지화학 자료의 확률론적 불확실성 및 위험성 분석을 위한 지시자 지구통계학의 응용

박 노 욱*

인하대학교 지리정보공학과, 402-751, 인천광역시 남구 용현동 253

Application of Indicator Geostatistics for Probabilistic Uncertainty and Risk Analyses of Geochemical Data

No-Wook Park*

Department of Geoinformatic Engineering, Inha University, Incheon 402-751, Korea

Abstract: Geochemical data have been regarded as one of the important environmental variables in the environmental management. Since they are often sampled at sparse locations, it is important not only to predict attribute values at unsampled locations, but also to assess the uncertainty attached to the prediction for further analysis. The main objective of this paper is to exemplify how indicator geostatistics can be effectively applied to geochemical data processing for providing decision-supporting information as well as spatial distribution of the geochemical data. A whole geostatistical analysis framework, which includes probabilistic uncertainty modeling, classification and risk analysis, was illustrated through a case study of cadmium mapping. A conditional cumulative distribution function (ccdf) was first modeled by indicator kriging, and then e-type estimates and conditional variance were computed for spatial distribution of cadmium and quantitative uncertainty measures, respectively. Two different classification criteria such as a probability thresholding and an attribute thresholding were applied to delineate contaminated and safe areas. Finally, additional sampling locations were extracted from the coefficient of variation that accounts for both the conditional variance and the difference between attribute values and thresholding values. It is suggested that the indicator geostatistical framework illustrated in this study be a useful tool for analyzing any environmental variables including geochemical data for decision-making in the presence of uncertainty.

Keywords: indicator kriging, geochemical data, uncertainty, risk

요 약: 지화학 자료는 환경 관리를 위한 중요한 환경 변수중 하나로 인식되어 왔다. 지화학 자료는 보통 공간적으로 산재되어 수집되기 때문에, 샘플링 되지 않은 지점에서의 속성값 예측과 더불어 부가적인 분석을 위해 예측에 수반되는 불확실성을 추정할 필요가 있다. 이 논문은 지시자 지구통계학이 지화학 자료의 공간적인 분포값의 제시뿐만 아니라 의사결정을 보조할 수 있는 정보를 제공하기 위해 유용하게 사용될 수 있는지를 예시하고자 한다. 카드뮴 자료의 추정 사례 연구를 통해 확률론적 불확실성 모델링, 위험성 분석 등 지구통계학적 분석의 틀을 제시하였다. 지시자 크리깅을 통해 조건부 누적 분포 함수를 모델링한 후에, 기대값 추정치와 조건부 분산을 카드뮴의 추정값과 정량적 불확실성 추정을 위해 각각 계산하였다. 그리고 확률 임계치와 속성 임계치의 적용을 통해 오염/비오염 지역을 구분하였다. 또한 조건부 분산과 속성값과 임계치값의 차이를 모두 설명할 수 있는 변동 계수를 통해 추가적인 샘플링 지점을 추출하였다. 이 연구에서 적용한 지시자 지구통계학적 분석 틀은 불확실성을 고려한 의사 결정과 관련하여 지화학 자료를 포함한 환경 변수의 분석에 유용하게 사용될 수 있을 것으로 기대된다.

주요어: 지시자 크리깅, 지화학 자료, 불확실성, 위험성

*Corresponding author: nwpark@inha.ac.kr

Tel: 82-32-860-7607

Fax: 82-32-863-1506

서론

토양, 하상 퇴적물, 수계 등에서 수집되는 지화학 원소는 초기에는 자원탐사의 기초자료로 활용되어 왔으나, 1990년대 이후에는 자원탐사 이외에 환경보전 및 관리 등 다양한 분야에서 활용이 가능한 기초 자료로 인식되어 왔다(이진수 외, 1998). 국내에서는 한국지질자원연구원에서 과거 한국자원연구소 시절부터 태백산 광화대 지역에서 하상 퇴적물로부터 채취한 지화학 원소의 분포도와 분석 결과를 제공하고 있으며, 2000년대에 8개 도의 지구화학 지도책을 발간하였다(진명식, 2007). 또한 특정 지질 암상에 대한 지화학적 분석, 지하수 및 하천 오염 분석 등의 연구가 많이 진행되어왔다(오강호 외, 2003; 황상기와 이평구, 2005; 조규성 외, 2007; 오인숙 외, 2008; 위수민과 김은호, 2009; Ko et al., 2009)

지화학 원소의 특성상 분석은 공간적인 함량 분포에 기반을 두고 있는데, 대부분 대상 지역의 일부 지역에서 샘플링이 이루어지기 때문에 샘플링 되지 않은 지점에서의 값을 추정하는 공간 추정 혹은 공간 내삽의 과정을 거치게 된다. 그러나 미지의 지점에서의 값 추정에는 불확실성이 수반되며, 이러한 불확실성은 추후 오염 위험성 분석, 오염 지역의 판단 및 정화 작업 등의 의사결정 과정에 영향을 미치게 된다. 따라서 지화학 원소의 공간적 분포의 정확한 추정과 더불어 불확실성 및 위험성 분석이 실질적인 활용을 위해 필요하다.

이러한 환경 변수의 공간 추정, 불확실성 및 위험성 분석에 지구통계학은 유용한 방법론을 제공한다. 국내에서 고급 내삽 기법으로 알려진 크리깅으로 대표되는 지구통계학은 확률 변수의 개념을 도입하여 조건부 누적 분포 함수(conditional cumulative distribution function: ccdf)의 계산을 통해 예측과 관련된 불확실성을 정량적으로 제시할 수 있다(Goovaerts, 1997; Deutsch and Journel, 1998). 외국에서는 토양 오염, 지하수 오염 등의 환경 관리를 목적으로 공간 분포도 작성뿐만 아니라, 불확실성 모델링을 통한 오염 지역의 구분, 위험성 분포도 작성 등에 지구통계학을 많이 활용해 왔다(Goovaerts et al., 1997; Juang and Lee, 1998; Barabas et al., 2001; Van Meirvenne and Goovaerts, 2001; Saito and Goovaerts, 2003; Saisana et al., 2004; Goovaerts et al., 2005; Goovaerts et al., 2008). 그러나 아직까지 국내에서 지화학 원소를

대상으로 한 불확실성 및 위험성 분석은 수행되지 못하였으며, 물리탐사 분야에서 오석훈과 서백수(2007)에 의해 암반등급의 확률론적 평가 연구가 수행되었다.

지구통계학은 불확실성 모델링을 위해 모수적 방법론과 비모수적 방법론을 제공한다(Goovaerts, 1997). 모수적 방법론은 ccdf의 모델링을 위해 원 자료가 다중 정규 분포를 이룬다는 가정하에 다중 가우시안 크리깅을 통해 얻어진 추정치와 크리깅 분산을 정규 분포의 평균과 분산으로 대체한다. 비모수적 방법론은 ccdf 모델링을 위해 특정 모수를 추정하는 방식이 아닌 이진 변환을 이용하여 ccdf를 직접적으로 모델링한다. 연속형 변수의 경우 특정 임계치의 초과 혹은 미만 여부에 따라 원 자료값을 0과 1로 이진 변환한 후에 지시자 크리깅을 수행하여 특정 임계치보다 작거나 같게 될 조건부 확률을 계산함으로써 ccdf를 모델링한다. 일반적으로 지화학 자료는 정규 분포를 따르지 않고 양으로 치우친 비대칭 분포를 나타내는 경우가 많으며, 관측 센서의 탐지 분해능의 한계로 인해 탐지 한계를 벗어난 관측값이 모두 특정 한계값으로 관측되는 경우가 많다(박노옥, 2009). 이렇게 정규 분포를 따르지 않는 비대칭 분포 자료에 대해 모수적 방법론은 정규 점수 변환(normal score transform)을 수행하여 강제로 정규 분포를 따르게 하나, 탐지 한계값은 임의의 서로 다른 정규 점수를 가지게 되는 한계가 있다. 또한 선행 연구로 박노옥(2009)은 비대칭 분포를 이루는 지화학 자료를 대상으로 정규 크리깅, 로그 정규 크리깅, 다중 가우시안 크리깅과 지시자 크리깅의 성능을 비교하였는데, 지시자 크리깅의 예측 능력이 가장 우수한 것으로 나타났다. 이러한 배경에서 본 연구에서는 불확실성 모델링에 비모수적인 방법론인 지시자 크리깅을 적용하였다.

이 연구에서는 지화학 자료의 공간 내삽, 불확실성 모델링 및 오염 위험도 작성 등의 일련의 과정을 지시자 지구통계학(indicator geostatistics) 기법을 이용하여 제시하고자 한다. 지시자 크리깅을 지화학 원소에 적용하여 우선적으로 ccdf 계산을 통한 불확실성 모델링 후에, 이러한 불확실성 모델링의 부산물로 지화학 원소 위험도를 확률론적으로 정량화하였다. 분포도 작성 이후의 의사 결정을 위해 오염 지역과 비오염 지역을 구분하고자 하였으며, 추가적인 샘플링 지점 또한 추출하고자 하였다(Fig. 1). 태백산 광화대

Fig. 1. Work flow of indicator approach for geochemical data analysis.

일부 지역에서 획득한 지화학 자료를 대상으로 한 사례 연구를 통해 이러한 일련의 분석과정을 예시하였다.

지시자 지구통계학 이론

일반적으로 지구통계학에서 불확실성은 식 (1)과 같이 특정 위치(u)에서의 미지의 속성값 $z(u)$ 을 확률 변수 $Z(u)$ 로 간주하여 특정값 z 보다 작을 확률인 cdf $F(u; z|(n))$ 를 통해 모델링된다.

$$F(u; z|(n)) = \text{Prob}\{Z(u) \leq z|(n)\} \quad (1)$$

여기서 “(n)”은 n개의 샘플링된 주변값 $z(u_\alpha)$ 를 나타낸다.

비모수적 방법론인 지시자 접근법(Journel, 1983)은 원 자료의 비선형 변환에 기반을 두고 있다. 우선 개별 관측값은 미리 정한 K개의 임계치 값에 따라 K개의 지시자 값들로 변환된다. 식 (2)와 같이 샘플링 지점의 속성값이 특정 임계치 z_k 보다 작거나 같으면 1로, 크면 0을 가지도록 지시자 변환을 수행한다.

$$i(u_\alpha; z_k) = \begin{cases} 1 & \text{if } z(u_\alpha) \leq z_k \\ 0 & \text{otherwise} \end{cases} \quad k=1, \dots, K \quad (2)$$

식 (2)에 의해 지시자 변환된 변수들에 대해 실험적 베리오그램을 구한 후에, 베리오그램 모델링을 수행하고, 지시자 크리깅을 K개의 임계치별로 적용하면 일련의 임계치 z_k 에 대한 조건부 확률을 얻게 된다(식 (3)).

$$\begin{aligned} [i(u; z_k)]^* &= [F(u; z_k|(n))]^* = \text{Prob}\{Z(u) \leq z_k|(n)\}^* \\ &= \sum_{\alpha=1}^{n(u)} \lambda_\alpha(u; z_k) i(u_\alpha; z_k) \end{aligned} \quad (3)$$

지시자 크리깅은 임계치별로 개별적으로 수행되고, 크리깅이 non-convex 추정자이기 때문에 확률의 일반적인 특성인 다음 2가지 조건을 만족시키지 않을 수 있다.

$$\begin{aligned} 0 &\leq [F(u; z_k|(n))]^* \leq 1 \\ [F(u; z_k|(n))]^* &\leq [F(u; z_l|(n))]^*, \quad z_k \leq z_l \end{aligned} \quad (4)$$

이러한 order violation 문제와 관련해서 첫번째 조건은 0보다 작거나 1보다 큰 추정값은 0과 1로 재조정하여 해결할 수 있으며, 두번째 조건은 cdf값의 상향, 하향 보정의 결과의 평균값을 통해 만족시킬 수 있다(Deutsch and Journel, 1998).

이렇게 지시자 크리깅을 통해 얻어진 임계치별 조

Fig. 2. Sample data; (a) location map, (b) histogram with descriptive statistics.

건부 확률을 통해 ccdf를 구성하여 미지의 위치 $\mathbf{u}$ 에서의 불확실성을 확률론적으로 표현하게 된다. 일단 지시자 크리깅을 통해 ccdf 모델링이 수행되면, 이 ccdf를 통해 다양한 추후 분석이 가능하다.

우선 특정 임계치보다 클 확률 즉 오염 기준치보다 클 확률은 ccdf를 이용하여 간단히 계산 가능하다.

$$\text{Prob}\{Z(\mathbf{u}) > z_c\} = 1 - F(\mathbf{u}; z_c | (n)) \quad (5)$$

또한 전반적인 연구지역에서의 추정값은 보통 ccdf의 평균값으로 식 (6)과 같이 계산된다.

$$z^*(\mathbf{u}) = \frac{1}{2K} \sum_{k=1}^K \left[F^{-1}\left(\mathbf{u}; \frac{k}{K} \middle| (n)\right) + F^{-1}\left(\mathbf{u}; \frac{k-1}{K} \middle| (n)\right) \right] \quad (6)$$

연구 지역 및 사용 자료

이 연구에서는 지화학 자료에 대한 지시자 크리깅의 적용 가능성을 예시하기 위해 한국자원연구소에 의해 강원도 태백산 광화대의 일부 지역의 198개 지점에서 획득된 하상 퇴적물 기반 카드뮴(Cd) 자료에 대한 사례연구를 수행하였다. 이 사례연구의 목적은 특정 대상 지역의 카드뮴 오염 분포도 작성 및 특성 분석이 아닌, 지시자 크리깅을 이용한 지화학 자료 분석 절차를 예시하는데 있다. 1차 수계에서 패닝(panning)에 의해 채취된 하상 퇴적물의 분석을 통해 얻어진 샘플링 자료를 이용하였다(Fig. 2a). 오염과 관련된 정보가 직접적으로 드러나기 때문에 연구 지역의 명칭과 위치 정보는 제시하지 않았다.

Fig. 2a의 샘플링 자료의 위치를 살펴보면, 대상 지

역의 북쪽과 서쪽에 상대적으로 높은 값들이 존재하고 있다. 기본 통계값을 살펴보면(Fig. 2b), 평균은 3.73 ppm이지만 표준편차가 4.51 ppm으로 높게 나타나고 있으며, 양으로 치우친 분포로 1.26의 양의 왜도(skewness)를 나타내고 있다. 또한 탐지 한계인 0.02 ppm을 나타낸 샘플링 지점이 42개로 전체 198개의 샘플링 지점 중에서 약 21%를 차지하고 있는 것으로 나타났다. 이러한 극단적인 양의 왜도 값과 탐지 한계 값에 많은 자료들이 분포하고 있는 기본 통계치 분석 결과는 이후 분석에서 다중 정규 분포 가정을 근간으로 하는 모수적 방법론보다는 비모수적 방법론의 적용이 더 효율적임을 지시한다. 연구지역에서 구할 수 있는 추가적인 조사 자료가 존재하지 않은 관계로, 198개 샘플링 지점을 2개의 독립적인 집단으로 나누어서 60%에 해당하는 119개 지점은 테스트 자료로, 40%에 해당하는 79개 지점은 검증용 자료로 사용하였다. 이후 분석은 모두 테스트 자료를 대상으로 수행되었다.

지구통계학적 분석 결과

지시자 변환 및 베리오그램 모델링 결과

지시자 크리깅을 적용하기 이전에, 테스트 자료를 대상으로 지시자 변환과 베리오그램 모델링을 수행하였다. 지시자 변환을 위한 임계치는 하위 20%의 자료값이 0.02임을 감안하여 백분위수 8개(20, 30, -, 80, 90%)로 설정하였다. 이러한 8개 임계치별로 실험적 지시자 베리오그램을 계산한 후에 베리오그램 모델링을 수행하였다. Fig. 3은 8개 임계치에 대해 지

Fig. 3. Experimental indicator variograms with the model fitted for each threshold value.

시자 변환된 샘플링 자료의 베리오그램 모델링 결과인데, 임계치값이 커지더라도 공간적 자기상관성은 유지되는 것으로 나타났다.

Ccdf 모델링 및 예측능력 검증 결과

임계치별로 베리오그램 모델링을 수행한 후에, 정규 지시자 크리깅을 수행하여 연구지역의 모든 지점에서 임계치별 조건부 확률을 계산함으로써 ccdf를 모델링하였다. 정규 지시자 크리깅 수행시 최대 사용 자료의 개수는 자료의 공간적 분포를 고려하여 12개로 설정하였으며, 최대 탐색 반경은 각 베리오그램의 상관거리를 고려하여 최대 9,000미터로 하였다. 또한 이산적 ccdf로부터 임의의 속성값에 해당하는 ccdf값을 계산하는 내삽 및 외삽에는 선형 모델을 이용하였다.

Fig. 4는 대상 지역의 특정 2개 지점에서의 지시자 크리깅으로부터 얻어진 ccdf를 나타내고 있다. Point

#1은 주변에 상대적으로 높은 값들이 몰려 있는 지점에 해당하며, point #2는 샘플링 지점과의 거리가 상대적으로 먼 곳에 위치하고 주변에 높고 낮은 값들이 존재하고 있다. Point #1의 경우, 주변의 상대적으로 높은 값들의 영향으로 7번째 임계치값인 7.5 정도에서 ccdf 값이 거의 1에 근접한 양상을 나타내고 있다. 반면, 주변에 샘플링 자료가 존재하지 않는 point #2는 ccdf의 모양이 넓게 퍼져 나타나고 있으며, 8번째 임계치값인 12.248에 도달하여도 ccdf값이 0.71 정도로 작게 나타나고 있다. 즉 이러한 ccdf의 분포 모양은 주변 샘플링 자료값의 크기와 밀도에 영향을 받으며, ccdf의 폭에 의해 예측의 불확실성 정도를 파악할 수 있다. Point #1의 경우 급격하게 ccdf의 값이 증가하는 양상이 나타나고 있어서 상대적으로 ccdf가 넓게 퍼져 있는 point #2에 비해 예측의 불확실성이 작음을 확인할 수 있다. 이러한 ccdf

Fig. 4. Ccdf models for two points marked on the location map of sample data.

의 분포에 따른 불확실성의 판단 여부는 기존 예측 값과 크리깅 분산을 제시하는 일반적인 크리깅과는 차별화된 정보에 해당한다. 단순 크리깅이나 정규 크리깅의 적용 결과 제공되는 크리깅 분산은 샘플 위치로부터 멀어질수록 커지게 되며 값의 분포와는 무관한 정보를 제공한다. 반면, 불확실성 모델링이 가능한 지시자 크리깅을 통해 예측값의 추정뿐만 아니라 ccdf의 분포 모양을 통해 예측에 수반되는 불확실성의 강도를 파악할 수 있다.

이러한 ccdf 모델링을 통해 다양한 정보 추출이 가능한데, 일반적인 관심은 샘플링 되지 않은 지점에서의 추정에 있기 때문에 이 연구에서는 이후 분석 이전에 식 (6)을 이용하여 예측값을 계산하였다. 또한 전술한 ccdf의 분포 모양을 통해 정성적으로 파악이 가능한 예측의 불확실성을 ccdf의 조건부 분산 $\sigma^2(\mathbf{u})$ 을 식 (7)을 이용하여 계산하여 정량적인 불확실성 정보를 함께 제공하였다.

$$\sigma^2(\mathbf{u}) = \sum_{k=1}^{K+1} [\bar{z}_k - z^*(\mathbf{u})][F(\mathbf{u}; z_k | (n)) + F(\mathbf{u}; z_{k-1} | (n))] \quad (7)$$

여기서 $\bar{z}_k$ 는 구간 (z_{k-1}, z_k) 사이의 평균값을 나타낸다.

Fig. 5는 식 (6)과 식 (7)을 이용하여 계산된 연구 지역의 카드뮴 분포도와 이러한 예측과 관련된 조건부 분산 분포도를 나타내고 있다. 카드뮴 분포도를 살펴보면, 연구 지역에서 지질학적으로 대석회암통과 평안계가 분포하는 북부와 서부에서 상대적으로 높은 값들이 나타나고 있으며, 크리깅의 일반적인 특성인

평활화 효과로 값들이 퍼진 분포를 나타낸다. 조건부 분산 분포도에서는 카드뮴 분포값이 작게 나타나는 곳에서는 분산값이 작게 나타나지만, 카드뮴 값이 높고 낮은 값들의 경계 부분에서 상대적으로 분산 값이 높게 나타나고 있다. 즉 주변에 상대적으로 높고 낮은 값들이 존재할 경우, 그 사이에 존재하는 지점에서의 값 예측에 상대적으로 높은 불확실성이 수반됨을 나타낸다. 카드뮴 분포의 추정과 관련된 불확실성과 수계 및 지질 경계와의 연관성을 살펴보기 위해 대상 지역의 수계망, 지질도를 중첩하여 연관성 분석을 수행하였다. 분석결과, 연구지역 북쪽의 카드뮴 분포값이 높게 나타나면서 조건부 분산이 낮게 나타나는 지역은 화강암과 대석회암통에 해당하였지만, 다른 지역에서는 지질 경계와의 뚜렷한 연관성은 나타나지 않았다. 수계망과의 중첩분석 결과, 연구지역의 북서쪽과 서쪽에서 나타나는 불확실성이 높은 지역은 대부분 수계가 끝나는 지역에 해당되었으나, 뚜렷한 경계와의 일치성은 드러나지 않았다. 다만 전술한 것과 같이 불확실성이 높게 나타나는 지역은 대부분 샘플링 값이 아주 높거나 낮은 값들이 서로 근접하여 나타나는 지역에 해당된다. 서쪽 지역은 대석회암층군이 연구지역을 관통하는 대하천에 의해 나뉘어져서 일부 지역은 높은 값들이 나타나고, 일부 지역은 아주 낮은 값들이 나타나서 불확실성이 높게 나타나는 것으로 판단된다. 즉 지질 암상별로 카드뮴 분포의 특성이 다르게 나타나지만, 연구지역의 지형적인 특성으로 분포의 경계가 나타나게 되고 이러한 영향으로 높은 값과 낮은 값의 경계가 형성되어 이

Fig. 5. (a) E-type estimates of Cd, (b) conditional variance.

Table 1. Comparison of prediction capabilities of three kriging algorithms

	MAE	MSE
Ordinary kriging	1.70	7.70
Multi-Gaussian kriging	1.88	11.74
Indicator kriging	1.61	7.31

부근에서 예측의 불확실성값이 크게 나타난 것으로 판단되며, 추후 지질/지형학적 추가 분석이 요구된다.

부가적으로 지시자 크리깅에 의한 카드뮴 예측 정확도를 기존 크리깅 방법론과 비교하기 위해 테스트 자료를 이용하여 정규 크리깅과 다중 가우시안 크리깅을 별도로 적용하였다. 예측 능력의 정량적 비교를 위해 검증 자료에 대한 예측을 수행하는 잭나이프 방법을 적용한 후에, 정량적 척도로 평균 절대 오차(mean absolute error: MAE)와 평균 제곱 오차(mean square error: MSE)를 계산하였다. 검증 결과는 Table 1에 제시되어 있는데, 지시자 크리깅이 정규 크리깅과 다중 가우시안 크리깅에 비해 낮은 평균 절대 오차와 평균 제곱 오차를 나타내어 상대적으로 높은 예측 능력을 나타내었다. 이러한 결과는 박노옥(2009)에 의해 수행된 연구결과와 일치하며, 비대칭 분포를 이루는 자료의 예측에 지시자 크리깅이 우수함을 다시 확인할 수 있었다.

오염 위험성 분포도 작성 결과

지화학 원소와 같은 환경 변수는 공간 분포도 작성에 최종 목적이 아니라 작성된 공간 분포도로부터 오염/비오염 지역의 구분에 따른 오염지역의 정화 등

과 같은 추가적인 의사 결정 과정이 뒤따르게 된다. 이러한 의사 결정 과정을 보조하기 위한 다양한 정보를 지시자 크리깅 결과로부터 추출하였다.

불확실성을 고려한 상태에서 오염/비오염 지역을 구분하는 방법은 크게 3가지로 구분할 수 있다(Saisana et al., 2004). 첫 번째는 지시자 접근에 의한 ccdf 모델링을 통해 확률 계산이 가능하므로, 특정 임계치보다 추정값이 크게 될 확률 분포도를 계산한 후에 확률 임계치를 적용하여 오염/비오염 지역을 구분하는 것이다. 두 번째 방법은 기존 크리깅과 마찬가지로 공간 분포도를 제작한 상태에서 속성값이 특정 임계치 값 이상을 오염 지역으로 구분하는 것이다. 이러한 오염/비오염 지역의 구분은 결정론적으로 이루어지지만 공간 분포도 작성의 불확실성 혹은 오차가 따르기 때문에 이러한 구분에 위험성이 따르게 된다. 세 번째 방법은 오염/비오염 지역의 구분에 따른 의사 결정에 수반되는 비용 효과를 고려하여 기대 손실을 최소화하는 방법이다.

이 연구에서는 처음 2가지 방법론을 적용하여 의사 결정 보조 정보를 추출하였다. 세 번째 방법론의 경우, 오염 지역 혹은 비오염 지역으로 구분하는 의사 결정 과정에서 낭비되는 정화 비용과 거주민의 건강에 미치는 영향에 따른 비용을 모델링해야 한다. 이 때 손실 비용은 토지 이용 상태에 따라 달라질 수 있으며, 경제적인 손실 모델 자체가 타당성이 있어야 한다. 이러한 경제적인 손실 모델의 선정에 어려움이 존재할 수 있어서 이 연구에서는 적용하지 않고 상대적으로 적용이 간편한 처음 2가지 방법론을 적용하였다.

Fig. 6. (a) Probability of exceeding the threshold value of 4 ppm, (b) variation of false classification rates with respect to the change of probability threshold value, (c) classification result on the basis that the probability of exceeding the threshold value of 4 ppm is larger than the probability threshold value of 0.577.

우선 첫 번째 방법론인 확률 임계치를 초과할 경우 오염지역으로 구분하는 방법론을 적용하였다. 지시자 크리깅을 통해 대상 지역의 모든 위치에서의 ccdf 모델을 알고 있기 때문에 특정 임계치 값을 초과할 확률을 계산은 쉽게 수행 가능하다. 이 연구에서는 시범적으로 토양 오염 우려 기준치인 4 ppm을 오염 임계치 값으로 설정하여 4 ppm을 초과할 확률 분포도를 작성하였다(Fig. 6a). 이러한 확률 분포도에서 오염/비오염 지역을 구분하기 위해서는 확률 임계치를 설정해야 하는데, 이 연구에서는 주관적으로 설정하지 않고, 교차 검증을 이용하여 최소 분류 오류를 나타내는 확률 임계치를 계산하였다. 우선 leave-one-out 기반 교차 검증을 통해 119개의 테스트 지점에서의 ccdf를 모델링한 후에, 그 지점에서의 참값이 4 ppm을 초과할 경우 오염 지역으로 구분하고 확률 임계치의 변화에 따라 오염/비오염 지역을 잘못 구분하는

분류 오류가 최소화되는 확률 임계치를 최종 확률 임계치로 설정하였다. 분류 오류는 오염 지역을 안전하다고 구분하는 false negative 오류와 비오염 지역을 오염 지역으로 구분하는 false positive 오류로 구분할 수 있는데, 이 연구에서는 이러한 2가지 오류의 합이 최소가 되는 확률 임계치 값을 최종적인 확률 임계치로 설정하였다. Fig. 6b는 확률 임계치의 변화에 따른 분류 오류의 변화 양상을 나타내고 있는데, 0.577에서 최소 분류 오차를 나타내어 이 값을 확률 임계치로 적용하여 Fig. 6c의 오염 분류도를 작성하였다. 분류 결과, 연구 지역의 25.59%가 오염지역으로 구분되었다. 이 분류 방법론의 정확도를 평가하기 위해 검증 자료를 바탕으로 분류 정확도를 계산한 결과, false false positive 오류는 발생하지 않았으며, false negative 오류만 79개 검증 지점중에서 6개 지점에서 나타나 분류 오류는 7.59%로 계산되었다.

Fig. 7. (a) Classification result on the basis that the e-type estimates exceeds the threshold value of 4 ppm, (b) risk α , (c) risk β .

두 번째 분류 방법론으로 Fig. 5a의 ccdf의 평균값이 임계값 4 ppm을 초과할 경우 오염지역으로 구분하였다(Fig. 7a). 분류 결과, 연구 지역의 36.58%가 오염지역으로 구분되어 첫 번째 방법론에 비해 비율이 높게 나타났다. 이 방법론을 적용할 경우 카드뮴 분포도를 우선적으로 작성해야 하므로, 이 단계에서 불확실성이 따르고 분류 과정에 오류가 발생할 수 있다. 이러한 분류 과정의 오류로 2가지가 존재하는데, 첫 번째는 비오염 지역을 오염 지역으로 잘못 구분하는 false positive 오류를 범할 제 1종 오류 혹은 α 오류가 있으며, 두 번째 오류로는 오염 지역을 비오염 지역으로 잘못 구분하는 false negative 오류를 범할 제 2종 오류 혹은 β 오류가 있다. 이미 ccdf 모델링을 통해 4 ppm을 초과할 확률 분포를 알고 있기 때문에 이러한 α 오류와 β 오류를 식 (8)에 의해 계산할 수 있다(Goovaerts, 1997, p. 348).

$$\begin{aligned} \text{risk } \alpha &= \text{Prob}\{Z(\mathbf{u}) \leq z_c | Z^*(\mathbf{u}) > z_c\} = F(\mathbf{u}; z_c | (n)) \\ \text{risk } \beta &= \text{Prob}\{Z(\mathbf{u}) > z_c | Z^*(\mathbf{u}) \leq z_c\} = 1 - F(\mathbf{u}; z_c | (n)) \end{aligned} \quad (8)$$

여기서 z_c 는 고려하는 임계치값을 나타내며, 이 연구에서는 4 ppm에 해당한다.

α 오류(Fig. 7b)와 β 오류(Fig. 7c) 분포도를 살펴보면, α 오류를 가지는 지역과 β 오류를 가지는 지역은 상호 배타적이기 때문에 α 오류가 존재하는 지역에서는 β 오류가 존재하지 않으며, Fig. 7a의 분류결과와 경계 부근에서 오류값이 상대적으로 높게 나타나고 있다.

이 분류 방법론의 정확도를 평가하기 위해 검증 자료를 바탕으로 분류 정확도를 계산하였다. 계산 결과, false negative 오류와 false positive 오류는 4개 지점과 2개 지점에서 발생하여 각각 5.06%와 2.53%로 총 분류 오류는 7.59%로 나타났다. 이러한 분류

Fig. 8. (a) Coefficient of variation, (b) locations of 50 secondary sampling candidates based on the large coefficient of variation.

오류가 첫 번째 분류 방법론에 의한 총 분류 오류와 동일하게 나타나서, 특정 분류 방법론이 우수하다는 결론을 내릴 수는 없었다. 확률 임계치와 속성 임계치에 의한 분류 결과 오분류가 일어난 지점은 각각 6개이었는데, 공통적으로 4개 지점에서 false negative 오류가 발생하였으며, 2개 지점씩에서 서로 다른 오류가 발생하였다. 이러한 8개 지점에 대해서 α 오류와 β 오류를 살펴보니 false positive가 발생한 2개 지점에서는 α 오류값이 0.51과 0.62로 상대적으로 높게 나타났으며, false negative가 발생한 6개 지점에서의 β 오류값은 최소 0.20에서 최대 0.56까지의 범위에서 평균값이 0.41로 나타났다. 즉 이렇게 분류 오류가 발생하는 지역에서는 분류 위험도가 상대적으로 높게 나타나서 α 오류와 β 오류가 분류 결과의 해석과 관련된 유용한 정보를 제공함을 확인할 수 있었다.

추가 샘플 지역 추출 결과

지금까지의 분석을 통해 얻어진 오염 분포도로부터 오염/비오염 지역의 구분과 같은 최종 의사 결정을 내리기 이전에 추가 샘플링과 같은 부가 정보의 획득이 필요한 경우가 있다. 지시자 접근에 의한 확률론적 분석 결과는 이러한 추가 샘플 지역 추출에 유용하게 사용될 수 있다.

추가 샘플 지역 추출과 관련하여 전술한 두 번째 분류 방법론과 같이 공간 분포도의 정확도에 오염/비오염 지역의 구분이 좌우될 수 있기 때문에 공간 분포도 작성 자체의 불확실성이 높은 지역을 우선적으로 고려할 수 있다. 이 때 불확실성이 높은 지역의 추출에는 ccd의 조건부 분산이나, 오염/비오염 지역

의 잘못된 구분에 수반되는 위험도를 고려할 수 있다. 그러나 지화학 원소 분석은 예측 정확도의 향상보다는 실제 오염 지역의 구분에 수반되는 정화작업과 같은 의사 결정을 최종 목표로 간주할 수 있기 때문에 특정 임계치 값을 초과하는 정도에 대한 불확실성을 최소화할 필요가 있다. 이러한 배경에서 Van Meirvenne and Goovaerts(2001)은 다음 식 (9)와 같이 조건부 분산과 임계치 값과의 차이를 고려한 변동 계수(coefficient of variation: CV)를 제시하였다.

$$CV = \frac{\sqrt{\sigma^2(\mathbf{u})}}{|z^*(\mathbf{u}) - z_c|} \quad (9)$$

식 (9)를 살펴보면, 조건부 분산이 커질수록 즉 값 추정에 수반되는 불확실성이 커질수록 변동 계수는 커지며, 또한 임계치와 추정값 사이의 차이가 작을수록 크게 나타난다.

연구지역에서 이미 198개의 샘플을 구할 수 있으므로, 앞선 잭나이프 기반의 분석과 달리 추가 샘플링 지역의 선정에는 198개를 모두 이용한 변동 계수를 계산하였다. 즉 식 (9)의 조건부 분산과 추정값의 계산에 198개의 모든 샘플을 이용하였다. Fig. 8a는 임계치 4 ppm을 기준으로 계산한 변동 계수의 분포도를 나타내고 있다. 변동 계수의 값이 높은 지역은 Fig. 6c와 Fig. 7a의 오염 지역과 비오염 지역의 경계 부분에 집중적으로 존재한다. 즉 임계치 4 ppm 근방의 값들로 추정된 지역에서 높은 변동 계수 값을 나타내고 있다.

이러한 변동 계수 분포도를 기반으로 변동 계수값

이 높은 지역을 대상으로 추가 샘플 지점을 추출하였다. Van Meirvenne and Goovaerts(2001)에 의해 제안된 변동 계수값을 이용한 추가 샘플 지점 추출 방식을 보완하여 기존 샘플링 위치와의 근접성과 제한된 변동 계수값을 고려한 방법을 제안하여 시범적으로 적용하였다.

우선 변동 계수값이 가장 높은 지점부터 추출을 시작하는데, 제한조건으로 다음 사항들을 고려하였다. 기존 198개의 샘플링 자료와의 근접성을 고려하여 기존 샘플링 자료 사이의 최소 거리인 200미터 이내의 지점은 배제하였다. 추가 샘플 지점의 개수는 예시를 목적으로 실험적으로 50개로 한정하였으며 또한 이러한 지점 사이의 거리는 원자료의 베리오그램 상 상관거리를 고려하여 최소 4 km 이상 떨어지도록 하였다. 즉 가장 변동 계수값이 높은 지점이 기존 샘플링 자료와 200미터 이상 떨어져 있다면 추가 샘플링 지점으로 선정을 하고, 이 지점으로부터 4 km 이내의 변동 계수값이 높은 지점들은 제외하고, 그 다음 변동 계수값이 높은 지역을 탐색하는 과정을 반복하였다. 그러나 4 km로 제한할 경우 22개 지점만 추출이 가능하여 이 제한 조건을 완화하여 거리를 줄여나가면서 50개 지점이 추출되도록 하였다. 3 km 정도로 제한할 경우 50개 지점이 추출되었지만, 상당히 작은 변동 계수를 가지는 지점도 추출되는 문제가 발생하여, 3 km 미만으로 다시 완화해나가면서 새로운 제한조건으로 연구지역의 전체 변동 계수값의 분포를 고려하여 75백분위수인 1.98 이상이 되는 지점을 추출하도록 하였다. 이러한 새롭게 추가된 제한 조건하에서는 1.5 km 정도에서 모든 조건을 만족하는 샘플 지점의 추출이 가능하였다. 이렇게 추출된 추가 샘플링 지역은 Fig. 8b에 제시되어 있다. Fig. 8b의 샘플링 지점에서의 평균 추정값은 대부분 임계값인 4 ppm 근방의 값으로 추정되어 변동 계수 값이 아주 높은 지점인 것으로 나타났다. 즉 이 연구에서 변동 계수는 식 (9)에서 예측 자체에 수반되는 불확실성보다는 임계치와 유사한 값으로 추정된 지점에서 높게 나타났다. 추출된 지점들의 토지 이용 현황은 거의 대부분 산림 지역으로 나타났는데, 만약 추가 조사 위치의 환경과 오염 피해의 영향이 토지 이용별로 상이한 점을 고려한다면 토지 이용도상의 도심지와 농경지 등으로 한정시켜 제안 방법론을 추가적으로 적용이 가능하다.

결론

지화학 자료와 같은 환경 변수들은 공간적인 분포 자체도 중요하지만 이러한 분포값의 분석을 통한 환경영향 평가나 각종 의사결정 행위와 직접적으로 연관된다. 따라서 공간적인 분포값 추정과 더불어 이에 수반되는 불확실성이 이후 분석 과정에 직접적으로 영향을 미치기 때문에 불확실성의 정량적 추정과 이와 관련된 부가 정보 추출이 중요하다.

이 논문에서는 지화학 자료를 대상으로 지시자 지구통계학이 공간 추정과 더불어 확률론적 불확실성 추정 및 오염도 분석에 유용하게 사용될 수 있음을 사례 연구를 통해 예시하고자 하였다. Ccdf의 모델링 과정에서 비모수적인 방법론인 지시자 크리깅과 모수적 방법론인 다중 가우시안 크리깅을 함께 적용하였는데, 예측 능력에서 지시자 크리깅이 우수한 것으로 나타나서 지화학 자료 분석에 지시자 크리깅이 유용함을 다시 확인할 수 있었다. 일단 연구지역의 모든 위치에서 ccdf가 모델링되면, 값 추정 및 불확실성의 정량적 추정이 가능해지는데, 이 연구에서는 이 후 과정으로 오염 지역과 비오염 지역을 구분하는 분류 과정을 확률론적으로 예시하였다. 사례 연구에서는 잭나이프 기반 검증 결과, 확률 임계치 방법과 물리적 속성값 임계치 방법에 의한 분류 분류 정확도가 동일하게 나타났으며, 오류가 크게 나타나지 않았다. 그러나 속성 임계치 방법의 경우 α 오류와 β 오류를 해석 보조 정보를 제공할 수 있는 장점이 있다. 또한 사후 의사결정을 내리기 이전에 추가적인 현장 조사를 위한 샘플링 지점 추출에 불확실성 척도가 유용하게 사용될 수 있음을 확인하였다. 그러나 환경오염 지시자로 활용가능한 지화학 자료의 분석의 경우, 지질, 수계 경계 등의 영향을 받기 때문에 분포도 작성 시 이러한 영향을 고려한 현실적인 분포도 작성이 필요하다. 이 경우, 이러한 정보들을 부가자료로 사용하는 다변량 지시자 크리깅의 적용이 가능하기 때문에 이 방법론의 적용을 통한 지질학적 해석이 병행되어야 할 것으로 판단된다.

앞으로의 연구로 현재 고려하지 않은 경제적인 비용 효과 모델을 고려한 위험도 분석 연구와 ccdf로부터 속성값을 추정하는 단계에서 손실 함수의 개념을 도입한 최적의 값 추정 연구도 병행되어야 할 것으로 판단된다. 또한 이 연구에서는 국소적 불확실성에

기반하여 오염/비오염 지역을 구분하는 분석을 수행하였으나, 실제 의사결정은 대상 지역의 공간해상도보다 큰 불규칙한 모양의 구역 단위로 이루어지는 경우가 많다. 크리깅과 더불어 지구통계학의 또다른 주요 방법론인 추계론적 시뮬레이션은 전술한 향후 과제에 효율적으로 적용 가능할 것으로 판단되며, 적용 연구를 수행할 계획이다.

감사의 글

이 연구는 한국지질자원연구원 기관목적사업 “GIS 기반 국토지질정보시스템 실용화 기술 개발”의 위탁 과제 연구비 지원으로 수행되었습니다. 논문의 표현과 질을 높이는데 도움을 주신 익명의 심사위원께 감사드립니다.

참고문헌

- 박노옥, 2009, 현장 조사 자료를 이용한 GIS기반 주제도 작성을 위한 단변량 크리깅 기법의 비교. 대한원격탐사학회지, 25, 321-338.
- 오강호, 김주용, 고영구, 윤석태, 신상은, 박배영, 문병찬, 김해경, 2003, 광주광역시 하천의 표층퇴적물에 대한 지구화학적 특성과 오염. 한국지구과학회지, 24, 346-360.
- 오석훈, 서백수, 2007, 탄성과 속도와 전기비저항 자료의 지구통계학적 복합해석에 의한 암반등급의 확률적 평가. 물리탐사, 10, 293-298.
- 오인석, 고경석, 공인철, 구민호, 2008, 금산 매립장 주변 대수층의 수리지화학적 특성 및 오염 확산 평가. 자원환경지질, 41, 657-672.
- 이진수, 서효준, 황인호, 1998, 지화학조사연구 (1:250,000 강릉도폭 광역 지화화도). 한국자원연구소, 147 p.
- 위수민, 김은호, 2009, 여수 지역에 분포하는 백악기 화강암류에 대한 지화학적 연구. 한국지구과학회지, 30, 267-281.
- 조규성, 노열, 정덕호, 2007, 당진화력발전소의 석탄회 연안 매립과 중금속 원소의 용출에 대한 생지화학적 연구. 한국지구과학회지, 28, 112-122.
- 진명식, 2007, 국내 지화학탐사 발전사. 자원환경지질, 40, 691-698.
- 황상기, 이평구, 2005, 지화학자료를 이용한 금·은 광산의 배태 예상지역 추정. 베이시안 지구통계학과 의사나무 결정기법의 활용. 자원환경지질, 38, 663-673.
- Barabas, N., Goovaerts, P., and Adriaens, P., 2001, Geostatistical assessment and validation of uncertainty for three-dimensional dioxin data from sediments in an estuarine river. Environmental Science & Technology, 35, 3294-3301.
- Deutsch, C.V. and Journel, A.G., 1998, GSLIB: Geostatistical Software Library and User's Guide. Oxford University Press, NY, USA, 369 p.
- Goovaerts, P., 1997, Geostatistics for Natural Resources Evaluation. Oxford University Press, NY, USA, 483 p.
- Goovaerts, P., Webster, R., and Dubois, J.-P., 1997, Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics. Environmental and Ecological Statistics, 4, 31-48.
- Goovaerts, P., Trinh, H.T., Demond, A., Franzblau, A., Garabrant, D., Gillespie, B., Lepkowski, J., and Adriaens, P., 2008, Geostatistical modeling of the spatial distribution of soil dioxins in the vicinity of an incinerator. 1. Theory and application to Midland, Michigan. Environmental Science & Technology, 42, 3648-3654.
- Goovaerts, P., Avruskin, G., Meliker, J., Slotnick, M., Jacques, G.M., and Nriagu, J., 2005, Geostatistical modeling of the spatial variability of arsenic in groundwater of Southeast Michigan. Water Resources Research, 41, W07013, doi:10.1029/2004WR003705.
- Journel, A.G., 1983, Non-parametric estimation of spatial distributions. Mathematical Geology, 15, 445-468.
- Juang, K.-W. and Lee, D.-Y., 1998, Simple indicator kriging for estimating the probability of incorrectly delineating hazardous areas in a contaminated site. Environmental Science & Technology, 32, 2487-2493.
- Ko, K.-S., Lee, J.-S., Kim, J.-G., and Lee, J., 2009, Assessments of natural and anthropogenic controls on the spatial distribution of stream water quality in Southeastern Korea. Geosciences Journal, 13, 191-200.
- Saisana, M., Dubois, G., Chaloulakou, A., and Spyrellis, N., 2004, Classification criteria and probability risk maps: Limitations and perspectives. Environmental Science & Technology, 38, 1275-1281.
- Saito, H. and Goovaerts, P., 2003, Selective remediation of contaminated sites using a two-level multiphase strategy and geostatistics. Environmental Science & Technology, 37, 1912-1918.
- Van Meirvenne, M. and Goovaerts, P., 2001, Evaluating the probability of exceeding a site-specific soil cadmium contamination threshold. Geoderma, 102, 75-100.

2010년 7월 6일 접수
2010년 8월 4일 수정원고 접수
2010년 8월 10일 채택