Sborník vědeckých prací Vysoké školy báňské – Technické univerzity Ostrava
Řada hornicko-geologická
Volume LI (2005), No.2, p. 79-88, ISSN 0474-8476

František STANĚK[*], Vlastimil KAJZAR[**]

VLIV TYPU STATISTICKÉ DISTRIBUCE LOŽISKOVÝCH ÚDAJŮ NA MODEL LOŽISKA NEROSTNÝCH SUROVIN

EFFECT OF NATURE OF STATISTICAL DISTRIBUTION OF DEPOSIT DATA ON DEPOSIT MODEL OF MINERAL RAW MATERIALS

Abstrakt

Pouze na základě správného popisu statistické distribuce zpracovávaných ložiskových údajů lze odhadnout statistické charakteristiky a realizovat následné zpracování. Je známým faktem, že empirické distribuce většiny veličin popisujících geologická tělesa nevyhovují běžně uvažovanému normálnímu rozdělení, ale že mají distribuci asymetrickou. Přitom je ale normální distribuce základní podmínkou použití mnoha dalších matematických postupů. V článku je popsáno programové řešení tohoto problému při modelování uhelné sloje v Interaktivním programovém systému pro aplikaci moderních metod hodnocení uhelných ložisek a jejich dílčích částí v komplikovaných podmínkách (vyvíjeném na IGI HGF VŠB-TU Ostrava v rámci řešení projektu GA ČR č. 105/03/1417) v případě, že je statistickým testem zjištěna jiná než normální distribuce. V tomto procesu se využívá tzv. kvantilová transformace vstupních údajů. V několika ukázkách je demonstrováno, jaké nepřesnosti bez použití této transformace vznikají při modelování atributů uhelné sloje.

Abstract

A correct formulation of statistical distribution of processed deposit data is a necessary condition to estimate statistical parameters and to realize subsequent data processing. It is a well-known fact that empirical distributions of most quantities describing geological bodies do not comply with a normal distribution, they are often asymmetric. However, most common mathematical procedures assume these quantities as normal distributed ones. We describe a method of transforming the initial data distributed asymmetrically into the normal distributed ones, applied in the field of coal seam modeling with use of Interactive software system for application of modern methods of evaluation of coal deposits and their parts under complicated conditions (This software is developed by IGI HGF VŠB-TU Ostrava within the frame of GA ČR project No 105/03/1417.). We use the so-called quantile transformation of input data. On several case studies we show which inexactitudes arise without application of such transformation when modeling attributes of coal seam.

Key words: statistical data processing, frequency distribution, skewed distributions, lognormal distribution, data transformation, quantile transformation procedure, geomodeling, mineral deposits, coal deposits.

Úvod

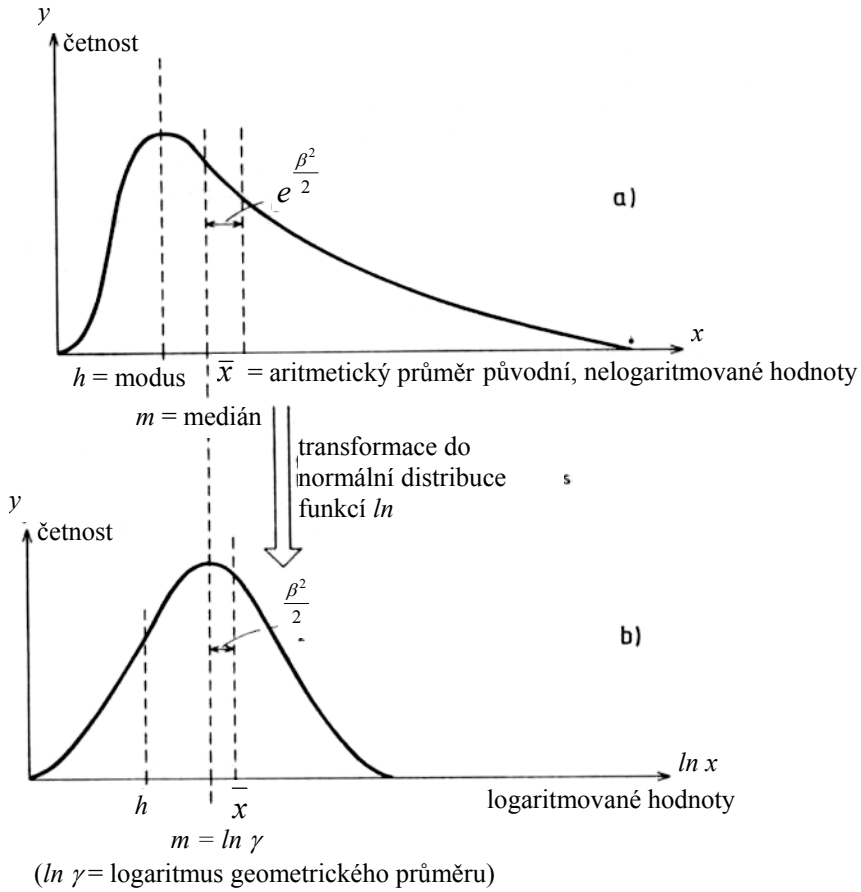
Jednou ze základních úloh statistického rozboru i volby dalších metod zpracování je studium charakteru statistické distribuce. Jen na základě správného popisu distribuce lze odhadnout statistické charakteristiky a realizovat další zpracování. Je známým faktem, že empirické distribuce většiny veličin popisujících geologická tělesa nevyhovují běžně uvažovanému normálnímu rozdělení $N(U;\mu,\sigma^2)$, ale že mají distribuci asymetrickou (převážně kladně) – nejčastěji lognormální. Přitom je ale normální distribuce základní podmínkou použití mnoha dalších matematických postupů. Bez správného přístupu k „zešikmeným“ datům není možné dělat například geostatistické analýzy a odhady, neboť nejlepší lineární odhad je ten, který je získaný z experimentálních hodnot, řídících se normálním Gaussovým rozdělením (Vizi, Timčák 2002).

[*] Doc. RNDr., Ph.D., Institut geologického inženýrství, VŠB-TU Ostrava, e-mail: frantisek.stanek@vsb.cz
[**] Ing., interní doktorand, Institut geologického inženýrství, VŠB-TU Ostrava,
e-mail: vlastimil.kajzar.hgf@vsb.cz

Vzniká proto otázka, jaká chyba vznikne při modelování ložiska nerostných surovin - například ložiska uhlí, pokud se neprovede transformace vstupních údajů pro zešikmená data.

Transformace ložiskových údajů

Pro popis zpravidla asymetricky rozložených veličin popisujících geologická tělesa se využívají různé transformace, např. obecná logaritmická $v = ln(a+bu)$, která vede k velmi často používanému obecnému lognormálnímu rozdělení $LN(U;\mu,\sigma^2,a,b)$, jejíž konstanty lze stanovit minimalizací šikmosti a koncentrace distribuce. Tento proces je schematicky demonstrován na obr. 1. Zešikmená distribuce je ukázána na obr. 1 a) s nejčetnější hodnotou modu *h*, mediánu *m* a střední hodnotou $\bar{x}$. Obr. 1 b) ukazuje rozdělení četnosti stejných hodnot po jejich logaritmické transformaci. Vrchol křivky četnosti nyní představuje medián. Jestliže je distribuce přesně lognormální, potom je medián *m* identický s logaritmem geometrického průměru γ. β^2 označuje rozptyl logaritmovaných hodnot.



Obr. 1: Převod doprava zešikmené distribuce do normální distribuce transformací jednotlivých hodnot funkcí ln(x) (Wellmer 1998, upraveno)

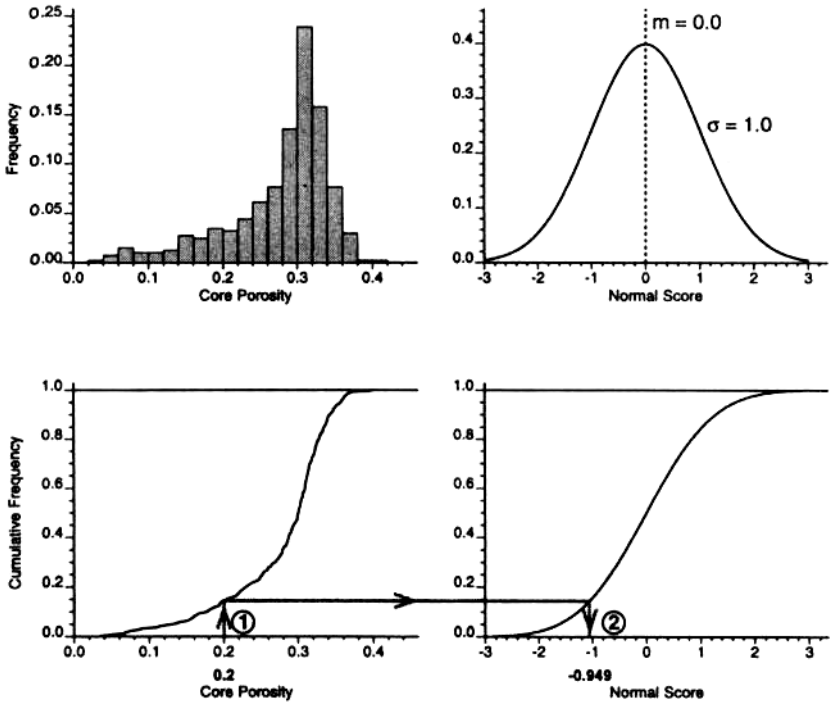
Figure 1: Transfer of a distribution skewed to the right into a normal distribution by transfer of the single values into logarithms (Wellmer 1998, adapted)

Je třeba si ale uvědomit, že lognormální distribuce je jedním z mnoha typů možných distribucí souboru hodnot. Existuje praktická cesta pro transformaci libovolné distribuce do distribuce normální a zpět – použití tzv. kvantilové (nebo také grafické) transformace vstupních údajů.

V případě, že vstupní soubor dat nevyhovuje normální distribuci, lze provést grafickou transformaci vstupního souboru tak, že výsledný soubor má normované normální rozdělení (se střední hodnotou nula a směrodatnou odchylkou 1). Tento proces ilustruje obr. 2. Tuto transformaci realizuje například program *nscore* v geostatistickém toolboxu GSLIB (Deutsch, Journel 1998). Tato transformace je oboustranná pod podmínkou, že histogram neobsahuje tzv. „spikes“ – interpopulační, konstantní hodnoty. Tyto je nutno před samotnou transformací vyhladit (Deutsch 2002).

V případě modelování uhelné sloje jsou z nepravidelně rozmístěných průzkumných bodů interpolovány hodnoty jednotlivých ložiskových atributů do pravidelné sítě bodů – tzv. gridu. Jelikož se jedná o lineární matematické postupy, je nutné, aby vstupní údaje byly rozloženy normálně. Aby se modelovaná sloj co nejvíce blížila realitě, je nutné použít pro interpolaci normálně rozložené vstupní hodnoty jednotlivých ložiskových atributů.

Proto je v programovém řešení Interaktivního programového systému pro aplikaci moderních metod hodnocení uhelných ložisek a jejich dílčích částí v komplikovaných podmínkách (dále IPSHUL) vyvíjeného v rámci řešení projektu GA ČR č. 105/03/1417 v případě, že je statistickým testem zjištěna jiná než normální distribuce, programově provedena výše popsaná kvantilová grafická transformace vstupních údajů do normovaného normálního rozdělení (dále NNR), následně se provede interpolace vybranou interpolační metodou a hodnoty gridu pak budou programově zpětně transformovány opět pomocí kumulovaných četností (empirické distribuční funkce).



Legenda:

Frequency – četnost,
Normal Score – normovaná normální distribuce,
Cumulative Frequency – kumulativní četnost.

Obr. 2: Postup transformace hodnot do normální distribuce. Pro transformaci jsou použity kumulativní četnosti (vlevo dole) histogramu (vlevo nahoře). Příklad transformace hodnoty 0,2: 1. zjištění kumulativní četnosti pro hodnotu 0,2, 2. odečtení odpovídající hodnoty distribuční funkce normovaného normálního rozdělení (vpravo dole) a odpovídající hodnoty (-0,949) (Deutsch 2002)

Figure 2: Procedure for transforming values to normal score values. The histograms are shown at the top of the figure. The cumulative distributions, shown at the bottom, are used for transformation. To transform 0.2 value: 1. read the cumulative frequency corresponding to the 0.2 value, 2. go to the same cumulative frequency on the normal distribution and read the normal score value (-0.949) (Deutsch 2002)

Tento proces je pro atribut obsah popela A^d vybrané části dubňanské sloje jihomoravského lignitového revíru znázorněn na obr. 3 a pro jednu vybranou hodnotu obsahu popela je transformace tam – do NNR demonstrována v tabulce 1 a transformace zpět – z NNR v tabulce 2.

Tabulka 1: Příklad transformace - pro hodnotu 50 % A^d do NNR

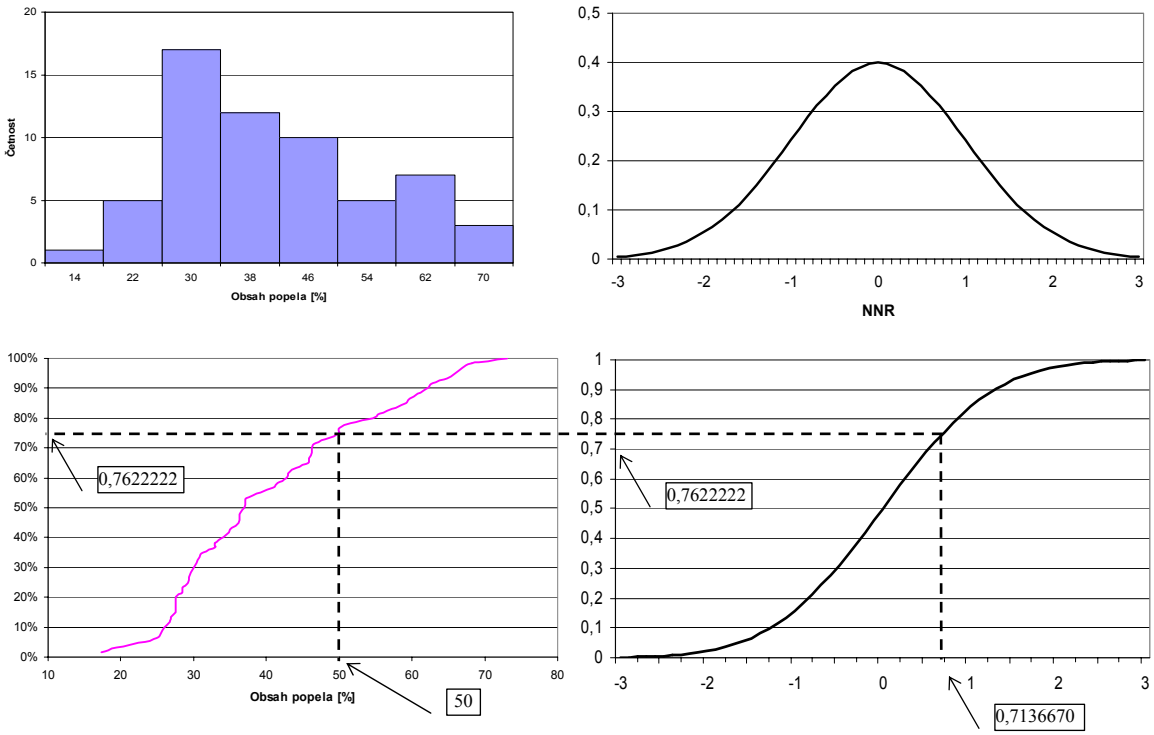
Table 1: Example of transforming of ash content – for value 50% A^d to standardized normal distribution (SND/NNR)

	Ve třídě od [%]; [odpovídající empirická kumulativní četnost]	*Ve třídě do* [%]; [odpovídající empirická kumulativní četnost]	*Transformovaná hodnota* 50 [%]; [odpovídající empirická kumulativní četnost]	*Transformovaná hodnota (NNR)*
Transformovaná hodnota 50 %	49.89; 0.75	50.04; 0.7666667	50; 0,7622222 (lineární interpolace)	**0,7136670**

Tabulka 2: Příklad transformace obsahu popela - pro hodnotu 50 % A^d z NNR

Table 2: Example of transforming of ash content – for value 50 % A^d from SND

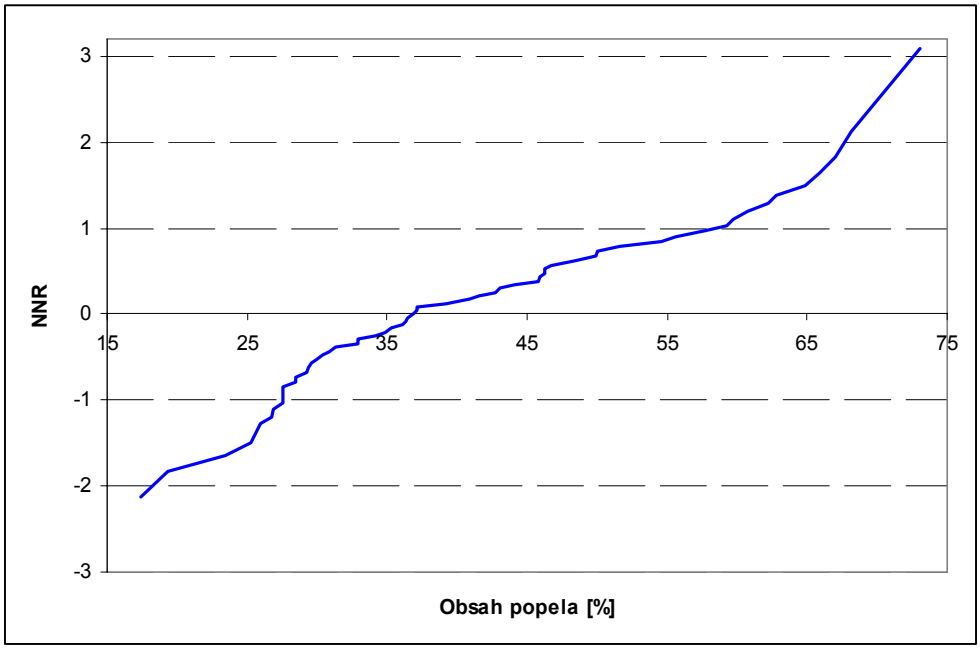
	Distribuční funkce NNR	*Ve třídě od* [empirická kumulativní četnost]; [odpovídající %]	*Ve třídě do* [empirická kumulativní četnost]; [odpovídající %]	*Zpětně transformovaná hodnota [%]*
Transformovaná hodnota (NNR) 0,7136670	0,7136670	0.6744897; 49,89	0.7279133; 50.04	**50,0000005** (lineární interpolace)



Obr. 3: Postup transformace obsahu popela části dubňanské sloje JLR do normální distribuce a zpět

Figure 3: Procedure for transforming ash content of apart of Dubňany coal seam in South Moravia lignite coalfield to normal distribution and back

Tato grafická transformace je v podstatě přiřazení odpovídajících kvantilů dvou typů distribucí. Na obr. 3 kvantil $q_{0,7666667}$ původních hodnot – obsahu popela (na obr. vlevo) odpovídá kvantilu $q'_{0,7136670}$ NNR (na obr. vpravo). Kvantilová závislost mezi vzorky obsahu popela vybrané části sloje (původní hodnoty) a odpovídajícími hodnotami NNR je graficky zobrazena na obr. 4.



Obr. 4: Grafická závislost mezi obsahem popela (původní hodnoty) a odpovídajícími hodnotami NNR ze všech vstupních údajů

Figure 4: Graphic dependence between ash content (original values) and corresponding SND values from all input data

Technický popis realizace transformace

Aplikace byla vyvíjena pomocí programovacího jazyka Microsoft Visual Basic 6.0 s využitím programových objektů programu Golden Surfer 8 a programu Microsoft Excel 2003.

Vstupní data o vrtech ve formátu MS Excel obsahují tyto základní údaje – lokalizaci vrtu v prostoru pomocí souřadnic souřadného systému JTSK (převedených do kartézské souřadné soustavy), název vrtu a hodnoty sledovaného ložiskového atributu (například obsahu popela). Vstupní údaje jsou seřazeny vzestupně podle zpracovávaného atributu a následně jsou dopočteny další hodnoty – kumulativní četnost, relativní četnost a hodnota inverzní funkce k distribuční funkci NNR vypočítaná s pomocí funkce MS Excel NORMINV. V tabulce 3 je ukázka takto připravených vstupních údajů.

Tabulka 3: Ukázka doplněných vstupních dat

Table 3: Specimen of prepared input data

X	Y	Nazev	A^d	Kumul	Relat	NNR
-570794	-1205218	BP14	17.34	1	0.016667	-2.12805
-571964	-1205273	BP30	19.37	2	0.033333	-1.83391
-572959	-1204477	HB26	23.39	3	0.050000	-1.64485
…	…	…	…	…	…	…
-580679	-1204208	HB75	66.97	58	0.966667	1.833915
-581762	-1208451	HB128	68.20	59	0.983333	2.128045
-582528	-1207836	HB127	73.11	60	0.999000	3.090232

Na základě těchto vstupních údajů se vybranou interpolační metodou stanoví hodnoty sledovaného ložiskového atributu v pravidelné síti bodů – tzv. gridu.

Pro usnadnění výběru vhodné interpolační metody se v IPSHUL využívá modul tzv. bumerangové metody (Cross Validation), kdy se pro bod se stanovenou hodnotou provede výpočet lokálního odhadu z ostatních hodnot. Výsledkem je vypočtená hodnota v místě, kde známe skutečnou hodnotu. Můžeme tedy

stanovit chybu odhadu v tomto místě a následně pro všechny body (Deutsch 2002, Staněk 1999). Testování se provádí pro konečnou množinu interpolačních metod s různými parametry.

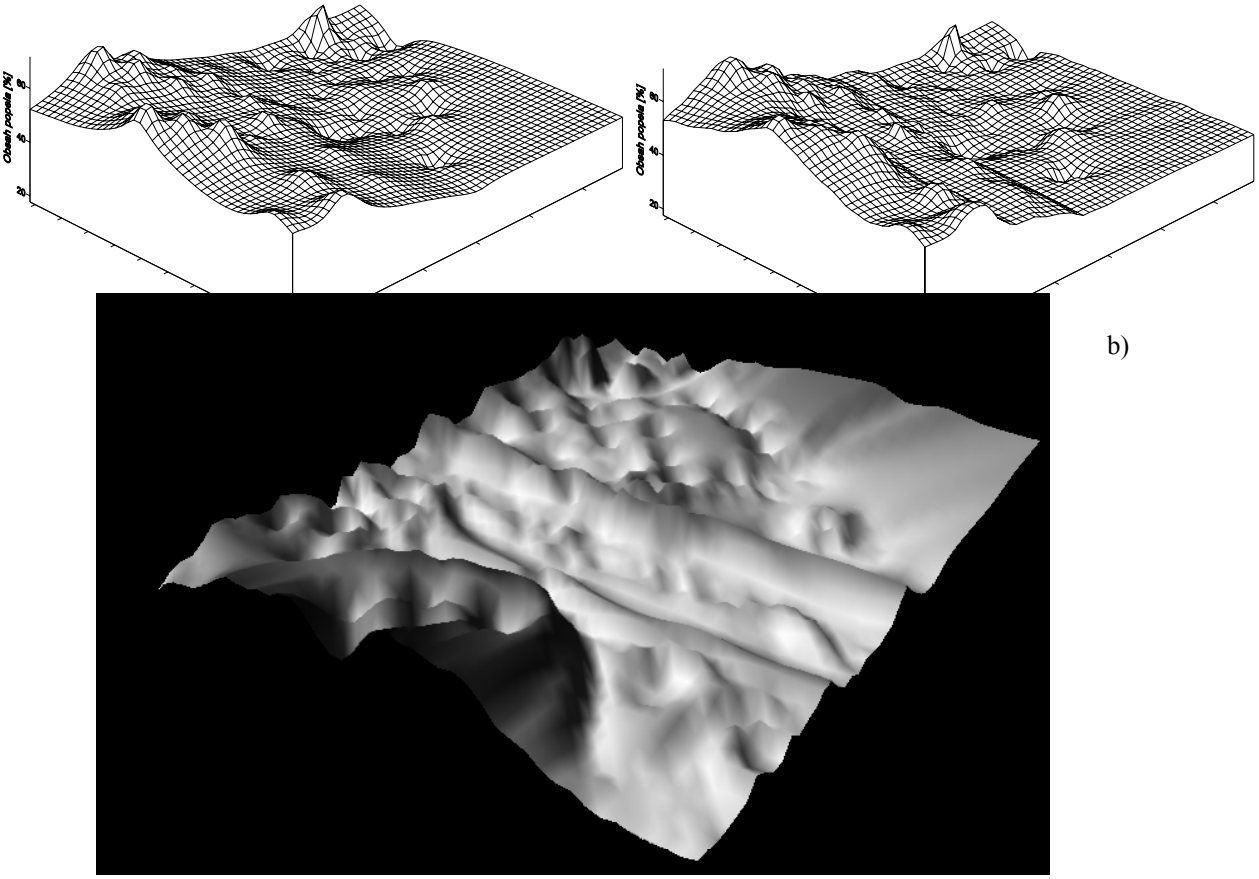
Po výběru jedné z možných interpolačních metod (viz tabulka 4) jsou v zadané oblasti vytvořeny pomocí objektu GridData (Surfer) dva gridy:

- Grid A - na základě původních hodnot (sloupec A^d – obsah popela – v tabulce 3),
- Grid B - na základě hodnot vypočtených pomocí funkce NORMINV (sloupec NNR v tabulce 3), přičemž jsou hodnoty takto vypočteného gridu následně pomocí zpětné transformace (viz tabulka 2) převedeny z NNR do původní distribuce.

Tabulka 4: Interpolační metody implementované v programu Surfer [6]

Table 4: Interpolation methods implemented within Surfer software [6]

Hodnota parametru SrfGridAlgorithm (GridData)	Popis
srfInverseDistance	Inverse Distance to a power
srfKriging	Kriging
srfMinCurvature	Minimum Curvate
srfShepards	Modified Shepard's Method
srfNaturalNeighbor	Natural Neighbor
srfNearestNeighbor	Nearest Neighbor
srfRegression	Polynomial Regression
srfRadialBasis	Radial Basis functions
srfTriangulation	Triangulation with Linear Interpolation

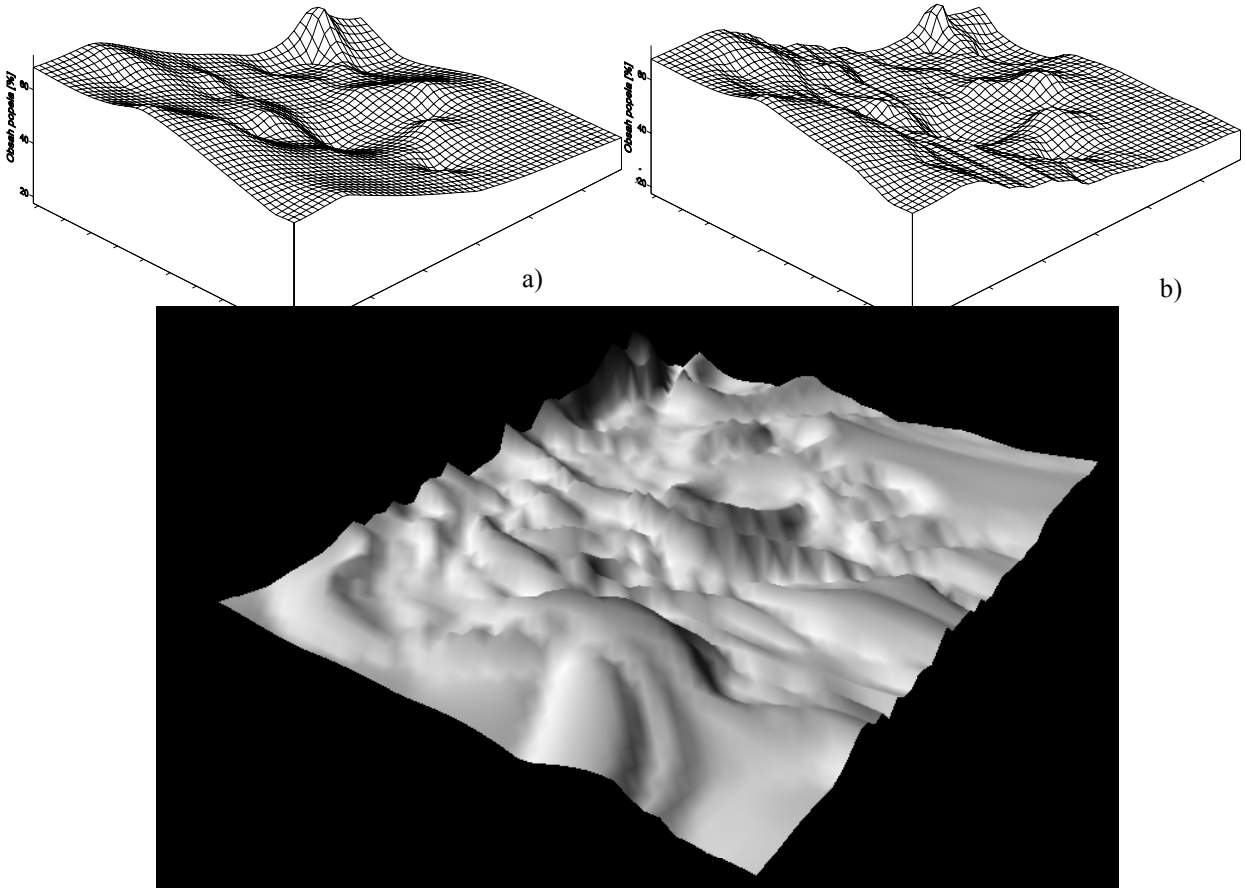


b)

Obr. 5: Metoda inverzních vzdáleností – a) grid vytvořený z původních hodnot, b) grid vytvořený z transformovaných hodnot, c) zobrazení gridu rozdílů absolutních hodnot gridů vytvořených z původních a transformovaných hodnot

Figure 5: Inverse Distance to a Power gridding method - a) grid formed from original values, b) grid formed from transformed values, c) grid of differences of absolute values between grids formed from original values and of transformed values

Zajímavé je vizuální porovnání gridů A a B vytvořených stejnou interpolační metodou z původních, resp. transformovaných hodnot. Hodnoty odpovídajících vybraných gridů obsahu popela interpolovaných metodou inverzních vzdáleností z původních hodnot a hodnot převedených do NNR ukazuje prostorové zobrazení na obr. 5a a 5b. Absolutní hodnoty rozdílů těchto gridů jsou názorně zobrazeny na obr. 5c. Hodnoty odpovídajících vybraných gridů obsahu popela interpolovaných metodou kriging z původních hodnot a hodnot převedených do NNR ukazuje prostorové zobrazení na obr. 6a a 6b. Absolutní hodnoty rozdílů těchto gridů jsou názorně zobrazeny na obr. 6c.



Obr. 6: Metoda krigování – a) grid vytvořený z původních hodnot, b) grid vytvořený z transformovaných hodnot, c) zobrazení gridu rozdílů absolutních hodnot gridů vytvořených z původních a transformovaných hodnot

Figure 6: Kriging method – a) grid formed from original values, b) grid formed from transformed values, c) grid of differences of absolute values between grids formed from original and transformed values

Závěr

Empirické distribuce většiny veličin popisujících geologická tělesa nevyhovují běžně uvažovanému normálnímu rozdělení, ale mají distribuci asymetrickou. V článku je teoreticky popsáno, že pouze na základě správného popisu statistické distribuce zpracovávaných ložiskových údajů lze odhadnout statistické charakteristiky a realizovat následné zpracování. Jedním z možných postupů pro transformaci údajů do normální distribuce, která je, jak známo, základní podmínkou použití mnoha statistických a numerických postupů, je právě kvantilová grafická transformace vstupních údajů do normovaného normálního rozdělení.

V článku je prezentováno programové řešení tohoto procesu při modelování uhelné sloje v Interaktivním programovém systému pro aplikaci moderních metod hodnocení uhelných ložisek a jejich dílčích částí v komplikovaných podmínkách (vyvíjeném na IGI HGF VŠB-TU Ostrava v rámci řešení projektu GA ČR č. 105/03/1417). Na příkladu tvorby gridu obsahu popela uhelné sloje jsou demonstrovány rozdíly interpolovaných hodnot s použitím kvantilové grafické transformace a bez jejího použití v případě, že je zjištěna

jiná než normální distribuce vstupních údajů. V grafických ukázkách je názorně demonstrováno, jaké nepřesnosti vznikají bez použití této transformace při modelování uhelné sloje.

Poděkování

Autor děkuje Grantové agentuře České republiky za podporu prací prezentovaných v tomto článku projektem reg. č. 105/03/1417.

Acknowledgement

The authors would like to thank to Grant Agency of ČR (GA ČR) for support of their research presented in this paper within project No 105/03/1417.

Literatura

[1] Deutsch, C., V., Journel, A., G.: GSLIB – Geostatistical Software Library and User's Guide. Second Edition. New York, *Oxford University Press, Oxford, 1998, 369 s*.

[2] Deutsch, C., V.: Geostatistical Reservoir modeling. Oxford, *Oxford university press, 2002, 376 s*.

[3] Vizi, L., Timčák, G., M.: Význam štúdia lognormálneho rozdelenia v geológii a baníctve. In *Sb. věd. prací VŠB-TU Ostrava, řada hornicko-geologická, Ostrava, 1, 2002, s. 29-39*.

[4] Wellmer, F., W.: Statistical Evaluations in Exploration for Mineral Deposits. Berlin, *Springer, 1998, 379 s.*

[5] Staněk, F.: Vliv výběru interpolační metody na přesnost výpočtu zásob uhelného ložiska. *Věstník Českého geologického ústavu 74, 2, 1999d, s. 211-222.*

[6] Surfer 8 – User's Guide. *Golden (California, USA) : Golden Software, Inc., 2002, 640 s.*

Summary

One of main tasks of statistical data analysis is to study the nature of statistical distribution. We need to describe a data distribution correctly to be able to estimate statistical parameters and to realise subsequent data processing. Use of normal distribution $N(U;\mu,\sigma^2)$ is often not relevant. Empirical distributions of quantities describing geological bodies are typically asymmetric (predominantly positively skewed). Many mathematical procedures can be applied only on normal distributed data. Without a correct approach to „skewed" data it is impossible to make for instance geostatistical analysis and assessments, because the best linear assessment is the one obtained from experimental values driven by Gauss distribution (see Vizi, Timčák, 2002).

The question is what error would occur when modeling a raw material deposit (i.e. coal deposit) with non-transformed input data.

For description of asymmetrically distributed quantities describing geological bodies we can apply various transformation methods, e.g. general logarithmic method $v = ln(a+bu)$ leading to the frequently used general lognormal distribution $LN(U: \mu, \sigma^2, a, b)$. Its constants can be determined by minimization of skewness and kurtosis. This process is demonstrated schematically in Figure 1. A skewed distribution is indicated in Figure 1a with the most frequent mode value h, median value m, and arithmetic mean value x. Figure 1b indicates a frequency distribution of identical values after their logarithmic transformation. The peak of frequency curve is now represented by a median. If the distribution is lognormal, the median m is identical with logarithm of geometric mean γ. By β^2 we indicate the variance of logarithmized values.

The lognormal distribution is only one among all the types of potential data distributions. There exists a way how to transform an arbitrary distribution into normal distribution and vice versa - application of so-called quantile (graphic) transformation of input data.

If the input data set does not comply with the normal distribution, it can be transformed by means of the quantil transformation into a data set with a standardized normal distribution (SND) with a zero mean value and a standard deviation of value 1. This process is illustrated in Figure 2. Such transformation can be realized e.g.

with use of program *nscore* in geostatistical toolbox GSLIB (see Deutsch, Journel, 1998). This transformation is bilateral if the histogram does not contain so-called „spikes"- interpopulation constant values. These values have to be eliminated before the transformation takes place (see Deutsch, 2002).

In case of modeling of coal seam the values of individual deposit attributes are interpolated from irregular deployed survey points into a regular grid. Because linear mathematical procedures are used, their input data should be normally distributed.

That is why the quantile transformation of non-normally distributed input data into the normally distributed ones take place in program solution of Interactive software system for application of modern methods of evaluation of coal deposits and their individual parts under complicated conditions (IPSHUL). Subsequently interpolation with use of a selected interpolation method is performed and the grid values are transformed back, again by means of cumulated frequencies (empirical distribution functions).

This process is represented in Figure 3 for the attribute of ash content A^d of selected part of Dubňany coal seam in South Moravia lignite coalfield. Transformation of a single selected ash content value the „hither" into SND is demonstrated in Table 1 and the back transformation is demonstrated in Table 2. This quantile transformation is in fact an assigning of corresponding quantiles of two distribution types. In Figure 3 the quantile $q_{0.7666667}$ of original values – ash content (in the left figure) corresponds to the quantile $q'_{0.7136670}$ SND (in the right figure). The quantile dependence between samples of ash content from a chosen coal seam part (original values) and the corresponding SND values is graphically illustrated in Figure 4.

The application was developed by means of programming language Microsoft Visual Basic 6.0 with use of program objects of Golden Surfer 8 software and of Microsoft Excel 2003 software. The input data of survey holes in MS Excel format contain spatial location of drill hole by means of co-ordinates of JTSK coordination system (transformed into Cartesian co-ordination system), name of survey hole, and values of observed deposit attribute (e.g. ash content). The input data are arranged in ascending order according to the processed attribute (e.g. ash content) and subsequently other values are calculated - cumulative frequency, relative frequency, and inverse function value of SND distribution function using function NORMINV of MS Excel. In Table 3 sample input data prepared in this way are shown.

Based on such input data the values of observed deposit attribute are determined by a selected interpolation method in a regular grid. To make the selection of proper interpolation method in IPSHUL easier a module of so-called cross validation method is used. For each point with known value a local estimate from the other known values is calculated. The result is an estimate of the value in the point where the accurate value is also known. For all points with known values it is possible to compute the error of estimate (Deutsch, 2002; Staněk, 1999). Testing is applied on a finite set of interpolation methods with various parameters.

After selection of one interpolation methods (see Table 4) the following two grids are formed in a considered area using GridData (Surfer) object:

- Grid A – based on original values (column A^d – ash content - in Table 3),
- Grid B – based on values calculated by means of NORMINV function (column NNR/SND in Table 3). The values of the grid calculated in this way are subsequently transformed from SND to original distribution by means of the back transformation (see Table 2).

Let us make a visual assessment of grids A and B formed with use of the same interpolation method from original and transformed values, respectively. The values of corresponding grids of ash content interpolated by method of inverse distances from original values and from values transformed into SND are depicted in Figure 5a and Figure 5b, respectively. Absolute values of differences between these grids are depicted in Figure 5c. The values of corresponding selected grids of ash content interpolated with use of kriging method from original values and from values transformed into SND are shown in Figure 6a and Figure 6b, respectively. The absolute values of differences between these grids are depicted in Figure 6c.

Empirical distributions of most quantities describing geological bodies do not comply with usually considered normal distribution, but they are asymmetric. We show that only based on correct description of statistical distribution of processed deposit data it is possible to estimate statistical characteristics and to realize a subsequent data processing. One of possible procedures for transforming data to normal distribution, which, as it is well-known, is a principal condition for applying many statistical and numerical procedures, is the quantile (graphic) transformation of input data to standard normal distribution and back to original distribution.

In the paper a program solution of this process is presented when modeling a coal seam within Interactive program system for application of modern methods of evaluation of coal deposits and their parts under complicated conditions (developed by IGI HGF VŠB-TU Ostrava within frame of solving GA ČR project N^{o} 105/03/1417). Differences between values obtained with use of quantile transformation and values obtained without this transformation are shown on example of forming grid of ash content. It is shown graphically, which inexactitudes would arise without application of this transformation method when modeling a coal seam.

Recenzenti: Prof. Ing. Gejza Timčák, Ph.D., FBERG, Košice,
Doc. RNDr. Jarmila Doležalová, CSc., VŠB-TU Ostrava.