자원환경지질, 제38권, 제6호, 663-673, 2005
Econ. Environ. Geol., 38(6), 663-673, 2005

지화학자료를 이용한 금·은 광산의 배태 예상지역 추정-베이시안 지구통계학과 의사나무 결정기법의 활용

황상기[1]* · 이평구[2]

[1]배재대학교 토목환경공학과, [2]한국지질자원 연구원 지질환경재해 연구부

Prediction of the Gold-silver Deposits from Geochemical Maps - Applications to the Bayesian Geostatistics and Decision Tree Techniques

SangGi Hwang[1]* and PyeongKoo Lee[2]

[1]Department of Civil & Geotechnical Engineering, Paichai University, Daejeon 302-735, Korea
[2]Geological & Environmental Hazards Division, Korea Institute of Geoscience and Mineral Resources, Daejeon 305-350, Korea

This study investigates the relationship between the geochemical maps and the gold-silver deposit locations. Geochemical maps of 21 elements, which are published by KIGAM, locations of gold-silver deposits, and 1:1,000,000 scale geological map of Korea are utilized for this investigation. Pixel size of the basic geochemical maps is 250m and these data are resampled in 1km spacing for the statistical analyses. Relationship between the mine location and the geochemical data are investigated using bayesian statistics and decision tree algorithms. For the bayesian statistics, each geochemical maps are reclassified by percentile divisions which divides the data by 5, 25, 50, 75, 95, and 100% data groups. Number of mine locations in these divisions are counted and the probabilities are calculated. Posterior probabilities of each pixel are calculated using the probability of 21 geochemical maps and the geological map. A prediction map of the mining locations is made by plotting the posterior probability. The input parameters for the decision tree construction are 21 geochemical elements and lithology, and the output parameters are 5 types of mines (Ag/Au, Cu, Fe, Pb/Zn, W) and absence of the mine. The locations for the absence of the mine are selected by resampling the overall area by 1km spacing and eliminating any resampled points, which is in 750m distance from mine locations. A prediction map of each mine area is produced by applying the decision tree to every pixels. The prediction by Bayesian method is slightly better than the decision tree. However both prediction maps show reasonable match with the input mine locations. We interpret that such match indicate the rules produced by both methods are reasonable and therefore the geochemical data has strong relations with the mine locations. This implies that the geochemical rules could be used as background values of mine locations, therefore could be used for evaluation of mine contamination. Bayesian statistics indicated that the probability of Au/Ag deposit increases as CaO, Cu, MgO, MnO, Pb and Li increases, and Zr decreases.

Key words : gold-silver deposit, prediction of mining location, geochemical maps, Bayesian statistics, decision tree technique

지화학 자료의 공간적 분포와 금은광산의 공간적 분포사이의 상관관계를 조사하였다. 활용된 자료는 한국자원연구소에서 발간된 지화학도 중 21개 원소에 대한 도면과, 현재까지 파악된 광산의 위치도면 및 1 : 100만 지질도이다. 지화학도는 250 m 등간격의 격자형 화소로 제작된 도면 중 통계분석을 위하여 1km 간격의 자료를 추출하여 분석하였으며, 광산위치의 지화학 자료 역시 250 m 간격의 화소에서 추출하여 분석을 수행하였다. 광산과 지화학자료의 공간적인 상관분석은 베이시안 중첩법과 의사결정나무 기법을 활용하였다. 베이시안 통계기법은 각 지화학도에 분포하는 원소의 화소값을 올림차순으로 정열한 후 자료의 개수가 각기 5, 25, 50, 75, 95, 100%에 해당하는 등급을 나누어 모든 지화학도를 6개의 등급을 갖는 도면으로 재분류 하였다. 각 등급에 속한 광산의 개수를 대상으로 광산이 발생할

*Corresponding author: sghmap@pcu.ac.kr

확률이 계산되었으며, 이 확률을 취합하여 최종 사후확률이 계산되었으며, 사후확률로 광산이 배태될 예측 도면이 작성되었다. 금/은, 동, 철, 납/아연, 텅스텐광산 및 광산이 존재하지 않는 위치에 해당하는 지화학 자료와 암상을 기준으로 의사결정나무를 학습시키고, 학습된 결과를 전체 자료에 적용하여 예측도면을 작성하였다. 광산이 존재하지 않은 지역을 추출하기 위하여 지화학도의 화소를 1 km 간격으로 추출한 후 이들 중 광산과 750 m 이내에 있는 자료는 제외시키는 알고리듬을 활용하였다. 예측결과 베이시안 방법에 의한 광산의 위치 예측이 의사결정나무에 의한 예측보다 상대적으로 정확함이 확인되었다. 그러나 두 방법 모두 공히 기존의 광산위치를 적절히 예측하고 있어서 지화학 자료는 광산의 위치와 밀접한 관계를 갖고 있음이 확인되었다.

주요어 : 금.은광산, 광산의 예측, 지화학도, 베이시안통계기법, 의사결정나무기법

1. 서 언

광산이 배태될 수 있는 위치를 통계적으로 예측하는 연구가 베이시안 통계와 GIS를 이용하여 시도되고 있다(Bonham-Carter *et al.*, 1994; Harris *et al.*, 1995; Wright and Bonham-Carter, 1996; Carranza and Hale, 1999; Asadi and Hale, 2001). 이러한 연구는 광산을 배태시키는데 영향력이 큰 지질학적 요인들을 선별하여 광역적인 도면으로 작성하고, 그 도면에 현재 광산의 위치를 중첩한 후, 중첩된 광산의 빈도를 정량화된 통계치로 계산한 후 통계치를 중첩하여 광산이 배태될 가능성에 대한 최종 확률을 구하는 방법으로 진행된다. 그러므로 광산과 관련된 지질학적 요인들이 적절히 선택되고 정확히 도면화 되었을 경우, 광역적인 예측이 가능할 수 있는 우수한 연구방법이다.

지구통계학을 이용한 광산의 예측에 관한 연구에서 중요한 요인으로 고려된 항목들은 지질도, 중력탐사도와 같은 지구물리 자료, 광역적 구조지질 자료 및 위성의 분광자료 등이다. 구조지질 자료를 제외하고는 대부분 적절한 정밀도로 구할 수 있는 광역적인 도면이므로 이들이 널리 활용되어 왔다. 지화학도는 여기에 첨가 할 수 있는 중요한 항목일 수 있다. 지화학 분포도는 지질도와 매우 밀접한 관계를 갖고 있으며(Hwang *et al.*, 2005) 금속광산의 경우 광산의 성인과도 밀접한 관계가 있을 것이 예상된다. 그러므로 광역적 지구통계 분석에 매우 적절한 항목일 수 있다. 그러나 광역적인 지화학 분포도를 광역적 지구통계학 분석에 이용한 사례는 극히 드문 실정이다.

선진국을 중심으로 활발히 진행 중인 지화학도 atlas의 발간 현황(IGS, 1978; Webb *et al.*, 1978; Meyer *et al.*, 1979; Bolviken *et al.*, 1986; Lahermo *et al.*, 1996)은 지화학도의 체계적인 응용방안에 관한 새로운 연구를 필요로 하고 있다. 국내에서도 한국지질자원연구원에서 전국의 지화학도가 21개 원소를 대상으로 작성된바 있으며(신성천, 2001), 이와 같은 광역적 자료는 지구통계학을 이용한 분석을 위한 매우 중요한 자료라 할 수 있다.

본 연구에서는 전국을 대상으로 한 21개 원소의 분포도, 지질도, 713개의 광산 위치 도면 등을 취합, 분석하여 광산과 화학원소 및 지질의 상관관계를 유추하고, 가능성 높은 광산 위치의 예측 도면의 작성을 목적으로 한다. 자료의 분석은 기존에 흔히 사용되어온 베이시안 기법과 데이터마이닝 기법을 활용하였다.

본 연구에 활용된 자료는 전국 규모의 21개 원소에 관한 지화학도(신성천, 2001)와 713개의 금·은광산 위치가 점기된 광산위치도(Fig. 1) 및 1 : 1,000,000 축척 지질도(강필종, 1995)이다. 모든 도면은 1 km 간격의 그리드 자료로 처리하여 통계계산을 수행하였다. 이 과정에서 광산의 위치는 점원의 위치를 1 km^2 단위의 면적으로 변환하여 면적에 관련된 확률로 계산을 수행하였다. 자료처리에 활용된 s/w는 도면처리를 위해 Arc View와 Map Object가 활용되었으며, 데이터베이스 및 통계계산은 Visual Basic과 Excel Macro Language로 제작된 알고리듬을 활용하였다.

2. 분석과정 및 결과

2.1. 베이시안 분석

일반적으로 광산의 위치를 이용해 조사지역에 대한 사전확률을 구하고 이에 영향을 미치는 다양한 조건부 독립확률을 중첩하여 광산에 배태될 가능성에 대한 최종 사후확률을 구하는 과정이다. 예를 들어 광산이 존재하는 전체영역에서 광산의 발생확률을 사전확률 P{D}라 하고, 광산에 영향을 미칠 수 있는 요인의 등급별 확률들을 $P\{B_1|D\}$, $P\{B_2|D\}$...$P\{Bn|D\}$이라 가정하면, 사후확률 P{D|B}는 사전확률 P{D}에 조건부 확률을 가산한 결과이다. 여기에서 사전확률은 연구지역(남한지역)에 금은광산이 존재할 전체적인 확률을 의

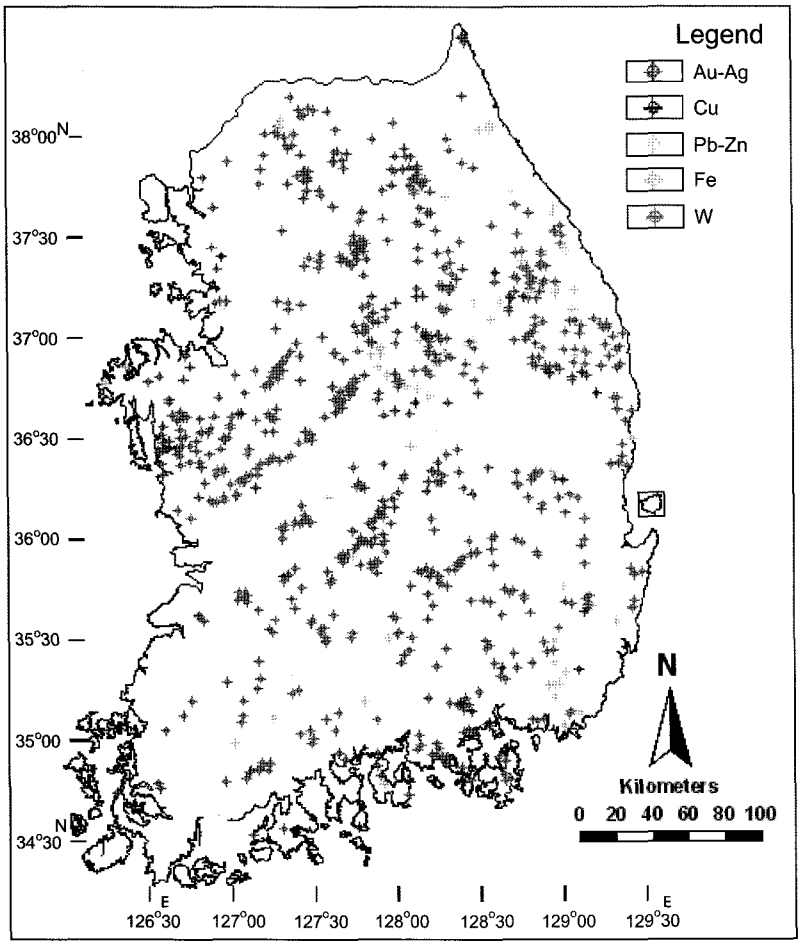

Fig. 1. Localities of the 5 kinds of metallic mines. Note that nearly 80% of them are gold-silver deposits.

미하며 조건부 확률들은 지질도의 암상, 지화학 원소 분포도에서 각 등급별 광산이 존재할 확률을 의미한다. 이들의 계산은 (식 1)과 같이 표현되며, 이를 로짓 형태로 변형하면 (식 2)와 같이 표현된다(Bonham-Carter *et al.*, 1994).

$$P\{D|B_1\cap B_2\ldots\cap B_n\} = P\{D\}\frac{P\{B_1|D\}}{P\{B_1\}}\frac{P\{B_2|D\}}{P\{B_2\}}\cdot\cdot\frac{P\{B_n|D\}}{P\{B_n\}} \quad (식\ 1)$$

$$\log_e O\{D|B_1\cap B_2\cap B_3\ldots\cap B_n\} = \log_e O\{D\} + \sum_{j=1}^{n} W_j^+ \quad (식\ 2)$$

(식 2)의 조건부 확률 W^+는 요인의 가중치로서 아래와 같이 정의되는 각 요인의 sufficiency ratio LS (식 3)를 양의 우도 W^+로 변환한 값(식 4)이다.

$$LS = P\{B|D\}/P\{B|\bar{D}\} \quad (식\ 3)$$

$$W^+ = \log_e LS \quad (식\ 4)$$

양의 우도는 요인의 중요도를 존재할 확률로 표현한 가중치이다. 한편 요인이 존재하지 않을 확률로 표현한 음의 가중치는 아래와 같은 necessity ratio LN으로 표현되며 이를 로그화 한 지수가 음의 우도인 W-가 된다. 즉 음의 우도는 요인의 중요도를 존재하지않을 확률로 표현한 가중치이다.

$$LN = P\{\bar{B}|D\}/P\{\bar{B}|\bar{D}\} \quad (식\ 5)$$

$$W^- = \log_e LN \quad (식\ 6)$$

사후확률에 영향을 미치는 가중치는 양의 우도가 높으면 커지며 음의 우도가 낮으면 커지게 된다. 이러한 관점에서 광산배태 가능성을 상대적으로 표현하는데 가치 있는 척도인 contrast 값은 양의 우도와 음의 우도

의 차이로 아래와 같이 정의된다.

$$C=\log(LS/LN)=W^{+}-W^{-} \quad (식\ 7)$$

본 연구에서 계산된 각 화소의 가중치는 (식 7)에 의해 계산된 contrast 값으로서 일반적인 사후 확률의 값은 아니다. 그러나 Bonham-Carter *et al.* (1989)가 지적하였듯이 contrast 값은 본 연구와 같이 광산배태의 상관관계를 상대적으로 비교하는데 매우 유용하게 활용될 수 있으므로 이를 적용하였다.

지화학 자료를 등급별로 분류하여 각 등급에 해당되는 광산의 개수를 파악함으로 등급별 확률을 계산하였다. 자료의 등급분류는 자료를 오름차순으로 정렬한 후 전체 자료의 개수 중 특정 비율(퍼센트)에 속하는 자료값을 분류의 경계로 지정하는 퍼센타일 기법을 활용하였다. 선택된 분류경계는 자료개수의 비율이 각기 5, 25, 50, 75, 95, 100%에 해당하는 자료값을 경계로 설정하였다. 21개 원소(Al_2O_4, Ba, CaO, Co, Cr, Cu, Fe_2O_3, K_2O, Li, MgO, MnO, Na2O, Ni, Pb, Rb, SiO_2, Sr, TiO_2, V, Zn, Zr)에 대한 퍼센타일 등급분류의 경계는 Table 1과 같다.

공간적인 통계분석은 Fig. 1에 점기 된 713개의 금·은광의 광산위치를 이용하여 수행 되었다. Table 1의

Table 1. Boundary values of 6 classis of each elements (in ppm scale).

	Range	Al_2O_3	Ba	CaO	Co	Cr	Cu	Fe_2O_3	K_2O
Class 1	0-5%	11.72	730.00	0.40	4.72	25.20	9.92	3.05	1.99
Class 2	6-25%	14.12	1050.00	0.73	8.99	45.80	16.00	4.48	2.60
Class 3	26-50%	15.57	1267.00	1.18	13.10	71.02	23.00	5.71	3.04
Class 4	51-75%	17.11	1507.02	1.90	17.40	104.00	32.62	7.13	3.51
Class 5	76-95%	19.34	1920.00	4.05	26.50	167.89	57.00	9.41	4.36
Class 6	96-100%	>19.34	>1920	>4.05	>26.5	>167.89	>57	>9.41	>4.36

	Li	MgO	MnO	Na_2O	Ni	Pb	Rb	SIO_2	Sr	TiO_2	V	Zn	Zr
Class 1	19.50	0.52	0.05	0.50	8.07	18.20	70.21	50.32	51.00	0.47	29.02	45.50	33.74
Class 2	30.40	0.91	0.08	1.04	15.00	24.36	111.00	56.60	96.00	0.68	49.00	78.50	53.00
Class 3	42.00	1.33	0.10	1.45	22.09	28.17	141.00	60.43	133.50	0.81	67.00	107.00	69.59
Class 4	57.00	1.82	0.14	2.01	32.00	32.20	178.00	64.48	192.50	0.95	86.00	149.00	90.00
Class 5	89.60	2.96	0.22	3.12	50.79	42.00	255.00	70.85	341.61	1.31	130.41	272.00	137.00
Class 6	>89.6	>2.96	>0.22	>3.12	>50.78	>42	>255	>70.85	>341.6	>1.31	>130.41	>272	>137

Table 2. Likelihood ratios of the each classes of chemical elements with respect to the localities of the gold-silver deposits.

A. positive likelihood ratio

	Al_2O_3	Ba	CaO	Co	Cr	Cu	Fe_2O_3	K_2O	Li
Class 1	0.27	0.11	-0.35	-0.06	-0.17	-0.53	-0.19	0.01	-0.37
Class 2	-0.06	-0.09	-0.38	-0.28	-0.16	-0.50	-0.23	-0.15	-0.05
Class 3	-0.07	-0.08	-0.02	-0.07	-0.08	-0.16	-0.13	-0.06	-0.08
Class 4	0.09	0.05	-0.04	0.08	0.02	0.00	0.20	0.04	0.06
Class 5	0.04	0.08	0.16	0.18	0.25	0.36	0.02	-0.01	0.08
Class 6	-0.60	0.10	0.59	0.06	-0.09	0.50	0.18	0.58	0.21

B. negative likelihood ratio

	Al_2O_3	Ba	CaO	Co	Cr	Cu	Fe_2O_3	K_2O	Li
Class 1	-0.01	0.00	0.01	0.00	0.01	0.01	0.01	0.00	0.01
Class 2	0.01	0.02	0.06	0.05	0.03	0.08	0.04	0.03	0.01
Class 3	0.02	0.03	0.01	0.02	0.03	0.05	0.04	0.02	0.03
Class 4	-0.04	-0.02	0.02	-0.03	-0.01	0.00	-0.09	-0.02	-0.02
Class 5	-0.01	-0.02	-0.05	-0.05	-0.07	-0.12	-0.01	0.00	-0.02
Class 6	0.02	0.00	-0.04	0.00	0.00	-0.03	-0.01	-0.03	-0.01

MgO	MnO	Na_2O	Ni	Pb	Rb	SIO_2	Sr	TiO_2	V	Zn	Zr
0.01	0.01	-0.01	0.01	0.01	-0.02	0.00	0.00	0.01	0.00	0.01	-0.01
0.06	0.04	0.01	0.05	0.06	0.08	0.01	0.04	0.00	0.06	0.06	-0.03
0.06	0.01	0.00	0.04	0.03	0.01	-0.05	0.03	-0.01	0.00	0.03	-0.08
-0.02	0.00	-0.03	-0.05	0.02	0.00	-0.05	-0.06	0.01	-0.04	-0.03	0.04
-0.08	-0.04	0.00	-0.06	-0.07	-0.05	0.06	-0.02	0.01	-0.04	-0.02	0.05
-0.03	-0.02	0.01	0.00	-0.04	-0.02	0.01	0.00	-0.02	0.01	-0.05	0.01

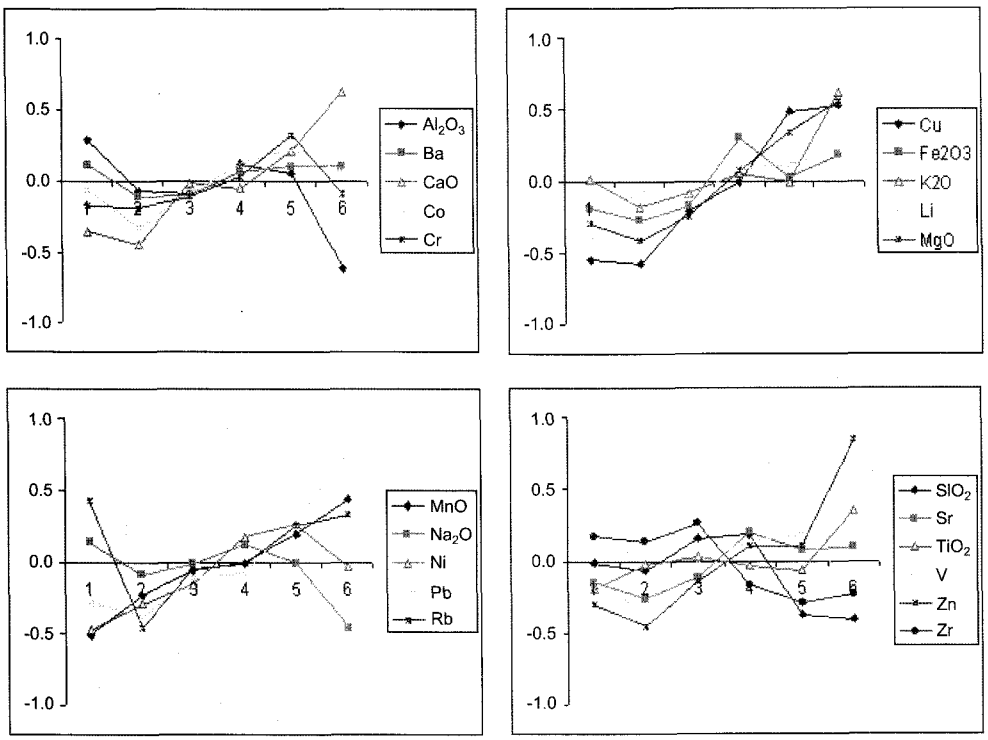

Fig. 2. Variation of the contrast values of each chemical elements. Vertical axes represents the contrast values, whereas the horizontal axes are the 6 divisions.

Table 3. Likelihood ratio and contrast values of the lithologic unites of the 1:1,000,000 scale geological map.

Symbol	Geological legend	Area (m^2)	Number of Mines	W^+	W^-	C
AR1	Rangrim group, gneiss, migmatitic gneiss	15,502,150,632	150	0.30	-0.07	0.37
AR2	Granitic gneiss	5,789,131,141	44	0.06	0.00	0.07
AR3	Porphyroblastic gneiss	2,930,144,372	14	-0.40	0.01	-0.41
C	Chosun super group, Hwangju group	471,734,693.2	2	-0.52	0.00	-0.52
E	Pyeongan group	899,567,306.7	13	0.70	-0.01	0.71
J1	Daedong group	775,391,860.1	5	-0.10	0.00	-0.10
Jgr	Daebo granite, Tancheon complex	21,394,495,258	167	0.09	-0.03	0.11
Jgr1	Foliated granite	4,230,541,983	23	-0.27	0.01	-0.28
K1	Sindong group, Hanbongsan group, Pakcheon group., Pongcheon-bong group	3,311,139,284	19	-0.22	0.01	-0.23
K2	Hayang group, Ponghwasan group, Neungju group, Jinan group	8,993,569,409	58	-0.10	0.01	-0.11
K3	Yucheon group, Jaedeok group	11,806,924,356	53	-0.46	0.05	-0.51
Kgr	Amnokgang complex, Bulguksa granite	5,094,265,827	31	-0.16	0.01	-0.17
O	Great limestone group, Singok group, Mandal group, Sangsori group	1,951,788,551	27	0.66	-0.02	0.68
Og1	Ogcheon group	886,580,166.2	9	0.35	0.00	0.36
Og2	Ogcheon group	1,094,813,832	10	0.25	0.00	0.25
Og3	Ogcheon group	512,244,833	3	-0.20	0.00	-0.20
P	Tuman group, Janggi group	458,446,551.5	4	0.20	0.00	0.20
PALgr	Namgang complex, Chongjin complex, Tumangang complex	493,381,469	7	0.69	0.00	0.69
PALv	Paleozoic basic volcanic rocks	180,503,367.8	1	-0.25	0.00	-0.25

Table 3. Continued.

Symbol	Geological legend	Area (m^2)	Number of Mines	W^+	W^-	C
PR1	Seosan group, Yulri group, Macheonryong group, Hwanghae group	2,186,588,549	29	0.62	-0.02	0.64
PR2	North type of Sangwon group, Yecheon group, Jangrak-Euiam group	2,783,106,008	5	-1.38	0.02	-1.40
PR3	South type of Sangwon group, Taean formation	1,573,474,866	8	-0.34	0.00	-0.34
PRan	Sancheong anorthosite	237,045,684.6	1	-0.53	0.00	-0.53
PRgr	Bucheon, Hongjesa granite, Seosan granite gneiss	2,753,238,029	25	0.24	-0.01	0.25
Tgr	Hyesan complex, Pyonggang complex	726,719,068.9	2	-0.95	0.00	-0.96

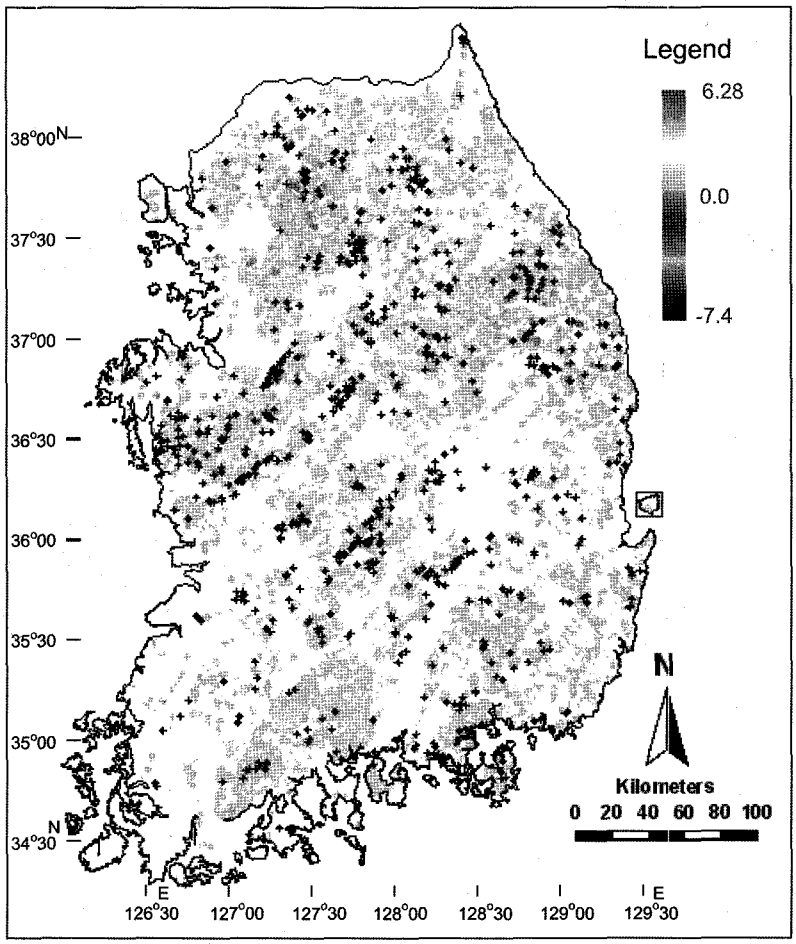

Fig. 3. Constructed probability map by the Bayesian method. Cross points in the map represent the localities of the gold-silver deposits.

기준 등급에 해당하는 화소의 개수와 동일등급 내에 해당되는 광산의 개수를 이용해 계산된 양과 음의 우도는 Table 2와 같으며 contrast 값은 Fig. 2와 같다.

위와 동일한 방법으로 1 : 1,000,000 축척 지질도(강필종, 1995)의 암상을 분류하여 각 암상에 속한 광산의 개수를 고려하여 Table 3과 같은 우도와 contrast 값이 계산되었다. 전국 지형도의 1 km 단위 화소에서 지화학의 등급별 우도와 지질도 암상의 등급별 우도를 식 3-7의 과정으로 취합하여 각 화소의 contrast 값이 계산되었다. 이와 같이 계산되어 도면화 된 결과는 Fig. 3과 같다.

2.2. 의사결정나무

의사결정나무기법은 다양한 데이터마이닝 기법 중 하나로, 데이터마이닝은 대형 데이터베이스 내부에 존재하는 자료의 구조(patterns, statistical models, relationships)를 추출하는데 매우 유용하다. 즉 데이터마이닝은 다량의 자료 내부에 존재하는 규칙들(패턴이나 변수들이 규칙성 등)을 전산으로 추출하고, 추출된 규칙을 이용해 무지의 상황에서 결론을 추정하는 기법이다. 이 기법은 의학, 유통학 등 많은 분야에서 성공적인 적용결과를 발표하고 있어 그 응용분야가 급속히 팽창하고 있다. 그러나 지구과학이나 공학분야의 적용사례는 아직 미약한 실정이다.

의사결정나무는 자료의 분류를 계층적(hierarchical)으로 수행하며, 각 분류단계에서 하나 이상의 가지 구조로 자료의 특성을 분류한다. 자료가 다양한 속성을 갖고 있을 때, 분류에 가장 큰 영향을 미치는 속성을 선택하여 가지로 분류하며, 각 가지에서 다시 가장 적합한 속성을 선택하여 분류를 진행한다. 여기에서 가지가 생성되는 마디(node)가 정의되며, 첫 번째의 마디를 뿌리마디(root node)라 정의한다. 의사결정나무를 이용한 분류의 목적은 속성의 특성에 따라 입력된 자료를 특정 카타고리로 분류하는데 있다. 그러므로 입력 자료가 주어지면, 속성에 값에 의해 계층적으로 분류해 나가서 자료의 카다고리를 설정하는 것을 목적으로 한다. 여기에서 분류될 최종 카타고리는 나무의 잎사귀(leaf)라는 의미에서 잎이라 정의한다. 의사결정나무의 각 단계에서 가지로 분류된 자료는 다음 단계의 가지로 분류되기도 하지만, 가지 자체가 잎일 수도 있다. 이와 같이 의사결정나무는 입력된 자료가 최종 분류단계인 잎이 될 때 까지 위에서 아래로 분류를 지속해 나가는 것이다.

의사결정나무를 제작하는 기법은 매우 다양하며, ID3 (Quinlan, 1986), classification and regression tree (CART) (Breiman, 1984), Chi-Square Automated Interaction Detection (CHAID) (Kass, 1980), C4.5 혹은 See5/C5.0 (Quinlan, 1993) 등이 있다. 본 연구에서 사용하는 C4.5 알고리듬은 마디에 지정될 속성의 선택을 위해서 정보이론(information theory) (Shannon, 1949)을 이용한다. 정보이론은 자료의 분포양상을 정량화 한 엔트로피라 불리는 지수를 활용한다.

$$\text{Entropy}(p_1, p_2, ..., p_n) = -p_1\log p_1 - p_2\log p_2 ... - p_n\log p_n$$

위에서 "p_n"은 하나의 속성군에서 자료의 집합이 갖는 확률이다. 의사결정나무의 마디에서 가지의 분할은 정보이득(gain)을 분류정보(split info)로 나눈 정보이득 비율(information gain ratio)을 계산하여 결정한다.

$$\text{gain ratio}(X) = \frac{\text{gain}(X)}{\text{split info}(X)} \quad \text{(식 9)}$$

정보이득은 각 클래스의 각 마디에 할당될 적정성을 계산한 인포메이션 값(info(T))과 마디에서 n개의 가지가 분할될 경우 이 분할로 인해 파생되는 인포메이션 값($\text{info}_x(T)$)의 차이로 정의된다.

$$\text{gain}(X) = \text{info}(T) - \text{info}_x(T) \quad \text{(식 9)}$$

인포메이션 값은 아래와 같이 계산된다.

$$\text{info}(T) = -\sum_{j=1}^{k} \frac{\text{freq}(C \cdot T)}{|T|} \times \log_2 \frac{\text{freq}(C_j \cdot T)}{|T|} \quad \text{(식 10)}$$

$$\text{info}(T) = \sum_{j=1}^{k} \frac{|T_i|}{|T|} \times \text{info}(T_i) \quad \text{(식 11)}$$

(식 10)에서 $\text{freq}(C_j \cdot T)$ 는 전체 자료 군에서 클래스에 속하는 자료의 개수를 의미하며, $|T|$ 자료 군 전체의 개수를 의미한다. (식 11)은 마디에서 n개의 가지가 분할될 경우, 이 분할로 인해 파생되는 인포메이션을 계산하는 과정이다.

의사결정나무를 제작하는 단계는 전체 자료를 이용해 각 마디에서 위의 과정을 이용해 가장 적절한 클래스와 가지를 설정하고, 다음 가지로 가서 가지를 마디화하여 위의 과정을 반복하는 과정으로 진행된다. 이러한 반복과정은 가지가 하나의 잎으로 설정되거나 정보이득의 값이 0에 근접하여 더 이상 가지분류가 필요없을 때까지 계속된다.

일반적으로 의사결정나무를 만드는 과정에서 학습자료의 모든 구성원이 특정 분류에 속하도록 분류 알고리듬이 만들어져 있으므로 처음에 제작된 나무구조는 가지가 많을 수밖에 없으며 당연히 "overfitting"이 이루어진다. 이러한 overfitting 문제를 해결하기 위해 정보의 분할이 뚜렷하지 않은 가지를 제거하여 의사결정나무를 단순화 하는 과정이 가지치기이며, 흔히 의사결정나무의 마지막 제작단계에서 활용되는 기법이다.

광산의 위치에 해당하는 21종의 지화학 정보(Al_2O_3, Ba, CaO, Co, Cr, Cu, Fe_2O_3, K_2O, Li, MgO, MnO, Na_2O, Ni, Pb, Rb, SiO_2, Sr, TiO_2, V, Zn, Zr)와 지질도의 암상정보를 학습시켜 의사결정나무를 제작하였다. 학습의 입력 자료는 전기한 지화학과 지질자료이며 출력자료로서 금·은, 동, 철, 납·아연, 텅스텐 및 광산이 배태되지 않는 위치의 6가지 분류를 설정하였다. 광산이 배태되지 않은 위치를 하나의 출력 분류로 설정한 이유는 학습과정에서 지화학 자료가

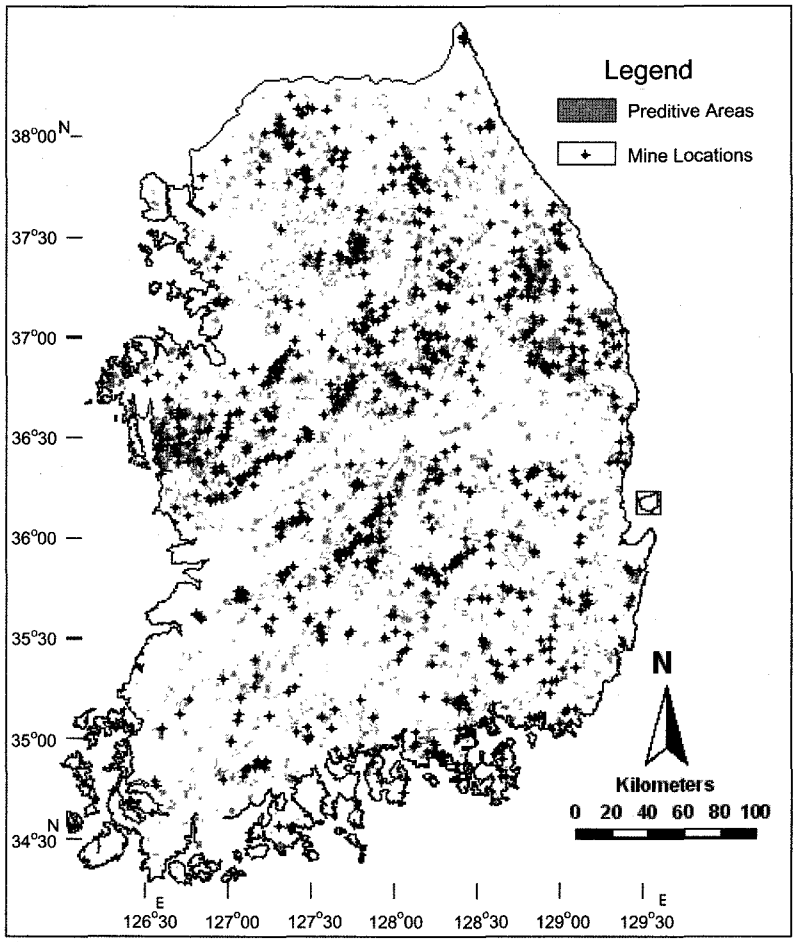

Fig. 4. Predictive Au-Ag map using decision tree technique. Cross points in the map represent the localities of the gold-silver deposits.

너무 광범위하게 적용되어 입력 자료의 영역이 지나치게 확대되는 것을 방지하기 위함이다. 광산이 배태되지 않은 지역의 설정을 위하여 전국의 자료를 1 km 등 간격으로 분할하여 분할 지점의 지화학과 암상 자료를 비교하였다. 광산의 위치에서 750 m 이내에 존재하는 지화학 자료는 광산이 배태되지 않은 영역의 자료군에서 제외시켰다. 이때 750 m의 영역은 소유역의 크기와 지화학 자료가 200 m 간격으로 보간 되었음을 감안하여 임의로 결정한 광산의 영향권이다.

분석에 활용된 s/w는 weka로서, J48 알고리듬을 이용하였다. 입력 자료에 의해 제작된 의사결정나무는 총 572개의 마디에서 333개의 잎을 갖는 구조이다. 가지치기를 수행하여 보았으나, 가지치기 후 광산위치 예측에 적용할 때, 분류가 지나치게 단순화 되어 문제를 발생시키므로, 본 연구에서는 가지치기의 과정을 생략한 원 분류 구조를 예측에 적용하였다. 학습된 나무구조는 금 · 은, 동, 철, 납 · 아연, 텅스텐 및 광산이 배태되지 않는 위치의 6가지 분류 결과를 갖고 있다. 전국의 1 km^2 단위의 지화학자료를 학습된 의사결정나무에 적용하여 각 화소에 해당되는 광산의 종류를 분류하였다. 제작된 도면은 6종류의 주제를 갖고 있으나 자료 분석의 동질성을 유지하기 위하여 이들 중 금.은광의 예측결과만을 선별하여 Fig. 4에 도식하였다.

2.3. 계산결과에 대한 검증

본 연구는 알려진 광산 위치를 이용하여 지화학 원소와 지질도의 요인이 광산에 배태에 미치는 영향을 전산학습 시켜서 학습된 규칙을 적용하여 광산의 배태 가능성을 예측한 것이다. 그러므로 예측된 결과의 정확한 검증은 새로운 광산이 예측된 지역에서 발견되는

Table 4. Likelihood ratio calculations for the Bayesian method.

Range	Legend	Area	Area (%)	Number of mines	%	Likelihood Ratio
-5.08 ~ -2	Unfavourable	272606	17.520	63	8.836	0.504
-2 to 0	Unfavourable	554515	35.638	197	27.630	0.775
0 to 2	Low susceptibility	595697	38.284	299	41.935	1.095
2 to 4	High susceptibility	131002	8.419	140	19.635	2.332
4 ~ 5.79	Extremely high susceptibility	2155	0.138	14	1.964	14.177

것을 확인하는 것이다. 그러나 이러한 검증을 위하여 광산의 탐사를 수행할 수 없는 한계가 있어서 본 연구에서는 예측된 지역에 기존광산이 어느 정도 포함되어 있는지를 확인하기로 하겠다. 즉 기존자료의 입력에 의해 예측된 규칙이 기존자료와 어느 정도의 연관성이 있는지를 확인하는 것이다.

베이시안 예측의 경우는 연관성이 contrast 값으로 계산되며, 그 값은 최하 -5.08에서 최고 5.79까지로 분포한다. 예측의 성공률을 개략적으로 분석하기 위하여 contrast 값을 Table 4와 같이 5개의 등급으로 분류하고 각 등급의 면적비율을 구하였다. 또한 예측된 자료의 개수를 등급별로 분류하여 그 비율을 구하였다. 각 등급별 likelihood ratio는 등급의 면적비율을 예측자료의 비율로 나눠 준 값으로 Table 4에 기술된 바와 같다. 일반적으로 contrast 값이 0 이하일 경우는 광산의 배태 가능성이 없는 것이며 0 이상일 때 값의 크기가 커지는 만큼 예측의 가능성이 높아진다고 생각할 수 있다. 계산된 결과는 전체 자료의 개수 713개 중 36.5%인 260개가 contrast 값 0 이하의 영역에 분포하며 63.5%인 453개가 0 이상의 영역에 분포한다. 한편 면적비로 계산된 likelihood ratio의 경우는 0 이하여 영역에서는 1 이하의 값을 갖으며, 0 이상의 영역에서는 contrast 값이 증가하면 3개의 등급에서 각기 1.09, 2.33, 14.17의 순으로 증가함을 알 수 있다 (Table 4).

의사결정나무에 의해 예측된 영역은 베이시안과는 달리 단일 예측영역으로 분류되며, 전체 913개 자료 중 69.85%인 498개의 자료가 예측영역에 속하며 30.15%인 215개의 자료가 예측영역 밖에 속한다(Table 5).

Table 5. success rate of the mine prediction by decision tree method.

Legend	No of Mines	%
Favourable	498	69.846
Unfavourable	215	30.154

3. 토의 및 결론

인간이 갖고 있는 사물에 대한 분석력 중 대표적인 기능이 기존의 자료와 특정 규칙에 의해 광역적인 현상을 예측하는 것이다. 본 연구와 같이 지화학 자료, 지질자료 등을 이용해 광산을 발생 가능지역을 예측하는 등의 행위가 전형적인 인간의 분석기능이다. 그러나 자료의 종류와 양이 지나치게 많을 경우는 단순 비교와 관찰만으로는 이러한 분석을 수행하기 어렵게 된다. 이러한 문제를 보완하기 위해 통계적, 전산 학습적 방법들의 활용이 필수적이다.

예측된 영역에 포함되는 광산의 개수 비율만으로 비교할 경우 의사결정나무는 69.8%, 베이시안 기법은 63.5%의 적중률을 보이고 있어서 의사결정나무가 다소 우수한 결과를 보이고 있다(Table 4, 5). 한편 강원도와 경기도 일원의 경우 광산의 위치와 예측된 광산의 영역이 잘 일치하는 반면에 충청도 내륙과 경상도 일원의 일부 광산지역이 예측지역과 불일치하는 곳이 관찰된다(Fig. 5).

현행광산이 예측된 지역을 벗어난 이유는 여러 가지일 수 있다. 금광의 경우, 소규모 맥암에 배태된 경우가 많아서 이 맥암의 성분이 하상퇴적물에 반영이 되지 않을 수 있다. 또한 집수유역이 고려되지 않아서 예측지역이 광산에서 조금씩 벗어날 수도 있다. 그러나 더욱 중요한 것은 하상퇴적물의 자료인 지화학도가 지질상황을 전반적으로 반영할 수 있을 것 인가 라는 점이다. 이 문제에 대한 간접적인 확인방법은 없다. 그러나 60% 이상의 광산위치가 예측지역에 포함되는 것을 고려하면, 지화학자료와 금·은 광산의 배태위치는 어느 정도의 상관관계를 갖고 있으며, 베이시안과 의사결정나무 공히 이러한 예측에 활용될 수 있는 적절한 기법이라 할 수 있겠다.

본 연구의 결과를 확실히 검증하는 방법은 현재 제작된 예측도면을 이용하여 광상을 탐사하여 새로운 금은광을 발견하면서 적중확률을 확인하는 것이나 이는 단기간의 연구로 가능하지 않다. 그러므로 본 연구는 이와 같은 예측의 검증을 지속적으로 수행하여야 할 과제로 남기고 있다.

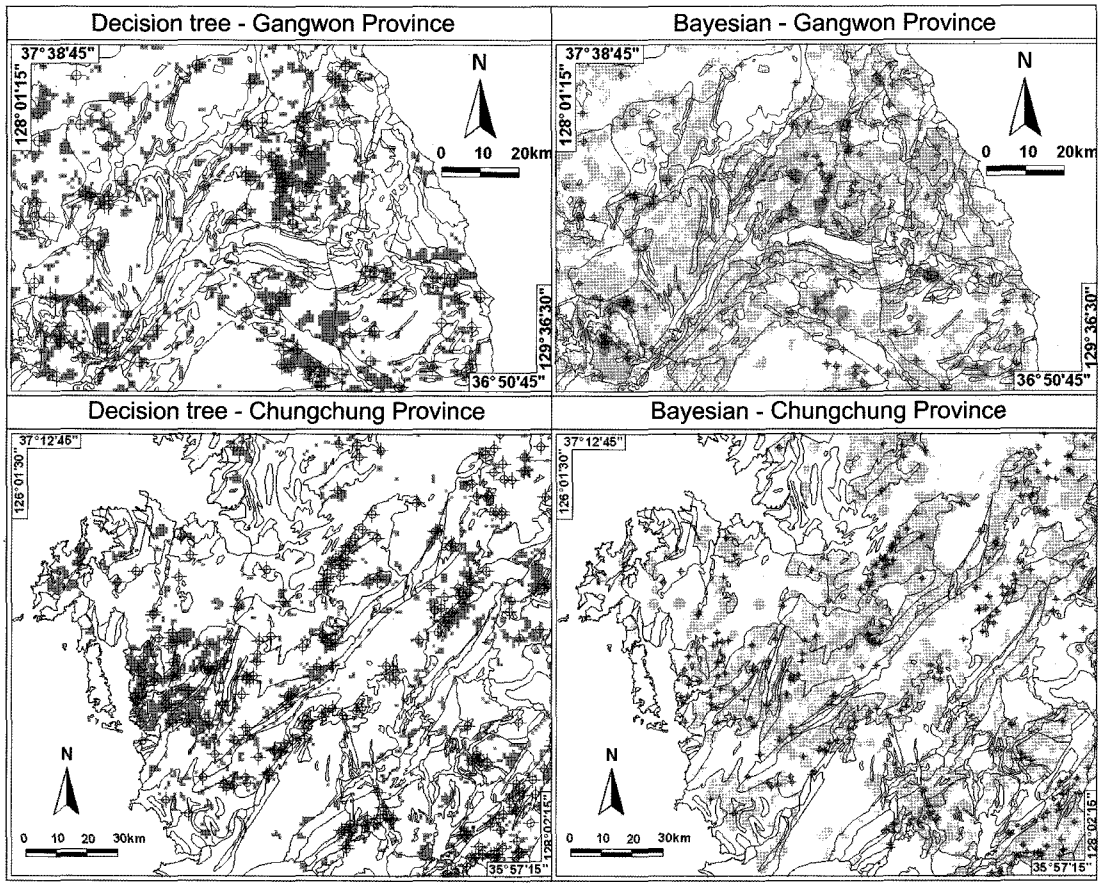

Fig. 5. Detailed prediction maps for the gold-silver deposits and the known mine locations in Chungchung and Gangwon provinces.

본 연구의 목적이 통계기법을 이용하여 광산의 배태 가능 지역을 예측하는 데 있으나 활용된 의사결정나무의 경우 현행광산의 위치에 해당되는 지화학자료를 분석하여 자료의 패턴을 유추한 것이므로 금·은광과 지화학의 관계를 Fig. 2와 같은 상관관계를 통해 유추할 수 있다. Fig. 2의 가로축은 분류 등급으로서 화학성분의 함량이 6단계로 증가함을 나타낸다. 이 도표의 세로축은 contrast 값으로서 이 값이 높으면 금·은광의 배태 확률이 높아진다 할 수 있다. 그러므로 특정원소가 등급이 높아지면서 점이적으로 contrast 값이 증가할 경우는 이 원소는 금·은광의 배태지역과 높은 상관관계를 갖는다고 할 수 있다. 한편 점이적인 감소 역시 음의 상관관계를 나타내는 지표가 될 수 있다. 이러한 관점에서 Fig. 2에 나타난 Ba, CaO, Cu, Li, Mn, Pb는 금·은광과 양의 상관관계를 보이며, Zr은 음의 상관관계를 보이고 있음을 알 수 있다. 이러한 화학원소의 상관관계와 광상학의 지화학적 의미는 본 연구의 법위 밖에 있다. 그러나 이러한 연관성에 대한 광상학적 지화학적인 추가 연구는 과소평가 할 수 없는 중요한 당면 과제라 할 수 있다.

사 사

이 연구는 한국지질자원연구원 기관고유사업(전국 금속광화대 권역별 중금속재해 전과정 평가 및 자연안정화기술 개발, KR-04(연차)-13-1)의 지원으로 수행되었으며 이에 감사드립니다. 세심한 조언을 하여 논문의 발전에 도움을 주신 두 심사위원께 감사드립니다.

참고문헌

강필종 (1995) 한국 지질도, 축척 1 : 1,000,000. 한국자원연구소, 성지문화사.

신성천 (2001) 한국 지구화학 지도책(1:700,000), 한국지질자원연구원.

Asadi H.H. and Hale M. (2001) Apredictive GIS model for mapping potential gold and base metal mineralization in Takab area, Iran. Computer & Geosciences v.27, p. 901-912.

Bolviken, B. (1986) Geochemical atlas of northern Fennoscandia, scale 1:4.000.000. Geological Survey of Sweden, 19pp, 155 maps.

Bonham-Carter, G.F., Agterberg, F.P. and Wright, D.F. (1989) Weights of evidence modelling: a new approach to mapping mineral potential. In: Agterberg, F.P., Bonham-Carter, G.F., (Eds.), Statistical Applications in the Earth Sciences. Geological Survey of Canada Paper, 89-9, p. 171-183.

Bonham-Carter, G.F. (1994) Geographic Information Systems for geoscientists, modeling with GIS. Pergamon Press, Oxford, 398pp.

Breiman, L., Friedman J.H., Olshen R. and Stone C. (1984) Classification and Regression Trees, Wadsworth International Group, California.

Carranza, E.J.M. and Hale, M. (1999) Geological-constrained probabilistic mapping of gold potential, Baguio District, Philippines. Geocomputation 99, July 2528, Fredericksburg, Virginia, Conference Volume on CD-ROM.

Debes J.D. and Urrutia R. (2004) Bioinformatics tools to understand human diseases. Surgery 135, v. 6, p. 579-585.

Harris, J.R., L. Wilkinson, J. Broome, and S. Fumerton, (1995) Mineral exploration using GIS-based favourability analysis, Swayze greenstone belt, northern Ontario, in Proceedings of 1995 Canadian Geomatics, Ottawa, Ontario Canada.

Hwang, S.G., Nguyen Q.P. and Lee P.K. (2005) Reproducibility of a regional geological map derived from geochemical maps, using data mining techniques: with application to Chungbuk province of Korea. Environmental Geology, in press.

IGS. (1978) Geochemical atlas of Gt. Britain: Shetland Islands. Institute of Geological Sciences, London.

Kass, G.V. (1980) An exploratory technique for investigating quantities of categorical data. Applied Statistics 29 v. 2, p. 119-127.

Lahermo, P., Vaananen P., Tarvainen T. and Salminen R. (1996) Geochemical Atlas of Finland, Part 3: Environmental geochemistry - Stream waters and sediments. Geological Survey of Finland, Espoo, 149p.

Meyer, W.T., Theobald P.K., Bloom H. (1979) Stream sediment geochemistry. In: Hood P.J. (ed) Geophysics and geochemistry in the search for metallic ores. Geol Surv Can. Econ. Geol. Rep., v.31, p 411-434

Quinlan, J.R. (1986) Induction of decision trees. Machine Learning 1, v.1, p. 81-106.

Quinlan, J.R. (1993) C4.5: Programs for machine learning. San Francisco: Morgan Kaufmann Publishers.

Shannon, C.E. and Weaver W. (1949) The Mathematical Theory of Communication, University of Illinois Press.

Webb, J.S., Thornton I., Howarth R.J. and Thompson M. (1978) The Wolfson Geochemical Atlas of England and Wales. Clarendon Press, United Kingdom, 69pp.

Wright, D.F. and Bonham-Carter G.F. (1996) VHMS favourability mapping with GIS-based integration models, Chisel Lake-Anderson Lake area, in : Bonham-Carter, Galley, and Hall (eds.): EXTECHI:A multidisciplinary approach to massive sulfide research in the Rusty Lake-Snow Lake greenstone belts, Manitoba. Geolocical Survey of Canada, Bulletin, v. 426, p. 339-376.

2005년 7월 19일 원고접수, 2005년 11월 29일 게재승인.