmanni collects seeds for food and builds a nest composed mainly of seed husks. C. bicolor is a heat-tolerant desert foraging ant which eats dead insects and other arthropods. Interest focuses on whether there is evidence in the data for intra-species competition between Messor nests (i.e. competition for resources) and for preferential placement of Cataglyphis nests in the vicinity of Messor nests.

  Some references:
  Handbook of Spatial Statistics, p. 325–327.
  Møller and Waagepetersen (2006), Modern statistics for spatial point processes. p. 44.

  Source:
  R/spatstat, data(ants)

- Oak-beech tree data.

  Locations of oak-beech trees in a window 80 m $\times$ 80 m together with observed diameter at breast height in cm.

  - How are oaks distributed?
  - How are beeches distributed?
  - Are oaks and beeches distributed independently of each others or is there some dependence between them?

  Source:
  http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470014911.html,
  http://users.jyu.fi/ penttine/ppstatistics/

- Hamster kidney data.

  Point pattern of cell nuclei in hamster kidney, each nucleus classified as either 'dividing' or 'pyknotic'. A multitype point pattern.

  - Are the dying cells distributed completely spatially randomly? Or are they clustered?
  - Are the dividing cells distributed completely spatially randomly? Or are they clustered or regular?
  - Is there any relationship between the two types of cells?

  Some references:
  Diggle (2003), p. 100.

  Source:
  R/spatstat, data(hamster)

- Gold particles.

  Source:
  http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470014911.html, see p. 8,
  http://users.jyu.fi/ penttine/ppstatistics/

- Nerve fiber data from a foot/calf/forearm/tight of a healthy/diseased subject (several possibilities)

  Kennedy et al. (1999) report that nerve fiber loss due to neuropathy (for example diabetic neuropathy) results in a more clustered epidermal nerve fiber (ENF) pattern across the epidermis, the outermost layer of the skin, in subjects with small fiber neuropathies than in non-diseased subjects. Before studying differences of the spatial pattern of ENFs or ENF entry points between healthy subjects and subjects with small fiber neuropathy, it is important to understand the spatial structure of ENFs within healthy subjects.

  Waller et al. (2011) found out a suitable point process model for a sample from tight of a healthy subject of their study. Point patterns available from the lecturer of the course gives locations of ENF entry points taken from a healthy/diseased subject from foot/calf.

  Some research questions:
  1) Is the point pattern of ENF entry points of a subject similar in foot, calf, forearm and thight? Is the point pattern in calf, foot or forearm more clustered or regular than in thight?
  2) How are the ENF entry points distributed in calf/foot? How can you fit a model for

replicated point patterns from calf/foot of a subject?
3) Is the ENF entry point data from calf/foot of a healthy subject more clustered than from a diabetic subject?

Some references:
Olsbo (2008). Spatial Analysis and Modelling Motivated by Nerve Fiber Patterns. PhD thesis.

Waller, L. A, Särkkä, A., Olsbo, V., Myllymäki, M., Panoutsopoulou, I. G., Kennedy, W. R. and Wendelschafer-Crabb, G. (2011). Second-order spatial analysis of epidermal nerve fibers. *Statistics in Medicine* 30(23), 2827-2841.

Source:
Data available from the lecturer of the course by request.

- Choose some subset of the rainforest dataset, see also below.

  Source:
  R/spatstat, data(bei)
  http://ctfs.arnarb.harvard.edu/webatlas/datasets/bci/

- Lancing

  Locations and botanical classification of trees in Lansing Woods.

  One question of interest is whether the species are evenly mixed across the forest plot or are segregated into different subregions where one species is predominant over the others. (Handbook, p. 373)

  Some references:
  Handbook of spatial statistics, p. 373

  Source:
  R/spatstat, data(lansing)

**Data sets for the second round:**

- A forest inventory data set from Zambia. The data comes from Prof. Erkki Tomppo, Metla (Finnish Forest Research Institute). Ask about details from the lecturer.

  What shape, size and spatial layout should the sample plot have in order to obtain good estimates for the total volume of wood?

- One species, e.g. Beilschmiedia pendula, in a rainforest

  A point pattern giving the locations of 3605 trees in a tropical rain forest. Accompanied by covariate data giving the elevation (altitude) and slope of elevation in the study region.

  - Does the distribution of the trees depend on elevation and/or slope?

  Some other interesting species pointed out by biologists: Trichilia tuberculata (TRI2TU), Alseis blackiana (AlSEBL), Ocotoea whitei (OCOTWH).

  Source:
  R/spatstat, data(bei)
  http://ctfs.arnarb.harvard.edu/webatlas/datasets/bci/
  - Any other species than B. pendula on a fixed time (e.g. 1990), same covariates as for B. pendula.

Some references:
B. pendula used as an example in many articles. Others not so much.
http://www.r-inla.org/, http://code.google.com/p/inla/source/browse/ (methods)

- Rainforest data, two species

  - Is there any relationship between two particular species?

  Some interesting species pointed out by biologists: Trichilia tuberculata (TRI2TU), Alseis blackiana (AlSEBL), Ocotoea whitei (OCOTWH)

  Source:
  http://ctfs.arnarb.harvard.edu/webatlas/datasets/bci/

- Rainforest data

  - Any other interesting questions.

- Gorillas

  Does the spatial locations of nests of gorillas depend on various spatial covariates?

  Some references:
  Funwi-Gabga, N. and Mateu, J. (2012) Understanding the nesting spatial behaviour of gorillas in the Kagwene Sanctuary, Cameroon. Stochastic Environmental Research and Risk Assessment, in press.

  Source:
  R/spatstat, data(gorillas)

- Lancing

  Locations and botanical classification of trees in Lansing Woods.

  One question of interest is whether the species are evenly mixed across the forest plot or are segregated into different subregions where one species is predominant over the others. (Handbook, p. 373)

  Some references:
  Handbook of spatial statistics, p. 373

  Source:
  R/spatstat, data(lansing)

- Copper data.

  These data come from an intensive geological survey of a 70 x 158 km region in central Queensland, Australia. They consist of 67 points representing copper ore deposits, and 146 line segments representing geological 'lineaments'. Lineaments are linear features, visible on a satellite image, that are believed to consist largely of geological faults (Berman, 1986, p. 55). It would be of great interest to predict the occurrence of copper deposits from the lineament pattern, since the latter can easily be observed on satellite images. Does the locations of deposits occur close to lineaments?

  Some references:
  Handbook of spatial statistics, p. 349-350
  Illian et al. p. 443

  Source:
  R/spatstat, data(copper)

- Chorley-Ribble Cancer Data.

  The data give the precise domicile addresses of new cases of cancer of the larynx (58 cases) and cancer of the lung (978 cases), recorded in the Chorley and South Ribble Health Authority of Lancashire (England) between 1974 and 1983. The supplementary data give the location of a disused industrial incinerator.

  Is there raised incidence of larynx in the vicinity of the disused industrial incinerator?

  Some references:
  Diggle, P. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. Journal of the Royal Statistical Soc. Series A 153, 349–362.

  Source:
  R/spatstat, data(chorley)

**More point pattern data sets available e.g. from:**
http://www.lancs.ac.uk/staff/diggle/pointpatternbook/Datasets/