情報地質 第26巻 第2号 052-053頁 2015年
Geoinformatics, vol.26, no.2, pp.052-053, 2015

取得地球情報の評価：エントロピーと分散の違い

正路　徹也

Evaluation of Obtained Geoinformation: Comparison of Entropy and Variance

Tetsuya Shoji*

* 東京大学　The University of Tokyo, Tokyo 113-8639, Japan.
E-mail: t-t_shoji@jcom.home.ne.jp

キーワード：地球情報，調査，評価，エントロピー，分散
Key words: Geoinformation, Survey, Evaluation, Entropy, Variance

1．はじめに

地球科学的調査は，調査そのものが目的ではなく，その過程で得られた情報にもとづいて適切な行動を選択・採用するために行なわれる．例えば，原子力発電所(原発)周辺で最近行われた活断層の調査は，再稼働後の原発の安全性の評価が目的である．この課題で，筆者には理解できない点が 1 つある．通常，安全性が完全には評価できないことに起因するリスクは損害保険で補償する．このとき，保険料はリスクの発生確率に基づいて算定され，その算定は調査の精度を関数に行われる．したがって，保険料率の算定には調査の精度に関するデータが必要である．原発に限らず何らなの操業に対する保険の掛金はコストに算入される．一方，調査精度を向上させるためにはより多くのコストが必要である．したがって，最適な調査は両コストをバランスさせることによって実現される．ところが，筆者が知る限り，原発の安全性に関する技術的・経済的議論では，最も肝心な安全性，保険，調査の間が結びつけていない．その理由の 1 つとして，調査精度が，保険料率の算定に役立つ定量的数値として示されていないことが考えられる．

地球科学的調査の過程で取得された地球情報は，次の式(1)で定義されるエントロピーs あるいは式(2)で定義される分散 V を使って評価できる(正路, 2003 ; Shoji, 2006 ; 正路, 2014).

$$s_{ik/K} = -\sum_{j=1}^{m} p_{jik/K} \ln p_{jik/K} \tag{1}$$

$$S_{k/K} = \sum_{i=1}^{N_{k/K}} s_{ik/K}$$

$$\bar{s}_{k/K} = \sum_{i=1}^{N_{k/K}} s_{ik/K} \Big/ N_{k/K}$$

$$V_{ik/K} = \sum_{j1=1}^{m-1} p_{j1ik/K} \sum_{j2=j1+1}^{m} p_{j2ik/K} \tag{2}$$

$$\bar{V}_{k/K} = \sum_{i=1}^{N_{k/K}} V_{ik/K} \Big/ N_{k/K}$$

ここで，添字中の k/K は調査が段階 K まで進んだときの途中の段階 k (k = 1,･･･,K)を表し，$p_{jik/K}$ はセル ik/K (総数 $N_{k/K}$) における地質体 j (総種類数 m)の存在度である．また，$S_{k/K}$ と $\bar{s}_{k/K}$ はそれぞれ地域全体のエントロピーと平均セルエントロピー，$\bar{V}_{k/K}$ は地域全体の平均分散である．

式(1)のエントロピーと式(2)の分散は，添字のik/Kには関係なく常にそれぞれ$\partial^2 s/\partial p_j^2 > 0$と$\partial^2 V/\partial p_j^2 > 0$ (j = 1, 2,･･･,m) が成り立つので，その形状はこれらの軸の正の方向に対して凸である．この性質により，エントロピーあるいは分散による情報の評価は，ほぼ同じである．しかし，厳密に考えると違う．その違いを以下で検討する．

2．エントロピーと分散の関係

式(1)のエントロピーsと式(2)の分散Vは，地質体の存在度pで結びつけられている．この両者の関係を第1図に示す．

第1図で，最も太い線は地質体の種類数mが2の場合を示す．この場合，地質体1の存在度をp_1で表すと地質体2の存在度p_2は，$p_2 = 1-p_1$である．したがって，独立変数の数が1すなわち自由度が1なので，sとVは一対一対応する．図では $0 \leqq p_1 \leqq 1/2$の範囲が示してある($p_2 = 1$となる点をA1，$p_2 = 1$となる点をA2としたとき，sもVも線分A1A2の重心≡中点G12に対して対象である)．これに対して，$m = 3$の場合は，独立変数はp_1とp_2で$p_3 = 1-(p_1+p_2)$であるため自由度が2となる．そこで，第1図では，三角形A1A2A3 (A3は$p_3 = 1$となる点)の重心を点G123として，p_1，p_2，p_3の関係を点G123と点A1および点G123と点G12を結ぶ2本の線分に限ったときのsとVの関係を2番目に太い線で表している．さらに，第1図には，$m = 4$，$m = 5$，$m = 6$の場合をそれぞれ3番目，4番目，5番目に太い線で示してある．図示された各線において，p_1，p_2，…，p_mは$m = 3$の場合を拡張した関係にある．例えば，$m = 6$の場合，5次元の単体A1A2A3A4A5A6の重心G123456と①点A1，②点G12，③点G123，④3次元の単体A1A2A3A4(四面体)の重心G1234，⑤4次元の単体A1A2A3A4A5の重心G12345を結ぶ関係にある場合が示してある．

第1図で明らかなように，地質体の種類数mの増加とともに，エントロピーは順調に増加するのに対し，分散はほとんど増加しない(図の横軸の最大値は2であるのに対し，縦軸のそれは0.5である)．これは，次の式(3)と(4)で示すように，mの増加に伴ってエントロピーの取り得る最大値$s_{m\mathrm{MAX}}$は発散するのに対し，分散の取り得る最大値$V_{m\mathrm{MAX}}$は1/2に収束するためである．

$$s_{m\mathrm{MAX}} = \ln m \tag{3}$$

$$V_{m\mathrm{MAX}} = \frac{1}{2}\left(1-\frac{1}{m}\right) \tag{4}$$

なお，これらの最大値は，$p_1 = p_2 = \cdots = p_m = 1/\mathrm{m}$ のとき得

られる．

3．エントロピーと情報量

日本語の情報学の教科書では，「情報エントロピー」と「情報量」がほぼ同義語として使われている．しかし，以下に述べるように両者は明確に使い分けた方がよい．

Wikipedia (2015/1/25)では，”entropy (information theory)”を”entropy is a measure of *unpredictability* of *information content*”(斜体は原文，下線は引用者)と説明している．この項のLanguagesで日本語に移ると「情報量」が出てきて，「情報量(じょうほうりょう，エントロピーとも)は，情報理論の概念で，あるできごと(事象)が起きた際、それがどれほど起こりにくいかを表す尺度である」(下線部引用者)と説明されている．筆者にとって「情報量」という語は，取得した情報に含まれる何らかの量を表していると思える．これに対し，エントロピーは”unpredictability (予測不可能性 ≒ 不確かさ)”の尺度なので，取得しようしている情報に含まれる何らかの量である．

上記の情報エントロピーの定義を，正路(2014)の調査進行のモデルに適用すると，調査段階(*k*−1)と*k*におけるセルエントロピーをそれぞれ $\bar{s}_{(k-1)/K}$ と $\bar{s}_{k/K}$ と書いた場合，それらの差 $\Delta s_{(k-1)<k}=\bar{s}_{(k-1)/K}-\bar{s}_{k/K}$ が段階*k*の調査で除去したエントロピーあるいは取得した情報量と定義できる．調査が始まる前の段階0におけるエントロピーは，取り得る値の最大値 $s_{m\mathrm{MAX}}$ と定義すると，$\Delta s_{k/K}=s_{m\mathrm{MAX}}-\bar{s}_{k/K}$ を段階*k*までに除去されたエントロピーあるいは取得した情報量と定義できる．

ところで，エントロピーは熱力学で導入された概念で，その統計力学的説明が式(1)である．地球では太陽からの電磁波エネルギーや地球内部からの熱エネルギーにより生物活動を含む各種の変化が起き，この変化に伴ってエントロピーが発生する．熱力学によると，この変化が持続するためには，地球は発生したエントロピーを外界(＝宇宙空間)に放出しなければならない．今，符号を取得は正，放出は負と定義すると，当然のことながら，地球が放出するエントロピーと宇宙空間が受け取るエントロピーは符号が逆で，絶対値は等しい．これと同じ関係が，上述した情報エントロピーと取得情報量の間でも成り立つ． したがって，次の式(3)が成り立つ．

$$\Delta s_{k/K}=\Delta s_{0<1}+\Delta s_{1<2}+\cdots+\Delta s_{(K-1)<K} \qquad (5)$$

4．分散の逆数と情報の精度

いま変量*x*に関して*n*個の推定値 $\bar{x}_1$， $\bar{x}_2$，…， $\bar{x}_n$ とそれらの分散それぞれ $V(\bar{x}_1)$，$V(\bar{x}_2)$，…，$V(\bar{x}_n)$ があるとき，これらからより良い推定値 $\bar{\bar{x}}$ は，次式のように重み付き平均として求めることができる．

$$\bar{\bar{x}}=\sum_{i=1}^{n}a_i\bar{x}_i \qquad \left(\sum_{i=1}^{n}a_i=1\right)$$

分散の加法性により，$\bar{\bar{x}}$ の分散 $V(\bar{\bar{x}})$ は，

$$V(\bar{\bar{x}})=V\left(\sum_{i=1}^{n}a_i\bar{x}_i\right)=\sum_{i=1}^{n}a_i^2V(\bar{x}_i)$$

である．この $V(\bar{\bar{x}})$ はλを Lagrange の未定乗数とすると，

$$a_1V(\bar{x}_1)=a_2V(\bar{x}_2)=\cdots=a_nV(\bar{x}_n)=\lambda/2$$

のとき最小になる(例えば，正路，1972)．したがって，より良い推定値を求めるには，各推定値の分散の逆数を重みとすればよい．

上述の数学は，分散は取得された情報に含まれるあいまいさを表し，その逆数は取得された情報の精度に対応することを示している．これを正路(2014)の調査進展のモデルに当てはめると，調査段階(*k*−1)と*k*における分散をそれぞれ $\bar{V}_{(k-1)/K}$ と $\bar{V}_{k/K}$ と書いた場合，それらの比 $r_{(k-1)<k}=\bar{V}_{k/K}\big/\bar{V}_{(k-1)/K}$ は段階*k*の調査におけるあいまいさの減少率を，その逆数の$1/r_{(k-1)<k}$は精度の向上率を表す．また，$r_{k/K}=\bar{V}_{k/K}\big/V_{m\mathrm{MAX}}$ と$1/r_{k/K}$はそれぞれ段階*k*までにおけるあいまいさの減少率と精度の向上率を表している．このように定義すると，分散に対しては次の式(6)が成り立つ．

$$r_{k/K}=r_{0<1}\cdot r_{1<2}\cdots r_{(K-1)<K} \qquad (6)$$

5．まとめ

地球科学的調査の過程で取得された情報にもとづく各地質体の存在度から求められるエントロピーあるいは分散は，ともに正の値で，調査の進行に伴う絶対値の減少は不確実性の減少に対応している．両者を比較すると，それらの最大値は，前者が地質体の種類の増加に対し敏感で無限大に発散するのに対し，後者は鈍感で1/2に収束する．不確実性の逆，すなわち確実性あるいは精度を示す因子として，前者は取得情報量(＝最大エントロピーからの差)，後者は分散の逆数(＝最大分散に対する比)が定義できる．

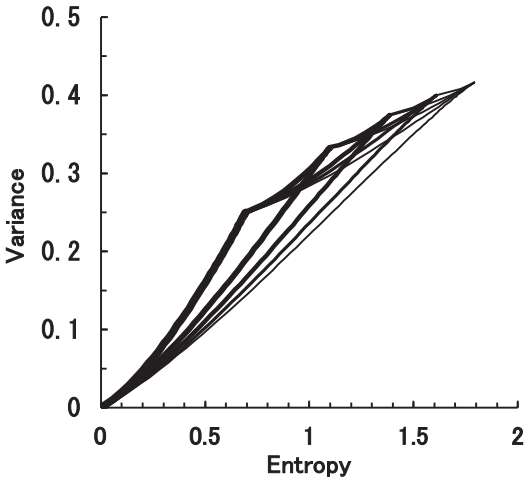
引用文献

正路徹也 (1972) X線粉末法によるアルカリチョウ石の組成および構造(Al/Si 秩序無秩序)の決定．鉱物学雑誌, vol. 10, no. 6, pp.413-425.

正路徹也 (2003) 地球科学情報の評価. 情報地質, vol. 14, no. 4, pp.285-299.

Shoji, T. (2006) Optimal systems of geoscience surveying – a preliminary discussion. Computers & Geosci., vol. 32, no. 1, pp.1128-1138.

正路徹也 (2014) 探査の進行に伴う地球情報エントロピーの変化．情報地質，vol. 25, no. 2, pp.68-71.



第 1 図．エントロピーsと分散 V の関係．各線は，m 成分からなる系で，(m−1)次元の単体の頂点と重心を結ぶ線に沿って存在度が変化した場合に対応し，太い方から細い方へ m が 2，3，4，5，6 の場合．