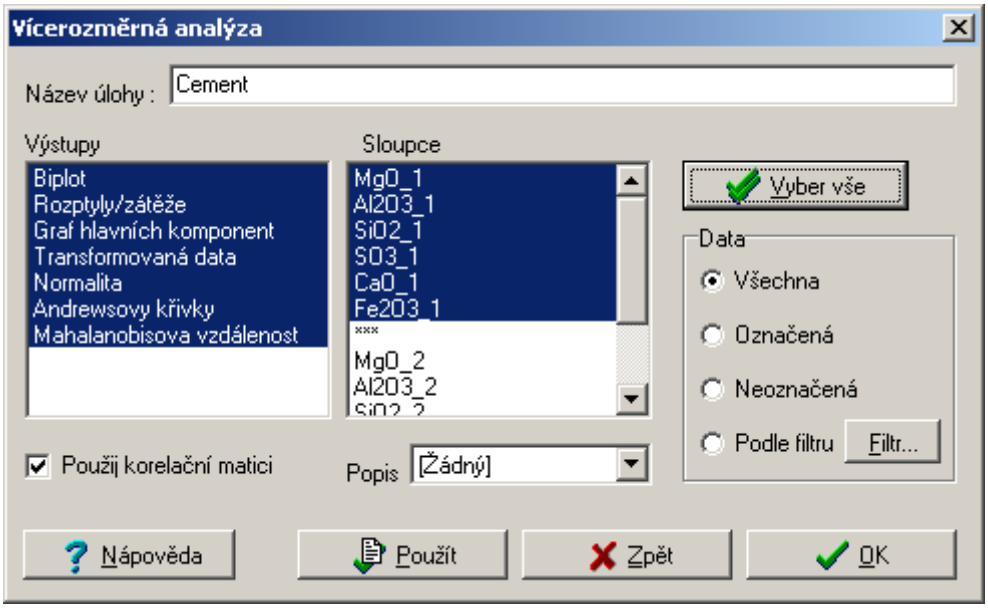
Vícerozměrná analýza

Menu:	QCExpert	Vícerozměrná analýza

Modul Vícerozměrná analýza je určen k posouzení struktury, exploratorní analýze kvantitativních vícerozměrných dat a analýze hlavních komponent. Vícerozměrná data (vícerozměrný náhodný výběr) jsou výsledkem současného měření několika (*m*) veličin (proměnných), například chemických a fyzikálních vlastností jednotlivých vzorků, několika rozměrů jedné součástky, nebo různých údajů o zaměstnancích. Počet takto získaných dat označujeme *n*. K posouzení vícerozměrné normality dat, která se předpokládá například při konstrukci Hotellingova diagramu, slouží graf vícerozměrné normality a graf symetrie s možností identifikace odlehlých dat pomocí interaktivních grafů. K posouzení struktury dat je určen Andrewsův graf a graf Biplot. Analýza hlavních komponent transformuje data do nových ortogonálních souřadnic tak aby se co nejvíce informace v původních datech dalo vyjádřit menším počtem proměnných. Míra variability vysvětlená komponentami je uveden v grafu Vysvětlený rozptyl. Zastoupení původních proměnných v jednotlivých komponentách je znázorněno v grafech zátěží. Další charakteristikou je Mahalanobisova vzdálenost (MV), pravděpodobnostní vzdálenost od průměru. Dá se přirovnat ke vzdálenosti od střední hodnoty jednorozměrného výběru v jednotkách sigma. Velká hodnota MV je málo pravděpodobná, a tedy podezřelá. Spolehlivější pro diagnostiku vybočujících dat je robustní MV založená na M-odhadu střední hodnoty místo průměru, která není ovlivněna vybočujícími daty.

Data a parametry

Data jsou ve sloupcích, každý sloupec odpovídá jedné proměnné. Počet dat v jednotlivých sloupcích by měl být stejný. Řádky s chybějícími hodnotami budou z výpočtu vyloučeny. Minimální počet sloupců je 2. Minimální počet řádků je 4. Názvy sloupců by měly odpovídat názvům analyzované proměnné, např. Obsah_Cr, Obsah_Mn, Tažnost. Výběr sloupců z aktuálního listu se může provést v poli *Sloupce* dialogového panelu *Vícerozměrná analýza*, Obrázek 1. Implicitně jsou vybrány všechny sloupce obsažené v aktuálním listu. Podle vybraných položek v poli *Výstupy* se provedou odpovídající analýzy. Je-li zaškrtnuto políčko *Použij korelační matici*, provede se analýza hlavních komponent na základě korelační matice, jinak se provede na základě kovarianční matice. Použití korelační matice se doporučuje zvláště jsou-li hodnoty proměnných řádově odlišné.



Obrázek 1 Dialogový panel Vícerozměrné analýzy

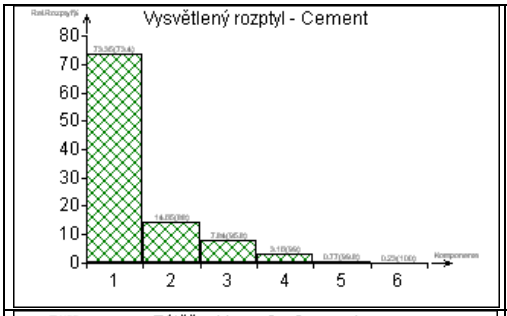
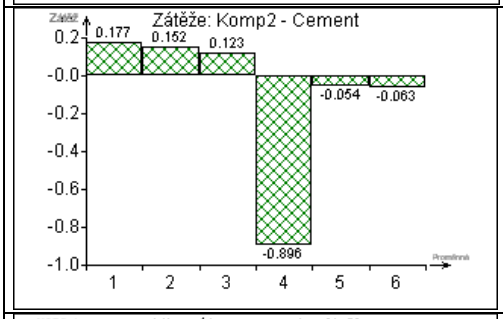
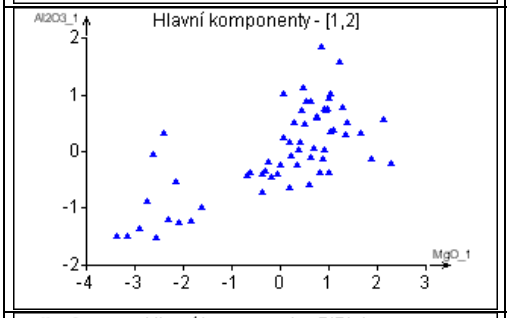
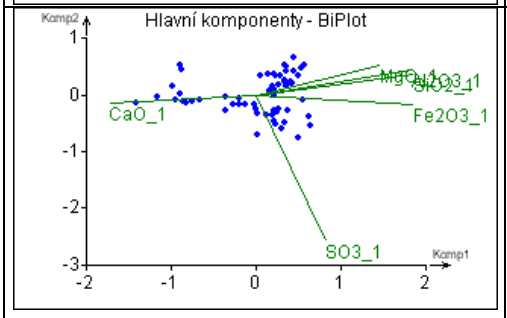
Protokol

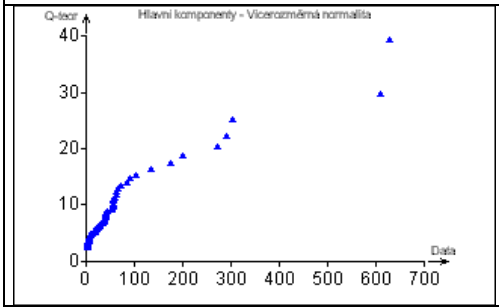
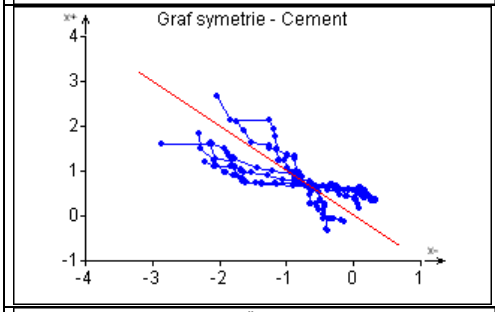
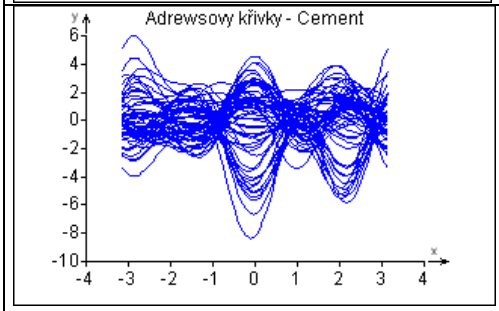
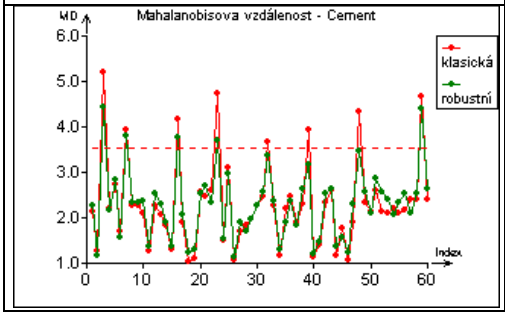
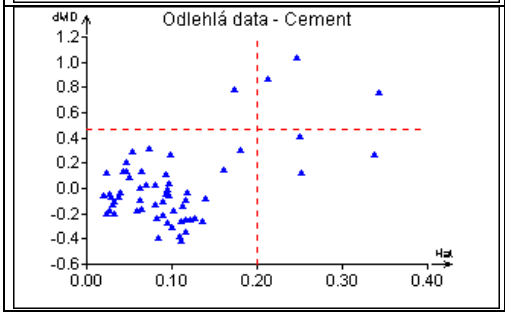
Název úlohy	Název úlohy z dialogového panelu.

Počet proměnných Počet dat	Počet proměnných (sloupců). Počet platných řádků.
Základní charakteristiky	
Proměnná Průměr Rozptyl Směr. odchylka Minimum Maximum	Název proměnné. Aritmetický průměr proměnné. Rozptyl proměnné. Směrodatná odchylka proměnné. Minimální hodnota proměnné. Maximální hodnota proměnné.
Korelační matice	Párové korelační koeficienty mezi jednotlivými proměnnými uspořádané do korelační matice. Na diagonále jsou jedničky. Tato matice je použita pro výpočet vlastních čísel, vlastních vektorů a analýze hlavních komponent, je-li při výpočtu zaškrtnuto políčko *Použij korelační matici*. Viz též modul Korelace.
Proměnná	Název proměnné.
Kovarianční matice	Kovariance mezi jednotlivými proměnnými uspořádané do matice. Na diagonále jsou rozptyly proměnných. Tato matice je použita pro výpočet vlastních čísel, vlastních vektorů a analýze hlavních komponent, není-li při výpočtu zaškrtnuto políčko *Použij korelační matici*.
Proměnná	Název proměnné.
Variabilita vysvětlená hl. komponentami	Čtyři míry udávající kolik variability je vysvětleno jednotlivými komponentami. Komponenty jsou vždy seřazeny tak, že první komponenta vysvětluje největší díl variability, poslední komponenta nejmenší díl.
Komponenta Vlastní číslo Rozptyl Směr. odchylka Rel. variabilita,% Kum. variabilita,%	Pořadové číslo komponenty. Vlastní čísla korelační nebo kovarianční matice, podle volby v dialogovém panelu vícerozměrné analýzy. Rozptyl vysvětlený danou komponentou, tj. rozptyl průmětu původních dat do dané komponenty. Odmocnina z rozptylu v dané komponentě. Relativní vyjádření vysvětleného rozptylu v procentech. Kumulativní relativní vyjádření vysvětleného rozptylu v procentech.
Vlastní vektory korelační/kovarianční matice	Sloupce obsahují vlastní vektory korelační nebo kovarianční matice, podle volby v dialogovém panelu vícerozměrné analýzy.
Sloupec	Název příslušné proměnné.
Zátěže	Vlastní vektory vynásobené odmocninou příslušného vlastního čísla. Zátěže udávají strukturu jednotlivých komponent. Někdy vystihují komponenty různé rysy dat určené jednou nebo několika proměnnými. Tyto proměnné pak mají v dané komponentě výrazně vyšší absolutní hodnotu zátěže.
Sloupec	Název proměnné.
Robustní M-odhady	Alternativní odhad vektoru středních hodnot určený iterativní metodou M-odhadu. Od klasického průměru (viz Základní charakteristiky) se liší

	svou robustností, to znamená, že tyto odhady nejsou ovlivněny vybočujícími daty.
Mahalanobisova vzdálenost	Vzdálenost od střední hodnoty určená pravděpodobností výskytu. Je uvedena klasická MV, založená na průměru a robustní MVm, založená na robustním M-odhadu střední hodnoty.
Klasická MV	Vzdálenost *i*-tého bodu $\mathbf{x}_i$ od průměru $\mathbf{x}_p$ vzhledem ke konfidenčnímu elipsoidu, určenému kovarianční maticí **S**, $(\mathbf{x}_i - \mathbf{x}_p)^T(\mathbf{S})^{-1}(\mathbf{x}_i - \mathbf{x}_p)$.
Robustní MVm	Robustní modifikace Mahalanobisovy vzdálenosti založená na robustním M-odhadu polohy $\mathbf{x}_M$, $(\mathbf{x}_i - \mathbf{x}_M)^T(\mathbf{S})^{-1}(\mathbf{x}_i - \mathbf{x}_M)$. Vybočující body mají velkou hodnotu MVm.
Transformovaná data	Původní data vyjádřená v souřadnicích hlavních komponent.

Grafy

	Graf vysvětleného rozptylu, relativní variabilita vysvětlená jednotlivými komponentami. Komponenty jsou vždy seřazeny tak, že první komponenta vysvětluje největší díl variability, poslední komponenta nejmenší díl. Na ose *x* jsou pořadová čísla komponenty, na ose *y* jsou procent**a rozptylu vztažená na celkový rozptyl. Čísla nad sloupci udávají rozptyl a kumulativní rozptyl v procentech. Absolutní hodnoty jsou uvedeny v protokolu.
	Grafické vyjádření zátěží pro jednotlivé komponenty. Zátěže udávají strukturu jednotlivých komponent. Někdy vystihují komponenty různé rysy dat určené jednou nebo několika proměnnými. Tyto proměnné pak mají v dané komponentě výrazně vyšší absolutní hodnotu zátěže.
	Grafy hlavních komponent jsou rozptylové grafy pro všechny kombinace komponent. Tyto grafy mají často větší vypovídací schopnost, než párové rozptylové grafy v původních souřadnicích generované např. v modulu *Korelace*. Mohou sloužit k posouzení homogenity dat.
	Biplot je projekce vícerozměrných dat do plochy (optimální z hlediska nejmenších čtverců). Body reprezentují řádky, paprsky odpovídají sloupcům. K identifikaci řádků lze použít označení bodů v interaktivním grafu. Při interpretaci grafu se vychází z toho, že aproximace původních dat úměrná vektorovému součinu jednotlivých bodů a úseček (bod reprezentuje konec vektoru s počátkem v bodě (0,0)). Z toho plyne, že blízké vektory řádků (body) nebo sloupců (paprsky) budou zřejmě vzájemně korelované. Vektory řádků, ležící ve směru některého vektoru sloupce budou mít v tomto sloupci vyšší, resp. nižší hodnoty. Znaménko (smysl vektoru) při tom nehraje roli. Je třeba brát v úvahu, že vzhledem k drastickému snížení počtu rozměrů jsou zvláště pro větší *m* tyto informace orientační a slouží spíše

	ke globálnímu posouzení struktury a možných souvislostí v datech.
	Graf vícerozměrné normality slouží k posouzení shody dat s vícerozměrným normálním rozdělením. Je obdobou Q-Q grafu v jednorozměrné analýze. V ideálním případě leží body na přímce. Využívá se F rozdělení Mahalanobisovy vzdálenosti. Při konstrukci grafu je použita korekce metodou jackknife pro zajištění nezávislosti $\mathbf{x}_i$ na průměru $\bar{\mathbf{x}}$.
	Graf symetrie je obdobou grafu polosum v jednorozměrné analýze. Jeden bod v grafu odpovídá vždy dvojici dat v datovém listu, proto se při označení v interaktivním grafu označí vždy dvě buňky v tabulce. V případě ideální symetrie rozdělení dané proměnné leží body zhruba na vyznačené přímce $y=-x$
	Andrewsovy křivky jsou účinným nástrojem k posouzení struktury vícerozměrných dat. Každá křivka reprezentuje jeden řádek dat (bod v *m*-rozměrném prostoru). Svazky křivek s podobným průběhem představují skupinu podobných dat. Jednotlivé křivky silně odlišné od ostatních jsou zřejmě vybočující měření, svazky křivek lišící se od ostatních představují shluky dat s odlišným chováním. K označení křivek se použije v interaktivním grafu kliknutí, resp. tažení myši. Se stisknutou klávesou *Ctrl* se označení zruší. Tímto postupem lze vybrat libovolné křivky a pak označit příslušné buňky v datovém listu.
	Klasická a robustní Mahalanobisova vzdálenost. Robustní vzdálenost (červeně) není ovlivněna vlivnými daty, může proto s výhodou posloužit k jejich diagnostice. Klasická vzdálenost (zeleně) je uvedena pro srovnání. Body nad vodorovnou přímkou (95% kvantilem T-rozdělení) lze považovat za podezřelé z vybočení. Označit lze pouze červené body (robustní).
	Alternativní diagnostika vybočujících a podezřelých bodů. Na ose x jsou hodnoty diagonálních prvků projekční matice (viz kapitola Lineární regrese), na ose y jsou rozdíly robustní a klasické Mahalanobisovy vzdálenosti z předchozího grafu. Body nad vodorovnou linií a vpravo od svislé linie lze považovat za podezřelé z vybočení.