# Indikátory vo vzťahu k regionálnej premennej

***Ladislav Vizi***[1]

***The indicators related to a regionalized variable***

*In many areas, one has to deal not only with quantitative variables but also very often with categorical variables. Then a common problem consists of estimating the probability for each category to prevail at any particular location. In the seventies, in order to make prediction for future selective mining, geostatistics had to deal with the problem of estimating reserves above the cut-off grades. In 1982 at the 17th APCOM symposium in Colorado, Andre Journel presented the earliest concepts of "Indicator approach to estimation of spatial distribution". Using an indicator approach these probabilities can by established. Since then practitioners have used techniques based on kriging of indicators. The indicators above a threshold define the ore at this cut-off or, in a non-mining context, define the geometric set of values above the cut-off. In the last decade of the past century the interest in geometric problems has increased, in particular for simulation. For this reason, the development of techniques for simulating random function Z(x) from indicators I[Z(x)<z] at different thresholds z was started. The concept of indicator transforms is one of simplest and (possibly) most elegant in modern geostatistics. In many such cases several thresholds and indicators corresponding to different random sets related to the variable under study. These sets depend on each other and their mutual arrangement is an important structural characteristic of the variable. The indicator approach is as follows. Select discriminator cut-off value, which should not be confused with an economic cut-off or some critical level of toxicity. All samples with measurements above this value are coded as "1". All samples below selected cut-off value are coded as "0". This new measurement is the "indicator" – yes/no, presence/absence, etc. Each pair of samples that goes into the experimental semi-variogram will be {0,0} if both samples are bellow the cut-off value, giving a difference of zero or {1,1} if both samples are above the cut-off, also giving a difference of zero or finally {1,0} or {0,1} if one sample is above and the other is below. The calculated graph will be the average of these differences and represents the predictability of being above or below cut-off. Resulted kriged map of indicator values is interpreted as the probability that unknown value is above the cut-off value. The paper deals with theory of mathematical background and application of indicator approach to estimation of probability unknown values above selected cut-off.*

***Key words:*** *indicator, cut-off, geostatistics, variogram, kriging, probability*

## Úvod

V mnohých oblastiach štúdia regionálnych premenných sa je možné stretnúť nie len s kvantitatívnymi, ale veľmi často s kategoriálnymi premennými (Goovaerts, 1993). V takomto prípade vzniká problém odhadu pravdepodobnosti výskytu pre každú kategóriu v rámci premennej v študovanej oblasti. Použitím indikátorového prístupu, ktorý navrhol v roku 1982 Andre Journel, je možné odhadnúť takúto pravdepodobnosť. Indikátory pre danú prahovú hodnotu definujú pravdepodobnosť výskytu študovanej premennej nad túto prahovú hodnotu. Z dôvodu zvýšenia záujmu o takéto geometrické problémy sa začali vyvíjať techniky pre odhady alebo simulácie náhodnej funkcie $Z(x)$ pomocou indikátora $I[Z(x) \geq z] = 1 - I[Z(x) < z]$ pre rôzne "cut-off" prahové hodnoty $z$. Koncept indikátorových transformácií je jeden z najjednoduchších a (pravdepodobne) najelegantnejší v modernej geoštatistike. Hlavnou myšlienkou tohoto konceptu je určiť výberové "cut-off" kritérium – zvyčajne hodnoty nášho záujmu. Táto hodnota nemusí byť nevyhnutne ekonomický cut-off ťažobného podniku alebo kritická úroveň toxicity pri oceňovaní parametrov životného prostredia, ale napríklad aj hodnota vplývajúca na ložiskový mechanizmus študovanej premennej (Clark, 2000).

## Štruktúrna analýza

Rozdelením náhodnej funkcie $Z(x)$ rôznymi cut-off kritériami dostávame náhodné súbory, ktorých počet je rovný počtu aplikovaných cut-off indikátorov a ktorých štruktúra sa vzťahuje k štruktúre medzi pármi bodov $(x, x+h)$ náhodnej funkcie $Z(x)$. Tieto súbory sú vzájomne závislé a ich vzájomné priestorové rozloženie je veľmi dôležitou štruktúrnou charakteristikou študovanej premennej (Goovaerts, 1993).

Predpokladajme, že náhodná funkcia $Z(x)$ je stacionárna s dvojrozmerným rozdelením $(Z(x), Z(x+h))$ a kovarianciou $C(h)$. Cut-off hodnoty $z$ rozdeľujú priestor do náhodných súborov bodov s hodnotami $\geq z$, pričom očakávaná hodnota (matematická nádej) daného indikátora je:

$$E\{I[Z(x) \geq z]\} = P[Z(x) \geq z] = T(z). \qquad (1)$$

Kovarianciu indikátora pre dané cut-off $z$ potom môžeme zapísať:

[1] *Ing. Ladislav Vizi*, Katedra geológie a mineralógie, Fakulta BERG Technickej univerzity v Košiciach, Park Komenského 15, 040 01 Košice
(Recenzované a revidovaná verzia dodaná 13.6.2001

$$C_z(h) = Cov\{I[Z(x) \geq z], I[Z(x+h) \geq z]\} =$$
$$= P[Z(x) \geq z, Z(x+h) \geq z] - T(z)^2 . \qquad (2)$$

Z vyjadrenia vyplýva, že takáto pravdepodobnosť je vypočítaná z dvojíc bodov *(Z(x), Z(x+h))*, oddelených vektorom *h*. Príslušný variogram pre danú cut-off hodnotu *z* bude:

$$\gamma_z(h) = 0{,}5E\{I[Z(x) \geq z] - I[Z(x+h) \geq z]\}^2$$
$$= 0{,}5(P\{I[Z(x) \geq z] \neq I[Z(x+h) \geq z]\}). \qquad (3)$$

Potom $0{,}5(P[Z(x) < z, Z/x+h) \geq z] + P[Z(x) \geq z, Z(x+h) < z])$ sú funkcie dvojrozmerného rozdelenia *(Z(x), Z(x+h))*. Z toho vyplýva, že je možné priamo vypočítať štruktúru daného indikátora pod hypotézou daného dvojrozmerného rozdelenia (Goovaerts, 1993). Štruktúra indikátora *I[Z(x) ≥ z]* sa mení zároveň so zmenou hodnoty cut-off *z*: okrem rozptylu sa mení aj tvar štruktúry.

Treba poznamenať, že integráciou indikátorových variogramov pre všetky nami zvolenými cut-off *z* hodnôt dostávame variogram prvého rádu náhodnej funkcie *Z(x)*: $\int \gamma_z(h)dz = 0{,}5E[Z(x+h) - Z(x)]$.

Každá dvojica vzoriek vstupujúcich do výpočtu experimentálneho variogramu bude:

- {0,0} ak sú obe vzorky menšie ako diskriminačná cut-off hodnota: ich rozdiel je rovný 0,
- {1,1} ak sú obe vzorky väčšie ako cut-off hodnota: ich rozdiel je rovný 0,
- {0,1} alebo {1,0} ak je jedna vzorka väčšia a druhá menšia ako daný cut-off: ich rozdiel je rovný 1.

Výpočet experimentálnych bodov variogramu spriemerňuje tieto rozdiely a umožňuje predpovedať výskyt hodnoty nad alebo pod danou hodnotou cut-off (Clark, 2000).

**Odvodenie systému krigovacích rovníc pre indikátorové odhady**

Nech $z_k$ ; *k=1,...,K* bude súbor *K* vzájomne oddeliteľných kategórií pozorovaných na študovanej oblasti *O*. V každej priestorovej pozícii vzorky $x_i$ môže byť vytvorený vektor *K* indikátorových hodnôt $i(x_i, z_k)$ kde:

$$i(x_i, z_k) = \begin{cases} 1 & \text{ak} \quad x_i \notin I_k . \\ 0 & \text{ak} \quad x_i \in I_k , \end{cases} \qquad (4)$$

Algoritmus odhadu začína odhadom kondičných pravdepodobností *K* kategórií v každom bode $x \in O$, čiže:

$$p(x, z_k) = P\{x \in z_k \mid (N)\} = P\{I(x, z_k) = 1 \mid (N)\}, \qquad (5)$$

kde označenie "*N*" reprezentuje počet použitých kondičných informácií. Odhadovaná podmienená pravdepodobnosť indikátorového krigovania pre danú kategóriu $z_{k_O}$ je vypočítaná ako lineárna kombinácia okolitých indikátorových dát $i(x_i, z_k)$ vstupujúcich do výpočtu a ich príslušných váh $\lambda_k$:

$$p^*_{IK}(x_O, z_{k_O}) = \sum_{i=1}^{N} \lambda_{k_O}(x_i, z_{k_O}).i(x_i, z_{k_O}) . \qquad (6)$$

Riešenie tohoto systému si vyžaduje *K* smerových indikátorových modelov variogramov $\gamma_I(h, z_k)$, kde "*h*" je vzdialenosť medzi známymi bodmi $x_i$ a $x_j$, ($x_j = x_i + h$), patriacich do súboru indikačných hodnôt, t.j. do súboru obsahujúceho kondičné informácie spĺňajúce podmienky kategórie $z_{k_O}$. Výsledný krigovací systém bude:

$$\begin{cases} \sum_{j=1}^{N} \lambda_j(x_j, z_{k_O}) \bar{\gamma}_I(x_i - x_j, z_{k_O}) - \mu_k = \bar{\gamma}_I(x_i - x_O, z_{k_O}) & \forall \ i = 1,...,N , \\ \sum_{j=1}^{N} \lambda_j(x_j, z_{k_O}) = 1 \end{cases} \qquad (7)$$

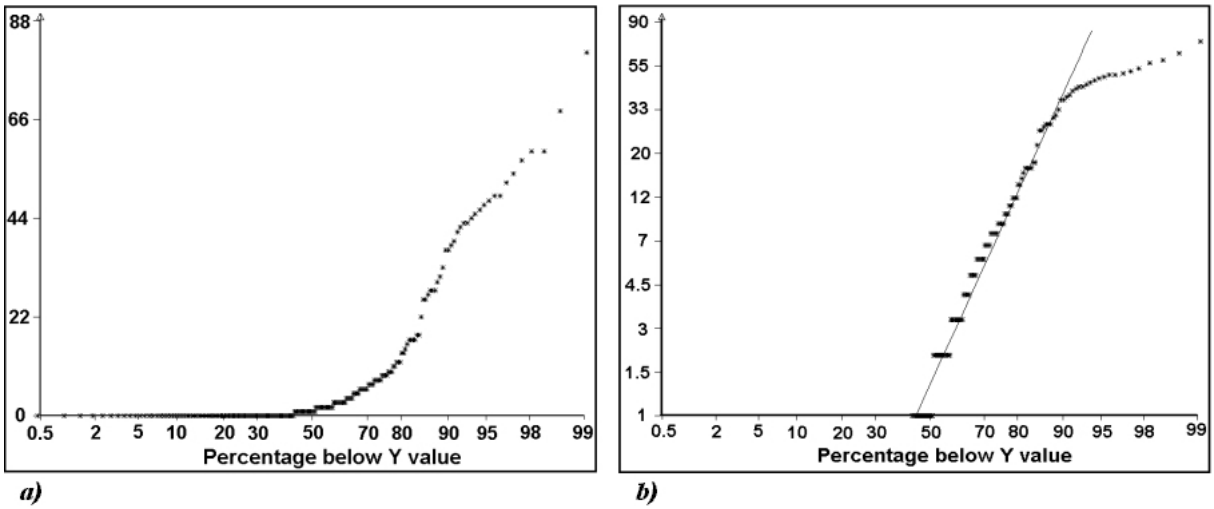
a príslušný krigovací rozptyl bude:

$$\sigma_{IK}^2 = \sum_{i=1}^{N} \lambda_i(x_i, z_{k_O})\bar{\gamma}(x_i - x_O, z_{k_O}) - \bar{\gamma}(x_O - x_O, z_{k_O}) + \mu_k \,, \qquad (8)$$

kde $\mu_k$ sú Langrangové multiplikátory minimalizujúci rozptyl odhadu za podmienky, že suma váh musí byť rovná 1, zavedením ktorého dostávame *N+1* lineárnych rovníc nazývaných krígovací systém odhadu.

**Príklady použitia indikátorového prístupu na regionálnej premennej**

Predpokladajme dátový súbor so súradnicami odberu vzoriek a spojitou študovanou premennou. Keď sa pozrieme na zobrazený pravdepodobnostný graf (Geostokos Toolkit, 2000) premennej (*obr.č.1a*) môžeme z jeho priebehu predpokladať jej lognormálne rozdelenie študovanej premennej.



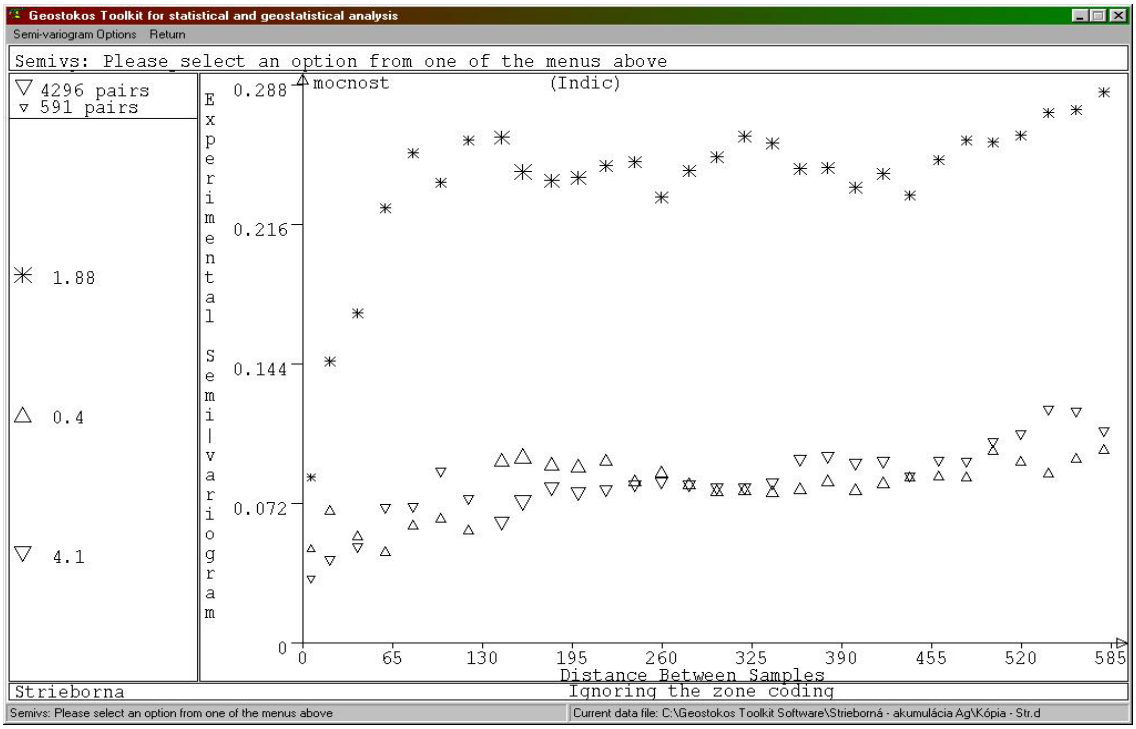

*Obr.1. Pravdepodobnostný graf študovanej premennej: a) bez nastaveného modelu distribúcie, b) s nastaveným modelom distribúcie.*
*Fig.1. Probability plot of variable under study: a) without a fitted model of the distribution, b) with a fitted model of the distribution.*

Po nastavení logaritmického modelu (*obr.1b*) sa objavuje viac ako len jedna zložka rozdelenia premennej. Z toho vyplýva, že rozdelenie premennej má aspoň dve oddelené distribúcie, tvoriace pravdepodobnostný graf, ktorý je rozdelený približne v hodnote 40. Nad touto hodnotou graf znižuje svoj sklon, z čoho sa dá usudzovať, že druhá zložka logaritmického rozdelenia obsahujúca vyššie hodnoty má nižšiu štandartnú odchýlku ako zložka nižších hodnôt. Po zakódovaní hodnôt premennej podľa diskriminačnej hodnoty 40, spriemernením rozdielov zakódovaných hodnôt podľa predošlej teórie dostávame semivariogram, odrážajúci podiel výskytu hodnôt nad alebo pod daný cut-off študovanej populácie.

Na *obr. č. 2* je vykrigovaná mapa študovanej premennej, zakódovanej podľa diskriminačnej hodnoty 40. Kolorované kontúry takejto mapy reprezentujú určitú odhadovanú pravdepodobnosť výskytu populácie s hodnotami vyššími ako daný cut-off. Na základe takejto interpretácie je ďalej možné vyvodiť záver, že vo východnej časti študovanej oblasti v smere sever – juh je veľká pravdepodobnosť výskytu hodnôt premennej nad 40 (populácia nad 40 je v tejto oblasti prevládajúca) a táto časť môže byť ďalej bližšie študovaná.

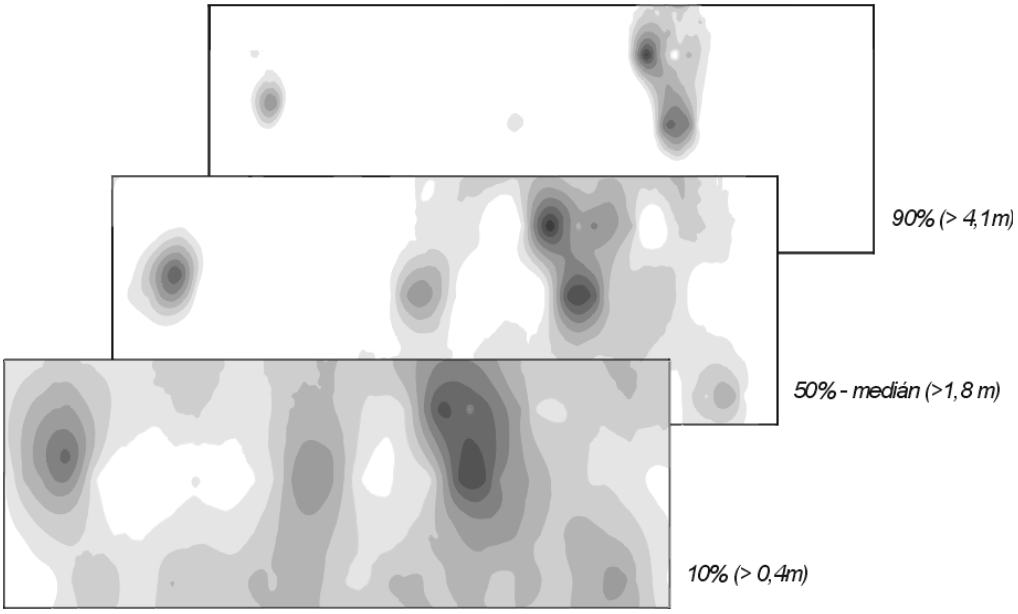
Na *obr. č. 3* sú zobrazené experimentálne semivariogramy premennej – mocnosti žily Strieborná v Rožňavskom rudnom poli pre percentily 10, 50 a 90. Tieto experimentálne semivariogramy zobrazujú štruktúru pravdepodobnosti výskytu mocnosti nad 0,4 m (pre percentil 10), nad 1,88 m (pre percentil 50) a nad 4,1m (pre percentil 90). Je dôležité poznamenať, že prah indikátorového semivariogramu nie je v skutočnosti rozmerom variability, ale skôr odráža podiel populácie nad a pod daným cut-off. Z toho vyplýva, že najväčší prah bude dosahovať indikátorový semivariogram skonštruovaný na hodnote mediánu študovanej premennej, čiže pre percentil 50 (v tomto prípade pre diskriminačnú hodnotu 1,88 m), kde sa 50% vzoriek stáva "0" (50% populácie je pod daným cut-off) a 50% "1" (50% populácie je nad daným cut-off), a preto má maximálny počet rozdielov.


*Obr.2. Mapa pravdepodobnosti výskytu hodnôt premennej nad 40.*
*Fig.2. Probability map of the occurrence of variable values above 40.*




*Obr.3. Experimentálne semivariogramy pre percentily 10 (>0,4 m), 50 (>1,88 m) a 90 (>4,1 m) mocnosti žily Strieborná.*
*Fig.3. Experimental semivariograms for percentiles 10 (>0.4 m), 50 (>1.88 m) and 90 (>4.1m) of the Strieborná vein thickness.*

Na *obr. č. 4* sú výsledné krigované mapy pravdepodobnosti rozloženia študovaných percentilov mocnosti Striebornej žily:




*Obr.4. Mapy pravdepodobností rozloženia mocnosti Striebornej žily podľa študovaných percentilov.*
*Fig.4. Probability maps of the thickness distribution of Strieborná vein by the percentiles under study.*

## Záver

Z predložených príkladov je jednoznačne zrejmé, že v praxi sa je možné stretnúť nie len s kvantitatívnymi premennými, ale často aj s premennými, ktoré je možné zakódovať podľa určitých kritérií do príslušných kategórií. Z toho dôvodu sa je možné stretnúť s problémami odhadu pravdepodobnosti každej kategórie prevládajúcej na určitej lokalite. Použitím indikátorového prístupu je možné tieto pravdepodobností odhadnúť. Indikátorové krigovanie sa dnes už udomácnilo nielen v banskej praxi pre ktorú boli tieto postupy vytvorené, ale aj v iných oblastiach výskumu ako je oceňovanie parametrov životného prostredia podľa rôznych kritérií znečistenia alebo v lesníckom či poľnohospodárskom priemysle pre odhad menej produktívnych oblastí, či vykreslenie oblastí najviac infikovaných škodlivou populáciou rastlín. Nepochybne veľkej popularite sa indikátorové krigovanie (či skôr indikátorový ko-kriging) teší v geologickom modelovaní vrstiev geologického prostredia alebo sedimentárnych procesov.

## Literatúra

CLARK, I.: Practical Geostatistics 2000. *Greyden Press,* Columbus, Ohio, U.S.A., 2000, pg.329-333.

GOOVAERTS, P.: Comparison of co-indicator kriging, multiple indicator kriging and median indicator kriging performances for modelling conditional probabilities of categorical variables. In: Roussos Dimitrakopoulos (editor): Geostatistical for next century: an international forum in honour of Michel David´s contribution to Geostatistics, Montreal 1993, pg.18-29.

JOURNEL, A.: Indicator Approach to Estimation of Spatial Distribution. 17th APCOM symposium, Colorado, 1982.

Geostokos Toolkit for Windows – software for 2D or 3D statistical and geostatistical analysis of spatial data. Geostokos (Ecosse) Limited, Alloa, Central Scotland, March 2000.