Journal of the Korean Society of Remote Sensing, Vol.15, No.2, 1999, pp.91~105

GIS-based Spatial Integration and Statistical Analysis using Multiple Geoscience Data Sets: A Case Study for Mineral Potential Mapping

Ki-Won Lee*, No-Wook Park, Kwang-Hoon Chi***, and Byung-Doo Kwon****

ETRI/CSTL-GIS*, Dept. of Earth Sciences, Seoul National University**, Div. of Geo-Environmental Research, KIGAM***

다중 지구과학자료를 이용한 GIS 기반 공간통합과 통계량 분석 : 광물 부존 예상도 작성을 위한 사례 연구

이기원* · 박노욱** · 지광훈*** · 권병두**

한국전자통신연구원 GIS연구팀*, 서울대학교 지구과학교육과**, 한국자원연구소 지구환경정보연구부***

Abstract : Spatial data integration using multiple geo-based data sets has been regarded as one of the primary GIS application issues. As for this issue, several integration schemes have been developed as the perspectives of mathematical geology or geo-mathematics. However, research-based approaches for statistical/quantitative assessments between integrated layer and input layers are not fully considered yet. Related to this niche point, in this study, spatial data integration using multiple geoscientific data sets by known integration algorithms was primarily performed. For spatial integration by using raster-based GIS functionality, geological, geochemical, geophysical data sets, DEM-driven data sets and remotely sensed imagery data sets from the Ogdong area were utilized for geological thematic mapping related by mineral potential mapping. In addition, statistical/quantitative information extraction with respective to relationships among used data sets and/or between each data set and integrated layer was carried out, with the scope of multiple data fusion and schematic statistical assessment methodology. As for the spatial integration scheme, certainty factor (CF) estimation and principal component analysis (PCA) were applied. However, this study was not aimed at direct comparison of both methodologies; whereas , for the statistical/quantitative assessment between integrated layer and input layers, some statistical methodologies based on contingency table were focused. Especially, for the bias reduction, jackknife technique was also applied in PCA-based spatial integration. Through the statistic analysis with respect to the integration information in this case study, new information for relationships of integrated layer and input layers was extracted. In addition, influence effects of input data sets with respect to integrated layer were assessed. This kind of approach provides a decision-making information in the viewpoint of GIS and is also exploratory data analysis in conjunction with GIS and geoscientific application, especially handling spatial integration or data fusion with complex variable data sets.

Received 26 February 1999 ; Accepted 29 April 1999.

Key Words : Certainty factor estimation, Contingency table, Jackknife, Principal component analysis, Spatial integration, Spatial statistics

요 약 : 최근 다중 지질정보의 통합적 해석은 GIS의 중요한 응용 분야중 하나로 인식되고 있다. 공간통합을 위하여 지구통계학적 방법들이 개발되어 있지만, 통합결과와 입력 주제도들 사이의 관계에 대한 통계적, 정량적 분석방법론의 개발은 아직까지 체계적으로 정립되어 있지 못한 상황이다. 본 연구에서는 지질도, 지화학자료, 항공지구물리자료, 지형자료 및 원격탐사 영상등 다양한 지질정보등이 보고된 옥동지역을 대상으로 하여 광물 부존 예상도 작성 사례연구를 수행하여 기존에 이용되고 있는 여러 공간 통합 방법 중 확실인자 (Certainty Factor: CF) 추정방법과 다변량 통계 분석방법중 하나인 주성분분석을 시험적인 통합방법으로 우선적으로 적용한 뒤, 입력 자료와 통합결과에 대한 정량적인 통계량 정보를 추출하고자 하였다. 입력 주제도와 통합 결과사이의 관계 규명에는 통계 분할표를 이용한 통계처리를, 편의 분석에는 잭나이프 방법을 적용하였다. 통합정보에 대한 통계량 분석을 통하여, 통합결과와 입력자료 사이의 정량적 관계를 추출할 수 있었으며, 부가적으로 입력자료의 상태수준에 대한 판단정보를 얻을 수 있었다. 이러한 결과는 GIS 관점에서 통합결과 해석에 중요한 결정보조자료로 활용될 수 있으며, 복잡한 다중정보를 다루는데 공간 통합문제에서도 입력정보 검증을 위한 일반적인 처리과정으로도 발전할 수 있을 것으로 생각된다.

주요어 : 공간 통합, 공간 통계량, 주성분분석, 잭나이프기법, 통계 분할표, 확실 인자 (Certainty Factor) 추정

1. Introduction

Since the early 1990s, GIS(Geographic Information System) is regarded as one of important tools for geo-based spatial data analysis. Especially, spatial integration, one of GIS-based data fusion approaches, with multiple source geo-registered data sets has been studied in the GIS perspectives. In spatial integration, it covers geological data sets of catchment geochemistry, (airborne) geophysical, and geological map, remotely sensed imagery, and DEM, as data types utilized. In these days, several useful methodologies towards spatial data integration based on geo-mathematical approaches have been developed and applied to site-specific researches such as mineral exploration/deposit modeling, geo-based hazard modeling, environmental vulnerability mapping given conditions and so forth. This integration schemes have the basis or application of : Weight of evidence (Bonham-Carter *et al.*, 1988), Dempster-Shafers theory (Moon, 1990), multivariate statistical approach, (Vulkan and Duval. 1993 ; Lee *et al.*, 1995), Fuzzy set theory (Wright and Bonham-Carter, 1996), Certainty Factor(CF) estimation approach (Chi *et al.*, 1997) or Bayes theory (Rostirolla *et al.*, 1998). Therefore, integrated or fused layer in the form of thematic/favorable mappable information, whatever any methodology is applied, can be regarded as decision-making information for a given purpose. However, influence of input layers and error propagation/assessment with respect to each integration method are not considered yet; moreover, there are a few researches or unpublished ones on this problem. While, though statistical analysis languages and statistical analysis methods directly linked with commercial

or public-based GIS tools are recently released, it is somewhat insufficient to quantitatively interpret the spatially integrated mappable information due to differences between spatial data and general variables. But most of them are currently developing from simple statistical expression or thematic mapping related to statistical estimation to more sophisticated problem-solving.

The main purpose of this study is not just comparison with various spatial integration methodologies mentioned before. Rather, quantitative assessments between resultant layer and input layers, independent on spatial integration approaches, are more emphasized. It is also possible to consider to statistical information extraction in the viewpoint of data fusion, or exploratory data analysis.

As an actual case study for this approach, mineral potential mapping by detection of favorable mineral occurrence zone was carried out. In this case study, multiple geo-based data sets collected from the Ogdong area were used for application of integration scheme (CF estimation and PCA). After the resultant integrated layer was obtained, some statistical methodologies (chi-square statistics coefficient, Cramers coefficient, contingency coefficient, entropy, Yule coefficient, and odds ratio) for quantitative assessment or extraction of statistic information were applied.

2. Applied Methodologies in the Case Study

In this study, CF estimation approach and multivariate statistical approach were applied for data integration scheme.

CF estimation, widely used in a rule-based system, measures certainty level of conditional probability with respect to priori probability given a certain evidence. This method is known as the effective method in case that there is much statistical data. A certainty factor (CF) at p for the kth layer, denoted by CF$k(p)$, is defined as the change in certainty that the proposition (a pixel p contains deposits of type D) is true, from without the evidence at p to given the evidence at p in the kth layer (Chung *et al.*, 1993). CF ranges between -1 and +1. Positive numbers for CF correspond to an increase in certainty in a proposition after the evidence is observed, whereas negative numbers correspond to a decrease in certainty.

The definition discussed by Heckerman(1986) is followed.

According to this formulation, CF is equal to zero if the conditional probability is equal to the priori probability; the absolute value of CF

$$CF_k(p) = \begin{cases} \dfrac{Prob_k\{T_p \mid v_k(p)\} - Prob_k\{T_p\}}{Prob_k\{T_p \mid v_k(p)\}\,(1 - Prob_k\{T_p\})} & \text{if } Prob_k\{T_p \mid v_k(p)\} > Prob_k\{T_p\} \\ \dfrac{Prob_k\{T_p \mid v_k(p)\} - Prob_k\{T_p\}}{Prob_k\{T_p\}(1 - Prob_k\{T_p\})} & \text{if } Prob_k\{T_p \mid v_k(p)\} > Prob_k\{T_p\} \end{cases}$$

where,

$Prob_k\{T_p\}$: the priori probability that a pixel p contains a deposit before any evidence(layers) is not given

$Prob_k\{T_p \mid v_k(p)\}$: the conditional probability that a pixel p contains at least one deposit given the evidence $v_k(p)$ at p.

increases if the conditional probability is far from the priori probability. Thus, CF can be utilized as a measure of certainty with respect to the priori probability only.

While, Principal Component Analysis (PCA), one of the multivariate statistical approach, is the method to reduce dimension between correlated variables (Jolliffe, 1986). The general PCA scheme is based on the eigen-analysis. To accomplish PCA, Z-scored transformation, the process of the standardization of variables, is needed. Z-scores with zero mean and unit variance mean statistical centering and standardization of each data set. It is necessary to handle observations with different unit. Principal component loading value reflects the relative importance of a variable within a principal component and principal component scores constitute orthogonal projections of the given data values onto the axes defined by principal components (Davis, 1986).

To measure association between two thematic maps with multiple classes, contingency table is commonly used (Bonham-Carter, 1994). Contingency table for cross-tabulation is the table showing discrete frequency or cell-counting in the matrix style and is similar to error/confusion matrix commonly used to calculate the classification accuracy of remote sensing image. After contingency table is calculated, several statistics related to measure of association can be obtained to quantitatively assess integrated results or layers; chi-square statistics coefficient, Cramers coefficient, contingency coefficient, entropy, kappa coefficient, Yule coefficient, or odds ratio. Among them, kappa coefficient is limited to the situation of comparing maps with the same number of matched classes. Yule coefficient and odds ratio is mainly utilized to the comparison of binary maps. However, measures of associations between binary patterns can be applied to the comparison of multi-class maps by treating each combination of map classes as a binary case. As for some statistics mentioned before, let the table between map A and B be called matrix T, with elements T_{ij}, where there are i=1,2,⋯,n classes of map B (rows) and j=1,2,⋯,m classes of map A (columns). The marginal totals of T are defined as $T_{i.}$ for the sum of the ith row, $T_{.j}$ for the sum of the jth column, and $T_{..}$ for the grand total summed over rows and columns.

Then, chi-square statistic is defined as

$$x^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(T_{ij} - T^*_{ij})^2}{T^*_{ij}}, \text{ where } T^*_{ij} = \frac{T_{i.}\, T_{.j}}{T_{..}}.$$

The meaning of large chi-square statistic value is that the association between two maps is strong. Two commonly quoted coefficients of association based on chi-squre values are the Cramers coefficient(V), and the contingency coefficient(C).

$$V = \sqrt{\frac{x^2}{T_{..} \min(i-1, j-1)}}, \quad C = \sqrt{\frac{x^2}{T_{..} + x^2}}$$

Two coefficients vary between 0 (indicating no correlation) to a maximum value less than 1.

To compute entropy statistics(U) which varies between 0 and 1, proportions, by dividing each element by the grand total, are used. Assuming that proportions matrix for map A and map B has been determined from T,

$$U = 2\left[\frac{H(A) + H(B) - H(A, B)}{H(A) + H(B)}\right]$$

where,

$$\text{entropy of A}: H(A) = -\sum_{j=1}^{m} \frac{T_{.j}}{T_{..}} \ln \frac{T_{.j}}{T_{..}}$$

$$\text{entropy of B}: H(B) = -\sum_{i=1}^{m} \frac{T_{.i}}{T_{..}} \ln \frac{T_{.i}}{T_{..}}$$

$$\text{joint entropy}: H(A,B) = -\sum_{i=1}^{m} \sum_{j=1}^{m} \frac{T_{ij}}{T_{..}} \ln \frac{T_{ij}}{T_{..}}.$$

Yule coefficient (α) and odds ratio(O_R) are defined as follows ;

$$\alpha = \frac{\sqrt{T_{11}/T_{21}} - \sqrt{T_{12}/T_{22}}}{\sqrt{T_{11}/T_{21}} + \sqrt{T_{12}/T_{22}}}, \quad O_R = \frac{T_{11}/T_{22}}{T_{12}/T_{21}}$$

Where,

$$A \cap B = T_{11},\ A \cap \overline{B} = T_{21},\ \overline{A} \cap B = T_{12},\ \overline{A} \cap \overline{B} = T_{11}.$$

Yule coefficient ranges in value between –1 and +1 like a correlation coefficient, and odds ratio is always positive, being greater than 1 for patterns that are positively associated, 1 if the two patterns are independent and less than 1 if they are negatively associated.

As another applicable statistics, Jackknife technique, as the approximation of bootstrapping, is a statistical method to convert a given estimation into a revised estimation which is less biased to original one (Cressie, 1993 ; Efron and Tibshirani, 1993), and this methodology is tentatively applied to PCA scheme with the following rationale.

Let $\hat{E}$ and $\hat{E}_{-i}$ denote the estimation of E on all n observation, and partial estimation obtained by deleting ith sample and estimation E from remaining (n-1) observations, respectively. Then, pseudo-value, $\hat{E}_i$, combination of the partial estimates with whole sample estimations, is

$$\hat{E}i = \hat{E} - bias = n\hat{E} - (n-1)\,\hat{E}_{-i}, \text{ where } i = 1, 2, \ldots, n.$$

The average of pseudo-value is the jackknifed estimate of E with regards to $\hat{E}$,

$$\hat{E} = \frac{1}{n}\sum_{i=1}^{n} \hat{E}_i .$$

3. Case Study for Mineral Potential Mapping

1) Data Sets for the Case Study

The general geology of the study area is shown in Fig. 1. Geologically, the study area is covered with the Precambrian metasediments, pegmatitic migmatite, Joseon supergroup of Cambro-Ordovician, Pyongan supergroup of late Carboniferous to Triassic, formation of Jurassic and igneous rocks such as granites, porphyritic intrusives and dikes. The most part of the study area is composed of metasediments.

Most polymetallic mines are located at western

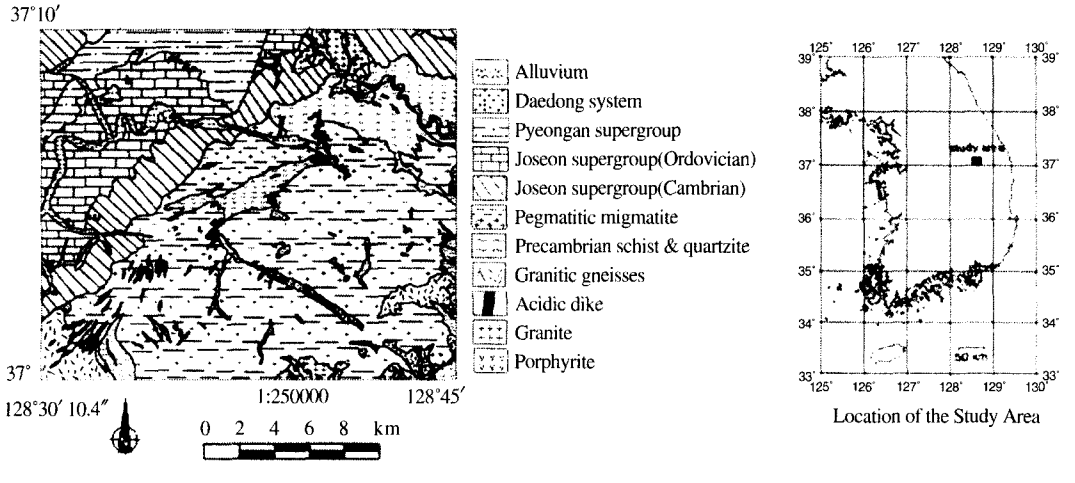

Fig 1. Geological map, as one of GIS layers, and location map of the study area.

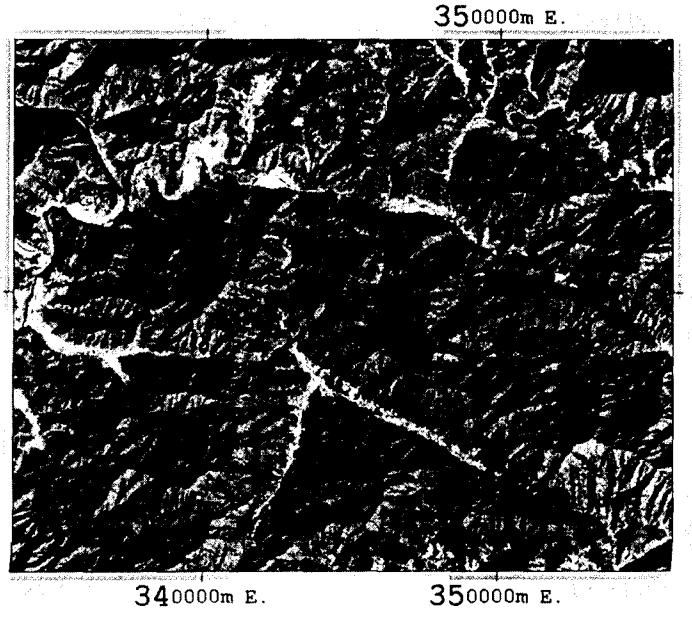

Fig. 2. False color composite image using LANDSAT TM, covering the whole study area.

and eastern area. Ore deposits are mainly located at great limestone series of Joseon supergroup and granite, ; rich elements of ore deposits are Fe, Pb and Zn.

As for reconnaissance survey of this case study area prior to data integration task, remotely sensed imagery was preprocessed and classified. A false color composite(532 bands) of the geometrically corrected Landsat TM imagery in Oct., 1989 of the study area is presented in Fig. 2. As shown in Fig. 2, the topography of the study area is steep and hilly mountainous region. The Ogdong stream which flows from east to west and drainages are well shown. In the north- west part, main drainage creeks intersect the direction of mountain ridges and in the eastsouth part, main drainages are parallel to the directions of ridges.

As for actual applications, 15 types of geo-based data sets were utilized (Table 1). As for geophysical data, airborne surveyed data composed of residual magnetic anomaly and radiometric anomaly in originally vector format were converted to raster format. As for geochemical data, ground surveyed data sets from widely distributed stream rock samples were resampled and interpolated to obtain grid data. DEM (Digital Elevation Model) was produced by vector to raster conversion process. Then, DEM-driven slope and aspect map sets were used as the topographic data sets. Supervised classification was used for Landsat TM imagery as remotely sensed data sets. As a preprocessing, geometric correction was carried out. Because the study area is mountainous region, the selection of GCP was somewhat difficult. The RMS(Root Mean Square) error was 0.94 pixel, less than 1 pixel. For spatial integration, cell resolution of all raster data sets was converted to 30 meter, that of Landsat TM imagery. Therefore, error less than 1 pixel had no problem for quantitative analysis. Maximum likelihood method, one of the supervised classification methods, was applied for

Table 1. Data sets for spatial integration in the case study

Input Layers		Number of Recorded Class	Applied Data Integration Scheme
Geological map*		11	CF
Geochemical Surveyed data sets**	Ag (Silver)	6	CF & PCA
	Cd (Cadmium)	6	CF & PCA
	Cu (Copper)	6	CF & PCA
	Pb (Lead)	6	CF & PCA
	U (Uranium)	6	CF & PCA
	Zn (Zinc)	6	CF & PCA
Airborne Geophysical data sets ***	gammaray	8	CF & PCA
	K (Potassium)	8	CF & PCA
	Th (Thorium)	8	CF & PCA
	U (Uranium)	8	CF & PCA
	Residual magnetic intensity	8	CF & PCA
Remote sensing Image (Landsat TM)	Supervised Classification image	5	CF
DEM	Slope	9	CF
	Aspect	8	CF
Ore deposit map		8 ore deposits	CF

(CF : Certainty factor estimation, PCA : Principal Component Analysis)

Note : The type of data used in PCA scheme is continuous variable.

Data Sources :

* Geological map of Ogdong area scaled by 1:50,000, Geological Survey of Korea, 1966.

** Geochemical Maps for Ogdong Sheet in the Taebaegsan Mineralized Belt, KIER(Korean Institute of Energy and Resources), 1984

*** Aerial gamma ray and magnetic survey map at Jungseon, Samcheok, Yemi, Jangsung (1:50,000), KIER, 1988

classification. The training area was specified by 5 classes ; water body, forest, alluvium & barren land, agricultural land, and shadow zone. To enhance the classification accuracy, shadow zone was additionally added. As a result of classification, average, overall accuracy and kappa coefficient were 94.14%, 95.03% and 0.9308, relatively. Geological map was also fully geo-registed into GIS with geometric features and their database attributes.

To integrate with categorical variables (geological data, supervised classification image), continuous variable data sets were converted to ordinal variables, additionally. Through histogram analysis, geochemical, geophysical and DEM-driven data sets, which were originally continuous variables, were reclassified for reflecting the feature of them. In that, minimum class value and maximum class value represent "very low" and "very high", respectively.

Also, the location map of known eight mines was used as prior evidence in spatial integration by CF estimation.

2) Spatial Integration and Statistics Analysis Results

As for the integration of continuous variable form data sets, geochemical and geophysical data sets, PCA scheme was applied ; As for the integration of all multiple geo-based data sets,

which were ordinal and categorical variables. towards detection of favorable mineral occurrence zone, CF estimation scheme was applied, in addition to known ore deposit map, as the ground truth.

(1) PCA and Jackknifed Estimation Results

Z-scores transformed matrix of geochemical and geophysical data sets, as continuous variable form, were used for application of the PCA scheme. Principal component Ⅰ is associated with 29.1% of the total information of principal components (Table 2 (a)). In general, PC Ⅰ is associated with most information of principal components. To consider from PC Ⅰ to PC Ⅳ, about 72% of the total information of principal components is associated. It is partly affected that attribute of geochemical data sets reflects only effect of surface anomaly, whereas that of geophysical data sets reveals the complex effect of both surface and sub-surface anomaly.

Principal component loading (Table 2 (b)), which reflects the relative importance of a variable within a principal component, were used in analysis of principal component score. Absolute value of PC loadings with respect to geochemical data sets of PC Ⅰ is larger than that of geophysical data sets, except residual magnetic intensity and PC loadings with respect to most element of PC Ⅰ are negative values, except Ag, U elements showing positive values and Potassium showing neutral (Tables 2(b)). High PC Ⅰ scores indicate low amounts of all of the variables (except Ag, U, Potassium) (Fig. 3(a)). Negative PC scores indicate higher percentages of the listed elements. Base metal ore bodies in this area is associated with spatial pattern of geochemical elements and residual magnetic intensity, due to minus high

Table 2-a. Eigenvalues and percent with respect to the first four PC axes as PCA result

Eigen vector	Eigen value	Percent	Cumulative percent
Ⅰ	3.227	29.10	29.10
Ⅱ	2.146	19.35	48.45
Ⅲ	1.474	13.29	61.74
Ⅳ	1.120	10.10	71.84

Table 2-b. PC loadings withe respect to each element of the first four PC axes

Input Layers		Ⅰ	Ⅱ	Ⅲ	Ⅳ
Airborne geophysical data sets	gammaray	-0.304	0.889	0.068	0.179
	K	0.002	0.823	0.089	0.118
	Th	-0.439	0.372	-0.362	-0.165
	U	-0.401	0.460	0.297	0.314
	Residual magnetic intensity	-0.697	0.106	-0.292	-0.380
Geochemical data sets	Ag	0.473	0.135	-0.729	0.040
	Cd	-0.645	-0.090	0.413	-0.492
	Cu	-0.628	-0.354	-0.160	0.549
	Pb	-0.685	-0.064	-0.063	-0.274
	U	0.559	0.377	-0.285	-0.397
	Zn	-0.691	-0.132	-0.587	0.165

loadings or saturation of geochemical elements (except Ag, U) and residual magnetic intensity into axis Ⅰ.

As the total spatial pattern of PC Ⅰ, northeast area and northwest area show multi-element anomalous zones. Those anomalous zones may be affected to the known polymetallic mines. In that, northeast area results from compounds originated from the Imog, Yujeon, Dohwadong mines located in Imog granite contact zone and northwest area is mainly originated from Fe ore deposit of Ogdong mine. Also, Fe component and non-ferrous metals such as Pb, Zn are mixed in those anomalous zones.

Relatively, PC loadings with respect to airborne geophysical elements of PC Ⅱ are positive values and absolute value of PC loadings with respect to airborne geophysical data sets is larger than that of geochemical data sets (Table 2. (b)). So PC Ⅱ score mainly contains surface feature of geophysical data as high score(Fig. 3(c)). Mostly, PC Ⅱ scores show high values in granite, granitic gneisses and rocks containing biotite. It is reflection of high response of radiometric data with respect to acidic rocks.

Jackknifed estimation in PCA-based integration processing was also applied for unbiased estimation to each PC scores. Applied jackknife estimation to PCA scheme, spatial pattern of integrated layers seems to slightly be changed as the variation of PC score range, due to bias reduction (Fig. 3(b), (d)). However, overall distribution features of estimation is consistently kept, compared to estimation without jackknifing. Normally, bias reduction method influences variance of estimation. In this case, variance of jackknifed estimation decreases. It can be explained; jackknifed estimation method causes a good result with respect to "smooth" estimation function. "smooth" estimation function means that the small change of observations causes also small change of estimation result. Airborne geophysical data showing radiometric property of bedrock and geochemical catchment data showing anomalous distribution of mineralization show somewhat difference variance distribution with respect to each data. According to the histogram analysis and correlation result (though not quoted in this paper), geochemical data shows large variation and dilution effect between background values and anomalous values than airborne geophysical data. Also, each data set is relatively correlated. Though a sample is deleted by partial estimation, some outliers in data sets are smoothened by other data sets with large variation and remaining samples. As result, it has information with respect to that sample deleted due to correlation. PCA scheme is to transform to a new set of variables which are uncorrelated, based on correlation of variables. Therefore, small changed quantity of data sets does not cause large change of estimation result and result of jackknifed estimation shows as applied smoothing effect with respect to PCA result.

Furthermore, because PCA scheme is not just multiple inputs-to-one target, but multiple inputs-to-one or more response, it may be insufficient to interpret the potential mapping by the PCA scheme. In spite of these drawbacks, the major trend of PC images applied to jackknifed scheme can be utilized as a kind of supporting or supplementary layer for interpretation of other integration method result.

(2) CF Estimation and Statistical Analysis Results

Spatial integration using CF estimation was performed using the whole data sets towards favorable mapping of mineral occurrence, one of

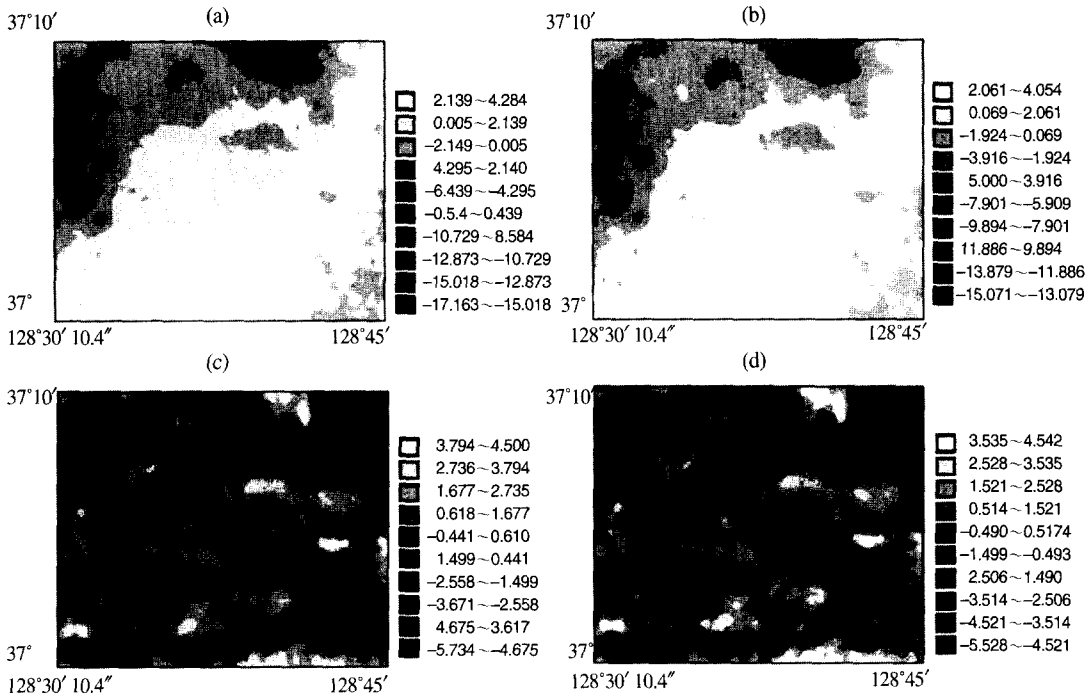

Fig. 3. Spatially integrated model by PCA: (a) PC I image, (b) Jackknifed PC I image (b) PC II image, and (d) Jackknifed PC II image.

main goals of this study. The resultant layer can be fitted real situation showing in-situ mineral occurrence (Fig. 4). It is basically caused by CF estimation method based on probabilistic relations between known occurrence event and input layers. Especially, there is a zone represents somewhat high potentiality (85% ~ 90%). This zone lies in granite zone. It results from the fact main ore deposits lie in granite and limestone. Also, low PC sore zone of PC I and high PC score of PC II are partly related to high potential zone of CF estimation result. For actual field exploration of this large scale area, high ranked zone in CF value can be considered as new target zones for mineral exploration.

Finally, to analyze quantitatively spatial integration result, chi-square statistic coefficient (χ^2), Cramer coefficient (C), contingency coefficient (V), entropy (U) were computed to reveal the relationship between original input layers and resultant layer. Also, Yule coefficient (α), odds ratio(O_R) were computed to reveal the most dominant class value within classified zones over 95% of CF estimation layer. While, those coefficients had somewhat different sensitivities in same situation, so that all coefficients were considered. This statistic can be utilized for quantitative assessment of spatially integrated information.

Using χ^2, C, V and U, overall spatial relationship between original input layers and resultant layer can be extracted. As for these statistics (Table 3 (a)), overall tendency of each layer is consistent, showing the highest rank of Pb, Cu, Zn and geology map. It is due to ore deposit major element and host rock. Also, geomorphological

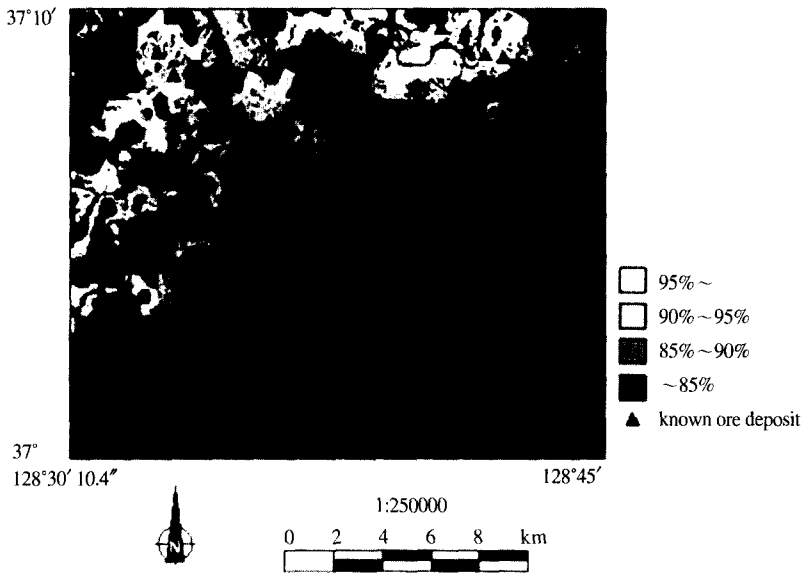

Fig. 4. Spatially integrated model by Certainty Factor estimation method.

aspect extracted from topographic data sets and supervised classification image are not highly affected to integration result. Relatively, the integrated result is more affected to geochemical data sets and geology map than airborne geophysical data (except, residual magnetic intensity). Considering the mineral occurrence, this fact means that the integrated/merged result is more highly affected to suface catchment data directly originated by ore deposits and residual magnetic intensiy, related to Fe element, regionally important and dominant element. Therefore, if there exists outlier in geochemical data sets, that severely may affect to the integration result.

Among the geochemical data sets, Ag, U elements are relatively small effect, but, effect of those elements is larger than airborne geophysical data sets. The high effect of Ag, U showing more response with respect to precambrian meta-sedimentary rocks than sedimentary rocks may attribute to non-diagonal deviation caused by outlier which may exist in data sets.

While, in α, O_R (Table 3 (b)), most dominant class of each data sets with respect to high mineral potential zone is revealed from the statistical result. Airborne magnetic intensity, Ag, Pb and Zn show relatively high value.

As for the airborne geophysical/radiometric data sets, most elements, except Potassium, show high class value. The meaning of these high class values is that granite zone and black shale containing coal seam related to Pyongan supergroup correspond to high potential zone. It is reasonable that the high class value of residual magnetic intensity is due to Fe component which results from Fe ore deposits. In spite of high response of Potassium with respect to granite, two cases that potassium shows low class value and that low association in granite zone results from the large non-diagonal deviation probably effect in computing a binary case.

As for the geochemical data sets, low class values of Ag, U correspond to pattern of the PC loading with respect to PC I. Especially, the extreme large value of Ag reflects on large effect of supplementary classes containing outliers. Therefore, it is thought that Potassium, Ag, and U elements are needed to assess field data by EDA (Exploratory Data Analysis), although it is not discussed in this study, and further research of this approach is useful to extend to GIS-based

Table 3. Summary of statistical analysis

(a) Chi Square(χ^2), Cramer(C), Contingency(V), Entropy(U) Coefficients

Input Layers		χ^2	C	V	U
Airborne geophysical data sets	gammaray	14903.72	0.105	0.179	0.0159
	K	47822.95	0.188	0.309	0.0557
	Th	44153.16	0.180	0.298	0.0634
	U	25285.42	0.137	0.230	0.0243
	Residual magnetic intensity	76838.59	0.238	0.381	0.066
Geochemical data sets	Ag	59790.51	0.210	0.341	0.081
	Cd	80075.26	0.243	0.388	0.070
	Cu	122641.1	0.300	0.462	0.151
	Pb	147770.7	0.330	0.496	0.152
	U	69416.89	0.226	0.365	0.096
	Zn	105290.0	0.279	0.434	0.135
Geology		144378.9	0.326	0.492	0.145
Landsat TM classification image		4160.51	0.055	0.096	0.006
DEM	Slope	4445.05	0.057	0.099	0.006
	Aspect	14159.45	0.102	0.174	0.018

(b) Yule coefficient(α), Odds Ratio(O_R)

Input layers		Main class	α	O_R
Airborne geophysical data sets	gammaray	7	0.350	4.313
	K	2	0.260	2.903
	Th	5	0.393	5.269
	U	6	0.399	5.414
	Residual magnetic intensity	7	0.607	16.739
Geochemical data sets	Ag	2	0.822	105.295
	Cd	5	0.572	13.500
	Cu	3	0.335	4.036
	Pb	6	0.618	17.972
	U	2	0.488	8.460
	Zn	4	0.551	11.938
Geology		granite	0.518	9.909
Landsat TM classification image		alluvium & barren land	0.158	1.894
DEM	Slope	1	0.097	1.474
	Aspect	2	0.117	1.599

EDA.

As results, Table 3 shows another useful information, not represented at integrated layer of Fig 4. But, in computing statistical coefficients related to association, those results may not perfectly reflect the relationship between each class, but the relationship between non-diagonal components. Notwithstanding this problem, based on the results of this case study, it is thought that these statistical/quantitative information is helpful to outline inter-relation between input layers and/or integrated layer and each input layer and even data quality assessment by spatial pattern which may be overlooked by visual interpretation of the resultant layer

4. Concluding Remarks

GIS-based data integration in the geoscience application can be realized as spatial analytical functionality; however, in this case study, post processing and/or interpretation of spatially merged layer, new generated layer, based on specific methodologies for data integration were dealt with.

In the case study, spatial integration by using PCA and CF estimation was implemented for mineral potential mapping. As a result, spatial pattern of PCA (PC I) and CF estimation results showed the pattern well fitted to actual ground truth representing actual mine or mineral ore deposits. Also, through statistical analysis based on contingency table, besides the mineral potential distribution, quantitative information between integrated layer and input layers was extracted. As a result, Pb, Cu, Zn of geochemical data sets, residual magnetic intensity of airborne geophysical data sets, geological map were revealed as input layers which highly affect to integrated layer. Additionally, it were thought that there might be outliers in Potassium, Ag, U and processing of assessing field data was required.

Geological interpretation of potential mapping composed of PC scores is not simple and mostly depends on computed results such as the cumulative percent of PC axes or the PC loading trend. In spite of the intrinsic complexity, the result of PCA scheme can be considered with relevant ground truth directly/indirectly related to mineral occurrence as a GIS-type decision-supporting information. In applying jackknife estimation technique for the bias reduction, because the applied scheme is not the model function, but the estimation technique, the resultant layer shows the local difference; however, the jackknife technique will efficiently be used as a tool for significant evaluation of initial estimation. Moreover, as for spatial reasoning techniques such as Dempster's rule of combination and fuzzy set theory handling semantics, it will be used as an effective tool. As well, because the jackknife technique is based on the relationships between whole estimation and partial estimations obtained by deleting one sample, comparison between whole estimation and partial estimations will be developed as a data quality assessment tool.

Though various spatial data fusion methodologies have been developed, data interpretation with respect to newly generated layer by various methodologies is somewhat different according to applied scheme; furthermore, influence or confusion effect of input layers severely may affect to the result.

In this present state, spatial integration with post-processing step of statistical analysis in this

case study provides significant information for assessment of integrated/merged layer, overlooked at normal interpretation of spatially integrated data layer. The statistical information extracted from merged/integrated layer and input layers provides quantitatively important information with respect to relationships among used data sets and even between each data set and integrated layer, in addition to the potential distribution of mineral occurrence. This information revealed by this methodologies are regarded as supporting information with quantitative additonal evidences for detailed qualitative interpretation.

While, this statistical analysis of this case study has more applicable or potential aspects : input data quality assessment and validation of data merging methodology. Actually, the interpretation of merged/integrated layer needs prerequisite concerning quality assessment of surveyed data themselves in the scope of EDA. However, this approach can be conversely utilized to assess multiple geo-based data, if provides efficient geological evidences in the actual field. Through the interpretation of whole spatial pattern and cell-valued aspect associated with EDA, validation process of data merging methodology can be assessed and the comparison with various merging methodologies can be possible. Also, though this approach and case study are performed in detecting task of mineral occurrence, the methodologies applied in this case study can be developed as an applicable general scheme, handling general multi-sources data.

Finally, this case study is towards geoscience application of GIS in spatial integration perspectives, but it can be applicable for the following general or specific GIS schemes using over tens of multiple data sets: site selection task in urban application, site characterization in environmental application, suitable range determination of land in precise farming and so forth.

References

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 and Remote Sensing, 54(11): 1585-1592.

Bonham-Carter, G. F., 1994, Geographic Information Systems for Geoscientists : Modeling with GIS, Pergamon.

Chi, K. H. , Seo, J. Y. and Han, J.K., 1997, Study on the Quantitative Evaluation of Mineral Resource Potentiality using Remote Sensing and Spatial Geoscience Data (Ⅱ), KR-97(T)-1.

Chung, F. C. and Fabbri A. G., 1993, The Representation of Geoscience Information for Data Integration, Nonrenewable Resources, 2(2): 122-139.

Cressie, N. A. C., 1993, Statistics for Spatial Data, Wiley Series.

Davis, J. C., 1986, Statistics and Data Analysis in Geology, John Wiley & Sons, Inc.

Efron, B. and Tibshirani R.J., 1993, An Introduction To the Bootstrap, Chapman & Hall.

Heckerman, D., 1986, Probabilistic Interpretations for MYCIN's Certainty Factors, in Kanal, L.N., and Lemmer, J.F., eds., Uncertainty in artificial intelligence: New York, Elsevier, 167-196.

Jolliffe, I. T. (1986) Principal Component Analysis, Springer-Verlag.

Lee, K., Kwon, B. D. and Chi, K. H., 1995, Multivariate Analysis of Geochemical Data for Mineral Potential Mapping in the Taebaek area, Jour. Geol. Soc. Korea. 31(6): 567-575.

Moon, W. M., 1990, Integration of Geophysical and Geological Data using Evidential Belief Function, IEEE Trans. on Geoscience and Remote sensing, 28(4): 711-720.

Park, N. W., Lee, K., Chi, K. H. and Kwon, B.D., 1999, Application to Statistical Analysis Method for Multiple Geoscience Data Integration: Perspectives of GIS Spatial Analytical Functionality, Proc. Of KSRS spring meeting, 99-104.

Rostirolla, S.P., Soares, P. C. and Chang, H. K., 1998, Bayesian and Multivariate Methods Applied to Favorability Quantification in Reconvavo Basin and Ribeira Belt, Brazil, Nonrenewable Resources, 7(1): 7-23.

Vulkan, U. and Duval J. S., 1993, Multivariate Statistical Analysis of Geophysical Data in Nevada, Geophysics, 58: 749-755.

Wright, D.F. and Bonham-Carter, G. F., 1996, VHMS Favourability Mapping with GIS-based Integration Models, Chiisel Lake-Anderson Lake area, Geological Survey of Canada, Bulletin 426: 339-376, 387-401.