Points clustering, classification, and density estimation GP tool using the MCLUST package

Konstantin Krivoruchko, July 7, 2015

Cluster analysis helps to identify groups of points on the plane that are separated from other points. The simplest clustering algorithm, nearest neighbor clustering, repeats the following action:

find the two nearest pairs of points and combine them into a group

until the required number of groups is reached or the distance between the remaining pair of points becomes larger than the specified distance. This and many other clustering algorithms are deterministic. They perform poorly when clusters are not well separated and they cannot answer questions like these:

- How many clusters are there?
- How certain that a point belongs to a particular cluster?
- How should we deal with outliers and noise?
- Which cluster model parameters should be used?

Probabilistic model-based clustering implemented in the *mclust* package written by Chris Fraley and Adrian Raftery is based on the idea that the observed data are a mixture of several populations with different ellipsoidal areas, shapes, orientations and possible noise. The model provides the uncertainty of the points classification on clusters. Partitions are determined by the maximum likelihood algorithm. The fitted models are compared using the Bayesian information criterion, which allows comparison of several models at the same time.

How to use this GP tool

The model-based clustering GP tool analyzes relative distribution of points and the only required input data for the tool is the points feature layer. The model predictions are reliable in the close vicinity to the outer points locations and, therefore, it is a good idea to clip the output by the user-defined polygon. Therefore, the clipping polygon is the second required input parameter for the model-based clustering GP tool.

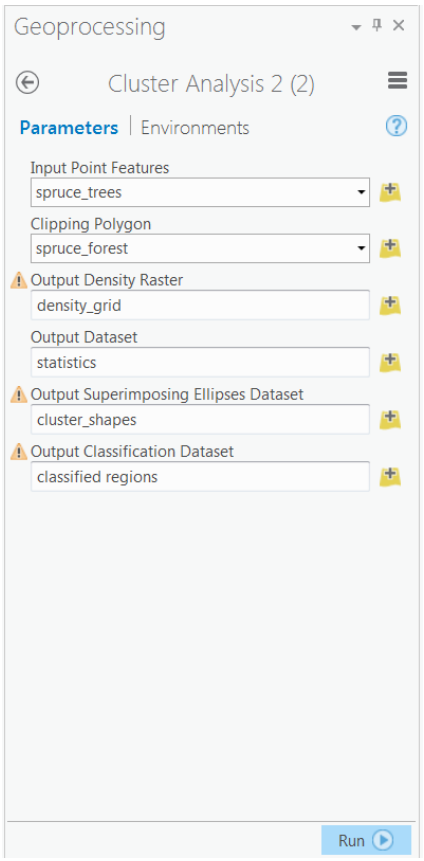

Figure 1. The model-based clustering GP tool.

I suggest the following changes to the tool:
Tool name: model-based clustering
Output Dataset -> classification uncertainty
Output superimposed Ellipses Dataset -> Standard deviation ellipses
Output classification dataset -> Classified regions
New output option with default value of 0: Number of simulated point sets and density rasters

Input points feature: the point feature class with observed points locations.
Clipping polygon: a border around the observed points locations.
Output density raster: estimated density of the observed points.
Classification uncertainty: conditional probability that a point belongs to each of the estimated clusters.
Standard deviation ellipses: the projections of the standard deviation of each Gaussian component.
Classified regions: partitioning of the clipping polygon into regions based on the points classification to the most likely cluster.
Number of simulated point sets and density rasters: the alternative point densities estimated using points simulated from the fitted clustering probability model. The default value is equal to zero, meaning that the alternative points densities are not produced.

Case study

Location of a tree in the forest depends on positions of other trees, soil characteristics, slope, and forest management in the past. Forest stands are usually inhomogeneous, and their characteristics vary in space.
Manual collection of the tree characteristics in the forest is expensive. Additionally, the limited number of samples that can be collected manually is not sufficient for the reliable forest features summary. An application of aerial photo interpretation to tree species identification and the assessment of disease and insect infestation allow incorporating statistical modeling into management strategies for maximizing forest yield, since thousands of samples can be collected and identified.
Figure 2 illustrates the individual tree identification from combined LIDAR and multispectral airborne imagery. A semiautomatic method allowed researchers to recognize nearly all dead trees, about 89 percent of the live spruce and about 62 percent of the deciduous crowns, and this performance can be improved.

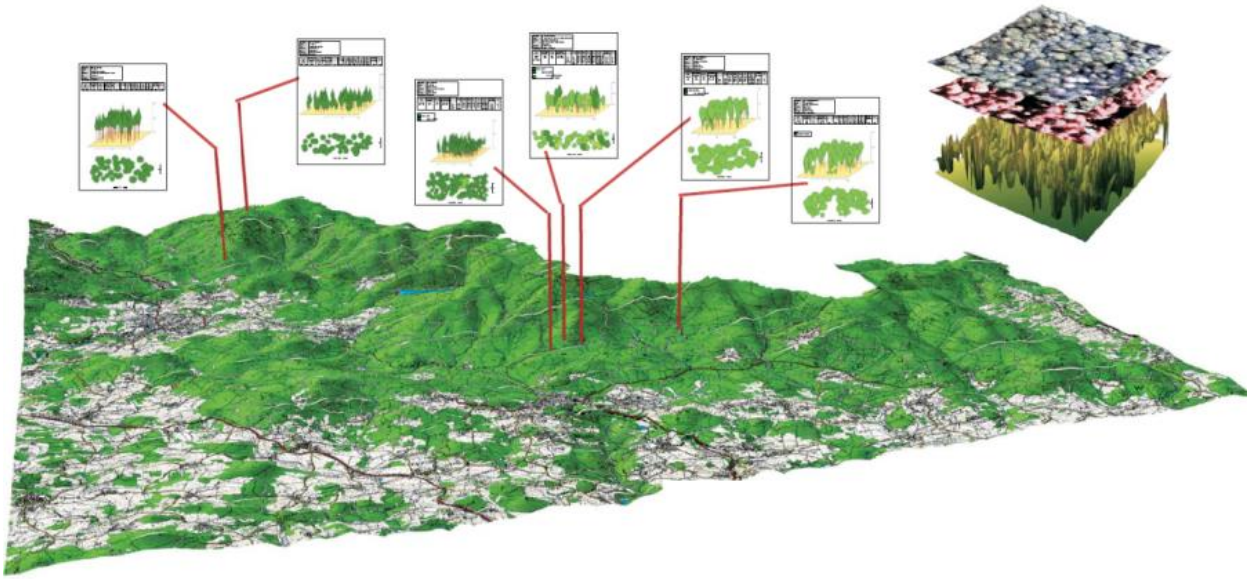

Figure 2. The individual tree identification from combined LIDAR and multispectral airborne imagery [4].

Figure 3 show two images and ground data collected in the Bavarian Forest National Park. These points are used as input to the model-based clustering GP tool. The clipping polygon was created using minimum bounding geometry GP tool with convex hull option.

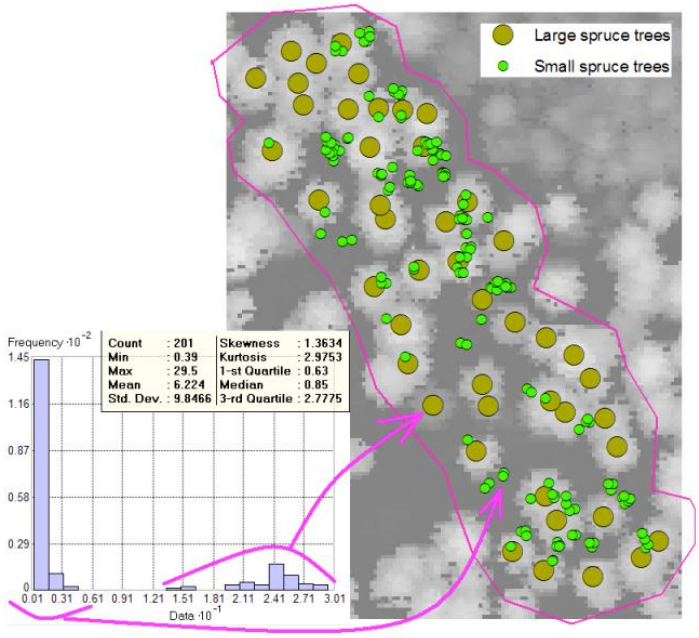

Figure 3. Locations of spruce trees of different sizes. A histogram shows distribution of the trees' heights (units are in meters) [4].

Figure 4 shows the output from the model-based clustering GP tool:

- The data region partitioning based on the clustering classification.
- The conditional probability that points belong to one particular cluster.
- The classification uncertainty.
- The projections of the standard deviation of each Gaussian component (black ellipses).
- The points density.

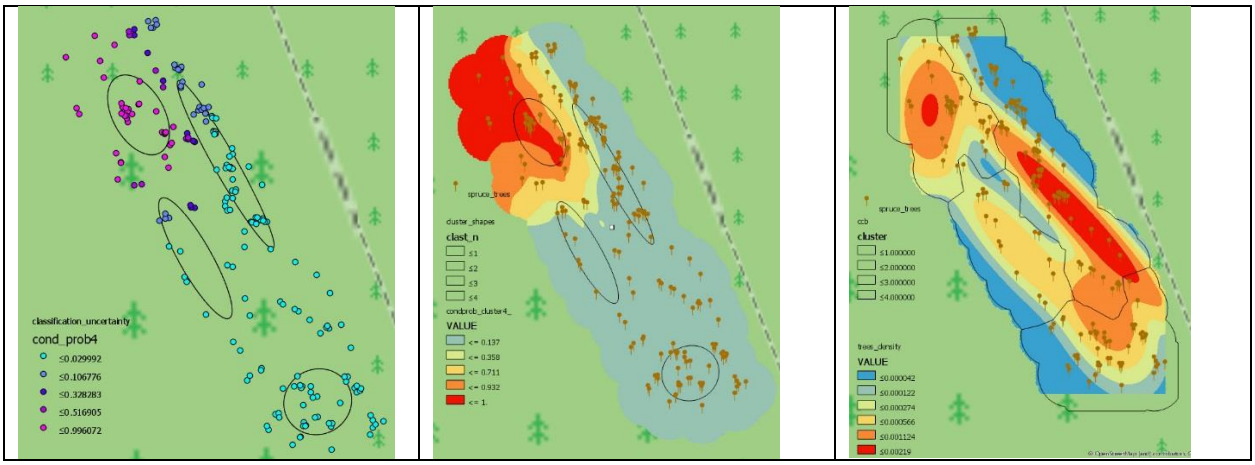

Figure 4. Outputs from the Model-based Clustering GP tools. Left: classification uncertainty and the projections of the standard deviation of each Gaussian component (black ellipses). Center: The ellipses and the conditional probability that points belong to one particular cluster. Right: points density and the data region partitioning based on the clustering classification.

The main use of the output maps is exploratory forest trees locations analyses, including comparison with typical and well-studied forest stands in the region.

Figure 5 shows additional output from the GP tool, three simulations of the tree locations and their densities using the fitted point clustering model. Note that since the trees clustering model is probabilistic, the fitted model each time generates different number of trees.
If the model was fitted correctly, we may find such or very similar trees distributions in the nearby areas of the forest. Then the fitted model can help with the forest management. For example, if the study area is typical for the studied forest and the forest area is 1000 times larger than the study area, the forest management can be based on the statistics from 1000 simulated outputs from the probabilistic clustering model.

<to be added>
Figure 5. Three simulations of the tree locations and their densities using the fitted point clustering model.

Further reading

[1] Fraley, C. and Raftery, A.E. (2003). Enhanced model-based clustering, density estimation and discriminant analysis software: MCLUST. Journal of Classification, 20, 263-286.
[2] Fraley C. and Raftery A.E. (2007). Model-based methods of classification: Using the mclust software in chemometrics. *Journal of Statistical Software*, 18, paper i06.
[3] Koch B., Heyder U., and Weinacker H. (2006) Detection of Individual Tree Crowns in Airborne Lidar Data. Photogrammetric Engineering & Remote Sensing, Vol. 72, No. 4, pp. 357–363.
[4] Krivoruchko, K. (2011) *Spatial Statistical Data Analysis for GIS Users*. ESRI Press, 928 pp:

- Cluster detection methods in regional data, chapter 11.
- Cluster analysis, chapter 13.
- Cluster analysis using the mclust02 package, chapter 16.
- Point pattern analysis in forestry, chapter 6.