Cluster analysis of geological point processes with R free software

Authors

- **Marj Tonini**, University of Lausanne, *Switzerland*
- **Antonio Abellán**, University of Lausanne, *Switzerland*
- **Andrea Pedrazzini**, University of Lausanne, *Switzerland*

KEYWORDS : landslides, spatial point pattern, R free software, Ripley's K-function, Nearest Neighbour clutter removal

Introduction

Landslides, as many other geological events (e.g. earthquakes, volcanoes, etc), are normally not randomly distributed but grouped in clusters both in space and in time. The analysis of their spatial properties and distribution is fundamental to understand their predisposing factors, and for prevention and forecasting purposes. From a statistical point of view, geologic events can be represented as «point processes» whose spatial distribution can be analysed using mathematical models for irregular or random point pattern.

Pattern recognition and specifically cluster analysis includes algorithms aiming at grouping objects showing similar properties into the respective categories. Spatial clusters can be identified whenever the observed distance among groups of point locations in space is lower than the expected distance for a random distribution. This assumption can be accepted or rejected based on the results of statistic tests. For geological events, which intensity is not constant and clusters are often not isolated, their detection is not evident. A vast literature exists on the spatial analysis of landslide distribution, especially for susceptibility map purpose, [4,7,11,12,13,16] while their spatial characterisation by means cluster algorithms is less investigated [18].

Two main types of spatial cluster algorithms can be outlined: 1) global cluster indicator, allowing to measure and test for the randomness of the point process; 2) local cluster algorithms allowing to identify clusters in space and/or in time. The present study illustrates two examples of application of the two abovementioned kinds of analysis by means of two distinct case studies: 1) the Ripley's K-function was applied to assess the global spatial attraction (clustering) among mapped landslides; 2) Nearest neighbour clutter removal method was applied to automatically detect landslides in LiDAR (Light Detection and Ranging) point clouds.

Computations were carried out using R free software for statistical computing and graphics [14]. R is a free software environment integrating facilities for data manipulation, calculation and graphical display. The R base can be extended via packages available through the Comprehensive R Archive Network (CRAN) which covers a very wide range of modern statistics. More specifically, the spatial point pattern analyses of the geological events considered in the present study and their cluster detection were supported by the package spatstat [3].

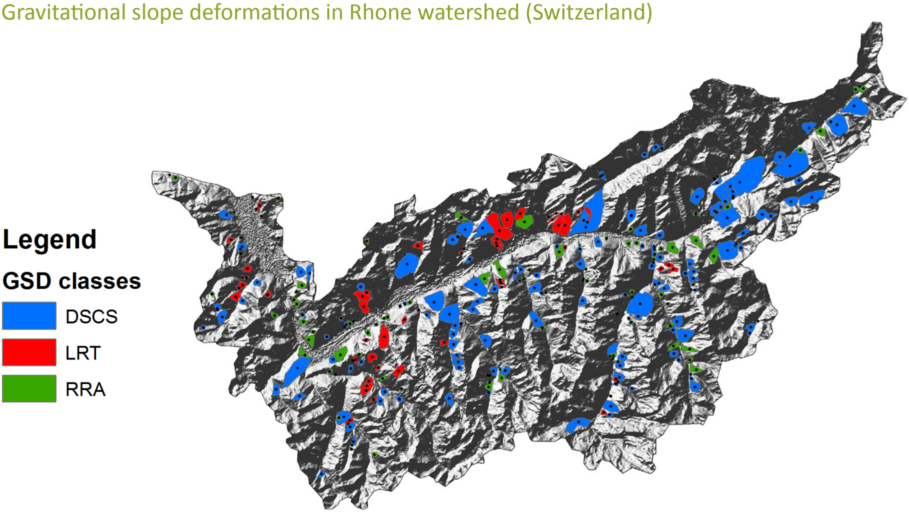
Case studies

1. The use of K-function to detect spatial pattern of landslides

In the present study the spatial pattern of mapped landslides inventoried in the Rhone watershed (Switzerland) was analysed. A total of 294 gravitational slope deformations (GSD) were detected and classified based on their typology. The complete GSD geo-database was implemented at the Institute of Geomatics and Analysis of Risk (Lausanne University, Switzerland). The identification of the landslides was based on different sources of information such as geological maps, aerial photos and orthophotos, digital elevation model. The inventoried GSDs were furthermore classified following the modified version of the Hutchinson (1988) [9], resulting in three main classes: Rockslide and Rock-Avalanches (RRA), Deep Seated Creep/Sagging (DSCS) and Large Roto-Translation slides (LRT). The main objective of the present study is to verify if these events show a clustered or random distribution, and if it exists a spatial attraction (i.e. cluster) at a specific distance among the different classes of landslides.

FIGURE 1

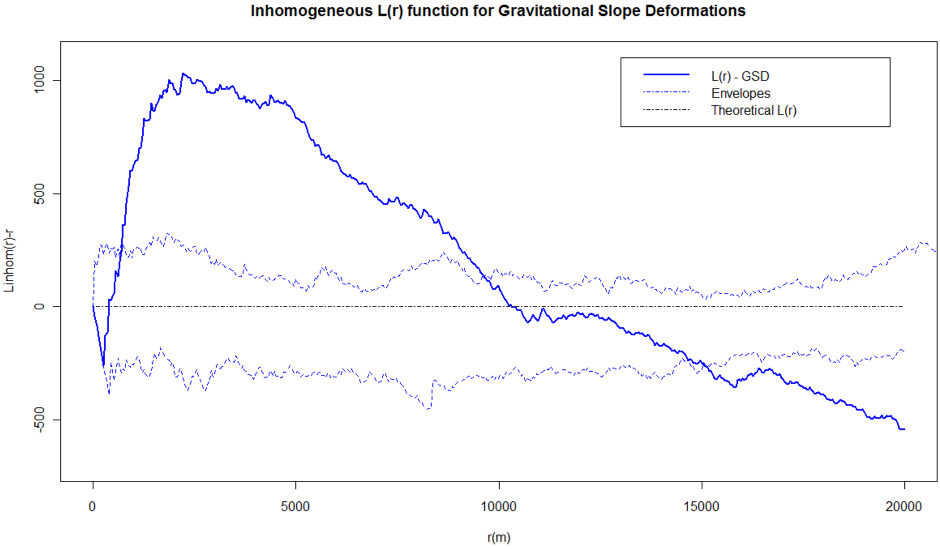
Gravitational slope deformations in Rhone watershed (Switzerland)



To test for randomness, the Ripley's K-function [15] was computed: the difference among the K-functions calculated for each dataset helped to compare the individual spatial pattern of each type of landslide and to reveal if they display similar cluster behaviour. Analytically ʌK(r) (where ʌ is the intensity of the point process) equals the expected number of additional points within a distance r from a randomly distributed event. Under complete spatial randomness (CSR) the theoretical K(r) is equal to πr2. The estimated K(r) can be plotted against the distance r and compared with the theoretical one: if the estimated function at a given distance r is higher than πr2, events are spatially clustered, whilst smaller values indicate repulsion between events. That way allows finding out at which range of distance data perform a non-random pattern distribution. To account for the natural non-uniform distribution of the geological events along the study area, a generalisation of K(r) for a spatial inhomogeneous distribution was applied [2]. Edge correction was also introduced in the computation.

FIGURE 2

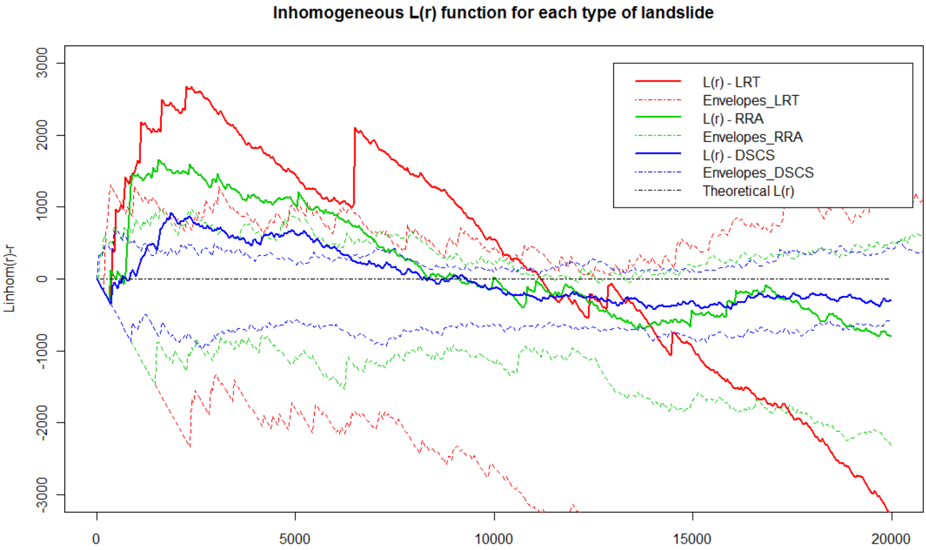
Inhomogeneous L(r) function for Gravitational Slope Deformations



In the present study we used a transformation of the Ripley's K-function, namely the L-function [5] that makes easier to compare the estimated with the theoretical curve and to evaluate departures from this last one. L(r) equals the square root of K(r) over π, minus r, so that the L(r) theoretical value is zero at every distance. To test for spatial randomness, 999 Monte Carlo simulations of a realisation of an inhomogeneous random point process were performed to provide confidence envelopes. Results show that GSDs are not randomly distributed over the study area: indeed they are clustered at a distance ranging from about 75 m up to about 10 km, dispersed above about 15 km and included between the upper and the lower simulated curves in between, meaning a random distribution in this range.

FIGURE 3

Inhomogeneous L(r) function for for each type of landslide



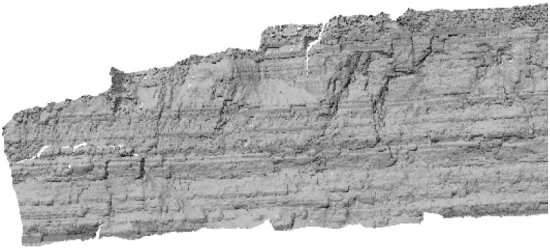
The cluster behaviour shows a maximum at about 2.5 km: this value can be retained for future local cluster analyses aiming to locate clusters in space. The L(r)-functions computed individually over the single datasets show that the three classes of landslides have similar pattern behaviour, with a cluster tendency at a distance ranging from 500 m up to about 10 km, and that events belonging to LRT are more clustered than the events belonging to the RRA and DSCS landslide types.

2. Landslide recognition in LiDAR point clouds using NN-clutter removal

Light Detection and Ranging (LiDAR) is a remote sensing technique that allows obtaining the geometry of the terrain through the detection of the distance from the sensor to a given target. When mounted over a ground-based sensor, the so called Terrestrial Laser Scanner (TLS) is able to acquire

FIGURE 4

LiDAR points cloud

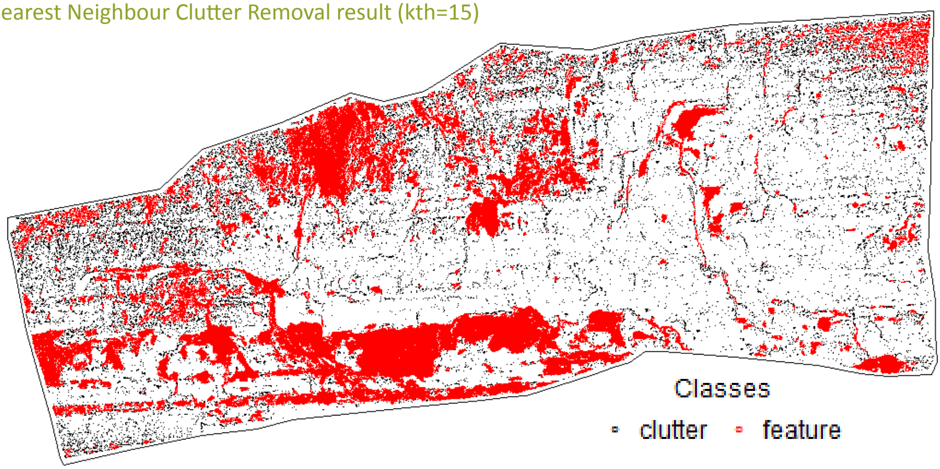


great resolution 3D information of the terrain (e.g. over 100 point per square meter) which results in a more or less homogeneously distributed point cloud. These points can be used to extract features' information of vegetation canopy, rivers, canyons, etc. Nevertheless, the automatic extraction of these features is not an evident task and different methods are currently being developed in different domains [8,17,10]. In the present study, we analysed TLS point clouds in order to automatically detect and extract landslides occurred in a pilot study area (Puigcercos, Catalonia, Spain).

The study area corresponds to a cliff affected by a high number of rockfalls per year [1]. The almost vertical geometry of the cliff allowed georeferencing points based on X,Z coordinates and to adopt a 2D approach.

FIGURE 5

Nearest Neighbour Clutter Removal result (kth=15)



The method is based on the Nearest Neighbour Clutter Removal (NNCR) in combination with the Expectation–Maximization (EM) algorithm [6]. The NNCR algorithm consists in compute the distance to the kth nearest neighbour for each point in the pattern: intuitively the points inside regions of higher density (i.e. in the features) have a smaller kth distance than the points inside regions of lower density (i.e. in the clutter).

Then the EM algorithm fits a mixture distribution (the feature and the clutter) to the nearest neighbour distances, which is used to classify each point as belong to the class "feature" or "clutter". The degree on neighbour (k) has to be fixed from the user and this choice can affect the result of the analysis. In the present study several increasing values for k were applied and then the one based on an entropy type measure of separation was retained.

This method was applied after the pre-filtering of the LiDAR pint cloud in order to remove points affected by instrumental error. The density of points belonging to landslides is higher than the density of points falling outside and the applied technique allowed assigning them to the class "feature" with their estimated probabilities. This preliminary result can be used to calculate landslide volume or to analyse landslide spatio/temporal patterns. The proposed method allowed automatically detecting landslides gaining in times and in accuracy compared to the visual technique.

Further development of the research will consist in the custom implementation of the algorithms in the R environment to be able to analyse multi-dimensional point patterns, including the third spatial dimension and/or the temporal dimension.

[1] ABELLÁN, A., CALVET, J., VILAPLANA, J.M., AND BLANCHARD, J. Detection and spatial prediction of rockfalls by means of terrestrial laser scanner monitoring. *Geomorphology, 119* (2010), pp. 162-171.

[2] BADDELEY, A., MOLLER, J., WAAGEPETERSEN, R. Non- and semiparametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica 54*, (2000), pp. 329–350.

[3] BADDELEY A., AND TURNER R. Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software,* 12 (6) (2005), pp. 1–42.

[4] BAI, S.B., WANG, J., LÜ, G., ZHOU, P., HOU, S.S., XU. S.N. GIS-based logistic regression for landslide susceptibility mapping of the Zhongxian segment in *the Three Gorges area, China*. Geomorphology, 115 (2010), pp. 23–31.

[5] BESAG, J. Discussion of Dr Ripley's paper. *Journal of the Royal Statistical Society*, Series B, 39, (1977), pp. 193–195.

[6] BYERS, S., AND RAFTERY, A. E. Nearest-Neighbor Clutter Removal for Estimating Features in Spatial Point Processes. *Journal of the American Statistical Association*, 93, (1988), pp. 577-584.

[7] ERENER, A., AND DÜZGÜN, H.S.B. Landslide susceptibility assessment: what are the effects of mapping unit and mapping method? *Environmental Earth Sciences*, June 2012, Volume 66, Issue 3, pp. 859-877.

[8] GIGLI, G, AND CASAGLI N. Semi-automatic extraction of rock mass structural data from high resolution LiDAR point clouds. *International Journal of Rock Mechanics and Mining Sciences*, Volume 48, Issue 2, February 2011, pp. 187-198.

[9] HUTCHINSON, J. N. General Report: Morphological and geotechnical parameters of landslides in relation to geology and hydrogeology. *Proceedings of the Fifth International Symposium on Landslides*. Edited by: Bonnard, C., Balkema, Rotterdam, pp.3–35 (1988).

[10] KABOLIZADE, M., EBADI, H., AHMADI, S. An improved snake model for automatic extraction of buildings from urban aerial images and LiDAR data. *Computers, Environment and Urban Systems*, Volume 34, Issue 5, August 2010, pp. 435-441.

[11] LEE, S., RYU, J.-H., KIM, I.-S. Landslide susceptibility analysis and its verification using likelihood ratio, logistic regression, and artificial neural network models: Case study of Youngin, Korea. *Landslides*, 4 (2007), pp. 327–338.

[12] NANDI, A., AND SHAKOOR A. A GIS-based landslide susceptibility evaluation using bivariate and multivariate statistical analyses. *Engineering Geology*, 110 (2010), pp. 11–20.

[13] OH, H.J., AND LEE, S. Landslide susceptibility mapping on Panaon Island, Philippines using a geographic information system. *Environmental Earth Sciences*, 62 (2011), pp. 935–951.

[14] R DEVELOPMENT CORE TEAM (2012). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. *Retrieved September 7*, 2012 from http://www.R-project.org/

[15] RIPLEY, B.D. Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society*, Series B, 39, (1977), pp. 172-212.

[16] VAN DEN EECKHAUT, M., POESEN, J., GULLENTOPS, F., VANDEKERCKHOVE, L., HERVÁS J. Regional mapping and characterisation of old landslides in hilly regions using LiDAR-based imagery in Southern Flanders. *Quaternary Research*, Volume 75, Issue 3, May 2011, pp. 721-733.

[17] ZHAO, K., POPESCU, S., MENG, X., PANG, Y., AGCA, M. Characterizing forest canopy structure with LiDAR composite metrics and machine learning. *Remote Sensing of Environment*, Volume 115, Issue 8, 15, August 2011, pp. 1978-1996.

[18] ZUO, R., AGTERBERG, F.P., CHENG, Q., YAO, L. Fractal characterization of the spatial distribution of geological point processes. *International Journal of Applied Earth Observation and Geoinformation*, Volume 11, Issue 6, December 2009, pp. 394-402.