第30回日本情報地質学会講演会 講演要旨集 001-002頁 2019年

調査の進行に伴う分散の増加

正路　徹也

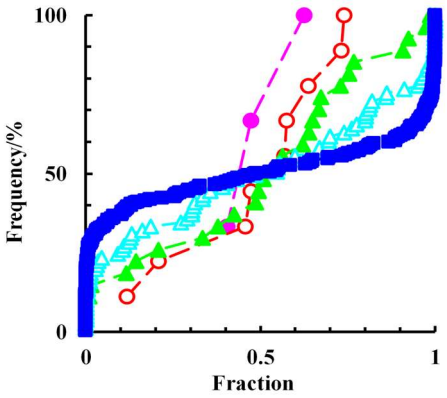
Increase of Variance with Progressing Survey

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：地球科学，調査，情報，評価，分散
Key words: Geoscience, Survey, Information, Evaluation, Variance

地球科学的調査の進行に伴い対象地域を構成する地質体の分布およびそれらの接触状態に関する地球情報が増加する．したがって，対象地域を構成する各セルにおける各地質体の存在度の推定値の精度が向上して，エントロピーあるいは分散の値が小さくなる．この調査の進行に伴うエントロピーあるいは分散が減少する様子を，正路(2018)は次の単純な数値モデルで示した．調査は直線上で行われ，段階 1 では全範囲の中点の地質体を同定する．段階 2 では，全体を 3 つのセルに分け，各セルの中点の地質体を同定する(このうち中間のセルの中点は，段階 1 で調査済み)．段階 3 では各セルをさらに 3 つのセルに分け，各セルの中点の地質体を同定する．以下，この操作を繰り返し，段階 6 でセルの大きさが目的の分解能に達し調査は終了する．条件を変えたいくつかのシミュレイションのうち，地質体の種類数が 2 (他は 4)で，存在割合が準二値(他は完全二値)で与えられ，存在割合の比が 50:50 (他は 20:80，10:90，5:95．2:98)の場合について，調査段階をパラメーターとして存在割合の累積頻度を第 1 図に示す(5 本の点列のうち，構成点数が少ない方から段階 2，3，4，5，6)．一見して明らかなように，調査の進行とともに分散が増加している．実際，分散の値は，この順序で 0.013，0.047，0.096，0.149，0.196 と増加している．これは正路(2018)の結論と逆である．その理由は，以下に詳述するように分散の定義の違いによる．



第 1 図．地質体の種類数が 2 で，存在割合が準二値で与えられ，存在割合の比が 50:50 の場合における存在割合の累積頻度．各点列は調査段階に対応し，構成点数が少ない方から段階 2，3，4，5，6．分散の値は，この順序で 0.013，0.047，0.096，0.149，0.196.

正路(2018)では，1 つのセルi (総数N)に注目し，そこにおける地質体j (種類数m)の存在割合がp_{ij}の場合，当該セルの分散$V_{i\mathrm{Indv}}$ は$(m-1)$次元空間の正単体 $\mathrm{A_1A_2\cdots A}_m$ の頂点A_j に重みp_{ij}がありp_{ij} $(j=1,\cdots,m)$で与えられる単体内の 1 点 P_iの周りの 2 次モーメントとして次式で定義する．

$$V_{i\mathrm{Indv}}=\sum_{j=1}^{m}p_{ij}\left\{\sum_{j1=1}^{m}{p_{ij_1}}^2+\sum_{j1=1}^{m-1}p_{ij_1}\sum_{j2=j1+1}^{m}p_{ij_2}-p_{ji}\sum_{j1=1}^{m}p_{ij_1}\right\}$$
$$=\sum_{j1=1}^{m-1}p_{ij_1}\sum_{j2=j1+1}^{m}p_{ij_2} \qquad (1)$$

ここで，添え字の Indv＝Individual は個別のセルに注目していることを意味し，{}内は点P_iと頂点A_jの距離の二乗である．これより全体の分散V_{Indv}は次式で与えられる．

$$V_{\mathrm{Indv}}=\sum_{i=1}^{N}V_{i\mathrm{Indv}}\Big/N$$

なお，式(1)で与えられる分散は，セルiのデータのみを使って計算されているので，以後個別分散(individual variance)と呼ぶ．

二元系$(m=2)$の場合，地質体 1 と 2 の存在割合をそれぞれp_{1i}とp_{2i} $(p_{1i}+p_{2i}=1)$とすると，点P_iと頂点A_1との距離はp_{i2}，頂点A_2との距離はp_{i1}であるから，式(1)は次式(2)のように表される．

$$V_{i\mathrm{Indv}}=p_{1i}{p_{2i}}^2+p_{2i}{p_{1i}}^2=p_{1i}p_{2i}(p_{2i}+p_{1i})$$
$$=p_{1i}p_{2i}=p_{i1}(1-p_{i1}) \qquad (2)$$

第 1 図に示されている累積頻度の分散V_{Bulk} (添え字 Bulk の意味は後述)は，セルiの存在割合をp_i (二元系なので地質体の種類を示す添え字は省略)として，次式(3)で与えられる．

$$V_{\mathrm{Bulk}}=\sum_{i=1}^{N}(p_i-\bar{p})^2\Big/N$$
$$=\sum_{i=1}^{N}\left(p_i-\sum_{i=1}^{N}p_i\Big/N\right)^2\Big/N \qquad (3)$$

ここで，$\bar{p}$は全データ(bulk data)の平均である．式(2)に相当するセルiのバルク分散$V_{i\mathrm{Bulk}}$は次式(4)で与えられる．

$$V_{i\mathrm{Bulk}} = (p_i - \bar{p})^2 \qquad (4)$$

式(4)には，$\bar{p}$というセルi以外のセルの値を必要とする全体の平均が入っている．すなわち，1 つのセルのバルク分散を求めるとき，個別分散とは違い，当該セルのみの値では不十分で，他のセルの値も必要とする．これがバルク(bulk)の意味である．

第 2 図の左右にそれぞれ調査の進行に伴う個別分散とバルク分散の変化を示す．両図を比較すると，1) 個別分散は段階 0 に値があるが，バルク分散にはない，2) バルク分散は個別分散の上下を反転したパターンを示す．段階 0 における個別分散が定義できるのは，対象地域全体の地質体の存在割合が最終段階における存在割合の平均に等しいと推定し，その値に相当する重みが地質体を表す両端に掛かっているとするためである．バルク分散は個別分散の上下を反転したパターンについては後述する．

第 1 図の累積頻度図を微分すると，微分型頻度図が得られる．上記の分散の説明はこの微分型頻度図に基づいている．三元系では，累積する方向によって得られる累積頻度図が変る．しかし，微分型頻度図を考えれば，分散の計算は可能である．例えば，m元系の場合，セルiの地質体の存在割合を$(m-1)$次元の正単体内の 1 点P_iとして表し，その点と平均の存在割合を表す点$\mathrm{P_M}$との距離$\overline{\mathrm{P}_i\mathrm{P_M}}$の二乗を${l_i}^2$とすれば，それが次式のようにバルク分散となる．

$$V_{i\mathrm{Bulk}} = {l_i}^2$$

三元系$(m=3)$で，高さが 1 の正三角形 $\mathrm{A_1A_2A_3}$ 内の点P_iの位置を辺 $\mathrm{A_2A_3}$，$\mathrm{A_1A_3}$，$\mathrm{A_1A_2}$ に下した垂線の長さそれぞれp_{1i}, p_{2i}, p_{3i} $(p_{1i}+p_{2i}+p_{3i}=1)$で，点$\mathrm{P_M}$の位置も同様に$\overline{p_1}$, $\overline{p_2}$，$\overline{p_3}$で表すと，長さ$\overline{\mathrm{P}_i\mathrm{P_M}}$の二乗${l_i}^2$は次式(5)で与えられ，これがセル$i$のバルク分散となる．

$$V_{i\mathrm{Bulk}} = {l_i}^2 = {\Delta p_1}^2 + \Delta p_1 \Delta p_2 + {\Delta p_2}^2 \qquad (5)$$

ここで，$\Delta p_1 = p_{i1} - \overline{p_1}$，$\Delta p_2 = p_{i2} - \overline{p_2}$で，

$$\overline{p_1} = \sum_{i=1}^{N} p_{i1} \Big/ N$$

$$\overline{p_2} = \sum_{i=1}^{N} p_{i2} \Big/ N$$

である．これより全体のバルク分散V_{Bulk}は次式で与えられる．

$$V_{\mathrm{Bulk}} = \sum_{i=1}^{N} V_{i\mathrm{Bulk}} \Big/ N$$

前述のよう，個別分散とバルク分散は上下を反転したパターンをなす(第 2 図)．したがって，両者は負相関を示している．さらに，二元系で両者の和は次式で示すように，$\left(\overline{p_1} - \overline{p_1}^2\right)$と一定値である．

$$\begin{aligned} V_{\mathrm{Indv}} + V_{\mathrm{Bulk}} &= \left\{ \overline{p_1} - \sum_{i=1}^{N} {p_{i1}}^2 \Big/ N \right\} \\ &\quad + \left\{ \sum_{i=1}^{N} {p_{i1}}^2 \Big/ N - \overline{p_1}^2 \right\} \\ &= \overline{p_1} - \overline{p_1}^2 \end{aligned}$$

すなわち，両者は完全負相関$(r=-1)$をなす．三元系においても，次式に示すように完全負相関をなす．

$$\begin{aligned} V_{\mathrm{Indv}} + V_{\mathrm{Bulk}} &= (\overline{p_1} + \overline{p_2}) \\ &\quad - \left(\overline{p_1}^2 + \overline{p_1 p_2} + \overline{p_2}^2\right) \end{aligned}$$

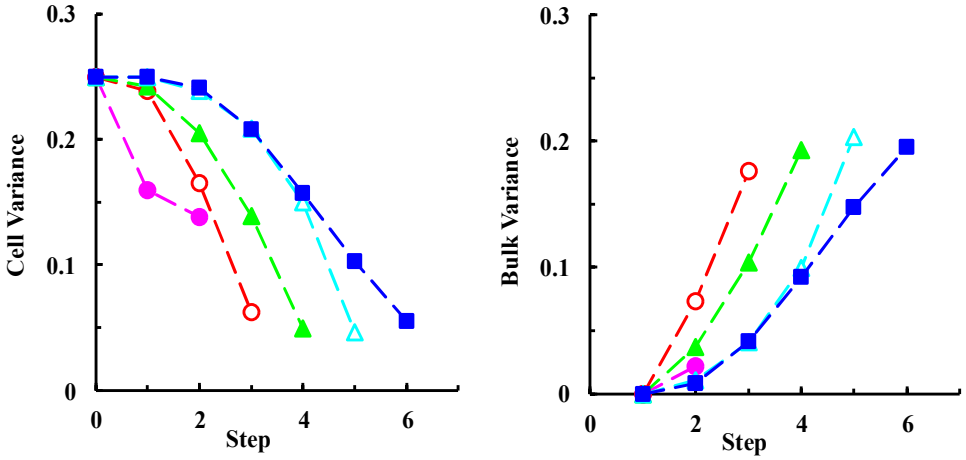
しかし，セルiの個別分散とバルク分散は次式のような関係にあり，負相関はなすが，完全負相関ではない．

$$V_{i\mathrm{Indv}} + V_{i\mathrm{Bulk}} = p_{i1} - 2p_{i1}\overline{p_1} + \overline{p_1}^2$$

以上より次のまとめが得られる．正路(2018)は，調査の進行に伴う分散の減少の様子を簡単な数値モデルで示した．しかし，通常の統計処理で用いられている分散を使うと，地質体の存在割合の分散は調査の進行に伴って増加する．前者の分散は個別セルの値のみで求められるので個別分散，後者の分散を求めるために全データの平均が必要なのでバルク分散と呼ぶと，両者は完全負相関の関係にある．

引用文献

正路徹也(2018)：地球科学的調査の進行に伴う地球情報エントロピーと分散の減少．情報地質，**29**(*2*)，61-75.



第 2 図．調査の進行に伴うセル分散(左)とバルク分散(右)の変化．各点列の右端が最終段階に対応する．