Quantitative Comparison of Probabilistic Multi-source Spatial Data Integration Models for Landslide Hazard Assessment

No-Wook Park, Kwang-Hoon Chi
Geoscience Information Center, KIGAM
30 Gajeong-dong, Yuseong-gu, Daejeon 305-350, Korea
nwpark@kigam.re.kr

Chang-Jo F. Chung
Geological Survey of Canada

Byung-Doo Kwon
Department of Earth Sciences, Seoul National University
San 56-1, Shillim-dong, Kanwak-gu, Seoul 151-742, Korea

Abstract: This paper presents multi-source spatial data integration models based on probability theory for landslide hazard assessment. Four probabilistic models such as empirical likelihood ratio estimation, logistic regression, generalized additive and predictive discriminant models are proposed and applied. The models proposed here are theoretically based on statistical relationships between landslide occurrences and input spatial data sets. Those models especially have the advantage of direct use of continuous data without any information loss. A case study from the Gangneung area, Korea was carried out to quantitatively assess those four models and to discuss operational issues.

Keywords: Integration, Multi-source Spatial Data, Landslide Hazard.

1. Introduction

Nowadays, the rapid growth of GIS and remote sensing techniques had made multi-source/sensor spatial data more available for geoscientific applications. Since most geoscientific phenomena are connected to various physical variables, it is necessary to analyze multi-source data in an integrated way. To deal with enormous amount of multi-source spatial data, more systematic management of spatial data is essential to obtain the best reasonable interpretation. Since the late 1980s, several methods designed for multi-source spatial data integration (e.g. Bayesian probabilistic models, fuzzy logic, evidential reasoning, neural network) have been proposed and tested to various site-specific applications such as mineral potential mapping, geological hazard mapping and land-cover classification [1], [2], [3].

Though much research has been conducted to integrate multi-source spatial data, some operational issues still arise. Most geoscientific applications generally include heterogeneous types of data such as categorical and continuous data. In traditional spatial data integration methods, continuous data have been converted into categorized layers with several classes. By converting the continuous data to categorical data, however, so much information was lost. In relation to this problem, [4] incorporated the non-parametric density estimation procedure into the fuzzy logic framework. They reported that direct use of the continuous data could improve the prediction capability.

This paper aims to extend our previous fuzzy logic approach to probabilistic multi-source spatial data integration models. Within a probabilistic framework, four spatial data integration models are proposed and applied to landslide hazard assessment. The main focus is on the continuous data representation by adopting statistical relationships between landslide occurrences and input continuous data sets. The empirical likelihood ratio model based on Parzen window estimation, logistic regression, generalized additive model and predictive discriminant model are presented. A cross-validation approach based on random partitioning is applied to quantitatively assess and compare those four models. The effects of those proposed models are evaluated through a case study from the Gangneung area, Korea.

2. Multi-source Spatial Data Integration Models

1) Problem Formulation

Suppose that there are k spatial data E_i (i = 1, 2, $\cdots$, m) related to landslide occurrences for a specific future landslide type in a study area A. The k spatial data consist of m categorical data and n continuous data (k=m+n). In landslide hazard assessment, the target is that at each pixel p, it will be affected by future landslides, denoted by T_P.

In a probabilistic framework, our final goal is to compute the joint conditional probability at each pixel p, denoted by Prob{ $T_P | E_1, E_2, \cdots, E_k$ }. In this paper, instead of computing the joint conditional probability directly, the likelihood ratio function was used. Unlike a traditional conditional probability approach, the likelihood ratio function has the advantage of considering the relative risk [4].

The joint likelihood ratio λ at p, is defined as:

$$\lambda = \frac{\mathrm{Prob}\{E_1,E_2,\Lambda,E_k \mid T_p\}}{\mathrm{Prob}\{E_1,E_2,\Lambda,E_k \mid \overline{T}_p\}} \qquad (1)$$

where $\overline{T}_p$ denotes the proposition that at each pixel p, it will not be affected by future landslides.

If the conditional independence assumption is adopted, the joint likelihood ratio can be expressed as a product of the joint likelihood ratios of categorical and continuous data:

$$\lambda = \frac{\mathrm{Prob}\{E_1,\Lambda,E_m \mid T_p\}}{\mathrm{Prob}\{E_1,\Lambda,E_m \mid \overline{T}_p\}} \cdot \frac{\mathrm{Prob}\{E_1,\Lambda,E_n \mid T_p\}}{\mathrm{Prob}\{E_1,\Lambda,E_n \mid \overline{T}_p\}} \qquad (2)$$

Under the conditional independence assumption, the joint likelihood ratios can also be a product of the likelihood ratio of the bivariate likelihood ratio at each layer:

$$\lambda = \prod_{i=1}^{k} \lambda_i , \quad \lambda_i = \frac{\mathrm{Prob}\{E_i \mid T_p\}}{\mathrm{Prob}\{E_i \mid \overline{T}_p\}} \qquad (3)$$

For categorical data representation, a frequency ratio based method proposed by [4] can be adopted. Four models that will be discussed in the following subsections provide specific formulation related to the continuous data representation.

2) Empirical Likelihood Ratio Estimation using Parzen Window Estimation

In this model, the likelihood ratio of continuous data is estimated by using a Parzen window estimation approach. This approach computes the empirical frequency distribution functions in a non-parametric manner. The predefined kernel functions are centered at past landslide locations and the density estimations at each location are derived from the average contribution of each of the kernels at that location.

This paper adopts the kernel function of the Gaussian type, as defined by the following equation:

$$\mathrm{Prob}\{E_i \mid T_p\} = \frac{1}{\mathrm{N(L)}} \sum_{\alpha=1}^{\mathrm{N(L)}} \frac{1}{h\sqrt{2\pi}} \cdot \exp[-\frac{(E_i(X)-E_i(X_\alpha))^2}{2h^2}]$$

$$\mathrm{Prob}\{E_i \mid \overline{T}_p\} = \frac{1}{\mathrm{N(A)-N(L)}} \sum_{\alpha=1}^{\mathrm{N(A)-N(L)}} \frac{1}{h\sqrt{2\pi}} \cdot \exp[-\frac{(E_i(X)-E_i(\overline{X}_\alpha))^2}{2h^2}] \qquad (4)$$

where is $E_i(X_\alpha)$ a set of values of the continuous data layer at a landslide location X_α and $E_i(X)$ refers to a value of the continuous data layer at location. And h and $\overline{X}$ represent the spread parameter value of the Gaussian kernel and the locations of areas not affected by landslides, respectively. N(A) and N(L) also denote the total number of pixels in the study area and the number of pixels affected by past landslides, respectively.

3) Logistic Regression

Unlike traditional linear regression, logistic regression that models the logit of the response probability with a linear form is appropriate when the dependent variable (i.e. known landslide occurrences) is dichotomous, such as the occurrence or non-occurrence of landslides [5]. To estimate the regression coefficients, logistic regression generally uses maximum likelihood estimation, rather than least square estimation.

$$\lambda = \frac{\mathrm{Prob}\{E_1,\Lambda,E_n \mid T_p\}}{\mathrm{Prob}\{E_1,\Lambda,E_n \mid \overline{T}_p\}} = C \cdot \exp\{\beta_0 + \sum_{i=1}^{n} \beta_i E_i\} \qquad (5)$$

where C=(1-prior probability)/prior probability.

4) Generalized Additive Model

The generalized additive model extends generalized linear models by replacing the linear form with the additive one [5].

$$\lambda = \frac{\mathrm{Prob}\{E_1,\Lambda,E_n \mid T_p\}}{\mathrm{Prob}\{E_1,\Lambda,E_n \mid \overline{T}_p\}} = C \cdot \exp\{\beta_0 + \sum_{i=1}^{n} f_i(E_i)\} \qquad (6)$$

where C=(1-prior probability)/prior probability. f_i are unspecified smoothed functions for each of the predictors. This paper applied cubic B-splines to fit smooth relationships between the predictors and the response.

5) Predictive Discriminant Model

In this model, the frequency distribution functions are computed by assuming multivariate t-density distribution functions. This approach adopts a vague prior distribution for the unknown mean and covariance values and the estimated density is a weighted average of all members of the class of multivariate normal distributions [6]. Details of theoretical backgrounds can be referred to [6].

3. Case Study

1) Study Area and Data Sets

To assess and compare the proposed spatial data integration models, a case study for landslide hazard assessment was carried out for the Gangneung area, Korea. The study area had much landslide damage following typhoon RUSA and heavy rainfall early in September, 2002. Landslides triggered by intense rainfall resulted in both extensive damage to property and the loss of life.

To detect the locations of past landslides, multi-temporal high-resolution remote sensing images (i.e. IKONOS and QUICKBIRD acquired on 14 October, 2001 and 20 July, 2003, respectively) were used for change detection analysis. After applying a spectral normalization algorithm based on regression for reducing the spectral discrepancy caused by differences in acquisition dates the unsupervised change detection algorithm based on 3D block segmentation [7] was applied to obtain a map showing changed areas and non-changed areas. The landslide locations detected from change detection analysis were then verified by fieldwork and a total of 337 landslides were finally mapped (Fig. 1).

As for multi-source spatial data sets related to landslide occurrences, 5 data layers were considered in this study. For the categorical data sets, forest type and soil material maps were used. Continuous data derived from the DEM of the study area included elevation, slope and aspect maps.

2) Results

In the likelihood ratio estimation model, a value of 4% of data range of the spread parameter *h* in the Gaussian kernel functions was selected experimentally. To implement the logistic regression and generalized additive models, we used the S-PLUS statistical package.

After integrating all multi-source spatial data sets using formulas discussed in Section 2, final landslide hazard maps were generated wherein each pixel contains the hazard level measure mapped in the range of 0 to 200 (Fig. 2). This procedure was done to show the relative hazard levels in the study area. First, all the pixel values were sorted in descending order and the ordered pixel values were then classified per high rank 0.5%. This means that the lowest fuzzy membership value was mapped as 0 and the highest as 200.

To evaluate the prediction capabilities of the proposed models, a cross-validation approach based on spatial random partitioning was carried out. First, the past landslides were randomly divided into 2 disjoint groups (i.e. estimation and validation groups). The integrated maps were generated 2 times using the estimation group, each time with the remaining group held out as the validation group. Using rank order statistics, each integrated map was expressed in terms of relative landslide hazard values in the study area. Finally, prediction rate curves [2] were computed by comparing the hazard values with all past landslides. It should be noted that the prediction rate curves should be used to interpret the landslide hazard maps generated by all 337 past landslides shown in Fig.2.

By interpreting the prediction rate curves shown in Fig. 3, it is possible to quantitatively compare the proposed models. As the validation results, the prediction rates of the application of the empirical likelihood ratio estimation and predictive discriminant models were higher than ones by the logistic regression and generalized additive models. For the empirical likelihood estimation and predictive discriminant models, the most hazardous 10% of the area contains about 38% of future landslides (i.e. 128 landslides). Whereas, for the logistic regression and generalized additive models, about 32% of future landslides (i.e. 107) were contained in the most hazardous 10% class. The similar prediction capabilities of the likelihood ratio and predictive discriminant models means that the frequency distribution functions of three continuous data may be expressed in terms of both non-parametric (the empirical likelihood ratio estimation model) and parametric (the predictive discriminant models) forms. Whereas, the continuous data sets in the study area may not follow the assumption of linear relationships between the landslide occurrences and the data.

4. Conclusion

To effectively integrate multi-source spatial data, this paper presented four probabilistic integration models and applied to landslide hazard assessment. Those models can directly use the original continuous data without any categorizing procedure that results in loss of information. Each model proposed here adopts different assumptions and problem formulation, though it is theoretically based on the likelihood ratio function. This means that those models should be assessed by adopting a proper validation procedure. As the results from a case of the Gangneung area, Korea, the prediction capability of the empirical likelihood ratio estimation and predictive discriminant models were higher than those of the logistic regression and generalized additive models. However, these results do not indicate that those two models would always show the higher prediction capabilities. The integration models heavily depend on the data sets used. Thus, we will carry out more case studies to verify the results identified from this case study. Another future work will include the incorporation of socio-economic data into multi-source spatial data integration tasks for landslide risk analysis.

Acknowledgement

This work was supported by the Korean Ministry of Science and Technology.

References

[1] Bonham-Carter, G.F., F.P. Agterberg, and D.F. Wright, 1988. Integration of geological data sets for gold exploration in Nova Scotia, *PE&RS*, 54(11): 1585-1592.

[2] Chung, C.F. and A.G. Fabbri, 1999. Probabilistic prediction models for landslide hazard mapping, *PE&RS*, 65(12): 1389-1399.

[3] Benediktsson, J.A. and I. Kanellopooulos, 1999. Classification of multisource and hyperspectral data based on decision fusion, *IEEE Trans. Geosciences and Remote Sensing*, 37(3): 1367-1377.

[4] Park, N.-W., K.-H. Chi, and B.-D. Kwon, 2004. Application of fuzzy set theory for spatial prediction of landslide hazard, *Proc. IGARSS 2003*, Toulouse, France, CD-ROM publication.

[5] Hastie, T.J. and R.J. Tibshirani, *1990. Generalized additive models*, Chapman & Hall/CRS, NY.

[6] McLachlan, G., 1992. *Discriminant analysis and statistical pattern recognition*, John Wiley, NY.

[7] Yamamoto, T., H. Hanaizumi, and S. Chino, 2001. A change detection method for remotely sensed multispectral and multitemporal images using 3-D segmentation, *IEEE Trans. Geosciences and Remote Sensing*, 39(5): 976-985.

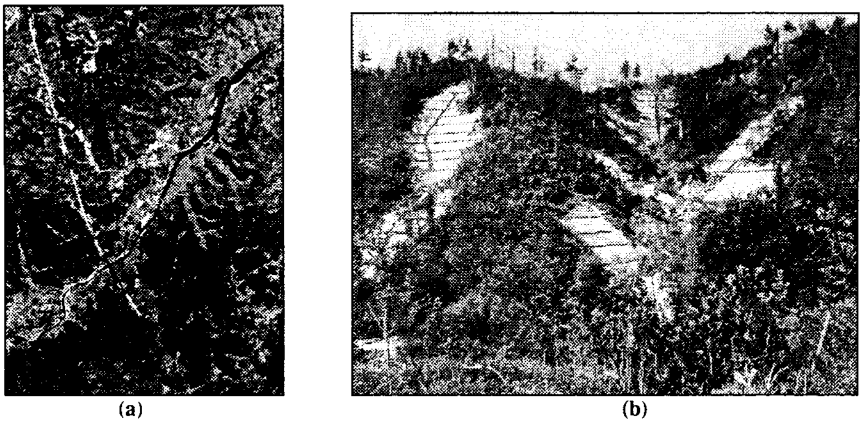

Fig 1. (a) Landslide locations draped over the QUICKBIRD imagery, (b) landslide scars that occurred in the study area.

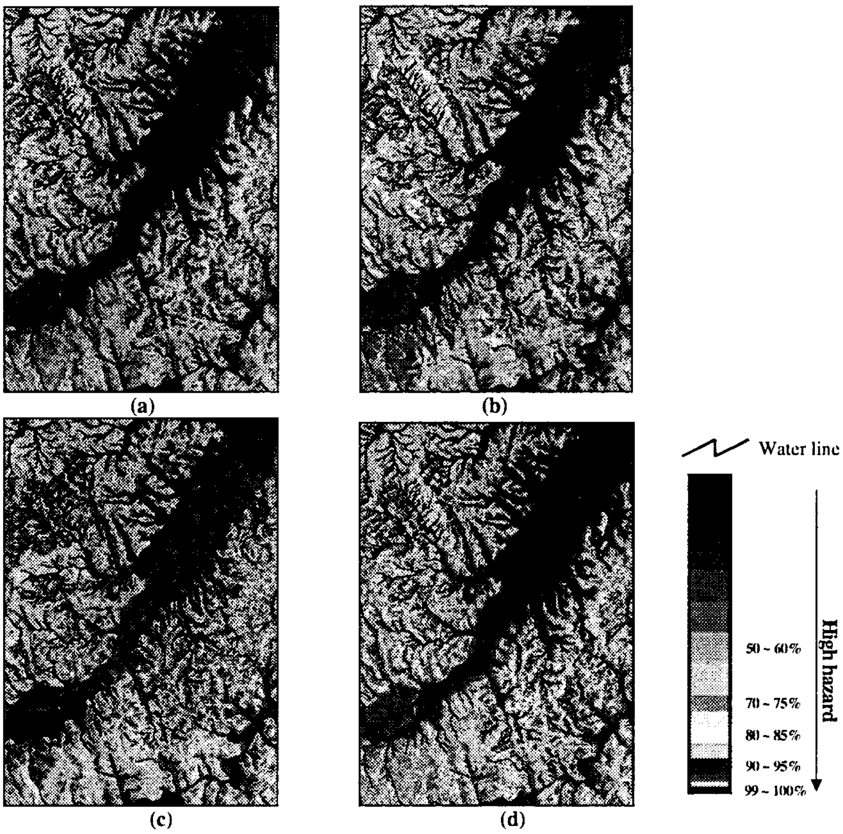

Fig. 2. Landslide hazard maps in the study area. (a) Empirical likelihood ratio estimation model, (b) Logistic regression model (c) Generalized additive model, (d) Predictive discriminant model.

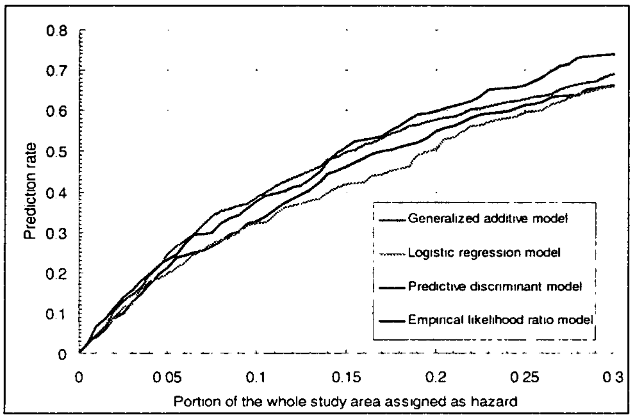

Fig. 3. Prediction rate curves.