Ing. Lubomíra Mejstříková, CSc., VÚHU, a.s., Most

Využití statistických metod v praxi

Ausnutzung der statistischen Methoden in Praxis

Die Datei von fünf vorgelegten Beiträgen schildert die Anwendungsmöglichkeiten von ausgewählten statistischen Methoden in Praxis. Auf den angeführten Beispielen sind die statistische Bewertung eines Komplexes von eindimensionalen Dateien, Bildung von linearen und unlinearen Modellen, Ausnutzung Streuungsanalysetechnik und mehrdimensionale statistische Analyse dargestellt.

Durch die Methode von eindimensionalen Dateien ist das statistische Testen des Ergebniskomplexes der Festigkeitsbestimmung im reinen Gesteinedruck durchgeführt, durch die Bildung von linearen Regressionsmodell für eine einheitliche Klassifikation der Sedimente (JKS) wurde die Beziehung zwischen den Ergebnissen von Laborproben erklärt, die die Gesteinebeschaffenheiten darstellen, und dem Index JKS. Die Demonstration der Bildung von unlinearen regressiven Modellen wurde einmal auf dem Modell für Berechnung des Filtrationskoeffizienten durchgeführt, der aus physikalischen Erwägungen ausgegangen ist, und einmal wurden die Abhängigkeit zwischen dem Beryliumgehalt in Kohle und der Asche in Kohle durch die Bildung eines empirischen Modells erfasst. Durch die Streuungsanalyse wurde der Einfluß der Temperatur auf längliche Änderung durch Brennen von Ziegelrohstoff untersucht und weiter wurde das Testen des Einflusses im Labor und der Methode für die Bestimmung des Schwefelgehaltes in Kohle durchgeführt. Als ein praktisches Beispiel der Ausnutzung der mehrräumigen statistischen Analyse ist dem Leser die Bewertung von keramischen Rohstoffen vorgelegt.

Utilization of statistical methods in practice

The set of five submitted articles shows the possibilities of technically directed research workers to apply the selected statistical methods in practice. At solved examples it is executed the statistical evaluation of the set of one-dimensional data, formation of linear and non-linear models, further the utilization of technique of scattering analysis and more-dimensional statistical analyses.

By the method of one-dimensional data it is executed the statistical testing of the result set, the statement of unconfined compression strength of rocks, formation of linear regression model for uniform sediment classification (JKS) it was expressed the relation between the results of laboratory tests which represent the rock properties, and by JKS index. The demonstration of the formation of non-linear regression models was executed partly on the model for calculation of filtration coefficient which came out of the physical considerations, and partly it was determined the dependence between the beryllium contents in coal and ash in coal by the formation of empirical model. By the scattering analysis it was tested the temperature influence to the length change by the burning of brick raw materials and further it was executed the testing of laboratory influence and method for determination of sulphur contents in coal. It is submitted the evaluation of ceramic raw materials to reader as a practical example of utilization of more-dimensional statistical analysis.

Использование статистических методов на практике

Собрание пяти предтавляемых статей показывает возможности технически оринтированных научно-исследовательских работников по практическому применению выбранных статистических методов. На примерах решения проводится статистическая оценка совокупности одноразмерных данных, создание линейных и нелинейных моделей, дальше использование техники анализа рассеяния и многоразмерного статистического анализа.

При помощи метода одноразмерных данных осуществлено статистическое испытание совокупности результатов установления прочности пород при простом сжатии, созданием линейной регрессивной модели для Единой

классификации седиментов /ЕКС/ было выражено соотношение между результатами лабораторных испытаний, репрезентирующих свойства пород, и индексом ЕКС. Демонстрирование образования нелинейных регрессивных моделей проводилось с одной стороны на модели расчета коэффициента фильтрации, который исходил из физических соображений, с другой стороны устанавливалась зависимость между содержанием бериллия в угле и золой за счет создания эмпирической модели. Помощью анализа рассеяния испытывалось влияние температуры на изменение длины при сжиганию кирпичного сырья, а также проводилось испытание влияния лаборатории и метода на установление содержания серы в угле. В качестве практического образца применения многоразмерного статистического анализа читателю предлагается оценка керамического сырья.

Využití statistických metod v praxi

Soubor pěti předložených článků ukazuje možnosti technicky zaměřených výzkumných pracovníků aplikovat vybrané statistické metody v praxi. Na řešených příkladech je provedeno statistické hodnocení souboru jednorozměrných dat, tvorba lineárních a nelineárních modelů, dále využití techniky analýzy rozptylu a vícerozměrné statistické analýzy.

Metodou jednorozměrných dat je provedeno statistické testování souboru výsledků stanovení pevnosti v prostém tlaku hornin, tvorbou lineárního regresního modelu pro jednotnou klasifikaci sedimentů (JKS) byl vyjádřen vztah mezi výsledky laboratorních zkoušek, které reprezentují vlastnosti hornin, a indexem JKS. Demonstrace tvorby nelineárních regresních modelů byla provedena jednak na modelu pro výpočet koeficientu filtrace, který vycházel z fyzikálních úvah, a jednak byla zjišťována závislost mezi obsahem beryllia v uhlí a popelem v uhlí tvorbou empirického modelu. Analýzou rozptylu byl testován vliv teploty na délkovou změnu pálením cihlářské suroviny a dále bylo provedeno testování vlivu laboratoře a metody na stanovení obsahu síry v uhlí. Jako praktická ukázka využití vícerozměrné statistické analýzy je čtenáři předloženo hodnocení keramických surovin.

Obsah

Úvod

I. Statistické zpracování jednorozměrných dat

II. Lineární regresní modely

III. Nelineární regresní modely

IV. Analýza rozptylu

V. Vícerozměrná statistická analýza

ÚVOD

Aplikace statistických metod do klasických i nových technických oborů patří spíše mezi mladé směry vědních disciplin. Vede ke vzniku oborů jako chemometrie, biometrie, psychometrie, ekonometrie apod. Statistická analýza nabývá stále většího významu a stává se jedním ze základních přístupů v řadě přírodovědných, technických a sociálních věd.

Předložený soubor pěti článků se zabývá využitím statistických metod pomocí počítačového software při zhodnocení výsledků laboratorních zkoušek hornin a uhlí severočeské a sokolovské pánve.

Články jsou rozděleny na dvě části. V první části jsou jen velmi stručně popsány základní statistické postupy, protože cílem nebylo ukázat čtenářům náročnost statistického testování, ale možnosti jeho použití v praxi. V druhé části článků je pak provedeno řešení vybraných příkladů.

Problematika ***jednorozměrných dat*** je diskutována ***v prvním článku***. Cílem analýzy jednorozměrných dat je určit povahu dat z hlediska symetrie, homogenity a stability, odhadnout jak je analyzovat, zda lze datům věřit a podobně. Na základě této diagnostiky je teprve možné rozhodnout o metodě výpočtu odhadu střední hodnoty, intervalu spolehlivosti, směrodatné odchylky. V řešeném příkladu prvního článku testování souboru výsledků stanovení pevnosti v prostém tlaku hornin je diskutováno, jakých chyb se lze dopustit, není-li dodržen uvedený postup statistické analýzy.

V druhém článku je demonstrován postup tvorby ***lineárního regresního modelu***. Vytvořením lineárního regresního modelu je objasňován vztah mezi výstupní, závisle proměnnou veličinou a vstupními nezávisle proměnnými, veličinami v řadě technických oborů. Řešený příklad z praxe se zabývá hledáním vztahu mezi výsledky laboratorních zkoušek hornin a indexem jednotné klasifikace zemin JKS.

Třetí článek se zabývá tvorbou ***nelineárních regresních modelů***, pomocí nichž lze prakticky řešit řadu technických a přírodovědných úloh, např. konstrukci kalibračních modelů, vyjádření základních fyzikálně-chemických zákonitostí, tvorbu empirických modelů apod. V řešených příkladech je předložen čtenáři jednak nelineární regresní model, vycházející z fyzikálních úvah, a jednak model empirický, který objasňuje vztah mezi obsahem popela v uhlí a obsahem beryllia v popelu uhlí.

Čtvrtý článek popisuje možnosti využití ***analýzy rozptylu*** v praxi. Analýza rozptylu, tzv. ANOVA, je technika, která umožňuje posouzení významnosti jednotlivých zdrojů variability v datech. Pomocí ANOVY lze např. určit vliv typu přístroje, lidského faktoru a obsluhy na výsledek měření. Tato technika je dále vhodná při zpracování mezilaboratorních experimentů a určení významnosti rozdílů mezi laboratořemi na výsledek analýz apod. Jako ukázka využití techniky ANOVY je čtenáři předloženo posouzení vlivu teploty na délkovou změnu pálením cihlářské suroviny a testování vlivu laboratoře a metody na stanovení obsahu síry v uhlí.

V pátém článku jsou diskutovány některé postupy ***vícerozměrné statistické analýzy***. V praxi se vícerozměrné statistické metody používají pro řešení problémů v situaci, kdy u souboru objektů je sledována více než jedna proměnná. S touto situací se lze dnes setkat prakticky ve všech vědních oborech, přírodními a technickými vědami počínaje a společenskými vědami konče. Jako praktická ukázka využití vícerozměrné statistické analýzy je uvedeno hodnocení keramických surovin.

I. STATISTICKÉ ZPRACOVÁNÍ JEDNOROZMĚRNÝCH DAT

1. DŮLEŽITÉ POJMY

Data. Data je soubor naměřených či jinak získaných hodnot, zatížených chybou. Ze statistického hlediska jsou data považována za realizaci náhodné veličiny, popř. náhodný výběr. Data se dále rozlišují na ***jednorozměrná*** a ***vícerozměrná*** (jednorozměrný a vícerozměrný výběr). Jednorozměrná data jsou např. opakované měření jedné veličiny (pevnosti v tlaku, vlhkosti, zisku apod.) v různém čase nebo na různých místech. U vícerozměrných výběrů se provádí opakovaná měření několika veličin současně.

Náhodná veličina, náhodné rozdělení, náhodný výběr. Veličina, která může nabývat jakoukoliv hodnotu z určité množiny hodnot a s níž je spojeno nějaké rozdělení pravděpodobnosti, se nazývá ***náhodná veličina***. Náhodná veličina, která může nabývat pouze izolované hodnoty, se nazývá ***diskrétní***. Náhodná veličina, která může nabývat jakékoliv hodnoty z konečného nebo nekonečného intervalu, se nazývá ***spojitá***. Existuje nekonečné množství všech možných hodnot, které lze naměřit. Tyto všechny hodnoty je možné chápat jako ***náhodné rozdělení*** nebo statistický model náhodné veličiny. Každé náhodné rozdělení je definováno např. ***hustotou pravděpodobnosti, distribuční funkcí nebo kvantilovou funkcí.*** Tyto tři funkce jsou co do informace o rozdělení ekvivalentní, jedna se dá spočítat z druhé. Obecně může mít takové rozdělení složitý tvar, k jehož popisu je třeba velké množství parametrů. V praxi je k dispozici jen omezený počet možných hodnot, které nám vybrala náhoda, tedy jde o ***náhodný výběr.*** Protože není známo jak spravedlivě jsou tyto hodnoty „vybrány", provádí se rekonstrukce původního rozdělení, obvykle pomocí odhadů parametrů rozdělení jako jsou průměr a směrodatná odchylka normálního rozdělení apod.

Distribuční funkce (kumulativní hustota) F(x) je pravděpodobnost, že náhodná veličina z daného rozdělení bude menší než x. F je zdola omezeno nulou, shora jedničkou. Na ose x je pravděpodobnost, na ose y je kvantil daného rozdělení.

Kvantil je hodnota, odpovídající zvolené pravděpodobnosti na distribuční funkci. Kvantil je v podstatě každá možná hodnota měřené veličiny, k níž lze přiřadit jisté procento $100.\alpha$ (resp. pravděpodobnost α). Mluví se pak o α - kvantilu neboli o α-procentním kvantilu.

Hustota pravděpodobnosti f(x) je derivace distribuční funkce. Na první pohled dává lepší představu o rozdělení dat než distribuční funkce. Je z ní obvykle patrná nesymetrie (sešikmení) rozdělení nebo přítomnost více maxim (nehomogenita). Její hodnota na ose y však nemá význam pravděpodobnosti. V intervalu kolem maxima se bude vyskytovat více naměřených hodnot, než ve stejném intervalu jinde.

Statistika je jednak označení matematické disciplíny, jednak hodnota získaná jednoznačně z náhodných experimentálních dat neboli funkce náhodného výběru. Statistikou je aritmetický průměr, směrodatná odchylka, medián, nejmenší a největší naměřená hodnota, směrnice regresní přímky, ale také korelační matice.

Normální rozdělení (neboli Gaussovo) je nejčastěji definováno svou hustotou pravděpodobnosti. Má dva parametry – střední hodnotu a rozptyl. Odmocnina z rozptylu se nazývá směrodatná odchylka. Normální rozdělení je symetrické, data s normálním rozdělením mohou nabývat hodnot od $-\infty$ do $+\infty$. Již z této věty vyplývá, že používat

modelu normálního rozdělení pro reálná data je většinou teoreticky a často i prakticky nesprávné. Reálná data nemohou nabývat libovolné hodnoty. Hmotnost, délka aj. veličiny nemohou být záporné, koncentrace je omezena na interval 0-100%, podobná omezení má většina přírodních, technologických i ekonomických veličin. Přesto se předpoklad o normalitě dat používá. Za prvé proto, že matematické vlastnosti Gaussova rozdělení dovolují snadné odvození jeho důležitých vlastností a výpočetních metod (např. testy, metoda nejmenších čtverců, vlastnosti odhadů). Pro jiná rozdělení jsou takové výpočty složitější a někdy i nemožné. Za druhé se někdy data skutečně chovají tak, že jejich rozdíl od normality je zanedbatelný.

Statistický test má dát jednoznačnou odpověď na otázku na základě naměřených dat. Má obvykle formu hypotézy H_0, kterou test zamítne nebo příjme. Zamítnutí hypotézy H_0 znamená přijetí takzvané alternativní hypotézy H_A, která je negací H_0. Vzhledem k náhodné povaze dat nemůže být výsledek testu stoprocentní. Každý test se může zmýlit. Proto je u většiny testů nutné se rozhodnout jaké riziko omylu chceme připustit. Tímto rizikem je ***hladina významnosti***, obvykle označovaná jako α. Volba hladiny významnosti je v kompetenci uživatele. Obvykle se užívá hodnota $\alpha = 0.05$, tedy 5%. Použití jiné hodnoty by se mělo zdůvodnit a pak se musí dodržovat vždy.

2. STATISTICKÉ ZPRACOVÁNÍ JEDNOROZMĚRNÝCH DAT

2.1 Průzkumová analýza dat

Statistickému zpracování jednorozměrných výběrů předchází ***průzkumová (exploratorní) analýza***. Jejím účelem je odhalit zvláštnosti dat a ověřit předpoklady pro následné statistické zpracování. Při zpracování jednorozměrných výběrů, které pocházejí ze souborů o ne zcela známém rozdělení, je sledován pouze jeden znak, např. mechanicko-fyzikální vlastnost, chemický nebo mineralogický komponent hornin, uhlí nebo parametr charakterizující určité vlastnosti materiálů apod. Cílem statistického zpracování je z chování výběru usuzovat na chování celého souboru. Z různých typů výběru se v praxi uplatňují např. v laboratoři naměřené hodnoty, které jsou chápány jako realizace jisté náhodné veličiny. Reprezentativní náhodný výběr je charakterizován následujícími předpoklady, které tvoří základ statistických metod vyhodnocení výsledků měření:

a) Jednotlivé prvky výběru jsou vzájemně nezávislé.
b) Výběr je homogenní, tj. všechny prvky výběru pocházejí ze stejného rozdělení pravděpodobnosti s konstantním rozptylem.
c) Předpokládá se, že jde o normální rozdělení.
d) Všechny prvky souboru mají stejnou pravděpodobnost, že budou zařazeny do výběru.

2.1.1 Grafická průzkumová analýza

Ke komplexnímu posouzení zvláštností statistického chování dat se využívají grafické metody: kvantilový graf, rozptylový graf, graf polosum, symetrie, šikmosti a špičatosti, Q-Q graf, podmíněný Q-Q graf, graf hustoty pravděpodobnosti, krabicový graf s kvantily a kruhový graf.

Kvantilový graf umožňuje přehledně znázornit data a snadněji rozlišit tvar rozdělení, který může být symetrický, sešikmený k vyšším nebo nižším hodnotám. Dále

lze identifikovat lokální koncentrace dat, vybočující data. ***Diagram rozptýlení*** představuje jednorozměrnou projekci kvantilového grafu do osy x. Tento graf ukazuje na lokální koncentrace dat a indikuje i podezřelá a vybočující měření. ***Rozmítnutý diagram rozptýlení*** představuje rovněž projekci kvantilového grafu. Pro částečnou sumarizaci dat lze využít ***krabicového grafu***, který umožňuje znázornění robustního odhadu polohy mediánu, posouzení symetrie v okolí kvantilů, posouzení symetrie u konců rozdělení a identifikaci odlehlých dat. Obdobou krabicového grafu je ***vrubový krabicový graf***, který umožňuje i posouzení variability mediánu. Pro svoji jednoduchost a přehlednost se užívají krabicové grafy především k porovnání několika výběrů. Indikují dobře symetrii rozdělení a podezřelá měření. Snadné ověření symetrie umožňuje ***graf polosum*** a ***graf symetrie***. ***Q-Q graf*** pro normální rozdělení je jedním z nejpoužívanějších exploratorních grafů pro hodnocení normality.

Při statistickém zpracování se předpokládá, že jde o nezávislé, stejně rozdělené náhodné veličiny pocházející z normálního rozdělení. Rozsah výběru postačuje k určení dostatečně přesného odhadu parametrů polohy a rozptýlení. Pokud předpoklady o datech nejsou splněny, jejich analýza je značně složitá. Ověření všech předpokladů je možné při využití metod průzkumové analýzy a interaktivního přístupu na osobních počítačích.

2.1.2 Ověření základních předpokladů

Základním předpokladem kvalitních měření je vzájemná ***nezávislost jednotlivých výsledků***. Závislost měření je obvykle způsobena:

a) nestabilitou měřícího zařízení nebo změnou stavu měřícího zařízení
b) nekonstantností podmínek měření
c) zanedbáním faktorů, které významně ovlivňují výsledek měření, jako je objem vzorků, teplota, nečistota chemikálií
d) nesprávným, nenáhodným výběrem vzorků k měření

Pokud se uvedené faktory mění s časem, projeví se vznikem časové závislosti mezi prvky výběru, uspořádanými v časovém sledu. V případě skokových změn těchto faktorů vzniká heterogenní výběr. V obou případech se objeví zvýšený rozptyl oproti rozptylu homogenního výběru. Odhalení závislosti v datech vyžaduje obecně prověření celého procesu měření a sběru dat.

K základním předpokladům patří ***normalita*** výběrového rozdělení, neboť je na ní založena celá klasická analýza dat, testování vybočujících měření a testy nezávislosti prvků výběru. Existují dva základní typy testů normality, směrové testy a omnibus-testy. Testy jsou obecně méně citlivé na odchylky od normality než výše uvedené diagnostické grafy a navíc odchylka od normality může být mnohdy způsobena vybočujícími hodnotami.

Tam, kde se vyskytuje výrazná nestejnoměrnost měřených vlastností vzorků, kde se náhle mění podmínky experimentů a data obsahují vybočující měření, dochází k ***nehomogenitě*** naměřených dat. Pokud se v datech vyskytují odlehlé, vybočující hodnoty, lze je běžně identifikovat v grafech průzkumové analýzy. Vybočující měření silně zkreslují odhady parametrů, takže mohou zcela znehodnotit další statistickou analýzu. Na základě logické analýzy je třeba nejdříve zvážit, zda nejde o sešikmené rozdělení. Body, které se jeví vybočující pro symetrické rozdělení, mohou být pro sešikmená rozdělení naopak přijatelné. Pokud se jedná o vybočující pozorování, lze použít dvou možností. První možnost spočívá ve vyloučení vybočujících hodnot z další

analýzy, což nemusí být vždy vhodné řešení. Vybočující měření mohou být totiž výsledkem řídce se vyskytujících jevů a jejich vyloučením pak může dojít k úplné ztrátě informace. Druhá možnost spočívá v použití robustních metod. Tento postup nemusí být ale vždy korektní. Robustnost spočívá v přiblížení se k přijatému modelu bez ohledu na jeho platnost. Která měření vykazují evidentní hrubé chyby jako je selhání přístroje, špatný zápis dat apod. a která data jsou jen podezřelá, by měl rozhodnout experimentátor. Evidentní hrubé chyby je vhodné z další analýzy vyloučit, ale podezřelá měření je lépe ponechat. Robustními metodami se jejich vliv na odhady parametrů výrazně oslabí.

2.2 Statistické zpracování dat

Po předběžné analýze naměřených dat následuje další etapa statistické analýzy. Pro výběry malého rozsahu se vyčíslují charakteristiky přímo, u výběrů většího rozsahu se data nejdříve člení do tříd a pak se pracuje se skupinami. V praxi, např. v chemické disciplině, není soubor všech možných naměřených hodnot většinou známý. Statistická analýza se provádí na základě jeho reprezentativního vzorku, tzv. náhodného výběru. Pro reprezentativní náhodný výběr platí, že všechny prvky výběru se chápou jako náhodné veličiny, které se řídí stejným zákonem rozdělení, tj. výběr je homogenní a hodnoty zahrnuté do výběru jsou vybrány nezávisle na sobě.

Výběr je charakterizován informací o střední hodnotě a rozptýlení kolem střední hodnoty. Určit střední hodnotu sledované veličiny je prvním a zdánlivě jednoduchým úkolem, který lze s naměřenými daty udělat. ***Aritmetický průměr*** je nejpoužívanějším, a většině lidí také jediným známým odhadem střední hodnoty náhodné veličiny. Je to součet hodnot dělený jejich počtem. Málokdo si uvědomuje nedostatky takového odhadu, které mohou vést až k jeho nepoužitelnosti. Aritmetický průměr je totiž „správným" odhadem jen tehdy mají-li data normální rozdělení (Gaussovo). Lze tušit, že tento předpoklad je často falešný.

Doplňkem střední hodnoty a rozptýlení kolem střední hodnoty je pak informace o tvaru výběrového rozdělení. Statistické charakteristiky polohy, rozptýlení a tvaru výběru se nazývají výběrové charakteristiky. Z těchto charakteristik se pak usuzuje na statistické charakteristiky základního souboru.

Použití jednotlivých typů charakteristik je závislé na rozdělení základního souboru, ze kterého výběr pochází. Pokud platí předpoklady normality, nezávislosti a homogenity, vyčíslují se ***základní momentové charakteristiky***. Jsou-li v datech předpokládány i vybočující hodnoty, užívají se ***robustní odhady***, zejména kvantilové. Pokud bylo nalezeno jiné než předpokládané normální rozdělení, používají se ***maximálně věrohodné odhady.***

Základní charakteristikou polohy je ***výběrový průměr*** $\bar{x}$, který je zároveň maximálně věrohodným odhadem střední hodnoty pro normální rozdělení. Základní charakteristikou variability je ***výběrový rozptyl*** s^2, který je zároveň nevychýleným odhadem rozptylu pro normální rozdělení. Obě tyto charakteristiky se označují jako ***momentové*** a používají pro základní statistický popis výběrů pocházející z libovolného rozdělení.

Kvantilové a robustní charakteristiky jsou méně citlivé na vybočující hodnoty než momentové. Patří sem především ***medián,*** $\tilde{x}_{0,5}$. Jde vždy o 50%ní kvantil, kdy

polovina prvků leží pod a polovina prvků nad hodnotou mediánu. Medián je maximálně věrohodným odhadem polohy u Laplaceova (oboustranného exponenciálního rozdělení). Pro normální rozdělení však již není medián nejvhodnější. Pro případ rovnoměrného rozdělení je vydatným odhadem polohy ***polosuma*** $\hat{x}_p$, kdy součet minimálního a maximálního prvku ve výběru je vydělen dvěmi. Dále sem patří ***módus,*** $\hat{x}_M$, je definován jako lokální maximum na hustotě pravděpodobnosti. Jedním z nejefektivnějších a přitom jednoduchých robustních odhadů polohy je ***uřezaný průměr*** $\bar{x}$ ***(9),*** který využívá lineární kombinace pořádkových statistik.

Ze statistického hlediska mají bodové odhady malý význam, protože neříkají nic o tom kde leží skutečné hodnoty parametrů. Více informací poskytuje intervalový odhad, který určuje interval, v němž se bude se zadanou pravděpodobností nacházet skutečná hodnota daného parametru. Neznámý parametr je tedy odhadován nikoliv jednou, ale dvěma číselnými hodnotami, které tvoří meze tzv. *intervalu spolehlivosti* (konfidenčního intervalu). Interval spolehlivosti pokryje neznámý odhadovaný parametr daného rozdělení základního souboru s předem zvolenou, dostatečně velkou pravděpodobností (1-α), která se nazývá *koeficient spolehlivosti* (konfidenční koeficient, statistická jistota).

Pro intervaly spolehlivosti platí, že:

a) Čím je rozsah výběru větší, tím je interval spolehlivosti užší.
b) Čím je odhad přesnější a má menší rozptyl, tím je interval spolehlivosti užší.
c) Čím je větší statistická jistota (1-α), tím je interval spolehlivosti širší.

2.3 Matematická transformace dat

Pokud se na základě analýzy dat zjistí, že rozdělení výběru dat se příliš odlišuje od rozdělení normálního, vzniká problém, jak data vůbec vyhodnotit. V řadě případů lze nalézt vhodnou transformaci, která vede ke stabilizaci rozptylu, zesymetričtění rozdělení a někdy i k normalitě. Vychází z představy, že zpracovaná data jsou nelineární transformací normálně rozdělené náhodné veličiny x. Hledá se k nim pak inverzní transformace g(x). Nejčastěji používaná transformační funkce se nazývá ***Box-Coxova transformace***. Cílem Box-Coxovy transformace je nalézt takovou hodnotu, která zajistí maximální symetrii nebo lépe maximální normalitu dat.

3. PŘÍKLAD STATISTICKÉHO ZPRACOVÁNÍ SOUBORU JEDNOROZMĚRNÝCH DAT

3.1 Zadání:

U 26ti vzorků hornin, které reprezentují kvazihomogenní materiál a jejichž stav odpovídal podmínkám zkoušky, bylo provedeno stanovení pevnosti v prostém tlaku dle ČSN 72 1025 „Laboratorní stanovení smykové pevnosti zemin zkouškou v prostém tlaku". U souboru výsledků laboratorních zkoušek je požadováno určení bodových a intervalových odhadů pevnosti v prostém tlaku. Výsledky jsou uvedeny v MPa.

3.2 Řešení:

U dat byla nejprve provedena pomocí statistického softwaru Adstat 2.0 spojitá exploratorní analýza a ověření základních předpokladů. Vstupní data a vstupní podmínky jsou pro všechny výpočty stejné, proto jsou uvedeny pouze u číselného vstupu exploratorní analýzy.

I. EXPLORATORNÍ ANALÝZA

Číselný výstup ze statistického software

V číselném výsledkovém výstupu statistického software jsou základní statistické charakteristiky a parametry regresní přímky pro závislost charakteristiky šikmosti na čtverci kvantilu normovaného normálního rozdělení. Dále je zde tabulka kvantilů od 5 do 95% a hodnoty dalších významných kvantilů. Kvantily se používají při konstrukci diagnostických grafů.

Pro normální rozdělení je šikmost $g_1 = 0$ a špičatost $g_2 = 3$. Tvarové parametry, uvedené v číselném výstupu exploratorní analýzy (3) klasické odhady, pro šikmost a špičatost ukazují na exponenciální rozdělení souboru.

(1) VSTUPNÍ DATA A PODMÍNKY:

Počet dat	26
Hladina významnosti alfa	0.050

(2) VSTUPNÍ DATA:

1	1.7500E-01	8	9.4000E-01	15	1.6000E-01	22	2.3000E-01
2	1.1700E-01	9	3.7000E-01	16	2.9500E-01	23	1.8600E-01
3	1.8000E-01	10	3.3000E-01	17	1.4000E-01	24	5.7000E-01
4	2.7100E+00	11	3.1000E-01	18	2.7000E-01	25	1.5500E-01
5	1.2700E+00	12	1.8000E-01	19	2.6000E-01	26	5.1200E-01
6	1.1500E+00	13	2.2000E-01	20	2.6000E-01		
7	1.2500E+00	14	1.9000E-01	21	1.7800E-01		

(3) KLASICKÉ ODHADY PARAMETRŮ

Medián	2.6000E-01	Průměr	4.8492E-01
Rozptyl	3.2871E-01	Směrodatná odchylka	5.7333E-01
Špičatost	9.8557E+00	Šikmost	2.5807E+00

(4) KVANTILY A PÍSMENOVÉ HODNOTY:

Kvantilové míry:

Procento	Kvantil	Procento	Kvantil
5	1.4375E-01	10	1.5750E-01
15	1.7125E-01	20	1.7800E-01
25	1.8000E-01	30	1.8300E-01
35	1.8900E-01	40	2.2000E-01
45	2.3750E-01	50	2.6000E-01
55	2.6750E-01	60	2.9500E-01
65	3.1500E-01	70	3.5000E-01
75	4.7650E-01	80	5.7000E-01
85	9.9250E-01	90	1.2000E+00
95	1.2650E+00		

Písmenové hodnoty:

Kvantil	Písmeno	Pravděpodobnost	Spodní mez	Horní mez
Sedecil	D	0.0625	1.4844E-01	1.2587E+00
Oktil	E	0.1250	1.6187E-01	1.1237E+00
Kvartil	F	0.2500	1.8000E-01	4.7650E-01
Medián	M	0.5000	2.6000E-01	2.6000E-01

(5) KVANTILOVÉ MÍRY:

Kvantil	F (0.25)	E (0.125)	D(0.0625)
Rozsah	2.9650E-01	9.6187E-01	1.1103E+00
Polosuma	3.2825E-01	6.4281E-01	7.0359E-01
Délka konců	0.0000E+00	1.1768E+00	1.3203E+00
Šikmost	1.3506E-01	-1.6448E-01	-9.2629E-02
PseudoSigma	2.1996E-01	4.1821E-01	3.6285E-01

Diskuse ke grafickému výstupu exploratorní analýzy

Diagnostické grafy exploratorní analýzy jsou znázorněny na obr 1 – 3.

a) **Bodové a krabicové grafy** na obr. 1 signalizují jednak přítomnost pěti vybočujících hodnot směrem k vyšším hodnotám a dále ukazují na asymetrii rozdělení.
b) **Kvantilový graf** na obr. 1 ukazuje odchylky od normálního rozdělení a signalizuje jeden vybočující bod na straně vyšších hodnot.
c) **Graf symetrie** na obr. 1 ukazuje na asymetrii dat. Pro ideálně symetrická rozdělení

by body měly ležet na horizontální přímce s úsekem rovným mediánu. Nelineární průběh indikuje nehomogenitu výběru. Byl zjištěn jeden vybočující bod směrem k vyšším hodnotám.

d) **Graf polosum** na obr. 1 rovněž signalizuje asymetrii, protože data mají výrazný trend. Graf signalizuje přítomnost dvou vybočujících hodnot.
e) **Graf šikmosti** na obr. 2 je zobrazen pro indikaci šikmosti. Za předpokladu symetrie vyjde ideální horizontální přímka.Graf ukazuje na sešikmení k nižším hodnotám. Směrnice určuje odhad parametru šikmosti.
f) **Graf** na obr. 2 indikuje **špičatost**. Pro normální rozdělení je ideální horizontální přímka.
g) **Q-Q graf** je znázorněn na na obr. 2. Tímto grafem je ověřována shoda rozdělení výběru s normálním rozdělením. V ideálním případě normality leží jednotlivé body na přímce. Čarou je vyznačen teoretický průběh. Průběh testovaných dat naznačuje, že data nepocházejí z normálního rozdělení. Z toho plyne, že použití aritmetického průměru jako odhadu střední hodnoty, by bylo nesprávné. Graf indikuje přítomnost jedné vybočující hodnoty směrem k vyšším hodnotám.
h) **Podmíněný Q-Q graf** na obr. 2 je znázorněn rovněž pro ověření shody rozdělení výběru s normálním rozdělením. V ideálním případě normality leží body na přímce. Průběh dat indikuje, data nepocházejí z normálního rozdělení. Graf indikuje přítomnost jedné vybočující hodnoty směrem k vyšším hodnotám.
i) **Graf rozptýlení s kvantily** na obr. 3 umožňuje orientační posouzení symetrie, resp. přítomnosti odlehlých bodů podle symetrie a umístění obdélníků. Tento graf signalizuje asymetrii testovaného výběru.
j) **Graf hustoty pravděpodobnosti** na obr. 3 nabízí porovnání hustoty pravděpodobnosti rozdělení, z něhož pocházejí naše data (čárkovaná čára), s hustotou pravděpodobnosti normálního rozdělení (plná čára) jako vhodné vizuální posouzení normality dat. Z grafu je zřejmé, že testovaný výběr nepochází z normálního rozdělení.
k) **Kruhový graf** na obr. 3 slouží pro indikaci normality, resp. symetrie testovaného rozdělení. V ideálním případě má tvar kružnice se symetrickým svislým výsekem ve tvaru V. Testovaný soubor dat pevností v prostém tlaku vykazuje asymetrii rozdělení.

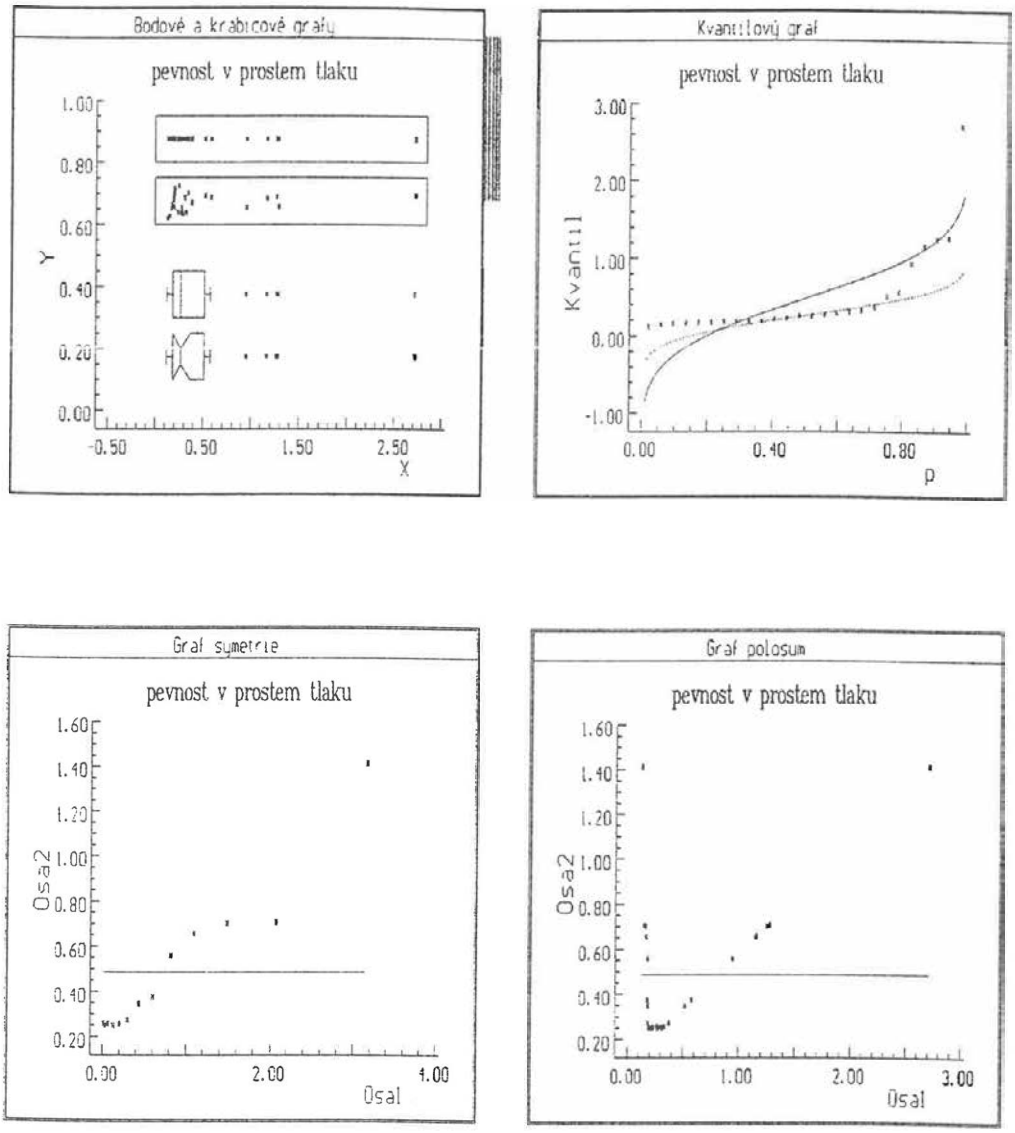
Závěr exploratorní analýzy:

Z průzkumové analýzy dat vyplývá, že data jeví asymetrii, nepocházejí z normálního rozdělení a v jednotlivých grafech byla zjištěna přítomnost vybočujících hodnot v počtu 1-5.

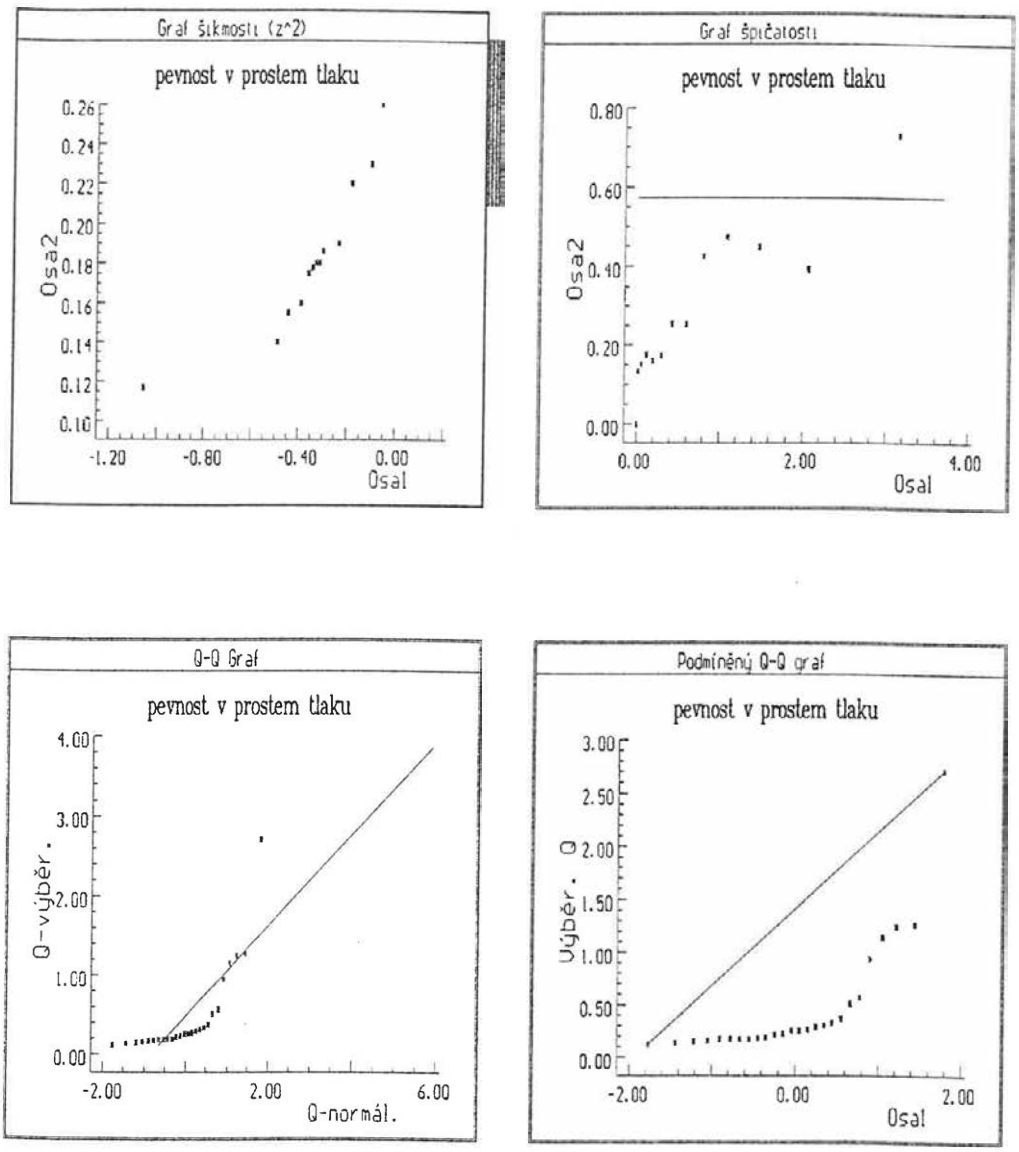
II. ZÁKLADNÍ PŘEDPOKLADY

Tato metoda obsahuje důležité diagnostiky pro ověřování předpokladů o datech. Ve výsledkovém souboru statistického software jsou po informaci o vstupním souboru a podmínkách uvedeny ve třech odstavcích základní diagnostiky.

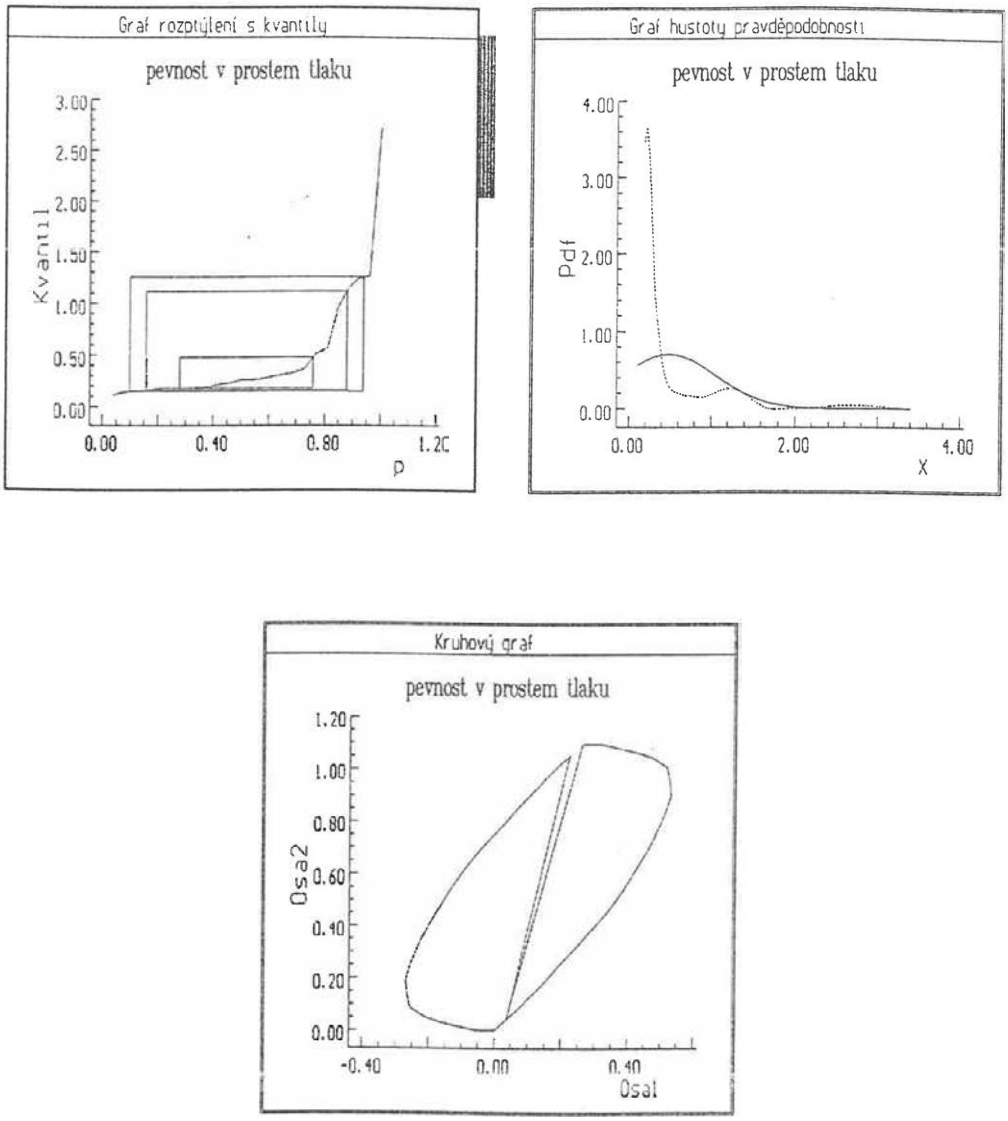
V prvním odstavci jsou uvedeny klasické odhady. Ve druhém odstavci je uveden závěr testu normality na základě porovnání vypočteného testovacího kritéria a tabulkového kvantilu. V testovaném případě byl předpoklad normality zamítnut. Ve třetím odstavci je provedeno ověření nezávislosti na základě srovnání testu autokorelace a tabulkového kvantilu. Testovaná data lze považovat za nezávislá. V případě, že jsou detekovány odlehlé body, jsou po jejich vyloučení ze souboru vypočteny ještě jednou klasické charakteristiky (4. odstavec) jako v prvním odstavci.



Obr. 1: Diagnostické grafy exploratorní analýzy



Obr. 2: Diagnostické grafy exploratorní analýzy



Obr. 3: Diagnostické grafy exploratorní analýzy

(1) TEST NORMALITY:

Tabulkový kvantil Chi^2(1-alfa,2)	5.9915E+00
Chi^2-statistika	1.3154E+02
Závěr	Předpoklad normality zamítnut
Vypočtená hladina významnosti	0.0000E+00

(2) TEST NEZÁVISLOSTI:

Tabulkový kvantil t(1-alfa/2,n+1)	2.0518E+00
Test autokorelace	1.1216E+00
Závěr	Předpoklad nezávislosti přijat
Vypočtená hladina významnosti	1.3595E-01

Předpoklad homogenity výběru:

Aritmetický průměr	4.8492E-01
Rozptyl	3.2871E-01
Směrodatná odchylka	5.7333E-01
Vnitřní meze	
Spodní mez	-5.2103E-01
Horní mez	1.2130E+00

(3) DETEKCE ODLEHLÝCH BODŮ:

Bod číslo 4 (horní)	2.7100E+00
Bod číslo 5 (horní)	1.2700E+00
Bod číslo 7 (horní)	1.2500E+00
Počet odlehlých bodů	3

Parametry s vynechanými odlehlými hodnotami:

Průměr	3.2079E-01	Rozptyl	6.5701E-02
Směrodatná odchylka	2.5632E-01	Šikmost	2.3135E+00
Špičatost	7.8612E+00		

Závěr z ověření základních předpokladů:

Závěr je shodný se závěrem exploratorní analýzy. Data nepocházejí z normálního rozdělení, předpoklad normality byl tedy zamítnut. Předpoklad nezávislosti byl přijat. Dále byly identifikovány tři odlehlé body.

III. MOCNINNÁ TRANSFORMACE

Základní předpoklady společně s exploratorní analýzou jsou první kroky při analýze neznámých dat. Na základě jejich výsledků je pak třeba rozhodnout jakým způsobem postupovat při další analýze. Provedením exploratorní analýzy a ověření základních předpokladů bylo zjištěno, že data testovaného souboru nepochází z normálního rozdělení, dále data jeví asymetrii a rovněž byla zjištěna přítomnost odlehlých bodů. Následkem toho nelze použít průměr jako odhad střední hodnoty. Z těchto výsledků vyplývá, že bude nutné provést nelineární transformaci dat. Data byla zpracována symetrizující mocninnou transformací a normalizující Box-Coxovou transformací. Jejich účelem je odstranit asymetrii a přiblížit se k normalitě s využitím vybraných průzkumových grafů. Po každé transformaci je vypočten opravený průměr.

Ve výsledkovém souboru použitého statistického software jsou uvedeny základní statistické charakteristiky původních dat. Tyto údaje jsou v předloženém článku vynechány, protože jsou již dokumentovány ve výše uvedených číselných výstupech statistického zpracování. Dále jsou ve výstupu uvedeny optimální hodnoty exponentu pro prostou mocninnou i pro Box-Coxovu transformaci, minimalizující různé míry špičatosti a asymetrie. Pro tento exponent jsou určeny transformovaný průměr a rozptyl spolu s opraveným (retransformovaným) průměrem. V posledním odstavci je jako ukázka uvedeno několik hodnot po transformaci.

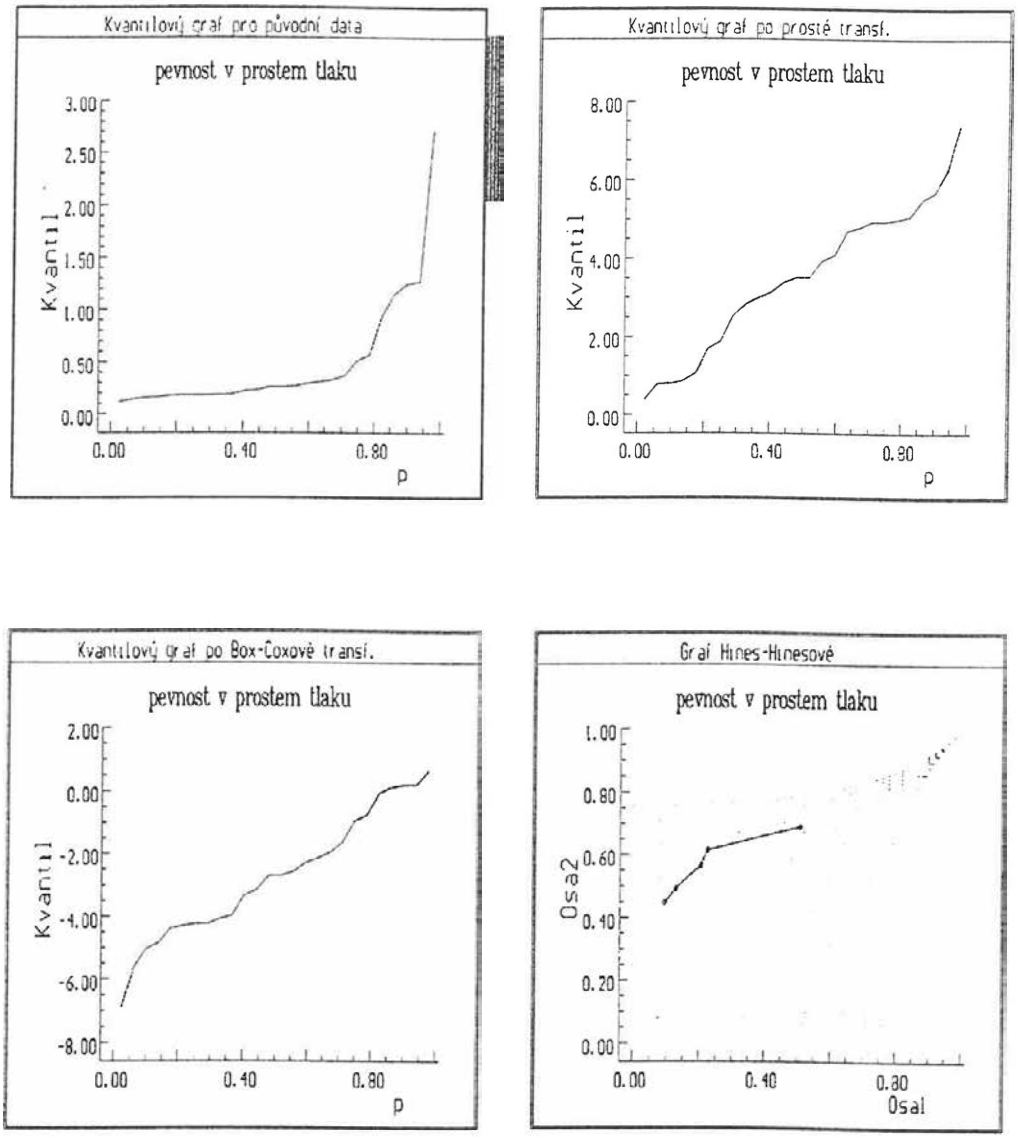
V grafickém výstupu jsou uvedeny čtyři typy grafů. Na obr.4 je uveden kvantilový graf pro původní data a pro data po prosté transformaci a po Box-Coxově transformaci. Na obr. 4 je uveden i graf Hines-Hinesové, který je konstruován pro výběr optimálního exponentu při prosté transformaci. Tečkované čáry odpovídají mocninám zleva. Lomená čára spojuje body odpovídající nalezeným optimálním exponentům pro různé části výběru (kvartil, oktil, sedecil). Pokud leží tato čára rovnoběžně s nejbližší izočárou, lze považovat transformaci za efektivní. Pokud leží napříč, nepochází asymetrie dat z mocninné nebo podobné transformace.

Na obr. 5 je uveden Q-Q graf pro původní data a pro data po prosté a Box-Coxově transformaci. Grafy jsou konstruovány pro ověření shody výběrového rozdělení s normálním. Dále je na obr. 5 uveden graf maximální věrohodnosti. Tento graf vyjadřuje závislost logaritmu věrohodnostní funkce pro Box-Coxovu transformaci proti hodnotě exponentu. Poloha maxima křivky na x-ové ose odpovídá optimální hodnotě exponentu z hlediska maximální věrohodnosti pro normální rozdělení.

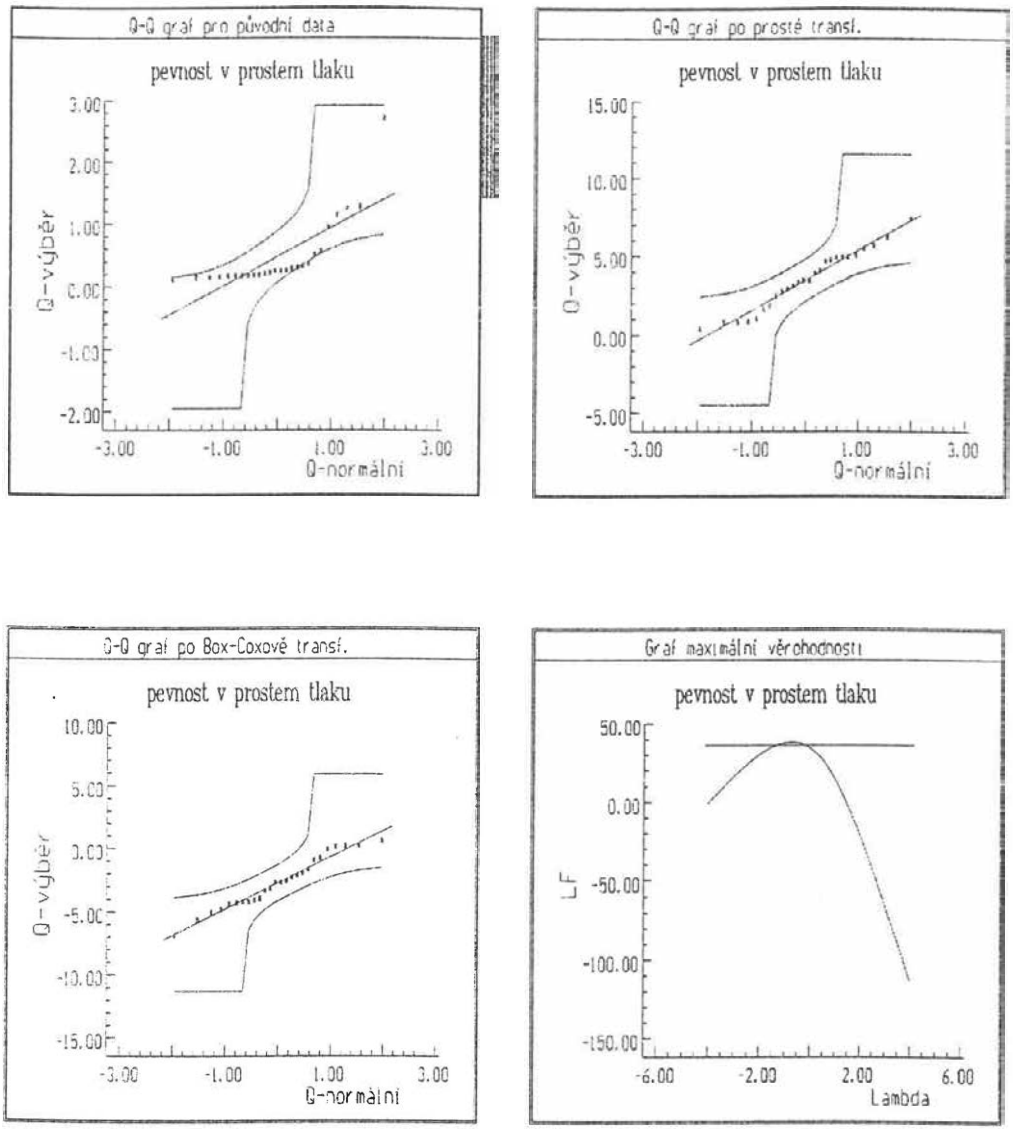
(1) PROSTÁ MOCNINNÁ TRANSFORMACE:

(A) Optimální hodnoty mocniny pro vybraná kritéria:

Optimální mocnina	-9.3333E-01	Pro šikmost	2.6720E-02
Optimální mocnina	-1.6000E+00	Pro špičatost	3.0398E+00
Optimální mocnina	-9.3333E-01	Pro asymetrii	2.4377E-03
Optimální mocnina	-1.2000E+00	Pro asymetrii, rob.	8.8679E-03
Optimální mocnina	-1.0667E+00	Pro Hinkley-asymetrii	7.4091E-02
Zvolená mocnina	-0.93		



Obr. 4: Grafy pro původní a transformovaná data



Obr. 5: Grafy pro původní a transformovaná data

Průměr	3.5329E+00
Rozptyl	3.5752E+00
Směrodatná odchylka	1.8908E+00
Šikmost	-2.6720E-02
Špičatost	2.1414E+00
Opravený průměr	2.5865E-01

(B) Kvantilové míry:

Kvantil P	Spodní mez	Horní mez	Polorozptyl
Medián 0.5	3.5158E+00	-	-
Kvartil 0.25	1.8234E+00	4.9684E+00	3.1450E+00

(C) Míry rozptylu

Kvantil P	Polosuma	Šikmost	Délka konců	Norm. d. konců
Kvartil 0.25	3.3959E+00	3.8132E-02	0.0000E+00	0.0000E+00

(2) BOX-COXOVA TRANSFORMACE:

(A) Optimální hodnoty mocniny pro vybraná kritéria:

Optimální mocnina	-9.3333E-01	pro šikmost	2.6720E-02
Optimální mocnina	-1.6000E+00	pro špičatost	3.0398E+00
Optimální mocnina	-9.3333E-01	pro asymetrii	2.4377E-03
Optimální mocnina	-1.2000E+00	pro asymetrii, rob.	8.8679E-03
Optimální mocnina	-1.0667E+00	pro Hinkley-asymetrii	6.9461E-02
Optimální mocnina	-8.0000E-01	pro věrohodnost	3.8355E+01
Zvolená mocnina	-0.93		

Průměr	-2.7138E+00
Rozptyl	4.1041E+00
Směrodatná odchylka	2.0259E+00
Šikmost	2.6720E-02
Špičatost	2.1414E+00
Opravený průměr	2.5865E-01

(B) Kvantilové míry:

Kvantil P	Spodní mez	Horní mez	Polorozptyl
Medián 0.5	-2.6955E+00	-	-
Kvartil 0.25	-4.2518E+00	-8.8218E-01	3.3697E+00

(C) Míry rozptylu

Kvantil P	Polosuma	Šikmost	Délka konců	Norm. d. konců
Kvartil 0.25	-2.5670E+00	-3.8132E-02	0.0000E+00	0.0000E+00

(4) SETŘÍDĚNÁ PŮVODNÍ A TRANSFORMOVANÁ DATA:

Původní	Po prosté transformaci	Po Box-Coxově-transformaci
1.1700E-01	3.9436E-01	-6.8656E+00
1.4000E-01	8.0005E-01	-5.6415E+00
1.5500E-01	8.1199E-01	-5.0331E+00
.	.	.
.	.	.
2.7100E+00	7.4079E+00	6.4890E-01

IV. URČENÍ ROBUSTNÍCH ODHADŮ PARAMETRŮ POLOHY A ROZPTÝLENÍ

Jak již bylo uvedeno, výběrový průměr $\bar{x}$ a výběrový rozptyl s^2 jsou efektivní odhady parametrů polohy rozptýlení jedině pro data, která pocházejí z normálního rozdělení. Jestliže však výběr pochází z jiného rozdělení než normálního nebo jsou-li v datech vybočující hodnoty, jak je tomu i v našem testovaném příkladě, efektivnost odhadů $\bar{x}$ a s^2 rychle klesá. Při narušení předpokladu normality způsobeném obyčejně vybočujícími měřeními lze získat efektivní odhady s využitím robustních metod.

(1) KLASICKÉ ODHADY PARAMETRŮ (za předpokladu normality):

Průměr	4.8492E-01
Směr. odchylka	5.7333E-01
Rozptyl	3.2871E-01
95.0% spolehlivost	
Spodní mez	2.5335E-01
Horní mez	7.1650E-01

(2) ROBUSTNÍ ODHADY PARAMETRŮ (pro neplatnost normality):

Medián	2.6000E-01
Směr. odchylka	1.5516E-01
Rozptyl	2.4074E-02
Směr. odchylka	2.0123E-03
Rozptyl (nepar.)	4.4859E-02
Směr. odchylka	2.0547E-03
Rozptyl (Marritz)	4.5329E-02
95.0% spolehlivost	
Spodní mez	1.6664E-01
Horní mez	3.5336E-01

Uřezání	**5% (pro P=0.05)**	**10% (pro=0.10)**	**40%(pro=0.40)**
Průměr	3.9991E-01	3.6192E-01	2.5558E-01
Směr. odchylka	4.3211E-01	4.6437E-01	1.5949E-01
Rozptyl	1.8672E-01	2.1564E-01	2.5436E-02
Průměr, winsor.	4.3042E-01	4.3004E-01	2.5712E-01
St.odch. winsor.	3.9736E-01	4.1044E-01	7.2868E-02
Rozptyl, winsor.	1.5790E-01	1.6846E-01	5.3098E-03
95.0% spolehlivost			
Spodní mez	2.2891E-01	1.7617E-01	1.7584E-01
Horní mez	5.7092E-01	5.4767E-01	3.3532E-01

3.3 ZÁVĚR

Soubor výsledků laboratorního stanovení pevnosti hornin zkouškou v prostém tlaku byl statisticky zpracován metodou jednorozměrných dat za účelem zjištění bodových a intervalových odhadů. Data byla nejprve analyzována pomocí spojité exploratorní analýzy, která pomocí diagnostických grafu odhaluje zvláštnosti dat. U testovaného souboru byly identifikovány odlehlé body, dále bylo zjištěno, že data nepochází z normálního rozdělení a jeví asymetrii.

Další analýza vstupních dat, tj. ověření základních předpokladů, spočívala v ověření předpokladu normality, nezávislosti a v detekci vybočujících hodnot. Jedním ze závěrů ověřování základních předpokladů bylo zamítnutí předpokladu normality, což znamená, že data nepochází z normálního rozdělení. Ověřením předpokladu nezávislosti dat byla testována vzájemná nezávislost výsledků. Předpoklad nezávislosti dat byl přijat, lze tedy učinit závěr, že se výsledky navzájem neovlivňují.

Vstupní analýzou dat, exploratorní analýzou a ověřením základních předpokladů, bylo zjištěno, že rozdělení výběru dat se odlišuje od rozdělení normálního, dále data jeví asymetrii a rovněž byly zjištěny odlehlé hodnoty. Na základě tohoto zjištění lze učinit nekompromisní závěr, že pro výpočet bodového parametru ***nelze použít klasický aritmetický průměr.***

Z výše uvedeného vyplynula nutnost zpracování dat matematickou transformací. Prostou mocninnou transformací a Box-Coxovou transformací byl vyčíslen opravený průměr, který je pro obě metody shodný. Při neplatnosti normality lze rovněž provést výpočet odhadů parametrů polohy i rozptýlení s využitím robustních metod. V následující tabulce je provedeno srovnání vypočtených jednotlivých odhadů polohy a rozptýlení.

Statistická charakteristika	Bodový odhad	Spodní mez	Horní mez
Průměr	4.8492E-01	2.5335E-01	7.1650E-01
Medián	2.6000E-01	1.6664E-01	3.5336E-01
Opravený průměr po prosté mocninné transformaci	2.5865E-01	-	-
Opravený průměr po Box-Coxově transformaci	2.5865E-01	-	-
Průměr pro P=0.05	3.9991E-01	2.2891E-01	5.7092E-01
Průměr pro P=0.10	3.6192E-01	1.7617E-01	5.4767E-01
Průměr pro P=0.40	2.5558E-01	1.7584E-01	3.3532E-01

Z přehledu v tabulce je zřejmé, že rozdíl mezi klasickým a opraveným průměrem, případně i mediánem, je asi 50%. Z praxe je známo, že u většiny souboru, zejména v oblasti technických, nelze zpravidla očekávat normální rozdělení. Jakých chybných závěrů se lze dopustit v rozhodovacích procesech při použití klasických odhadů není již třeba dále komentovat.

II. LINEÁRNÍ REGRESNÍ MODELY

1. Formulace lineárního regresního modelu

Cílem v řadě technických oborů je objasnění vztahu mezi měřenou, výstupní závisle proměnnou (vysvětlovanou) veličinou y a nastavovanými, vstupními nezávisle proměnnými (vysvětlujícími) veličinami x. Typ závislosti vyjádřený známou funkcí $y = f(x, \beta)$, záleží na tom, jaké povahy jsou veličiny y a x.

K odhadu parametru regresní funkce se používají mnohé metody a postupy, ale nejčastěji jde o různé varianty známé metody nejmenších čtverců. Podle této metody se získají odhady parametrů regresní funkce minimalizací součtu čtverců reziduí, tj. odchylek skutečných a vypočítaných hodnot vysvětlované proměnné y.

Metoda nejmenších čtverců má optimální vlastnosti za jistých předpokladů:

a) Regresní parametry β mohou nabývat libovolných hodnot. V technické praxi však často existují omezení parametrů, která vycházejí z jejich fyzikálního smyslu.
b) Regresní model je lineární v parametrech a platí tzv. aditivní model měření $y=x\beta+\varepsilon$, kde ε jsou náhodné veličiny, které zahrnují jak chyby měření, tak i chyby modelu, vzniklé tím, že funkce předpokládaného modelu neodpovídá skutečnému „teoretickému" modelu.
c) Žádné dva sloupce matice nenáhodných, nastavovaných hodnot, vysvětlujících proměnných nejsou kolineární, tj. rovnoběžné vektory.
d) Náhodné chyby mají nulovou střední hodnotu.
e) Náhodné chyby mají konstantní rozptyl, platí homoskedasticita.
f) Náhodné chyby jsou vzájemně nekorelované.
g) Náhodné chyby mají normální rozdělení.

Pokud platí prvních šest předpokladů, jsou odhady, získané minimalizací kritéria nejmenších čtverců, nejlepší nevychýlené lineární odhady regresních parametrů. V případě, že platí i poslední předpoklad, mají odhady normální rozdělení už pro konečné rozsahy výběru.

2. Statistické vlastnosti metody nejmenších čtverců

Bodové odhady jsou náhodné veličiny, které mají v praxi malý význam. Důležitější jsou ***konfidenční oblasti***, nazývané také ***oblasti nebo intervaly spolehlivosti***, ve kterých leží teoretická hodnota se zvolenou pravděpodobností $(1-\alpha)$. Stejně jako u jednorozměrných výběrů se volí hladina významnosti α 0.05 nebo 0.01. Této volbě odpovídají 95%ní nebo 99%ní intervaly spolehlivosti. Pro konstrukci konfidenčních pásů se používá ***Schéffeho metoda*** nebo ***metoda Working-Hottellinga***.

S konstrukcí intervalů spolehlivosti úzce souvisí ***testování významnosti parametrů β***. Standardním výstupem většiny programů regresní analýzy je závěr ***Fisherova-Snedecorova F-testu*** o významnosti koeficientu determinace a výsledky ***Studentova t-testu*** o významnosti jednotlivých parametrů β. F-test určuje zároveň simultánní významnost všech složek vektoru β kromě absolutního členu. Mohou nastat tyto případy:

a) F-test vychází nevýznamný, všechny t-testy vychází rovněž nevýznamné. Model se pak považuje za nevhodný, protože nevystihuje variability proměnné y.
b) F-test a všechny t-testy vychází významné. Model se považuje za vhodný vystižení variability proměnné y. Neznamená to ale, že je model správně navržen.
c) F-test vychází významný, ale t-testy nevýznamné u některých regresních parametrů. Model lze považovat za vhodný a provádí se případné vypouštění těch vysvětlujících proměnných x, pro které jsou parametry β nevýznamně odlišné od nuly.
d) F-test sice vychází významný, ale t-testy parametrů β indikují nevýznamnost všech vysvětlujících proměnných. To je paradox, protože formálně sice model jako celek vyhovuje, ale žádná z vysvětlujících proměnných není sama o sobě významná. Jde o důsledek ***multikolinearity.***

Multikolinearita neznamená v pravém slova smyslu porušení předpokladů klasické metody nejmenších čtverců, ale souvisí pouze s předpokladem o pozitivní definitnosti matice. Je to situace, kdy jeden ze sloupců x matice je lineární kombinací několika ostatních sloupců. V přítomnosti multikolinearity nelze odděleně sledovat vliv jednotlivých vstupních proměnných x. Multikolinearita se vyskytuje často i u modelů dobře popisujících data, u polynomických modelů a dat z neplánovitých experimentů.

Základní příčinou vzniku multikolinearity je, že regresní model obsahuje nadměrný počet nezávislých proměnných, které vyjadřují stejné faktory. Dále může být multikolinearita způsobena nevhodným rozmístěním bodů, hodnoty nezávisle proměnných kolísají jen v malém rozmezí a jsou proto kolineární s vektorem odpovídajícím absolutnímu členu. Multikolineritu může způsobit i fyzikální omezení v modelu nebo datech, kdy vznikají vazby mezi nezávisle proměnnými přímo ve studovaném systému.

K posouzení vhodnosti navrženého lineárního modelu s ohledem na možnou multikolinearitu navrhl ***Scott testační charakteristiku*** M_T. Na základě simulačních experimentů byla vytvořena následující pravidla k posouzení stupně multikolinearity:

1. Pokud je $\mathbf{M_T > 0.8}$, je model z hlediska multikolinearity nevyhovující a ***je třeba*** provést jeho případnou úpravu.
2. Pokud leží M_T v intervalu $\mathbf{0.33 \leq M_T \geq 0.88}$, je model z hlediska multikolinearity málo vyhovující a ***je vhodná*** jeho případná úprava.
3. Pokud je $\mathbf{M_T < 0.33}$, není model výrazně ovlivněn multikolinearitou a ***není třeba*** ho upravovat.

Test pomocí Scottova kritéria M_T je vhodný především tam, kde je zapotřebí stanovit ty vysvětlující proměnné, které významně přispívají k objasnění variability proměnné y. Pokud jde pouze o úlohu aproximace dat empirickým modelem, např. polynomem, není třeba k hodnotám M_T přihlížet.

3. Regresní diagnostika

Při vyhodnocení lineárních i nelineárních regresních modelů se často užívá metody nejmenších čtverců. Tato metoda však ještě nezajišťuje nalezení přijatelného modelu, a to jak ze statistického, tak fyzikálního hlediska. Zdrojem problému jsou složky tzv. regresního tripletu (data, model, metoda odhadu).

Metoda nejmenších čtverců poskytuje optimální výsledky jenom při současném

splnění předpokladů o datech a o regresním modelu. Pokud tyto předpoklady nejsou splněny, je metoda nejmenších čtverců nevhodná. *Regresní diagnostika* obsahuje postupy k identifikaci:

a) kvality dat pro navržený model,
b) kvality modelu pro navržená data,
c) splnění základních předpokladů metody nejmenších čtverců.

Základní rozdíl mezi regresní diagnostikou a klasickými testy spočívá v tom, že u regresní diagnostiky není třeba přesně formulovat alternativní hypotézu a jsou přitom odhaleny typy odchylek ideálního regresního tripletu. Tímto pojetím se naše regresní diagnostika blíží spíše k exploratorní regresní analýze, která vychází z faktu, že *„uživatel ví o analyzovaných datech přece jenom více než počítač"*. Počítač slouží jako nástroj analýzy dat, modelu a metody odhadu. Model je navrhován interaktivně spoluprací uživatele s programem na osobním počítači. Tím by měl být omezen vznik formálních regresních modelů, které nemají fyzikální smysl a jsou v technické praxi nepoužitelné.

3.1 Využití průzkumové analýzy dat

O metodách průzkumové (exploratorní) analýzy jednorozměrných dat bylo pojednáno v předchozím článku. V regresní analýze jsou využívány tyto metody:

a) pro určení statistických zvláštností jednotlivých proměnných nebo reziduí,
b) k posouzení „párových" vztahů mezi všemi sledovanými proměnnými,
c) k ověření předpokladů o rozdělení proměnných.

K orientačnímu posouzení vztahů mezi jednotlivými proměnnými se užívá rozptylových grafů, kde se na osy vynášejí přímo hodnoty sledovaných proměnných. Informace o multikolinearitě lze získat vynesením dvojic vysvětlujících proměnných. Přibližně lineární závislost zde indikuje silnou multikolinearitu. Na druhé straně však může vést vynášení y proti x k mylným závěrům o nelinearitě modelu, který je ve skutečnosti lineární.

K ověření normality dat se často používá Q-Q grafů. Mezi základní techniky průzkumové analýzy patří i stanovení volby rozsahu a rozmezí dat, jejich variability a přítomnosti vybočujících pozorování. K tomu lze např. využít grafů rozptýlení s kvantily a řady dalších postupů. Přes svoji jednoduchost umožňuje průzkumová analýza identifikovat ještě před vlastní analýzou:

a) nevhodnost dat, tj. malé rozmezí nebo přítomnost vybočujících bodů,
b) nesprávnost navrženého modelu (skryté proměnné),
c) multikolinearitu,
d) nenormalitu.

3.2 Posouzení kvality dat

Kvalita dat úzce souvisí s užitým regresním modelem. Při posuzování se sleduje zejména výskyt vlivných bodů, které jsou hlavním zdrojem řady problémů, především zkreslení odhadů a růst rozptylů až k naprosté nepoužitelnosti regresních odhadů parametrů. Vlivné body ovlivňují většinu výsledků regrese a lze je rozdělit do tří základních skupin:

a) ***Hrubé chyby***, které jsou způsobeny měřenou veličinou (vybočující pozorování) nebo nevhodným nastavením vysvětlujících proměnných

(extrémy). Jsou obvykle důsledkem chyb při manipulaci s daty.

b) ***Body s vysokým vlivem*** jsou speciálně vybrané body, které byly přesně změřeny a obvykle rozšiřují predikční schopnosti modelu.

c) ***Zdánlivě vlivné body*** vznikají jako důsledek nesprávně navrženého regresního modelu. Podle toho, kde se vlivné body vyskytují, lze provést dělení na:
   1. vybočující pozorování, které se na ose y výrazně liší od ostatních,
   2. extrémy, které se liší v hodnotách na ose x nebo v jejich kombinaci (v případě multikolinearity) od ostatních bodů.

Vyskytují se však i body, které jsou jak vybočující, tak i extrémní. O jejich výsledném vlivu však rozhoduje zejména to, že jsou extrémy. K identifikaci vlivných bodů typu vybočujícího pozorování se využívá především reziduí a k identifikaci extrémů pak diagonálních prvků projekční matice.

Statistická analýza reziduí vychází z předpokladu, že jde o odhady chyb. Pomocí reziduí se tak ověřují vlastnosti chyb, i když tento přístup není zcela správný, neboť rezidua nejsou nezávislá, i když chyby nezávislé jsou. Rezidua jsou rozlišována na ***klasická, normovaná, standardizovaná, Jackknife, predikovaná a rekurzívní.*** Různé typy reziduí se liší co do vhodnosti k diagnostickým účelům:

1) k identifikaci heteroskedasticity postačují standardizovaná rezidua,

2) k určení vybočujících bodů pak Jackknife rezidua nebo predikovaná rezidua,

3) k detekci autokorelace rekurzívní rezidua.

K analýze reziduí se užívá především různých typů grafů. Mezi klasické patří tři základní typy grafů reziduí, které mohou indikovat nesprávnost navrženého modelu, sezónní trendy, heteroskedasticitu nebo vlivné body v datech. Pokud se v těchto grafech reziduí objeví „mrak" bodů, je indikována správnost metody nejmenších čtverců. Různé druhy obrazců bodů v grafu indikují převážně nesprávnost v datech nebo nesprávnost modelu.

K identifikaci různých typů vlivných bodů se používá řada grafů, které kombinují rozličné typy reziduí s prvky tzv. projekční matice:

1) Graf predikovaných reziduí
2) Williamsův graf
3) Pregibonův graf
4) MCCullohův-Meeterův graf
5) Indexové grafy
6) Rankitové grafy

Při posuzování vlivných bodů je třeba mít na paměti, že mohou nestejně výrazně ovlivňovat různé charakteristiky regrese. Například body ovlivňující výrazně predikci nemusí být z hlediska rozptylu parametrů vůbec vlivné. Stupeň vlivu jednotlivých bodů je třeba posuzovat vždy s ohledem na to, které charakteristiky regrese ovlivňují. K identifikaci vlivných bodů existuje řada dalších diagnostik, které lze rozdělit podle dvou základních přístupů:

a) První je založen na sledování změn, ke kterým dojde při vypuštění jednotlivých bodů.

b) Druhý přístup vychází z platnosti lineárního regresního modelu se speciální strukturou rozptylů chyb.

Podrobný popis těchto diagnostik lze nalézt v příslušné odborné literatuře.

3.3 Posouzení kvality navrženého modelu

Kvalitu regresního modelu lze posoudit v případě jedné vysvětlující proměnné x přímo z rozptylového grafu závislosti y na x. V případě více vysvětlujících proměnných mohou však rozptylové grafy mylně indikovat nelinearitu i u lineárního modelu. Z řady různých grafů k posouzení vztahu y a x jsou dva základní:

a) parciální regresní grafy
b) parciální reziduální grafy

Parciální regresní grafy jsou zařazeny mezi základní nástroje počítačové interaktivní analýzy regresních modelů. Vyjadřují závislost mezi y a zvolenou vysvětlující proměnnou při statisticky neměnném vlivu ostatních vysvětlujících proměnných. Tyto grafy umožňují nejen posouzení kvality navrženého modelu, ale indikují i přítomnost vlivných bodů a nesplnění předpokladů klasické metody nejmenších čtverců.

Parciální reziduální grafy se označují také jako grafy „komponenta+reziduum". Jsou speciální analogií parciálního regresního grafu. Jedná se o závislost parciálních reziduí přímo na proměnné x. Parciální reziduální grafy se doporučují především k indikaci rozličných typů nelinearity v případě nesprávně navrženého regresního modelu.

Dalším testem, který vychází přímo z reziduí, je ***znaménkový test***. Nesprávnost navrženého regresního modelu se projeví nenáhodností reziduí a tuto nenáhodnost lze testovat právě znaménkovým testem, při němž se určuje počet sekvencí, kde mají rezidua stejná znaménka.

3.4 Ověření předpokladů nejmenších čtverců

Průvodním jevem řady měřených dat je ***heteroskedasticita***. Heteroskedasticita znamená ***porušení*** předpokladu ***o konstantnosti rozptylu chyb***. Je-li rozptyl chyb v datech konstantní, nazývají se pak data ***homoskedastická***. Rozptyl měření bývá rostoucí funkcí velikosti proměnné y, protože relativní přesnost měření je obyčejně konstantní. Tento typ heteroskedasticity lze odhalit v grafu závislosti čtverce reziduí na y. Vzniká obrazec s výrazným lineárním nebo nelineárním trendem.

Předpokladem pro správné vyhodnocení regresního modelu je nezávislost pozorování, tedy i vzájemná nezávislost chyb. Není-li tento předpoklad splněn, hovoří se o ***autokorelaci***. S autokorelací se lze setkat především v případech, kdy se výběrová pozorování vztahují k různým časovým okamžikům nebo intervalům (časové řady), v praxi například při sledování parametrů nějakého zařízení pracujícího v kontinuálním režimu. Autokorelaci lze testovat, nejpoužívanější je ***Durbinův-Watsonův test*** a ***Waldův test***. Vysoká hodnota autokorelace může být i důsledkem nesprávně navrženého regresního modelu.

K ověření normality chyb se užívá již zmíněného Q-Q grafu, v němž se na osu pořadnic vynáší pořádkové statistiky reziduí a na osu souřadnic kvantily normovaného normálního rozdělení pro pořadové pravděpodobnosti. Pokud je zjištěno, že chyby nemají normované normální rozdělení, může to být způsobeno jiným rozdělením chyb nebo častěji přítomností vlivných bodů.

4. Postupy při porušení předpokladů metody nejmenších čtverců

V kapitole 1. byly uvedeny předpoklady, za kterých vede metoda nejmenších čtverců k nejlepším nestranným lineárním odhadům. Z těchto předpokladů vychází také konstrukce intervalů spolehlivosti a testů významnosti. V praxi však bývají některé předpoklady porušeny. Nejdůležitější diagnostické postupy k odhalení porušení předpokladů metody nejmenších čtverců jsou uvedeny předchozí kapitole 3.

V řadě praktických úloh jsou na regresní parametry kladena omezení zajišťující jejich fyzikální smysl a interpretovatelnost. Nejčastějším omezením v praxi je požadavek, aby regresní model procházel počátkem. Dalším příkladem jsou často požadované kladné hodnoty parametrů. Do této skupiny patří tedy úlohy, kdy některé parametry musí nabývat zadaných hodnot, dále kdy některé parametry musí zachovávat předepsané vzájemné poměry, součty nebo rozdíly některých parametrů se musí rovnat zadanému číslu nebo regresní model musí procházet zadanými body o známých souřadnicích. Pro řešení těchto omezení se používá technika Lagrangeových multiplikátorů, neboli ***metoda podmínkových nejmenších čtverců (MPNČ).*** Při nekonstantnosti rozptylu se využívá ***metody vážených nejmenších čtverců (MVNČ)***. ***Metody zobecněných nejmenších čtverců (MZNČ)*** se využívá při autokorelaci, ***metody racionálních hodností*** u multikolinearity. ***Metoda rozšířených nejmenších čtverců (MRNČ)*** se používá v případě, že všechny proměnné jsou zatížené náhodnými chybami Jsou-li rozdělení dat jiná než normální a data s vybočujícími hodnotami a extrémy, pak se využívá robustních metod. Podrobný popis těchto metod lze nalézt v odborné literatuře.

5. PŘÍKLAD STATISTICKÉHO ZPRACOVÁNÍ MODELU LINEÁRNÍ REGRESE

5.1 Zadání: Pro hodnocení rozpojitelnosti je používána jednotná klasifikace sedimentů JKS. Hodnotícím kritériem je index JKS, jehož hodnota je stanovena na základě výsledků následujících laboratorních zkoušek: určení obsahu jílových minerálů a karbonátů, stanovení objemové hmotnosti a vlhkosti, zkoušky odporu v penetraci a pevnosti v prostém tlaku. Uvedené zkoušky jsou ve stejném pořadí uloženy ve vstupním souboru a představují nezávisle proměnné x_1 až x_6 a index JKS představuje závisle proměnnou y. Vstupní data nejsou pro svoji rozsáhlost v článku uvedena.

5.2 Řešení: U vstupních dat bylo provedeno statistické zpracování pomocí software Adstat

V S T U P

(1) ZVOLENÁ STRATEGIE REGRESNÍ ANALÝZY:

Omezení, P	1.0000E-34
Transformace	Ne
Váhy	Ne
Absolutní člen zahrnut	Ano

Uvedené podmínky se definují pro zajištění maximální flexibility lineární regrese.

Omezení P je omezení na vlastní čísla. Implicitní hodnota je 1E-34. Pro běžné úlohy odpovídá tato volba klasické metodě nejmenších čtverců. Čím je tento parametr vyšší, tím je také vyšší vychýlení odhadů.

(2) PODMÍNKY A KVANTILY PRO STATISTICKÉ TESTY:

Hladina významnosti, alfa	0.050
Počet bodů, n	67
Počet parametrů, m	6
Kvantil Studentova rozdělení t(1-alpha/2,n-m)	2.000
Kvantil rozd. Chí-kvadrát Chi-square(1-alpha,m)	14.067

V Ý S T U P

(1) PŘEDBĚŽNÁ STATISTICKÁ ANALÝZA:

Proměnná	Průměr	Směrodatná odchylka	Párový korelační koeficient	Spočtená hladina významnosti
Y	1.0276E+02	4.1375E+00	1.0000	-----
x1	6.1761E+01	1.2143E+01	-0.4050	0.001
x2	6.9104E+00	7.3808E+00	0.5123	0.000
x3	2.1184E+00	1.0798E-01	0.8257	0.000
x4	3.2096E+01	5.9861E+00	-0.7177	0.000
x5	2.2888E+02	1.0395E+02	0.9051	0.000
x6	2.4075E+00	3.7446E+00	0.5705	0.000

Pro všechny proměnné jsou vypočteny průměry, směrodatné odchylky a párové korelační koeficienty všech vysvětlujících proměnných (x_i) vzhledem k vysvětlované proměnné (y) a párové korelační koeficienty mezi vysvětlujícími proměnnými.

Párové korelační koeficienty mezi dvojicemi vysvětlujících proměnných		Spočtená hladina významnosti
x1 versus x2	-2.7478E-01	0.024
x1 versus x3	-6.6220E-01	0.000
x1 versus x4	6.2289E-01	0.000
x1 versus x5	-4.1477E-01	0.000
x1 versus x6	-1.1705E-01	0.346
x2 versus x3	3.4107E-01	0.005
x2 versus x4	-2.0311E-02	0.870
x2 versus x5	2.8591E-01	0.019
x2 versus x6	1.3679E-01	0.270
x3 versus x4	-8.2549E-01	0.000
x3 versus x5	7.8489E-01	0.000

x3 versus x6	3.1669E-01	0.009
x4 versus x5	-7.6367E-01	0.000
x4 versus x6	-2.8004E-01	0.022
x5 versus x6	3.4719E-01	0.004

(2) INDIKACE MULTIKOLINEARITY:

Č [j]	Vlastní čísla korel. matice l[j]	Čísla podmíněnosti K[j]	Variance inflation factor VIF[j]	Vícenás.korel. koef pro X[j]
1	1.0840E-01	3.0337E+01	2.2284E+00	0.7425
2	1.5502E-01	2.1214E+01	1.7191E+00	0.6468
3	5.3361E-01	6.1629E+00	5.4598E+00	0.9038
4	9.2525E-01	3.5542E+00	5.7537E+00	0.9090
5	9.8915E-01	3.3246E+00	3.6278E+00	0.8511
6	3.2886E+00	1.0000E+00	1.1569E+00	0.3683

Maximální číslo podmíněnosti K: 3.0337E+01

(K[j], K > 1000 indikuje silnou multikolinearitu)

(VIF[j] > 10 indikuje silnou multikolinearitu)

K identifikaci multikolinearity bylo použito číslo podmíněnosti K a faktor VIF. Jsou-li hodnoty K a VIF vyšší než uvedené meze, pak je mezi nezávisle proměnnými silná lineární závislost, která zhoršuje statistické vlastnosti odhadů a v některých případech může znehodnotit i celou analýzu. V našem případě nebyla multikolinearita pomocí K a VIF zjištěna.

(3) ODHADY PARAMETRŮ A TESTY VÝZNAMNOSTI:

Parametr	Odhad	Směrodatná odchylka	Test H_0: B[j] = 0 vs. H_A: B[j] <> 0		
			t-kriterium	hypotéza H_0 je	Hlad. výz.
B[0]	8.2360E+01	2.7951E+00	2.9466E+01	Zamítnuta	0.000
B[1]	6.2429E-02	6.3875E-03	9.7736E+00	Zamítnuta	0.000
B[2]	1.8137E-01	9.2303E-03	1.9649E+01	Zamítnuta	0.000
B[3]	7.3820E+00	1.1244E+00	6.5653E+00	Zamítnuta	0.000
B[4]	-1.6450E-01	2.0821E-02	-7.9006E+00	Zamítnuta	0.000
B[5]	1.8550E-02	9.5208E-04	1.9484E+01	Zamítnuta	0.000
B[6]	2.8530E-01	1.4925E-02	1.9116E+01	Zamítnuta	0.000

V tabulce (3) jsou uvedeny odhady parametrů, jejich směrodatné odchylky a t-testu pro testování významnosti individuálních parametrů. Verbálně je vyjádřen výsledek testu nevýznamnosti, nulová hypotéza je zamítnuta, parametry lze tedy považovat za významné. V posledním sloupci je pak zpětně spočítaná hladina významnosti t-testu.

(4) STATISTICKÉ CHARAKTERISTIKY REGRESE:

Vícenásobný korelační koeficient, R	9.9526E-01
Koeficient determinace, R^2	9.9054E-01
Predikovaný korelační koeficient, Rp^2	9.9304E-01
Střední kvadratická chyba predikce, MEP	2.3392E-01
Akaikeho informační kritérium, AIC	-1.0896E+02

V tabulce (4) je vyčíslen vícenásobný koeficient a koeficient determinace, což je čtverec vícenásobného korelačního koeficientu, dále je spočítán predikovaný korelační koeficient, střední kvadratická chyba predikce a Akaikeho informační kritérium.

Střední kvadratická chyba predikce, MEP se používá k ověření linearity modelu. Vhodný model indikuje nejnižší hodnota MEP. Nelinearitu lze dobře indikovat statistikou AIC. Za nejvhodnější je považován model, pro který je AIC minimální. Predikovaný koeficient determinace Rp^2 je vypočten pomocí statistiky MEP a jeho vysoká hodnota potvrzuje vhodnost zvoleného modelu.

(5) ANALÝZA KLASICKÝCH REZIDUÍ:

Bod	**Měřená hodnota**	**Predikovaná hodnota**	**Směrodatná odchylka**	**Klasické reziduum**	**Relativní reziduum**
i	yexp[i]	yvyp[i]	s(yvyp[i])	e[i]	er[i]
1	9.9000E+01	9.9302E+01	8.7963E-02	-3.0178E-01	-3.0483E-01
2	1.0100E+02	1.0105E+02	1.0576E-01	-5.3766E-02	-5.3234E-02
66	1.0400E+02	1.0456E+02	6.7685E-02	-5.6395E-01	-5.4226E-01
67	1.0100E+02	1.0113E+02	9.4521E-02	-1.3005E-01	-1.2877E-01

Rezidualní součet čtverců, RSC	1.0691E+01
Průměr absolutních hodnot reziduí, Me	3.1695E-01
Průměr relativních reziduí, Mer	3.0956E-01
Odhad reziduálního rozptylu, s^2(e)	1.7819E-01
Odhad směrodatné odchylky reziduí, s(e)	4.2212E-01
Odhad šikmosti reziduí, g1(e)	4.8786E-03
Odhad špičatosti reziduí, g2(e)	2.9913E+00

V tabulce (5) je provedena analýza klasických reziduí naměřených a vypočítaných (predikovaných) hodnot, směrodatné odchylky měření, klasická a relativní rezidua. Platí, že pokud je naměřená hodnota větší než vypočítaná, je reziduum kladné a naopak.

V tabulce (5) jsou uvedeny hodnoty pouze pro vzorky 1,2 a 66, 67 jako ukázka. Hodnoty zbývajících 63 vzorků jsou pro svoji rozsáhlost ze článku vypuštěny. Různé typy reziduí se liší co do vhodnosti k diagnostickým účelům na základě sestrojených a dále uvedených grafů.

V následující tabulce (6) jsou uvedeny výsledky testování regresního tripletu. Testy zahrnují statistiku F-testu významnosti vícenásobného korelačního koeficientu, statistika M pro testaci stupně multikolinearity, statistika SF pro test heteroskedasticity a statistika L pro ověření normality reziduí. Dále jsou provedeny testy autokorelace a znaménkový test trendu v reziduích. Normalita dat byla prokázána. Rezidua vykazují homoskedasticitu, nejsou autokorelována a nevykazují trend. Scottovo kritérium multikolinearity je větší než hodnota 0.33, ale menší než 0.8. V ADSTATu v bodě (2) není multikolinearita faktorem VIF indikována.

(6) TESTOVÁNÍ REGRESNÍHO TRIPLETU (DATA + MODEL + METODA):

Fisher-Snedocorův test významnosti regrese,F	1.0468E+03
Tabulkový kvantil, F(1-alpha,m-1,n-m)	2.2541E+00
Závěr	**Navržený model je významný.**
Spočtená hladina významnosti	0.000
Scottovo kriterium multikolinearity, M	4.8111E-01
Závěr	**Navržený model není korektní.**
Cook-Weisbergův test heteroskedasticity, Sf	5.9340E-02
Tabulkový kvantil, Chi^2(1-alpha,1)	3.8415E+00
Závěr	**Rezidua vykazují homoskedasticitu.**
Spočtená hladina významnosti	0.808
Jarque-Berraův test normality reziduí, L(e)	4.7628E-04
Tabulkový kvantil, Chi^2(1-alpha,2)	5.9915E+00
Závěr	**Normalita je prokázána.**
Spočtená hladina významnosti	1.000
Waldův test autokorelace, Wa	5.6716E-01
Tabulkový kvantil, Chi^2(1-alpha,1)	3.8415E+00
Závěr	**Rezidua nejsou autokorelována.**
Spočtená hladina významnosti	0.451
Znaménkový test, Dt	-1.2295E+00
Tabulkový kvantil, N(1-alpha/2)	1.6449E+00
Závěr	**Rezidua nevykazují trend.**
Spočtená hladina významnosti	0.109

V tabulce (7) je uveden přehled vlivných bodů, který se používá k identifikaci vlivných a odlehlých bodů. Ostatní hodnoty zde nejsou pro svoji rozsáhlost uvedeny. Dále jsou uvedeny věrohodnostní vzdálenosti pro první a poslední dva vzorky jako ukázka. Věrohodnostní vzdálenosti se rovněž používají při diagnostice vlivných a podezřelých bodů.

Z numerické analýzy dat z kritiky dat v bodě č. 7, tj. indikace vlivných bodů, byly

vyhodnoceny celkem 3 outliery určené *Jackknifeovým reziduem*. Jedná se o body č. 5, 6 a 54.

(7) **INDIKACE VLIVNÝCH BODŮ: (* indikuje odlehlý nebo vlivný bod)**

Bod I	Standardizované reziduum eS[i]	Jackknife reziduum eJ[i]	Predikované reziduum eP[i]	Diagonální prvky H[i, i]
5	-2.4288E+00	-2.5364E+00*	-1.0458E+00	3.8938E-02
6	2.5543E+00	2.6830E+00*	1.1055E+00	4.8705E-02
7	4.8362E-01	4.8051E-01	2.3166E-01	2.2339E-01*
24	-1.4886E+00	-1.5041E+00	-7.2368E-01	2.4609E-01*
33	1.5139E+00	1.5308E+00	1.0260E+00	6.1203E-01*
34	2.6367E-01	2.6161E-01	1.2854E-01	2.5028E-01*
35	4.1889E-01	4.1599E-01	2.0039E-01	2.2137E-01*
39	-6.7040E-01	-6.6730E-01	-1.4438E+00	9.6158E-01*
43	1.3059E-01	1.2952E-01	6.2138E-02	2.1296E-01*
54	-2.1581E+00	-2.2283E+00*	-9.2489E-01	2.9814E-02
62	-8.0610E-01	-8.0372E-01	-5.5338E-01	6.2190E-01*
67	-3.1612E-01	-3.1374E-01	-1.3692E-01	5.0140E-02

Bod I	Zobecněné diag. prvky Hm[i,i]	Cookova vzdálenost D[i]	Atkinsonova vzdálenost A[i]	Vliv na predikci DF[i]
24	2.7393E-01*	1.0332E-01	2.5159E+00*	-8.5935E-01*
33	6.2685E-01*	5.1652E-01*	5.6289E+00*	1.9226E+00*
34	2.5114E-01*	3.3154E-03	4.4253E-01	1.5115E-01
39	9.6187E-01*	1.6070E+00*	9.7740E+00*	-3.3385E+00*
62	6.2600E-01*	1.5269E-01	3.0178E+00*	-1.0308E+00*

Bod	Věrohodnostní vzdálenosti		
i	LD(b)[i]	LD(s^2)[i]	LD(b,s^2)[i]
1	2.7078E-02	1.2412E-03	2.8161E-02
2	1.2945E-03	7.2520E-03	8.5275E-03
.	.	.	.
66	5.3963E-02	8.5908E-03	6.3444E-02
67	5.8903E-03	5.9626E-03	1.1775E-02

5.3 DISKUSE KE GRAFICKÉMU VÝSTUPU

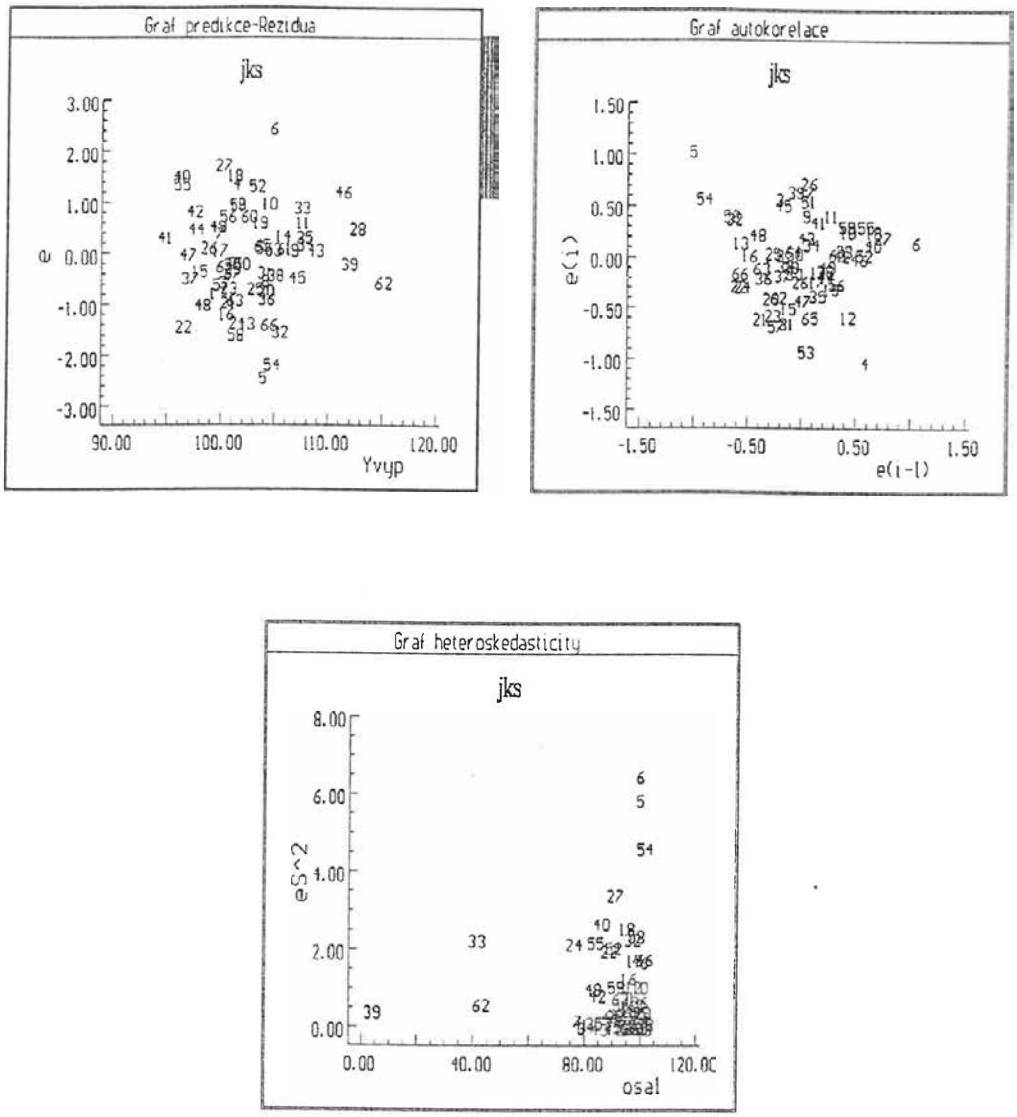
Na obr. 1 je uveden graf predikce rezidua, autokorelace a graf heteroskedasticity. Na grafu predikovaných reziduí je nakreslen graf reziduí proti hodnotám určeným z regresního modelu. V grafu autokorelace jsou proti sobě vynášena postupně sousední rezidua. Pokud vznikne přibližně lineární obraz, je indikována silná autokorelace prvního řádu. V našem testovaném případě nebyla autokorelace zjištěna. Heteroskedasticita je indikována pokud vznikne typický klínový obrazec. V našem případě není heteroskedasticita indikována.

Indexové grafy na obr. 2 mají informativní charakter, graf ***index-normalizovaná rezidua*** indikuje outliery a graf ***index-prvky hat matice*** indikuje přítomnost extrémů. Pro vlivné body jsou hodnoty těchto statistik výrazně odlišné od ostatních.

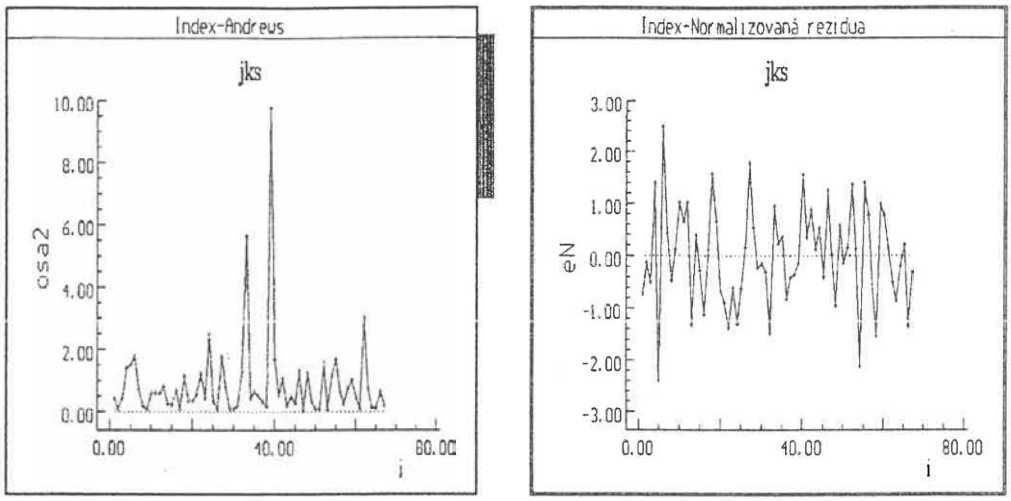
Rankitové grafy na obr. 3 upozorňují na vybočující hodnoty. Graf ***rankit-Andrews, rankit-predikovaná rezidua*** a ***rankit-Jackknife rezidua*** indikují shodně vybočující bod č.39, 33 a případně i 6.

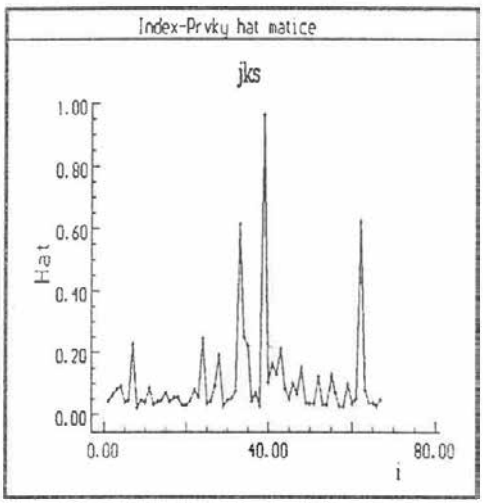
Při analýze grafů vlivných bodů na obr. 4 a 5 byly identifikovány outliery i extrémy. Graf ***predikovaných reziduí*** indikuje jeden extrém, bod č.33. ***Pregibonův graf*** indikuje outliery a extrémy dohromady, z tohoto grafu byly vyhodnoceny tři vlivné body, č.33, 39 a 62. ***Williamsův graf*** indikuje tři extrémy, bod č.33, 39 a 62 a tři odlehlé body, č. 55, 6 a 54. ***McCulloh-Meeterův graf*** určuje dva extrémy, č.33 a 39. ***L-R graf*** indikuje tři extrémy, body č. 33, 39 a 62. Z grafické analýzy lze vyhodnotit pouze extrémy, outliery nebyly prokázány.

Na základě výsledků jak numerické, tak grafické kritiky dat, nebude z výpočtu vypuštěn žádný odlehlý bod. Rovněž vzhledem ke způsobu získání dat (viz zadání) nelze vyloučit ze souboru žádnou hodnotu.

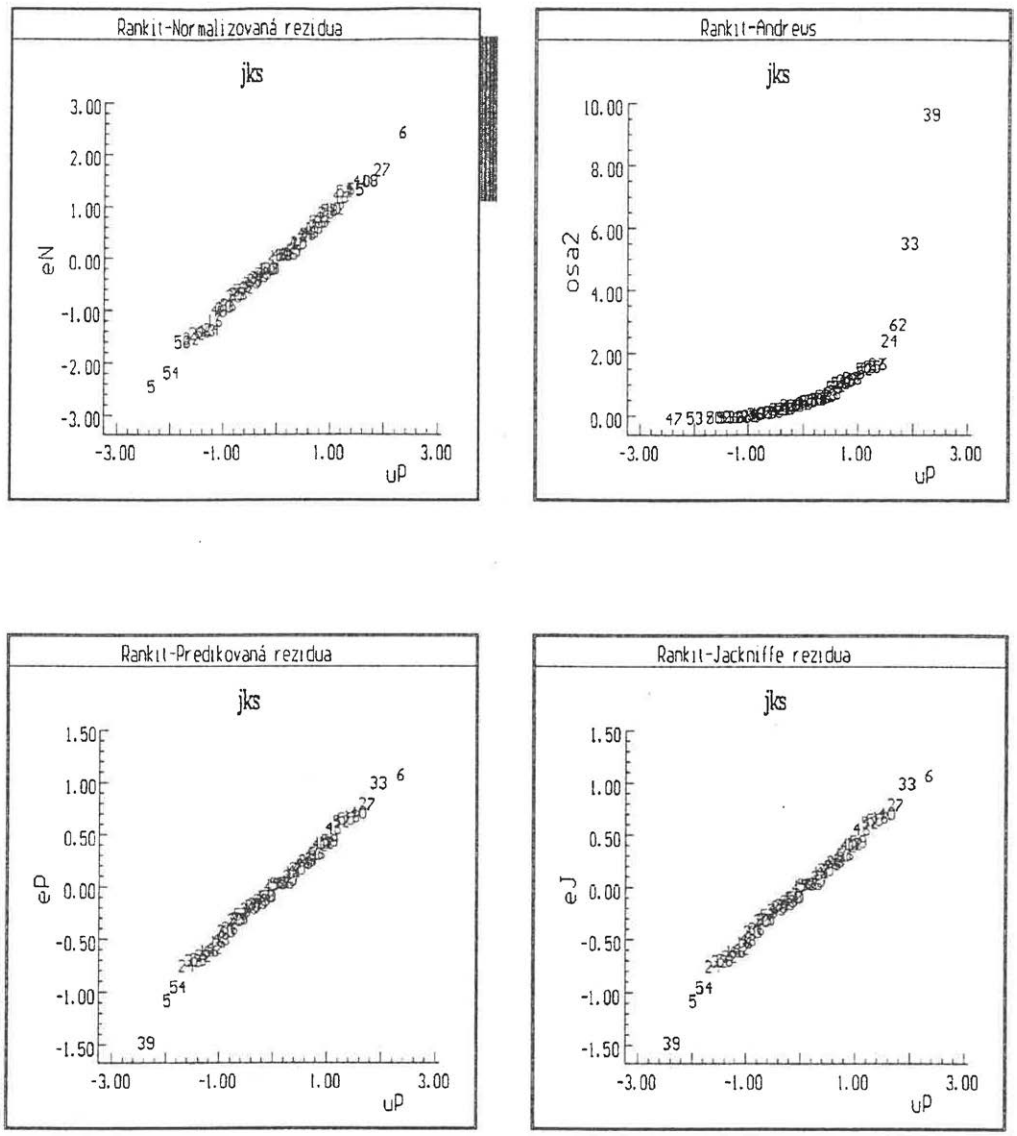


Obr. 1 Graf predikce – rezidua, autokorelace a heteroskedasticity

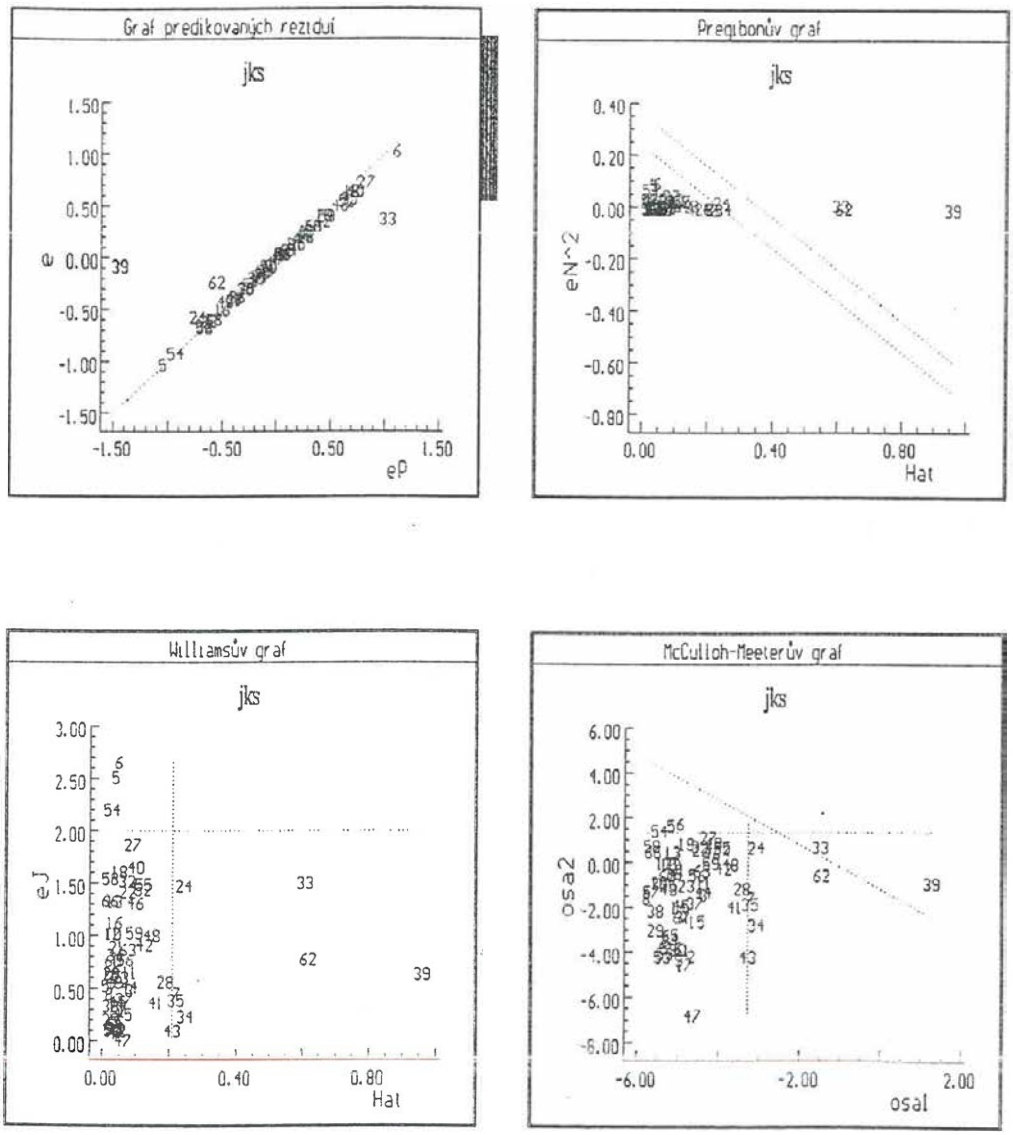




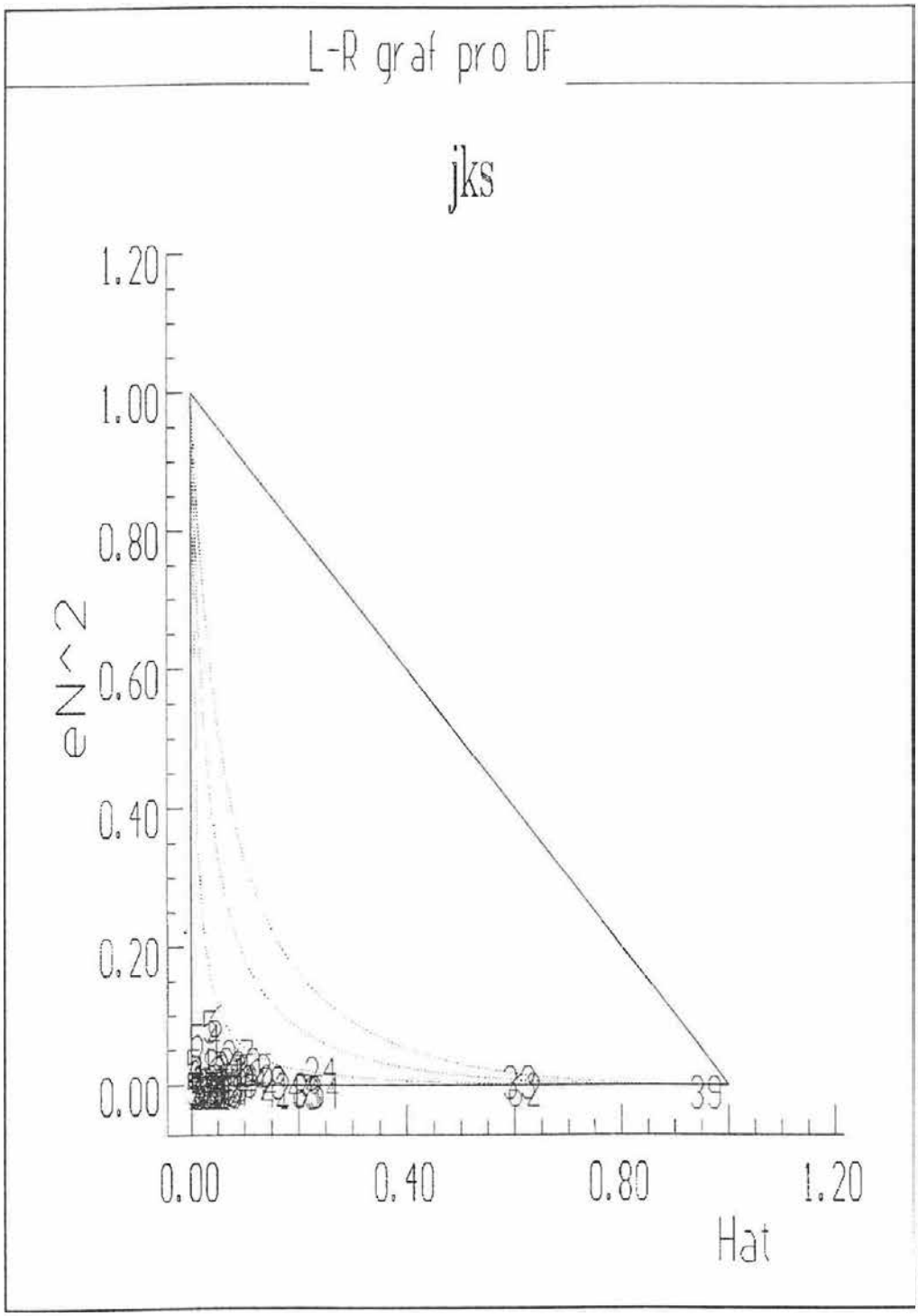
Obr. 2 Indexové grafy



Obr. 3 Rankitové grafy



Obr. 4 Grafy vlivných bodů



Obr. 5 Graf vlivných bodů

5.4 ZÁVĚR

Jako příklad lineární regrese byla vybrána ukázka z praxe, týkající se ověření vztahů mezi výsledky laboratorních zkoušek a indexem jednotné klasifikace sedimentů.

Statistickou analýzou byl vyčíslen ***vícerozměrný lineární regresní model***, který byl získán statistickým zpracováním vstupních dat po předchozím testování regresního tripletu, data + model + metoda, který obsahuje důležité statistické testy pro posouzení kvality dat a kvality vyčísleného regresního modelu.

Ze statistické analýzy vyplývá, že index jednotné klasifikace sedimentů je závislý na všech šesti zadaných parametrech x_1 až x_6 a vyčíslený regresní model má tvar:

$$y = 82.360(\pm 2.795) + 6.2429.10^{-2}(\pm 6.3875.10^{-3})\, x_1 + 1.8137.10^{-1}(\pm 9.2303.10^{-3})\, x_2 + 7.3820\,(\pm 1.1244)\, x_3 - 1.6450.10^{-1}(\pm 2.0821.10^{-2})\, x_4 + 1.8550.10^{-2}(\pm 9.5208.10^{-4})\, x_5 + 2.8530.10^{-1}(\pm 1.4925.10^{-2})\, x_6$$

kde y je index jednotné klasifikace sedimentů
x_1 obsah jílových minerálů (%)
x_2 obsah karbonátů (%)
x_3 objemová hmotnost ($g.cm^{-3}$)
x_4 vlhkost v % objemu (%)
x_5 odpor v penetraci ($N.cm^{-3}$)
x_6 pevnost v prostém tlaku (MPa)

Podobným způsobem lze zpracovat ***jednorozměrné lineární regresní modely*** a ***jednorozměrné lineární kalibrace***, kdy je hledán vztah jednou vstupní nezávisle proměnnou a výstupní závisle proměnnou. Dále lze provádět tímto postupem i ***validaci nových analytických metod.***

III. NELINEÁRNÍ REGRESNÍ MODELY

1. Formulace nelineárního regresního modelu

V praxi se pomocí nelineárních regresních modelů řeší řada technických a přírodovědných úloh. Mezi základní patří:

1. Konstrukce kalibračních modelů, jsou-li hodnoty y nelineární odezvou měřícího přístroje na změnu měřené fyzikální veličiny nastavitelné proměnné x.
2. Ověření teoretických modelů popisujících základní fyzikálně-chemické zákonitosti.
3. Tvorba empirických modelů, založená na hledání nelineární závislosti mezi vysvětlovanou proměnnou y a vysvětlujícími proměnnými x.

Podle typu úlohy se volí přístup k procesu tvorby regresního modelu $f(x, \beta)$. Vlastní úloha se formuluje s ohledem na ***regresní triplet***, zadaná data, navržený model a kritérium regrese, ***podobně jako u lineárních modelů***. Jde vlastně o hledání modelu $f(x, \beta)$ na základě zadaných dat a zvoleného kritéria regrese. V řadě případů je model $f(x, \beta)$ znám a regresní úloha je zjednodušena na konstrukci odhadů b parametrů β. Na rozdíl od lineárních regresních modelů mají parametry β v nelineárních regresních modelech většinou rozhodující roli.

Zatímco u lineárních modelů nemívají regresní parametry fyzikální smysl a jsou často pouhými numerickými koeficienty, mají parametry v nelineárních modelech často přesný fyzikální význam. Jejich číselné hodnoty jsou hlavním cílem regresní analýzy, například rovnovážné konstanty (disociační konstanty, součiny rozpustnosti, konstanty stability) reakčních produktů, rychlostní konstanty u kinetických modelů nebo neznámé koncentrace u titračních křivek apod.

Pojmem lineární regresní model se označuje model, který je lineární kombinací modelových parametrů. To znamená, že i lineární model může být z hlediska průběhu modelové funkce nelineární. Např. model $f(x, \beta) = \beta_1 + \beta_2 \sin(x)$ má sinusový průběh a přesto je vzhledem k parametrům β_1 a β_2 lineárním regresním modelem.

Pro lineární regresní modely platí podmínka:

$$g_j = \frac{\partial f(x, \beta)}{\partial \beta_j} = \text{konst.} \qquad j=1,\ldots,m \quad (1)$$

Pokud je aspoň jeden parametr β_j parciální derivace g_j jeho funkcí, jde o ***nelineární regresní model.***

Nelineární regresní modely se **člení *na neseparabilní, separabilní a vnitřně lineární modely***. Neplatí-li podmínka (1) pro žádný parametr modelu, pak se jedná o neseparabilní model. Platí-li podmínka (1) aspoň pro jeden modelový parametr, pak se jedná o separabilní model. Vnitřně lineární modely jsou sice nelineární, ale lze je korektní transformací neboli ***reparametrizací*** převést na lineární regresní model. Reparametrizace znamená transformaci parametrů β do nových γ, které jsou funkčně spjaty s původními parametry β. Vhodnou reparametrizací lze odstranit řadu numerických i statistických problémů regrese. Z neseparabilních modelů lze často reparametrizací získat separabilní modely a naopak. V praxi se vyčleňují ještě ***modely linearizovatelné***, které lze vhodnou

transformací převést na lineární regresní model. Bodové odhady z linearizovaných regresí se pak používají pouze jako výchozí odhady.

K bodovému odhadu regresních parametrů nelineárního modelu se nejčastěji používá metoda nejmenších čtverců, přičemž získané bodové odhady jsou nejlepší nestranné, jestliže za předpokladů platnosti uvedených v předchozím článku v první kapitole účelová funkce nabývá absolutního minima. Hodnota účelové funkce je označována jako globální extrém (minimum), jestliže se pro daná data jedná o nejmenší možnou hodnotu odpovídající bodovému odhadu parametrů z oboru platnosti regresní funkce. Kromě globálního extrému existují rovněž lokální extrémy a obecně není zaručeno, že nalezený extrém je globální. Proces hledání extrému účelové funkce se nazývá optimalizace a nalezené bodové odhady, vyhovující cíli optimalizace, se nazývají optimální odhady.

2. Optimalizační metody

Optimalizační metody se dělí podle způsobu hledání optimálních odhadů na ***derivační*** a ***nederivační (komparativní)***. Derivační metody vyšetřují okolí posledního odhadu prostřednictvím derivace účelové funkce podle parametrů a na základě kvantitativního vyhodnocení se vytváří nový odhad. Při komparativních metodách se parametry mění podle určitých pravidel a oprava se provádí pouze na základě porovnání změny účelové funkce (zlepšení, zhoršení) po poslední změně parametru. Zvláštní skupinu tvoří algoritmy pro speciální případy (jiná kritéria regrese, robustní odhady). Podle toho, kolik parametrů je optimalizováno, dělí se optimalizační metody na ***jednorozměrné*** a ***vícerozměrné***.

Při optimalizaci nastává několik problémů. Prvním problémem je tzv. ***renundance*** neboli přeurčenost modelu. Derivace g_j v rovnici (1) mají význam míry citlivosti parametru β_j v modelu $f(x,\beta)$. Míry citlivosti parametrů slouží především v předběžné analýze nelineárních regresních modelů k posouzení jejich kvality a odhalení renundance (přeurčení), způsobené nadbytečným počtem parametrů. Jde o analogii multikolinearity u lineárních regresních modelů.

Dalším problémem je ***rychlost konvergence*** optimalizační metody, která je spojena se spotřebou resp. nalezením extrému v rozumném čase. Rychlost konvergence souvisí se vzdáleností odhadu od extrému. Ve větší vzdálenosti konvergují lépe komparativní metody, blízko extrému naopak metody derivační.

Zvláštním problémem při optimalizaci jsou tzv. ***vázané extrémy***. V tomto případě je řešení podmiňováno splněním nějakých dalších podmínek, např. fyzikální reálnosti optimalizovaných parametrů (kladné výsledky měření) nebo dokonce možností výpočtu účelové funkce (kladné argumenty logaritmu). Tento problém se při výpočtu řeší použitím ***penalizační funkce*** (při čtení hodnoty značně zhoršující hodnotu účelové funkce) nebo návratového kódu (při nesplnění podmínek se hodnota účelové funkce nevyčíslí a informace o porušení podmínek se předá optimalizujícímu algoritmu, kde je tato situace ošetřena obvykle zmenšením změny odhadu).

Je-li možno model linearizovat, použijí se odhady z lineární regrese. V ostatních případech se použijí odhady získané studiem submodelů nebo vyplývající z fyzikálně-chemického rozboru problému. Kvalitní optimalizační algoritmy nejsou závislé na

počatečních odhadech, přesto je však vhodné se přesvědčit, zda výsledkem optimalizace s různými počátečními odhady jsou stejné výsledné odhady parametrů. Není-li tomu tak, je třeba podrobně analyzovat regresní model a příslušné citlivostní koeficienty. Vhodné je rovněž zkusit jiné optimalizační metody.

Mezi nejefektivnější metody z jednorozměrných optimalizačních metod patří ***Fibonacciho metoda***, která je metodou komparativní. Předpokladem použití této metody je, že v intervalu ve kterém vyšetřujeme extrém, neexistují mimo globálního extrému žádné další (lokální) extrémy. Potom se jedná o problém, jak zvolit počet bodů tak, aby se pro nejmenší počet vyčíslení účelové funkce zúžil počáteční interval na požadovanou délku obsahující hledaný extrém.

Mezi nejjednodušší komparativní vícerozměrné optimalizační metody patří ***metoda s náhodným přírůstkem*** (varianta metody Monte Carlo). Princip spočívá v generování rovnoměrně náhodných čísel v intervalu < 0,1 >, která se používají k odhadu parametrů podle příslušné rovnice. Další komparativní vícerozměrnou optimalizační metodou je ***metoda simplexová.***

Nejjednodušší derivační metodou je ***metoda největšího spádu.*** Po minimalizaci se nové odhady vypočítají podle příslušného vzorce. Metoda pracuje lépe ve větší vzdálenosti od extrému. ***Newtonova metoda*** využívá kvadratickou aproximaci hyperboly účelové funkce nad parametry v okolí odhadu, extrém této kvadratické aproximace je novým odhadem.Čím větší je shoda mezi skutečnou a aproximující hyperplochou, tím rychleji je extrém nalezen, obvykle tomu tak bývá v okolí extrému při použití účelové funkce z metody nejmenších čtverců. ***Levenberg-Marquardtova metoda*** je kombinací obou uvedených derivačních metod. Existují i další, obvykle složitější, ale současně i efektivnější metody, se kterými se lze seznámit i ve specializované literatuře.

3. Regresní diagnostika nelineární regrese

Regresní diagnostika výsledků nelineárních modelů je značně komplikovaná, proto jsou zde uvedeny jen některé aspekty. Především je to statistická analýza reziduí. U lineárních modelů slouží analýza reziduí k ověřování některých předpokladů o chybách, u nelineárních modelů pak především k posouzení dosažené těsnosti proložení vypočtené regresní křivky danými experimentálními body. Analýzou vlivných bodů se identifikují body, které silně ovlivňují odhadované regresní parametry.

3.1 Grafická analýza reziduí

Grafickou (předběžnou) analýzou, která spočívá v prostém zobrazení vektoru reziduí, lze snadno odhalit:

a) odlehlé (extrémní) hodnoty v souboru reziduí
b) trend v reziduích
c) nedostatečné střídání znaménka u reziduí
d) chybný model nebo vzájemnou závislost reziduí
e) heteroskedasticitu (nekonstantnost rozptylu) závisle proměnné veličiny y
f) náhlou změnu podmínek při měření hodnot y

3.2 Statistická (numerická) analýza reziduí

V řadě regresních programů aplikované nelineární regrese v laboratoři je analýza

souboru reziduí hlavní diagnostickou pomůckou při hledání a rozlišení modelu a navíc těsnost dosaženého proložení experimentálními body je mírou věrohodnosti nalezených odhadů. Mezi nejčastěji užívané statistiky patří především ***střední hodnota reziduí***, která by se měla rovnat nule, ***průměrné reziduum*** a ***směrodatná odchylka střední hodnoty reziduí*** a ***koeficient šikmosti*** a ***špičatosti.***

K uvedenému užívání analýzy reziduí je však třeba kriticky poznamenat, že diagnostické využívání není rigorózní, protože rezidua nemají nulovou střední hodnotu, jsou vychýlená a dále jsou přibližně lineární kombinací chyb. Jsou navíc závislá na skutečných hodnotách parametrů. K testování reziduí lze užít všech statistik užívaných u lineárních regresních modelů. Potíže zde činí pouze určení rozdělení testačních statistik, které jsou závislé na nelinearitě modelu.

3.3 Analýza vlivných bodů

U lineárních regresních modelů jsou všechny charakteristiky k odhalení vlivných bodů funkcí reziduí a diagonálních prvků projekční matice. Řada charakteristik vlivných bodů používá odhadů, určených ze všech bodů kromě i-tého.

U nelineárních regresních modelů je situace komplikována tím, že již nelze vyjádřit odhady parametrů a rezidua jako lineární kombinaci experimentálních dat. Použije se linearizace nelineárního modelu, je možné použít přímo všech technik odhalení vlivných bodů v lineárních modelech.

Vlivné body lze snadno identifikovat na základě ***Jackknife reziduí***. K vyjádření vlivu jednotlivých bodů na odhady parametrů lze použít i kvadratického rozvoje regresního modelu a vyčíslovat změny vektoru vychýlení při vynechání i-tého bodu nebo změny střední hodnoty i-tého rezidua při vynechání i-tého bodu. Mezi nelineární míry vlivu i-tého bodu na odhady parametrů patří ***věrohodnostní vzdálenost*** LD_i.

4. Hledání nejlepšího nelineárního regresního modelu

Kvalita nelineárního modelu se posuzuje podobně jako u lineárního s ohledem na následující kritéria:

a) Kvalita nalezených odhadů

Kvalita nalezených odhadů parametrů se posuzuje podle intervalu spolehlivosti parametrů, popř. podle testů hypotézy. Jsou-li intervaly spolehlivosti příliš široké, nebo některé parametry jsou statisticky nevýznamné, znamená to, že model je blízký předurčenému modelu a je tedy příliš složitý (vysoká korelace mezi parametry).

b) Kvalita dosažené těsnosti proložení

Kvalitu dosažené těsnosti proložení lze posoudit prostřednictvím ***koeficientu determinace, residuální směrodatné odchylky*** a ***regresní diagnostiky***. Stonásobek koeficientu determinace se nazývá ***regresní rabat***. K rozlišení těsnosti proložení mezi dvěma modely bylo navrženo ***Akaikovo kritérium***, které pro optimální model dosahuje minimální hodnoty.

c) Predikční schopnost modelu

Kritéria posouzení predikční schopnosti modelu vycházejí z rozdělení bodů do

dvou podskupin. Predikční schopnost modelu se pak vyjadřuje kritériem K, příslušný vzorec lze nalézt v odborné literatuře (6). Predikční schopnost modelu je tím vyšší, čím víc se hodnota K blíží k jedničce.

d) Kvalita experimentálních dat

Kvalita experimentálních dat je určována z regresní diagnostiky. K posouzení kvality experimentálních dat se užívá analýzy reziduí a vlivných bodů.

e) Správnost navrženého modelu

Pro správnost navrženého modelu navrhl White test C. Vzorec lze nalézt v odborné literatuře (6). Pro správný model je C=0.

f) Souhlas s požadavky fyzikálního smyslu

Důležitou podmínkou je fyzikální smysl nalezených odhadů. Modely, které nesplňují tuto podmínku jsou nepřijatelné, i když budou preferovány na základě statistických kritérií.

5. Příklady řešení nelineárního regresního modelu

5.1 Nelineární regresní model vycházející z fyzikálních úvah:

Propustnost zemin testována na propustoměru s řízeným stupněm sycení. Propustnost materiálů lze charakterizovat koeficientem filtrace **k**, přičemž platí vztah

$$\mathbf{k = v/i}$$

kde i - je hydraulický spád [1]
v - filtrační rychlost způsobená spádem i [m/s]

Hydraulický spád je definován vztahem

$$\mathbf{i = h_w/h}$$

kde h_w - je výška vodního sloupce [m]
h - výška vzorku [m]

Filtrační rychlost je definována rovnicí

$$\mathbf{v = V_w/A/t}$$

kde V_w - je objem vody, který v čase t proteče průřezem zkušebního tělesa [m^3 resp. $ml.10^{-6}$]
A - průřez zkušebního tělesa [m^2]

Na základě výše uvedeného je model pro výpočet koeficientu filtrace následující:

$$k = \frac{V_w \cdot h}{t \cdot h_w \cdot A} \quad [m.s^{-1}]$$

5.2 Model empirický

5.2.1 Zadání:

V uhlí je stanovován obsah popela a obsah beryllia. Mezi obsahem popela v uhlí a obsahem beryllia přepočteném na obsah beryllia v popelu uhlí, byl hledán empirický model.

5.2.2 Řešení:

Tvorba regresního modelu je demonstrována na výše uvedeném empirickém modelu, který řeší vztah mezi obsahem popela v uhlí a obsahem beryllia v popelu uhlí. Výstupní závisle proměnná **y** je obsah beryllia v popelu uhlí a vstupní nezávisle proměnná **x** je obsah popela. Při hledání závislosti mezi x a y bylo navrženo několik modelů, které byly testovány softwarem ADSTAT 2.0, modelem nelineární regrese. V řešení jsou uvedeny podmínky, vybraný testovaný model, vstupní data a graf vybraného regresního modelu. Výsledný model byl vybrán ve smyslu kritérií pro výběr vhodného regresního modelu.

V S T U P

Ve vstupních podmínkách (1) je zadána hladina významnosti, informace o tom, které parametry chceme zkonstantnit, počet iterací, tolerance pro parametry a tolerance pro součet čtverců. Poslední tři podmínky mají význam pro ukončení výpočtu. V tabulce (2) je uveden zadaný regresní model a v tabulce (3) vstupní data.

(1) PODMÍNKY:

Hladina významnosti, alfa	0.050
Počet bodů, n	24
Počet parametrů, m	2
Počet nezávislých proměnných	1
Minimální změna RSC [%]	1.000000E-01
Minimální změna parametrů [%]	0.000000E+00
Maximální počet iterací	150
Kvantil Studentova rozdělení t(1-alfa/2,n-m)	2.074
Kvantil Fisher-Snedecorova rozdělení F(1-alfa,n,n-m)	2.063
Kvantil Chi^2 rozdělení Chi^2(1-alfa,m)	5.991

(2) REGRESNÍ FUNKCE A POČÁTEČNÍ ODHADY PARAMETRŮ:

Regresní funkce:p[1]+p[2]*(1/x1)

p[1] :p 1.000000E+00 p[2] :p 1.000000E+00

(3) VSTUPNÍ DATA:

Celá matice editoru.

Poř.č.	1	2	3	4	5
x	4.5800E+00	3.3500E+00	2.5000E+00	3.2400E+00	5.1300E+00
y	9.1700E+00	1.2540E+01	1.6400E+01	1.2350E+01	7.9900E+00
Pořad.č.	**6**	**7**	**8**	**9**	**10**
x	8.1800E+00	8.9780E+01	4.5230E+01	5.5700E+00	2.9600E+01
y	5.0100E+00	4.6000E-01	9.3000E-01	7.5400E+00	1.3900E+00
Pořad.č.	**11**	**12**	**13**	**14**	**15**
x	1.7400E+01	1.4540E+01	3.7800E+00	6.6700E+00	8.7300E+00
y	2.3000E+00	2.0300E+00	1.0580E+01	6.3000E+00	4.7000E+00
Pořad.č.	**16**	**17**	**18**	**19**	**20**
x	5.7780E+01	2.1900E+00	1.8300E+00	2.3000E+00	2.1390E+01
y	6.9000E-01	1.8260E+01	2.5140E+01	1.7390E+01	1.9200E+00
Pořad.č.	**21**	**22**	**23**	**24**	
x	5.3000E+00	1.5800E+00	8.9140E+01	7.6220E+01	
y	8.4900E+00	2.8480E+01	4.7000E-01	5.4000E-01	

VÝSTUP:

Ve výstupní tabulce (1) a (2) jsou uvedeny bodové a intervalové odhady parametrů s pološířkami asymptotických intervalů spolehlivosti, vychýlení odhadu a relativní vychýlení odhadu. Pokud je relativní vychýlení menší než 3%, lze tyto intervaly akceptovat.

(1) BODOVÉ ODHADY PARAMETRŮ:

Parametr	Bodový odhad	Směrodatná odchylka	Absolutní Vychýlení	Relativní vychýlení[%]
p[1]	-2.997429E-01	2.092577E-01	2.733299E-11	-9.118814E-09
p[2]	4.345761E+01	7.799269E-01	1.996661E-10	4.594503E-10

(2) INTERVALOVÉ ODHADY PARAMETRŮ:

Parametr	Bodový odhad	Poloviční délka konfidenčního int. spočtená z	
		délky poloos	maxim
p[1]	-2.997429E-01	+- 4.214351E-01	+- 5.491454E-01
p[2]	4.345761E+01	+- 2.045441E+00	+- 2.046727E+00

V tabulce (3) je uvedena korelační matice s prvky, které odpovídají párovým korelačním koeficientům mezi odhady parametrů.

(3) KORELAČNÍ MATICE ODHADŮ:

	x[1,i]	x[2,i]
x[1,i]	1.0000E+00	-7.4423E-01
x[2,i]	-7.4423E-01	1.0000E+00

V tabulce (4) jsou uvedeny statistické charakteristiky regrese, zahrnující reziduální součet čtverců, regresní rabat (tj. stonásobek čtverce vícenásobného korelačního koeficientu) a Akaikeho informační kritérium. Jak již bylo uvedeno v kapitole 4, slouží Akaikovo kritétium k rozlišení těsnosti proložení mezi dvěma modely. Pro optimální model dosahuje AIC minimální hodnoty.

(4) STATISTICKÉ CHARAKTERISTIKY REGRESE:

Reziduální součet čtverců, RSC	1.031442E+01
Regresní rabat, D^2 [%]	9.929639E+01
Akaikeho informační kriterium, AIC	-1.626826E+01

Dále je statistickým software provedena analýza reziduí podobně jako u předchozího příkladu lineární regrese, směrodatné odchylky měření, první čtyři momenty reziduí, reziduální součet čtverců a statistiky reziduí. Výsledky analýzy nejsou v předloženém článku uvedeny. V rámci statistického testování je provedeno hledání vlivných bodů, podobně jako u lineární regrese. Vlivné body nebyly zjištěny, proto je výsledný číselný soubor v článku vypuštěn.

5.2.3 Grafický výstup

Grafický výstup software Adstat 2.0 předkládá 2 grafy. Na obr. 1 je zobrazena závislost reziduí na proměnné y_{vyp}. Na obr. 2 je nakreslen průběh regresního modelu spolu s konfidenčními pásy a zadanými daty. Mírně rozšířený konfidenční pás pro hodnoty x 30-80% popela je způsoben menším počtem experimentálních bodů.

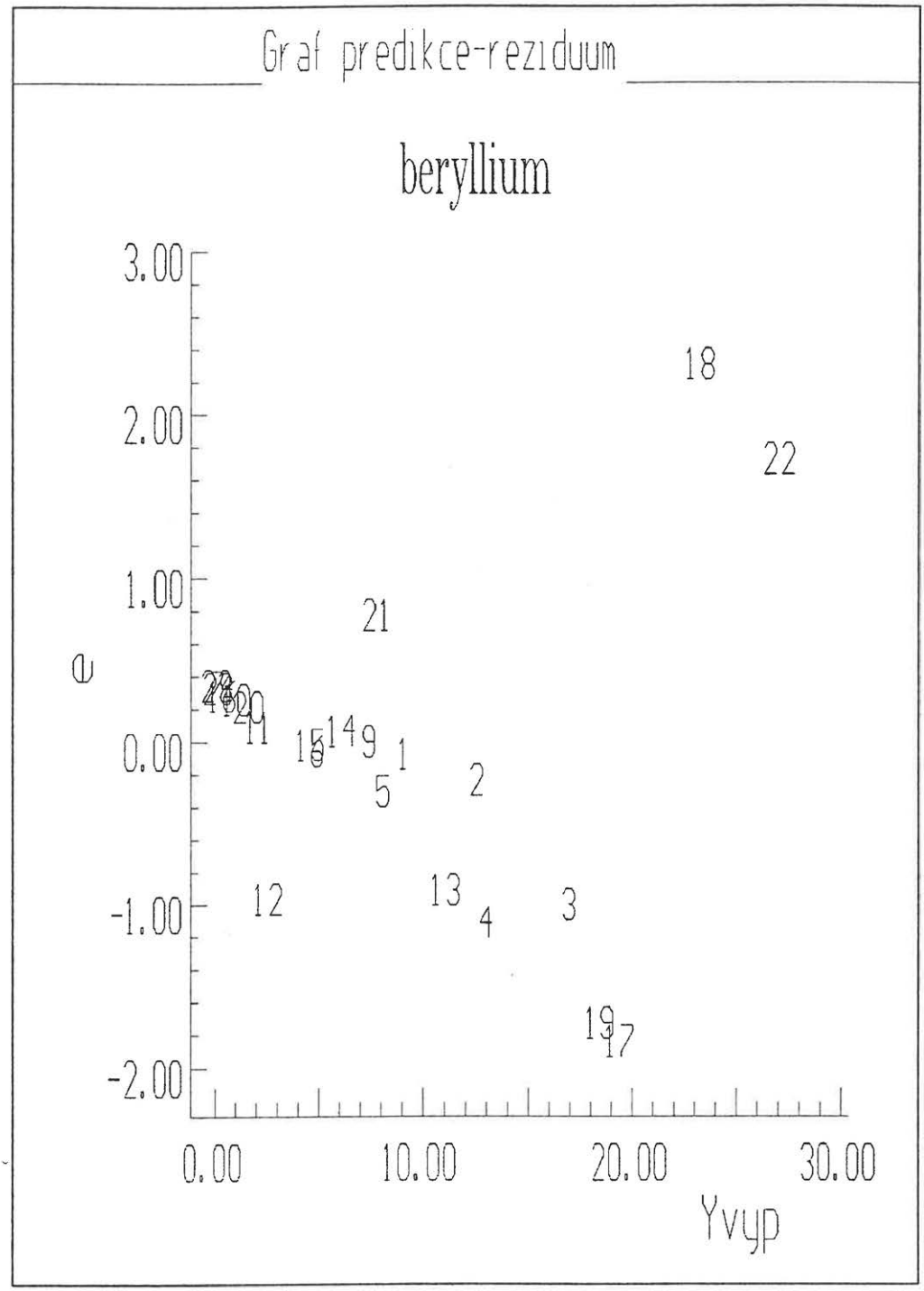
5.2.4 ZÁVĚR

Jako ukázka nelineární regrese byly uvedeny dva příklady. První příklad vychází z fyzikálních úvah pro hodnocení propustnosti zemin. Výsledkem je vzorec pro výpočet koeficientu filtrace.

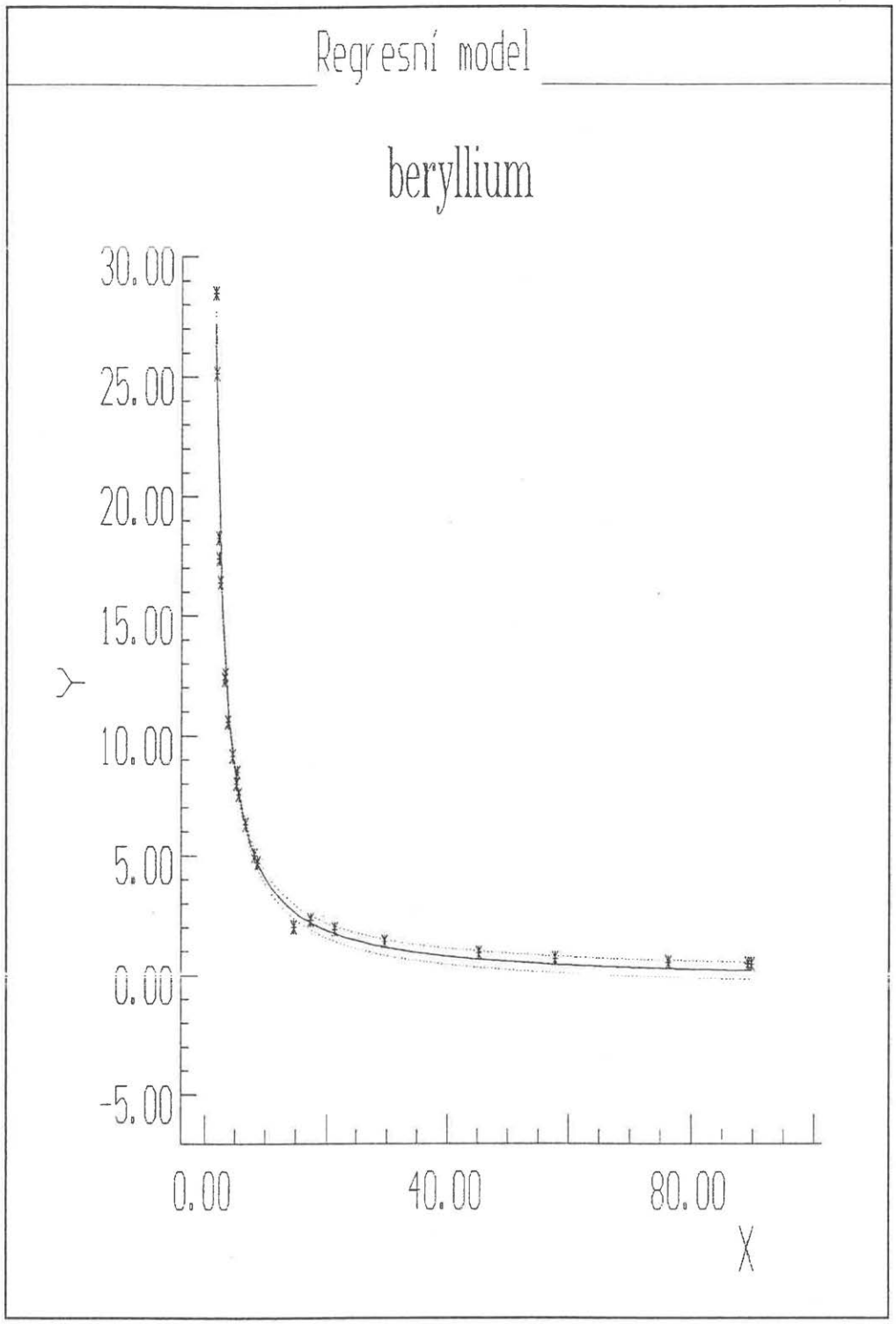
Druhý příklad je empirický. Cílem bylo nalézt matematický vztah mezi obsahem popela a obsahem beryllia přepočteném na obsah beryllia v popelu uhlí. Vyčíslený regresní vztah má tvar:

$y = -0.2997 + 4.3458/x,$

kde x je obsah popela v uhlí
y obsah beryllia v popelu uhlí



Obr. 1 Graf predikce rezidua



Obr. 2 Regresní model

IV. ANALÝZA ROZPTYLU

1. Základní pojmy analýzy rozptylu

Analýza rozptylu, označovaná také jako *ANOVA*, se v technické praxi používá buď jako samostatná technika nebo jako postup umožňující analýzu zdrojů variability u lineárních statistických modelů. Ze statistického hlediska lze analýzu rozptylu chápat jako speciální případ regresní analýzy, kde vysvětlující proměnné mají pouze binární charakter a mohou nabývat hodnot jen 0 nebo 1.

V předloženém článku je ANOVA použita jako samostatná technika, která umožňuje posouzení významnosti jednotlivých zdrojů variability v datech. V technické praxi se ANOVA uplatňuje v úlohách:

a) Určení významnosti způsobu přípravy vzorku na výsledek analýzy resp. experimentu.
b) Určení vlivu typu přístroje, lidského faktoru a obsluhy na výsledek měření.
c) Zpracování různých mezilaboratorních experimentů a určení významnosti rozdílů mezi laboratořemi na výsledek analýz.
d) Zpracování plánovaných experimentů, u kterých se systematicky sleduje vliv rozličných faktorů (teploty, času, koncentrace a dalších) na výsledek reakce či analýzy.

Podstatou analýzy rozptylu je rozklad celkového rozptylu dat na složky objasněné (známé zdroje variability) a složku neobjasněnou, o níž se předpokládá, že je náhodná.

Terminologie ANOVY je poněkud speciální. Jsou rozlišovány ***kvalitativní a kvantitativní faktory.*** Za kvalitativní faktor lze například považovat způsob přípravy vzorku. Kvantitativní faktor je například velikost upraveného vzorku nebo další fyzikální a chemické veličiny. Jednotlivé faktory se vyskytují na určitých ***úrovních***, které se označují jako ***zpracování***. Tyto úrovně mohou být opět kvantitativní nebo kvalitativní. Zdrojem variability výsledků měření jsou jednotlivé úrovně faktoru.

Účelem analýzy rozptylu je testování shody jednotlivých úrovní, neboli nulové hypotézy. Pokud jsou zjišťovány pouze rozdíly mezi danými úrovněmi, např. mezi způsoby přípravy vzorku, jde o ***modely s pevnými efekty***. Pokud jsou jednotlivé úrovně pouze výběrem z konečného či nekonečného souboru, jde o ***modely s náhodnými efekty***. O model s pevnými efekty půjde, např. bude-li vyšetřován vliv přípravy vzorku na výsledek analýzy. O model s náhodnými efekty jde tehdy, je-li např. zjišťován vliv průměrné velikosti částic vzorku na výsledek analýzy.

Pro volbu typu faktorů je možné použít následujících orientačních pravidel:

a) typické faktory, které odpovídají pevným faktorům, jsou obvykle způsob zpracování, typ měřícího přístroje, analytická metoda, druh chemikálie, typ suroviny apod.
b) typické faktory, které odpovídají náhodným faktorům, jsou laboratoře, pracovníci, dny apod.

Je-li sledován pouze jeden faktor, pak jde o ***jednofaktorovou analýzu rozptylu*** neboli třídění podle jednoho faktoru. Často se však sleduje i vliv několika faktorů, pak jde o ***vícefaktorovou analýzu rozptylu.***

Pro vlastní zpracování modelů analýzy rozptylu je důležité, zda je při všech kombinacích faktorů realizován stejný počet měření, resp. opakování, či nikoliv. Kombinace úrovní jednotlivých faktorů se označuje jako ***cela***. Pro stejný počet opakování se ***experimenty*** označují jako ***vyvážené*** a pro nestejný počet opakování jako ***nevyvážené.*** Z hlediska jednoduchosti zpracování a interpretace je výhodnější pracovat s vyváženými experimenty.

2. Jednofaktorová analýza rozptylu

V nejjednodušším případě se předpokládá, že na náhodnou veličinu má vliv pouze jeden faktor. Cílem jednoduchého třídění je především testování, zda jsou efekty α_i nulové, tedy zda jednotlivé úrovně daného faktoru vedou ke statisticky nevýznamným rozdílům ve výsledcích, což znamená, že je zkoumán vliv faktoru na výsledek experimentu. Nulová hypotéza H_0: $\alpha_i = 0$, i=1,..., K, se ověřuje proti alternativní hypotéze H_A: $\alpha_i \neq 0$, i = 1,..., K. Pro posouzení významnosti efektu se využívá testovací charakteristika F_e, která je porovnávána s kvantilem $F_{1-\alpha}$. Vyjde-li F_e větší než kvantil, je nutné na hladině významnosti α nulovou hypotézu H_0 vyloučit a efekty považovat za nenulové, neboli významné.

Vlastní analýza rozptylu závisí na tom, zda jde o modely s pevnými nebo náhodnými efekty. U modelů s náhodnými efekty se může projevit porušení předpokladu normality u náhodných veličin. Indikaci normality lze provádět s využitím rankitových grafů pro rezidua. Podobně jako u modelů s pevnými efekty lze i u modelů s náhodnými efekty zlepšovat normalitu vhodnou transformací dat. Problémem je však hodnocení rozptylů v transformovaném měřítku. Pokud není splněn předpoklad normality, je možné použít k určení odhadu rozptylu a testování jeho významnosti techniku Jackknife.

3. Dvoufaktorová analýza rozptylu

Při dvoufaktorové analýze rozptylu se provádí experimenty na různých úrovních dvou faktorů, A a B. Kombinace úrovní faktorů tvoří typickou mřížkovou strukturu, jejímž elementem je tzv. ***cela***. V každé cele je obecně n_{ij} opakování. Často se lze setkat s případem bez opakování, kdy je v každé cele pouze jediné opakování. Kromě řádkových a sloupcových efektů se zde vyskytuje také interakční člen, který je důsledkem různých kombinací sloupcových a řádkových efektů.

Je-li v cele jedno pozorování, je výhodné použití ***Tukeyova modelu interakce***. Složitější jsou ***řádkově lineární*** modely interakcí nebo ***sloupcově lineární.*** Kompletnější je ***aditivně-multiplikativní*** model. Tukeyho model interakce využívá statistik F_{AB}, F_A a F_B. Testuje se nulová hypotéza proti alternativní.

V chemické praxi je nejužívanější ***model s pevnými efekty.*** Umožňuje určit vliv dvou faktorů na výsledek chemické analýzy. Podle počtu opakování v jednotlivých celách lze úlohy dvojného třídění rozlišit. Je-li v každé cele jedna hodnota, jde o ***modely pro případ bez opakování měření***. Při testování je provedena úplná analýza rozptylu s interakcí Tukeyova typu. Pro odhad parametrů je možné použít buď vztahů vycházejících z předpokladů normality nebo robustní techniky mediánového vyhlazení. Kromě testů významnosti se určuje vhodnost mocninné transformace pro eliminaci neaditivity a ověřuje se předpoklad normality. Indikace aditivity se provádí pomocí grafu.

Tvoří-li body náhodný obrazec, pak se jedná o aditivitu, je-li obrazec přibližně lineární, pak se jedná o eliminovatelnou neaditivitu. Vhodná mocninná transformace pro eliminaci neaditivity se určí ze směrnice regresní přímky.

Je-li ve všech celách stejný počet opakování, jedná se o ***vyvážený model.*** Při testování nulové hypotézy proti alternativní se opět využívá statistik F_{AB}, F_A a F_B. Je hledán optimální model analýzy rozptylu, odhadovány jeho parametry a provedeny testy významnosti. Je ověřována vhodnost transformace dat a normality reziduí. Výhodou vyvážených experimentů je, že jednotlivé složky modelů analýzy rozptylu jsou vzájemné nezávislé.

Pro ***nevyvážené modely*** obecně platí, že v každé cele je jiný počet opakování. Pro nevyvážené experimenty je rovněž hledán optimální model analýzy rozptylu, odhadnuty jeho parametry a provedeny testy významnosti. Je ověřována vhodnost logaritmické transformace a normalita reziduí.

4. Příklady ANOVY

4.1 Jednofaktorová analýza: Vliv teploty na délkovou změnu pálením cihlářských surovin

Zadání: Při teplotě 950, 1000 a 1050°C byla sledována délková změna pálením zkušebních tělísek vytvořených z těsta cihlářské suroviny. Na hladině významnosti $\alpha = 0.05$ je ověřováno, zda má teplota výpalu vliv na délkovou změnu tělísek.

Řešení: Data byla zpracována jednofaktorovou analýzou rozptylu s pevnými efekty.

V S T U P

(1) DATA A PODMÍNKY:

Hladina významnosti alfa	0.050
Transformace	Ne
Počet úrovní faktoru A, k	3
Cel. Vel. n = n[1] + n[2] + ... + n[k]	27

Ve vstupních podmínkách (1) je zadána hladina významnosti, která se používá při všech testech. Dále se zadává zda budou data logaritmicky transformována či nikoliv. Potřeba logaritmické transformace vychází na základě grafu závislosti směrodatné odchylky na průměru. Počet úrovní znamená v předloženém příkladě tři různé teploty. Celkový počet naměřených hodnot je 27. V následující tabulce (2) je uvedena celá matice, kde jsou data seřazena podle jednotlivých úrovní. V prvním řádku je uveden maximální počet opakování, tj. 9 pro každou úroveň, a počet úrovní = počet sloupců.

(2) OPAKOVANÁ POZOROVÁNÍ NA RŮZNÝCH ÚROVNÍCH FAKTORU A:

Celá matice editoru.

Opakování	Úrovně efektu A Velikost		
	n[1] = 9	n[2] = 9	n[3] = 9
1	1.3000E+00	6.1000E+00	9.8000E+00
2	3.1000E+00	7.8000E+00	1.0300E+01
3	1.6000E+00	7.3000E+00	9.9000E+00
4	6.4000E+00	8.4000E+00	1.1200E+01
5	4.2000E+00	5.5000E+00	1.1200E+01
6	3.6000E+00	7.9000E+00	1.0700E+01
7	6.1000E+00	7.5000E+00	1.0800E+01
8	6.5000E+00	8.0000E+00	1.0200E+01
9	5.0000E+00	8.7000E+00	1.0900E+01

V Ý S T U P

(1) PRŮMĚRY A EFEKTY ÚROVNÍ:

Celkový průměr = 7.4074E+00

Reziduální rozptyl = 1.7534E+00

Úroveň	Průměr	Efekt	H_{ii}
1	4.2000E+00	-3.2074E+00	1.1111E-01
2	7.4667E+00	5.9259E-02	1.1111E-01
3	1.0556E+01	3.1481E+00	1.1111E-01

V odstavci (1) výstupu software Adstat 2.0 jsou uvedeny odhady parametrů jednotlivých úrovní. V dalším odstavci (2) jsou uvedeny výsledky analýzy vlivných bodů z Jakknife reziduí a extrémů z prvků projekční matice H_{ii}.

(2) DETEKCE VLIVNÝCH BODŮ (vybočující a odlehlé body)

	Odlehlé	Vybočující
Závěr	Žádné	Žádné

Z tabulky analýzy rozptylu (3) je zřejmé, že nulová hypotéza byla zamítnuta, efekty lze považovat za nenulové, neboli významné. Teplota výpalu má tedy vliv na délkovou změnu vypalovaných zkušebních tělísek. Stejné závěry lze učinit na základě Scheffeho procedury, viz tab (4).

(3) TABULKA ANOVA:

H_0: Efekty faktoru A jsou nulové, H_A: ... nejsou nulové

Kvantil F(1-alfa,k-1,n-k) = 3.403

Zdroj rozptylu	Stupně volnosti	Součet čtverců	Průměrný čtverec	Testovací kriterium	Závěr H_0 je	Spočtená hlad.výz.
mezi úrovněmi	k-1 = 2	1.8182E+02	9.0908E+01	51.846	Zamítnuta	0.000
rezidua	n-k = 24	4.2082E+01	1.7534E+00	-	-	-
celkový	n-1 = 26	2.2390E+02	8.6115E+00	-	-	-

(4) VÍCENÁSOBNÁ POROVNÁNÍ (Sheffeho procedura):

Hypotéza H_0	Průměrný párový rozdíl	Meze konfidenčního intervalu		Závěr
		dolní	**horní**	
P 1 = P 2	-3.267	-4.895	-1.638	Zamítnuta
P 1 = P 3	-6.356	-7.984	-4.727	Zamítnuta
P 2 = P 3	-3.089	-4.717	-1.460	Zamítnuta

(5) ZKOUŠKA TRANSFORMACE:

Korelační koeficientent R : -0.998

(Pokud je R blízké nule není transformace nutná.)

Protože se R blíží jedné, byla provedena logaritmická transformace. Výsledek je uveden v tabulce (6). Testovací kritérium nabývá hodnoty 20.624 a je tedy vyšší než kvantil $F_{1-\alpha}(K-1,N-1)$, který nabývá hodnoty 3.403. Závěry jsou shodné jako pro zpracování dat bez transformace.

(6) TABULKA ANOVA PO TRANSFORMACI:

H_0: Efekty faktoru A jsou nulové, H_A: ... nejsou nulové

Kvantil F(1-alfa,k-1,n-k) = 3.403

Zdroj rozptylu	Stupně volnosti	Součet čtverců	Průměrný čtverec	Testovací kriterium	Závěr H_0 je	Spočtená hlad.výz.
mezi úrovněmi	k-1 = 2	5.1635E+00	2.5818E+00	20.624	Zamítnuta	0.000
rezidua	a-k = 24	3.0044E+00	1.2518E-01	-	-	-
celkový	a-1 = 26	8.1679E+00	3.1415E-01	-	-	-

Závěr: Testovací kritérium F_e = 51.846 je vyšší než kvantil $F_{1-\alpha}(K-1,N-1)$ = 3.403, nulovou hypotézu H_0 lze zamítnout, teplota výpalu má tedy vliv na délkovou změnu vypalovaných zkušebních tělísek.

Při testovaní nebyly zjištěny vybočující a odlehlé body. Přiložený Q-Q graf na obr.1 je konstruován na základě Jackknife reziduí pro ověření jejich normality. Přibližně lineární závislost se považuje za potvrzení normality. Ve výstupu v Adstatu v bodě (5) je provedena zkouška transformace. Je-li R blízké nule, pak není transformace nutná. Protože se R blíží jedné, byla provedena logaritmická transformace, bod (6). Protože testovací kritérium F_e je vyšší než kvantil $F_{1-\alpha}(K-1,N-1)$, jsou tedy závěry shodné se závěry uvedenými v předchozím odstavci.

4.2 Dvoufaktorová analýza - ANOVA pro dvojné třídění s pevnými efekty se stejným počtem pozorování: Vliv laboratoře a metody na stanovení obsahu síry v uhlí

Zadání: Byl vyšetřován obsah veškeré síry v uhlí dvěma nezávislými metodami v 5ti laboratořích. Každé měření bylo dvakrát opakováno. Na hladině významnosti α = 0.05 je vyšetřováno, zda obsah síry v uhlí je ovlivněn analytickou metodou (faktor A) nebo laboratoří (faktor B), kde byla analýza provedena.

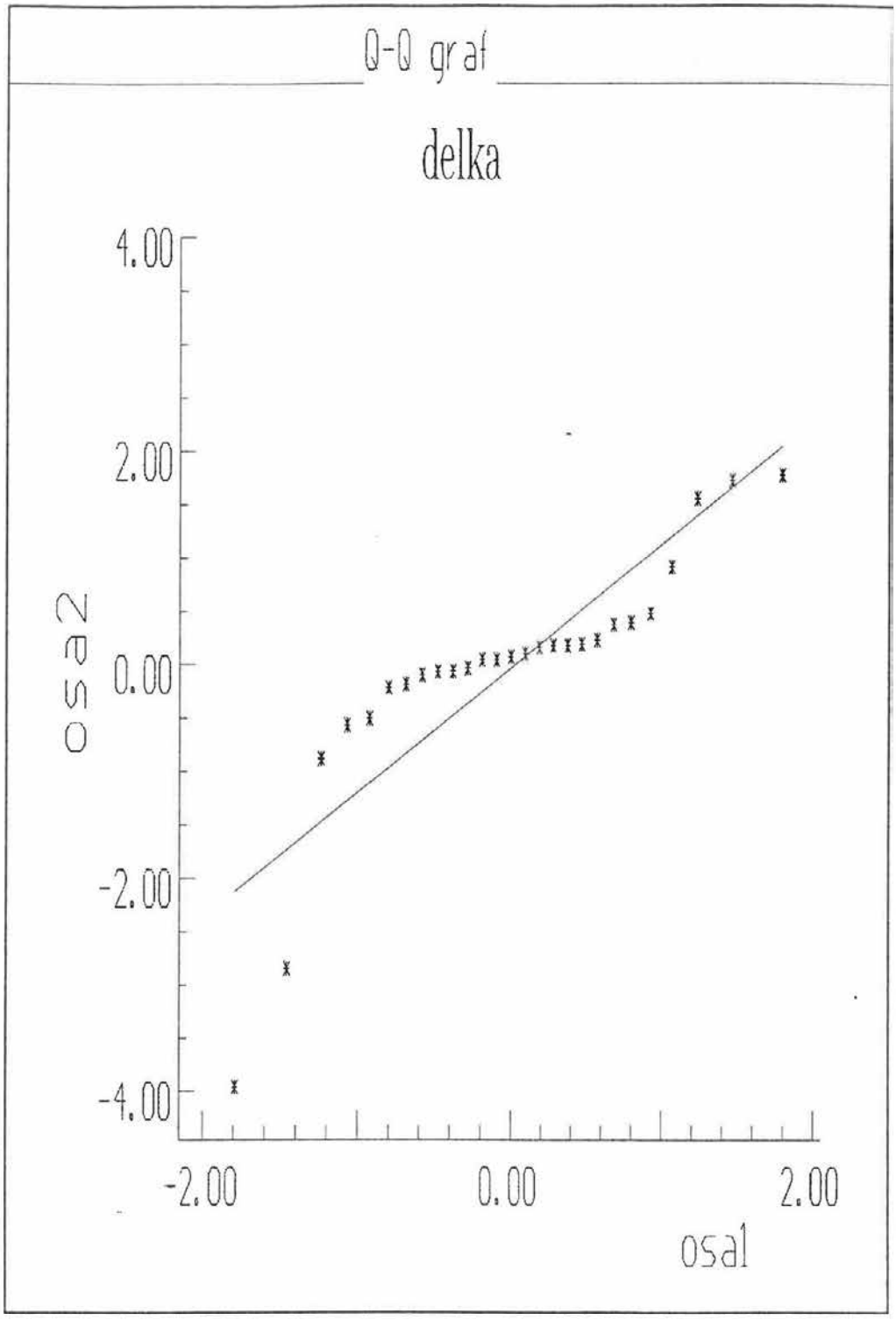
Řešení:

V S T U P

(1) DATA A PODMÍNKY:

Hladina významnosti alfa	0.050
Transformace	Ne
Počet úrovní parametru A, n	2
Počet úrovní parametrů B, m	7
Počet opakování v jedné buňce, o	2

Ve vstupních podmínkách (1) je zadána hladina významnosti stejně jako v předchozím příkladě. Dále se zadává zda budou data logaritmicky transformována či nikoliv. Potřeba logaritmické transformace vychází na základě grafu závislosti směrodatné odchylky na průměru. Počet úrovní znamená v předloženém příkladě tři různé teploty. Celkový počet naměřených hodnot je 27. V následující tabulce (2) je uvedena celá matice, kde jsou data seřazena podle jednotlivých úrovní.



OBR. 1

(2) MĚŘENÍ NA RŮZNÝCH ÚROVNÍCH FAKTORŮ A ,B:

Celá matice editoru. (pro každou úroveň faktorů A a B je měření 2 krát opakováno)

Úrovně Faktoru A	Úrovně faktoru B						
	1	2	3	4	5	6	7
1	1.0700E-01	1.2700E-01	1.1500E-01	1.0800E-01	9.7000E-02	1.1400E-01	1.5500E-01
	1.0500E-01	1.2200E-01	1.1200E-01	1.0800E-01	9.6000E-02	1.1900E-01	1.4500E-01
2	1.0500E-01	1.2700E-01	1.0900E-01	1.1700E-01	1.1000E-01	1.1600E-01	1.6400E-01
	1.0300E-01	1.2400E-01	1.1100E-01	1.1500E-01	9.7000E-02	1.2200E-01	1.6000E-01

V Ý S T U P

(1) PRŮMĚRY A ÚROVNĚ EFEKTŮ:

Celkový průměr = 1.1821E-01

Reziduální rozptyl = 1.4500E-05

FAKTOR A:			FAKTOR B:		
Úroveň	Průměr	Efekt	Úroveň	Průměr	Efekt
1	1.1643E-01	-1.7857E-03	1	1.0500E-01	-1.3214E-02
2	1.2000E-01	1.7857E-03	2	1.2500E-01	6.7857E-03
			3	1.1175E-01	-6.4643E-03
			4	1.1200E-01	-6.2143E-03
			5	1.0000E-01	-1.8214E-02
			6	1.1775E-01	-4.6429E-04
			7	1.5600E-01	3.7786E-02

Ve výsledkovém souboru v tabulce (1) jsou obsaženy odhady řádkových efektů (alfa) a sloupcových efektů (beta), celkový průměr a reziduální rozptyl. V tabulce (2) je uvedena analýza rozptylu. Z výsledků je zřejmé, že testovací kritérium pro faktor A má hodnotu 6.158 a tedy je větší než kvantil F, který nabývá hodnoty 4.600. Nulová hypotéza je tedy zamítnuta. Lze tedy konstatovat, že faktor A je statisticky významný a analytická metoda má vliv na stanovení obsahu síry v uhlí. Testovací kritérium pro faktor B nabývá hodnoty 94.750 a rovněž větší než kvantil F, který má hodnotu 2.848. Faktor B je tedy rovněž statisticky významný a laboratoř má vliv na stanovení obsahu síry. Interakce obou vyšetřovaných faktorů A a B je naopak statisticky nevýznamná, protože testovací kritérium je menší než kvantil.

(2) TABULKA ANOVA PRO MODEL S INTERAKCEMI FAKTORŮ A, B:

H_0: Efekty faktoru A jsou nulové, H_A: ... nejsou nulové

Kvantil F(1-alfa,n-1,mn(o-1) = 4.600

H_0: Efekty faktoru B jsou nulové, H_A: ... nejsou nulové

Kvantil F(1-alfa,m-1,mn(o-1) = 2.848

H_0: Interakce I je nulová, H_A: ... není nulová

Kvantil F(1-alfa,(n-1)(m-1),nm(o-1)) = 2.848

(Zde I znamená efekty interakcí A a B dohromady)

Zdroj rozptylu	Stupně volnosti	Součet čtverců	Průměrný čtverec	Testovací kriterium	Závěr H_0 je	Spočtená hlad.výz.
mezi úrovněmi A	n-1 = 1	8.9286E-05	8.9286E-05	6.158	Zamítnuta	0.026
mezi úrovněmi B	m-1 = 6	8.2432E-03	1.3739E-03	94.750	Zamítnuta	0.000
interakce	(n-1)(m-1)=6	1.9121E-04	3.1869E-05	2.198	Akceptována	0.105
rezidua	mn(o-1) = 49	2.0300E-04	1.4500E-05			
celkový	mno-1 = 27	8.7267E-03	3.2321E-04			

Z tabulky (3) je zřejmé, že korelační koeficient se blíží nule a logaritmickou transformaci není nutné provádět.

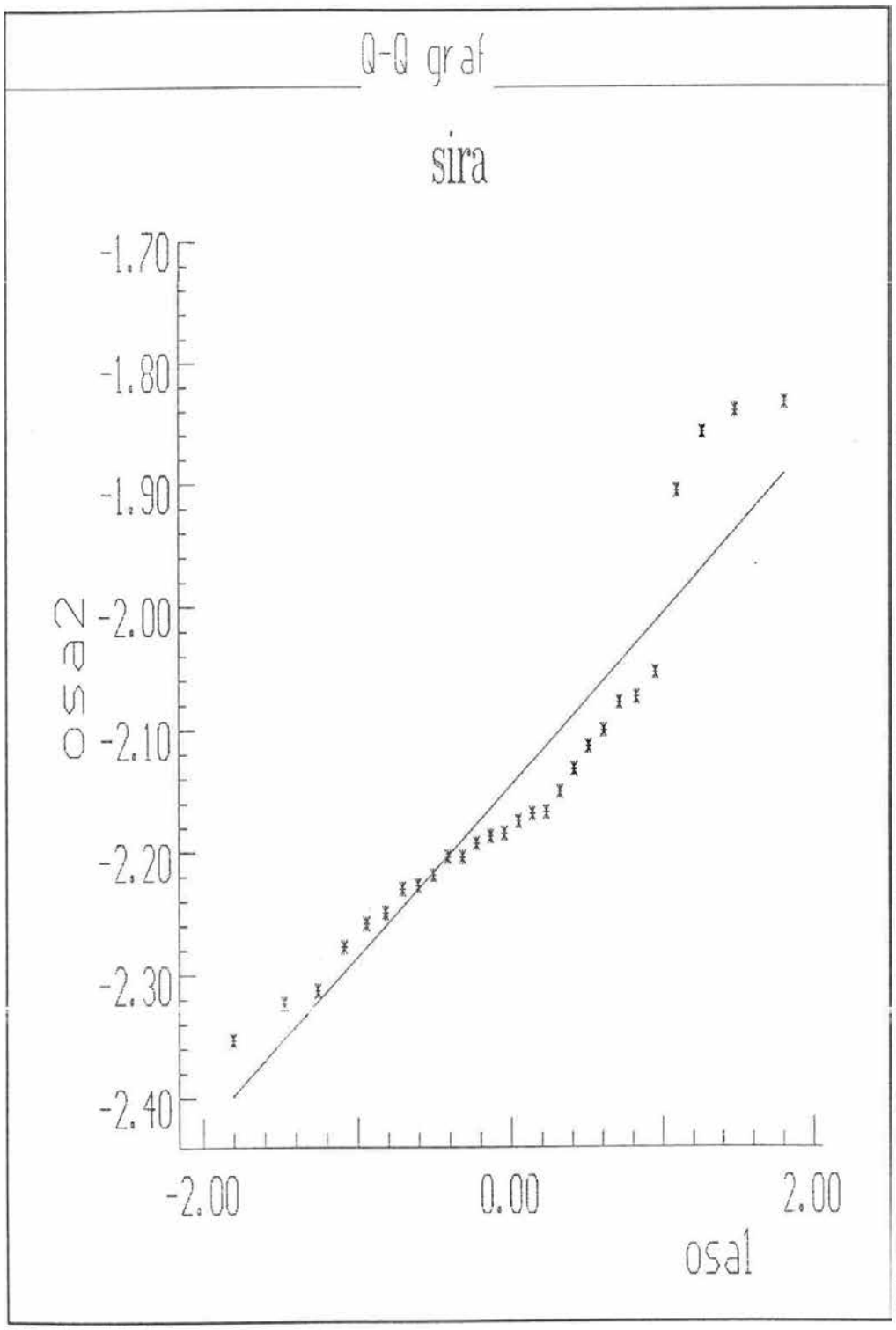
(3) ZKOUŠKA TRANSFORMACE:

Korelační koeficient, R : 0.287

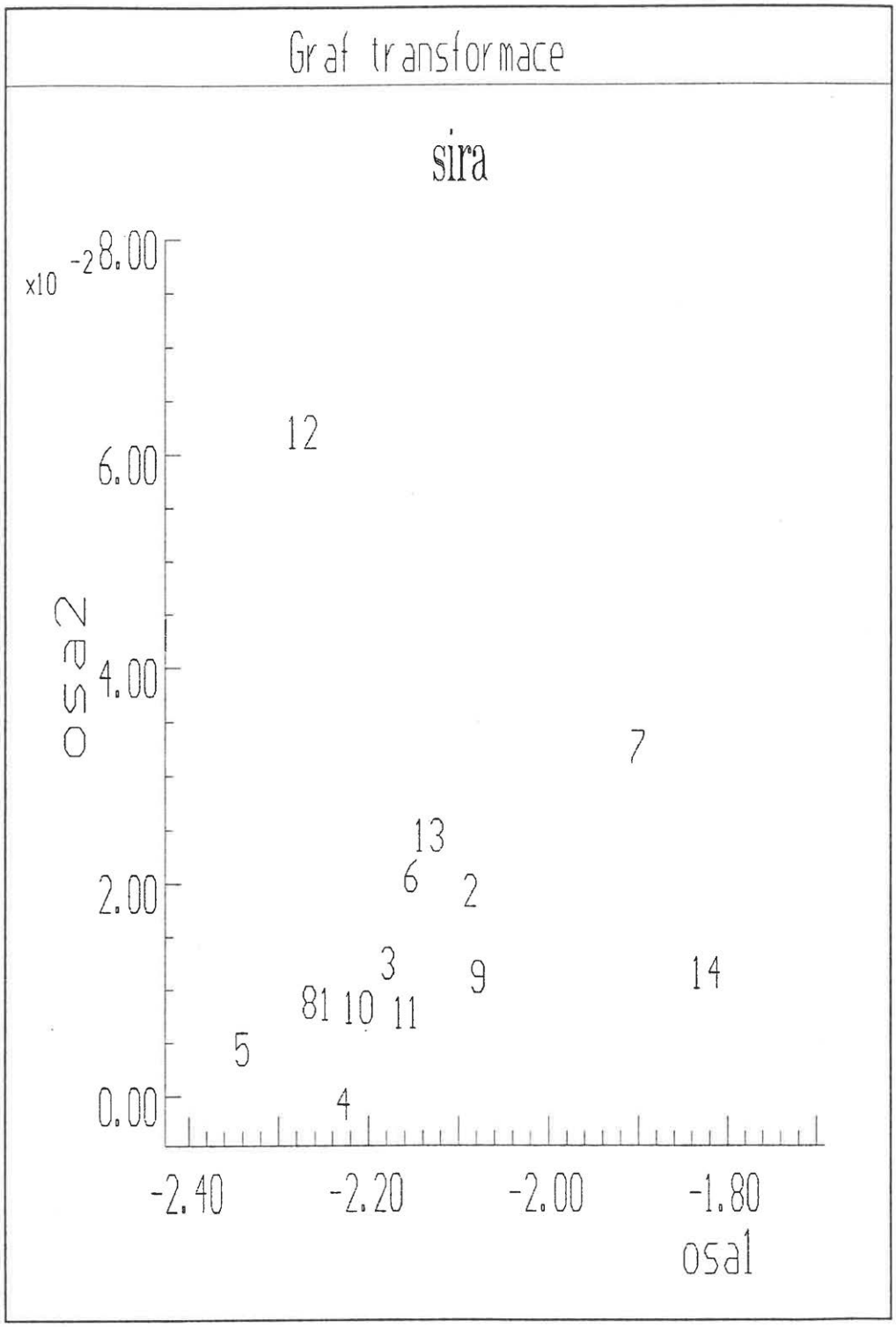
(Pokud je R blízké nule není transformace nutná.)

Závěr: Testovací kritéria pro faktor A i B jsou vyšší než jejich kvantily, což znamená, že nulovou hypotézu lze zamítnout a oba faktory jsou statisticky významné. Tyto závěry v praxi znamenají, že jak zvolená analytická metoda, tak i testující laboratoř mají vliv na stanovení obsahu síry v uhlí.

Přiložený Q-Q graf na obr. 2 je konstruován na základě Jackknife reziduí pro ověření jejich normality. Přibližně lineární závislost lze považovat za potvrzení tohoto předpokladu. Graf transformace na obr. 3 nevykazuje lineární závislost a potvrzuje, že není nutné provádět logaritmickou transformaci.


Q-Q graf
sira
-1.70
-1.80
-1.90
-2.00
-2.10
-2.20
-2.30
-2.40
osa2
-2.00
0.00
2.00
osa1

OBR. 2


Graf transformace
sira
x10 -2
8.00
6.00
4.00
2.00
0.00
osa2
12
7
13
6
2
3
9
14
81
10
11
5
4
-2.40
-2.20
-2.00
-1.80
osa1

OBR. 3

V. VÍCEROZMĚRNÁ STATISTICKÁ ANALÝZA

Vícerozměrnou statistickou analýzou jsou řešeny problémy, kdy u souboru objektů je sledováno více než jedna proměnná. V předložené práci byla data zpracována programovým systémem Statgraphics modulem Vícerozměrné metody. Z tohoto modulu byla použita analýza hlavních komponent, shluková analýza a grafické metody k vizuálnímu posouzení podobnosti vícerozměrných objektů.

1. Analýza hlavních komponent

Ve vícerozměrné statistice jsou výchozí datovou formou tzv. zdrojové matice, u nichž představují řádky objekty (měřené vzorky) a sloupce měřenou vlastnost (výsledek laboratorní zkoušky). Při velkém počtu měřených vlastností i objektů je poměrně obtížná interpretace. Pro zjednodušení analýzy a usnadnění interpretace je analýzou hlavních komponent zkoumána možnost nahrazení velkého počtu měřených vlastností menším počtem, aniž by došlo k podstatné ztrátě informace. Analýza hlavních komponent je metoda, která se pokouší nalézt skryté (latentní) proměnné, označované jako hlavní komponenty, jež maximálně reprezentují původní proměnné.

Protože platí, že součet rozptylů všech hlavních komponent je roven součtu rozptylů vstupujících proměnných, lze z podílů rozptylů jednotlivých hlavních komponent usuzovat na část celkové variability vstupujících proměnných vysvětlenou příslušnou hlavní komponentou. Jestliže součet prvních (nejvyšších) x těchto podílů je dostatečně blízký jedné, většinou se požaduje kolem 0,85-0,9, stačí vzít v úvahu těchto prvních x hlavních komponent pro vysvětlení chování původních proměnných.

Programový systém Statgraphics metody hlavních komponent předkládá v základním výstupu tabulku, kde jsou uvedena procenta variability vysvětlená jednotlivými hlavními komponentami. V grafických výstupech je graf komponentních vah, který zobrazuje komponentní váhy pro první dvě hlavní komponenty. Další graf, bodový diagram, zobrazuje komponentní skóre, tj. hodnoty prvních dvou hlavních komponent u jednotlivých objektů, navzájem. Dvojný graf je kombinací obou předchozích grafů. V číselných výstupech systému Statgraphics lze nalézt charakteristická čísla použité výběrové kovarianční nebo korelační matice, tedy rozptyly jednotlivých komponent. Jedním z dalších výstupů je matice komponentních vah (koeficientů lineární kombinace). Počet řádků i sloupců je roven počtu proměnných. Jednotlivé sloupce této matice představují charakteristické vektory, které odpovídají příslušným charakteristickým číslům (rozptylům komponent). Posledním z číselných výstupů jsou hodnoty hlavních komponent jednotlivých objektů. U výsledné matice je počet řádků roven počtu objektů a počet sloupců počtu proměnných.

2. Shluková analýza

Shluková analýza je metoda, která se zabývá zkoumáním podobnosti vícerozměrných objektů a jejich roztříděním do shluků. Cílem je v podstatě dosáhnout stavu, kdy objekty uvnitř shluku jsou si co nejvíce podobné a s objekty z různých shluků co nejméně. Podle způsobu shlukování se metody shlukové analýzy dělí na hierarchické a nehierarchické.

Hierarchické postupy jsou založeny na postupném spojování objektů a jejich

shluků do dalších (větších) shluků. Hierarchické metody se dále dělí na aglomerativní a divizní. Při aglomerativních postupech shlukování se vychází od jednotlivých objektů, které se postupně seskupují do tříd ke konečnému stavu, tj. spojení všech objektů do jedné třídy. Divizní postup je obrácený. Při tomto postupu se vychází z množiny všech objektů jako jediné třídy a jejím postupným dělením se získá systém tříd, až se skončí ve stádiu jednotlivých objektů. Výsledky hierarchických shlukovacích postupů lze zachytit graficky v podobě stromu - dendogramu.

Mezi aglomerativní hierarchické metody shlukové analýzy patří metoda průměrné vazby, kdy se vzdálenost dvou shluků počítá jako průměr z možných mezishlukových vzdáleností objektů. U metody nejbližšího souseda je kritériem pro spojování shluků minimum z možných mezishlukových vzdáleností objektů. Na opačném principu je založena metoda nejvzdálenějšího souseda, která počítá vzdálenost dvou shluků jako maximum z možných mezishlukových vzdáleností objektů. Další metodou je centroidní metoda, při níž se vzdálenost shluků počítá jako euklidovská vzdálenost jejich centroidů. mediánová metoda je v podstatě vylepšení centroidní metody, neboť se snaží odstranit rozdílné váhy, které centroidní metoda dává různě velkým shlukům. Mezi divizní metody patří rozklad nejkratší neuzavřené cesty a MacNaughtonova-Smithova metoda.

V nehierarchických shlukovacích postupech je počet tříd zadán předem, i když se v průběhu výpočtu může změnit. Zůstane-li počet tříd zachován, jde o nehierarchické metody s konstantním počtem shluků, v opačném případě to jsou nehierarchické metody s optimalizovaným počtem shluků. Nehierarchické metody lze rozdělit na dvě základní varianty, optimalizační metodu a analýzu modů. Analýza modů patří mezi metody, při kterých jsou hledány rozklady do tříd. Třídy jsou místa se zvýšenou koncentrací objektů. Optimalizační nehierarchické hledají optimální rozklad přiřazováním objektů z třídy do třídy s cílem minimalizovat nebo maximalizovat nějakou charakteristiku kvality rozkladů.

Při zpracování dat v této práci byla použita nehierarchická metoda typických bodů (Seeded) systému Statgraphics. Při použití této metody jsou na základě věcných znalostí zadány objekty, které jsou typickými představiteli nově vytvořených shluků, a systém rozdělí objekty do shluků podle jejich euklidovské vzdálenosti od těchto typických objektů.

3. Grafické metody

Grafické metody slouží pro vizuální posouzení podobnosti vícerozměrných objektů. Jednotlivé proměnné jsou znázorněny s ohledem na jejich konkrétní hodnoty do určitých geometrických tvarů nebo symbolů. Vlastnosti dat se posuzují s ohledem na vizuální rozdíly mezi symboly. Mezi základní typy symbolů patří profily, polygony, tváře, křivky a stromy. Profily představují dvourozměrné zobrazení m-rozměrných dat. Každý bod je charakterizován vertikálními úsečkami nebo sloupci. Profil pak vzniká spojením koncových bodů těchto úseček nebo sloupců. Tváře charakterizují každou složku x_{ij} vektoru x_i nějakým znakem, který je součástí schematizované tváře. Křivky využívají transformace každého bodu x_i na spojitou křivku, která je lineární kombinací všech jeho složek. Stromy jsou vhodné pro případy, kdy je počet složek vektoru **x** veliký. Jednotlivé složky x_j představují délku větví schematického stromu. Polygony jsou vlastně profily v polárních souřadnicích. Každá složka x_{ij} vektoru x_i odpovídá délce paprsku vycházejícího z jednoho středu. Paprsky jsou rozmístěny ve stejných vzdálenostech na kružnici.

V systému Statgraphics je k dispozici graf slunečních paprsků (Sun Ray Plot) a hvězdicový symbolický graf (Star Symbol Plot). Při využití metody Sun Ray Plot je systémem nakreslena pro každý objekt hvězdice, tzv. polygon, který je tvořen z paprsků spojených ve společném bodě a úseček mezi paprsky. Střed každého paprsku představuje průměr příslušné proměnné a jeho délka odpovídá 2n násobku směrodatné odchylky této proměnné. Polygon grafické metody Star Symbol Plot je tvořen z paprsků, které se pro každý objekt spojují v centrálním bodě. Stejně směřující paprsky u různých objektů datového souboru se liší svoji délkou. Nejdelší paprsek reprezentuje nejvyšší hodnotu příslušné proměnné a naopak nejkratší paprsek informuje, že u daného objektu nabývá příslušná proměnná nejmenší hodnoty z celého souboru.

4. Příklad vícerozměrné statistické analýzy

Zadání: V rámci technologického průzkumu v severočeské pánvi byl proveden odběr vzorků terciérních jílovců ze čtyř průzkumných vrtů V1, V2, V3 a V4 za účelem zjištění jejich použitelnosti v keramické výrobě. Odebrané vzorky byly dopraveny do laboratoře a na nich provedeny zkoušky: silikátová analýza včetně ztráty žíháním, zrnitostní zkouška, stanovení optimální vlhkosti keramického těsta, délkové změny sušením a pevnost v ohybu po vysušení. Zkoušky byly provedeny dle příslušných normovaných metodik.

Datový soubor:

Na základě provedených laboratorních zkoušek je v tab.1 zpracován přehled výsledků pro jednotlivé vzorky všech čtyř vrtů. Protože alkalické kovy, tj. Na a K, jsou v jílovcích severočeské pánve obsaženy v živcích a prvky alkalických zemin, tj. Ca a Mg, v karbonátech, případně v jílových minerálech v poměrně malém množství, byly při vyhodnocení oxidy alkalických kovů Na_2O a K_2O a oxidy prvků alkalických zemin CaO a MgO použity v podobě jejich součtu.

Řešení:

1. Analýza hlavních komponent

Metodou hlavních komponent byl zpracován výše uvedený soubor dat. V následující tab. 2 jsou uvedeny výsledky analýzy hlavních komponent. V prvním sloupci jsou uvedena pořadová čísla jednotlivých komponent. Ve druhém sloupci jsou uvedena procenta variability vysvětlená jednotlivými hlavními komponentami, tj. podíly rozptylů jednotlivých hlavních komponent na celkové variabilitě, v pořadí od nejvyššího k nejnižšímu. Ve třetím sloupci jsou tytéž podíly uvedeny kumulovaně a lze z nich určit optimální počet komponent. Obyčejně se požaduje, aby součet prvních nejvyšších komponent měl 85 - 92 %. Z tab. 2 vyplývá, že tento předpoklad splňuje 5 - 6 hlavních komponent. Na obr. 1 - 3 jsou znázorněny grafy hlavních komponent.

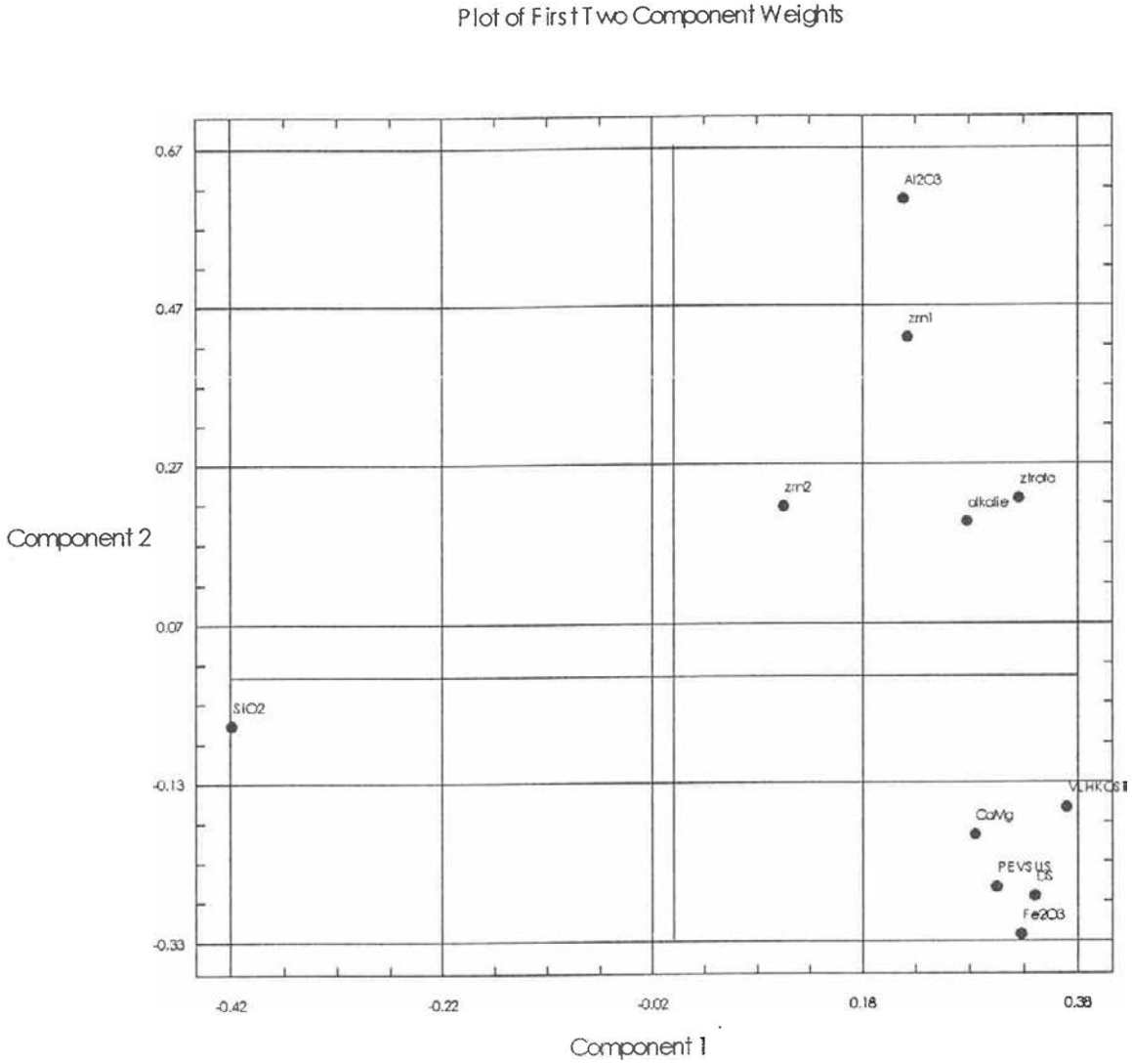
Tab. 2: Analýza hlavních komponent

Principal Components Analysis		
Component Number	Percent of Variance	Cumulative Percentage
1	47.46525	47.46525
2	16.39470	63.85995
3	11.95133	75.81128
4	8.09601	83.90729
5	4.86966	88.77695
6	4.29131	93.06826
7	3.20702	96.27527
8	1.73000	98.00527
9	1.27612	99.28139
10	0.68424	99.96563
11	0.03437	100.00000

V tab. 3 jsou uvedeny komponentní váhy pro první dvě hlavní komponenty a na obr. 1 je znázorněn **graf komponentních vah** pro první dvě hlavní komponenty u jednotlivých proměnných vstupujících do analýzy.

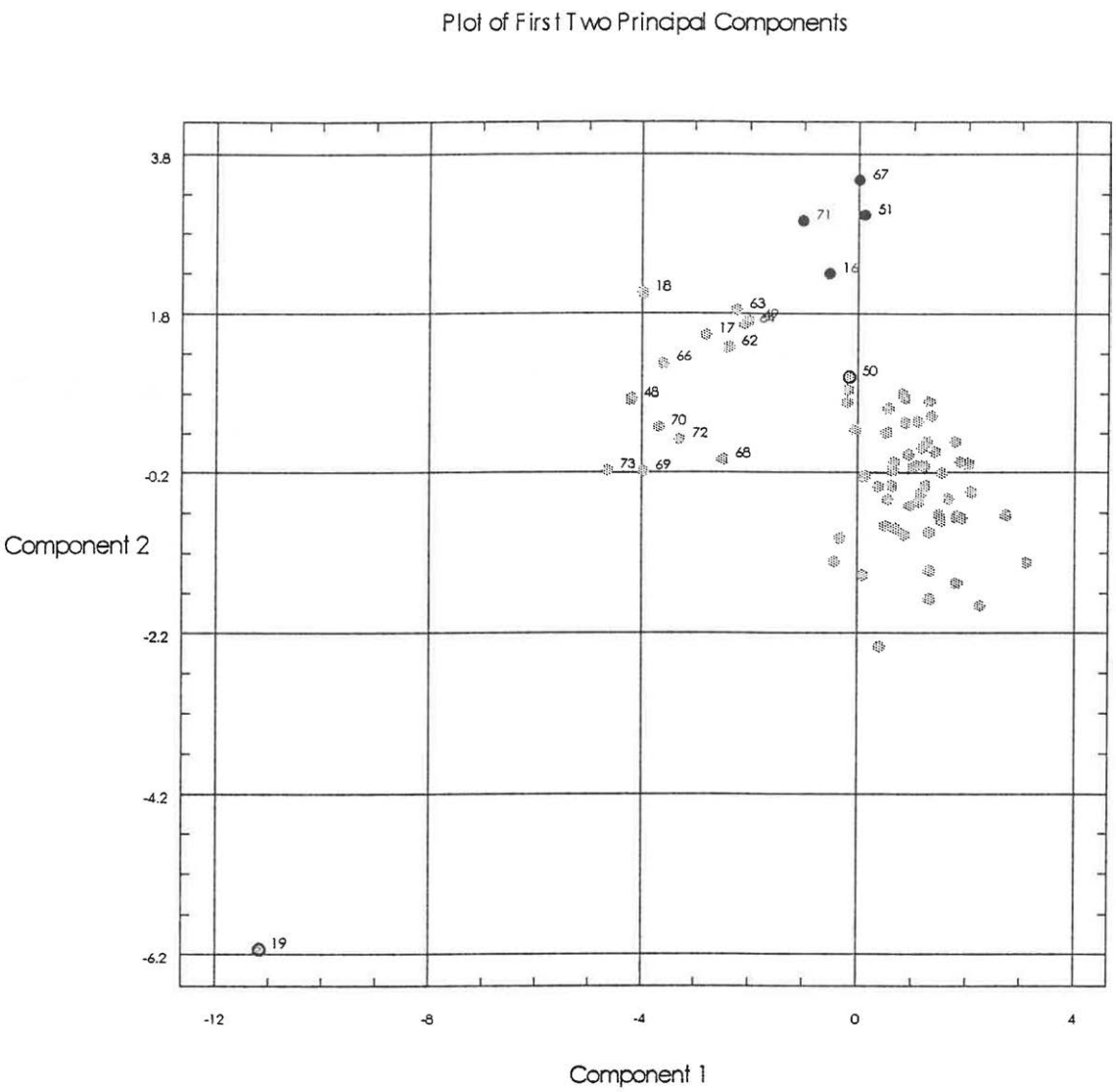
Tab. 3: Komponentní váhy (koeficienty lineární kombinace) pro první dvě hlavní komponenty

Číslo objektu	komponentní váhy		Číslo objektu	Komponentní váhy	
	pro 1. hlavní komponentu	pro 2. hlavní komponentu		pro 1. hlavní komponentu	pro 2. hlavní komponentu
1	0.36939	-0.165576	7	0.285145	-0.199643
2	0.339967	-0.277312	8	0.277501	0.193265
3	0.305127	-0.266153	9	-0.418784	-0.0609797
4	0.325519	0.222766	10	0.218275	0.601643
5	0.222033	0.425989	11	0.327449	-0.324889
6	0.105111	0.212836			



Obr. 1 Graf komponentních vah pro první dvě hlavní komponenty

Na obr. 2 je znázorněn **bodový diagram**, který zobrazuje komponentní skóre, tj. hodnoty prvních dvou hlavních komponent u jednotlivých objektů. V číselném výstupu systému Statgraphics modulu Vícerozměrná data byly uloženy hodnoty hlavních komponent jednotlivých objektů. V tab. 4 jsou uvedeny hodnoty prvních dvou hlavních komponent jednotlivých objektů.

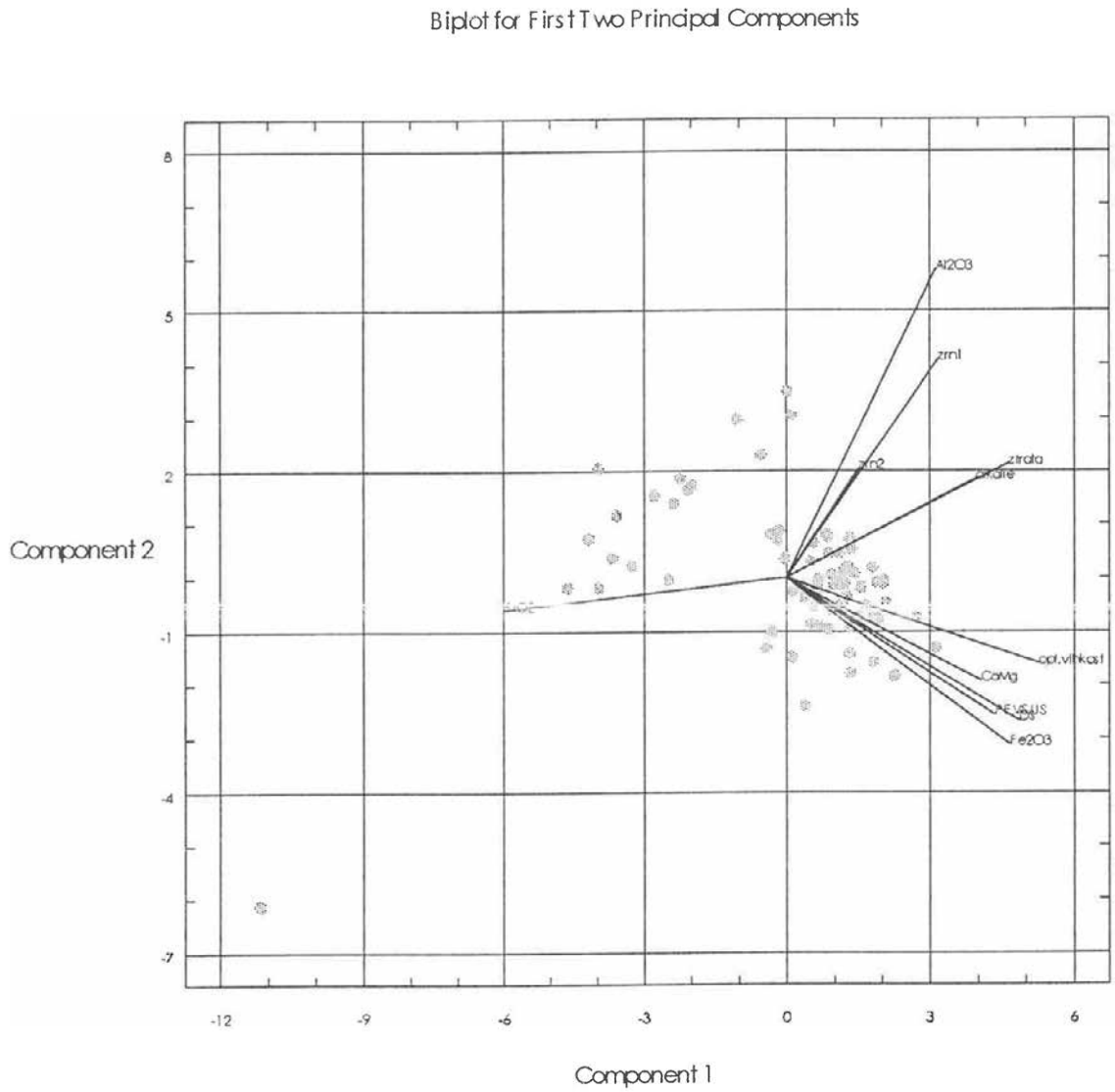


Obr. 2 Bodový diagram klasifikuje data dle prvních dvou hlavních komponent

Z grafu na obr. 2 je zřejmé, že lze soubor testovaných objektů rozdělit na 4 shluky. Nejpočetnější shluk tvoří skupina objektů, které jsou na grafu bez označení, vyjma objektu 50. Tento objekt je dalšími metodami zařazen do jiného shluku (viz kapitola 4.2 a 4.3). Druhý shluk představuje skupina objektů č. 17, 18, 49, 62 - 66, 68 - 70, 72 a 73. Třetí shluk tvoří čtyři objekty č. 16, 51, 67 a 71. Zcela ojediněle je v grafu na obr. 2 umístěn objekt č. 19, který představuje čtvrtý shluk. První shluk tvoří objekty resp. vzorky jílovců, které byly odebrány ve svrchní části vrtů V1, V2, V3 a V4. Ze spodní části uvedených vrtů jsou vzorky jílovců, které byly zařazeny do shluku 2, 3 a 4. Vzhledem k poloze vzorků ve vrtech lze očekávat i odlišné chemické a technologické vlastnosti vzorků. Svrchní části vrtů jsou tvořeny jílovcem s vyšším obsahem Fe_2O_3 než v materiálech spodní části vrtů, dále byla ve vzorcích svrchní části vrtů zjištěna i vyšší ztráta žíháním, vyšší optimální vlhkost keramického těsta, větší hodnota délkových změn sušením i vyšší hodnota pevnosti v ohybu po vysušení než mají vzorky ze spodní části vrtů.

Na obr. 3 je znázorněn **dvojný graf**, který kombinuje oba předchozí grafy. Body představují komponentní skóre jednotlivých objektů. Přímky spojující se v bodě (0,0) představují původní proměnné, délka každé přímky je úměrná svému příspěvku k prvním dvěma hlavním komponentám.

Z grafického výstupu na obr. 3 vyplývá, že za hlavní komponenty lze považovat obsah oxidu Si, Al, Fe, ztrátu žíháním, optimální vlhkost a velikost zrn pod 2 μm označenou zrn 1. Ostatní komponenty nebudou mít pro hodnocení suroviny zásadní význam, protože jsou v korelaci s některými proměnnými. Proměnná „alkálie" je v korelaci s proměnnou „ztráta", proměnné „DS" a „PEVSUS" jsou v korelaci s proměnnou „vlhkost". Lze uvažovat i proměnnou „zrn2" za dostatečně nezávislou, a proto i důležitou. Data byla standardizována, protože nejsou vyjádřena ve stejných jednotkách.



Obr. 3 Dvojný graf přináší kombinaci grafu komponentních vah a bodového diagramu

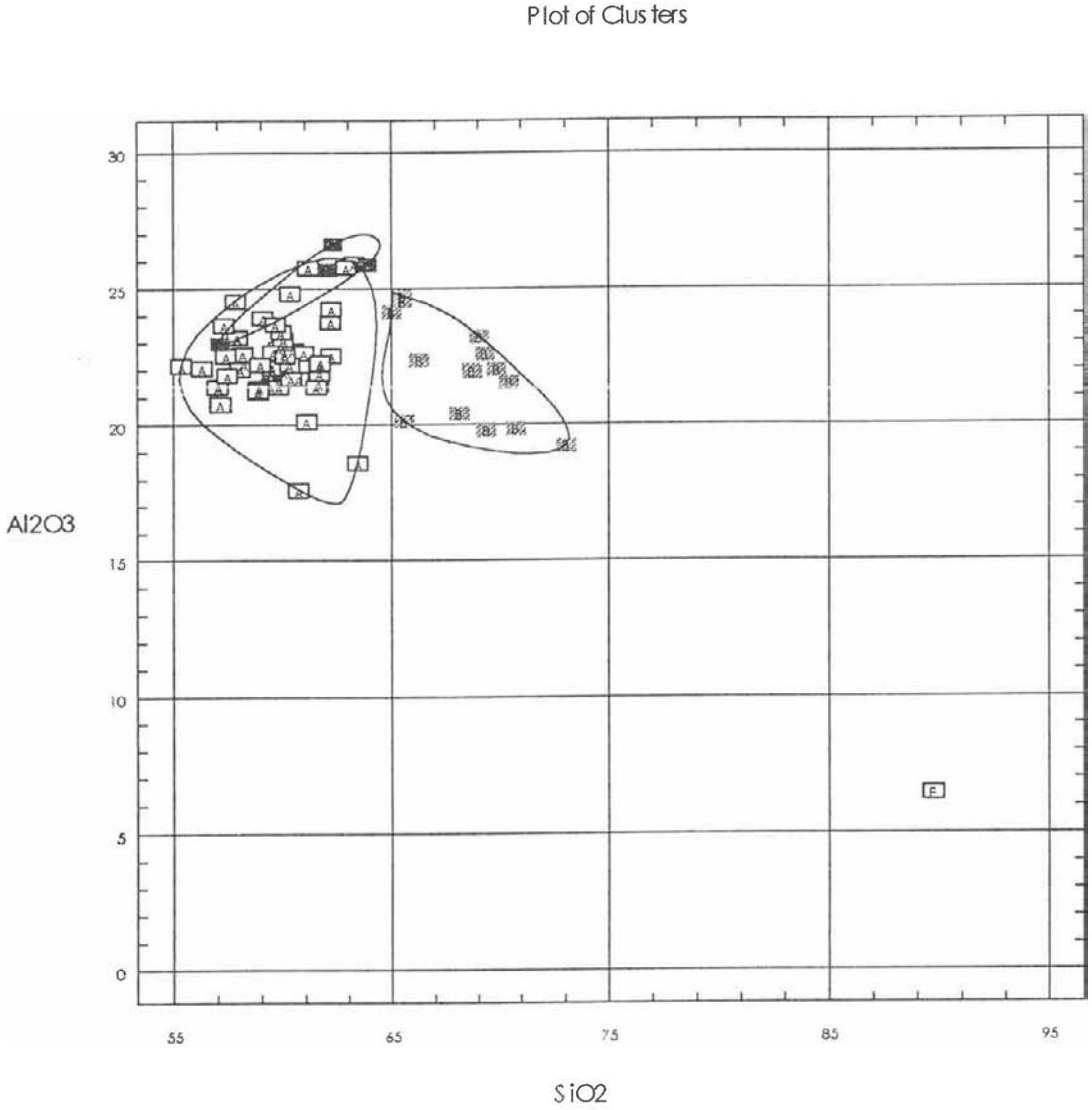
2. Shluková analýza

Shlukovou analýzou byla zkoumána podobnost vícerozměrných objektů. Na základě uvedeného byl datový soubor zpracován nehierarchickou metodou typických vzorků (metoda Seeded). Vzorky č.3, 66, 67 a 19 byly zadány jako objekty, které jsou typickými představiteli nově vytvořených čtyř shluků, tj. čtyř typů surovin. Programovým systémem byl proveden výpočet a rozdělení objektů do čtyř shluků. Přehled výsledků je uveden v tab. 5. Vstupní proměnné nejsou vyjádřeny ve stejných jednotkách, proto byla provedena standardizace.

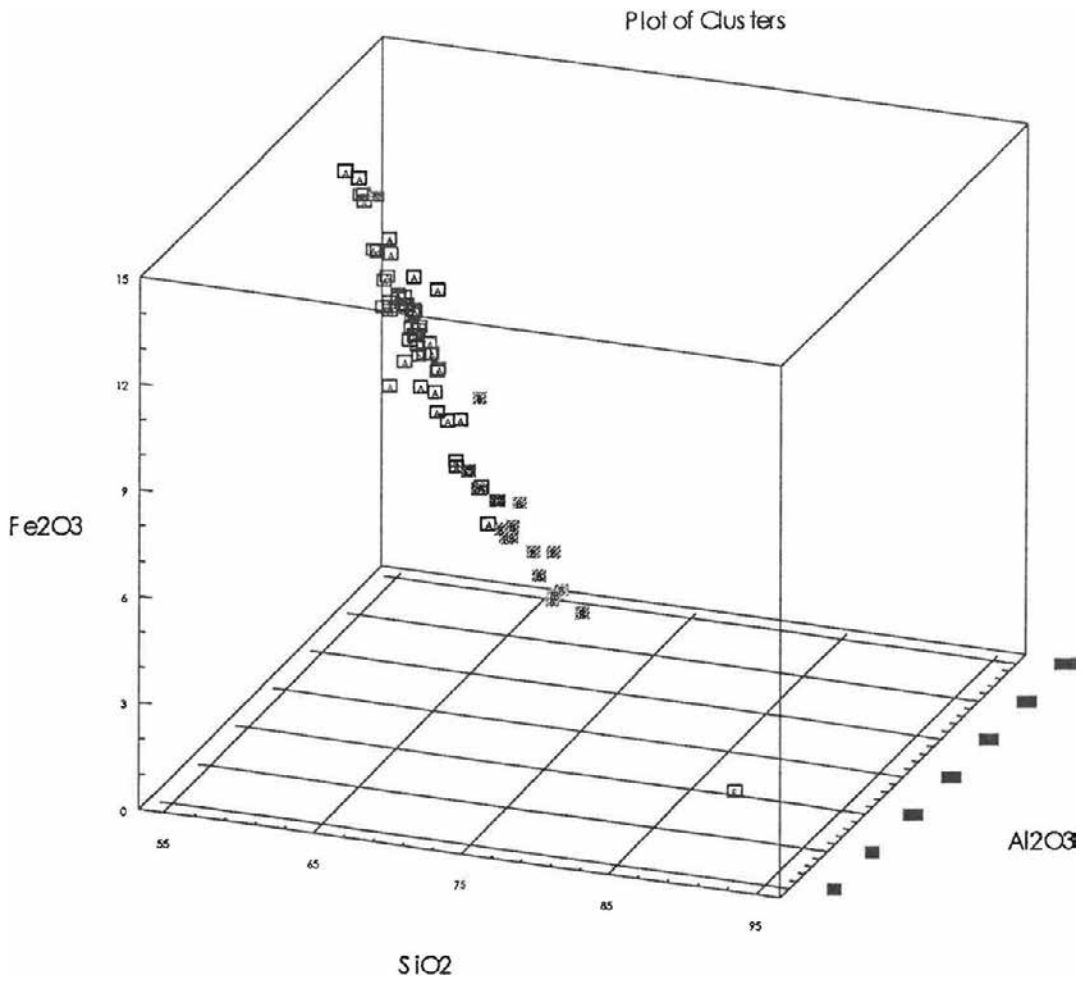
Tab. 5: Výsledky shlukové nehierarchické analýzy vzorků č. 1 - 73

Results of Clustering by Seeded Method				
Cluster	**Označení v grafech 4 a 5**	**Frequency**	**Percentage**	**Seed**
1 (ostatní vzorky)	**A**	54	73.9726	3
2 (vzorek č.17, 18, 48, 49, 62-66, 68-70, 72, 73)	**B**	14	19.1781	66
3 (vzorek č. 50, 51, 67, 71)	**C**	4	5.4795	67
4 (vzorek č.19)	**E**	1	1.3699	19

Jak vyplývá z tabulky 3, do skupiny č. 2 byly zařazeny objekty č. 17-19, 48, 49, 62-66, 68-70, 72 a 73. Tyto objekty představují asi 19 % z celkového počtu objektů. Do skupiny č. 3 byly zařazeny objekty č. 50, 51, 67 a 72 a představují asi 5.5 %. Do skupiny č. 4 byl zařazen pouze objekt č. 19. Zbývající objekty, které představují 74 %, byly zařazeny do skupiny č.1. Rozdělení objektů do shluků téměř koresponduje s rozdělením popsaném v předchozí kapitole 4.1. Výjimku tvoří pouze objekty č. 16 a 50. Shlukovou analýzou byl objekt č. 16 zařazen do shluku 1 a objekt č. 50 do shluku 3 a analýzou hlavních komponent bylo zařazení provedeno opačně, tj. objekt č. 16 byl zařazen do shluku 3 a objekt č. 50 do shluku 1. Grafické znázornění výsledků shlukové analýzy je provedeno na obr. 4 a 5. Na obr. 4 jsou hodnoty jednotlivých objektů znázorněny ve dvojrozměrném prostoru a na obr. 5 v trojrozměrném prostoru.



Obr. 4 Graf čtyř shluků v dvojrozměrném prostoru (SiO_2, Al_2O_3)

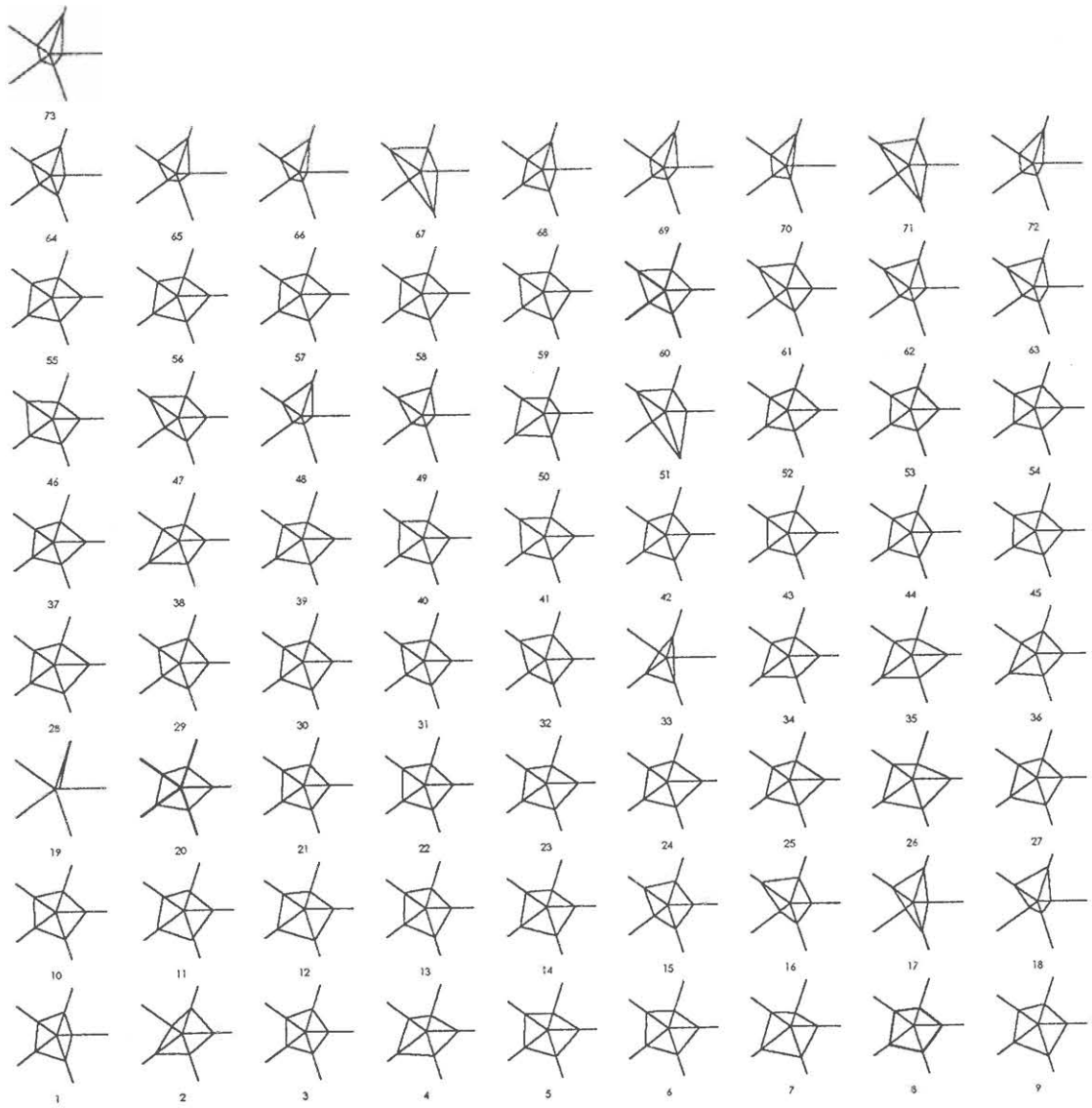


Obr. 5 Graf čtyř shluků v trojrozměrném prostoru (SiO_2, Al_2O_3, Fe_2O_3)

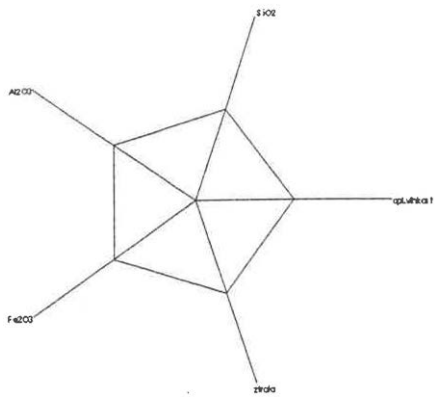
4.3 Grafické metody

4.3.1 Sun Ray Plot - graf slunečních paprsků

Grafickou metodou slunečních paprsků Sun Ray Plot bylo provedeno vizuální srovnání objektů z hlediska jejich porovnání s průměrným objektem. Pro každý objekt je nakreslen polygon z pěti paprsků, které představují 5 hlavních komponent, tj. SiO_2, Al_2O_3, Fe_2O_3, ztrátu žíháním a optimální vlhkost. Na obr. 6 je 73 polygonů testovaných vzorků a klíč je znázorněn na obr. 7. Dle vizuálního srovnání jednotlivých polygonů lze zaznamenat výraznější odlišnost tvaru polygonů u objektů č. 17 -19, 33, 48 - 51 a u objektů č. 62-73 od ostatních polygonů.



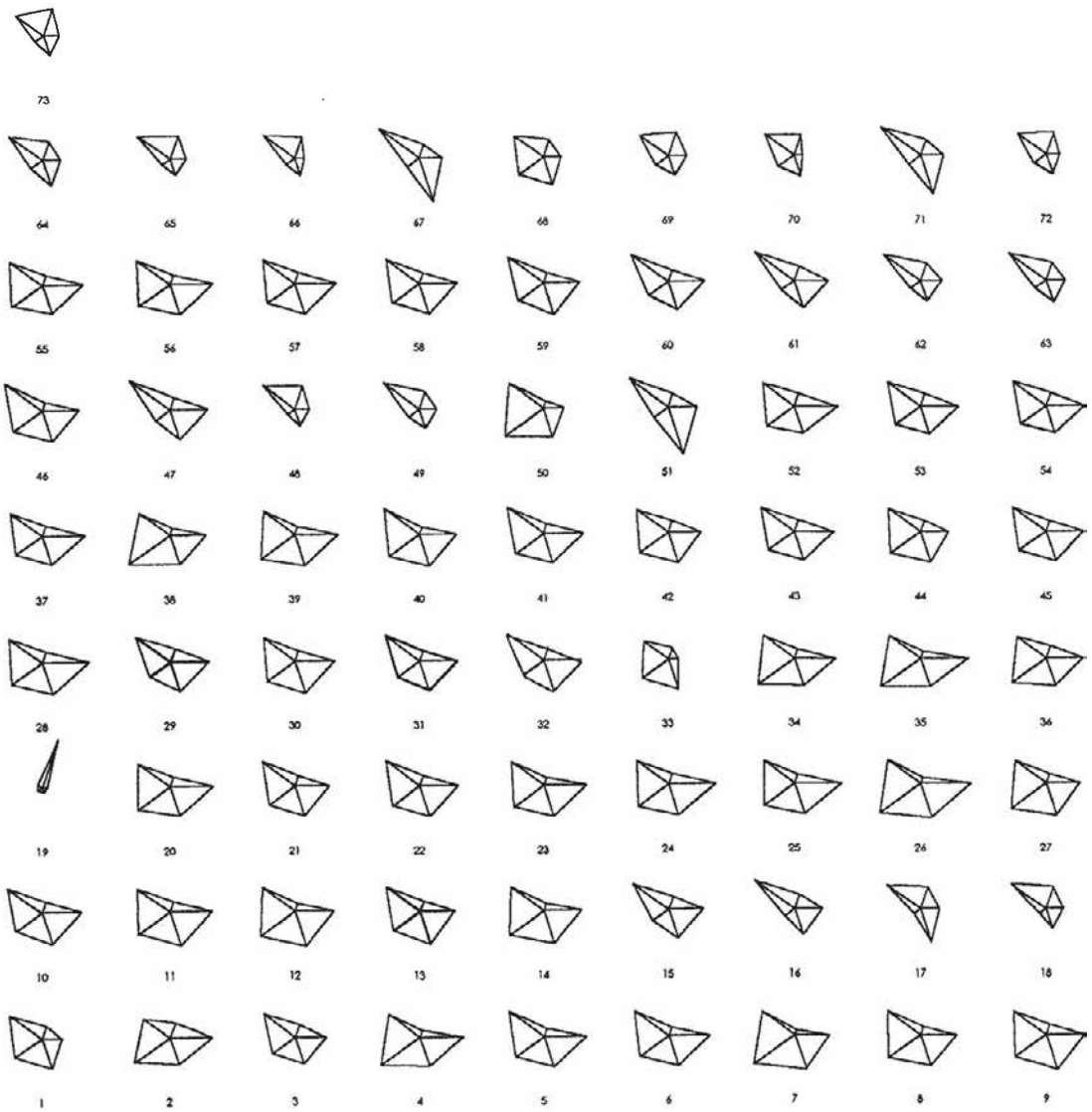
Obr. 6 Sun Ray Plot - graf slunečních paprsků pro proměnné: SiO_2, Al_2O_3, Fe_2O_3, ztrátu žíháním a optimální vlhkost



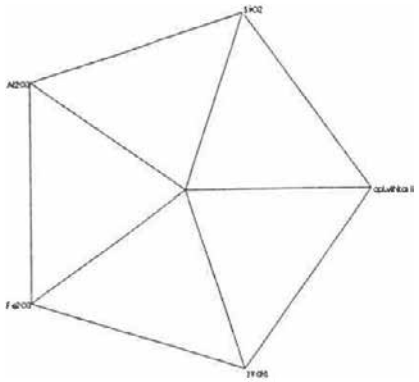
Obr. 7 Klíč ke grafu slunečních paprsků

4.3.2 Star Symbol Plot - hvězdicový symbolický graf

Graf, podobně jako předchozí, slouží k vizuálnímu zobrazení podobnosti vícerozměrných objektů. Pro každý objekt je nakreslen polygon z pěti paprsků, které představují 5 hlavních komponent, tj. SiO_2, Al_2O_3, Fe_2O_3, ztrátu žíháním a optimální vlhkost. Na obr. 8 je 73 polygonů testovaných vzorků a klíč je znázorněn na obr. 9. Dle vizuálního srovnání jednotlivých polygonů lze zaznamenat výraznější odlišnost tvaru polygonů i jejich velikosti od ostatních polygonů u testovaných objektů č. 16 - 19, u objektů č. 48 - 51 a u objektů č. 62-73.



Obr. 8 Star Symbol Plot - hvězdicový symbolický graf pro proměnné: SiO_2, Al_2O_3, Fe_2O_3, ztrátu žíháním a optimální vlhkost

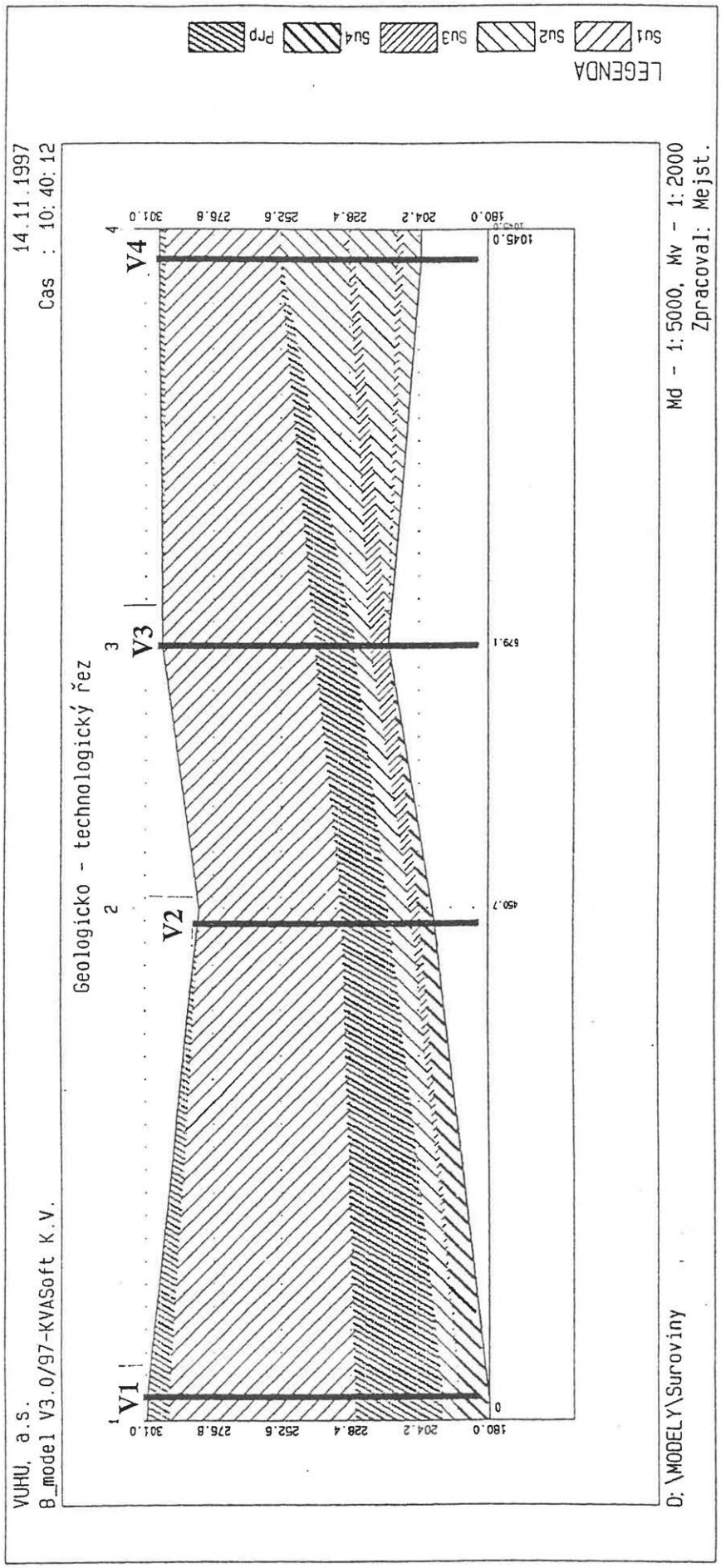


Obr. 9 Klíč k hvězdicovému symbolickému grafu

5.0 ZÁVĚR

Cílem předložené práce bylo zhodnocení vzorků nadložních sedimentů v severočeské pánvi metodou vícerozměrných dat na základě výsledků laboratorních zkoušek a rozdělení nadložních sedimentů do surovinových typů. Vzhledem k průzkumným pracím v minulých letech lze na vybrané lokalitě předpokládat přítomnost minimálně 2 typů surovin.

Na základě statistického zpracování bylo možné zařadit jednotlivé vzorky nadložních jílovců do čtyř surovinových typů. Zjištěné surovinové typy lze zhodnotit formou geologicko-technologického řezu, kde je provedeno grafické znázornění zjištěných typů surovin. Geologicko-technologický řez je veden čtyřmi vrty – V1, V2, V3 a V4 (viz obr. 10). Ve vrtu V1 tvoří prvních 7,5 m navážka, kterou lze považovat za nevhodnou surovinu. V geologicko-technologickém řezu jsou nevhodné suroviny označeny jako proplástek (Prp). V hloubce 7,5 m - 73,4 m byla zjištěna surovina typu 1. Od 73,4 m do 103,7 m se nachází vrstva uhelných jílovců, která je nevhodná pro keramické zpracování. Od 103,7 m do 109,4 m se nachází surovina typu 2 a v hloubce 109,4-119,0 m byla zjištěna surovina typu 4, reprezentovaná objektem č. 19. V hloubce 119,0 m končí vrt. Vrtem V2 byla od hloubky 0,0 do 50,7 m zastižena surovina typu 1. V této hloubce byly ukončeny i vrtné práce. Vrstva suroviny typu 2 ve vrtu V2, zakreslená v geologicko-technologickém řezu na obr. 10, je pouze idealizované dokončení na základě vyhodnocené suroviny v sousedních vrtech V1 a V3. Vrt V3 byl realizován do hloubky 79,8 m. Od 0,3 do 54,2 m byla ve vrtu vyhodnocena surovina typu 1. V hloubce 54,2 - 67,7 m se nachází nevhodná surovina jílovců s uhelnou příměsí. Surovina typu 2 byla zjištěna v hloubce 67,7 - 73,6 m. V hloubce 73,6 - 79,8 m leží vrstva suroviny typu 3. Vrt V4 byl realizován do hloubky 92,5 m. Ve svrchní části vrtu, v hloubce 2,0 - 41,0 m byla opět zjištěna surovina typu 1. Vrstva suroviny typu 2 se nachází v hloubce 41,0 - 67,2 m, 69,2 - 84,0 m a 86,3 - 92,5 m. Vzorky jílovců odebrané v hloubce 67,2 - 69,2 m a 84,0 - 86,3 m byly vyhodnoceny jako surovina typu 3.


LEGENDA
Su1
Su2
Su3
Su4
PrP
Geologicko - technologický řez
V1
V2
V3
V4
1
2
3
4
180.0
204.2
228.4
252.6
276.8
301.0
0
450.7
679.1
1045.0
14.11.1997
Cas : 10:40:12
Md - 1:5000, Mv - 1:2000
Zpracoval: Mejst.
VUHU, a.s.
B_model V3.0/97-KVASoft K.V.
D:\MODELY\Suroviny

Obr. 10

Jak již bylo uvedeno, předpokládala se přítomnost dvou typů surovin. Statistickým testováním byly určeny čtyři surovinové typy. Z geologicko-technologického řezu na obr. 10 vyplývá, že svrchní část řezu tvoří keramický materiál suroviny typu 1. Jedná se o keramzitovou surovinu, ze které je vyráběn keramzit rychlovýpalem. Surovinu typu 1 lze rovněž použít jako přísadovou surovinu pro výrobu cihlářských výrobků. Tuto surovinu je nutné ostřit, tzn. je nutný přídavek ostřící směsi, písku. Ve spodní části řezu byly jílovce vyhodnoceny jako surovinové typy 2, 3 a 4. Je nutné zdůraznit, že surovinový typ 3 je reprezentován čtyřmi vzorky a typ 4 pouze jedním vzorkem. Tyto tři typy surovin lze označit jako podskupiny kameninové suroviny.

Vyhodnocené suroviny jsou první etapou při průzkumu doprovodných surovin v severočeské pánvi. Na základě tohoto průzkumu je prováděna realizace doplňujících vrtů, odběry dalších vzorků a jejich laboratorní testování se zaměřením na zjištěné typy surovin. Geologicko-technologický řez bývá využíván případně i pro hodnocení báňsko-technologických podmínek.

6. POUŽITÁ LITERATURA

1) ČSN 01 5030: Stanovení zrnitosti sítovým rozborem.
ČSN 72 1127: Stanovení zrnitosti keramických látek sedimentací.
ČSN 72 1074: Stanovení optimální a pracovní vlhkosti keramických těst Pfefferkornovým přístrojem.
ČSN 72 1565, část 5: Stanovení délkových změn sušením.
ČSN 72 1565, část 7: Stanovení pevnosti v ohybu.

2) Hlaváč J. : Základy technologie silikátů. SNTL/ALFA, Praha, 1988.

3) Hebák P., Hustopecký, J.: Vícerozměrné statistické metody s aplikacemi. SNTL/ALFA, Praha, 1987.

4) Koschin F. a kol.: Statgraphics. GRADA, Praha, 1992.

5) Lach V.: Technologie. Cihlářska a žárovzdorná výroba. SNTL, Praha, 1981.

6) Meloun M., Militký J.: Statistické zpracování experimentálních dat. PLUS, Praha, 1994.

7) Mejstříková L.: Průzkum a hodnocení doprovodných surovin na DJŠ v rámci TP 1992. Výzkumná zpráva, VÚHU Most, 1993.

8) Pytela O.: Chemometrie pro organické chemiky. Univerzita Pardubice, 1995.