A Priori Study of Using Spatial Data Mining Technology with FORMOSAT-2 Imagery for Analyzing Potential Landslide-causing Factors

Chia-Cheng Yeh
Department of Electrical Engineering, National Taipei University of Technology
Research Assistant, National Science and Technology Center for Disaster Reduction
E-mail: andrew@ncdr.nat.gov.tw
Pai-Hui Hsu
Department of Civil Engineering, National Taiwan University
Email: hsuph@ntu.edu.tw
Yang-Lang Chang
Department of Electrical Engineering, National Taipei University of Technology
E-mail: ylchang@mail.ntut.edu.tw
Tzu-Yin Chang
National Science and Technology Center for Disaster Reduction
Email: geoct@ncdr.nat.gov.tw

KEY WORDS: Spatial Data Mining, FROMOSAT-2, Landslide, Jhuo-shuei River Basin

ABSTRACT: Mountainous region, steep slope, broken terrain, and together with the frequent earthquakes and heavy rainfall are easily to trigger severe geological hazards such as large-scale landslides and debris flows in Taiwan. Series property damages and life losses caused by natural hazards can be reduced effectively if modern technology and knowledge are introduced into early warning systems. In this study, the FORMOSAT-2 imagery over Jhuo-shuei River basin acquired from 2006 to 2012 was employed to classify the landslide area. Spatial data mining technology was applied to calculate the weightings of various spatial factors for relevant landslide sites. In addition, spatial auto-correlation analysis and spatial auto-regression analysis were used to achieve a disaster hot spot analysis. The dataset of hot spot concentrated area over Jhuo-shuei River basin can provide to the disaster support system to achieve disaster early warning.

1. INTRODUCTION

In recent years, severe disasters, such as large-scale landslide and debris flow, happen unceasingly in Taiwan. These hillside disasters happen frequently because of overused and abused. In order to reduce losses caused by natural disasters, it is necessary to enhance disaster prediction ability and develop a disaster management system. Since several decades ago, land use/land cover change has been acquired, the models have been built, and a great number of disaster data from field investigation and remote sensing technology were collected after disasters happened. Moreover, various types of geographic and spatial data were also collected, including aerial and satellite images, digital elevation model (DEM), topographic maps, vegetation maps, geological maps, river maps and road systems to circle the dangerous regions and evaluate losses after the disasters. However, how to integrate the data and further effectively to predict disasters was still a critical and complicated issue. Recently, geographic and spatial data mining technology are fast developed to retrieve information from the massive data especially disaster hot zones indication is useful for supporting to make decisions. This study, the spatial data mining technology was applied to evaluate the weightings of various spatial factors for disaster sites and to circle the hot spot zones over Jhuo-shuei River basin in Taiwan.

2. METHODOLOGIES

Factors in the occurrence of landslides can be divided into two types – potential factors and triggering factors. Potential factors represent internal environmental factors, such as the geological structure and terrain characteristics of hillsides. Triggering factors are external factors, such as typhoons, extreme rain, earthquakes. When triggering factors' magnitude excesses the threshold, the disasters will happen.

In previous studies, all sorts of spatial and geographic parameters were selected to apply in the model to find the critical conditions for landslide disasters. The mechanism of collapse is very complicated, not only the selection of collapse factors but also the employed methods and models influence the predictions of collapse. If physical models are applied to prediction, physical mechanisms should be used to establish physical models in advance, and assumptions and simplification inevitably influence the final prediction results. However, the non-physical models, such as statistic methods or data mining methods, are directly used to find out the critical conditions of collapse in order to predict landslide occurrence.

Data mining mothed means that it can retrieve unknown knowledge from massive data. When data mining is applied to data with spatial characteristics, it is called spatial data mining.

Spatial data mining can be divided into five categories (Ester, 1999):

1. Spatial Characterization Analysis: A spatial characterization analysis is to select suitable relevant data from an archive and establish spatial or non-spatial description. For instance, to analyze water quality, a spot with bad water quality is first selected, the region with bad water quality is then estimated, and eventually, possible characteristic factors are induced.
2. Spatial Discriminate (Outlier) Analysis: A spatial discriminate (Outlier) analysis indicates comparing the characteristics of a target and the contrast objects and discovering the boundaries of distinction.
3. Spatial Classification Analysis: In a spatial classification analysis, spatial data is classified by a certain feature or is used to find the causes. It is a method of supervisory categorization.
4. Spatial Clustering: Spatial data is categorized according to the similarity in order to increase the heterogeneity between different categories. It is a non-supervisory classification.
5. Spatial Association Rules: This analysis aims to discover the association and rules among massive spatial data.

In addition to the aforementioned classifications, data mining can be simply divided into two types according to purposes. One type is predictive data mining that establishes models which can predict future phenomena, styles, or changes in trend, and the other type is descriptive data mining, which describes data and phenomena and finds the relations and styles between them. It was expected in this investigation that predictive spatial data mining can be applied to the prediction of landslide and collapse destinations.

3. SPATIAL AUTOCORRELATION

The index that is most frequently used to calculate global spatial autocorrelation is Moran's I. The calculation is as follows (Moran, 1950):

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{s^2 \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}} \qquad (1)$$

Weight w_{ij} in Formula (1) represents the spatial relationship between Feature i and Feature j . Generally, it is assumed that $w_{ij}$and the distance between features are in an inverse ratio. That is, the farther the distance between two features is, the lower the correlation is. From the formula of Moran's I, it was found that the numerator is the product sum of the attribute value and average difference of neighboring features. In other words, *I* will be greater than 0 if the attribute values of Feature and Feature are both higher than the average value or lower than the average value, indicating that the neighboring areas own similar data attributes, namely clustering. If *I* is lower than 0, it indicates that the attributes of neighboring features are very different, that is, the spatial distribution of features is uneven. If *I* is close to 0, the correlation between neighboring features is low, namely the attribute values of neighboring features distribute randomly and irregularly.

Through Moran's I, only the similarity of the attribute values of spatial features can be obtained, and it is impossible to know whether the attribute values of the features in a cluster are simultaneously large or small. Therefore, it can not be applied to hot spot or cold spot analysis. Getis and Ord (1992) thus brought up statistic method called Getis, or G-Statistic Index. The formula for calculating General G, namely global spatial autocorrelation indices, is listed as follows:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij} x_j - \bar{x}\sum_{j=1}^{n} w_{ij}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{ij}^2 - \left(\sum_{j=1}^{n} w_{ij}\right)^2\right]}{n-1}}} \qquad (2)$$

$$\bar{x} = \frac{\sum_{j=1}^{n} x_i}{n} \qquad (3)$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \bar{x}^2} \qquad (4)$$

$\mathrm{G_i^*}$ is used to conduct a hot spot analysis, and a distance range, namely d, is given. Afterward, Weight $\mathrm{w_{ij}}$, which is used to describe the spatial relation between features $\mathrm{i, j}$, can be set as an inverse ratio with the square of distance. That is, the farther a distance is, the smaller the spatial correlation is. When Feature i 's $\mathrm{G_i^*} > 0$, and $\mathrm{G_i^*}$ is larger, it indicates that clusters occur, and the attribute value is high. It is called hot spot or hot zone. On the contrary, if $\mathrm{G_i^*} < 0$, and $\mathrm{G_i^*}$ is smaller, it indicates that clusters occur, but the attribute values are low. It is called cold spot or cold zone.

4. SPATIAL REGRESSION

When traditional regression models are used to analyze the features of spatial data, it is often assumed that spatial data are mutually independent. However, data with spatial features are usually mutually influential, which is called spatial effect. Anselin (1988) divided spatial effect into two types, that is, spatial dependence and spatial heterogeneity. Spatial dependence is spatial autocorrelation, which was described in the previous section. For example, phenomena observed in some spatial locations are probably caused by phenomena generated in another location. Furthermore, spatial heterogeneity indicates that spatial effect is not consistent. For instance, the seriousness of collapse varies according to the characteristic differences of regions. Consequently, in terms of the display of econometric models, the parameters of models or the models themselves may vary according to different spatial locations.

When phenomena related to space are discussed, if spatial influences are not considered, wrong assumptions of model variables and errors may cause model parameters to be misjudged, and they may further cause the subsequent analyses or predictions to be biased. To consider the relation between variables and other variables in the neighboring space as well as the relation between errors and the errors in other spatial locations, the fundamental formulas of spatial regression models are generally defined as follows:

$$\mathrm{Y} = \rho \cdot (w_1 \cdot Y) + \beta \cdot \mathrm{X} + \mu \tag{5}$$

$$\mu = \lambda \cdot (w_2 \cdot \mu) + \varepsilon \tag{6}$$

$$\varepsilon \sim N(0, \sigma_\varepsilon^2 I_n) \tag{7}$$

5. EXPERIMENTS

Jhuo-shuei River Basin is located in the mid-western Taiwan, and it originates from Sakuma Touge between the main peak and east peak of Hehuanshan, or Joy Mountain, among Central Mountains. The main stream is approximately 186.6KM in total, which makes it the longest river in Taiwan. The average gradient of the river is 1/55 from the mountains which the river originates from and the estuary. The basin is approximately 3156.9 KM2 (the ground whose elevation is less than 100M only accounts for 8.37% of the basin.), so it is the second largest in Taiwan in terms of river basin. The average height of the entire basin is about 1422M (Water Resources Agency, Ministry of Economic Affairs, 2011). The administrative areas in the basin include four counties, namely Nantou County, Chiayi County, Changhua County, and Yunlin County, and twenty-one townships. The average topographic height of the entire basin is approximately 1,422M. The basin is illustrated with Figure 1.

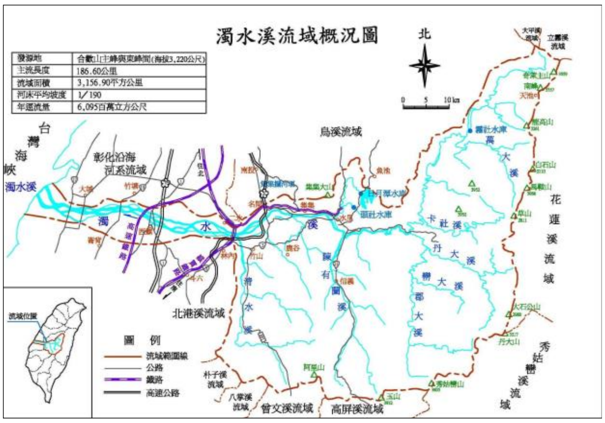
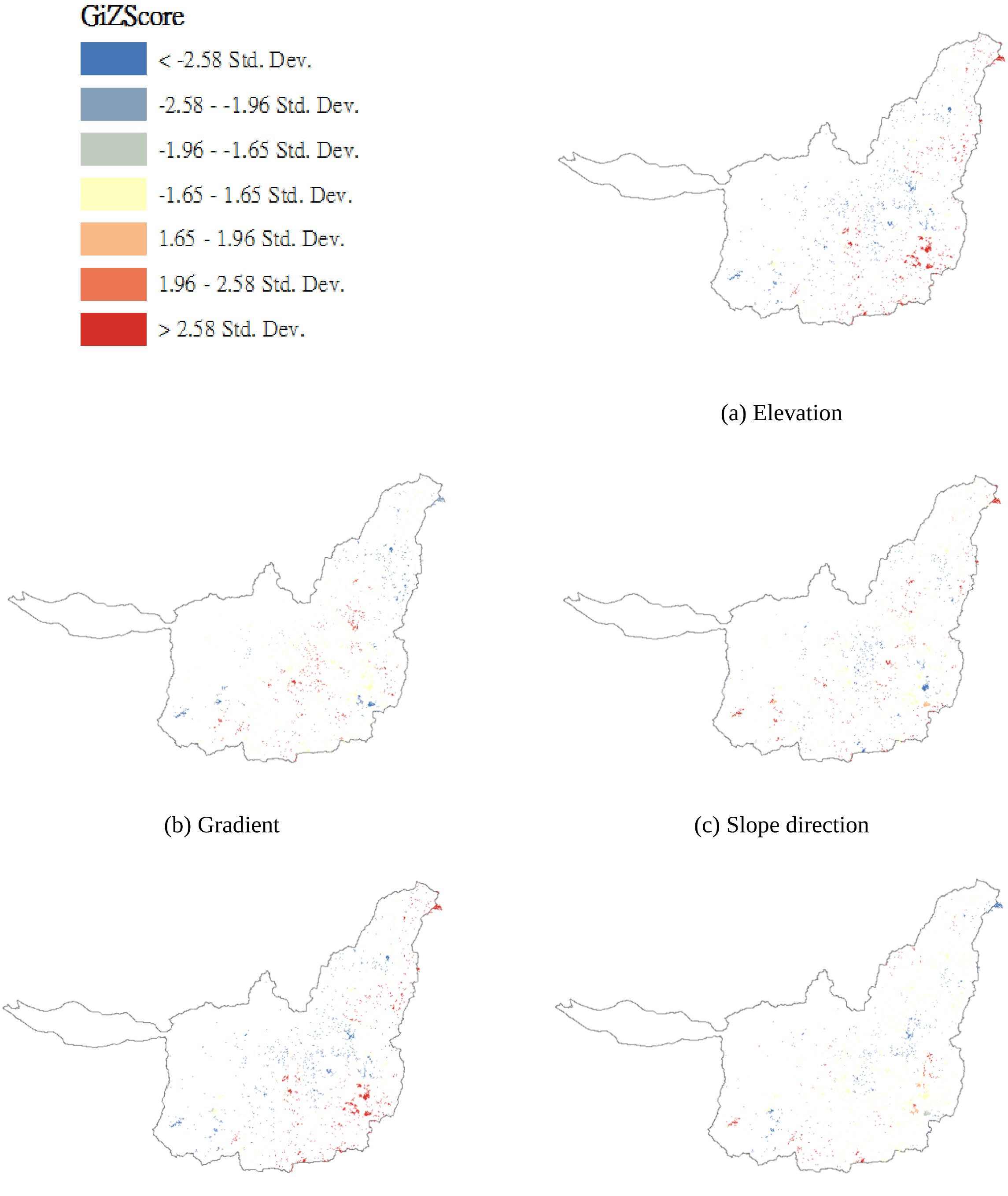


Figure 1. The test area

The midstream and upstream of Jhuo-shuei River Basin are geologically broken, so there is collapse easily. Moreover, the river transports a great deal of sand, and the difference between the high flow and low flow is huge, so the fluvial phases and water features vary significantly as the gradients change.

Hot spots and cold spots can be analyzed through G-Statistic. A so-called hot spot means that there is a clustering phenomenon, and the attribute value of the area is higher than the average attribute value. On the other hand, if there is a clustering phenomenon in an area, but the attribute value is lower than the average attribute value, the area is called a cold spot. Figure 2 shows the analysis result of the hot spot of each disaster factor. From the figure, it is found that hot spots and cold spots more or less exist in each attribute. This not only displays a clustering phenomenon but also means that the attribute value of the area is an extreme value. Characteristic areas obtained from a regional Moran 'I analysis are possible only the regional extreme values of an attribute. Nevertheless, the hot spot and cold spot obtained from G-Statistic are the absolute extreme values of the attribute.



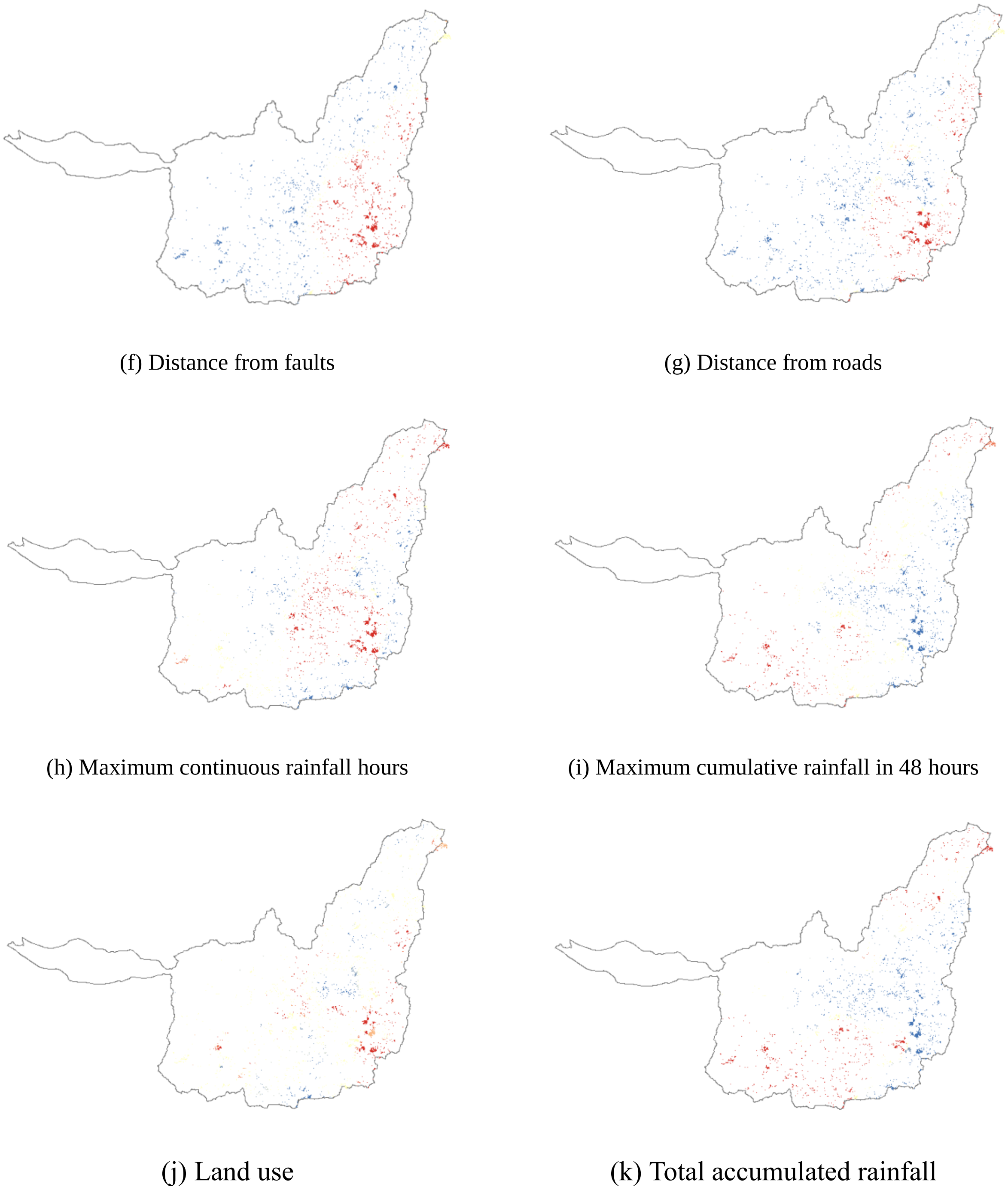
(a) Elevation

(b) Gradient

(c) Slope direction

(d) Lithology

(e) Historic landslide



(f) Distance from faults

(g) Distance from roads

(h) Maximum continuous rainfall hours

(i) Maximum cumulative rainfall in 48 hours

(j) Land use

(k) Total accumulated rainfall

Figure 2.The results of hot spot analysis

6. CONCLUSION

This study was aimed to use data mining technology to discover the critical rainfall of collapse from currently available massive spatial data in hope of combining the result with rainfall forecast to achieve collapse warning and effectively reduce life and property losses caused by disasters. Currently, there have been a great number of investigations on landslide and collapse, but spatial dependence has been rarely considered. When there is dependence among data, but the assumptions are independent, it will result in biased estimates in regression results.
Spatial dependence among spatial data should not be ignored, so spatial data mining is performed, spatial dependence should be accurately considered to effectively increase the outcome of spatial data mining. In addition, it is obvious in the cluster analysis of a single factor that the key factors and critical conditions of collapse can not be discovered through a single factor, indicating that the relationship between the mechanisms and factors of collapse is complicated,

so a more advanced approach should be employed, and, meanwhile, it should be considered applying different factors to the analysis.

7. REFERENCES

Dilley, M., Chen, R. S., Deichmann, U., Lerner-Lam, A. L., Arnold, M., Agwe, J., Buys, P., Kjekstad, O., Lyon, B. and Yteman, G. , 2005, Natural Disaster Hotspots: A Global Risk Analysis, Disaster Risk Management Series No.5, The World Bank, Washington, D. C..

Hsu, P.-H., Wu, S.-Y. and Lin, F.-T. , 2005, Disaster management using GIS technology: A case study in Taiwan, Proc. ACRS 2005, Hanoi, Vietnam.

Su, W. R., Hsu, P. H., Wu, S. Y., Lin, F. T. and Chou, H. C. , 2010, Development of Safe Taiwan Information System (SATIS) for typhoon early warning in Taiwan, Journal of Systemics, Cybernetics and Informatics 8(4), pp.48-52.

M. Yuan, 1998, Representing Spatiotemporal Processes to Support Knowledge Discovery in GIS databases. In T. K. Poiker and N. Chrisman (eds.), Proceedings: 8th International Symposium on Spatial Data Handling Spatial Data Handling, pp. 431-440.

M. Gahegan, M. Wachowicz, M. Harrower, and T.M. Rhyne, 2001, The integration of geographic visualization with knowledge discovery in databases and geocomputation. Cartography and Geographic Information Systems, special issue on the ICA research agenda.

H.J. Miller, and J. Han, 2009, Geographic data mining and knowledge discovery: An overview, In H.J. Miller and J. Han (eds) Geographic Data Mining and Knowledge, 2nd ed., CRC Press, Taylor & Francis Group