Jul 1st, 12:00 AM

Landslide Data Analysis with Gaussian Mixture Model

V. Timonin

S. B. Bai

J. Wang

M. Kanevski

A. Pozdnukhov

iEMSs 2008: International Congress on Environmental Modelling and Software
Integrating Sciences and Information Technology for Environmental Assessment and Decision Making
4th Biennial Meeting of iEMSs, http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings
M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)
International Environmental Modelling and Software Society (iEMSs), 2008

Landslide Data Analysis with Gaussian Mixture Model

V. Timonin[a], S.B. Bai[b], J. Wang[b], M. Kanevski[a] and **A. Pozdnoukhov[a]**

[a] *Institute of Geomatics and Analysis of Risk, University of Lausanne, Switzerland (Alexei.Pozdnoukhov@unil.ch)*

[b] *Nanjing Normal University, Nanjing, China*

Abstract: In this paper we present the approach for the analysis and modeling of landslide data using the Gaussian Mixture Model. We model the probability density of the landslide events in the high-dimensional space of parameters, conventionally used for predicting the landslide susceptibility. This work describes the application of the method for the area of Bailongjiang River, in northwest China. The available information includes the digital elevation model of the region, geological map and different GIS layers including land cover data obtained from satellite imagery. The landslides were observed with aerial imagery and documented during the field studies.

Keywords: Susceptibility mapping; Landslides; Gaussian Mixture Model.

1. INTRODUCTION

In this paper we present an approach to the analysis and modeling of landslide data using one of the baseline statistical and machine learning algorithms known as Gaussian Mixture Model. The information available for modeling includes the digital elevation model of the region with resolution of 30 meters, the geological map, the climatological information obtained as an output of large scale physical model, and several different GIS layers including land cover data obtained from satellite imagery. The preceding landslides were observed during the field studies, providing the coordinates and the shapes for 144 landslides.

Traditionally, the logistic regression approaches are commonly used in landslide susceptibility mapping. However, complex relations of the landslide phenomena with different factors ask for modeling with more adequate tools. These are, for example, the robust non-linear non-parametric data-driven algorithms of machine learning. The application of machine learning algorithms [Hastie et al., 2001, Haykin 1998] for the analysis and modeling of landslides, in particular for the landslides susceptibility analysis, have already demonstrated their predictive capability and robustness [Ermini et al., 2005; Melchiore et al., 2008; Yao et al., 2008].

Another important concern deals with the setting of the problem of landslide susceptibility mapping. Traditionally, it has been considered as a classification problem aiming to discriminate the dangerous zones from the safe ones. While it is relatively straightforward to define the dangerous zones as the ones where landslides has happened, the selection of the safe ones which are both reliably safe and brings useful discriminative information is much less evident. Having said that, we will study the setting of the susceptibility mapping as a probability density modeling problem, that is, the task of this study is the modeling of the probability density of the landslides in the space of input parameters.

A nonparametric approach based on Gaussian Mixture Models (GMM) is applied for nonlinear mapping of probability densities of landslide areas. Input information contains eleven features and tuning of GMM model was based on Expectation-Maximization algorithm.

2. GEOLOGICAL ENVIRONMENT OF THE STUDIED AREA

The study area covers 1361 km^2 mainly on the Bailongjiang River, in northwest China. The site includes 6 sub-basins and the landslide area covers an area of 19.67 km^2. The site lies in the middle south of the west wing of Qinling orogen. The area is formed by Qinghai-Tibet tectonic belt and Wudu arc structure and is affected by unlift of the Qinghai-Tibet plateau. When structure movement shift its forms from main horizontal movement to main vertical movement, it is a very profound effect to the wings of Qinling in this location. The unbalanced vertical movement creates the extreme development of folds, crushs, faults and joints in the location. Lithology is mainly phyllite, schist, slate, carbonate rocks and all kinds of clastic rocks. Today in China tectonic units can be divided into three teconic systems: Alps - Himalaya tectonic system, shortly as Tethys tectonic domain; the Pacific tectonic system and the new generation of Central Xibailie tectonic system (or ancient Asia tectonic domain). Two tectonic belts: Helan-Sichuan-Yunnan north and south of the belt and Kunlun-Qilian-Qinling-Dabieshan east and west of tectonic belt (central orogenic system).

2.1 Lithology of stratum

The case study area possesses much of loose soil layer due to all kinds of causes, such as the Silurian, Devonian, Carboniferous, Permian and Triassic in Paleozoic and the Triassic of Mesozoic and Quaternary. The oldest stratum is the Silurian in the area which mainly distributes in the two shores of Bailongjiang River and forms multiple anticlines. The stratum constructed mainly by shallow marine sedimentary metamorphic rocks and carbonate rocks. From east to west, the sediment thickness changes from thin to thick, sediment is from coarse to fine and carbon capacity is from low to high. Lower Silurian is composed by carbonate rocks and clastic rock and is the nuclear of multiple anticline of Bailongjiang river, at the same time, it lies in the south of Bailongjiang river. Between middle and upper Silurian is called Bailongjiang group maked up of clastic and carbonate rocks. It mainly distributes in the Bailongjiang northern shore of the east of two estuaries and southern shore from the west of two estuaries to Zhouqu. Devonian develops greatly in the area and mainly located in two wings of Bailongjiang multiple anticlines. Middle Devonian and upper Devonian have a close connection with landslide, while as we know Middle Devonian mainly distributes in the north wing of Bailongjiang multiple anticlines.

2.2 Geological Structure

The study area is located in the intersection of the new regional structure north-south and east-west tectonic activity zone in the north edge of the Qinghai-Tibet Plateau. In the tectonic, it is the West Qinling orogen ofa Qinling micro-plot .With the long-term impact of tectonic activity , displayed extrusion zone along in the western ;It is along a EW trend in the eastern ,including folds and faults,while the stratums also distribute to the NW-NWW as a band Since the Cenozoic, Indian plate and Eurasian plate collision developed mountains, the Qinghai-Tibet Plateau uplift ,and large-scale strike-slip of altiplano crust result in stress-strain field and tectonic activity exceptionally complex. Strong seismic activities and water system development of Bailong River drainage are the main reasons for the landslide and damming disasters.

2.3 Landform

In the landform, the study area is located in the southern of the Qinling Mountain, and has tall upright mountains, steepness terrain, deep valleys, fast-flowing river and presents V-shaped valley or canyon terrain features. It is the erosion middle and high mountain landform.

2.4 Hydrology and climate characteristics

The study area is the drainage of Bailong River, which is a tributary of the Changjiang River. The water system of Bailong River is plume-shaped elongated at NW-SE. The mainstream of Gansu is 475 km long and the both sides of river are asymmetric, South being wide and North narrow. Bailong River has many tributaries. The most lengthy tributaries are Min River, Beiyu River and Baishui River.

The rainfall is not balanced in the time and space. In time, the rainfalls mainly concentrates in the 4-10 months, from 6-9 months, Dangchang rainfall is 59.9 percent of the annual rainfall, Wudu is 65.5%, Wenxian is 62% and Zhouqu is 61.7%. Short and heavy rainfalls are typical for the region.

In the spatial distribution, rainfall trends to decrease from south to north such as from 900mm of the Wudu southeast to 500mm of the Wudu northwest. The rainfall intensity tends to increase with elevation. The average annual rainfall is 400-500mm below the elevation of 1500m, 500-600mm between 1500m and 2000m and more than 600 mm above 2000m.

Different climatic zones differentiated by elevation are easily distinguished in this region. The temperature decreases with altitude, providing the annual average temperature to be below 5°C at the elevations of more than 2500m. There are very severe differences between the extreme yearly maximum and minimum temperatures.

3. GAUSSIAN MIXTURE MODEL

3.1 Model description

First let us consider briefly some theoretical questions concerning Gaussian Mixture Models and mapping via density modeling. Usually Mixture Models are used for the *density estimation* of the data. Density estimation is the construction of an estimate, based on observed data, of an unobservable underlying *probability density function (p.d.f.)*. Mixture Model estimates density distribution in a form of a linear combination of some simple functions (called components, units, or kernels):

$$p(x) = \sum_{j=1}^{m} p(x \mid j) P(j) \tag{1}$$

Such representation of p.d.f. is called a *mixture distribution* [Titterington *et al.*, 1985]. *P(j)* are mixing coefficients. In a Bayesian framework, *P(j)* can be considered as *prior* probabilities of any data point having been generated from component *j* of the mixture. These priors, like any probability, should satisfy the following constraints:

$$\sum_{j=1}^{m} P(j) = 1, \qquad 0 \le P(j) \le 1 \tag{2}$$

The component density functions *p(x|j)* are also normalised such that

$$\int p(x \mid j) dx = 1 \tag{3}$$

In a Bayesian framework, one can introduce the corresponding *posterior* probabilities, which can be expressed using theorem of Bayes in the following form:

$$P(j \mid x) = \frac{P(x \mid j) P(j)}{p(x)}, \quad \sum_{j=1}^{m} P(j \mid x) = 1, \qquad 0 \le P(j \mid x) \le 1 \tag{4}$$

The value of *P(j|x)* – the *posterior* probability - represents the probability that a particular component *j* was responsible for generating data point *x*.

The most important property of such model is that it can approximate any continuous density with an arbitrary accuracy (if the number of components is large enough and the parameters of the model are chosen correctly, see, for example, [Silverman, 1986]).

The most often choice of the function for the component is a Gaussian one:

$$p(x \mid j) = \frac{1}{(2\pi\sigma_j^2)^{d/2}} \exp\left\{ -\frac{\left\| x - \mu_j \right\|^2}{2\sigma_j^2} \right\} \tag{5}$$

Here the σ^2 parameter for each *j* from *1...m* interval is a scalar and we deal with a model of *m isotropic* components. In more general case, parameter σ^2 may be presented as a full covariance matrix Σ. Covariance matrix is a squared symmetrical matrix of dimension *d* by *d* and the number of parameters is *d(d+1)/2*. In this case the (5) can be rewritten as:

$$p(x \mid j) = \frac{1}{(2\pi)^{d/2}(\det \Sigma_j)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\right\} \quad (6)$$

where *det* denotes determinant and Σ^{-1} is an inverse of Σ matrix.

A Mixture Model with Gaussian components is called Gaussian Mixture Model.

A model with *anisotropic* components is much more flexible but less stable and requres more calculations and patemeters tuning. It is vital especially in case of complex multidimensional data. So in the current study isotropic model will be used.

3.1 Tuning of parameters. Expectation-Maximisation algorithm

Various procedures have been developed for determining the parameters of the GMM from a given data set. One of the most famous is an *expectation-maximisation (EM)* algorithm [Dempster *et al.*, 1977].

At the first step, *E-step*, algorithm tries to detect which component is responsible for generating each point of the data set. At the next step, *M-step,* parameters of the mixture taking into account the results of the first step (by maximising the expected likelihood) are tuned.

Updating of parameters of the mixture model by EM algorithms is given by the following iterative formulas:

$$\mu_j^{new} = \frac{\sum_n P^{old}(j \mid x^n) x^n}{\sum_n P^{old}(j \mid x^n)} \quad (7)$$

$$(\sigma_j^{new})^2 = \frac{\sum_n P^{old}(j \mid x^n) \left\| x^n - \mu_j^{new} \right\|^2}{\sum_n P^{old}(j \mid x^n)} \quad (8)$$

$$P(j)^{new} = \frac{1}{N} \sum_n P^{old}(j \mid x^n) \quad (9)$$

where "*old*" means values from the previous step of the algorithm.

4. CASE STUDY

The GMM will be used to model the probability density function which, being applied to the previously unseen territories, can be used for mapping the landslide susceptibility, i.e. the potential areas where landslides can occur. As a training data set, the areas (scars and masses) where landslides were observed are used. The training data set is formed pixel-wise from inside the landslides polygons, which were outlined by experts. To validate the quality of mapping, the studied area was divided in two parts: west (76 polygons, training part) and east (68 polygons, validation one). In Figure 1 the digital elevation model (DEM) of the studied area with 144 polygons of observed landslides and data division into training and validation subsets are presented.

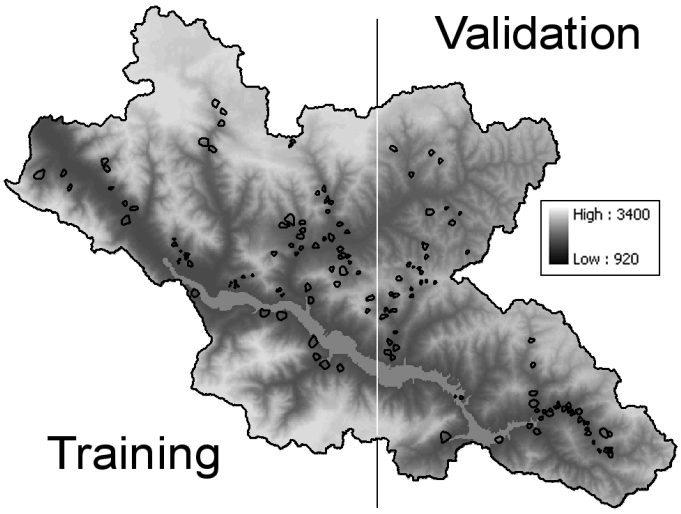

Figure 1. Digital elevation model (DEM) of the studied area. Polygons are the areas (scars and masses) of the observed landslides outlined by experts (total 144). Left (76 polygons, west) part is a training data set, right (68 polygons, east) is a validation one.

Eleven different features were presented as input information: DEM, slope, aspect, curvature (profile and plan), distance to inhabited locality, distance to river, distance to road, fall distance, lithofacies, and land use types. Note that last 2 features are categorical and cannot be presented to the model based on the distance calculation. So it is necessary to present such type of information in some encoded manner. The widely used schema for this purpose is to present *m*-category feature as an *m*-dimensional attribute with a single non-zero component corresponding to the category number. For example, the landuse type feature has six categories. So, it is necessary to define six-dimensional attribute as input information for the model. Category 1 will be presented as (1,0,0,0,0,0) vector, category 2 as (0,1,0,0,0,0), and so on. The lithofacies feature has seven categories, and the seven-dimensional attribute was defined. As a result, the original 11 features of information are presented for the model as 22-dimensional input vector.

Two GMM models with 12 and 40 isotropic Gaussian kernels (5) were explored. The density distribution (1) was calculated with EM-algorithm until convergence. The resulting model is combined as a model with the maximum likelihood from 50 tries with different initial locations of the units.

Both models are flexible enough to reproduce the training data, that is, they provide high probabilities at the areas where the landslides have occurred. The next step in model evaluation is to calculate the probability density values for all studied area. The validation of the model will be carried out using the areas of validation landslides. Given sufficient performance, the model could be used to explore the areas where high susceptibility values are predicted but historically the landslides have not yet occurred (or were not yet registered by experts).

5. RESULTS

The results as the maps of the probability density values obtained with two models are presented in Figure 2. Both maps are very similar instead of a different number of kernels (number of free, tuned parameters) in the models. The model with 12 kernels is somewhat smoother than the 40-kernels one. Both models cover the areas outlined by the training polygons in the validation (east) part of the area. Also there are some well-defined areas with high density values in the areas without training polygons. These areas can be treated as potentially landslides dangerous (where historically landslides were not occurred or not registered by experts).

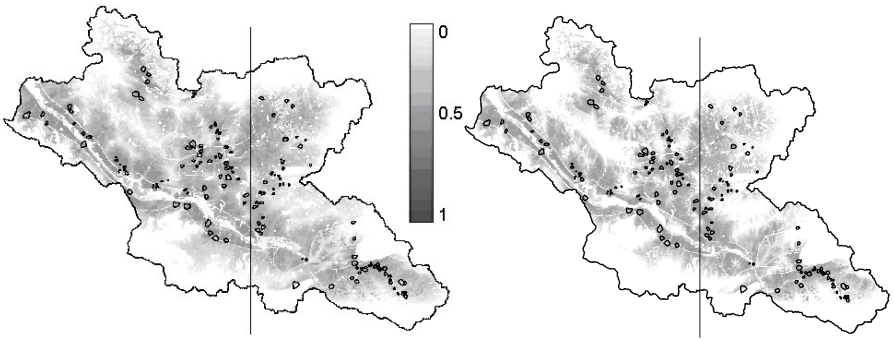

Figure 2. Map of mixture density estimation of the landslide susceptibility areas by 12-kernels model (left), and 40-kernels model (right. Darker colours present pixels with higher density values. Training and validation areas are outlined with polygons.

The analysis of the outputs can also be done by considering the obtained landslide susceptibility model with respect to the input physical parameters. The histogram of susceptibility values and slope inclination is illustrated in Figure 3 (left). It suggests that most landslides in the area are observed at the inclinations of 30-35 degrees.

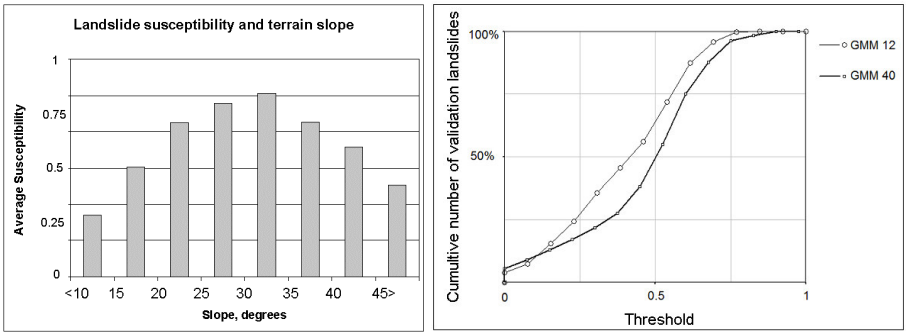

Figure 3. Left: Average susceptibility suggested by GMM model with respect to slope inclination. Right: cumulative number of landslides with respect to the selected threshold, computed pixel-wise on the validation dataset.

Figure 3 (right) illustrates the cumulative number of landslide cells (calculated pixel-wise) when changing the susceptibility threshold. One can notice that at the threshold level of 0.8 the model reproduced all of the validation landslides. By further taking several thresholds on the probability density values, one may select the conventional categorical hazard levels.

6. DISCUSSION AND CONCLUSIONS

The current study shows the feasibility of the GMM modeling as a tool for landslides predictions or susceptibility mapping. Model does not only reproduce the areas where landslides were occurred (both in the training and validation parts of the studied area) but also predict potentially hazardous areas. By selecting the thresholds for the continuous output of GMM model, one may delineate the conventional hazard class zones. While the results on validation dataset were quite promising in terms of landslide detection, more analysis is needed to analyse the false alarm rate. This would in turn require finding a non-trivial dataset of safe areas, which can be obtained with long and intensive field studies.

The same problem appears when trying to apply the conventional susceptibility mapping technique such as logistic regression. GMM avoids the latter obstacle by directly modeling the probability of the dangerous class in the input space of parameters. The results are robust from the point of view of model stability, that is, with respect to the number of units in the mixture. Though, GMM model is based on pair-wise distances in the input space of high dimension thus may be prone to the so-called “curse of dimensionality”.

The next steps in this research may be directed towards the selection of the most informative input features or their (generally, nonlinear) combinations for improving the quality of mapping. Other machine learning algorithms, including the one-class Support Vector Machines could also find interesting applications in landslide susceptibility mapping.

ACKNOWLEDGEMENTS

The authors wish to thank Swiss National Science Foundation (SNF) for the partial financial support of the projects “GeoKernels”(200021-113944).

REFERENCES

Dempster, A.P., N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society*, B 39 (1), 1-38, 1977.

Ermini L., Catani F., and Casagli N. Artificial neural networks applied to landslide susceptibility assessment. *Geomorphology*, 66, 327-343, 2005.

Hastie T., Tibshirani R. and Friedman J. (2001). The elements of Statistical Learning. Springer.

Haykin S. Neural Networks. A Comprehensive Foundation. (1998). Prentice Hall.

Melchiore C., Matteucci M., Azzoni A., and Zanchi A. Artificial neural networks and cluster analysis in landslide susceptibility zonation. *Geomorphology*, 94, 379-400, 2008.

Titterington, D.M., A.F.M. Smith, and U.E. Makov. Statistical Analysis of Finite Mixture Distributions. New York: John Wiley, 1985.

Silverman, B.W. Density Estimation for Statistics and Data Analysis. Chapman and Hall, New York, 1986.

Yao X., Tham L.G., and Dai F.C. Landslide susceptibility mapping based on support vector machines: A case study on natural slopes of Hong Kong, China. Geomorphology, doi: 10.1016/j.geomorph.2008.02.011, 2008.