VYSOKÁ ŠKOLA BÁŇSKÁ – TECHNICKÁ UNIVERZITA OSTRAVA

Hornicko – geologická fakulta

Katedra geoinformatiky

PROSTOROVÁ ANALÝZA DAT

Exploračí analýza dat

Lucie Orlíková

Ostrava, 2019

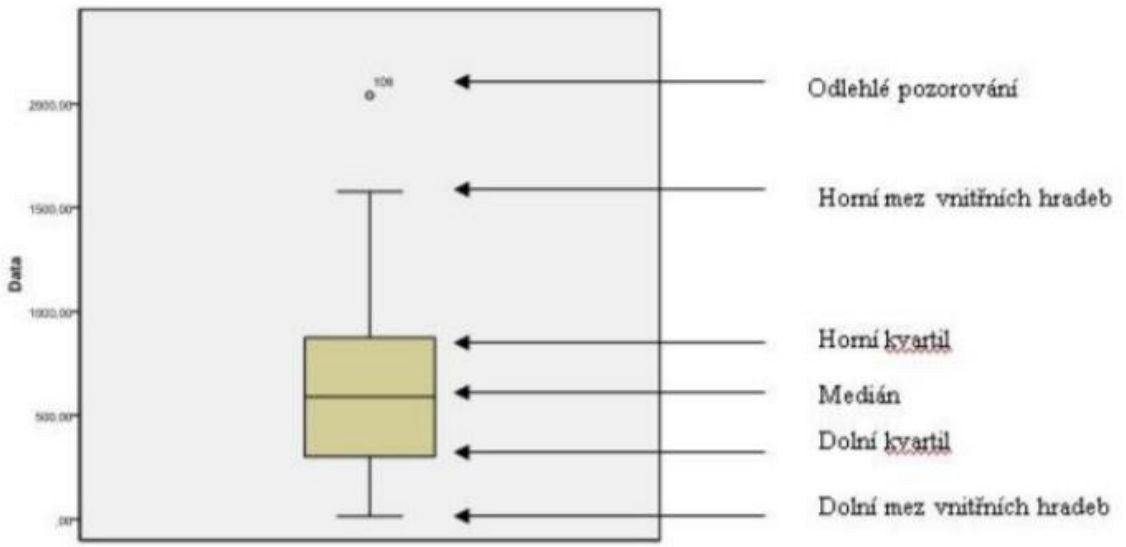
Úvod do problematiky

Explorační analýzou dat obecně rozumíme souhrn statistických metod používaných pro průzkum dat. Měla by být provedena vždy před přistoupením k samotné práci s daty. Mezi její základní operace ve vztahu k použití v GIS patří:

- studium statistického rozdělení datového souboru
- identifikace odlehlých hodnot, hledání jejich důvodu, případná eliminace
- identifikace globálních trendů ve vývoji hodnot

Před samotným statistickým zhodnocením dat je vhodné si data nejdříve znázornit v ploše, abyste si udělali hrubou představu o síti bodů a o plošném rozložení hodnot co do velikosti. Mezi základní statistické charakteristiky patří: minimum, maximum, průměr, medián, první kvartil, třetí kvartil, směrodatná odchylka, šikmost a špičatost.

Pro grafické znázornění kvantitativní proměnné se využívá několika metod – krabicový graf, histogram nebo kvantil-kvantilový graf. Krabicový graf se ve statistice využívá od roku 1977, kdy jej poprvé prezentoval americký statistik J. W.Tukey. Nazval jej "box with whiskers plot" – krabicový graf s vousama. Grafická podoba tohoto grafu se v různých aplikacích mírně liší. Jednu z jeho verzí vidíte na níže uvedeném obrázku.



Odlehlá pozorování jsou znázorněna jako izolované body, konec horního (popř. konec dolního) vousu představují maximum proměnné po vyloučení odlehlých pozorování, "víko"krabice udává horní kvartil, "dno" dolní kvartil, vodorovná úsečka uvnitř krabice označuje medián. Z polohy mediánu vzhledem ke "krabici" lze dobře usuzovat na symetrii vnitřních 50% dat a my tak získáváme dobrý přehled o středu a rozptýlenosti proměnné.

Histogram (angl. histogram) Histogram představuje grafické zobrazení intervalového členění kvantitativní proměnné. Umožňuje získat dobrou představu o struktuře dat. Odlehlá pozorování jsou hodnoty proměnné, které se výrazně liší od ostatních hodnot a ovlivňují tak vypovídající hodnotu průměru. Odlehlost pozorování může být způsobena hrubými chybami, překlepy, prokazatelným selháním techniky či lidí, vlivem chybného měření, technologických poruch apod. Jestliže známe příčinu

odlehlosti a předpokládáme, že již nenastane, můžeme tato pozorování vyloučit z dalšího zpracování. V ostatních případech je nutné zvážit, zda se vyloučením odlehlých hodnot nepřipravíme o důležité informace o jevech vyskytujících se s nízkou četností. Identifikovat odlehlá pozorování můžeme pomocí metody vnitřních hradeb: 1,5 násobek interkvartilového rozpětí.

Zadání programu:

- Vykreslete plochu studované oblasti s vyznačením hodnot sledované veličiny. Interpretujte vývoj hodnot.
- Proveďte základní statistické posouzení pro normální i logaritmované hodnoty.
- Vypočítejte základní charakteristiky za předpokladu normálního a lognormálního rozdělení (aritmetický průměr, rozptyl, koef. šikmosti a koef. špičatosti, kvantilové charakteristiky). Popište typ distribuce (jednoduchá nebo smíšená, v případě smíšené odhadnout rozpětí a střední hodnotu dílčích distribucí).
- Pro normální i logaritmované hodnoty vykreslete histogram a rankitový graf (probability plot). Interpretujte.

Postup práce v prostředí R

Data:

Data v souboru Doubrava.xls, v tomto případě budete pracovat v proměnnou **kal**

Data jsou z vrtů realizovaných v odkališti Pilňok u dolu Doubrava jako střední hodnoty příslušných vlastností kalu.

X a Y	relativní souřadnice X a Y v metrech,
Pisek	obsah písčité frakce (průměr zrn > 0.063 mm) v %,
ALPE	obsah prachové frakce (průměr zrn od 0.004 do 0.063 mm) v %,
kal	obsah aleuropelitické frakce (průměr zrn pod 0.004 mm) v %,
Ad	popelnatost v bezvodém vzorku v %,
W	vlhkost v %,
S	obsah síry v %.

#Nastavení pracovního adresáře na disk, kde máte uložená data

```
setwd("D:/PAD/Geostatistika")
```

#Načtení knihovny pro excel a import dat

```
File – Import Dataset – From Excel
```

#Zapnutí knihoven pro práci s prostorovými daty (nutné předem instalovat jako Pluginy)

```
library(ggplot2)

library(scales)

library(moments)
```

#Vykreslení histogramu hodnot a výpočet základní popisné statistiky.

```
hist(Doubrava$Kal)
summary(Doubrava$Kal)
```

#Výpočet koeficientu šikmosti a špičatosti. Koeficient šikmosti popisuje nesymetrii souboru. Pokud se pohybuje okolo 0, jsou hodnoty rozděleny rovnoměrně vpravo i vlevo. Kladná šikmost znamená, že v pravo od průměru se vyskytují odlehlejší hodnoty, většina hodnot je vlevo. Záporná šikmost znamená , že se hodnoty vyskytují většinou vpravo. Koeficient špičatosti pak porovnává dané rozdělení s norm álním rozdělením náhodné veličiny. Objektivní koeficient špičatosti je čtvrtým normovaným moment em zmenšeným o hodnotu 3 (nutné odečíst tedy hodnotu 3)

```
skewness(Doubrava$Kal)
kurtosis(Doubrava$Kal)
```

#Pro testování normality použijeme Shapiro-Wilkův test normality. Testuje se hypotéza, že náhodný výběr pochází z normálního rozdělení. Pokud je hodnota p-value menší než 0,05, pak zamítáme nulovou hypotézu. Dle výsledku koeficientu šikmosti a testu normality se zamítá nulová hypotéza o normálním rozdělení. Dalším krokem bude testování logaritmovaných hodnot.

```
shapiro.test(Doubrava$Kal)
```

#Pro vykreslení kvantil-kvantilového grafu je nutné mít nainstalovanou knihovnu **car.** Pravděpodobnostní křivka pro dané rozdělení není zcela přímková. Křivka mění svůj směr, což znamená změnu distribuce, z toho lze usoudit, že distribuce je smíšená.

```
library(car)
qqPlot(Doubrava$Kal)
```

#Vykreslení plochy studované oblasti. Z výsledného obrázku je patrné, že rozložení hodnot kalové frakce v odkališti není pravidelné. I když se některé hodnoty prolínají je vidět náznak růstu hodnot od severozápadu k jihovýchodu. V blízkosti středu a na severovýchodě je oblast bez hodnot.

```
Doubrava$cat <- cut(Doubrava$Kal, breaks = c(13,53,60,74,92))
hodnota.plot <- ggplot(aes(x = X, y = Y), data = Doubrava)
hodnota.plot <- hodnota.plot + geom_point(aes(color = cat))
hodnota.plot <- hodnota.plot + coord_equal()
hodnota.plot <- hodnota.plot + scale_color_brewer(palette = "YlGnBu")
hodnota.plot
```

#Převod hodnot do logaritmické škály a testovány normality.

```
Doubrava$log <- log(Doubrava$Kal)
hist(Doubrava$log)
summary(Doubrava$log)
shapiro.test(Doubrava$log)
qqPlot(Doubrava$log)
```

#Vzhledem k tomu, že oba koeficienty šikmosti i špičatosti vycházejí hůře než v případě hodnot bez logaritmizace, v dalších krocích se bude pracovat s původními hodnotami.