ГОДИШНИК НА МИННО-ГЕОЛОЖКИЯ УНИВЕРСИТЕТ “СВ. ИВАН РИЛСКИ”, Том 49, Св. I, Геология и геофизика, 2006
ANNUAL OF THE UNIVERSITY OF MINING AND GEOLOGY “ST. IVAN RILSKI”, Vol. 49, Part I, Geology and Geophysics, 2006

КАЛИБРИРАНЕ НА ГЕОСТАТИСТИЧЕСКИ МОДЕЛИ НА РУДНИ НАХОДИЩА ЧРЕЗ “CROSS-VALIDATION” ТЕХНИКА

Светлозар Бакърджиев

Минно-геоложки университет “Св. Иван Рилски”, София 1700; zarcobak@mgu.bg

РЕЗЮМЕ. Кригинг алгоритъмът минимизира пресметнатите несходства (разлики)по предварително определен модел на ковариация. Най-често за тази цел се използва линейна регресия. Основно и най-силно качество на метода е възможността модела да се калибрира по изходните данни. Правилният избор на вероятностния модел е гаранция за адекватното симулиране на стохастичната компонента на геостатистическия модел. В статията се предлага, калибрирането на модела се извършва чрез нелинейна регресия, която минимизира пресметнатите несходства (разлики) между изходните данни и прогнозите в техните координати (методът Cross-Validation). Резултатът е оценката на двата параметъра на характеристичната функция на “Устойчивото” разпределение. Методът е успешно апробиран по данни на рудни находища. Налице е подобряване на геостатистическите модели по отношение трите адитивни компоненти като вариациите на шума, на оценяваната променлива и квадратичната оценка на нейното отместване.

CALIBRATE OF GEOSTATISTICS ORE DEPOSIT MODELS THROUGH CROSS VALIDATION TECHNICS

Svetlozar Bakardjiev

University of Mining and Geology “St. Ivan Rilski”, Sofia 1700; zarcobak@mgu.bg

ABSTRACT. Kriging algorithm minimizes estimations of a dissimilarity (difference) on are preliminary set covariance model. More frequently for this purpose uses linear regress. The basic and strongest quality of this method is opportunities of model calibrate through the input data. The good choice model of probability is a guarantee for adequate of stochastic component of a geostatistical model. In work it is offered, calibrations of model it was made through linear regresses which minimizes the appreciated dissimilarities between the input data and prognosis in their coordinate (Cross-Validation Method). The result is estimation of two parameters of characteristics function of Stable distribution. The method is successes examination up to ore deposits data. It is available improvement of Geostatistical model concerning three additive parts, noise variance, and estimation variance, and squared estimation bias.

Въведение

Калибрирането на геостатистическите модели е свързано със следната статистическа задача. Нека са известни съдържанията в известен брой проби. Обектът се разбива на N еднакви по форма блокове V_i, всеки от който има обем V. За всеки блок е намерена оценка на средните съдържания $Z^*(V_i)$. Необходимо е да се определи за даден блок V_i, имащ оценка $Z^*(V_i)$, разпределението на честотите на съдържанията $Z(v)$ вътре в обема V_i. Освен това е желателно да се знаят стойностите в по-малки участъци, където стойностите не превишават допустимите минимални стойности. Средната стойност в по-малки участъци, може да превишава някаква гранична стойност. Следва, че размерът на блока не влияе съществено върху стандартната статистическа оценка, докато при метода „Кригинг” той е най-съществената част на оценката. При малко на брой наблюдения, характерни предимно при стадия на предварителното проучване, Кригинг оценките обикновено се отместват от истинското средно. В този случай процедурата може да се разглежда като начин за управление на извадката в процеса на обработката на данните.

В случай, че информацията за изследвания от нас обект е малко тогава неговите пространствени свойства могат да се моделират с известна доза неопределеност, която идва от недостатъчната информация, с което броят на възможните варианти на интерпретация е много голям. В този случай е най-логично изходните данни да се разделят случайно на две еднакви съвкупности. Едната от съвкупностите се използува за получаването на интересните за изследването свойства, а другата – за проверка на изместеността на оценката. Еволюцията на тази идея е довела до въпроса: защо да се дели извадката само веднъж, защо да се дели наполовина, не може ли по-малки части за участвуват в експеримента и т.н.?

Когато броят на данните не може да се увеличава, задачата за подобряването на геостатистическата оценка се свежда до подобряването на оценката, чрез максималното й приближаване към изходните данни. Естествено, това съвпада с идеята на „фитването” на регресионните модели, с тази разлика, че при геостатистическите модели „липсва” аналитичния модел – аналитична функция или диференциално уравнение, което обикновено се решава чрез методите на крайните

елементи, крайните разлики или граничните елементи[1]. В тази връзка, калибрирането на тези модели е силно и успешно развито в областта хидрогеологията, като теоретичните и практическите аспекти на калибрирането на хидрогеоложките геостатистически модели са развити и представени най-добре от Хил, Кули и Потлък (Hill, Cooley and Pollock, 1998), в съответния труд. За съжаление, предлаганият формализъм и конкретната компютърна реализация в програмните модули UCODE и MODFLOWP са пригодени и апробирани за работа единствено с хидрогеоложки данни.

Граници на възможната грешка

От позицията на геостатистическия формализъм е необходимо е да се пресметнат средните разлики в различни направления, представени от експерименталните вариограми и да се подбере подходящ теоретичен модел, който адекватно да описва природната структура на променливостта в използваните данни. В този контекст, първата стъпка в геостатистическия анализ е построяване на експериментална вариограма. Както е известно, тя представлява двумерна зависимост (двумерна X-Y точкова диаграма) изразяваща разстоянието (h) между съвкупност от пространствено ориентирани променливи (обикновено характеризиращи свойствата на някакъв пространствен обект) групирани по двойки (всяка с всяка), от средно аритметичното от квадрата на разликите в техните стойности (γ).

/1/ $\gamma(h) = \frac{1}{n}\sum_{i=1}^{n}\left(f_{1i} - f_{2i}\right)^2$.

С други думи при построяването на вариограма по оста X се нанася разстоянието (h) между дадена двойка стойности, за която се пресмята (γ), нанасящо се по оста Y. Получените точки се свързват помежду си с начупена линия, наречена вариографна крива.

Най-често при вариограмните анализи се използват т.н. "полувариограма"

/2/ $\gamma(h) = \frac{1}{2n}\sum_{i=1}^{n}\left(f_{1i} - f_{2i}\right)^2$, т.е. използва се половината от стойността на (γ), така че нанесените стойности стават еквивалентни на статистическата дисперсия, или логаритмичната полувариограма – еквивалент на полувариограмата с логаритмувани стойности по Y.

Подробности относно математическия апарат на метода може да се намерят в Deutsch and Journel, 1992, Isaaks and Srivastava, 1989, Journel and Huijbregts, 1978 и др. Съществена част от геостатистиката интерпретира и развива проблемите, свързани с пространствения анализ на първичните данни.

Пространствен анализ

Съгласно Ж. Матерон (Matheron, 1963; 1971) пространствената променлива не може да се дефинира строго математически, тъй като е преходно звено между дискретни и непрекъснатите случайни величини. Тя се изменя в пространството от едно местоположение до следващото по "непрекъснат" начин и по тази причина близко разположените точки имат висока степен на пространствена корелация, докато точки, които са значително отдалечени са статистически независими. В този аспект, вариограмният анализ е съществена част от геостатистическото моделиране. Неговата цел е изявата на закономерностите в пространствената променливост на изучавания геоложки показател (Matheron, 1971; Rendu, 1981) и др. По дефиниция, стойността на вариограмата 2γ (h) за дадено разстояние h в рудното тяло Ω е очакваната квадратична разлика между стойностите на пробите на разстояние h:

$2\gamma(h) = E\{[x(z) - x(z+h)]^2\}$.

Функцията γ(h) е полувариограма:

$$\gamma(h) = \frac{1}{2}E\{[x(z) - x(z+h)]^2\}$$

Оценката $\overset{щ}{\gamma}(h)$ на полувариограмата γ(h) е:

$$\overset{щ}{\gamma} = \frac{1}{2n(h)}\sum_{i=1}^{n(h)}[x(z_i) - x(z_i+h)]^2.$$

където $x(z_i) - x(z_i+h)$ са n(h) на брой разлики между стойностите (съдържанията на полезен компонент), които са измерени в проби намиращи се на разстояние h помежду си в дадено направление.

Получените стойности за γ(h) се нанасят на x-y диаграма, където по оста x се нанася разстоянието h, а стойностите на гама се нанасят на оста y. Следователно, основното предимство на вариограмния анализ е, че чрез него се отчитат като средно различни природни свойства, като анизотропия, характер и степен на зависимост между съседните проби при различни разстояния между тях, ниво на общата променливост, прекъснатост и др., които са характерни за изследвания обект.

Сведения за метода "Cross Validation"

Термините "кръстосано сравнение" (Cross-validation) и Бутстрап (bootstrapping) са идеологически близки методи за оценка на общата грешка на геостатистическия модел, основаваща се на преоценка на случайни извадки от общата съвкупност от данни - (Weiss and Kulikowski 1991; Efron and Tibshirani 1993, White 1994). Методът се основава на статистики, получени от K на брой извадки от изходната съвкупност от данни, които съдържат еднакъв брой данни. В зависимост от получената грешка на кригинга, извадките се класифицират на "добри" и "лоши", като се търсят причините за разликите в сумарната грешка. Кригинг параметрите се коригират (апроксимират) спрямо резултатите от добрите извадки. Смята се, че методът е по-ефективен при малко на брой изходни данни – виж Goutte (1997). В тази връзка Стоун (Stone, 1977) предлага при по-големи извадки да се оценява генералното отместване на кригинг оценката, вместо да се оценява общата грешка по случайни извадки, която очевидно ще нараства пропорционално на броя на данните от

[1] Изисква се наличието на фундаментално решение

съвкупността. Нека Z_i, Z_j, представят две случайно избрани половини на изходната съвкупност от данни. Ако съвкупността от данни е геостатистически еднородна, то $Z_i \sim Z_j$, т.е. облакът от данни трябва да е с минимални отклонения от линейната регресия. Всяко отклонение или липса на корелация свидетелства, че броят на проучвателните изработки е недостатъчен за коректното прилагане на кригинга. В повечето от случаите е по-важно грешката, с която се съгласяваме да бъде по-ниска от тази, която ни лимитират кондициите.

За съжаление, повечето от кригинг методите слабо отчитат влиянието на количеството от данни върху сумарната грешка. Чрез методите на случайните извадки е възможно да се предложат множество обработки, вместо една, т.е. вместо една да се получат множество оценки. По тях е възможно да се строят доверителни интервали и да се следи за асимптотическото поведение на оценките, т.е. към каква стойност клонят същите. По този начин може да се подобри и оцени качество на оценката.

Методът "джобно ножче" (Jackknife method)

Идеята на метода се свежда до изключване на едно наблюдение, обработка на цялата останала информация и предсказване на резултата в изключената точка (наблюдение). По този начин е възможно да бъдат получени разлики (отмествания) за всички наблюдения. Един път те дават информация за дисперсията, втори път може да се построи корелационна диаграма между оценките и накрая – да се построи карта на разликите, т.е. да се планира мястото на новите изработки. В конвенционалната статистика, този метод още е познат и под името "метод на джобното ножче (Jack Knife)" тъй като е удобен да замени всички останали техники, както джобното ножче е призвано да замени всички останали инструменти. Начините за ползването на метода са описани в специалната литература. По-нататък ще видим, че формализмът е основополагащ при калибрирането на геостатистическия модел виж Исакс и др., 1989 (Isaaks and Srivastava, 1989).

При търсенето на ефективност допълнително може да се използва и т. нар. метод на интерполация с тегла на стойностите обратно пропорционална на отдалечеността – IDW, Inverse Distance Weighted Interpolation". Методът е известен още и като метод на Шепард – виж по-подробно описание в (Goovaerts, 1997). Той е един от най-често използваните интерполационни методи поради неговата простота и ефективност, слабо зависещи от броя на изходните данни. Основава се на допускането, че при определяне значението на всяка точка от "равномерната" мрежа, стойностите на по-близките точки трябва да бъдат вземани с по-голяма тежест от по-отдалечените, което е по същество и основополагащата се идея на метода «Кригинг». В соответствие с метода, стойността на всяка точка от новата мрежа на прогнозни, в общия случай интерполационни стойности, се определя на база на "N" на брой най-близки точки, като стойността на всяка една, използвана при интерполацията точка, участва с обратно-пропорционално тегло в зависимост от отдалечеността й спрямо търсената. В сравнение с чистата интерполация този метод дава възможност за получаване на по-загладени повърхнини и за частично отстраняване на негативния отпечатък на грешни или аномални точки.

Методът има следва следния формализъм:

$$F(x,y) = \sum_{i-1}^{n} w_i . f_i ,$$

където n е броят на използваните точки от "неравномерната" (началната) мрежа, f_i е функцията, определяща разпределението на стойностите (prescribed function values) (стойностите на данните от "неравномерната" мрежа), а w_i е подходящо избрана функцията на отдалечеността, приложена за всяка точка от неравномерната мрежа. Класическата форма на тази функция може да се изрази чрез уравнението:

$$w_i = \frac{h_i^{-p}}{\sum_{j=1}^{n} h_j^{-p}}$$, където:

p е произволно положително реално число, определящо степенния параметър (това число определя в действителност степента на заглаждане на повърхнината). В конкретните случаи $p = 2$.

h_i е разстоянието от всяка точка от неравномерната до всяка изчисляваща се точка от равномерната мрежа.

/4.7/ $h_i = \sqrt{(x - x_i)^2 + (y - y_i)^2}$, където (x,y) са координатите на интерполираните (получените по равномерната мрежа) точки, а (x_i, y_i) са координатите на използваните точки от началната (неравномерната) мрежа.

Функцията на влиянието или т. нар. "тегловна функция" варира от единица до стойност, клоняща към нула при повишаване на разстоянието между пресмятащата се точка и тази от неравномерната мрежа. Тегловната функция е нормализирана, така че сумата от теглата е равна на единица.

Класическата форма на функцията има вида:

$$w_i = \frac{\left[\frac{R - h_i}{Rh_i}\right]^2}{\sum_{j=1}^{n}\left[\frac{R - h_j}{Rh_j}\right]^2}$$, където:

h_i е разстоянието от интерполиращата се точка до точка i от неравномерната мрежа;

R е разстоянието от интерполиращата се точка (участваща в изчисленията) до най-отдалечената точка от неравномерната мрежа;

n е броят на точките от неравномерната мрежа, участващи в изчисленията.

Тегловната функция е функция на "Евклидовото" разстояние (нарича се още Евклидова норма) е радиално симетрична на всяка една от началните точки от неравномерната мрежа. Като резултат получената повърхност е повлияна от всички използваните точки от неравномерната мрежа и следи тенденциите на изменение на последната, като притъпява ефекта на аномалните участъци. За тримерни определения формулата е абсолютно идентична на използваната в двумерното пространство и има вида:

$$h_i = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2}$$

Калибровка на модела

Както се спомена във въведението, калибровката представлява важна част от моделното решение. Тя цели чрез промяна на характеристиките на подходящо подбрани, предимно екстремни стойности на граничните условия на геоложката среда или на измерените в конкретни точки в масива стойности на изследваните компоненти, да се търси моделно решение, което максимално точно да интерпретира реалната геоложка обстановка.

Като изходни параметри, по които се извършва калибровката най-често служат измерени (опробвани) стойности на съдържанията на полезни и вредни компоненти в определени точки. Обикновено на корекция се подлагат характеристики на средата, където е налице по-голяма гъстота на геоложкото опробване.

За калибриране е подходящо да се използват и най-близките точки с налична информация. За калибрирането на геостатистическите модели на рудни находища се предлага следния формален подход – алгоритъм:

Предлагаме, калибрирането на модела да се извършва по следния начин:

Търси се минимума на функцията:

$$F = \sum_{i=1}^{N} \left(Z_i - \gamma Z^*_i \right)^{\alpha} \rightarrow \min$$

,

където Z_i е истинската (измерената) стойност на i-тото наблюдение;

Z^*_i е прогнозната стойност в координатите на i-тото наблюдение;

α е степенен показател, $0 < \alpha < 2$;

γ е мащабен показател, чрез който се търси намаляване на систематичната грешка.

Резултатът от калибрирането е намирането на такива стойности на α и γ при които $F \rightarrow \min$.

Кригинг коефициентите (w) се пресмятат по израза:

$$w = \left(\frac{\gamma^3}{\|d_1\| * \|d_2\| * \|d_3\|} \right)^{\alpha}$$

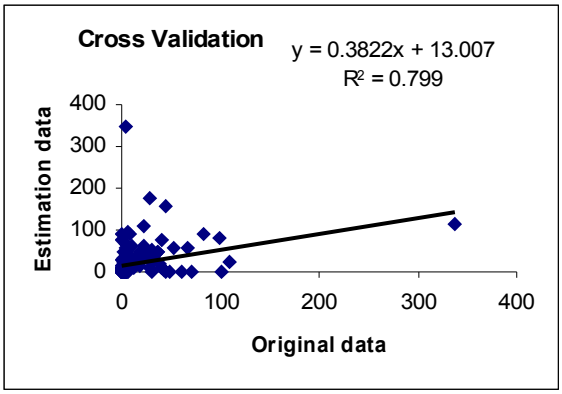
където $\|d_j\|$ е евклидовото разстояние между точката на прогноза и i-тото наблюдение, като $\|d_1\| * \|d_2\| * \|d_3\|$ проекциите на $\|d_j\|$ в тримерното пространство.

Тогава Z^*_i се получава чрез уравнението на Кригинга $Z^* = \sum w_i z_i$, като се спазва условието $\sum w_i = 1$.

Параметрите α и γ имат съответствието на параметрите на характеристичната функция на симетрично Устойчиво разпределение, виж Леви (Levy, P. 1925). Съгласно теоремата на Хинчин (Khinchin. A. 1938) параметърът α може да приема стойности в интервала 0–2, като при стойност равна на 2, дисперсията на данните има крайна дисперсия. При α>2 дисперсията на извадката става безкрайна и от „нормално" разпределението става „Устойчиво". При α=1 разпределението става разпределението на „Коши", където и средната стойност е безкрайност. Параметърът γ има смисълът на мащабен множител, който може да служи и за косвена оценка на систематичната грешка на калибрирането.

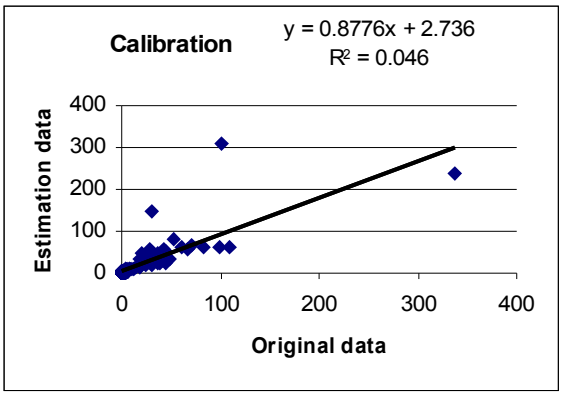
Апробация

Използвани около 2300 данни от експлоатационното проучване на един вече отработен участък на находище "Челопеч". На фигура 1 и 2 са представени получените графики по метода на кръстосаната проверка. На фиг. 1 се вижда се, че са налице много на брой екстремни стойности, които силно деформират посоката на тенденцията за съответствие на данните. Този резултат показва, че геостатистическия модел се нуждае от корекции и\или планиране на ново опробване. Тъй като последното е невъзможно, се построи карта на разликите между истинските стойности и съответните оценки. Коефициентите на линейния модел имат съответно стойности 0.3822 за наклона на тренд линията и 13.007 за отместването по ординатната ос. Съответствието е много лошо, стойността 0.3822 е много отдалечена от стойността единица на „идеала", а стойността 13.007 е драматично отдалечена от стойността нула на „идеала". Всъщност, тази стойност съвпада със стойността на систематичната грешка на модела.



Фиг. 1. Взаимовръзка между реалните съдържания на мед в пробите (Cu) и техните кригинг оценки

Следователно, калибрирането е значително подобрило общата геостатистическа оценка, което личи и от сравнението между стойностите на R^2 при некалибрирания и калибрирания модел.



Фиг. 2. Взаимовръзка между реалните съдържания на мед в пробите (Cu) и техните крайгинг оценки, коригирани след калибрирането на модела

След калибрирането на модела – виж фиг. 2, се вижда, че стойността 0.8776 на наклона е значително близка до единицата, а стойността 2.736 е значително по близо до нулевата стойност, отколкото 13.007 в не калибрирания модел.

Методиката на калибриране е апробирана успешно при масиви до 3000 наблюдения. Калибрирането при по-големи масиви ще се затруднява изчислително, поради огромния обем от сметки и от съществуването наличие на екстремни стойности – „ураганни проби".

Литература

Deutsch, C. V., A. G. Journel. 1998. *GSLIB: Geostatistical Software Library and User's Guide.* Second Edition Oxford University Press, New York.

Efron, B., R. J. Tibshirani. 1993. *An Introduction to the Bootstrap.* London, Chapman & Hall.

Geostatistics for Natural Resources.

Goovaerts, P. 1997 Evaluation of Geostatistics Methods. Oxford University Press, New York.

Goutte, C. 1997. Note on free lunches and cross-validation. – *Neural Computation*, 9, 1211-1215

Hill, M. C., R. L. Cooley, D. W. Pollock. 1998. A controlled experiment in ground-water flow model calibration using nonlinear regression. – *Ground Water, 36*, 520-535.

Isaaks, E. H., R. M. Srivastava. 1989. *An Introduction to Applied Geostatistics.* Oxford University Press, New York.

Journel, A. G., C. Huijbregts. 1978. *Mining Geostatistics*. Academic Press.

Khinchin, A. 1938. *Limited laws for sums independed random variables.* O.N.T.I., Moscow - St. Petersburg.

Levy, P. 1925. *Calcul des probabilities*. Paris, Gauthier-Villars et Cie, 350 p.

Matheron, G. 1963. Principles of geostatistics. – *Economic Geology, 58*, 1246-1266.

Matheron, G. 1971. The Theory of Regionalised Variables and its Applications. – Cahier No. 5, Centre de Morphologie Mathematique de Fontainebleau.

Rendu, J.-M. 1978. *An Introduction to Geostatistical Methods of Mineral Evaluation*. Monograph of the South African Inst. Min. Metall.

Stone, M. 1977. Asymptotics for and against cross-validation. – *Biometrika, 64*, 29-35

Weiss, S. M., C. A. Kulikowski. 1991. *Computer Systems That Learn*. Morgan Kaufmann.

Препоръчана за публикуване от
Катедра "Геология и проучване на полезни изкопаеми", ГПФ