Jour. Korean Earth Science Society, v. 26, no. 2, p. 114-128, March 2005

# Application of Fuzzy Information Representation Using Frequency Ratio and Non-parametric Density Estimation to Multi-source Spatial Data Fusion for Landslide Hazard Mapping

**No-Wook Park[1]*, Kwang-Hoon Chi[1] and Byung-Doo Kwon[2]**

[1]Geoscience Information Center, Korea Institute of Geoscience and Mineral Resources,
30 Gajeong-dong, Yuseong-gu, Daejon 305-350, Korea
[2]Department of Earth Science Education, Seoul National University,
San 56-1, Shillim-dong, Kwanak-gu, Seoul 151-748, Korea

**Abstract:** Fuzzy information representation of multi-source spatial data is applied to landslide hazard mapping. Information representation based on frequency ratio and non-parametric density estimation is used to construct fuzzy membership functions. Of particular interest is the representation of continuous data for preventing loss of information. The non-parametric density estimation method applied here is a Parzen window estimation that can directly use continuous data without any categorization procedure. The effect of the new continuous data representation method on the final integrated result is evaluated by a validation procedure. To illustrate the proposed scheme, a case study from Jangheung, Korea for landslide hazard mapping is presented. Analysis of the results indicates that the proposed methodology considerably improves prediction capabilities, as compared with the case in traditional continuous data representation.

Keywords: information representation, fuzzy logic fusion, density estimation, landslide hazard

## Introduction

During the last decade, advance in data acquisition techniques and state-of-the-art of computer technologies such as GIS have made multi-source data more available for geological applications. Since most geological phenomena are representatives of the combined results of various physical parameters or variables, it is reasonable to analyze multi-source data in integrated ways in order to address complex geological problems. This integrated analysis is known as "data fusion" and has attracted the attention of geologists, geographers and other scientists interested in the extraction of more information with higher accuracy and less uncertainty.

Since the late 1980s, several methods designed for multi-source geological data fusion (e.g. Bayesian probabilistic models, fuzzy logic, evidential reasoning) have been proposed. They are theoretically based on mathematical geology, statistics and even artificial intelligence, and have been tested for various site-specific applications such as nonrenewable resource exploration and geological hazard mapping (An et al., 1991; Bonham-Carter et al., 1988; Chung and Fabbri, 1993; Moon, 1990).

Among these methods, fuzzy logic can provide a tool for modeling knowledge from incomplete data sets and developing more reliable decision-making processes (Zadeh, 1965; Zimmerman, 1996). Although much research has shown the applicability of the fuzzy logic fusion model for geological applications (An et al., 1991; Carranza and Hale, 2001), some operational issues still arise, such as the assignment of fuzzy membership function values, or the selection of optimum fuzzy combination operators. As discussed in Moon (1998), those issues are dependent on the data sets used and the target proposition adopted. As with other data fusion methods, it is very important to quantitatively relate the quantitative interrelationship between the target occur-

*Corresponding author: nwpark@kigam.re.kr
Tel: 82-42-868-3287
Fax: 82-42-867-0421

rence data and input data to the target proposition. Most research into mineral potential mapping has constructed fuzzy membership functions through the application of an expert's reasoning and deposit models. This target-driven approach (Bonham-Carter, 1994) can be applied in a case where sufficient target occurrence data are not available. As the fuzzy membership function values are determined by the modeler based on her/his knowledge and experience, this approach is somewhat subjective. On the contrary, a data-driven approach is based on the quantitative relationships between input data sets and the target proposition. If sufficient target occurrence data are available, a more objective assignment of the fuzzy membership function values can be achieved when compared with the case of the target-driven approach (Park et al., 2003a).

Spatial data fusion tasks in geological applications generally include the different types of data such as categorical data that consist of mutually exclusive and specific labels, and continuous data that have no particular category label attached to any particular data value and take ratio, interval values for attributes. For example, the commonly used categorical data set for landslide hazard mapping includes the geology, forest and soil maps. As for the continuous data set, the elevation, slope and aspect maps are commonly used. Most researchers first convert the continuous data set into some categorized classes. Binary or multi-class representation is, however, inappropriate for continuous data, since it requires optimal discretization. In consequence, distortion and loss of valuable information could not be avoidable.

This paper aims to apply a data-driven fuzzy logic fusion scheme, which includes the information representation methods for the continuous data set, to landslide hazard mapping. The information representation method adopted in this paper came from a personal communication with Dr. Chang-Jo F. Chung of Geological Survey of Canada. It is based on the ratio of frequency distribution functions. However, detailed procedures for the assignment of fuzzy membership function values differ between categorical and continuous data sets. For the representation of the categorical data set, the frequency ratio based approach is applied. The Parzen window estimation, which is one of the several non-parametric density estimation methods, is applied to the representation of the continuous data set. The effects of those data representation methods are then evaluated through a case study of landslide hazard mapping undertaken for Janghe-ung, Korea.

## Methods

### Fuzzy set theory

Fuzzy set theory as proposed by Zadeh (1965) is conceptually different from conventional crisp set theory. In the conventional crisp set theory, one element either belongs to a set or it does not. Alternatively, fuzzy sets are defined as sets that do not have a crisply defined membership, but rather allow objects to have grades of membership from 0 to 1 (Zimmerman, 1996).

If $X=\{x\}$ denotes a universe of the attribute values, the fuzzy set A in the X is the set of ordered pairs:

$$A = \{x, \mu_A(x)\}, x \subset X \quad (1)$$

where $\mu_A(x)$ is known as the grade of membership of $x$ in the A.

Usually, $\mu_A(x)$ is an integer or a floating number in the range [0,1] with 1 representing full membership and 0 non-membership. The grade of membership reflects a kind of ordering that is not based on probability but on possibility. The value of $\mu_A(x)$ for the attribute value $x$ in A can be interpreted as the degree of compatibility of the predicate associated with set A and attribute value $x$.

In the framework of fuzzy set theory, multi-source spatial data fusion aims at combining or integrating the membership functions, in order to obtain a single fused membership value. Each data layer of target information denoted from fuzzy set theory can be integrated by using fuzzy combination operators.

More detailed theoretical backgrounds and operational issues for geological applications can be found in Moon(1998) and Park et al. (2003a).

### Information representation

The information representation step can be regarded as transforming spatial data into information with respect to a chosen target proposition. Within a fuzzy set framework, the certainty that the target proposition is true is based on the conceptual idea of expressing landslide hazard in terms of the fuzzy membership function.

In order to construct the quantitative model for future landslide hazard, some assumptions are required. Two assumptions adopted in this study are as follows. First, the characteristics of future landslides are the same as those of past landslides under similar circumstances. Second, the input causal factors provide useful information for landslide hazard mapping. In other words, information in input data sets is sufficiently representative of the typical conditions required for landslide occurrences.

In practice, the construction of fuzzy membership functions greatly depends on the problem to be solved. Furthermore, it is also very difficult to construct certain types of semantic model such as a linear, or bell-type, especially for categorical data. In this paper the fuzzy membership functions are constructed using an empirical estimation of the frequency ratio, instead of a certain type of semantic model. In contrast to the traditional fuzzy membership function representation, the functions are separately constructed for categorical and continuous data. Fortran programming and the Spatial Prediction Modeling System by SpatialModels Inc. was used for implementation of the method. Detailed description of the system can be referred to users guide of spatial prediction modeling system (SpatialModels Inc., 2004).

### Categorical data representation using frequency ratio

For categorical data representation, a frequency ratio based method is adopted to highlight the differing characteristics of those hazardous areas affected by landslides and the unaffected non-hazardous regions. Thereafter, the frequency ratio is converted to the fuzzy membership function.

A simply way of categorical data representation is to use a portion of landslide occurrence data in a particular class attribute. Good spatial data should sufficiently separate hazardous and non-hazardous areas. This implies that input spatial data should effectively reveal the different characteristics between the hazardous areas affected by landslides, and the non-hazardous, unaffected areas. When the two frequency distribution functions of the hazardous and the non-hazardous areas are compared, possible quantitative measures, which can highlight the contrast, may be a differencing or a ratio. A difference in the proportions of fixed size may have greater importance when both proportions are close to 0 or 1 than when they are near to the middle of the range (Agresti, 1990). For example, the difference between 0.010 and 0.001 may be more noteworthy than the difference between 0.510 and 0.501. In such a case, the ratio of proportions is a more useful descriptive measure than differencing.

Suppose that the spatial database includes $m$ spatial data related to landslide occurrences for a specific future landslide type in a study area A. Each layer of spatial data is regarded as evidence $E_i$ ($i$ = 1, 2, ···, $m$) for the target proposition such as "At each pixel p, it will be affected by future landslides", denoted by $T_p$.

For the $E_{ij}$, which is the $j$th class attribute of the evidence $E_i$, frequency distribution functions of the hazardous $(F(T_p)|_{E_{ij}})$ and that of the non-hazardous sub-areas $(F(\overline{T}_p)|_{E_{ij}})$ are defined as:

$$(F(T_p)|_{E_{ij}}) = \frac{N(L \cap E_{ij})}{N(L)} \tag{2}$$

$$(F(\overline{T}_p)|_{E_{ij}}) = \frac{N(E_{ij}) - N(L \cap E_{ij})}{N(A) - N(L)}$$

where $N(L \cap E_{ij})$ is the number of pixels in

landslides that occurred in $E_{ij}$. $N(L)$ is the total number of all landslides. $N(E_{ij})$ is the number of pixels in $E_{ij}$, and $N(A)$ is the number of pixels in the whole study area. The $\overline{T}_p$ denotes the proposition that at each pixel p it will not be affected by future landslides.

The frequency ratio $R_{ij}$ of above two frequency distributions is defined as:

$$R_{ij} = \frac{F(T_p)_{ij}}{F(\overline{T}_p)_{ij}} \tag{3}$$

The more the frequency ratio exceeds 1, the stronger the relationship between two patterns will be. The frequency ratio ranges from zero to infinity. However, the fuzzy membership function value should be a number in the range [0,1] with 1 representing full membership and 0 non-membership. To rescale the frequency ratio to the range [0,1], the three conditions below are assumed:

(1) The ratio value of 0 means non-membership with respect to the target proposition. Thus, the value should be returned a membership function value of 0.

(2) The ratio value of 1 means the independent relationship between data and the target proposition. By assuming that 0.5 is the neutral fuzzy membership value, the transformation procedure returns a membership function value of 0.5 for a ratio value of 1.

(3) A very large ratio value that converges to infinity, which means a strong relationship between data and the target proposition, returns the full membership value of 1.

To satisfy the above three conditions, the following relationship was used (SpatialModels Inc., 2004, p. 33):

$$\mu_{ij} = \frac{R_{ij}}{1 + R_{ij}} \tag{4}$$

where $\mu_{ij}$ is the fuzzy membership value of $E_{ij}$.

## Continuous data representation using Parzen window estimation

As for continuous data, the same ratio based representation method is adopted. However, original continuous data are used directly for the construction of fuzzy membership functions, without their conversion into categorized data. This procedure can prevent the loss of original meaning to the continuous data.

To generate a frequency distribution function from data, the two possible approaches are parametric and non-parametric. To compute the function in a parametric manner, certain statistical assumptions are made. The maximum likelihood technique is based on the assumption that the digital values to be processed are multi-dimensional and normally distributed. However, as many researchers have commented, spatial data, including those of geology, often show a skewed distribution or have multi-modal densities, and it is more difficult to construct a common statistical model in multi-source geological data fusion than in single source data processing.

An alternative way is to adopt a non-parametric approach that can be used with arbitrary distributions and without the assumption that the forms of the underlying densities are known (Duda et al., 2000). A simple non-parametric method is to produce a histogram, a bar chart in which a continuous variable is divided into discrete categories. The number of observations that fall into each category is represented by the area of the corresponding bar (Davis, 1986). The histogram is useful for showing discrete distributions; however, it is not employed in this study, despite its simplicity, because the interest in this paper is in the direct use of continuous data.

A Parzen window estimation approach, also called a kernel density estimate (Parzen, 1962) has been used herein to compute the empirical frequency distribution function in a non-parametric manner. This approach provides a consistent estimate of the related empirical frequency distribution function by using an appropriate kernel function (Fukunaga, 1990; Silverman, 1986).

In the case being considered, the predefined ker-

nel functions are centered at landslide locations, and the density estimations at each location are derived from the average contribution of each of the kernels at that location (Fig. 1). The shape of the distribution using the method depends on the shape of the kernel function. This study adopts the kernel function of the Gaussian type, as defined by the following equation:

$$F(\mathrm{T_p})_{E_i} = \frac{1}{\mathrm{N}(L)} \sum_{\alpha=1}^{\mathrm{N}(L)} \frac{1}{\mathrm{h}\sqrt{2\pi}} \exp\left[ -\frac{(E_i(\mathrm{X}) - E_i(\mathrm{X}_\alpha))^2}{2\mathrm{h}^2} \right]$$

$$F(\overline{\mathrm{T}}_\mathrm{p})_{E_i} = \frac{1}{\mathrm{N}(A) - \mathrm{N}(L)} \sum_{\alpha=1}^{\mathrm{N}(A)-\mathrm{N}(L)} \frac{1}{\mathrm{h}\sqrt{2\pi}} \exp\left[ -\frac{(E_i(\mathrm{X}) - E_i(\overline{\mathrm{X}}_\alpha))^2}{2\mathrm{h}^2} \right] \quad (5)$$

where $\{E_i(\mathbf{X}_\alpha),\ \alpha = 1, 2, \cdots \mathrm{N}(L)\}$ is a set of values of the continuous evidence $E_i$ at landslide location $\mathbf{X}_\alpha$ and $E_i(\mathbf{X})$ refers to a value of the continuous evidence $E_i$ at location $\mathbf{X}$. $h$ and $\overline{\mathrm{X}}$ also represent the spread parameter value of the Gaussian kernel and the locations of areas not affected by landslides, respectively.

The shape of the kernel depends only on the spread parameter ($h$) that controls how much to smooth. The extent of smoothing selected is of crucial importance in density estimation. If $h$ is too small, the estimate will be very spurious. If, on the other hand $h$ becomes large, important spatial variation may be lost, and all detail will be obscured due to over-smoothing. In the considered case, the main objective in applying the Parzen window approach is to determine how well the empirical frequency distributions can reflect quantitative relationships between continuous data and landslides, not the choice of an optimum value of $h$. Thus, various values of $h$ are tested, and the value of $h$ which has the optimum prediction powers or least uncertainties is determined through a validation procedure.

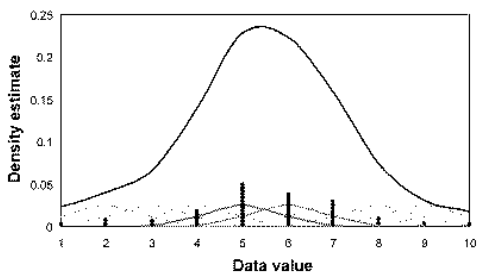


**Fig. 1.** An example of Parzen window estimation. The black dot denotes the frequency of each data value (modified from Hastie et al., 2001).

Once the two empirical frequency distributions of landslide areas and non-landslide areas are obtained, the fuzzy membership functions are computed by using Equation (4).

## Case Study

### Study area and data sets

To highlight the proposed data-driven fuzzy logic fusion scheme, a case study for landslide hazard mapping was carried out for the Jangheung area, Korea. The northern part of Kyeonggi province, including the study area, suffered much landslide damage due to intense rainfall that occurred between August 4 and 9, 1998. During that period, the maximum daily rainfall was about 620 mm and more than 20,000 ha of cultivated area and many houses were flooded. Landslides resulted in the loss of life of about 170 people (Kim, 2001). In the study area, the landslides were mainly flows that occurred during or shortly after three to four hours of high intensity rainfall.

The study area lies between 37°43'N and 37°46'N, and 126°56'E and 127°01'E, and covers approximately 37.29 km$^2$. Since the bedrock lithology of the district mainly consists of gneiss, and most land-

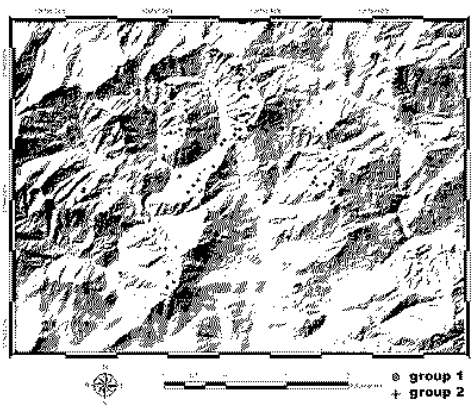


**Fig. 2.** Landslide scarp distributions draped over the shaded relief map of the study area. The symbols do not represent the actual size of the scarp but are enlarged for visibility.

slides occurred in that area, only gneissic areas were considered. As a result, 1,491,443 pixels with a size of 5 m by 5 m constitute the database. Five multi-source spatial data layers and one landslide occurrence layer are considered for this case study. For the categorical data set, forest type and soil maps were used. The continuous data set derived from the DEM of the area includes elevation, slope and aspect maps.

An inventory of landslides was examined through a systematic analysis of panchromatic satellite remote sensing images. Two IRS-1C images and one KOMPSAT-1 EOC image, acquired on 5 June 1998, 12 October 1998 and 2 February 1999, respectively, were used in change detection analysis. Two IRS-1C images were used for unsupervised change detection analysis (Park et al., 2003c) and the KOMPSAT-1 EOC image was used for additional information extraction. The landslide locations detected from remote sensing images were then verified by fieldwork and a total of 332 landslides were finally mapped. The 332 landslides were then divided into two random groups for validation analysis (Fig. 2). The topographically highest points of the landslide scars were regarded as trigger areas (i.e. scarp). One group of 166 landslides (252 pixels) was used as the estimation data set to construct quantitative relationships between the landslides and the input data set, and also used to generate the landslide hazard map. The landslide hazard map based on those relationships was then evaluated by comparing the susceptibility levels with the distribution of the other validation group of 166 landslides, assuming that the landslides had not yet occurred.

## Results

The frequency ratio values for each of the categorical data were estimated first, using a quantitative relationship between past landslides and the categorical data.

After calculating the frequency ratio values for all categorical data, the fuzzy membership functions that account for relative landslide hazard levels were constructed. To convert the frequency ratio to the fuzzy membership function ranging from [0,1], the conversion rule expressed by equation (4) was followed. To avoid numerical instabilities in the fuzzy combination step, the fuzzy membership function value for the frequency ratio value of 0 was set to 0.001.

Table 1 shows the relationships between landslides and categorical data such as the forest type and soil maps. As discussed before, if the ratio value is greater than one, it means a higher association. A value less than one means a lower association. In the map of forest type, the frequency ratio was the highest for Korean nut pine. This result is related to the amount of roots and their ability to maintain the water and soil pore pressure. In the soil map, the frequency ratio value was higher in the soil IV and V categories. According to the increase in rate, the soil has little moisture and has a low thickness. This is related to an increase of unit weight and of shear stress of soil due to pore-water increase.

For continuous data on elevation, slope, and aspect, four different $h$ values: 0.5%, 2%, 4% and 8% of the total range of data for the Gaussian ker-

**Table 1.** Frequency ratio and resultant fuzzy membership values of categorical data

| Data | Class | Landslide occurrence (a) | Portion of (a) | Remaining pixels (b) | Portion of (b) | Frequency ratio value (a/b) | Fuzzy membership value |
|---|---|---|---|---|---|---|---|
| Forest type | Non-forest | 2 | 0.008 | 162,501 | 0.109 | 0.073 | 0.068 |
| | Broad leaf tree | 78 | 0.310 | 635,000 | 0.426 | 0.727 | 0.421 |
| | Mixed broad leaf tree | 13 | 0.052 | 220,845 | 0.148 | 0.348 | 0.258 |
| | Korean nut pine | 156 | 0.619 | 332,171 | 0.223 | 2.779 | 0.735 |
| | Larch | 0 | 0.000 | 53,297 | 0.036 | 0.000 | 0.001 |
| | Rigida pine | 1 | 0.004 | 33,592 | 0.023 | 0.176 | 0.150 |
| | Pine | 0 | 0.000 | 2,575 | 0.002 | 0.000 | 0.001 |
| | Bare land | 2 | 0.008 | 51,160 | 0.034 | 0.231 | 0.188 |
| Soil | II | 0 | 0.000 | 51,546 | 0.035 | 0.000 | 0.001 |
| | III | 86 | 0.341 | 626,213 | 0.420 | 0.813 | 0.448 |
| | IV | 145 | 0.545 | 532,224 | 0.357 | 1.612 | 0.617 |
| | V | 19 | 0.075 | 61,025 | 0.041 | 1.842 | 0.648 |
| | Rocks | 0 | 0.000 | 36,155 | 0.024 | 0.000 | 0.001 |
| | Cultivated land | 0 | 0.000 | 6,538 | 0.004 | 0.000 | 0.001 |
| | Etc | 2 | 0.008 | 177,576 | 0.119 | 0.067 | 0.062 |

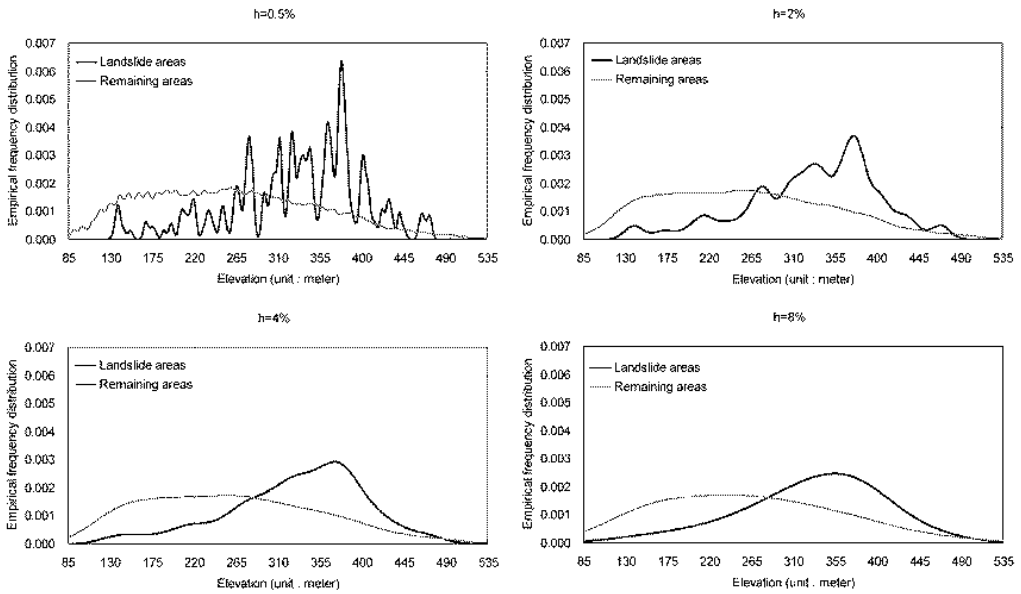


**Fig. 3.** Empirical frequency distributions of the elevation map with respect to four different values of $h$.

nel function, were tested for non-parametric density estimation. The two empirical distribution functions and the frequency ratio values of the elevation map are shown in Fig. 3. According to an increase of the value of $h$, the kernel density estimate tends to be smoothed and artifices of the data disappeared as expected. Those smoothing effects also resulted in impacts of the decrease of the maximum value.

Using those two empirical distribution functions, the frequency ratio values and the fuzzy membership functions were computed and are shown in Fig. 4. For comparison, the original continuous data were converted into categorized continuous data and then their frequency ratio and resultant fuzzy mem-

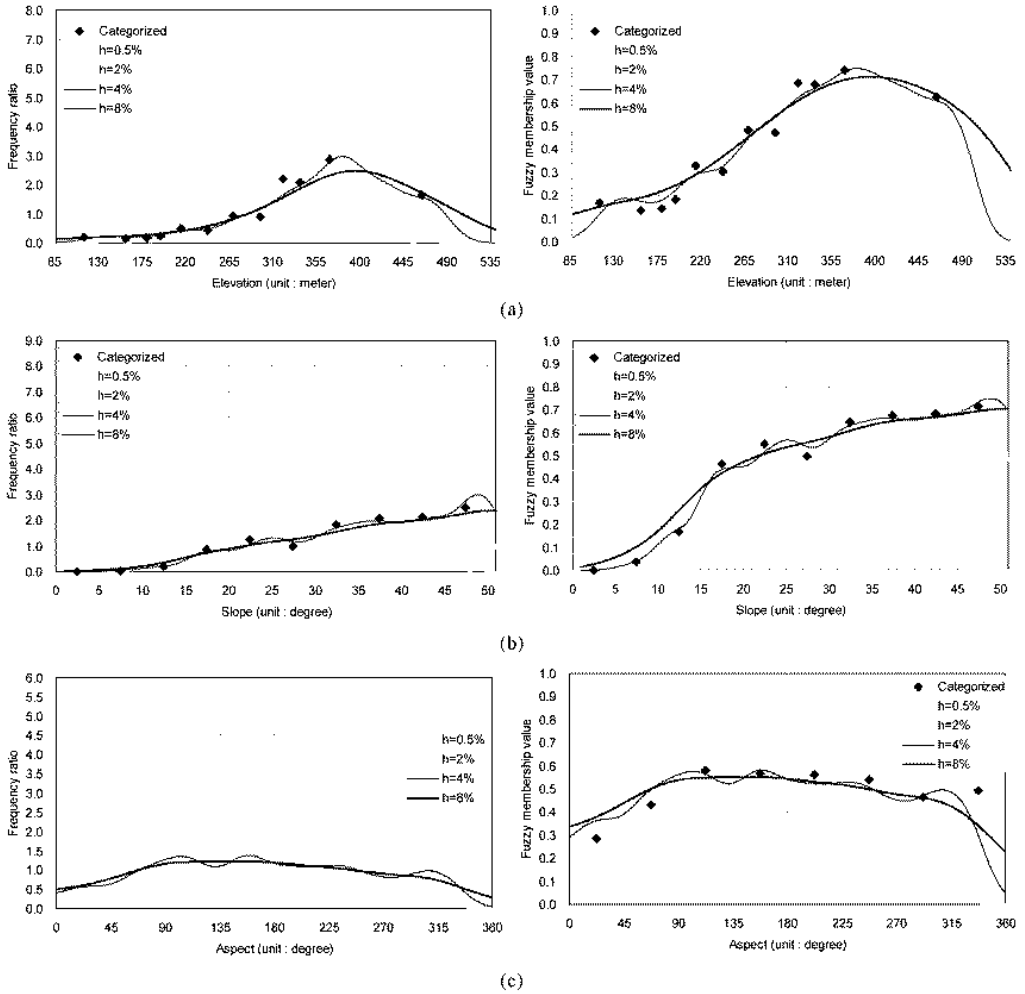


**Fig. 4.** The frequency ratio and fuzzy membership values of (a) elevation map, (b) slope map, (c) aspect map.

bership were computed. In the elevation map, the frequency ratio value was greater than one, in the range of elevation values between 285 m and 500 m. Most mountain areas are located between these altitudes and thus the possibility of landslide occurrences is high. However, the highest areas do not always indicate a high possibility of landslides because of a lack of superficial deposits.

In the slope map, the frequency ratio increases according to the slope angle. In particular, a slope angle higher than 22 degrees shows frequency ratio values greater than one. This means that most landslides occurred in areas where the slope angle is greater than 22 degrees. In the aspect map, the frequency ratio values on east-, southeast- and south-facing hill slopes were high. These results may be a function of the differing hours of sunshine experienced according to the aspect. In all cases, the fuzzy membership values of categorized continuous data lie in $h$ values between 0.5% and 8%.

After preparing the final fuzzy membership functions of all input spatial data, they were integrated

using a fuzzy algebraic sum operator. From our previous study on the effects of the fuzzy combination operator on the final integrated results (Park et al., 2003a), the fuzzy algebraic sum operator showed the highest prediction power. As the choice of an optimum fuzzy combination operator was not the focus of the case study herein, the fuzzy algebraic sum operator was experimentally selected.

Given a set of $n$ fuzzy membership functions $\{\mu_i,\ i=1, 2, \cdots, n\}$, the final integrated fuzzy membership function ($\mu_{integ}$) using the fuzzy algebraic sum operator is defined as:

$$\mu_{integ} = 1 - \prod_{i=1}^{n}(1-\mu_i) \tag{6}$$

Finally, landslide hazard maps were generated wherein each pixel contains the hazard level measure mapped in the range of 0 to 200. 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. The mapping function is similar to histogram equalization.

The main objective of this landslide hazard mapping is to estimate the relative hazard level within the study area. When interpreting the final combined fuzzy membership function, it should be noted that this is relative. A fuzzy membership function close to 1 does not necessarily mean certainty with respect to the target proposition. A final membership value is higher at one location than at another, however, means that the possibility for the target proposition is higher at the former location than at the latter (Moon, 1998).

Another problem in direct interpretation of the fuzzy membership functions arises when comparing one result generated by a certain combination operator with results derived through various other fuzzy combination operators. The combination obtained through the fuzzy algebraic product operator generally results in fuzzy membership function values that are lower than those given by the fuzzy algebraic sum operator. This does not mean that the result by the fuzzy algebraic sum operator is better.

From the rank-based presentation, there are approximately the same number of pixels at each class level. Thus, for example, to obtain the 10% most hazardous area, one can threshold the output image at 200X 90%. The pixels above this threshold should fall in the top 10% category. This presentation enables not only the expression of the relative hazard level but also a comparison with other integrated results generated by other fusion methods with different theoretical backgrounds.

Before integrating all input data, three continuous data were first considered. As shown in Fig. 5, the integrated map shows relatively smoother patterns with the increase of $h$ value. Though the rank-based presentation may act as a smoothing operator, some spurious discontinuities remain in the integrated result with an $h$ value of 0.5%.

To quantitatively evaluate those integrated results, the occurrence of landslides in the validation group is compared with relative hazard level classes in the integrated maps. First, the number of pixels of landslides in the validation group, which fall in a certain hazard level class was counted. Then the number was divided by the total number of landslides in the validation group. This normalized metric can be regarded as a quantitative measure for the prediction of future possible landslides, because landslides not used in the integration procedure were compared. From the normalized metric, a prediction rate curve showing the cumulative proportion of the normalized metric (Chung and Fabbri, 1999) was obtained. If a hazard map shows random patterns, it does not have any significance for prediction and its prediction rate curve will be a diagonal line. Thus, the larger the area between the prediction rate curve and the diagonal line becomes, the better the prediction capability will be.

The prediction rate curves of the integrated results using continuous data are shown in Fig. 6. An analysis of those curves led to the result that the pre-

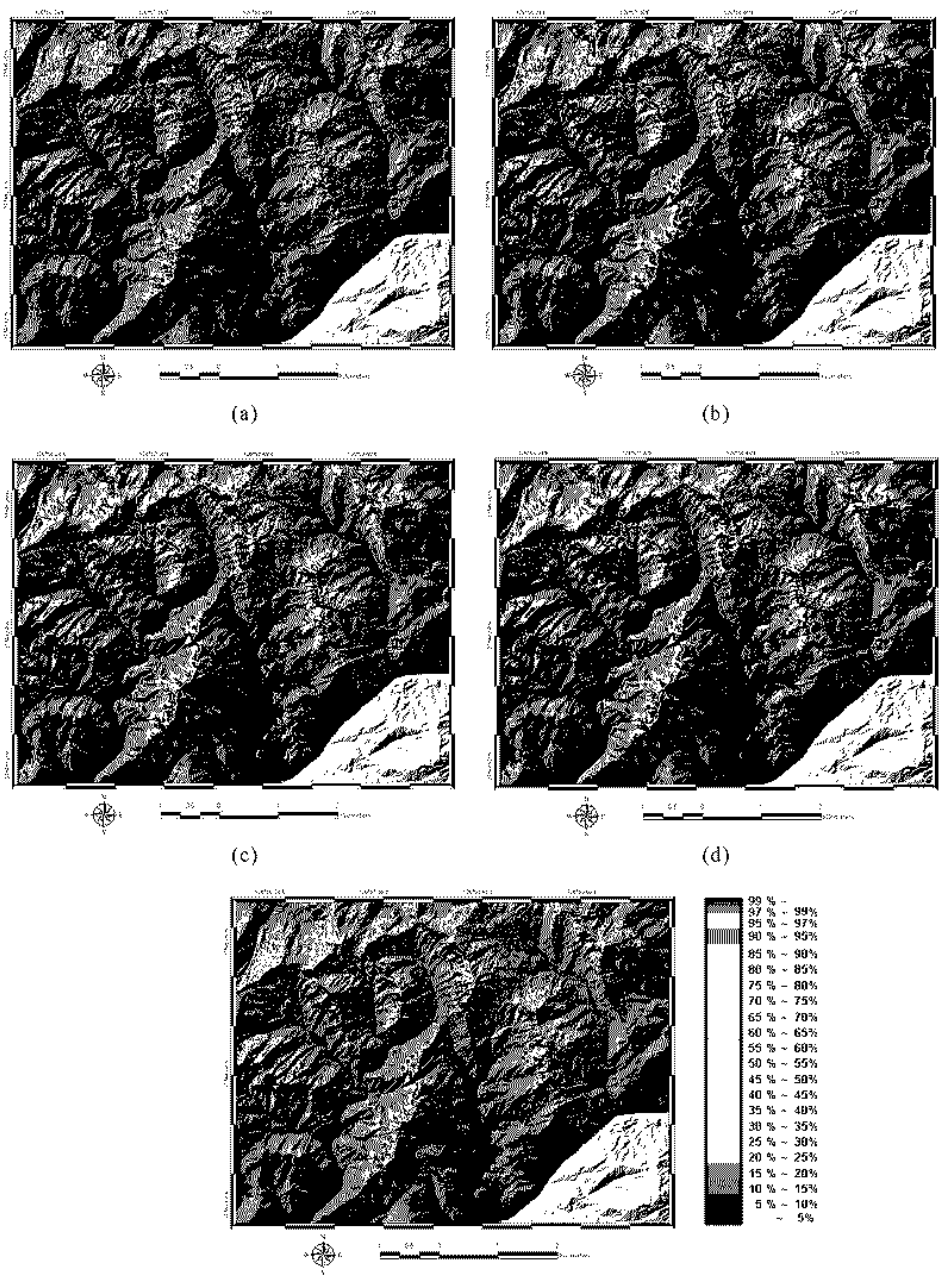


**Fig. 5.** Landslide hazard maps generated by using three continuous data; (a) result for $h$=0.5%, (b) result for $h$=2%, (c) result for $h$=4%, (d) result for $h$=8%, (e) result for categorized maps. The black dots denote landslides in the validation set and the background is a shaded relief map.

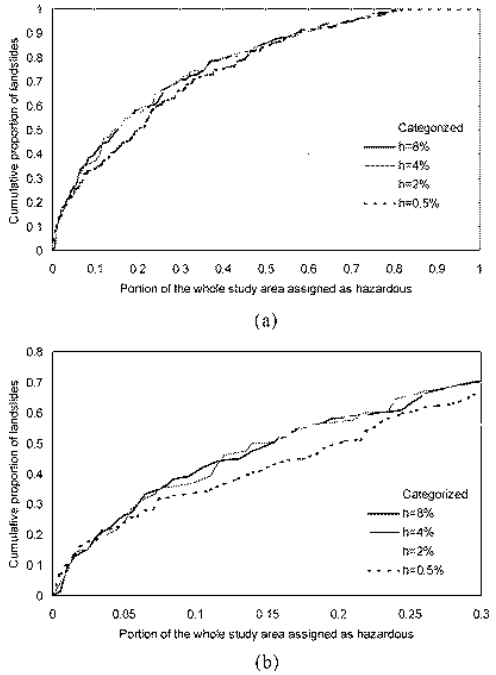


**Fig. 6.** (a) Prediction rate curves of the integrated maps shown in Figure 5, (b) enlargement of (a).

diction rates of the application of a Parzen window estimation for continuous data representation were higher than the one by traditional categorized representation of continuous data. In the uppermost categories, significant improvements in the prediction capability, of about 10~15% were achieved. Those improvements were observed for the top 60% of categories of hazard level, or proportion of the study area. The best prediction capabilities were obtained by $h=4\%$ and $h=8\%$. Through these $h$ values about 40% of the landslides in the validation group were predicted from the uppermost 10% class, which occupy 10% of the whole study area. In the case of $h=0.5\%$, the prediction rate was lower than for the results obtained with larger values of $h$.

As another useful quantitative measure for interpreting the prediction rate curve, slope values were computed for the curve for each 5% (Table 2). These slope values represent the increment of the prediction rate changes. A value of 1 means that the prediction pattern in that class is a random one and thus it has no significance. The more the slope value exceeds 1, the stronger is the significance of the prediction result.

For prediction rate curves to show reasonably significant results, the slope value corresponding to the most hazardous class should be much larger than that for the next lower hazard class. That is, the most hazardous class should include most of the landslides in it, and will occupy small areas throughout the study area.

The conclusion derived from visual and/or quantitative interpretations of the prediction rate curves was confirmed through the results shown in Table 2. The slope values for the most hazardous 5% class computed from the integrated results derived by Parzen density estimation were higher than those gained from categorized continuous data. Notably, $h=8\%$ gave the highest slope value for both the top 5% and the next lower hazard class.

The validation procedure was repeated for the data set including two categorical and three continuous categories (Fig. 7). As with the results for continuous data, the order of the prediction capability was preserved (Fig. 8 and Table 3). That is, in case

**Table 2.** Slope values of the top 30% classes (5% apart) for the integrated results using continuous data

| Class | $h$=0.5% | $h$=2% | $h$=4% | $h$=8% | Categorized |
|---|---|---|---|---|---|
| 0~5% | 4.77 | 4.77 | 4.69 | 5.27 | 3.12 |
| 5~10% | 1.97 | 1.81 | 2.72 | 2.80 | 1.90 |
| 10~15% | 1.65 | 2.72 | 2.63 | 1.81 | 1.64 |
| 15~20% | 1.73 | 1.98 | 1.40 | 1.81 | 1.24 |
| 20~25% | 1.89 | 1.81 | 1.65 | 0.90 | 1.82 |
| 25~30% | 1.24 | 0.91 | 1.07 | 1.48 | 1.56 |

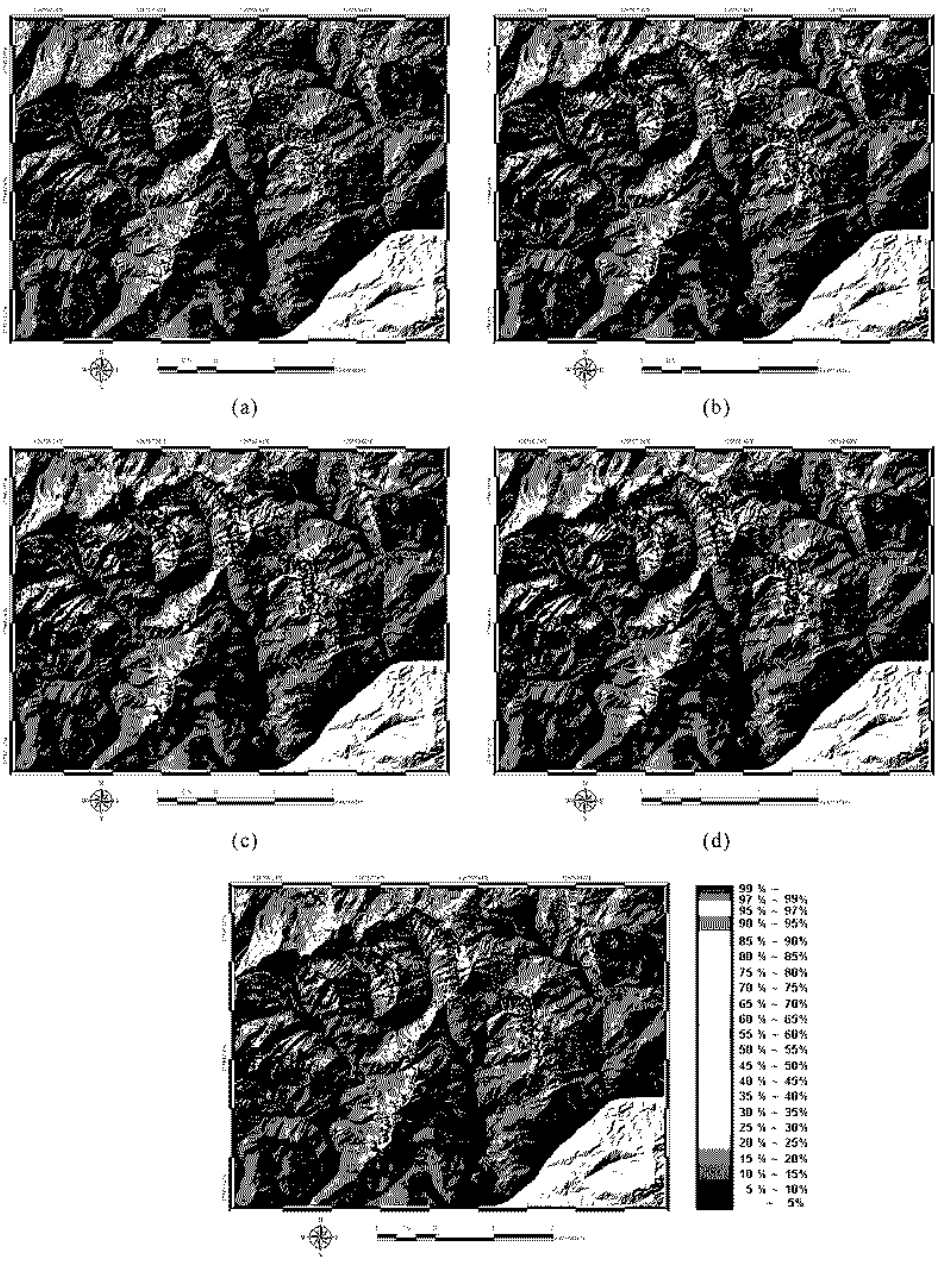


**Fig. 7.** Landslide hazard maps generated by using all data; (a) result for $h$=0.5%, (b) result for $h$=2%, (c) result for $h$= 4%, (d) result for $h$=8%, (e) result for categorized maps. The black dots denote landslides in the validation set and the background is a shaded relief map.

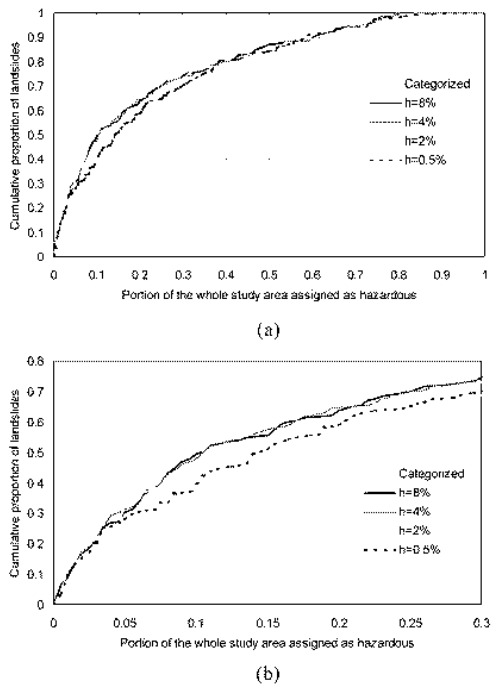


**Fig. 8.** (a) Prediction rate curves of the integrated maps shown in Figure 7, (b) enlargement of (a).

of $h = 8\%$, its prediction rate was the highest and the result obtained through categorized continuous data showed the worst prediction rate.

When comparing the results generated from different values of $h$, $h = 4\%$ and $h = 8\%$ showed the best prediction capabilities. It is worth noting that adding the categorical data allowed an increase in the overall prediction rate in all cases. In the case of $h = 8\%$, the prediction rate using continuous data only was 40% within the top 10% class. Whereas the integration results obtained by adding the categorical data showed a prediction rate value of 49% in the same uppermost 10% class.

This improvement of the prediction rates can be explained by the information content. The forest type and soil maps may provide useful information for landslide occurrences which are not fully explained by the continuous data of topographic information. Other complementary information may isolate or strengthen the relationships between the input data and landslide occurrences and thus help to improve the prediction capabilities. This result confirms the effectiveness and necessity of multi-source data fusion.

## Conclusions

In this paper, an effective methodology for data-driven fuzzy information representation of multi-source spatial data was applied to landslide hazard mapping. The main contribution of the methodology presented herein lies in the representation of continuous data by adopting a non-parametric density Parzen window estimation. This approach allows the direct use of the original continuous data without any categorizing procedure. The choice of an optimum value of $h$ in the kernel function, which is one of main issues arising from applying the Parzen window estimation method, was experimentally evaluated by the prediction rate curve and the slope value computed from validation analysis.

A case study in Jangheung, Korea, for landslide hazard mapping demonstrated that fuzzy information representation by a Parzen window estimation could directly use original continuous data and thus could prevent loss of information. As a result, the

**Table 3.** Slope values of the top 30% classes (5% apart) for the integrated results using all data

| Class | $h$=0.5% | $h$=2% | $h$=4% | $h$=8% | Categorized |
|---|---|---|---|---|---|
| 0~5% | 5.60 | 6.42 | 6.26 | 6.09 | 4.36 |
| 5~10% | 2.80 | 2.88 | 3.29 | 3.79 | 2.72 |
| 10~15% | 1.81 | 1.73 | 1.97 | 1.24 | 1.40 |
| 15~20% | 1.56 | 1.15 | 1.40 | 1.65 | 1.89 |
| 20~25% | 1.23 | 1.56 | 1.15 | 1.23 | 2.55 |
| 25~30% | 0.99 | 0.91 | 0.66 | 0.91 | 0.91 |

prediction capability was improved by more than 10%. The values of *h* showing the best prediction capabilities were 4% and 8%. In addition, it should be noted that the quantitative measures (i.e. increment and slope) from the prediction rate curves made it possible not only to compare the prediction capabilities but also to interpret the integrated results. To confirm the results derived from data-driven fuzzy information representation, more research will be done in several study areas.

When undertaking multi-source spatial data fusion, the reliabilities and uncertainties of data should be properly assessed in the information representation step. This paper only evaluated the uncertainties arising from continuous data representation. In the case of categorical data, uncertainties may arise from the crisp boundary representation. Recently, Park et al.(2003b) proposed the effective data representation methodology for categorical data using fuzzy boundary representation. Future work will investigate the combined effects of the fuzzy boundary representation for categorical data and the continuous data representation method proposed here.

## Acknowledgments

The authors thank Dr. Chang-Jo F. Chung of Geological Survey of Canada for providing the motivation of this work as well as for his constructive comments on the methodology development. Constructive comments by two anonymous reviewers also helped us improve the presentation of this paper. This work was partly supported by the Korean Ministry of Science and Technology.

## References

Agresti, A., 1990, Categorical data analysis, John Wiley & Sons, New York, 558 p.

An, P., Moon, W.M., and Rencz, A., 1991, Application of fuzzy set theory to integrated mineral exploration, Canadian Journal of Exploration Geophysics, 27, 1-11.

Bonham-Carter, G.F., 1994, Geographic information systems for geoscientists: modeling with GIS, Pergamon Press, New York, 398 p.

Bonham-Carter, G.F., Agterberg, F.P., and Wright, D.F., 1988, Integration of geological data set for gold exploration in Nova Scotia, Photogrammetric Engineering & Remote Sensing, 54, 1585-1592.

Carranza, E.J.M. and Hale, M., 2001, Geologically constrained fuzzy mapping of gold mineralization potential, Baguio district, Philippines, Natural Resources Research, 10, 125-136.

Chung, C.F. and Fabbri, A.G., 1993, The representation of geoscience information for data integration, Nonrenewable Resources, 2, 122-139.

Chung, C.F. and Fabbri, A.G., 1999, Probabilistic prediction models for landslide hazard mapping, Photogrammetric Engineering & Remote Sensing, 65, 1389-1399.

Davis, J.C., 1986, Statistics and data analysis in geology, John Wiley & Sons, New York, 656 p.

Duda, R.O., Hart, P.E., and Stork, D.G., 2000, Pattern classification, John Wiley & Sons, New York, 654 p.

Fukunaga, K., 1990, Introduction to statistical pattern recognition, Academic Press, San Diego, 592 p.

Hastie, T., Tibshirani, R., and Friedman, J., 2001, The elements of statistical learning, Springer, New York, 533 p.

Kim, K.-S., 2001, Prediction of landslide probability by geomorphic characteristics and soil properties, KIGAM Bulletin, 5, 29-41.

Moon, W., 1990, Integration of geophysical and geological data using evidential belief function, IEEE Transactions on Geoscience and Remote Sensing, 28, 711-720.

Moon, W., 1998, Integration and fusion of geological exploration data: a theoretical review of fuzzy logic approach, Geosciences Journal, 2, 175-183.

Park, N.-W., Chi, K.-H., Chung, C.F., and Kwon, B.-D., 2003a, GIS-based data-driven geological data integration using fuzzy logic: theory and application, Economic and Environmental Geology, 36, 243-255.

Park, N.-W., Chi, K.-H., Chung, C.F., and Kwon, B.-D., 2003b, Predictive spatial data fusion using fuzzy object representation and integration: application to landslide hazard assessment, Korean Journal of Remote Sensing, 19, 233-246.

Park, N.-W., Chi, K.-H., Lee, K.-J., and Kwon, B.-D., 2003c, Automatic estimation of threshold values for change detection of multi-temporal remote sensing images, Korean Journal of Remote Sensing, 19, 465-478.

Parzen, E., 1962, On the estimation of a probability density function and the mode, Annals of Mathematical Statistics, 33, 1065-1076.

Silverman, B.W., 1986, Density estimation for statistics and data analysis, Chapman and Hall, Florida, 175 p.

SpatialModels Inc., 2004, Users guide of spatial prediction modeling system, 168 p.

Zadeh, L.A., 1965, Fuzzy sets, Information and Control, 8, 338-353.

Zimmermann, H.J., 1996, Fuzzy set theory and its applications, Kluwer Academic Publisher, Massachusetts, 435 p.

Manuscript received, December 28, 2004
Revised manuscript received, February 15, 2005
Manuscript accepted, February 15, 2005