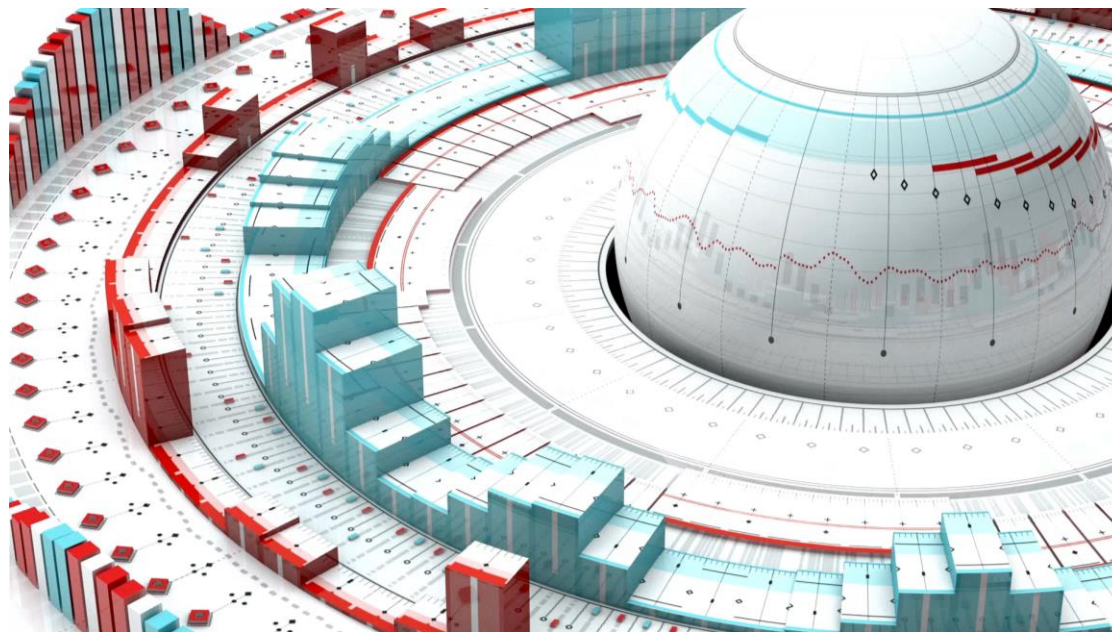




Kvalitatívne a kvantitatívne analytické metódy a techniky II.

Beáta Stehlíková

Bratislava

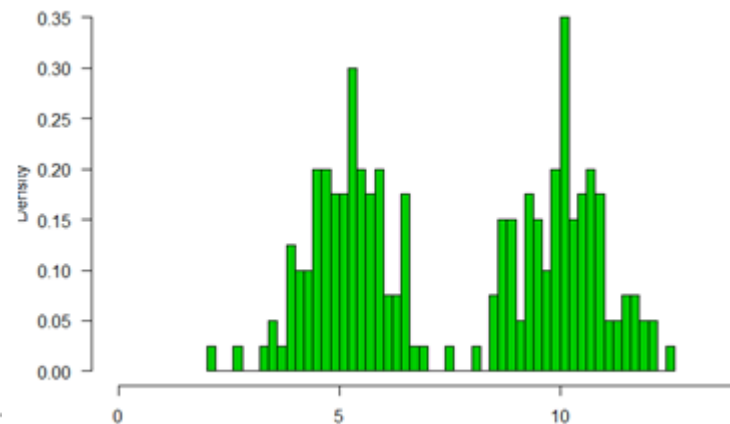
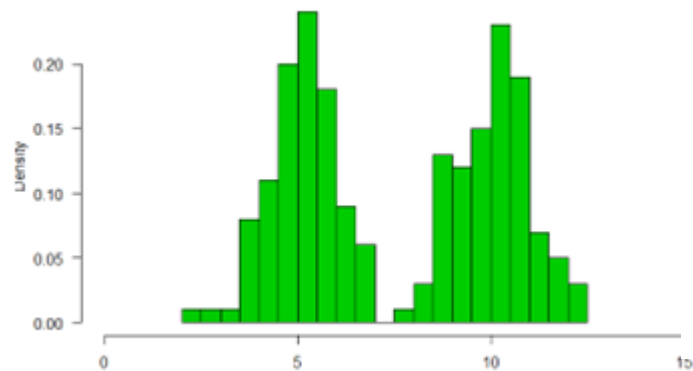
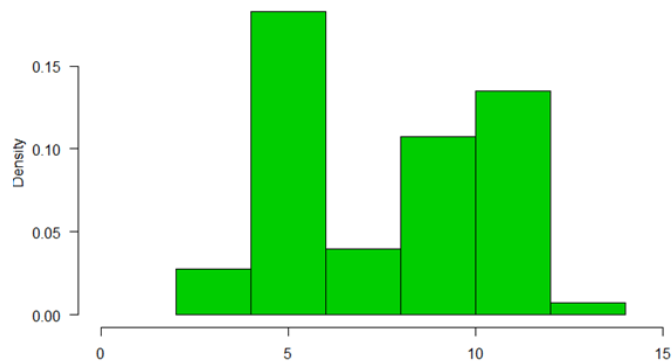
2024

Obsah

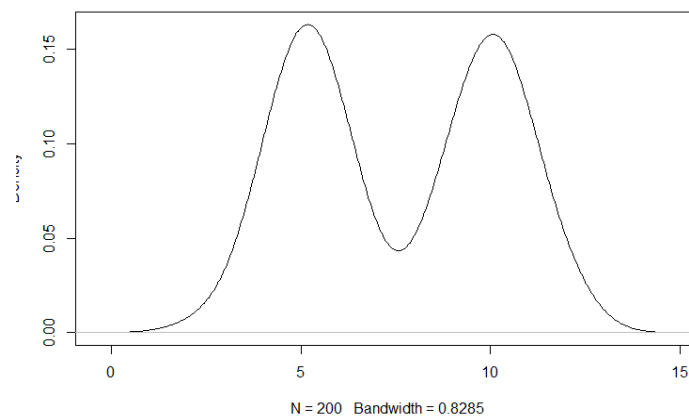
- Popis premenných
- Závislosti
- Podobné objekty



-histogram, jadrová funkcia hustoty

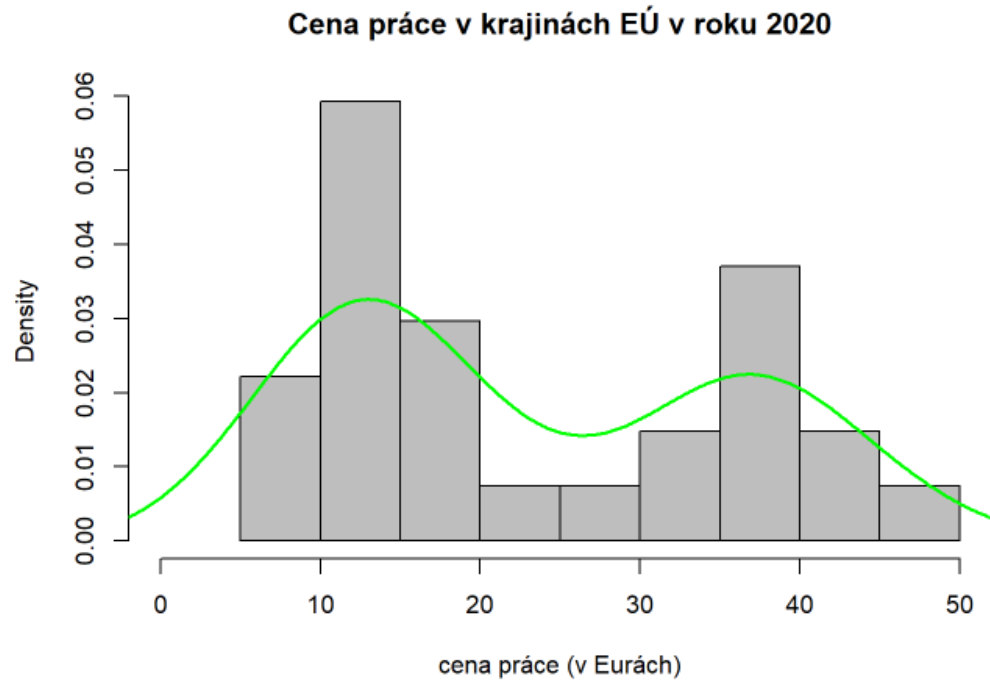


`density.default(x = normalne$c, na.rm = TRUE)`





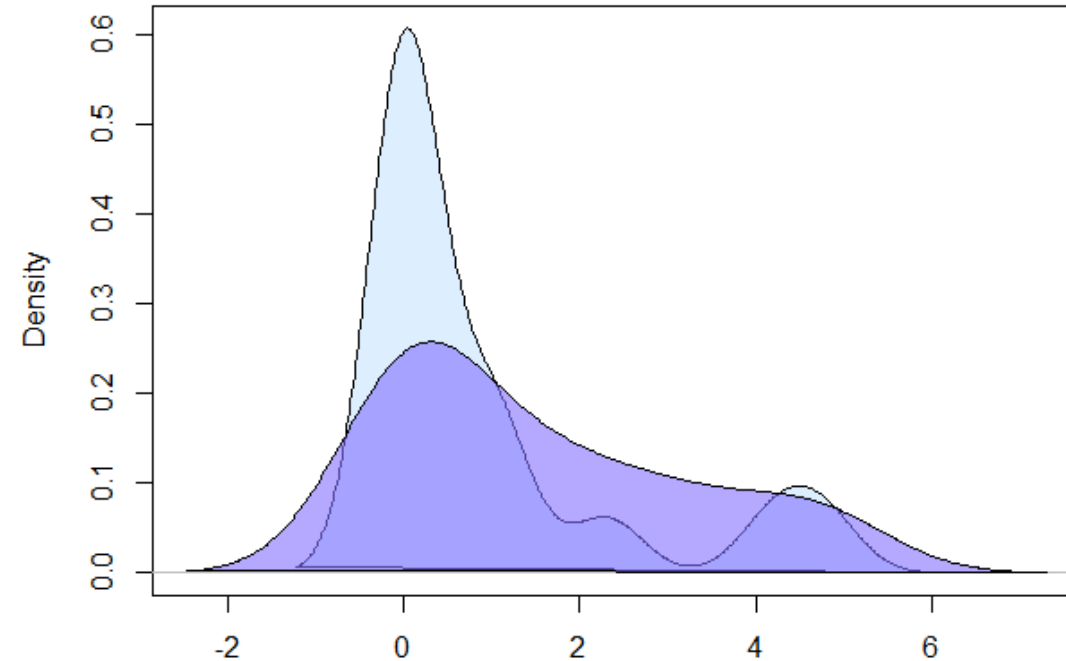
Jadrová funkcia hustoty



Ameijeirasov-Alonsov test

<https://rpubs.com/BeaSte/856160>

Global Terrorism Index (GTI)



Ujvary, K. (2022) POUŽÍVANIE KOMUNIKAČNÝCH PROSTRIEDKOV A INFORMAČNÝCH SYSTÉMOV NA IDENTIFIKÁCIU A ELIMINÁCIU TERORIZMU V SPOJITOSTI S NELEGÁLNOU MIGRÁCIU, dizertačná práca



Pravdepodobnostné rozdelenia

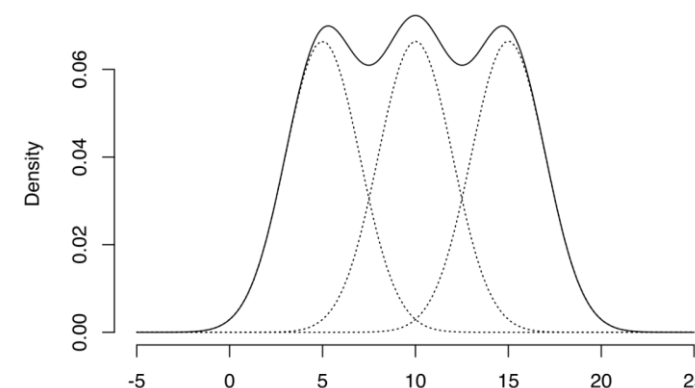
Pred aplikáciou ktorejkoľvek z techník, je dôležité vizuálne

preskúmať pravdepodobnostné rozdelenie

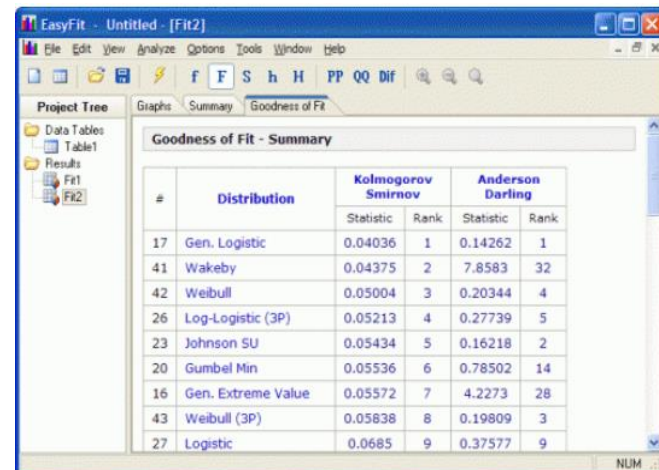
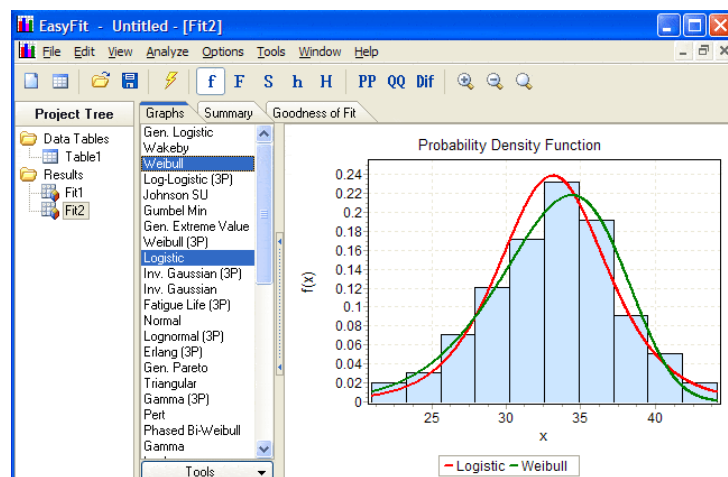
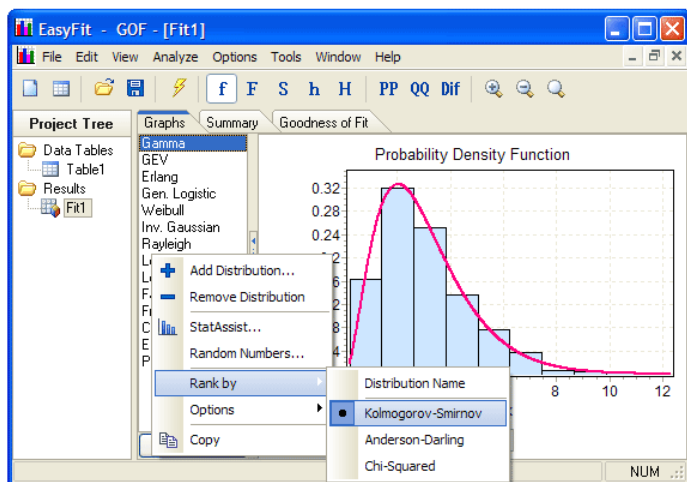
Distribučná funkcia *plne* charakterizuje pravdepodobnostné správanie náhodnej premennej. Pravdepodobnostné rozdelenia:

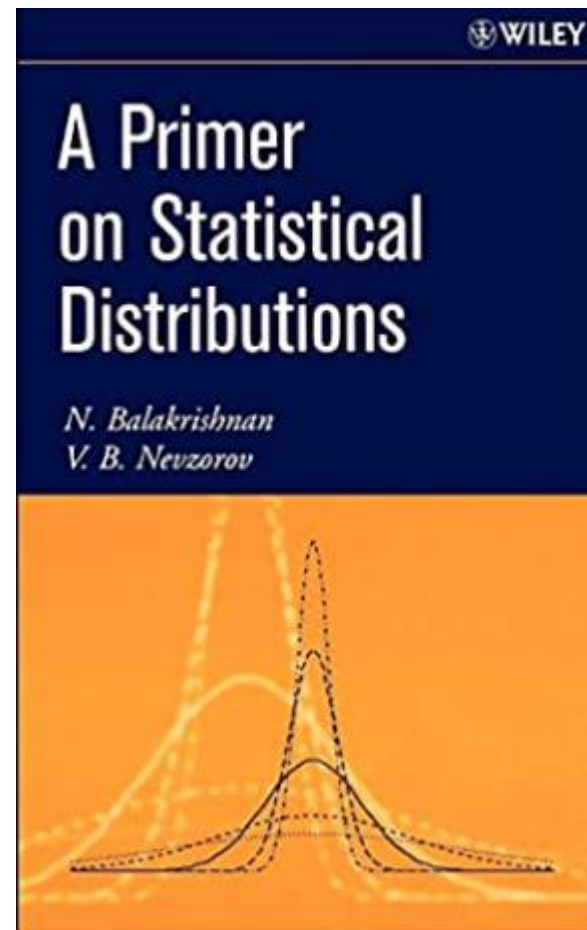
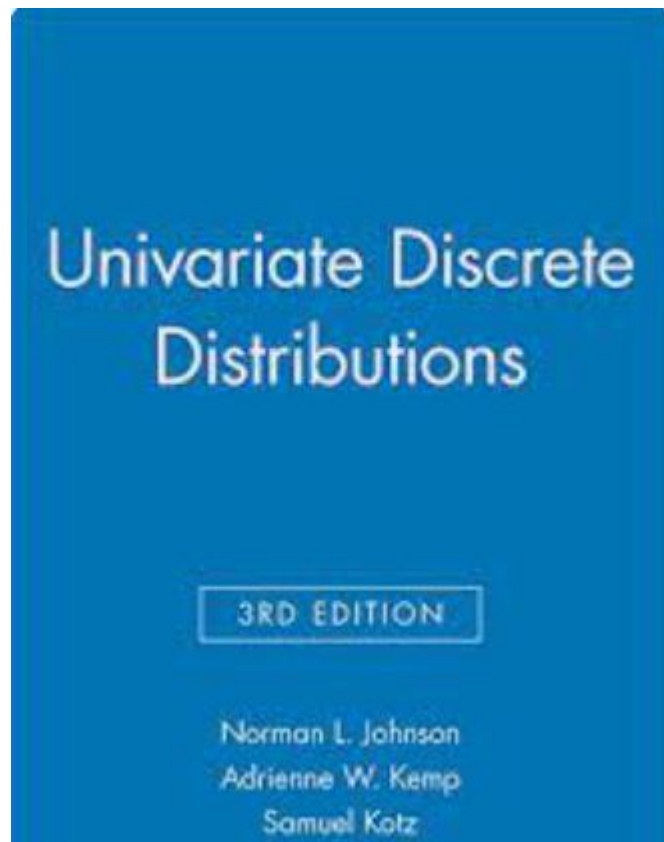
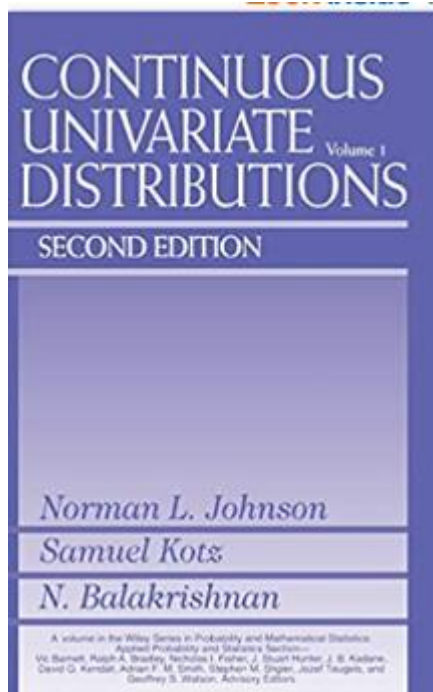
- diskkrétne (rovnomerné , Bernoulliho, binomické, Poissonovo, geometrické, hypergeometrické, negatívne binomické rozdelenie).
- spojité (rovnomerné rozdelenie na intervale, normálne, exponenciálne, Studentovo t-rozdelenie, chí-kvadrát rozdelenie),
- zmiešané náhodné premenné,
- zmesy rozdelení

Zmesy - ukážka



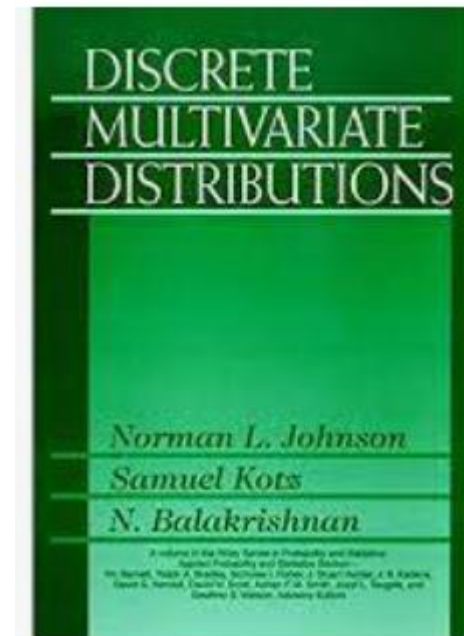
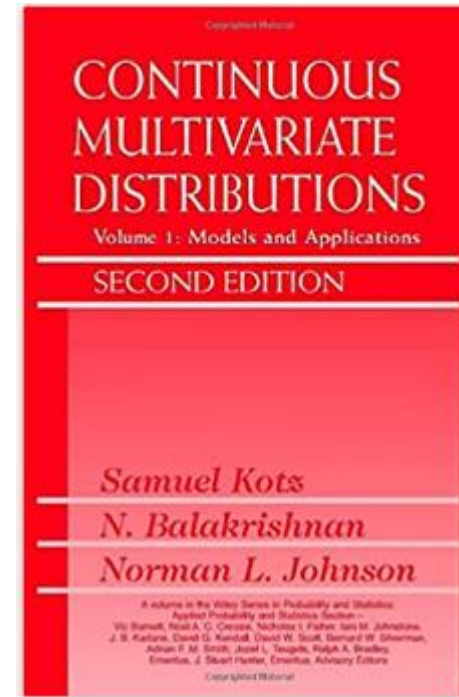
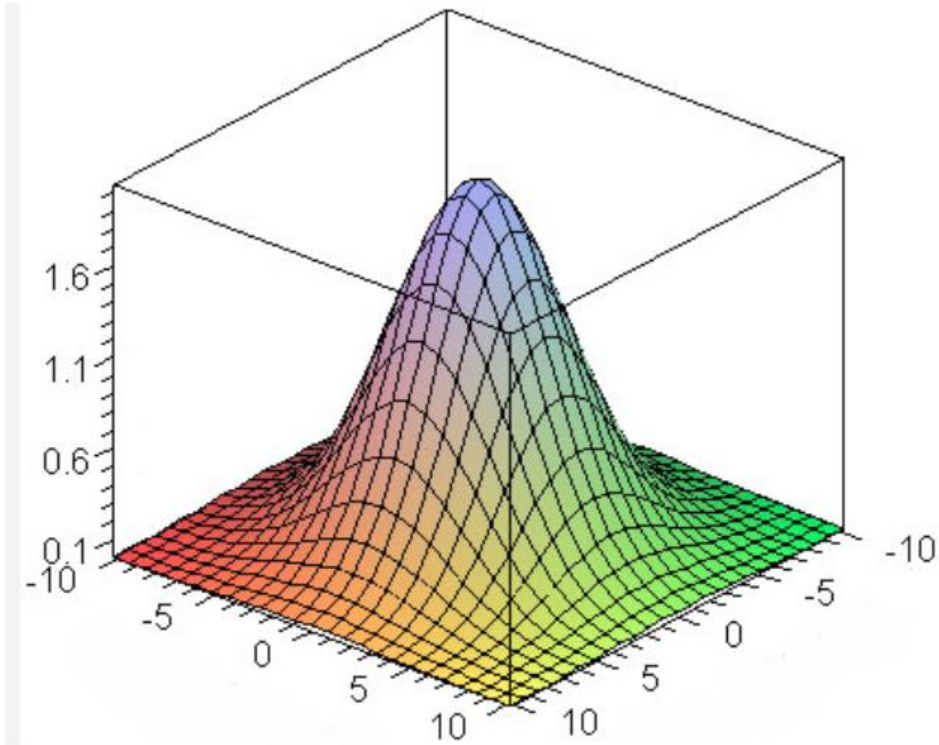
Testy dobrej zhody (EasyFit – 40 rozdelení)







Viacrozmerné rozdelenia



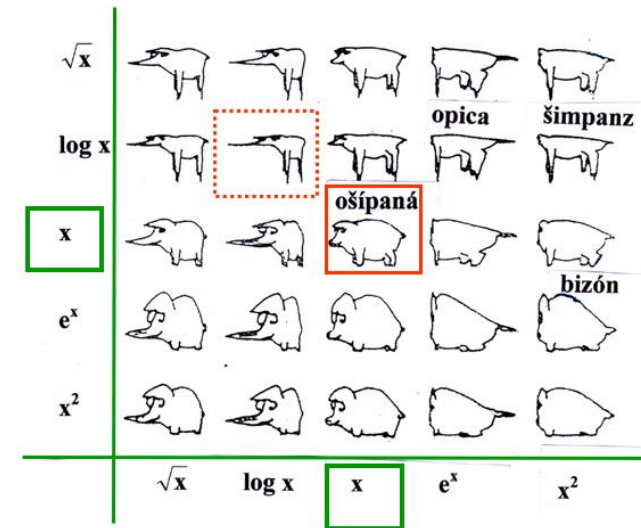
<https://ardianumam.wordpress.com/2017/10/19/deriving-gaussian-distribution/>



Transformácia údajov

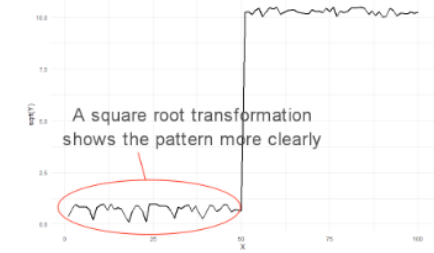
(spracovane podľa <https://www.spss-tutorials.com/normalizing-variable-transformations/>)

- Transformácia je matematická operácia, ktorá mení mierku merania premennej. Zvyčajne sa to robí, aby sa súbor stal použiteľným s konkrétnym štatistickým testom alebo metódou.
- Mnohé štatistické metódy vyžadujú údaje, určitý druh rozdelenia, zvyčajne normálne

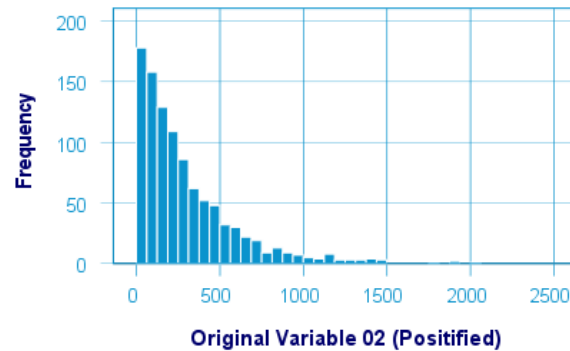




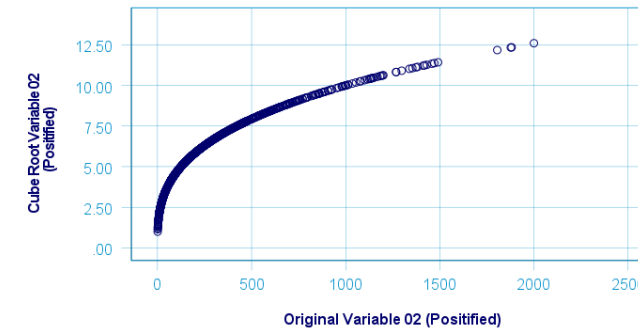
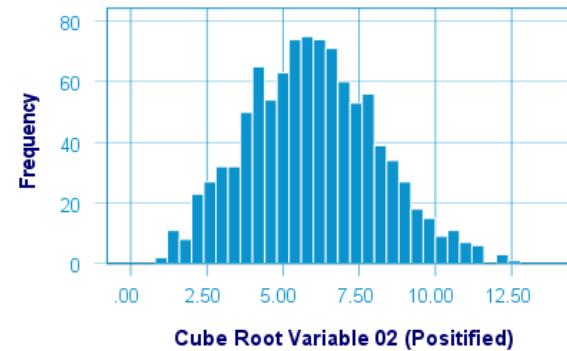
Odmocninová transformácia



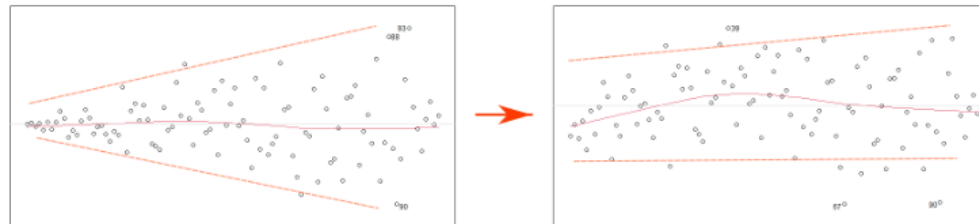
Original Variable 02 (Positifed)



Cube Root Variable 02 (Positifed)



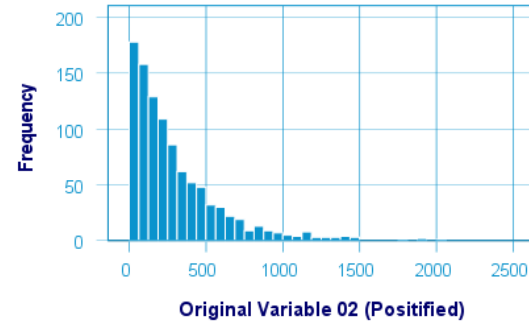
- **normalizácia zošikmeného rozdelenia** (premenné s ľavým zošikmením sa stanú najhoršími po transformácii druhej odmocniny)
- **zníženie heteroskedasticity**
- **linearizáciu vzťahu**
- **lepšiu vizualizáciu**



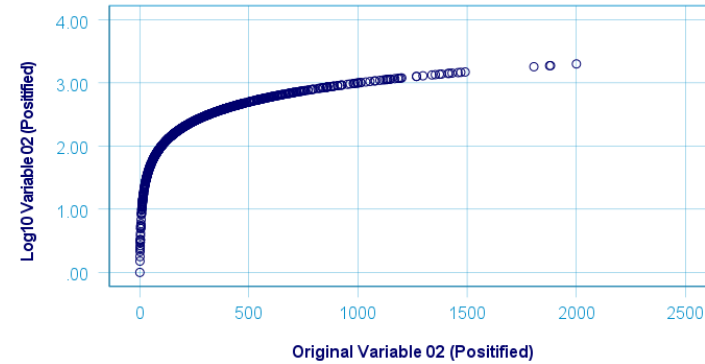
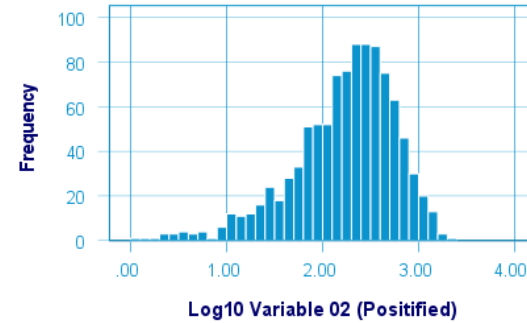
- **Poznámka:** pre záporné hodnoty: `if (pos01 < 0) curt01 = -abs(pos01)**(1/3)`

Logaritmická transformácia

Original Variable 02 (Positifed)



Log10 Variable 02 (Positifed)





$$Y = 5 + 0.03X$$

Priesečník v tejto rovnici je 5. To znamená, že keď je nezávislá premenná (x) 0, závislá premenná (Y) je 5. Koeficient x je 0,03, čo znamená, že závislá premenná (Y) by sa zvýšila o 0,03 pre každé zvýšenie o 1 jednotku nezávislej premennej (x).

Log — Level Regression

Logaritmicкая regresia je model, kde cieľová premenná je logaritmickey transformovaná, ale prediktorové premenné nie. Aby sme mohli interpretovať koeficienty logaritmickej regresie, musíme najprv umocniť koeficienty nezávislých premenných so základom e. Pozrime sa na nasledujúci výpočet, aby sme videli, ako zmena nezávislej premennej zmení závislú premennú.

$$\ln(Y_2) - \ln(Y_1) = (5 + 0.03x_2) - (5 + 0.03x_1)$$

$$\ln(Y_2) - \ln(Y_1) = 5 + 0.03x_2 - 5 - 0.03x_1$$

$$\ln(Y_2/Y_1) = 0.03(x_2 - x_1)$$

$$Y_2/Y_1 = e^{0.03(X_2 - X_1)}$$

$$Y_2/Y_1 = 1.03^{(X_2 - X_1)}$$

Takže na základe tejto rovnice môžeme dospieť k záveru, že pri každom zvýšení našej nezávislej premennej (x) o jednu jednotku sa závislá premenná (Y) zvýši o 3 % ($1,03^1 = 1,03$). Dôležité je, že toto zvýšenie musí byť **fixnou hodnotou**. Ak zväčšíme x o 10 jednotiek, zmena Y by sa rovnala 34 % ($1,03^{10} = 1,34$).

Level -Log Regression

Úroveň-log regresia je model, kde sa jedna alebo viac nezávislých premenných log-transformuje, ale závislá premenná zostáva pôvodná. Pred analýzou toho, čo znamená jeho koeficient, poďme sa ponoriť do výpočtu nižšie —

$$Y_2 - Y_1 = (5 + 0.03 \ln(x_2)) - (5 + 0.03 \ln(x_1))$$

$$Y_2 - Y_1 = 5 + 0.03 \ln(x_2) - 5 - 0.03 \ln(x_1)$$

$$Y_2 - Y_1 = 0.03(\ln(x_2) - \ln(x_1))$$

$$Y_2 - Y_1 = 0.03(\ln(x_2/x_1))$$

Podľa tejto rovnice môžeme predpokladať, že pri **fixnej percentuálnej** zmene hodnoty našej nezávislej premennej (x) sa naša závislá premenná (Y) zmení o $0,03 \cdot \ln(x_2/x_1)$ percentuálnej zmeny. Napríklad, ak sa x zvýši o 10 %, Y sa zvýši o 0,3 % ($0,03 \cdot \ln(1,1) = 0,003$). Ak sa x zvýši o 50 %, Y sa zvýši o 1,2 % ($0,03 \cdot \ln(1,5) = 0,012$). Pokiaľ je percentuálna zmena v x pevná, Y sa tiež zmení o pevné percento.

<https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9>

Ako interpretovať výsledky regresie transformovaných údajov

Log-Log

Log-log regresia je model, kde cieľová premenná a aspoň jedna prediktorová premenná sú log-transformované. Podobne ako pri regresii na úrovni logaritmu odstránime logaritmus z ľavej strany rovnice a pravú stranu umocníme základom e. Pozrime sa -

$$\ln(Y_2) - \ln(Y_1) = (5 + 0.03 \ln(x_2)) - (5 + 0.03 \ln(x_1))$$

$$\ln(Y_2) - \ln(Y_1) = 5 + 0.03 \ln(x_2) - 5 - 0.03 \ln(x_1)$$

$$\ln(Y_2/Y_1) = 0.03(\ln(x_2) - \ln(x_1))$$

$$\ln(Y_2/Y_1) = 0.03(\ln(x_2/x_1))$$

$$Y_2/Y_1 = e^{0.03(\ln(x_2/x_1))}$$

$$Y_2/Y_1 = e^{(\ln(x_2/x_1)) \cdot 0.03}$$

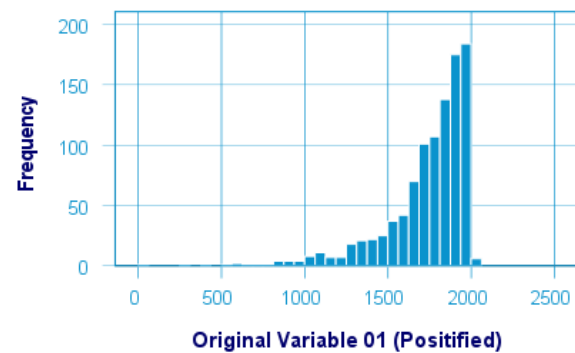
$$Y_2/Y_1 = (x_2/x_1)^{0.03}$$

Táto rovnica nám hovorí, že pre **pevné percento** zmenené v našej nezávislej premennej (x) by sa naša závislá premenná (Y) zmenila o x percentuálnu zmenu na mocninu 0,03. Ak sa x zmení o 10 %, y sa zmení o 0,3 % ($1,1^{0.03} = 1,003$). Ak sa x zmení o 20 %, y sa zmení o 0,5 % ($1,2^{0.03} = 1,005$).

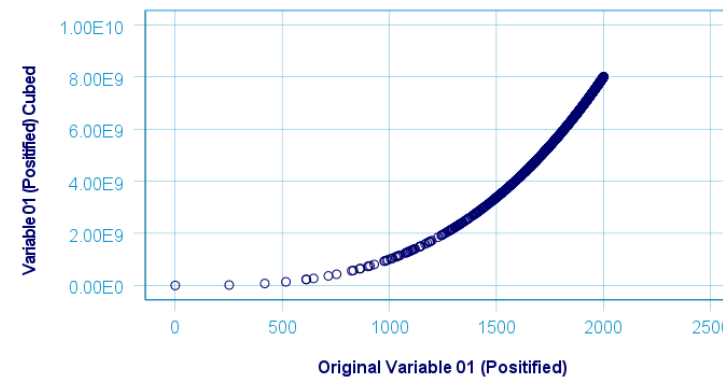
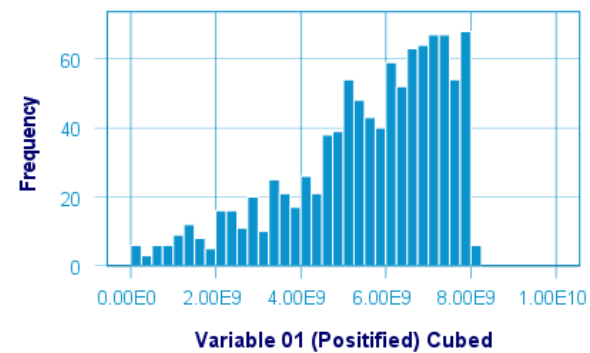


Mocninová transformácia

Original Variable 01 (Positifed)



Variable 01 (Positifed) Cubed





Box-Coxova transformácia

$$B(x, \lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \ln(x), & \text{if } \lambda = 0. \end{cases}$$

https://www.researchgate.net/publication/268412346_The_Box-Cox_Transformation_Technique_A_Review

Citované 1137-krát

<http://eprints.lse.ac.uk/103537/1/StatSciV4.pdf>

<https://www.le.ac.uk/users/dsgp1/COURSES/TOPICS/BoxCox.pdf>

Ďalšie informácie

<https://www.datanovia.com/en/lessons/transform-data-to-normal-distribution-in-r/>

Výber najvhodnejšej transformácie:

<https://cran.r-project.org/web/packages/bestNormalize/vignettes/bestNormalize.html>



Nie všetky dáta môžu byť transformované

- Symetrické dáta s „ťažšími chvostami" ako má normálne rozdelenie (väčšie množstvo dát na oboch koncoch rozdelenia v porovnaní s normálnym rozdelením).
- Bimodálne (multimodálne) rozdelenia
- Keď je prítomné väčšie množstvo identických pozorovaní

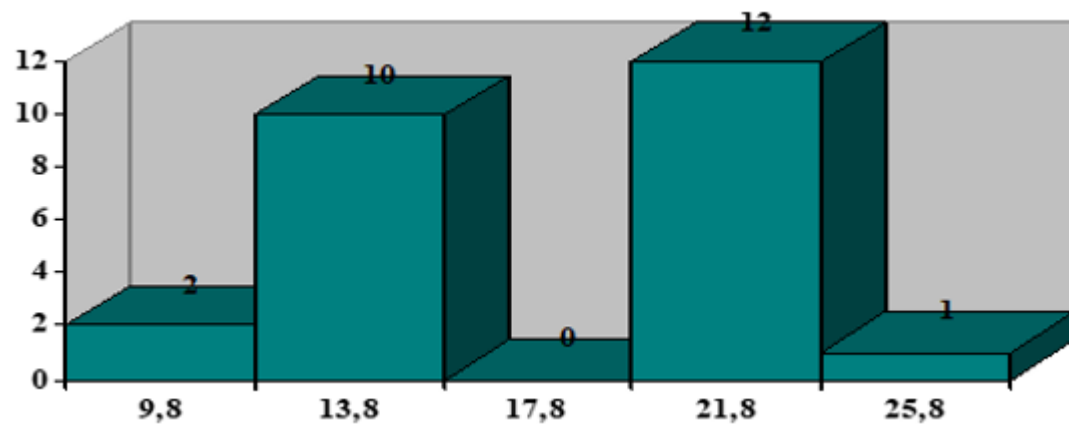
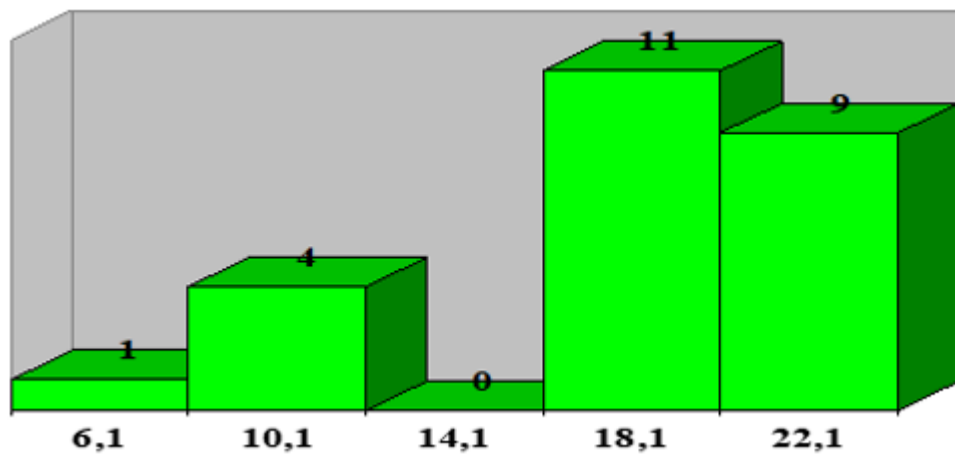
Ktorá je najvhodnejšia:

<https://cran.r-project.org/web/packages/bestNormalize/vignettes/bestNormalize.html>



Štatistické charakteristiky

ŠKÁLA	ŠTATISTICKÉ CHARAKTERISTIKY
Nominálna	Modus, absolútne a relatívne početnosti
Ordinálna	Modus, absolútne a relatívne početnosti medián, kvartily, kvantily, percentily
Intervalová	Modus, absolútne a relatívne početnosti medián, kvartily, kvantily, percentily aritmetický priemer, rozptyl, smerodajná odchýlka, špicatosť, šikmosť
Pomerová	Modus, absolútne a relatívne početnosti medián, kvartily, kvantily, percentily aritmetický priemer, rozptyl, smerodajná odchýlka, špicatosť, šikmosť Geometrický priemer

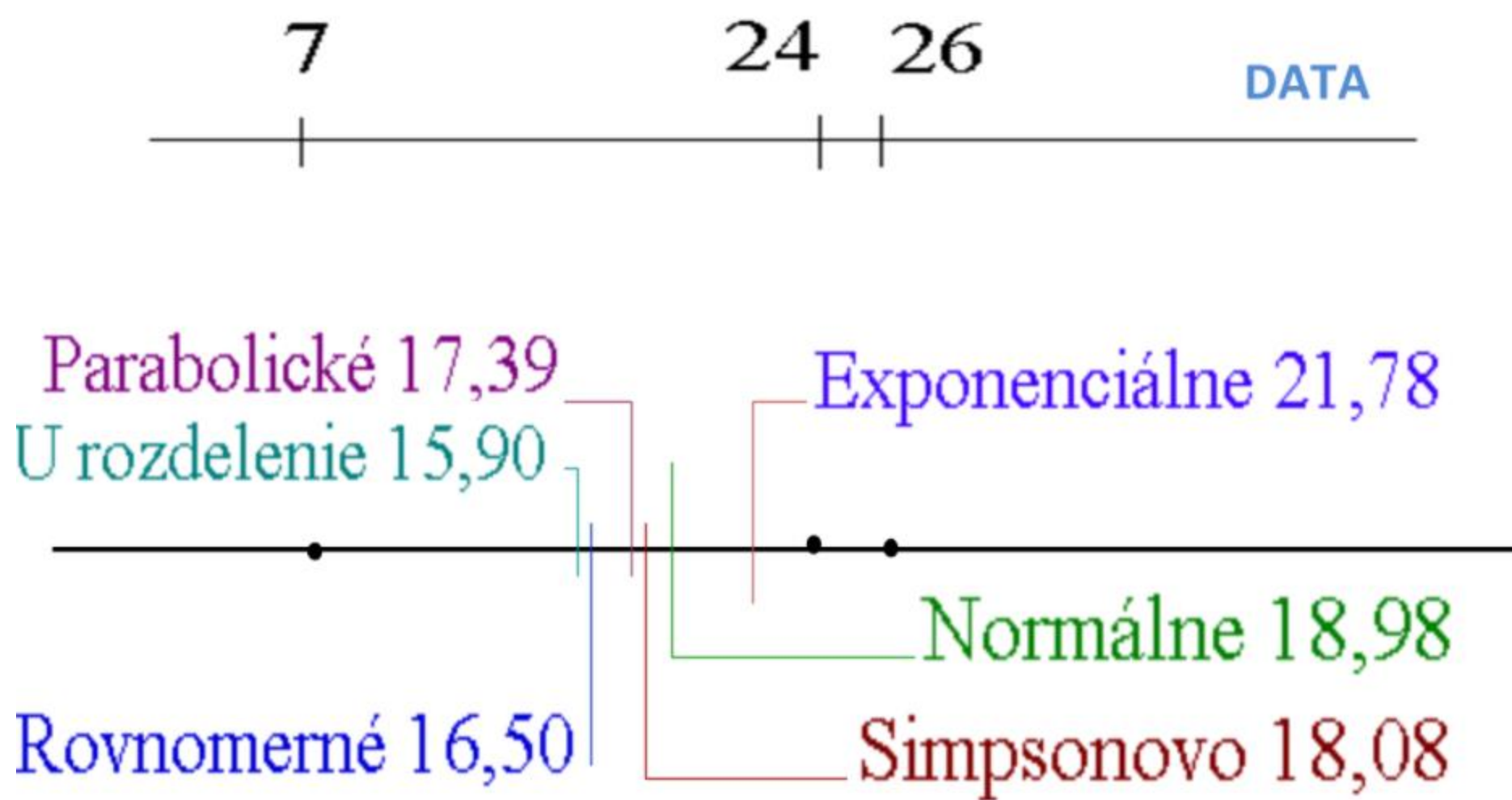


Rozsah výberu $n = 25$

Priemer 17,8

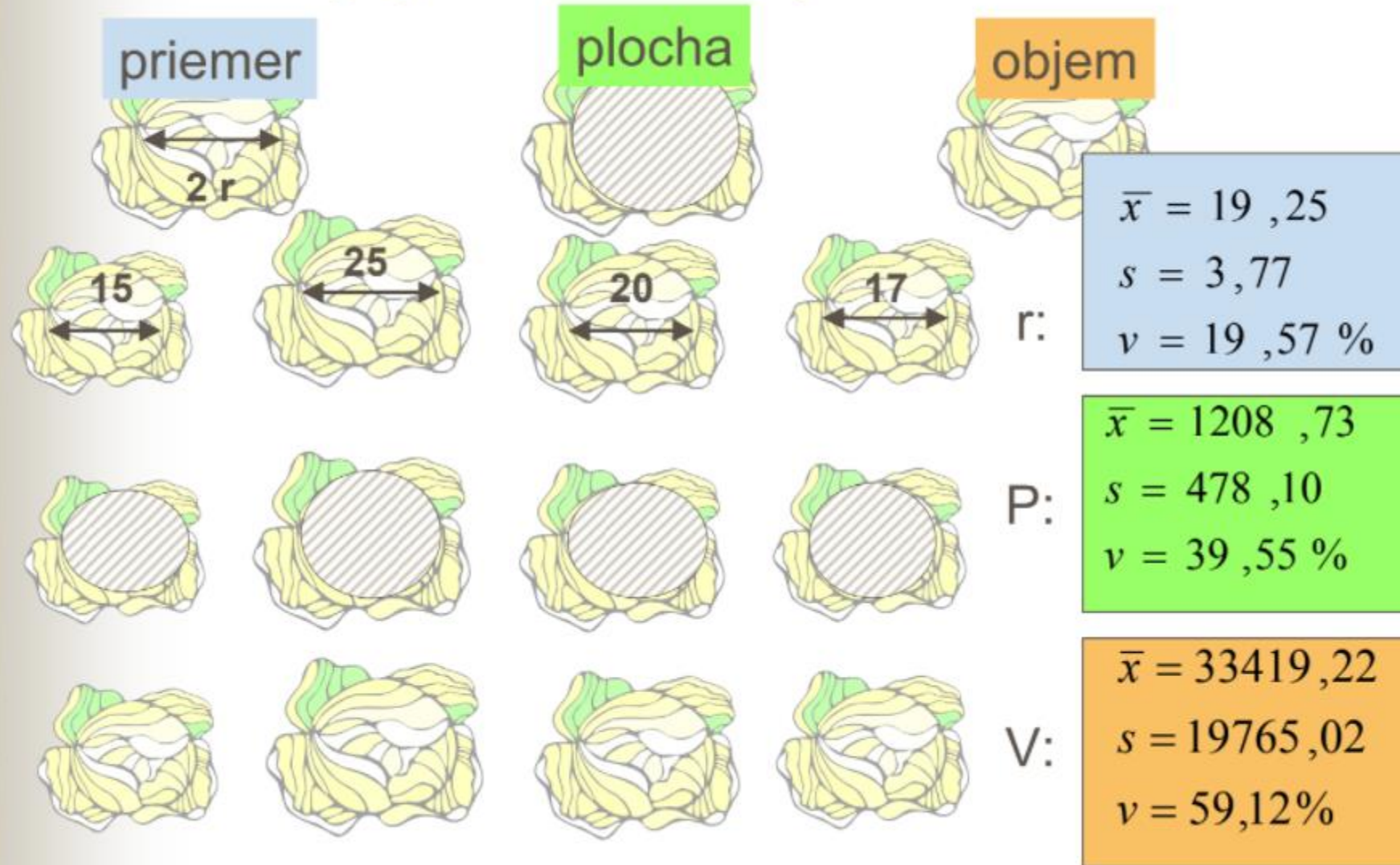
Štandardná odchýlka 4,7

Variačný koeficient 26,4



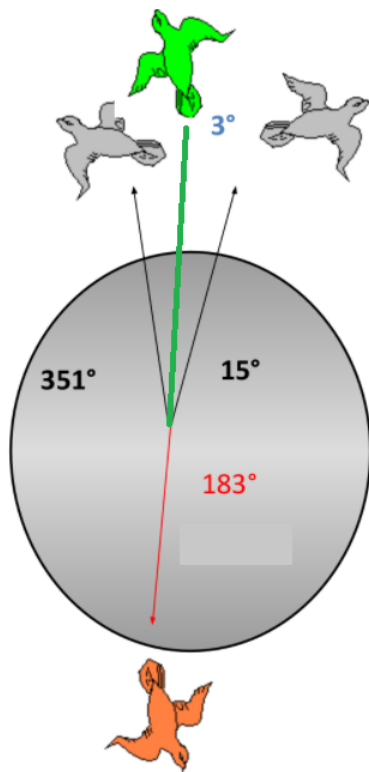


Nie je jedno, aké znaky meriame





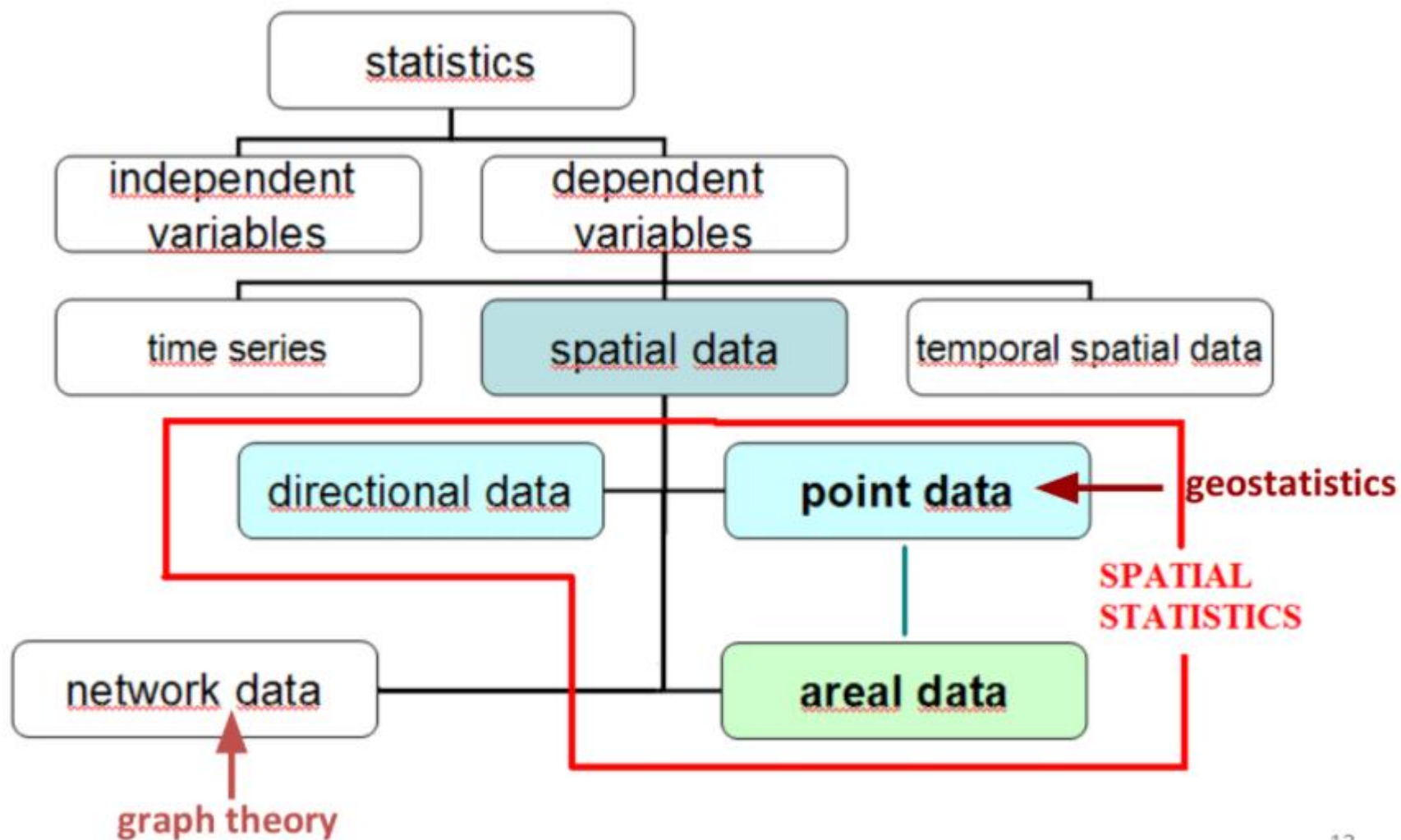
Smerové údaje



$$\bar{x} = \frac{351^\circ + 15^\circ}{2} = 183$$

$$\bar{x} = \arctg \frac{\sum \sin x_i}{\sum \cos x_i}$$

$$\bar{x} = 3^\circ \quad \checkmark$$





Kompozičné údaje (Compositional data)

- V štatistike sú kompozičné údaje kvantitatívnym opisom častí nejakého celku, ktorý poskytuje **relatívne informácie**.
- Matematicky sú kompozičné údaje reprezentované bodmi na simplexnej báze.

The arithmetic mean of the compositional data set is usually not representative of the set's 'centre'. The geometric mean

$$\text{cen}(\mathbf{X}) = \frac{(g_1, \dots, g_D)}{g_1 + \dots + g_D}, \quad (1)$$

where $g_j = \left(\prod_{i=1}^N x_{ij} \right)^{\frac{1}{N}}$ is the geometric mean of the j^{th} component (i.e. column-wise geometric mean) of the data set, is more representative.

Fernández *et al.*, 1998). We use centered log-ratio transformation (clr) of the composition, which is given by

$$\text{clr}(\mathbf{x}) = \left(\log\left(\frac{x_1}{g(\mathbf{x})}\right), \dots, \log\left(\frac{x_D}{g(\mathbf{x})}\right) \right), \quad (2)$$

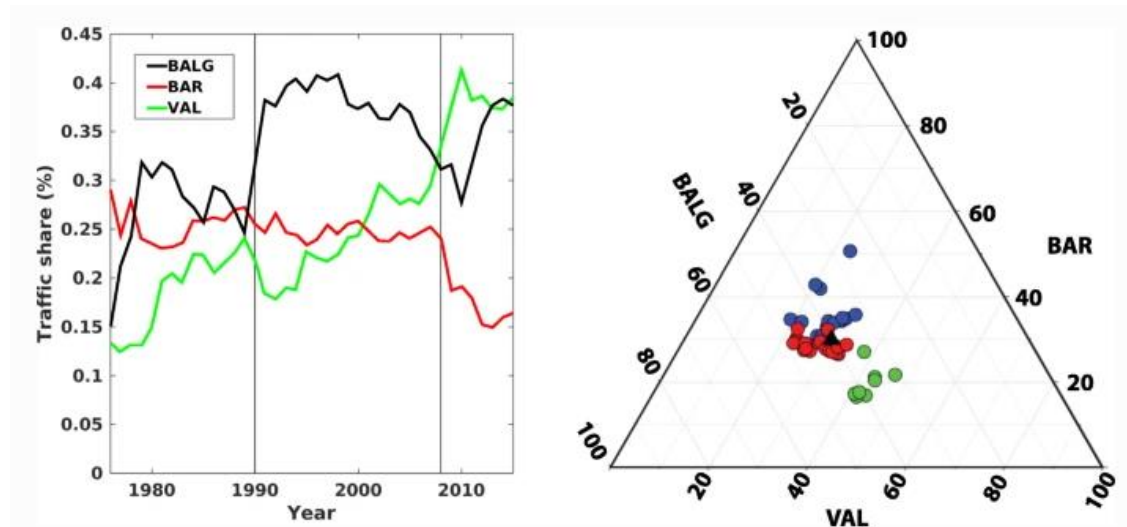
where $g(\mathbf{x})$ is the geometric mean of the composition (row-wise geometric mean).



Kompozičné údaje (Compositional data)

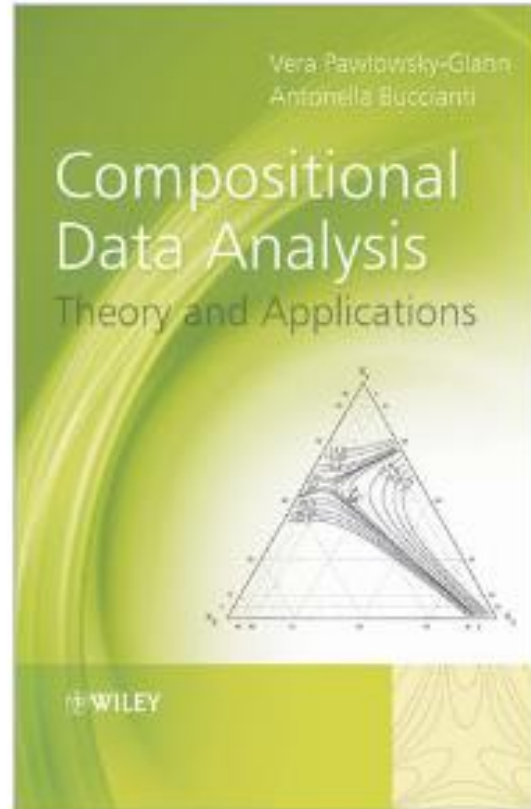
Na skúmanie vývoja podielu dopravy sa používajú štatistické techniky založené na údajoch o zložení priepustnosti kontajnerov v systéme s viacerými portami. Kompozičné vektory sú tie, ktoré obsahujú relatívnu informáciu častí nejakého celku. **Aplikácia konvenčných štatistických techník na kompozičné údaje môže viesť k chybným záverom a falošným koreláciám.**

tri najväčšie prístavy:
Algeciras Bay (BALG),
Barcelona (BAR) a
Valencia (VAL)



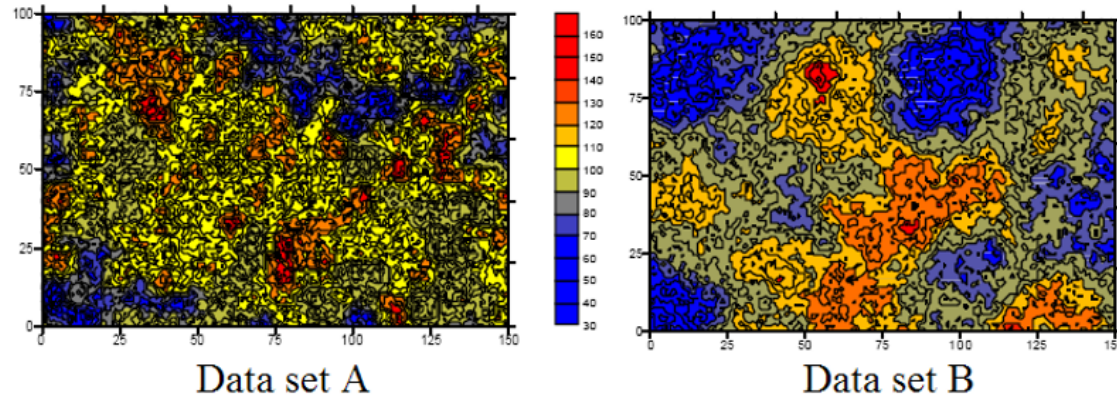
<https://etr.springeropen.com/articles/10.1186/s12544-019-0350-z>

Kompozičné údaje (Compositional data)

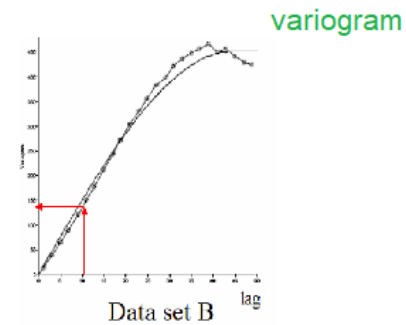
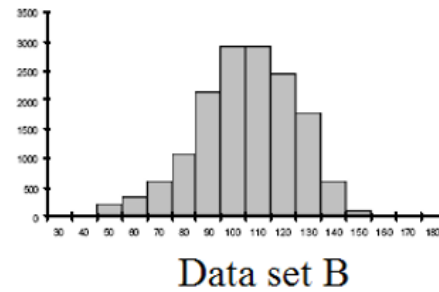
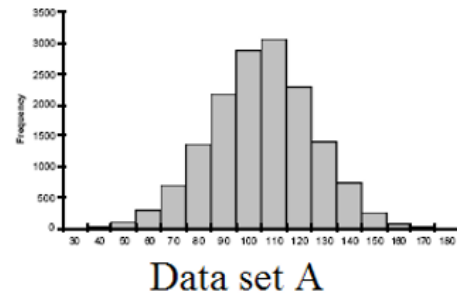


<https://www.wiley.com/en-us/Compositional+Data+Analysis:+Theory+and+Applications-p-9780470711354>

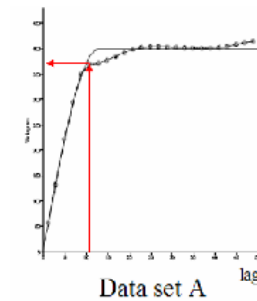
Priestorové údaje



	A	B
Count	15251	15251
Average	100.00	100.00
Standard Deviation	20.00	20.00
Median	100.35	100.92
10 Percentile	73.89	73.95
90 Percentile	125.61	124.72



Small values at short lags
indicate that data are
autocorrelated.

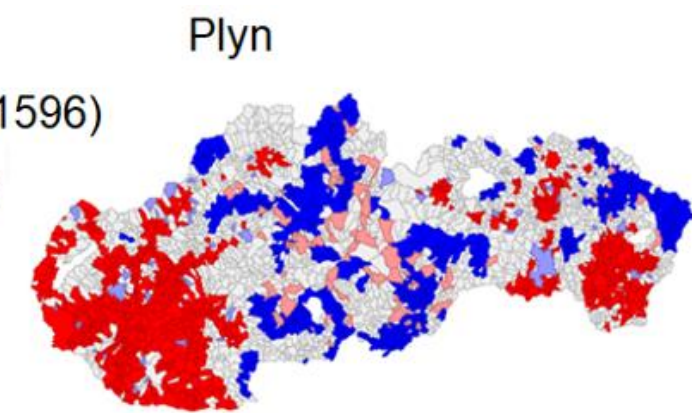
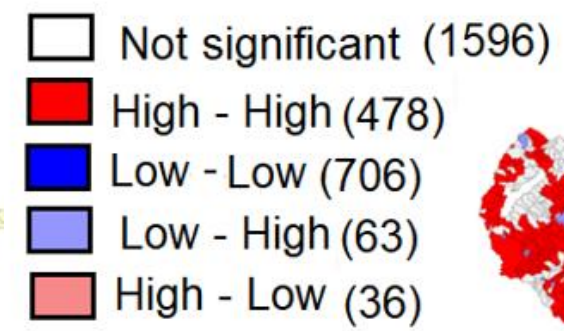
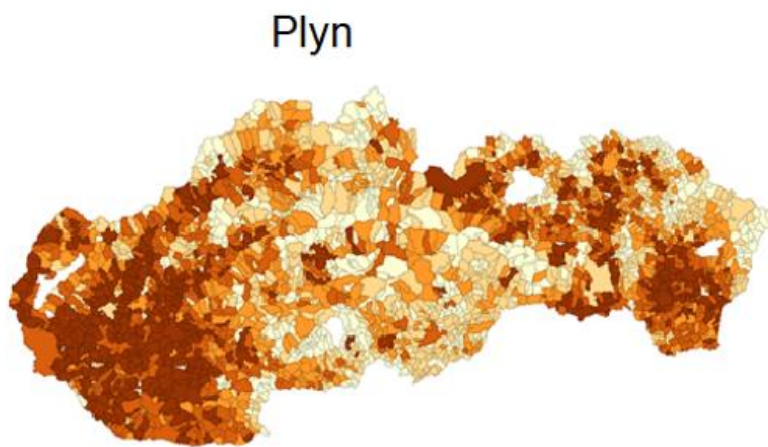
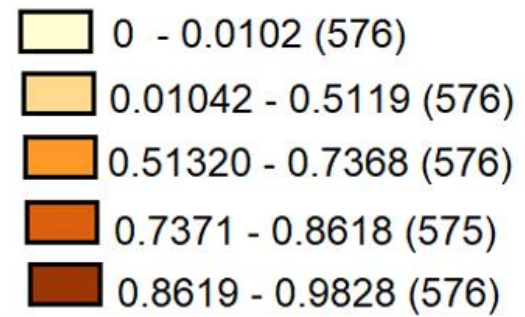
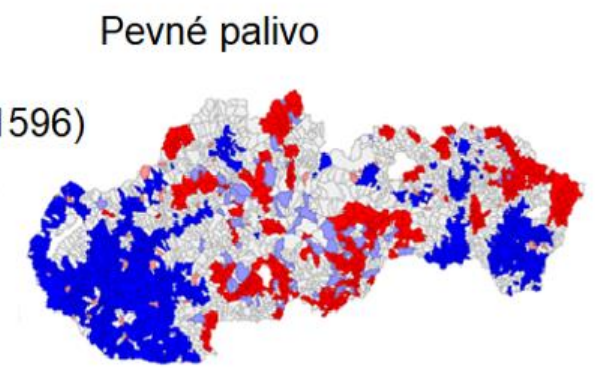
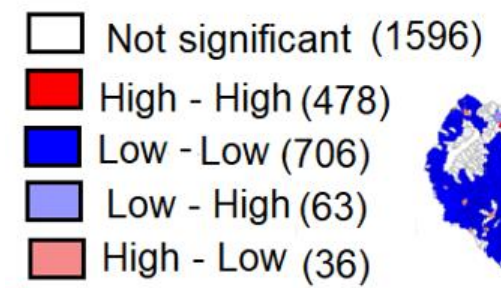
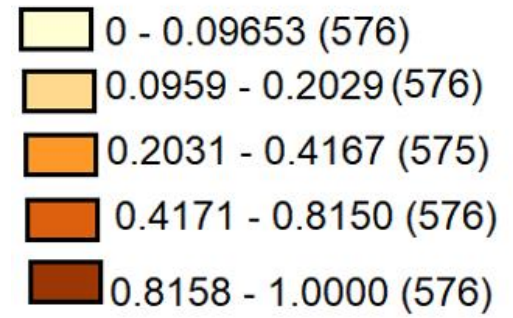


Large values at short lags,
indicate that the paired
samples are dissimilar and
more spatially discontinuous.



Priestorová autokorelácia

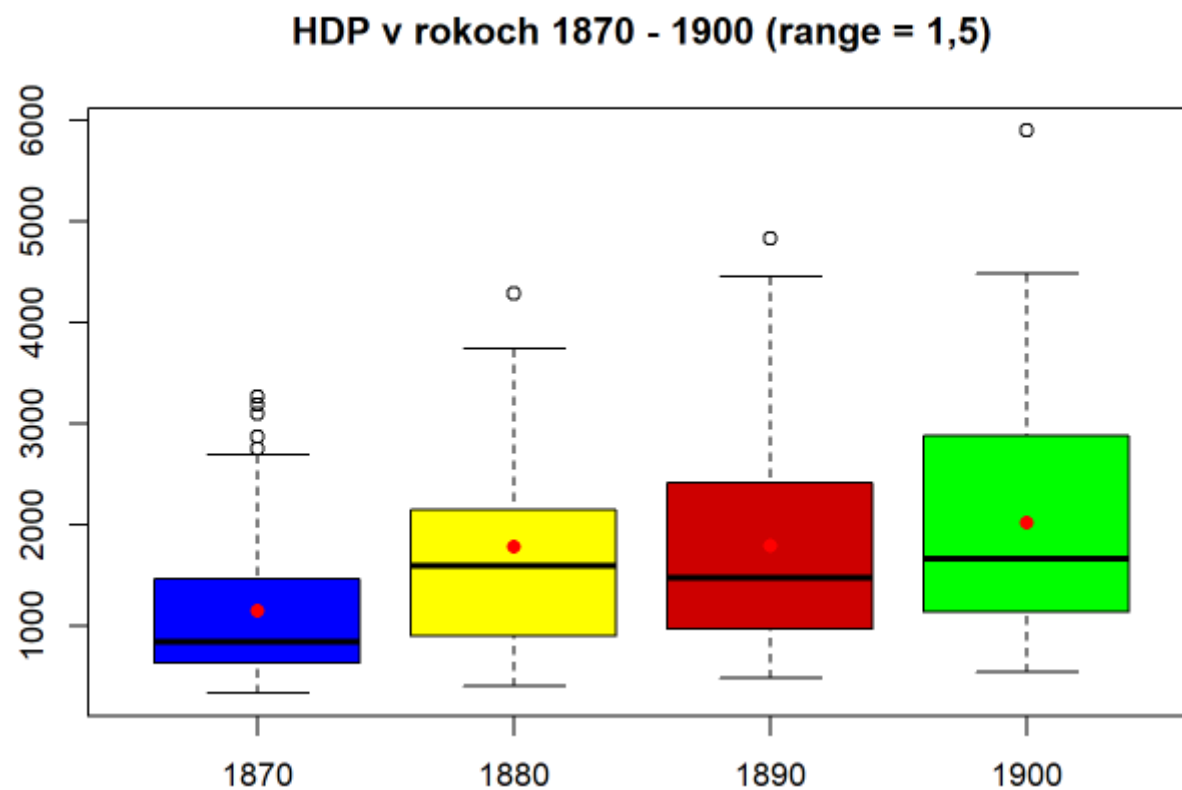
Indikátor	Moranov koeficient	P-hodnota
Podiel bytov vykurovaných plynom	0,5513	0,001
Podiel bytov vykurovaných pevným palivom	0,5571	0,001





Krabičkové grafy

Krabičkový graf je jedným zo spôsobov grafickej vizualizácie údajov.





Korelácia nie je kauzalita



Korelačné koeficienty

V štatistike používame štyri typy korelácií:

- Parametrické: Pearsonova korelácia
- Neparametrické:
 - Kendalova poradová korelácia,
 - Spearmanova korelácia
 - Bodovo biserialova korelácia.

V prospech Spearmanovho rho: V praxi sa zdá byť populárnejší.

V prospech Kendallovho tau: Spearmanov rho je citlivejší na chyby a nezrovnalosti v údajoch.

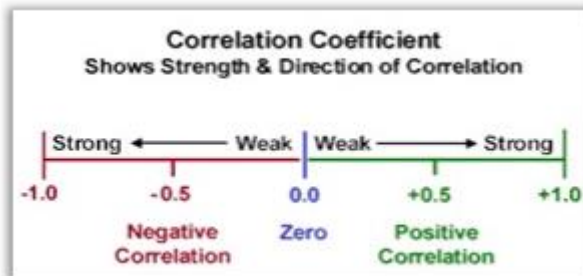
Keď sú údaje normálne, Kendalovo tau má menšiu citlivosť na hrubú chybu a menší asymptotický rozptyl.

Spearmanove hodnoty rho bývajú vyššie ako hodnoty tau Kendalla.



Pearsonov korelačný koeficient

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$



<i>r</i> value	Interpretation
+ .70 or higher	Very strong positive relationship
+ .40 to + .69	Strong positive relationship
+ .30 to + .39	Moderate positive relationship
+ .20 to + .29	weak positive relationship
+ .01 to + .19	negligible relationship
0	No relationship [zero order correlation]
- .01 to - .19	negligible relationship
- .20 to - .29	weak negative relationship
- .30 to - .39	Moderate negative relationship
- .40 to - .69	Strong negative relationship
- .70 or higher	Very strong negative relationship

<https://www.elegantjbi.com/blog/what-is-karl-pearson-correlation-analysis-and-how-can-it-be-used-for-enterprise-analysis-needs.htm>

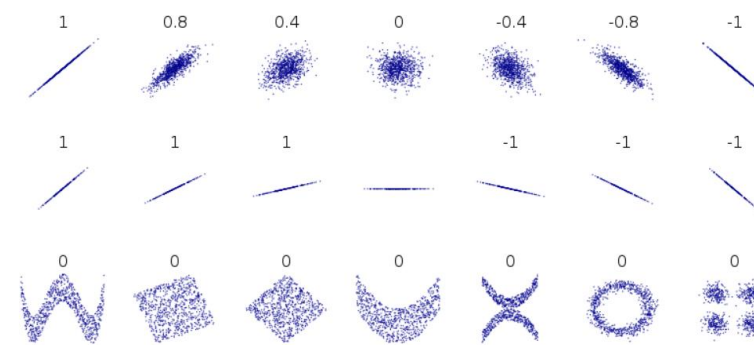
Size of Correlation	Interpretation
.90 to 1.00 (-.90 to -1.00)	Very high positive (negative) correlation
.70 to .90 (-.70 to -.90)	High positive (negative) correlation
.50 to .70 (-.50 to -.70)	Moderate positive (negative) correlation
.30 to .50 (-.30 to -.50)	Low positive (negative) correlation
.00 to .30 (.00 to -.30)	negligible correlation

<http://www.bioline.org.br/pdf?mm12018>

Signifikantnosť Pearsonovho korelačného koeficienta — test $H_0: r_0 = 0$ (p hodnota < hladina významnosti, potom H_0 zamietame)

Miera lineárnej závislosti

Hodnota veľkosti účinku
Pearsonovej r korelácie sa
**pohybuje medzi -1 (dokonalá
negatívna korelácia) až +1
(dokonalá pozitívna korelácia)**





Veľkosť efektu

Odporúča sa, aby **výskumníci uvádzali veľkosti efektov** (vid.

<https://academic.oup.com/innovateage/article/3/4/igz036/5560156>)

- Veľkosť efektu (effect size) nám **povie**,
 - ako významný je vzťah medzi premennými alebo
 - ako významný je rozdiel medzi skupinami

Softvér:

<https://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-R-main.php>



Veľkosť efektu pre Pearsonov korelačný koeficient

Cohen, J. (1988). Statistical power analysis for the behavioral sciences, 2nd ed. Hillsdale, NJ: Erlbaum

Podľa Cohena (1988, 1992) je veľkosť účinku nízka, ak sa hodnota r pohybuje okolo 0,1, stredná, ak sa r pohybuje okolo 0,3, a veľká, ak sa r mení o viac ako 0,5.

Strength of Association	Coefficient, r	
	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

ALE závisí od kontextu. Pre ekologickú terénnu štúdiu je r 0,1 dosť veľké a pre vzťah medzi dávkou a odozvou v experimente bunkovej kultúry je $r=0,9$ skôr nízke. „**Jedna veľkosť sa hodí pre všetko**“ nikdy nefunguje.“

Hodnota p nestačí

Nižšia p - hodnota sa niekedy interpretuje tak, že medzi dvoma premennými existuje silnejší vzťah. **Štatistická významnosť** však znamená, že je nepravdepodobné, že nulová hypotéza je pravdivá ($H_0: \rho = 0$) (menej ako 5 %).

Preto významná hodnota p nám hovorí, že zásah funguje, zatiaľ čo veľkosť účinku nám hovorí, do akej miery funguje.

Možno tvrdiť, že zdôraznenie veľkosti účinku podporuje vedeckejší prístup, keďže na rozdiel od testov významnosti je veľkosť účinku nezávislá od veľkosti vzorky.



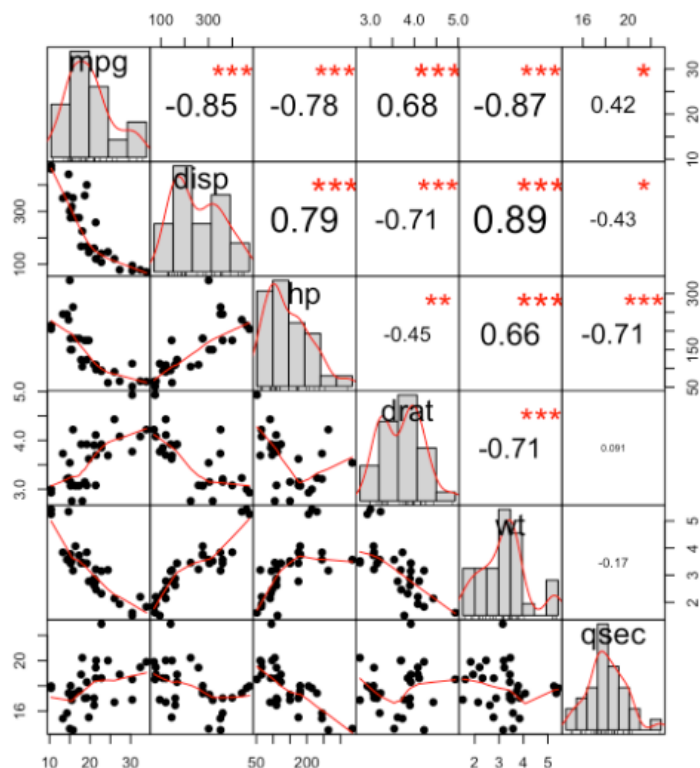
Pearsonov korelačný koeficient

Ak sú premenné nezávislé , Pearsonov korelačný koeficient je 0, ale opak neplatí, pretože korelačný koeficient detekuje iba lineárne závislosti medzi dvoma premennými.

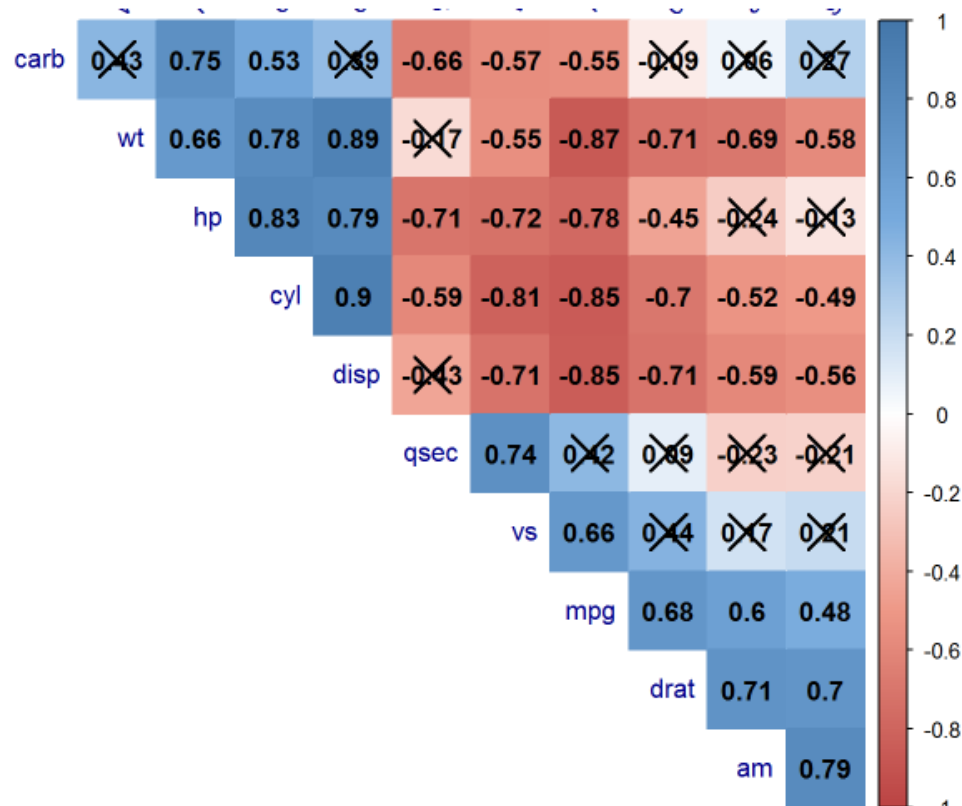
Avšak v špeciálnom prípade, keď (X,Y) majú dvojrozmerné normálne rozdelenie, nekorelácia je ekvivalentná nezávislosti.



Pearsonov korelačný koeficient



<http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software>



https://rstudio-pubs-static.s3.amazonaws.com/240657_5157ff98e8204c358b2118fa69162e18.html



Kontingenčné tabuľky

Kontingenčná tabuľka je tabuľka, ktorá sa používa na prehľadné zhrnutie vzájomného vzťahu dvoch kategoriálnych štatistických znakov.

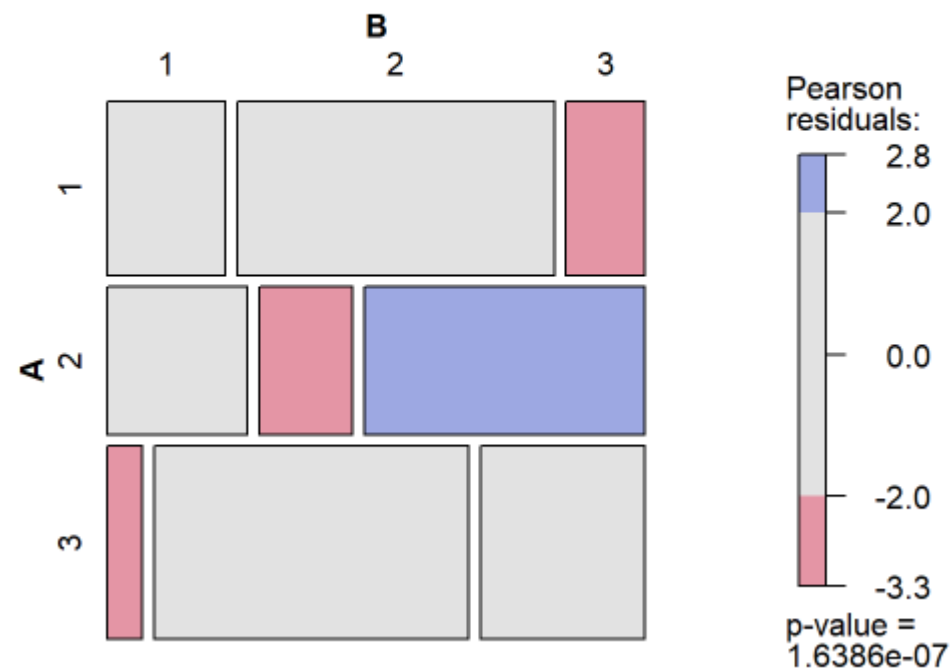


Kontingenčné tabuľky

Najvyššie dosiahnuté vzdelanie	Preferovaná politika		
	žiadne reštrikcie	fajčenie na vyhradených miestach	zákaz fajčenia
základná škola	15	40	10
stredná škola	15	100	30
vysoká škola	5	44	23

- Rozsah výberového súboru je 282. Hodnota testovacej štatistiky je 37,2 (počet stupňov voľnosti je 4). Výsledky výpočtov naznačujú signifikantný vzťah medzi vzdelaním a preferovaním druhu politiky fajčenia na verejných priestranstvách (p hodnota testovacej štatistiky je 0,0000001639). Veľkosť efektu je stredne vysoká 0,2955

Mozaikový graf





Viacrozmerne metody

Nezávislá premenná (premenné)		Závislá premenná (premenné)			
		Jedna		Viac ako jedna	
		Kvantitatívna	Kvalitatívna	Kvantitatívna	Kvalitatívna
Jedna	Kvantitatívna	Regresná analýza	Logistická regresia Diskriminačná analýza	Kanonická korelačná analýza	Viacskupinová diskriminačná analýza
	Kvalitatívna	t-test	Diskrétna diskriminačná analýza	MANOVA	Diskrétna viacskupinová diskriminačná analýza
Viac ako jedna	Kvantitatívna	Viacnásobná regresná analýza	Logistická regresia Diskriminačná analýza	Kanonická korelačná analýza	Viacskupinová diskriminačná analýza
	Kvalitatívna	ANOVA	Diskrétna diskriminačná analýza Conjoint analýza	MANOVA	Diskrétna viacskupinová diskriminačná analýza



Počet premenných	Typ údajov	
	Kvantitatívne	Kvalitatívne
Dve	Jednoduchá korelácia	Analýza dvojrozmerných kontingenčných tabuliek
		Loglineárne modely
Viac ako dve	Analýza hlavných komponentov	Analýza viacrozmerných kontingenčných tabuliek
		Loglineárne modely
	Faktorová analýza	Korešpondenčná analýza



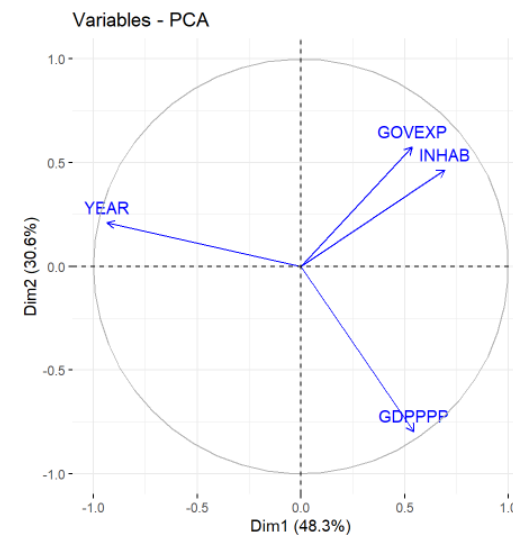
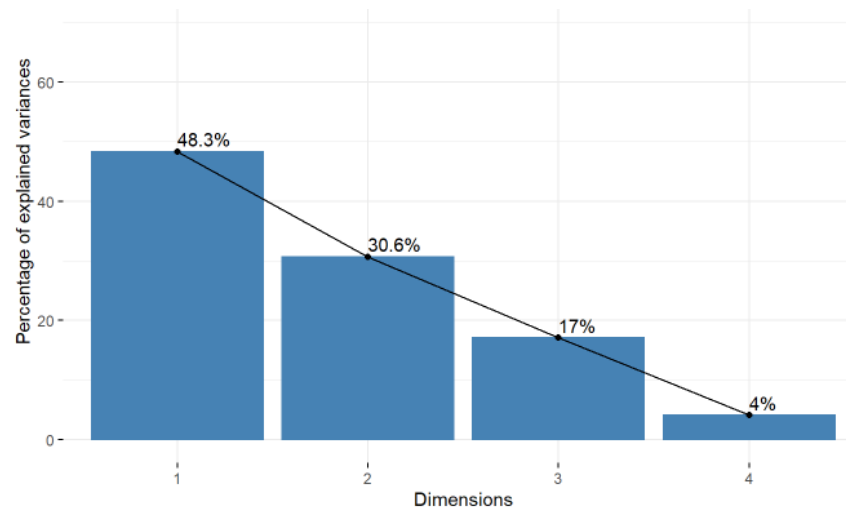
Analýza hlavných komponentov (PCA) a biplot

- Analýza **hlavných komponentov (PCA)** nám umožňuje sumarizovať a vizualizovať informácie v súbore údajov obsahujúcom pozorovania popísané viacerými vzájomne korelovanými kvantitatívnymi premennými. Keď každú premennú považujeme za dimenziu a ak náš súbor obsahuje viac ako tri premenné, vizualizácia je problematická.
- Analýzu hlavných komponentov používame na extrakciu dôležitých informácií z datovej matice s viacerými premennými a na vyjadrenie týchto informácií pomocou nových premenných, ktoré sa nazývajú hlavné komponenty. Počet hlavných komponentov je menší alebo rovnaký ako počet pôvodných premenných.



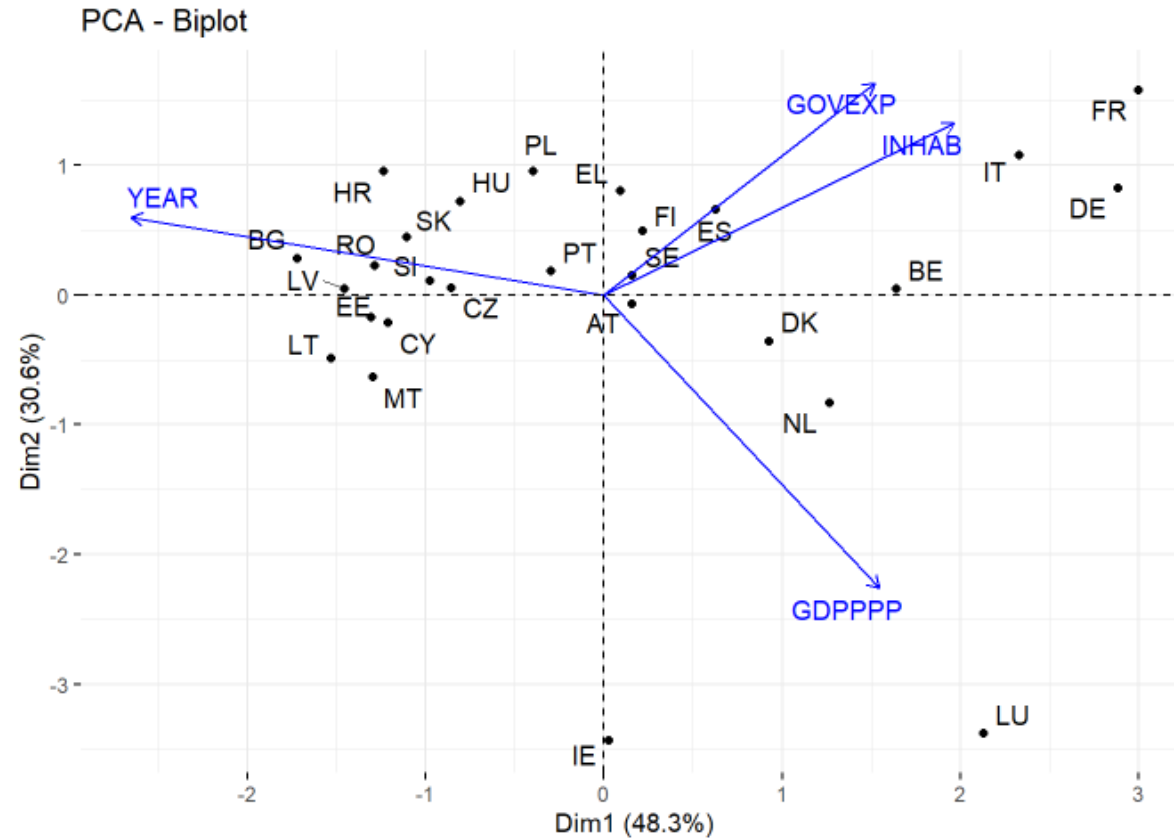
Analýza hlavných komponentov (PCA) a biplot

Štát	GDPPPP	INHAB	GOVEXP	YEAR
BE	36 700	11 455 519	52,1	1958
BG	16 500	7 000 039	36,3	2007
CZ	28 900	10 649 800	41,4	2004
DK	40 500	5 806 081	49,2	1973
DE	37 500	83 019 213	45,2	1958
EE	26 100	1 324 820	38,9	2004
IE	60 200	4 904 240	24,6	1973
EL	20 700	10 724 599	47,9	1981
ES	28 400	46 937 060	42,1	1986
FR	33 100	67 177 636	55,4	1958
HR	20 300	4 076 246	47,2	2013
IT	29 800	59 816 673	48,6	1958
CY	27 900	875 899	39,7	2004
LV	21 500	1 919 968	38,4	2004
LT	26 000	2 794 184	34,6	2004
LU	81 000	613 894	42,3	1958
HU	22 800	9 772 756	45,7	2004
MT	31 100	493 559	36,8	2004
NL	39 900	17 282 163	42,0	1958
AT	39 400	8 858 775	48,6	1995
PL	22 700	37 972 812	41,8	2004
PT	24 700	10 276 617	42,5	1986
RO	21 700	19 414 458	36,2	2007
SI	27 700	2 080 908	43,3	2004
SK	21 900	5 450 421	42,7	2004
FI	34 700	5 517 919	53,2	1995
SE	37 000	10 230 185	49,3	1995



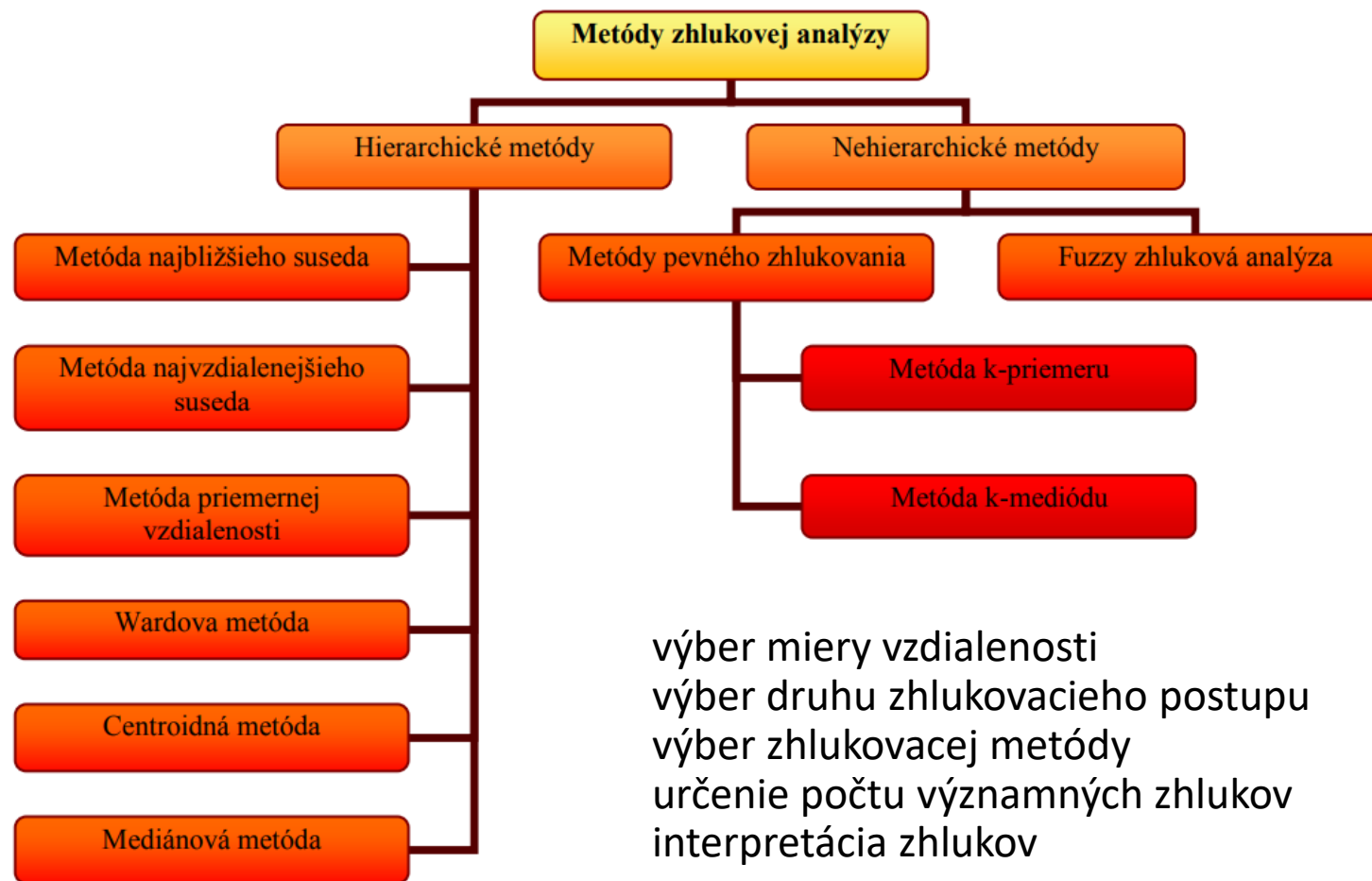


Analýza hlavných komponentov (PCA) a biplot





Klasifikácia metód zhlukovej analýzy

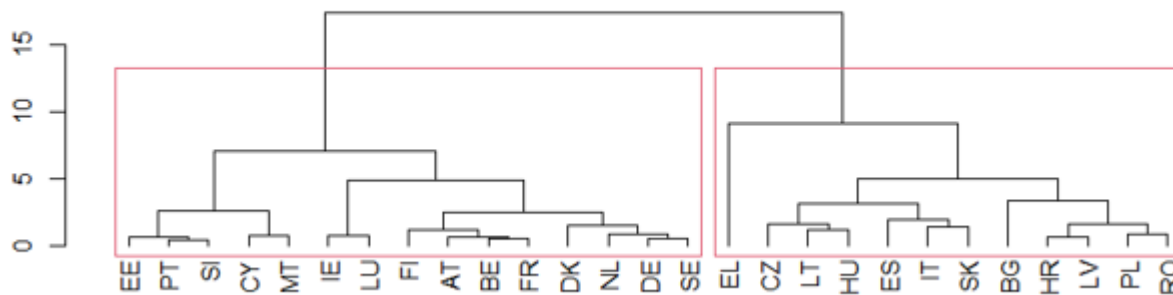




Dendrogram

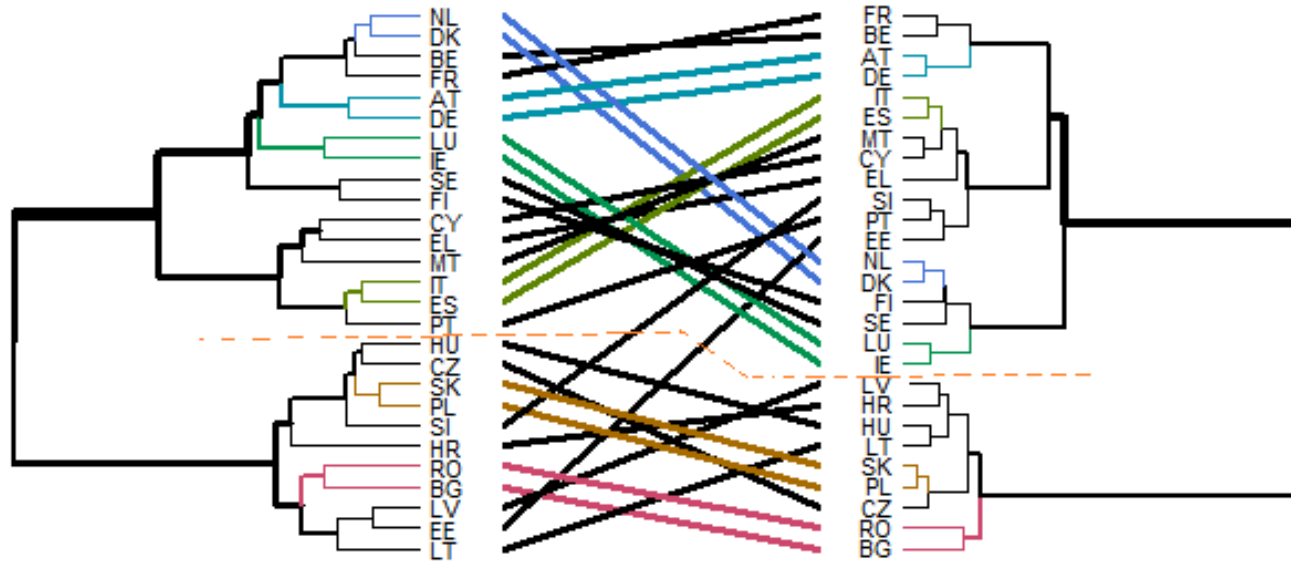
- Údaje majú tendenciu zhlukovať sa, lebo hodnota Hopkinsovej štatistiky je vysoká (0,7298). Na zhlukovanie sme použili Wardovu metódu, lebo hodnota aglomeratívneho koeficientu bola pre túto metódu najvyššia (average: 0,8073, single: 0,7898, complete: 0,8293, ward: 0,8518). Optimálny počet zhlukov je dva.

Obr. 7: Dendrogram

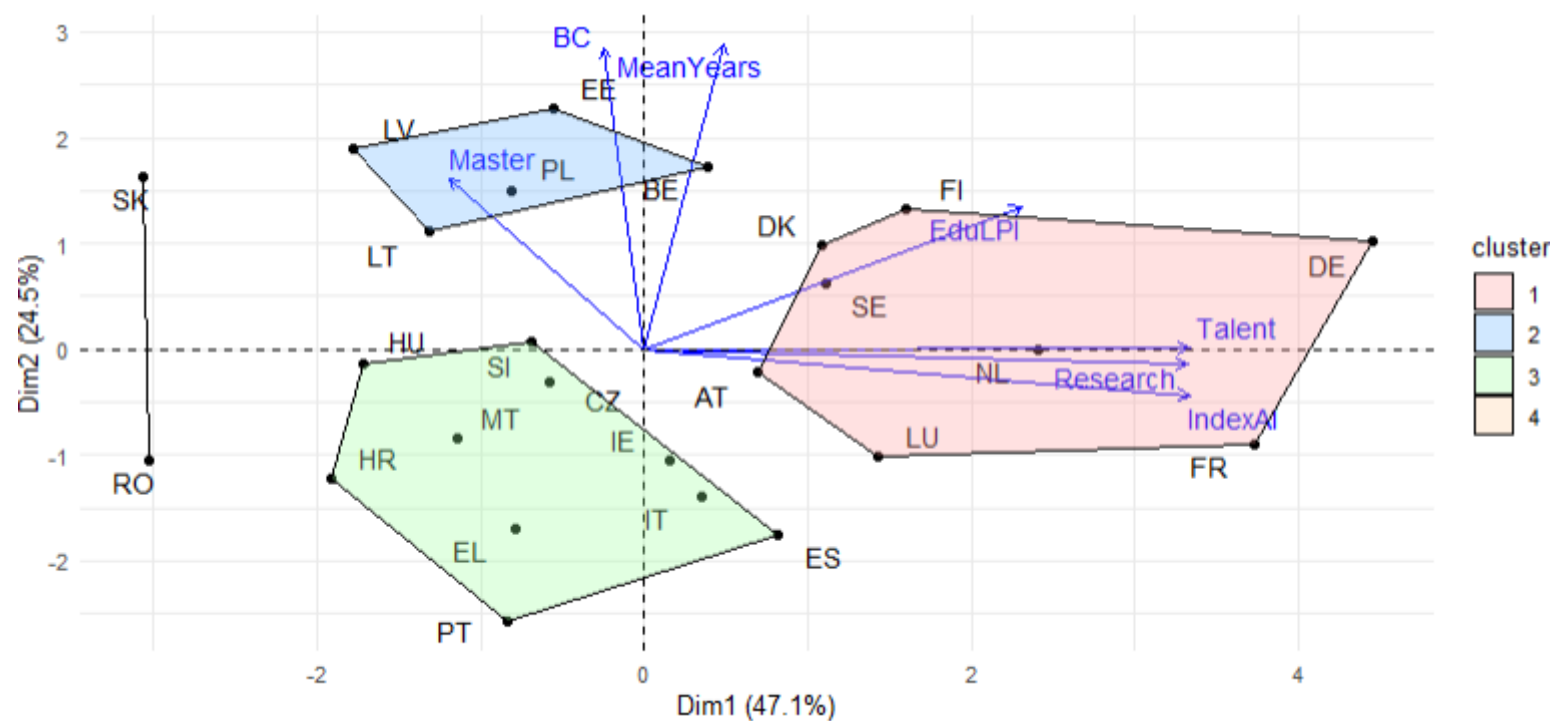




Tanglegram

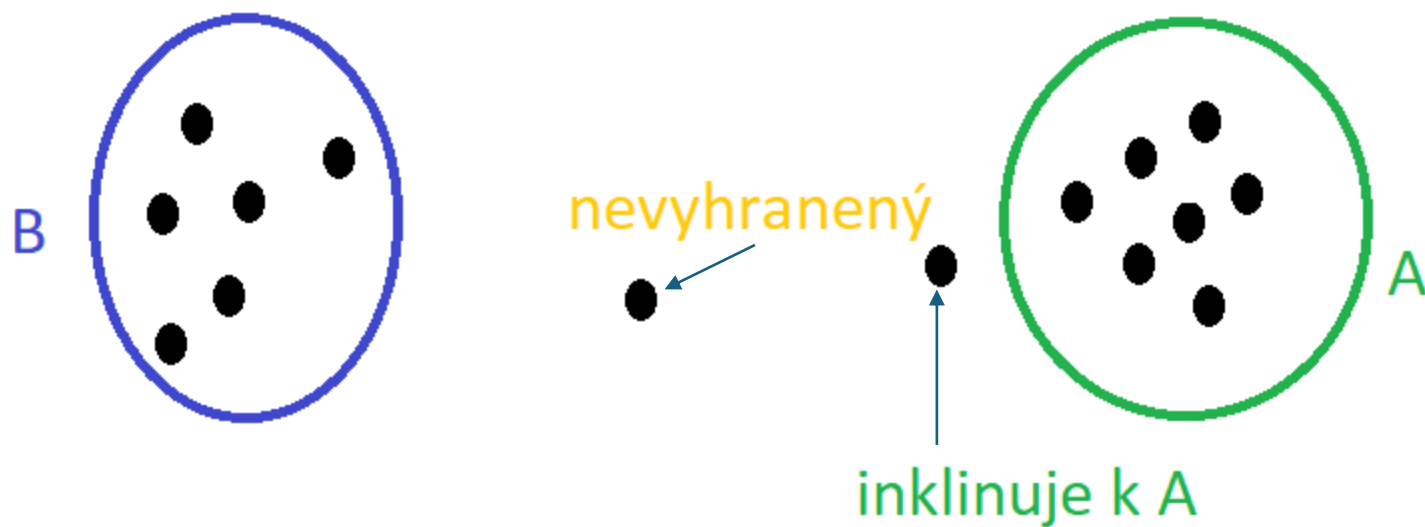


Zhluková analýza z PCA





Fuzzy zhukovanie





Mapové zobrazenie výsledkov fuzzy c-zhlukovej analýzy



zhluk 1 (681)

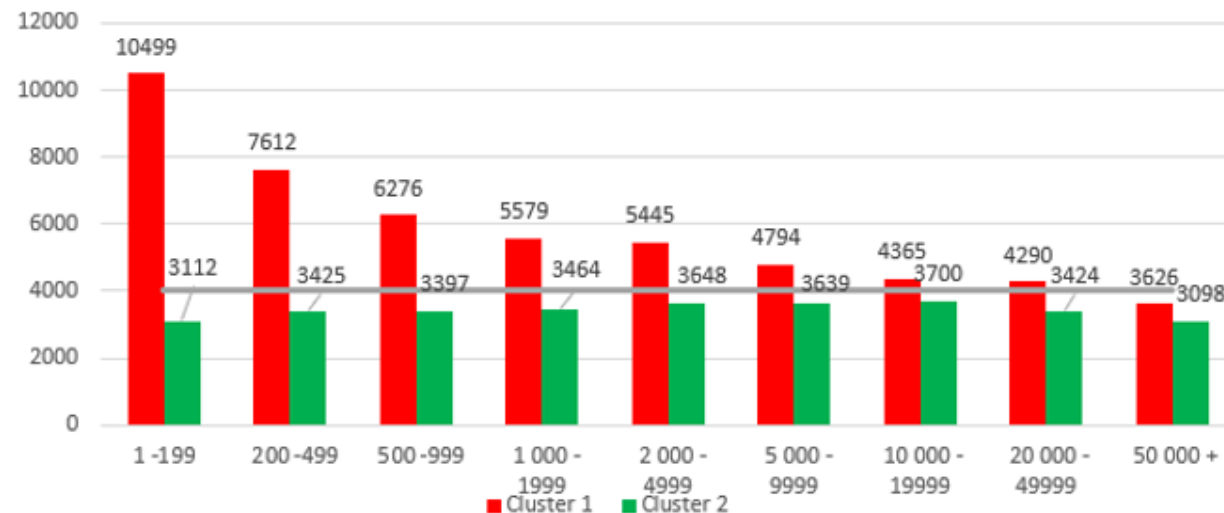
zhluk 2 (1586)

inklinuje k zhluku 1 (326)

inklinuje k zhluku 2 (286)

Moranov koeficient 0,1277 (P hodnota je 0,0001)

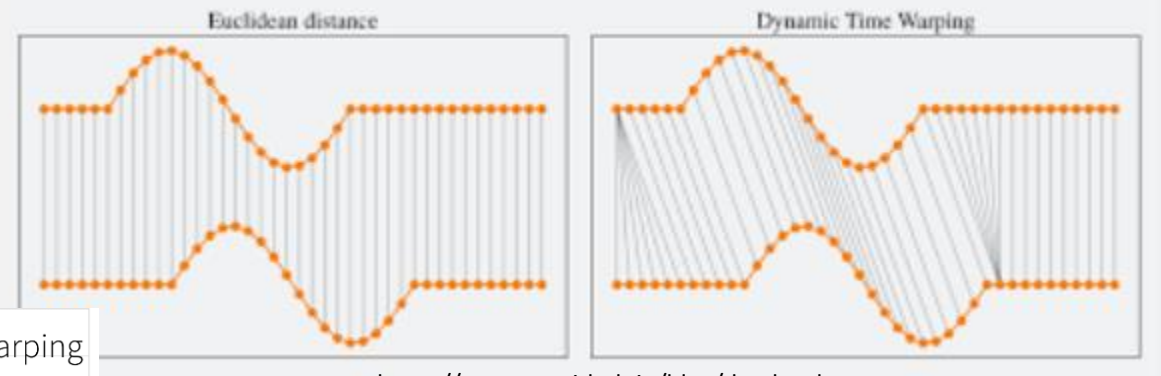
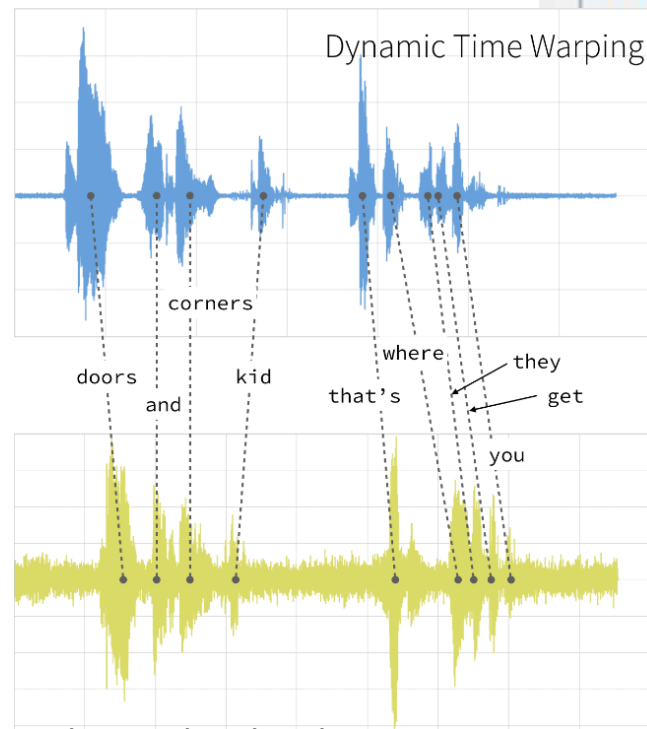
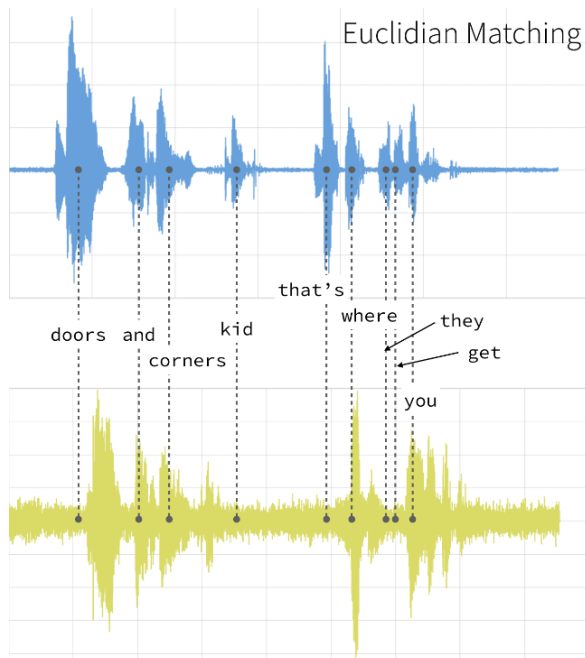
Hodnoty PYLL typických predstaviteľov zhlukov podľa veľkostných kategórií



PYLL je ukazovateľ predčasného úmrtia (pred 75. rokom života), ktorý dáva väčšiu váhu úmrtiam v mladšom veku ako úmrtiam v neskoršom veku. PYLL zdôrazňuje stratu potenciálneho prínosu, ktorý môžu mladší jedinci priniesť spoločnosti.



Dynamic Time warping



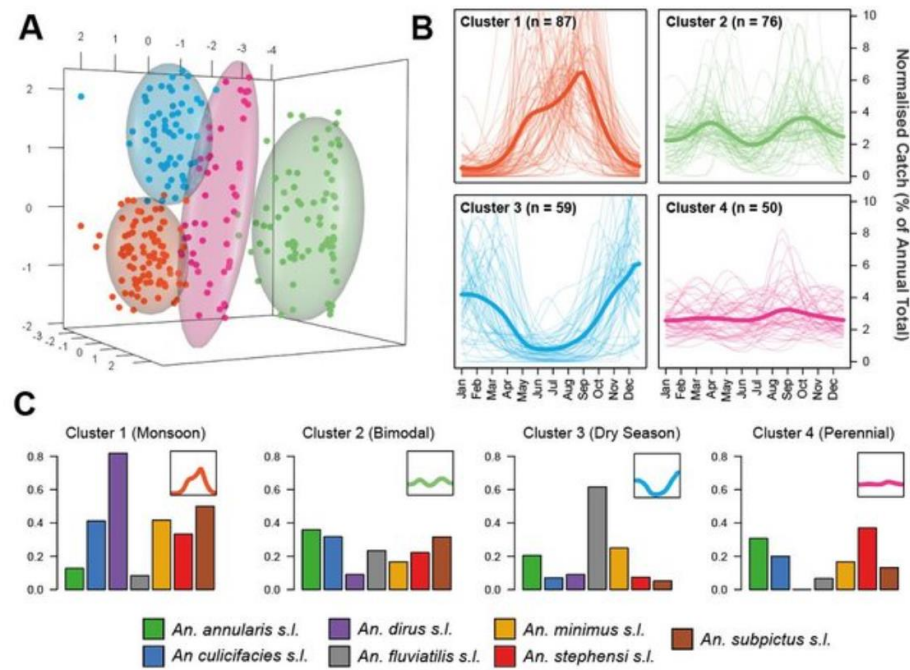
<https://rtavenar.github.io/blog/dtw.html>

<https://www.databricks.com/blog/2019/04/30/understanding-dynamic-time-warping.html>

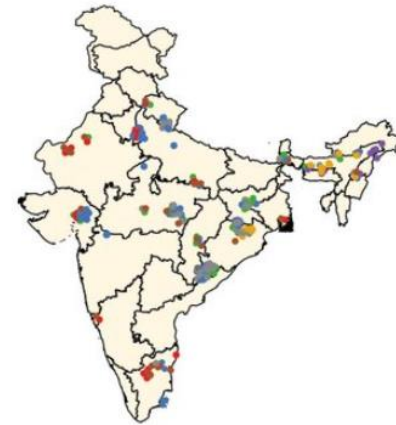


Dynamic Time warping

- Charakterizácia a zhlukovanie časových radov s podobnými časovými vlastnosťami



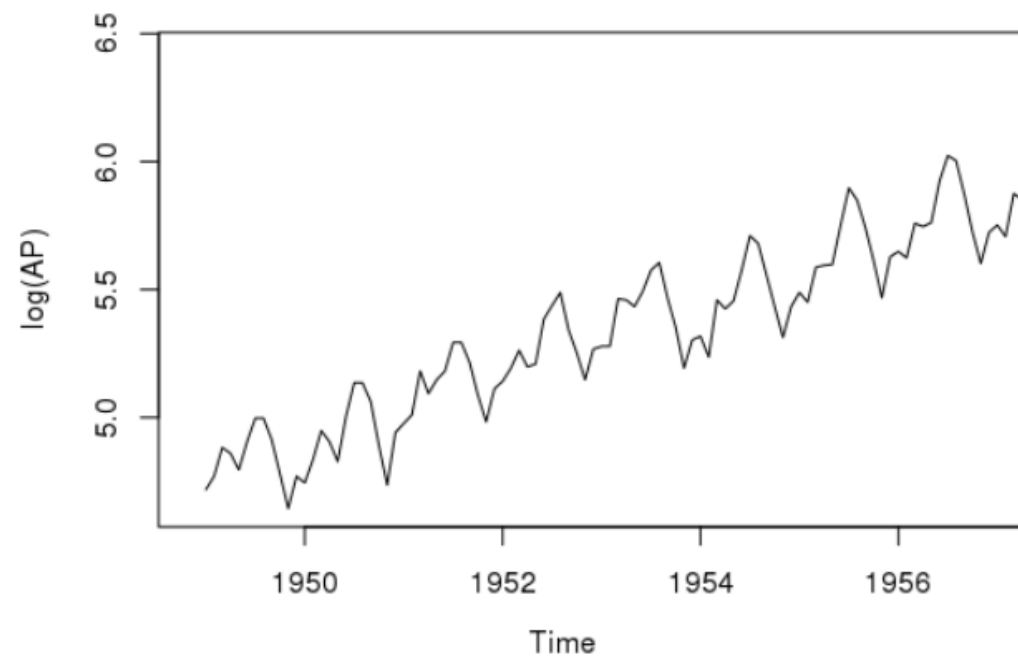
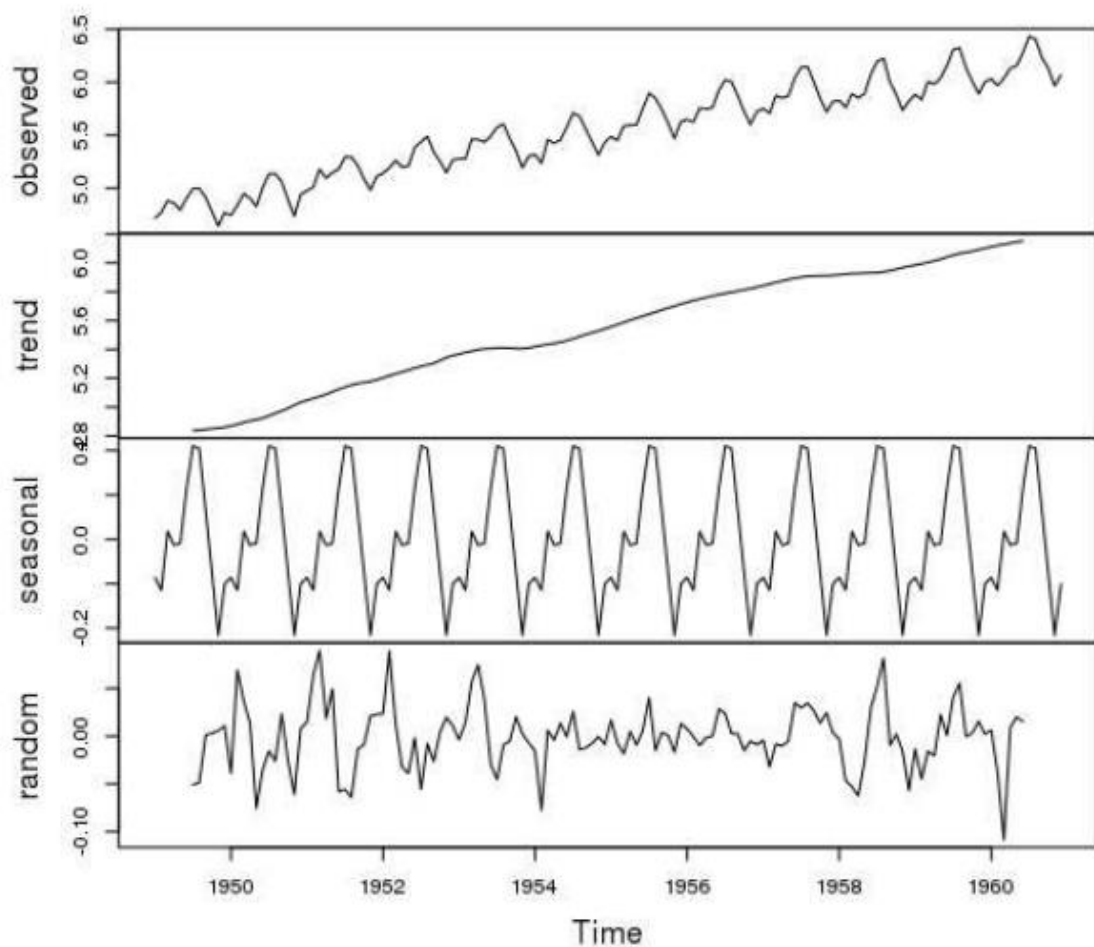
Skúmanie druhovo špecifických vzorcov populačnej dynamiky komárov .



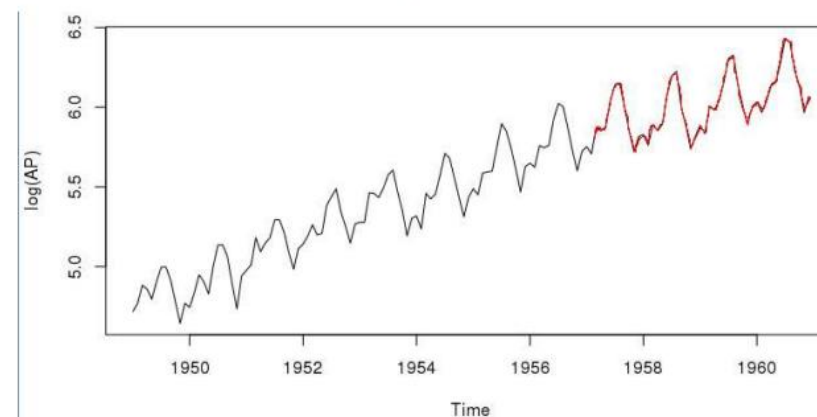
https://www.researchgate.net/publication/348494644_The_Ecological_Structure_of_Mosquito_Population_Dynamics_Insights_from_India/figures?lo=1



Časové rady



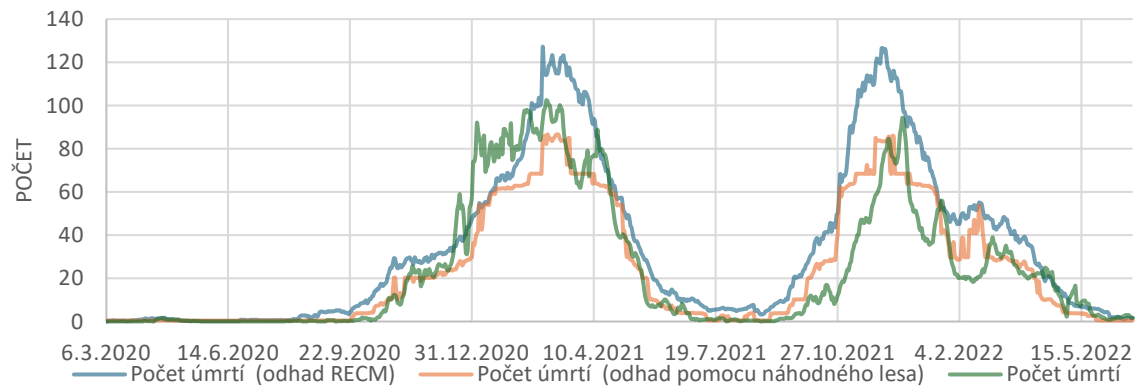
Intuitívne: ak nenastane nejaký šok, zachová sa rastúci trend a sezónnosť, teda nejako takto:





Časové rady

- **Kointegrácia** - V kointegrácii je zahrnutá informácia o dlhodobých vzťahoch pôvodných nestacionárnych radov, teda dlhodobá rovnováha medzi premennými.
- **ARDL modely**

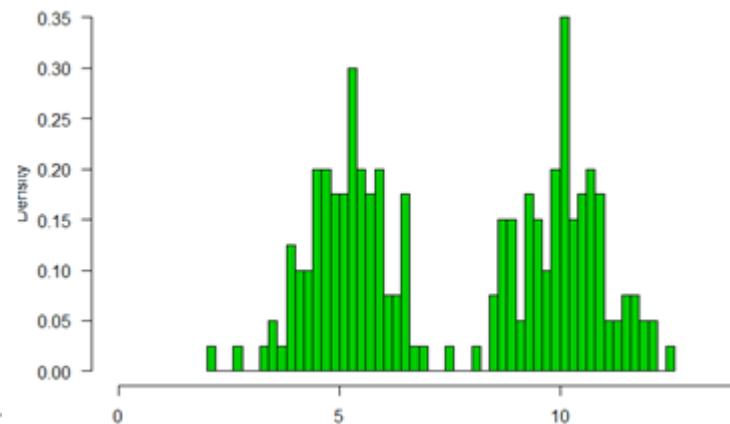
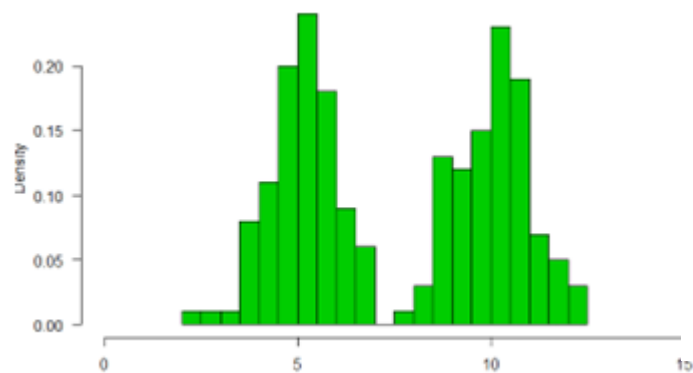
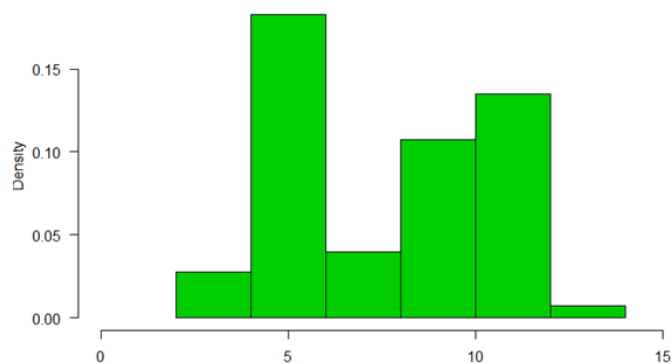


Obsah

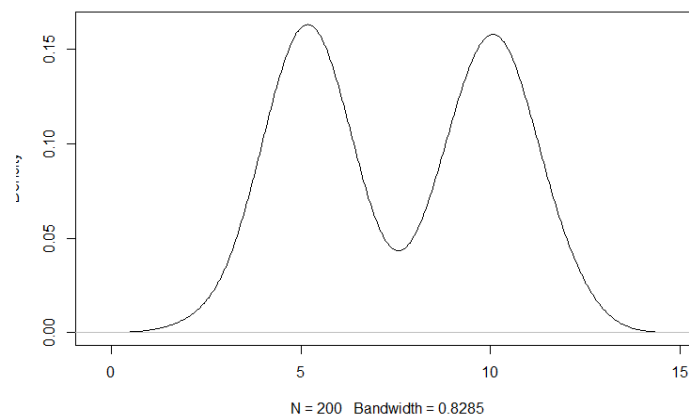
- Popis premenných
- Závislosti
- Podobné objekty



-histogram, jadrová funkcia hustoty

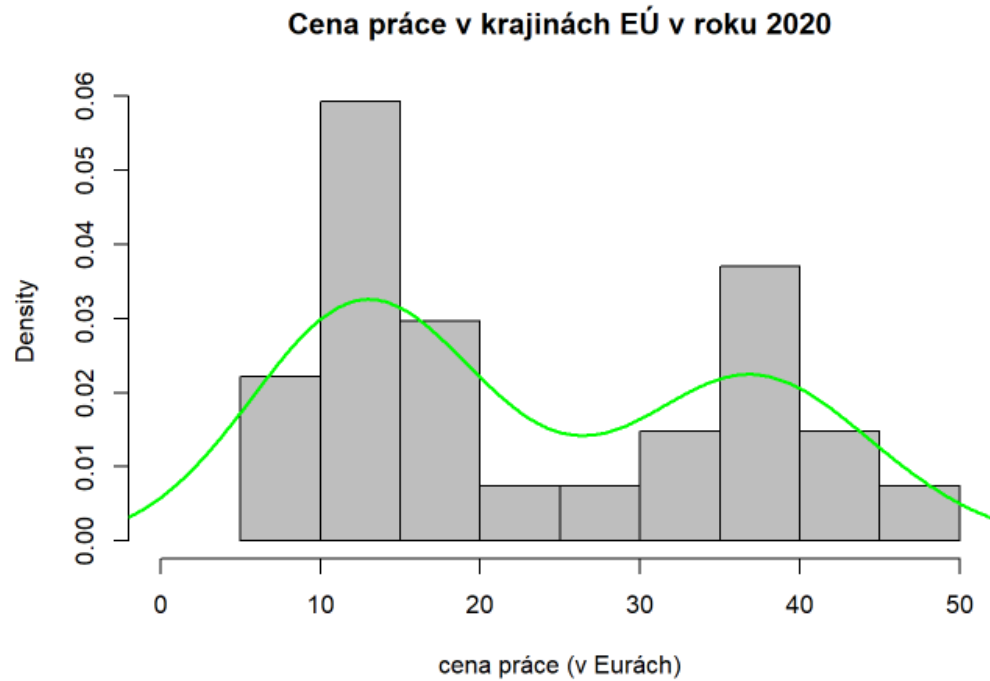


`density.default(x = normalne$c, na.rm = TRUE)`





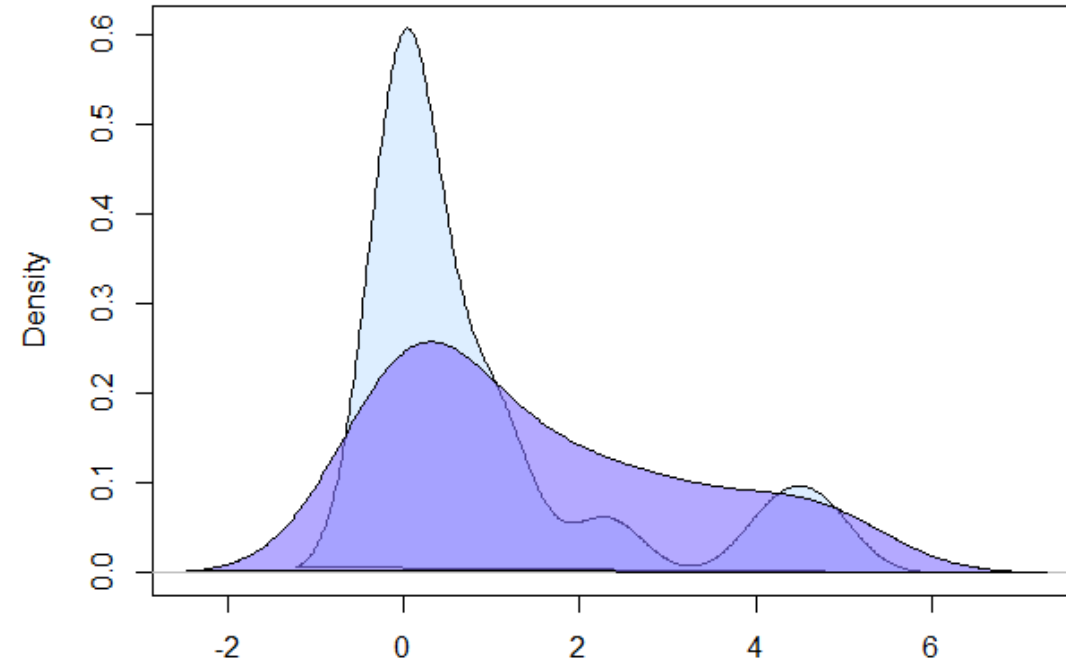
Jadrová funkcia hustoty



Ameijeirasov-Alonsov test

<https://rpubs.com/BeaSte/856160>

Global Terrorism Index (GTI)



Ujvary, K. (2022) POUŽÍVANIE KOMUNIKAČNÝCH PROSTRIEDKOV A INFORMAČNÝCH SYSTÉMOV NA IDENTIFIKÁCIU A ELIMINÁCIU TERORIZMU V SPOJITOSTI S NELEGÁLNOU MIGRÁCIU, dizertačná práca



Pravdepodobnostné rozdelenia

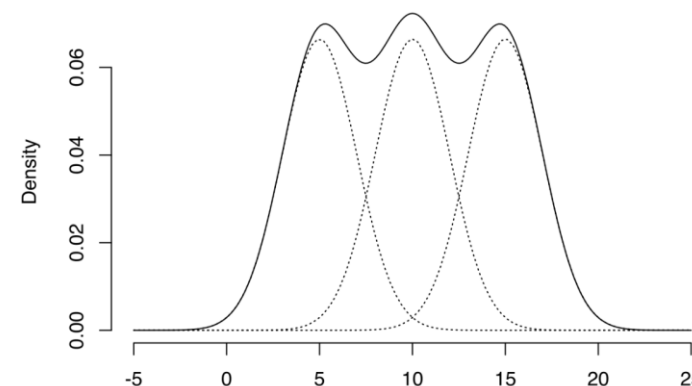
Pred aplikáciou ktorejkoľvek z techník, je dôležité vizuálne

preskúmať pravdepodobnostné rozdelenie

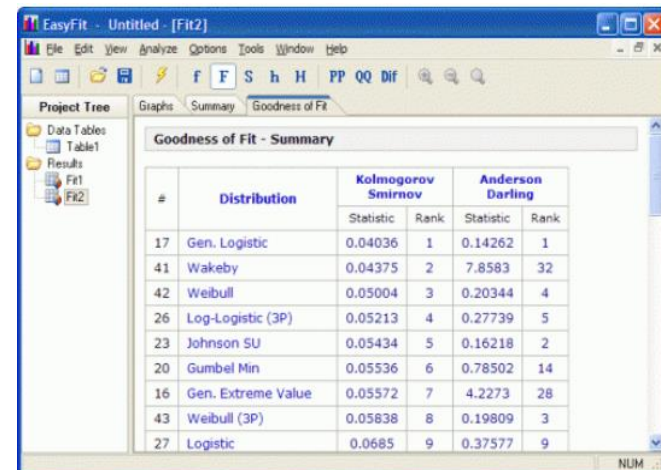
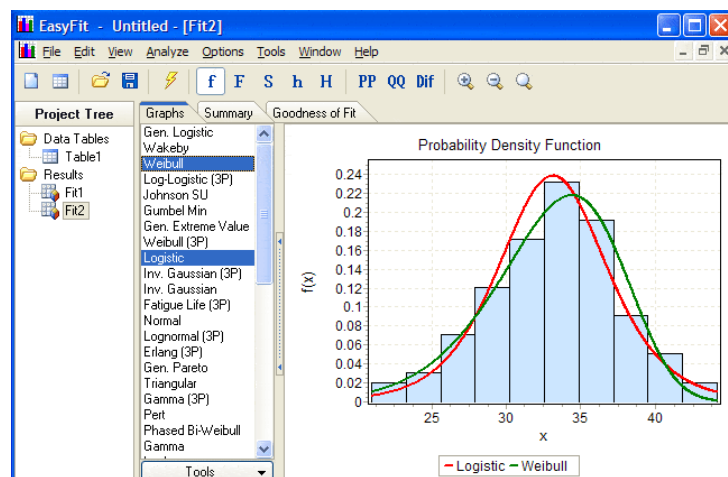
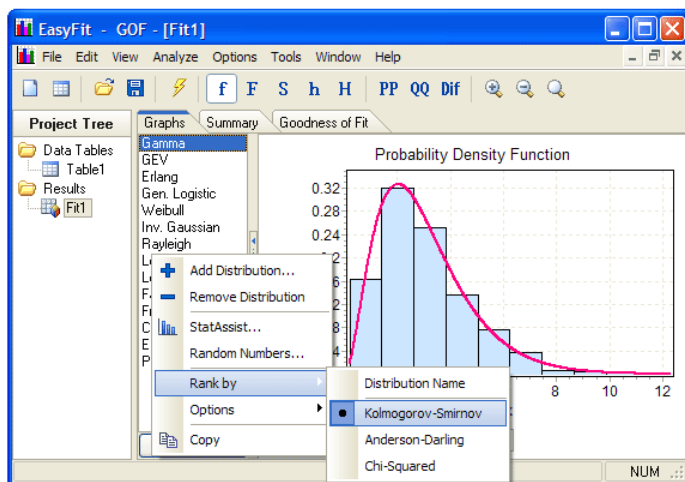
Distribučná funkcia *plne* charakterizuje pravdepodobnostné správanie náhodnej premennej. Pravdepodobnostné rozdelenia:

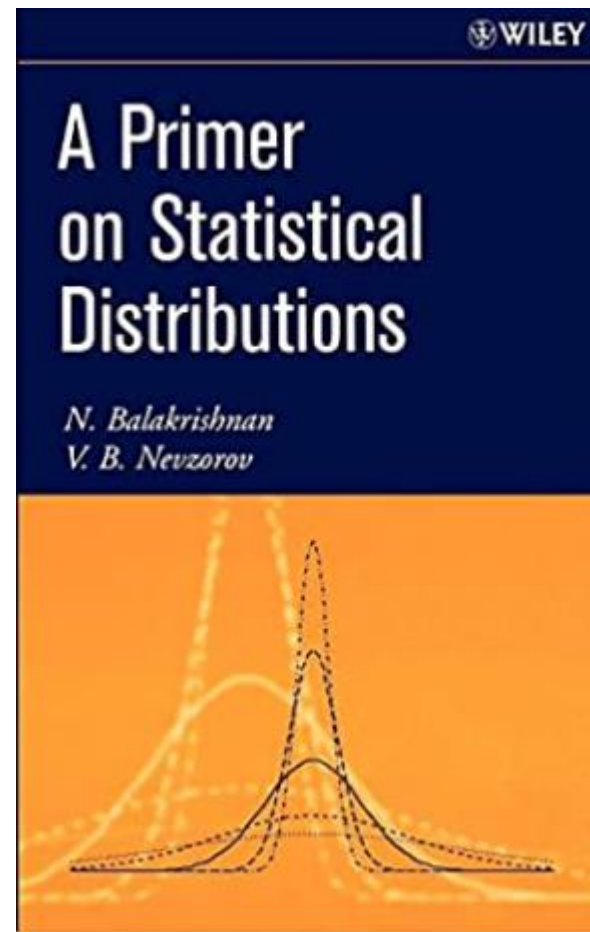
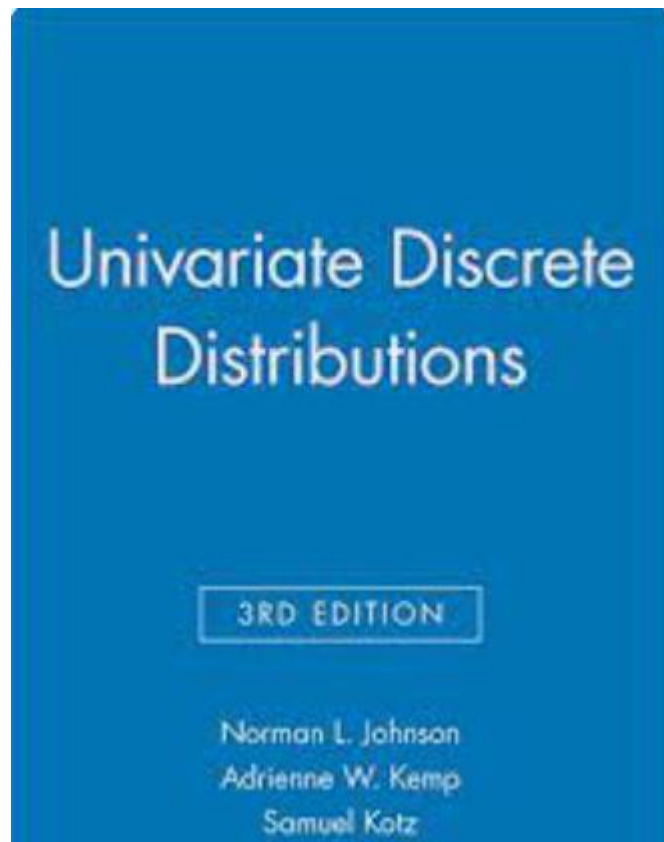
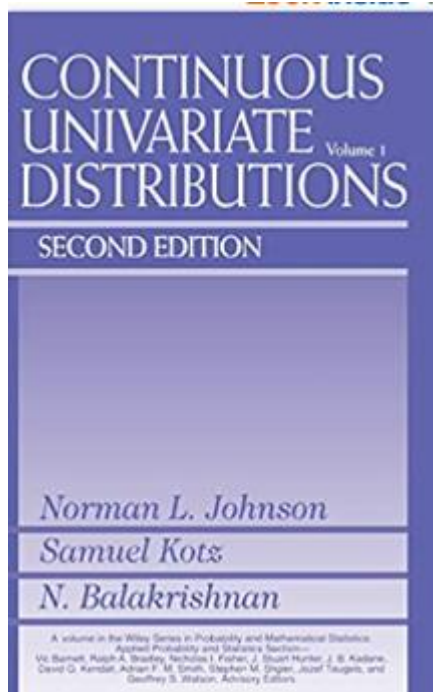
- diskkrétne (rovnomerné , Bernoulliho, binomické, Poissonovo, geometrické, hypergeometrické, negatívne binomické rozdelenie).
- spojité (rovnomerné rozdelenie na intervale, normálne, exponenciálne, Studentovo t-rozdelenie, chí-kvadrát rozdelenie),
- zmiešané náhodné premenné,
- zmesy rozdelení

Zmesy - ukážka



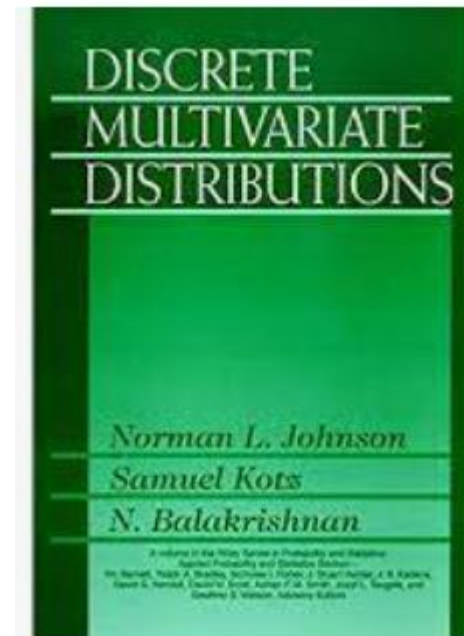
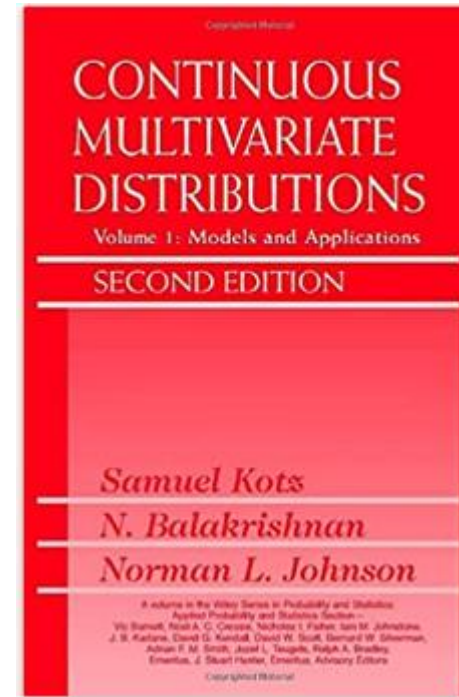
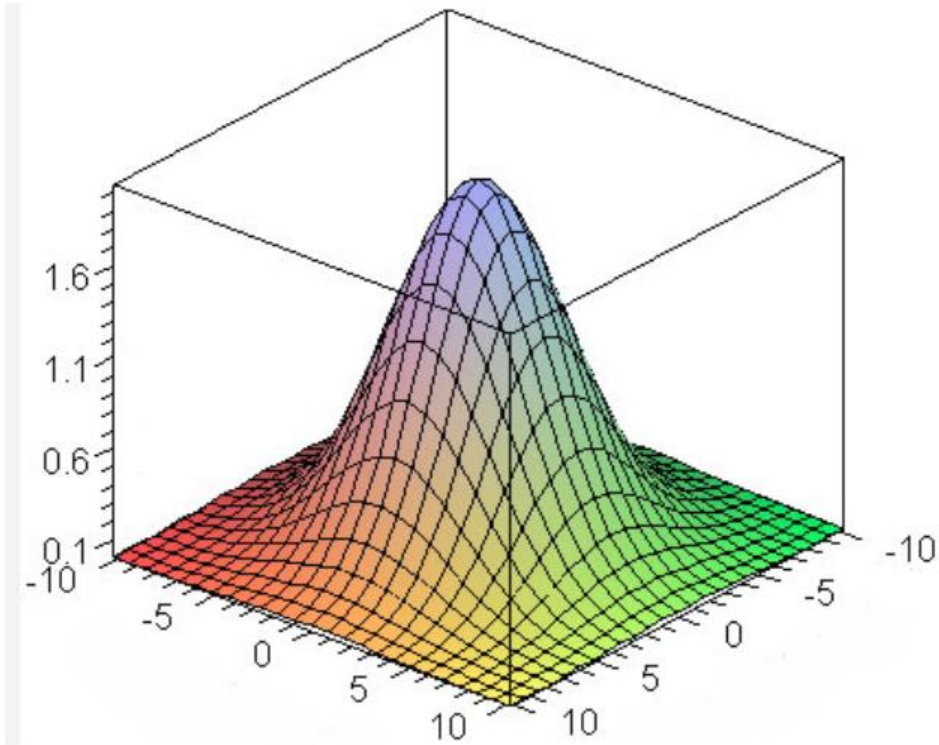
Testy dobrej zhody (EasyFit – 40 rozdelení)







Viacrozmerné rozdelenia



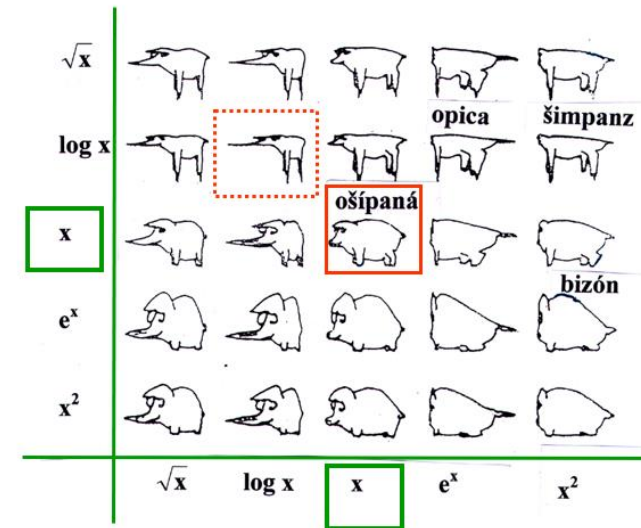
<https://ardianumam.wordpress.com/2017/10/19/deriving-gaussian-distribution/>



Transformácia údajov

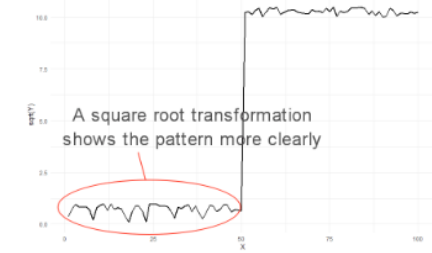
(spracovane podľa <https://www.spss-tutorials.com/normalizing-variable-transformations/>)

- Transformácia je matematická operácia, ktorá mení mierku merania premennej. Zvyčajne sa to robí, aby sa súbor stal použiteľným s konkrétnym štatistickým testom alebo metódou.
- Mnohé štatistické metódy vyžadujú údaje, určitý druh rozdelenia, zvyčajne normálne

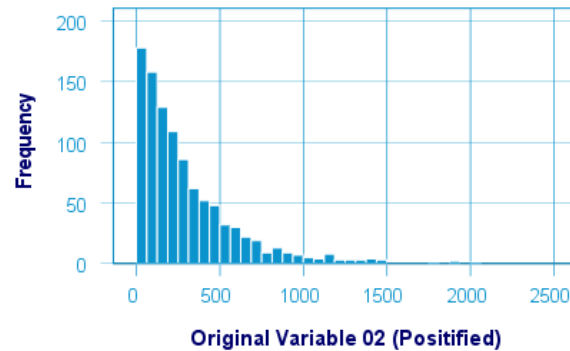




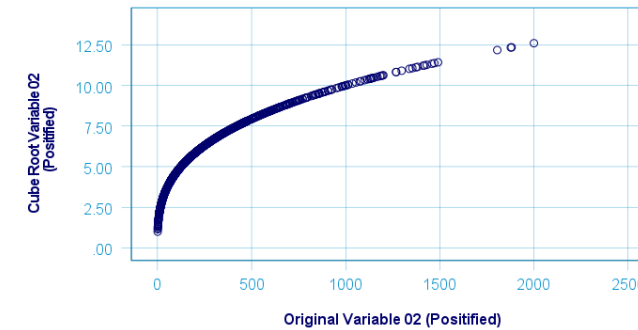
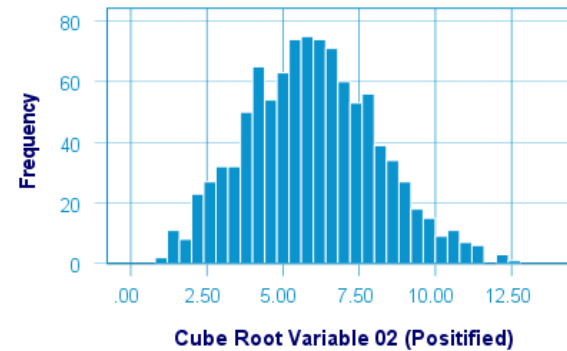
Odmocninová transformácia



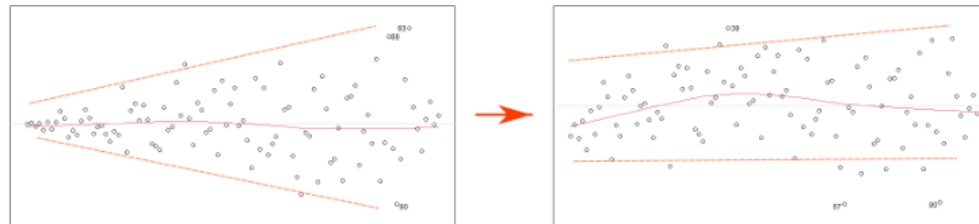
Original Variable 02 (Positifed)



Cube Root Variable 02 (Positifed)



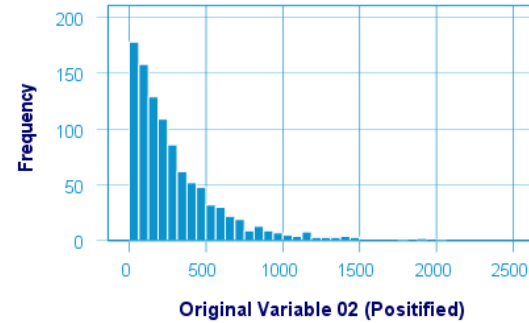
- **normalizácia zošikmeného rozdelenia** (premenné s ľavým zošikmením sa stanú najhoršími po transformácii druhej odmocniny)
- **zníženie heteroskedasticity**
- **linearizáciu vzťahu**
- **lepšiu vizualizáciu**



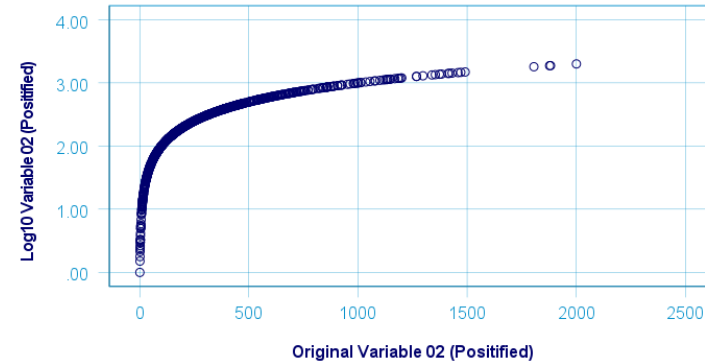
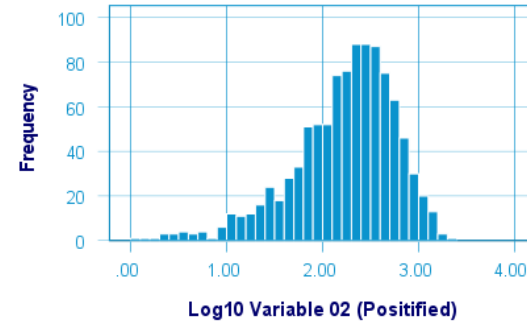
- **Poznámka:** pre záporné hodnoty: `if (pos01 < 0) curt01 = -abs(pos01)**(1/3)`

Logaritmická transformácia

Original Variable 02 (Positifed)



Log10 Variable 02 (Positifed)





$$Y = 5 + 0.03X$$

Priesečník v tejto rovnici je 5. To znamená, že keď je nezávislá premenná (x) 0, závislá premenná (Y) je 5. Koeficient x je 0,03, čo znamená, že závislá premenná (Y) by sa zvýšila o 0,03 pre každé zvýšenie o 1 jednotku nezávislej premennej (x).

Log — Level Regression

Logaritmicкая regresia je model, kde cieľová premenná je logaritmickey transformovaná, ale prediktorové premenné nie. Aby sme mohli interpretovať koeficienty logaritmickej regresie, musíme najprv umocniť koeficienty nezávislých premenných so základom e. Pozrime sa na nasledujúci výpočet, aby sme videli, ako zmena nezávislej premennej zmení závislú premennú.

$$\ln(Y_2) - \ln(Y_1) = (5 + 0.03x_2) - (5 + 0.03x_1)$$

$$\ln(Y_2) - \ln(Y_1) = 5 + 0.03x_2 - 5 - 0.03x_1$$

$$\ln(Y_2/Y_1) = 0.03(x_2 - x_1)$$

$$Y_2/Y_1 = e^{0.03(X_2 - X_1)}$$

$$Y_2/Y_1 = 1.03^{(X_2 - X_1)}$$

Takže na základe tejto rovnice môžeme dospieť k záveru, že pri každom zvýšení našej nezávislej premennej (x) o jednu jednotku sa závislá premenná (Y) zvýši o 3 % ($1,03^1 = 1,03$). Dôležité je, že toto zvýšenie musí byť **fixnou hodnotou**. Ak zväčšíme x o 10 jednotiek, zmena Y by sa rovnala 34 % ($1,03^{10} = 1,34$).

Level -Log Regression

Úroveň-log regresia je model, kde sa jedna alebo viac nezávislých premenných log-transformuje, ale závislá premenná zostáva pôvodná. Pred analýzou toho, čo znamená jeho koeficient, poďme sa ponoriť do výpočtu nižšie —

$$Y_2 - Y_1 = (5 + 0.03 \ln(x_2)) - (5 + 0.03 \ln(x_1))$$

$$Y_2 - Y_1 = 5 + 0.03 \ln(x_2) - 5 - 0.03 \ln(x_1)$$

$$Y_2 - Y_1 = 0.03(\ln(x_2) - \ln(x_1))$$

$$Y_2 - Y_1 = 0.03(\ln(x_2/x_1))$$

Podľa tejto rovnice môžeme predpokladať, že pri **fixnej percentuálnej** zmene hodnoty našej nezávislej premennej (x) sa naša závislá premenná (Y) zmení o $0,03 \cdot \ln(\text{percentuálnej zmeny})$. Napríklad, ak sa x zvýši o 10 %, Y sa zvýši o 0,3 % ($0,03 \cdot \ln(1,1) = 0,003$). Ak sa x zvýši o 50 %, Y sa zvýši o 1,2 % ($0,03 \cdot \ln(1,5) = 0,012$). Pokiaľ je percentuálna zmena v x pevná, Y sa tiež zmení o pevné percento.

<https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9>

Ako interpretovať výsledky regresie transformovaných údajov

Log-Log

Log-log regresia je model, kde cieľová premenná a aspoň jedna prediktorová premenná sú log-transformované. Podobne ako pri regresii na úrovni logaritmu odstránime logaritmus z ľavej strany rovnice a pravú stranu umocníme základom e. Pozrime sa -

$$\ln(Y_2) - \ln(Y_1) = (5 + 0.03 \ln(x_2)) - (5 + 0.03 \ln(x_1))$$

$$\ln(Y_2) - \ln(Y_1) = 5 + 0.03 \ln(x_2) - 5 - 0.03 \ln(x_1)$$

$$\ln(Y_2/Y_1) = 0.03(\ln(x_2) - \ln(x_1))$$

$$\ln(Y_2/Y_1) = 0.03(\ln(x_2/x_1))$$

$$Y_2/Y_1 = e^{0.03(\ln(x_2/x_1))}$$

$$Y_2/Y_1 = e^{(\ln(x_2/x_1)) \cdot 0.03}$$

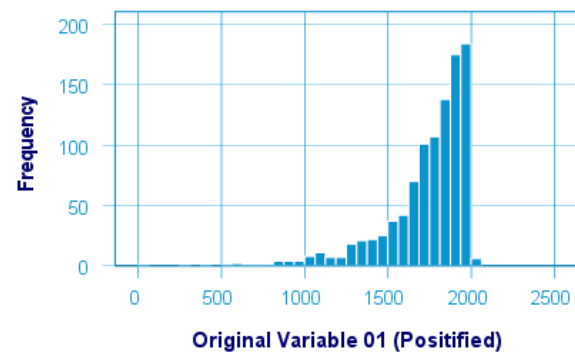
$$Y_2/Y_1 = (x_2/x_1)^{0.03}$$

Táto rovnica nám hovorí, že pre **pevné percento** zmenené v našej nezávislej premennej (x) by sa naša závislá premenná (Y) zmenila o x percentuálnu zmenu na mocninu 0,03. Ak sa x zmení o 10 %, y sa zmení o 0,3 % ($1,1^{0.03} = 1,003$). Ak sa x zmení o 20 %, y sa zmení o 0,5 % ($1,2^{0.03} = 1,005$).

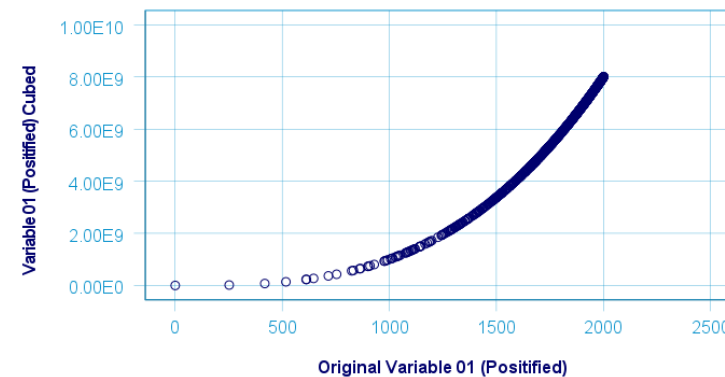
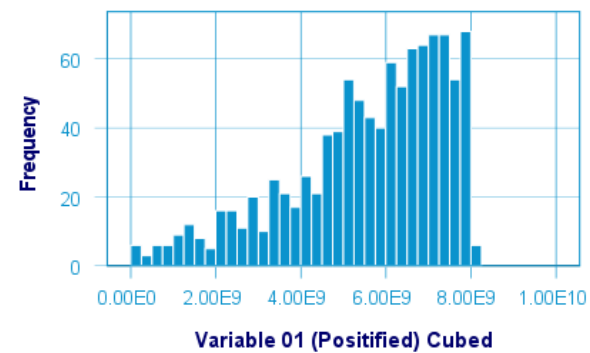


Mocninová transformácia

Original Variable 01 (Positifed)



Variable 01 (Positifed) Cubed





Box-Coxova transformácia

$$B(x, \lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \ln(x), & \text{if } \lambda = 0. \end{cases}$$

https://www.researchgate.net/publication/268412346_The_Box-Cox_Transformation_Technique_A_Review

Citované 1137-krát

<http://eprints.lse.ac.uk/103537/1/StatSciV4.pdf>

<https://www.le.ac.uk/users/dsgp1/COURSES/TOPICS/BoxCox.pdf>

Ďalšie informácie

<https://www.datanovia.com/en/lessons/transform-data-to-normal-distribution-in-r/>

Výber najvhodnejšej transformácie:

<https://cran.r-project.org/web/packages/bestNormalize/vignettes/bestNormalize.html>



Nie všetky dáta môžu byť transformované

- Symetrické dáta s „ťažšími chvostami" ako má normálne rozdelenie (väčšie množstvo dát na oboch koncoch rozdelenia v porovnaní s normálnym rozdelením).
- Bimodálne (multimodálne) rozdelenia
- Keď je prítomné väčšie množstvo identických pozorovaní

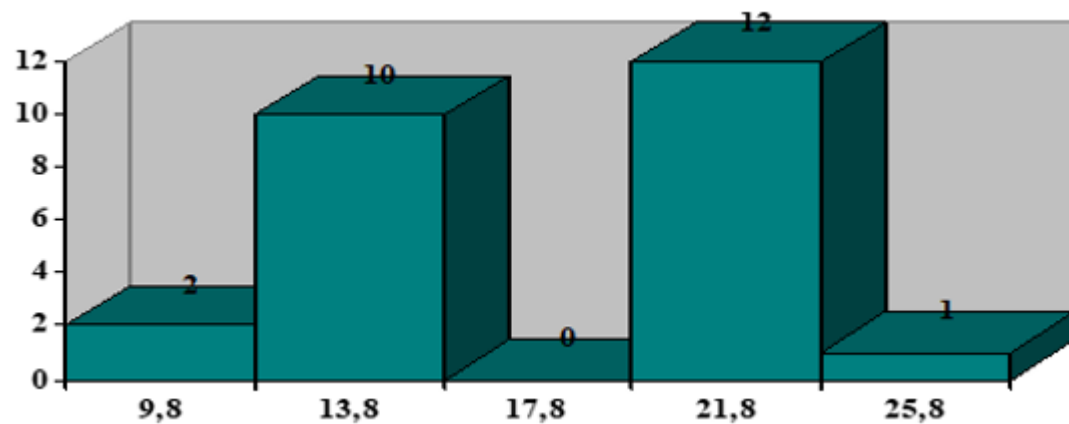
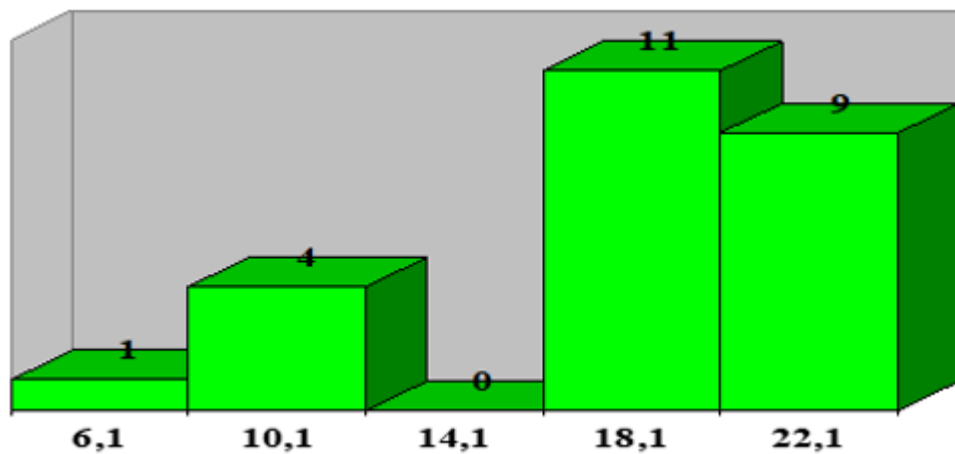
Ktorá je najvhodnejšia:

<https://cran.r-project.org/web/packages/bestNormalize/vignettes/bestNormalize.html>



Štatistické charakteristiky

ŠKÁLA	ŠTATISTICKÉ CHARAKTERISTIKY
Nominálna	Modus, absolútne a relatívne početnosti
Ordinálna	Modus, absolútne a relatívne početnosti medián, kvartily, kvantily, percentily
Intervalová	Modus, absolútne a relatívne početnosti medián, kvartily, kvantily, percentily aritmetický priemer, rozptyl, smerodajná odchýlka, špicatosť, šikmosť
Pomerová	Modus, absolútne a relatívne početnosti medián, kvartily, kvantily, percentily aritmetický priemer, rozptyl, smerodajná odchýlka, špicatosť, šikmosť Geometrický priemer

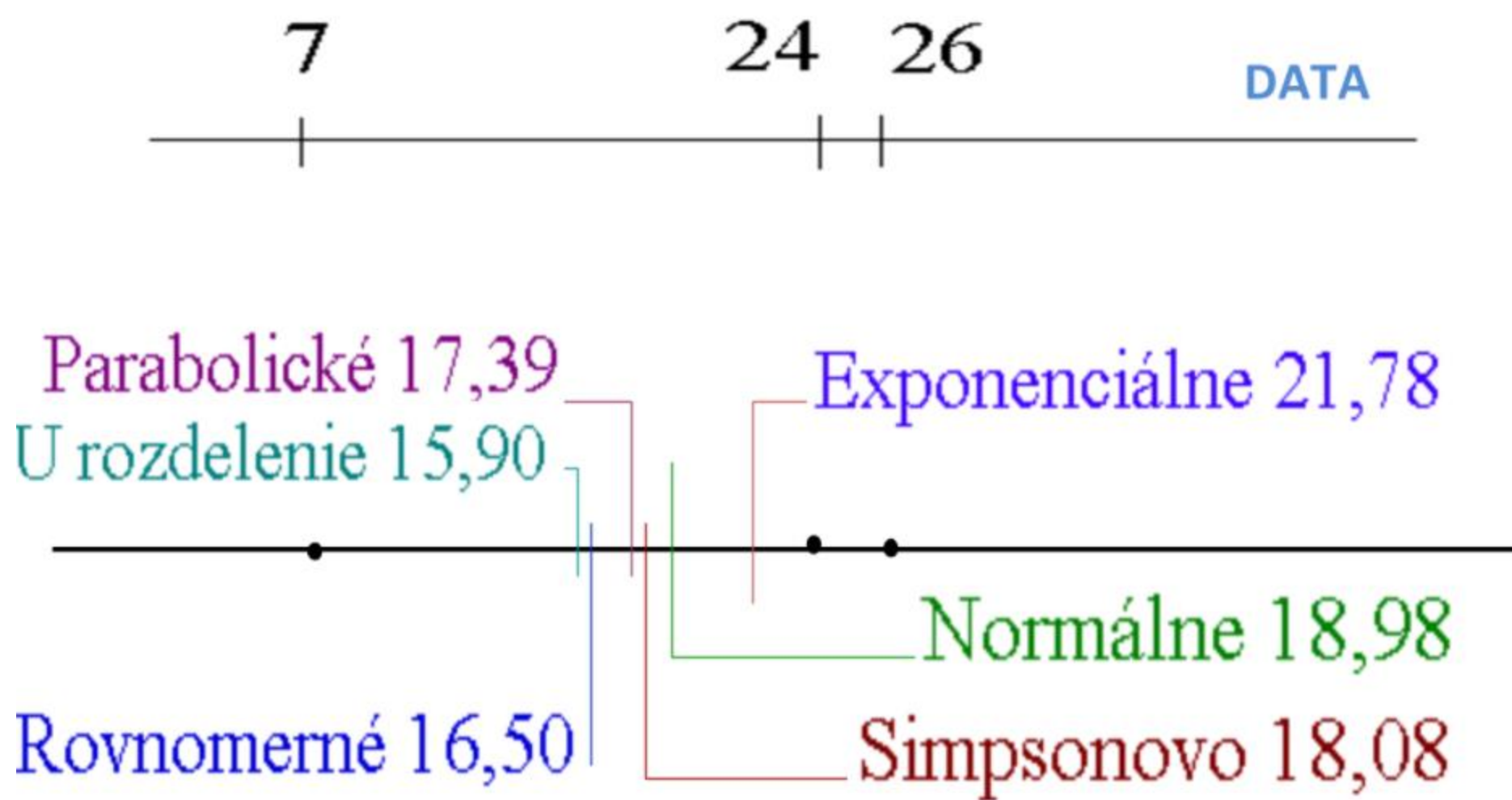


Rozsah výberu $n = 25$

Priemer 17,8

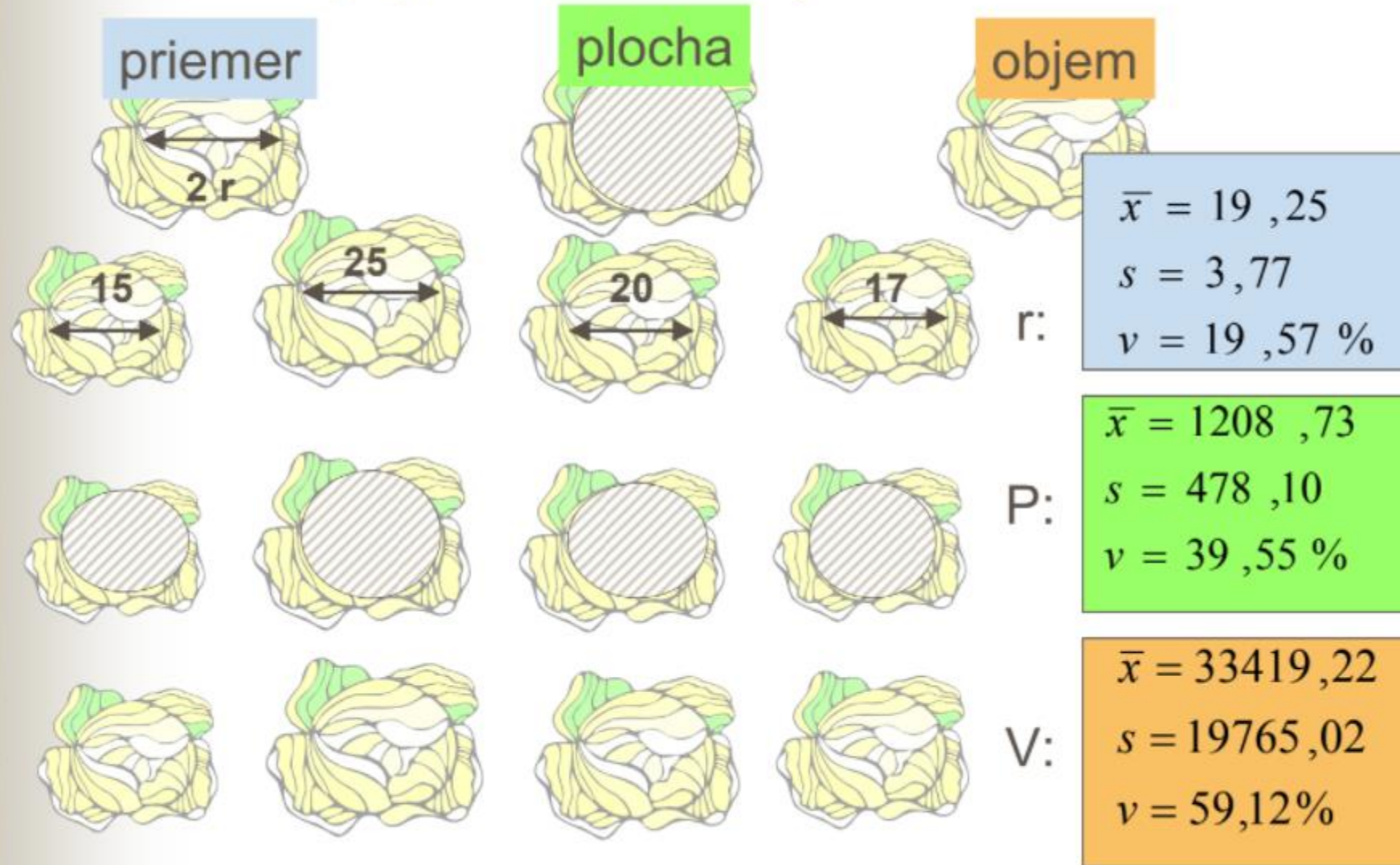
Štandardná odchýlka 4,7

Variačný koeficient 26,4



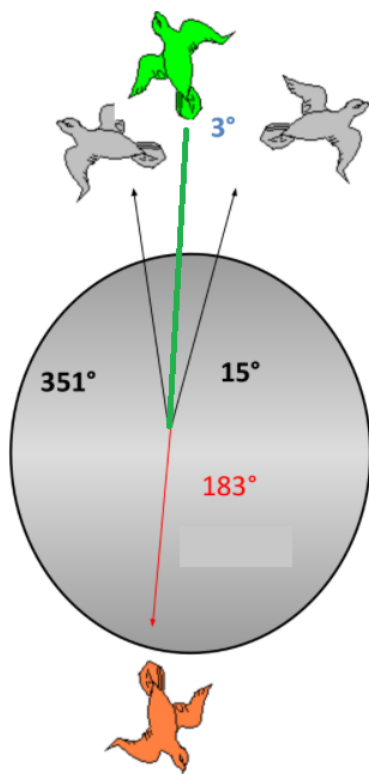


Nie je jedno, aké znaky meriame





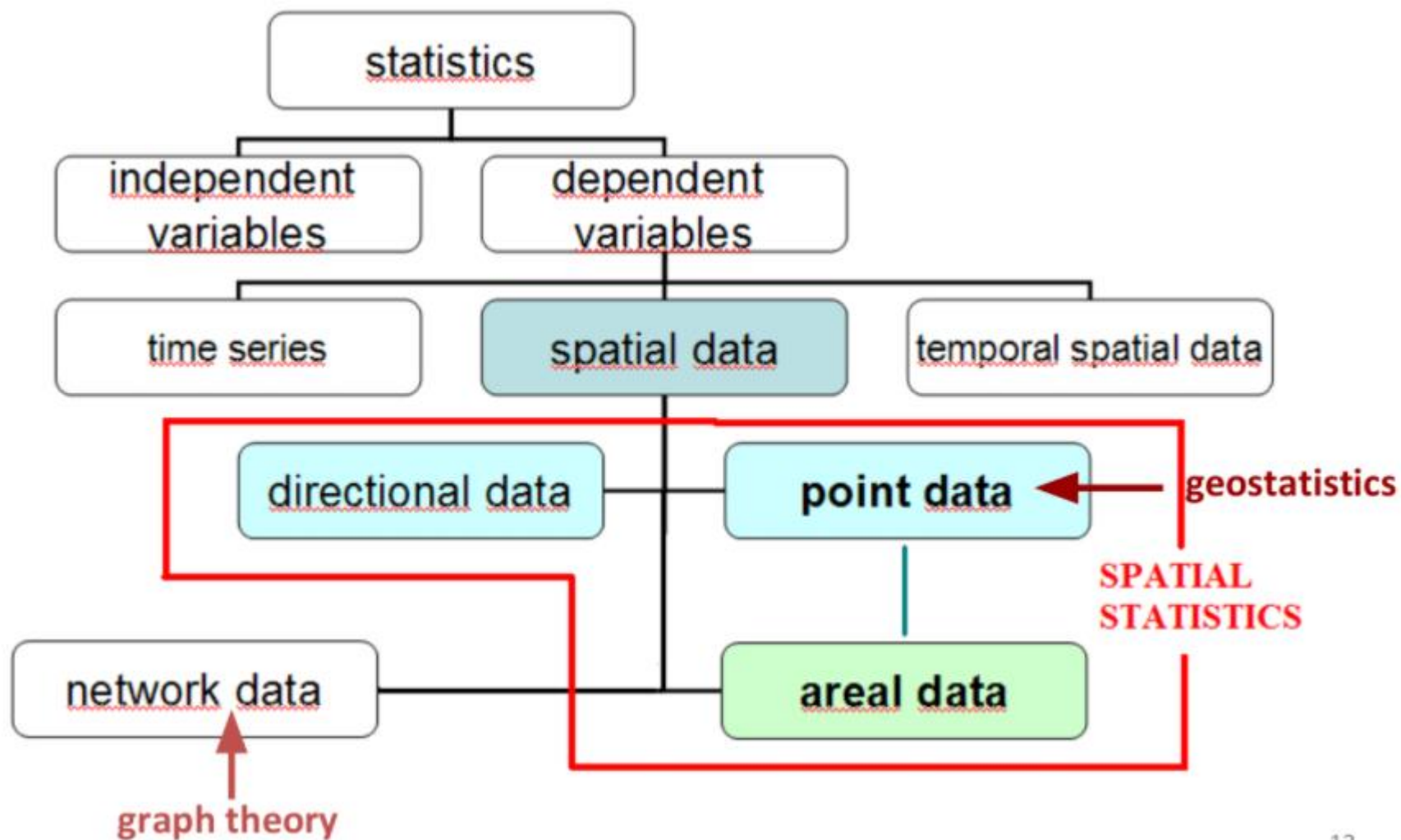
Smerové údaje



$$\bar{x} = \frac{351^\circ + 15^\circ}{2} = 183$$

$$\bar{x} = \arctg \frac{\sum \sin x_i}{\sum \cos x_i}$$

$$\bar{x} = 3^\circ \quad \checkmark$$





Kompozičné údaje (Compositional data)

- V štatistike sú kompozičné údaje kvantitatívnym opisom častí nejakého celku, ktorý poskytuje **relatívne informácie**.
- Matematicky sú kompozičné údaje reprezentované bodmi na simplexnej báze.

The arithmetic mean of the compositional data set is usually not representative of the set's 'centre'. The geometric mean

$$\text{cen}(\mathbf{X}) = \frac{(g_1, \dots, g_D)}{g_1 + \dots + g_D}, \quad (1)$$

where $g_j = \left(\prod_{i=1}^N x_{ij} \right)^{\frac{1}{N}}$ is the geometric mean of the j^{th} component (i.e. column-wise geometric mean) of the data set, is more representative.

Fernández *et al.*, 1998). We use centered log-ratio transformation (clr) of the composition, which is given by

$$\text{clr}(\mathbf{x}) = \left(\log\left(\frac{x_1}{g(\mathbf{x})}\right), \dots, \log\left(\frac{x_D}{g(\mathbf{x})}\right) \right), \quad (2)$$

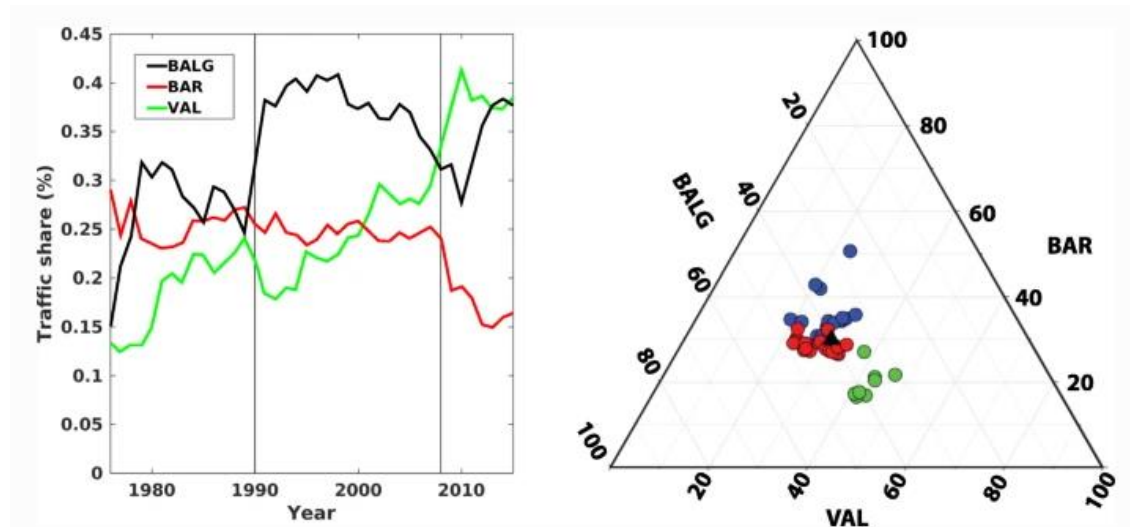
where $g(\mathbf{x})$ is the geometric mean of the composition (row-wise geometric mean).



Kompozičné údaje (Compositional data)

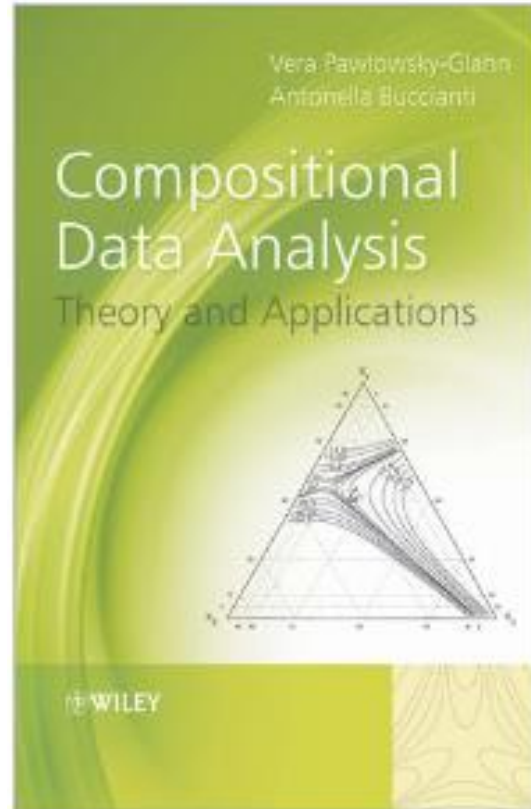
Na skúmanie vývoja podielu dopravy sa používajú štatistické techniky založené na údajoch o zložení priepustnosti kontajnerov v systéme s viacerými portami. Kompozičné vektory sú tie, ktoré obsahujú relatívnu informáciu častí nejakého celku. **Aplikácia konvenčných štatistických techník na kompozičné údaje môže viesť k chybným záverom a falošným koreláciám.**

tri najväčšie prístavy:
Algeciras Bay (BALG),
Barcelona (BAR) a
Valencia (VAL)



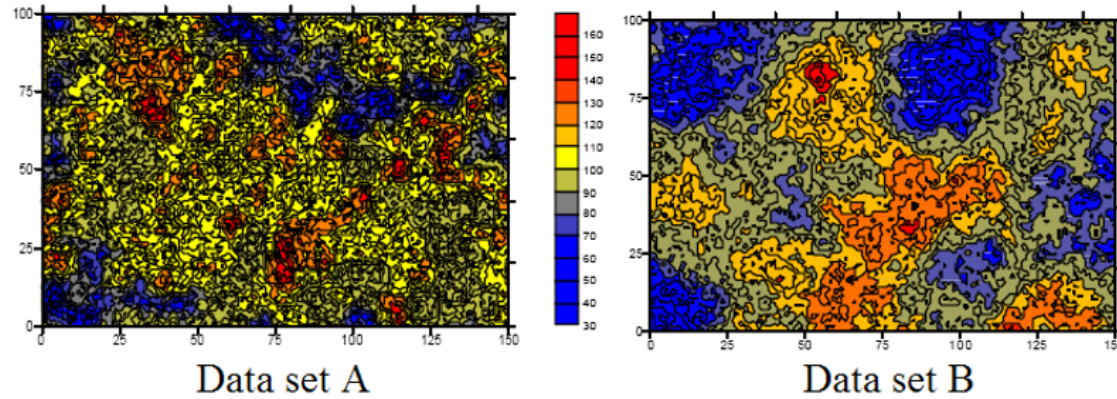
<https://etr.springeropen.com/articles/10.1186/s12544-019-0350-z>

Kompozičné údaje (Compositional data)

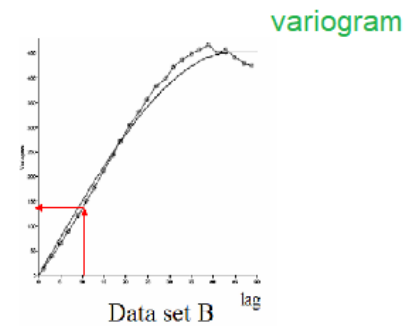
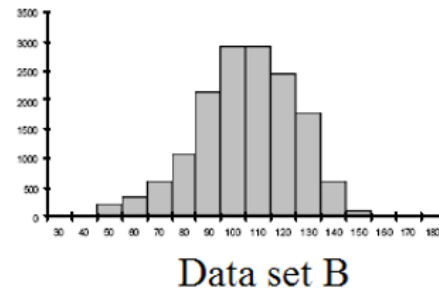
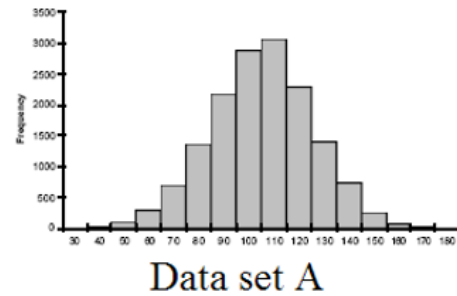


<https://www.wiley.com/en-us/Compositional+Data+Analysis:+Theory+and+Applications-p-9780470711354>

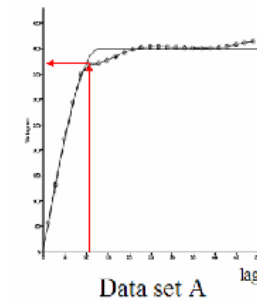
Priestorové údaje



	A	B
Count	15251	15251
Average	100.00	100.00
Standard Deviation	20.00	20.00
Median	100.35	100.92
10 Percentile	73.89	73.95
90 Percentile	125.61	124.72



Small values at short lags
indicate that data are
autocorrelated.

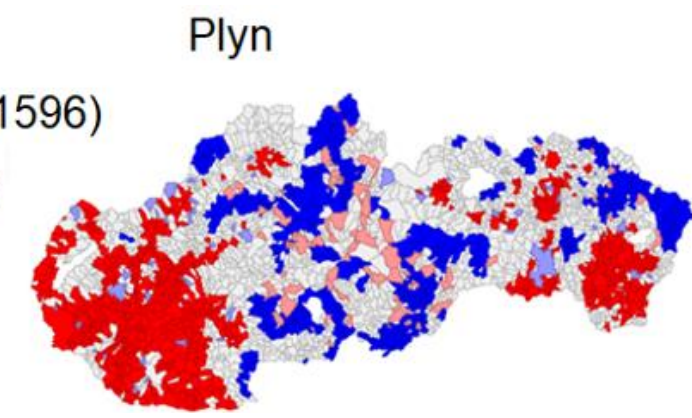
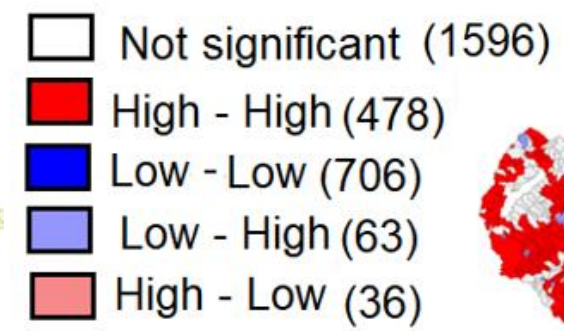
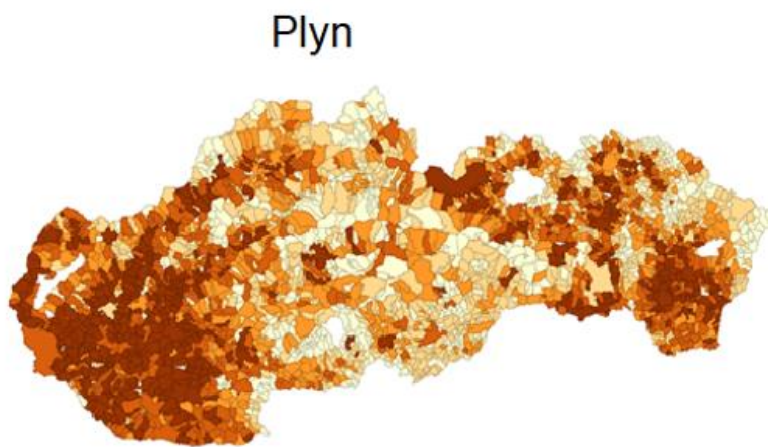
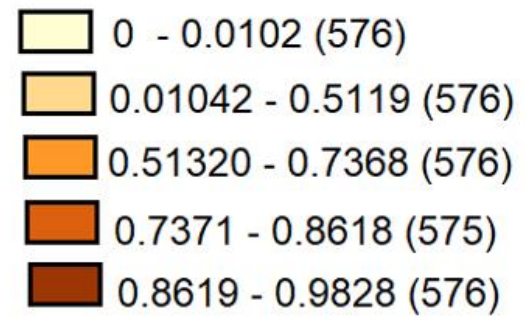
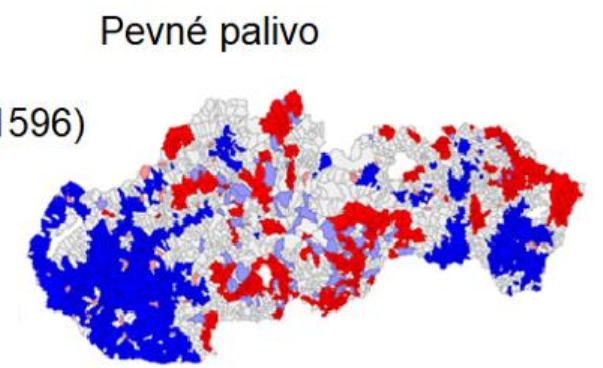
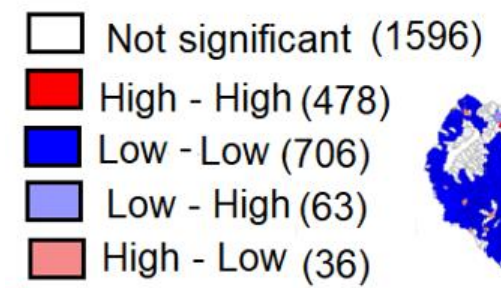
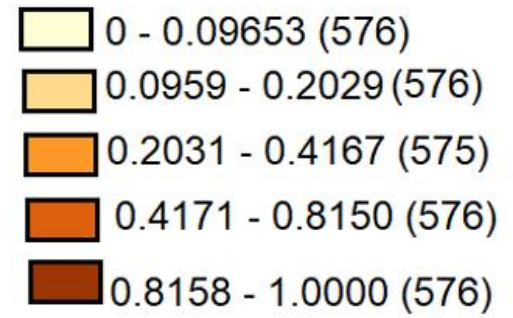


Large values at short lags,
indicate that the paired
samples are dissimilar and
more spatially discontinuous.



Priestorová autokorelácia

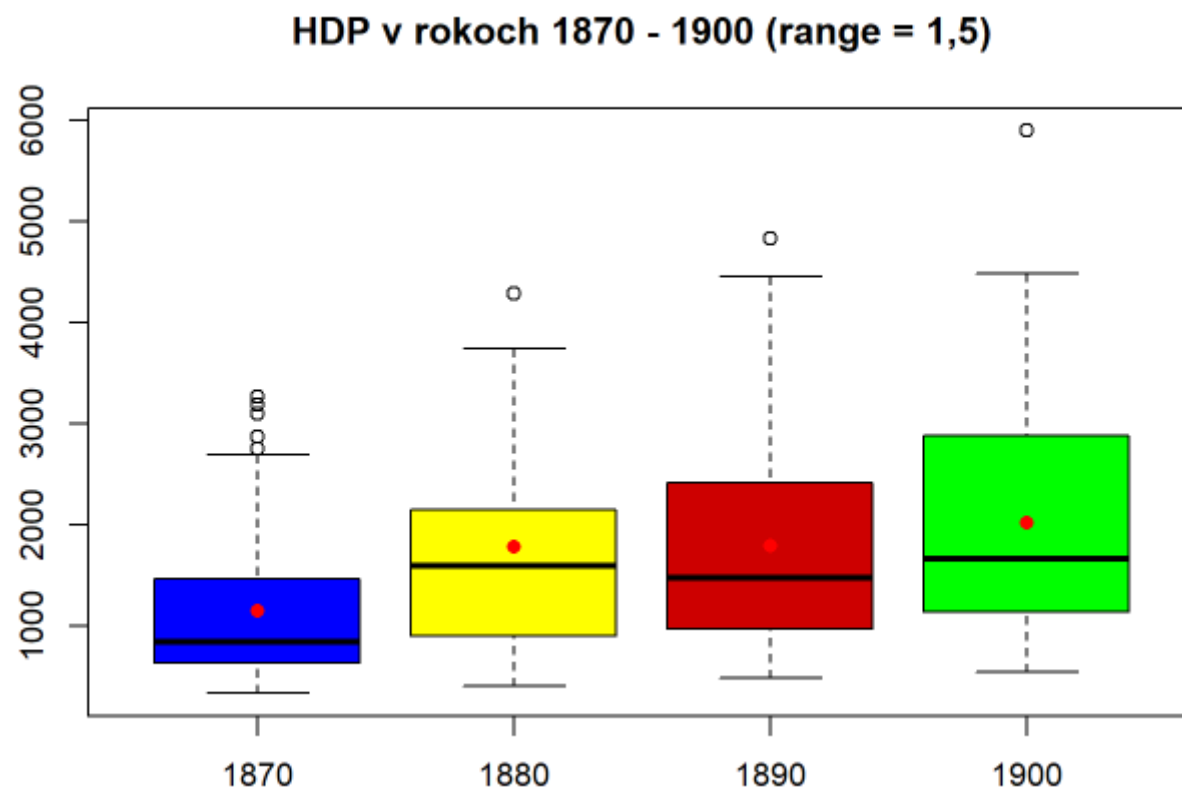
Indikátor	Moranov koeficient	P-hodnota
Podiel bytov vykurovaných plynom	0,5513	0,001
Podiel bytov vykurovaných pevným palivom	0,5571	0,001





Krabičkové grafy

Krabičkový graf je jedným zo spôsobov grafickej vizualizácie údajov.





Korelácia nie je kauzalita



Korelačné koeficienty

V štatistike používame štyri typy korelácií:

- Parametrické: Pearsonova korelácia
- Neparametrické:
 - Kendalova poradová korelácia,
 - Spearmanova korelácia
 - Bodovo biserialova korelácia.

V prospech Spearmanovho rho: V praxi sa zdá byť populárnejší.

V prospech Kendallovho tau: Spearmanov rho je citlivejší na chyby a nezrovnalosti v údajoch.

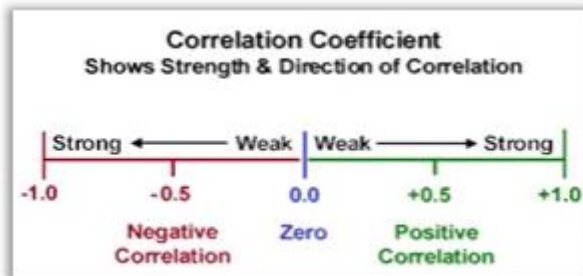
Keď sú údaje normálne, Kendalovo tau má menšiu citlivosť na hrubú chybu a menší asymptotický rozptyl.

Spearmanove hodnoty rho bývajú vyššie ako hodnoty tau Kendalla.



Pearsonov korelačný koeficient

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$



<i>r</i> value	Interpretation
+ .70 or higher	Very strong positive relationship
+ .40 to + .69	Strong positive relationship
+ .30 to + .39	Moderate positive relationship
+ .20 to + .29	weak positive relationship
+ .01 to + .19	negligible relationship
0	No relationship [zero order correlation]
- .01 to - .19	negligible relationship
- .20 to - .29	weak negative relationship
- .30 to - .39	Moderate negative relationship
- .40 to - .69	Strong negative relationship
- .70 or higher	Very strong negative relationship

<https://www.elegantjbi.com/blog/what-is-karl-pearson-correlation-analysis-and-how-can-it-be-used-for-enterprise-analysis-needs.htm>

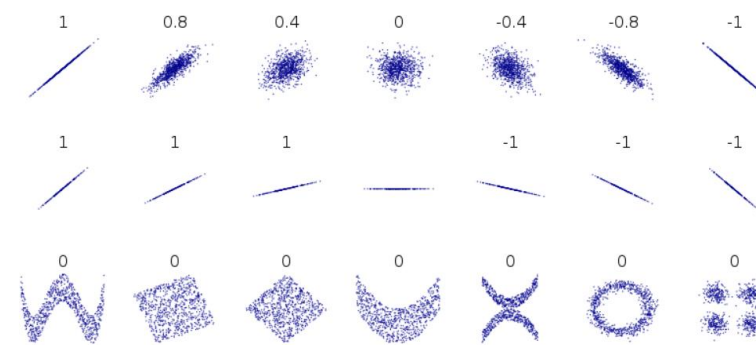
Size of Correlation	Interpretation
.90 to 1.00 (-.90 to -.1.00)	Very high positive (negative) correlation
.70 to .90 (-.70 to -.90)	High positive (negative) correlation
.50 to .70 (-.50 to -.70)	Moderate positive (negative) correlation
.30 to .50 (-.30 to -.50)	Low positive (negative) correlation
.00 to .30 (.00 to -.30)	negligible correlation

<http://www.bioline.org.br/pdf?mm12018>

Signifikantnosť Pearsonovho korelačného koeficienta — test $H_0: r_0 = 0$ (p hodnota < hladina významnosti, potom H_0 zamietame)

Miera lineárnej závislosti

Hodnota veľkosti účinku
Pearsonovej r korelácie sa
**pohybuje medzi -1 (dokonalá
negatívna korelácia) až +1
(dokonalá pozitívna korelácia)**





Veľkosť efektu

Odporúča sa, aby **výskumníci uvádzali veľkosti efektov** (vid.

<https://academic.oup.com/innovateage/article/3/4/igz036/5560156>)

- Veľkosť efektu (effect size) nám **povie**,
 - ako významný je vzťah medzi premennými alebo
 - ako významný je rozdiel medzi skupinami

Softvér:

<https://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-R-main.php>



Veľkosť efektu pre Pearsonov korelačný koeficient

Cohen, J. (1988). Statistical power analysis for the behavioral sciences, 2nd ed. Hillsdale, NJ: Erlbaum

Podľa Cohena (1988, 1992) je veľkosť účinku nízka, ak sa hodnota r pohybuje okolo 0,1, stredná, ak sa r pohybuje okolo 0,3, a veľká, ak sa r mení o viac ako 0,5.

Strength of Association	Coefficient, r	
	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

ALE závisí od kontextu. Pre ekologickú terénnu štúdiu je r 0,1 dosť veľké a pre vzťah medzi dávkou a odozvou v experimente bunkovej kultúry je $r=0,9$ skôr nízke. „**Jedna veľkosť sa hodí pre všetko**“ nikdy nefunguje.“

Hodnota p nestačí

Nižšia p - hodnota sa niekedy interpretuje tak, že medzi dvoma premennými existuje silnejší vzťah. **Štatistická významnosť** však znamená, že je nepravdepodobné, že nulová hypotéza je pravdivá ($H_0: \rho = 0$) (menej ako 5 %).

Preto významná hodnota p nám hovorí, že zásah funguje, zatiaľ čo veľkosť účinku nám hovorí, do akej miery funguje.

Možno tvrdiť, že zdôraznenie veľkosti účinku podporuje vedeckejší prístup, keďže na rozdiel od testov významnosti je veľkosť účinku nezávislá od veľkosti vzorky.



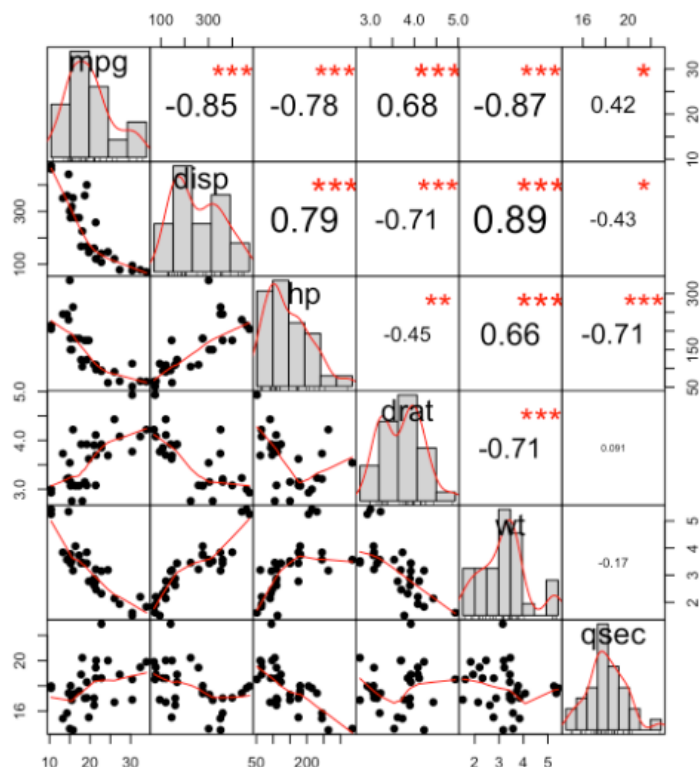
Pearsonov korelačný koeficient

Ak sú premenné nezávislé , Pearsonov korelačný koeficient je 0, ale opak neplatí, pretože korelačný koeficient detekuje iba lineárne závislosti medzi dvoma premennými.

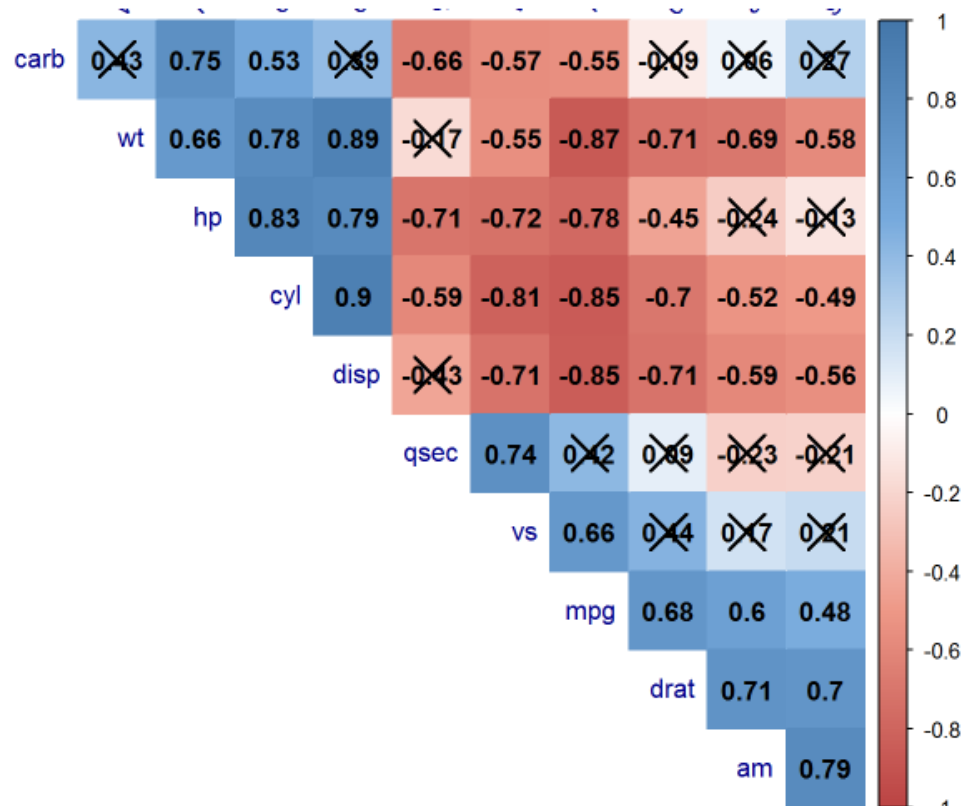
Avšak v špeciálnom prípade, keď (X,Y) majú dvojrozmerné normálne rozdelenie, nekorelácia je ekvivalentná nezávislosti.



Pearsonov korelačný koeficient



<http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software>



https://rstudio-pubs-static.s3.amazonaws.com/240657_5157ff98e8204c358b2118fa69162e18.html



Kontingenčné tabuľky

Kontingenčná tabuľka je tabuľka, ktorá sa používa na prehľadné zhrnutie vzájomného vzťahu dvoch kategoriálnych štatistických znakov.

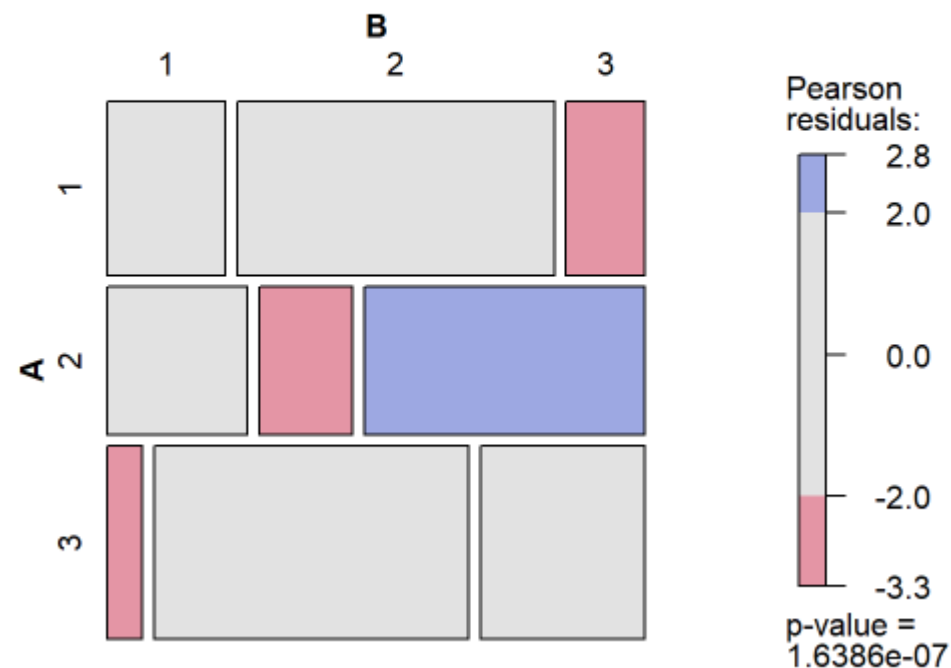


Kontingenčné tabuľky

Najvyššie dosiahnuté vzdelanie	Preferovaná politika		
	žiadne reštrikcie	fajčenie na vyhradených miestach	zákaz fajčenia
základná škola	15	40	10
stredná škola	15	100	30
vysoká škola	5	44	23

- Rozsah výberového súboru je 282. Hodnota testovacej štatistiky je 37,2 (počet stupňov voľnosti je 4). Výsledky výpočtov naznačujú signifikantný vzťah medzi vzdelaním a preferovaním druhu politiky fajčenia na verejných priestranstvách (p hodnota testovacej štatistiky je 0,0000001639). Veľkosť efektu je stredne vysoká 0,2955

Mozaikový graf





Viacrozmerne metody

Nezávislá premenná (premenné)		Závislá premenná (premenné)			
		Jedna		Viac ako jedna	
		Kvantitatívna	Kvalitatívna	Kvantitatívna	Kvalitatívna
Jedna	Kvantitatívna	Regresná analýza	Logistická regresia Diskriminačná analýza	Kanonická korelačná analýza	Viacskupinová diskriminačná analýza
	Kvalitatívna	t-test	Diskrétna diskriminačná analýza	MANOVA	Diskrétna viacskupinová diskriminačná analýza
Viac ako jedna	Kvantitatívna	Viacnásobná regresná analýza	Logistická regresia Diskriminačná analýza	Kanonická korelačná analýza	Viacskupinová diskriminačná analýza
	Kvalitatívna	ANOVA	Diskrétna diskriminačná analýza Conjoint analýza	MANOVA	Diskrétna viacskupinová diskriminačná analýza



Počet premenných	Typ údajov	
	Kvantitatívne	Kvalitatívne
Dve	Jednoduchá korelácia	Analýza dvojrozmerných kontingenčných tabuliek
		Loglineárne modely
Viac ako dve	Analýza hlavných komponentov	Analýza viacrozmerných kontingenčných tabuliek
		Loglineárne modely
	Faktorová analýza	Korešpondenčná analýza



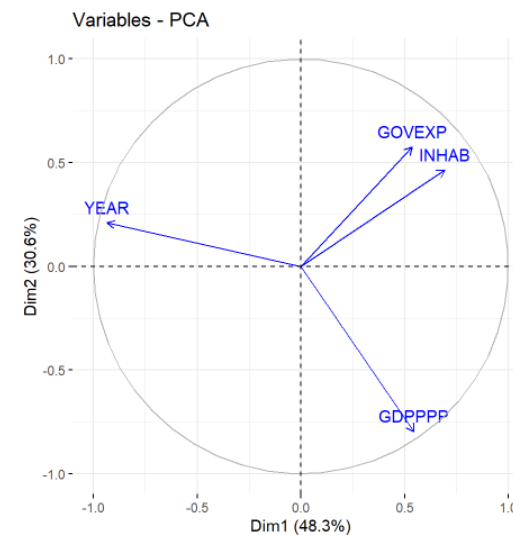
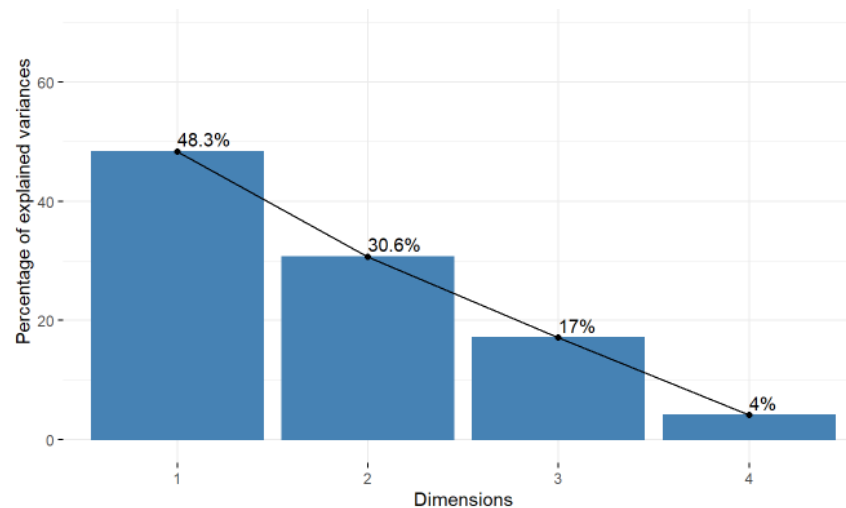
Analýza hlavných komponentov (PCA) a biplot

- Analýza **hlavných komponentov (PCA)** nám umožňuje sumarizovať a vizualizovať informácie v súbore údajov obsahujúcom pozorovania popísané viacerými vzájomne korelovanými kvantitatívnymi premennými. Keď každú premennú považujeme za dimenziu a ak náš súbor obsahuje viac ako tri premenné, vizualizácia je problematická.
- Analýzu hlavných komponentov používame na extrakciu dôležitých informácií z datovej matice s viacerými premennými a na vyjadrenie týchto informácií pomocou nových premenných, ktoré sa nazývajú hlavné komponenty. Počet hlavných komponentov je menší alebo rovnaký ako počet pôvodných premenných.



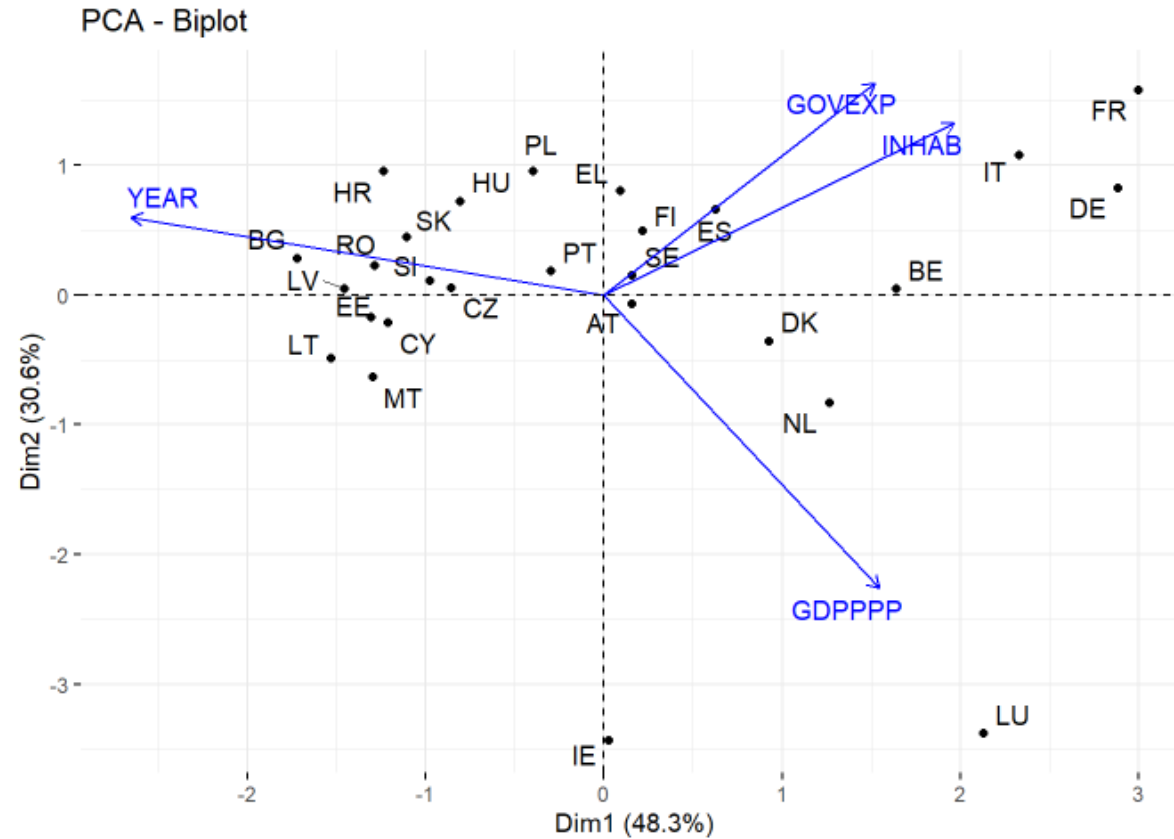
Analýza hlavných komponentov (PCA) a biplot

Štát	GDPPPP	INHAB	GOVEXP	YEAR
BE	36 700	11 455 519	52,1	1958
BG	16 500	7 000 039	36,3	2007
CZ	28 900	10 649 800	41,4	2004
DK	40 500	5 806 081	49,2	1973
DE	37 500	83 019 213	45,2	1958
EE	26 100	1 324 820	38,9	2004
IE	60 200	4 904 240	24,6	1973
EL	20 700	10 724 599	47,9	1981
ES	28 400	46 937 060	42,1	1986
FR	33 100	67 177 636	55,4	1958
HR	20 300	4 076 246	47,2	2013
IT	29 800	59 816 673	48,6	1958
CY	27 900	875 899	39,7	2004
LV	21 500	1 919 968	38,4	2004
LT	26 000	2 794 184	34,6	2004
LU	81 000	613 894	42,3	1958
HU	22 800	9 772 756	45,7	2004
MT	31 100	493 559	36,8	2004
NL	39 900	17 282 163	42,0	1958
AT	39 400	8 858 775	48,6	1995
PL	22 700	37 972 812	41,8	2004
PT	24 700	10 276 617	42,5	1986
RO	21 700	19 414 458	36,2	2007
SI	27 700	2 080 908	43,3	2004
SK	21 900	5 450 421	42,7	2004
FI	34 700	5 517 919	53,2	1995
SE	37 000	10 230 185	49,3	1995



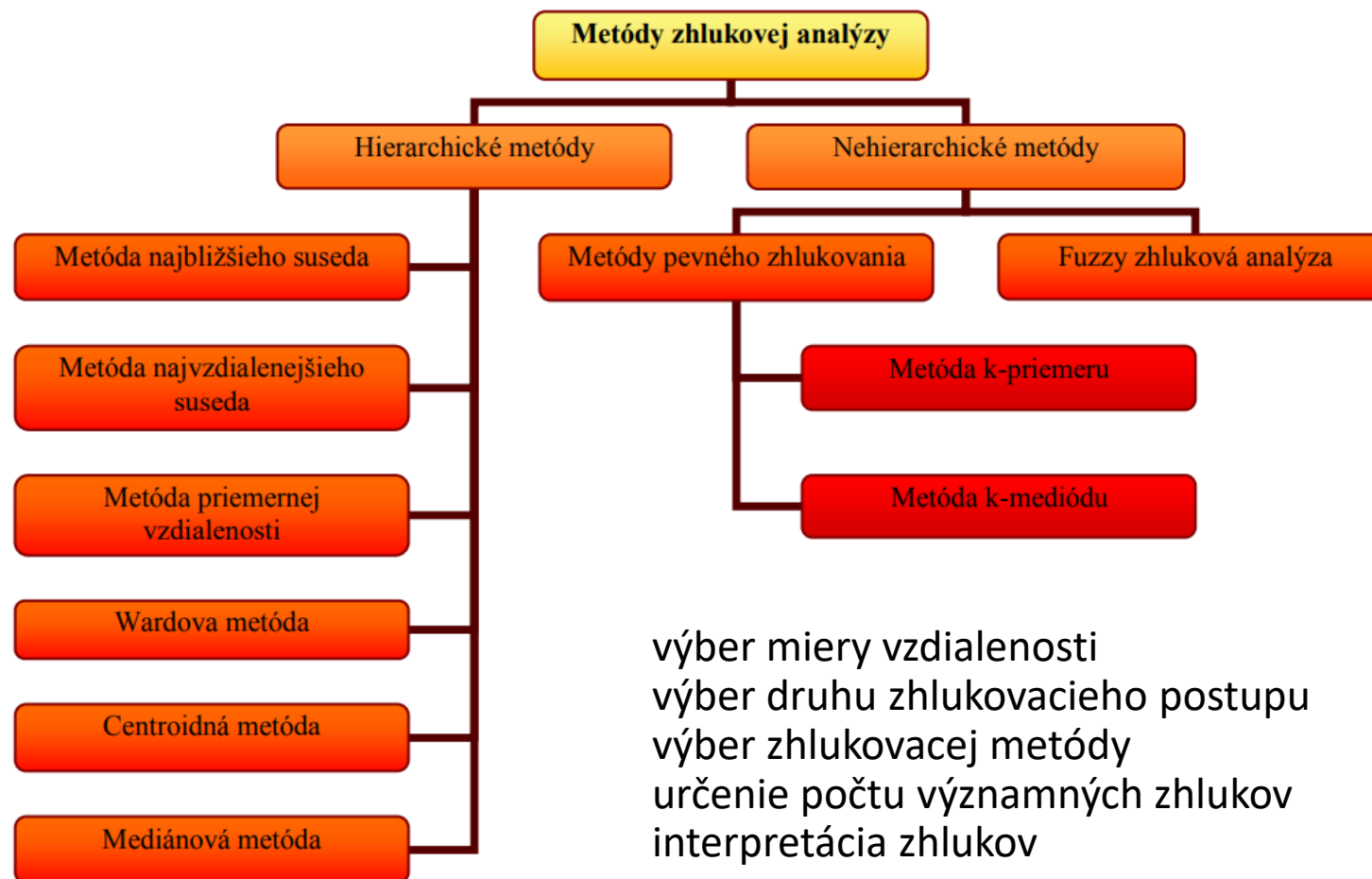


Analýza hlavných komponentov (PCA) a biplot





Klasifikácia metód zhlukovej analýzy

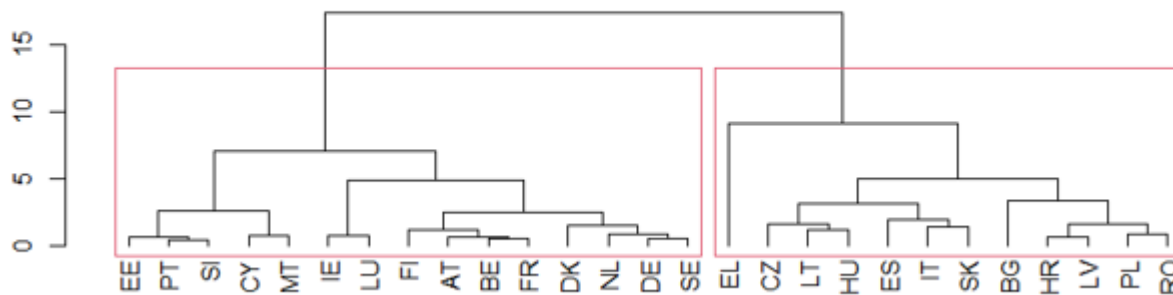




Dendrogram

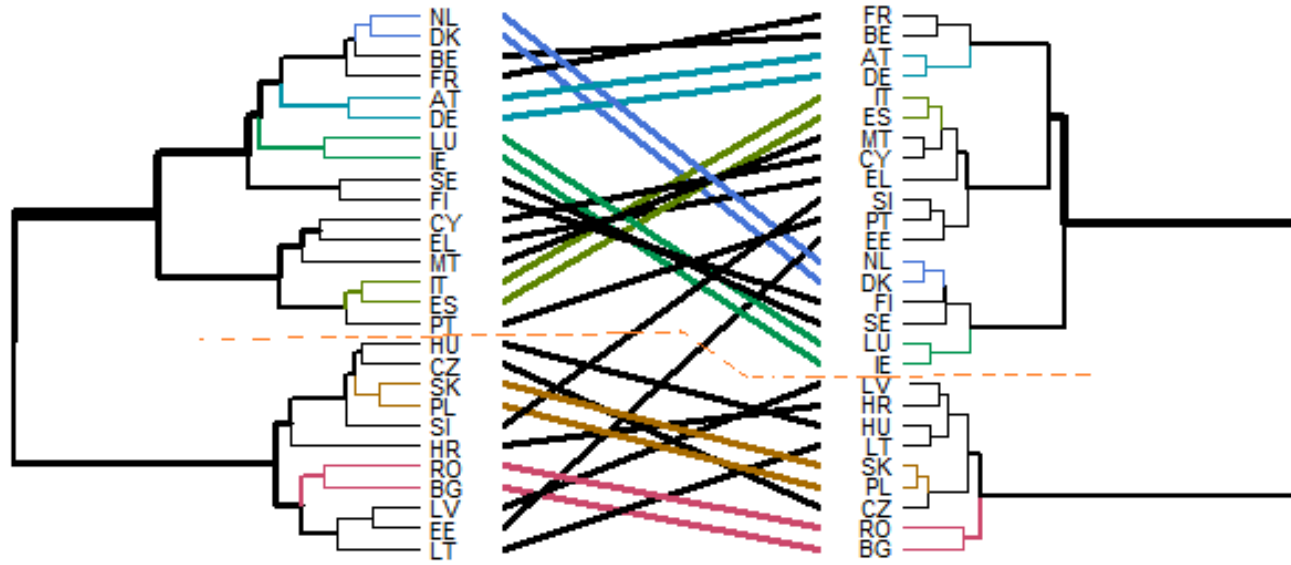
- Údaje majú tendenciu zhlukovať sa, lebo hodnota Hopkinsovej štatistiky je vysoká (0,7298). Na zhlukovanie sme použili Wardovu metódu, lebo hodnota aglomeratívneho koeficientu bola pre túto metódu najvyššia (average: 0,8073, single: 0,7898, complete: 0,8293, ward: 0,8518). Optimálny počet zhlukov je dva.

Obr. 7: Dendrogram



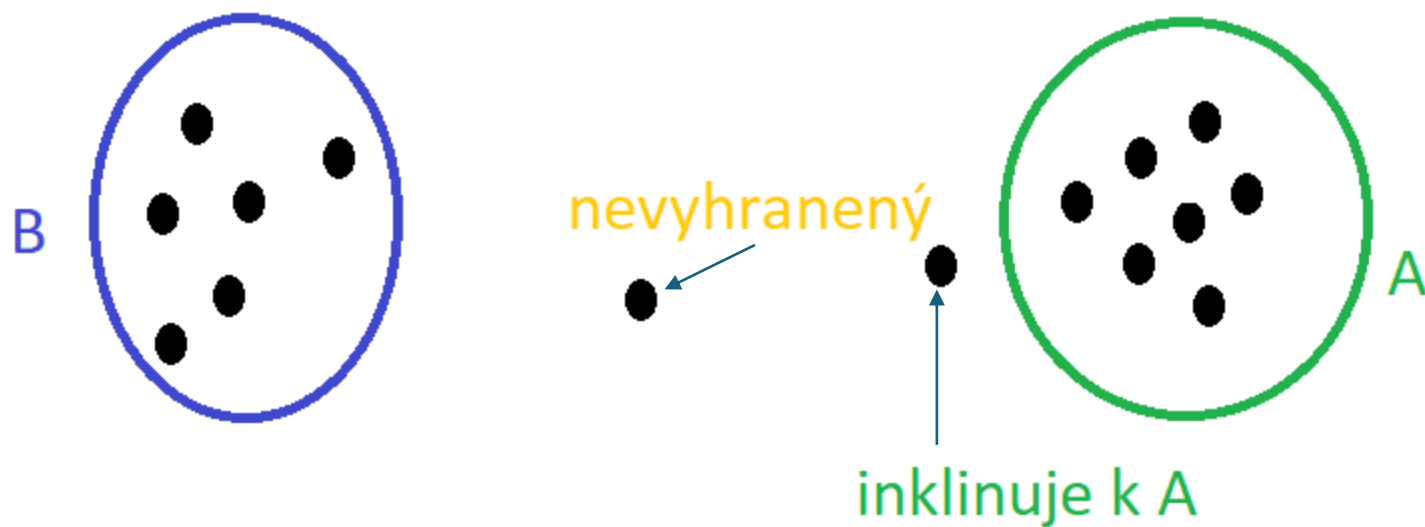


Tanglegram





Fuzzy zhukovanie





Mapové zobrazenie výsledkov fuzzy c-zhlukovej analýzy



zhluk 1 (681)

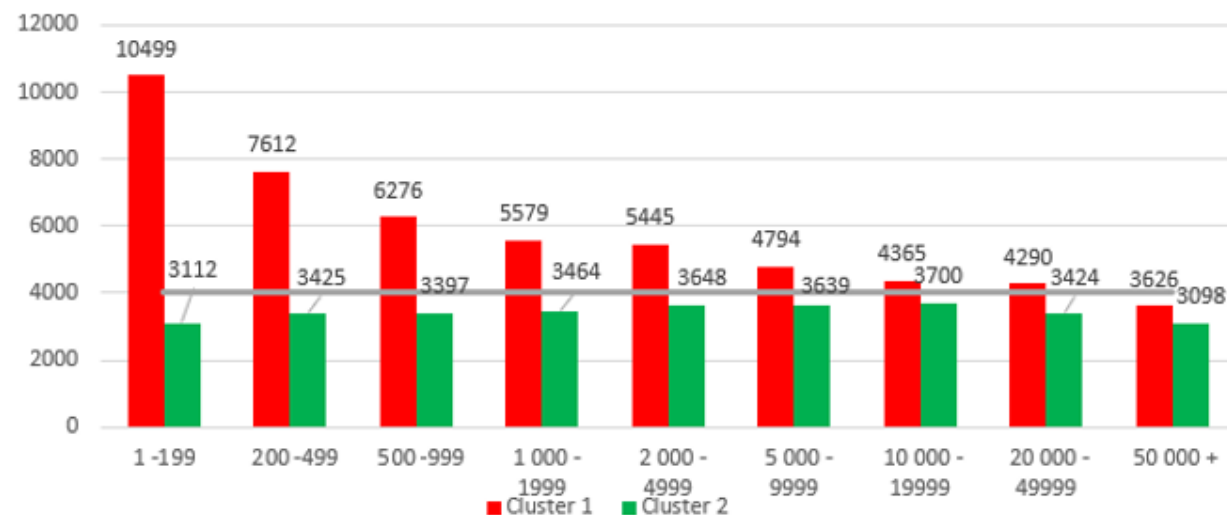
zhluk 2 (1586)

inklinuje k zhluku 1 (326)

inklinuje k zhluku 2 (286)

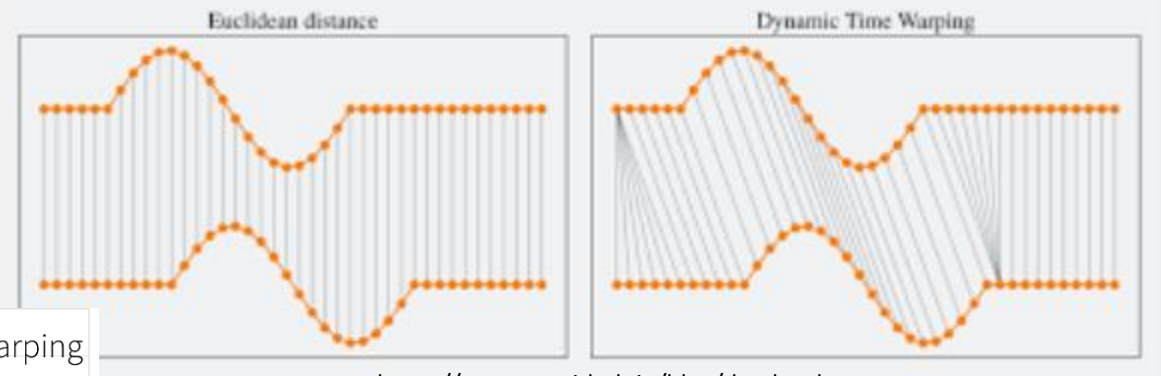
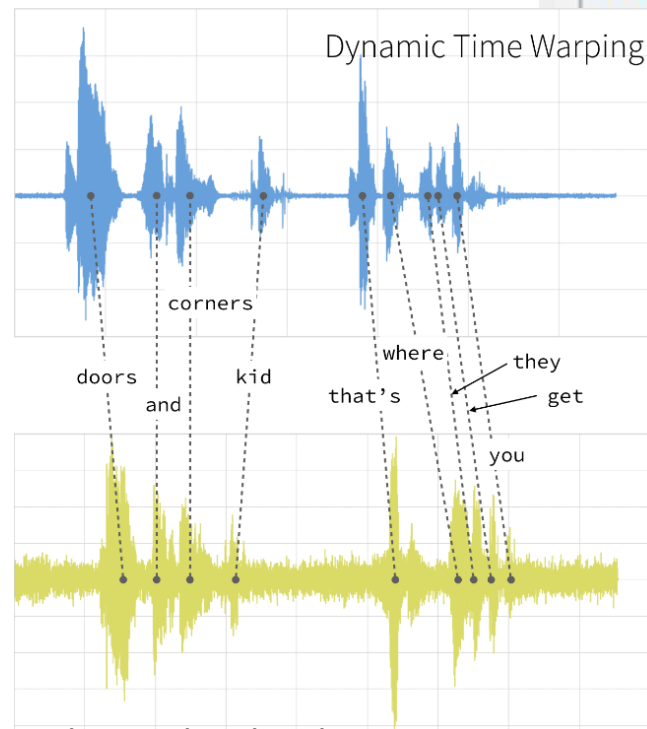
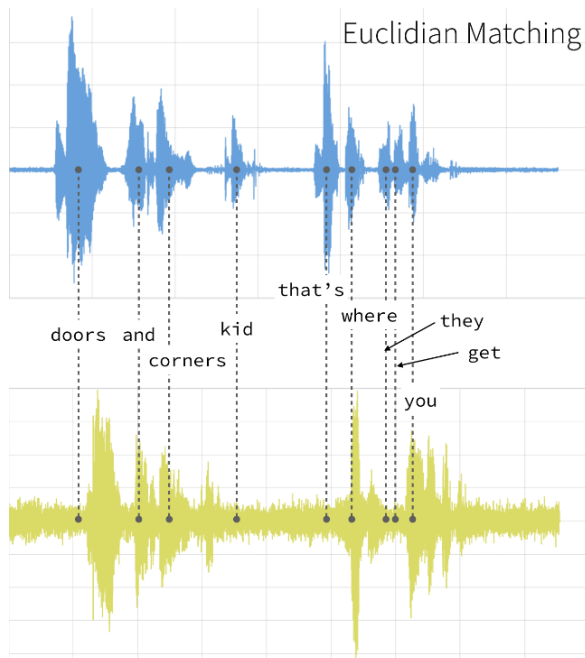
Moranov koeficient 0,1277 (P hodnota je 0,0001)

Hodnoty PYLL typických predstaviteľov zhlukov podľa veľkostných kategórií





Dynamic Time warping



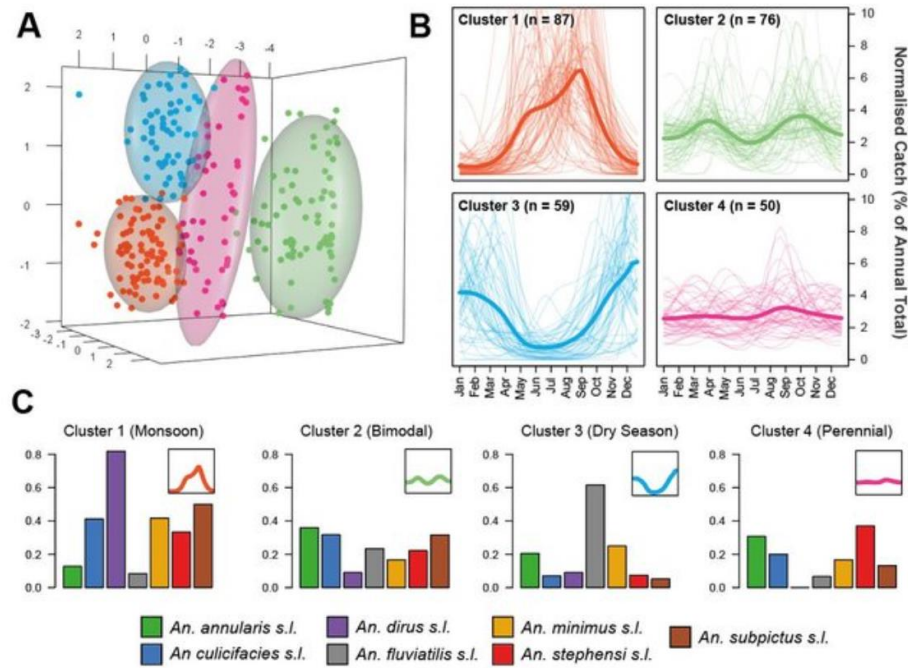
<https://rtavenar.github.io/blog/dtw.html>

<https://www.databricks.com/blog/2019/04/30/understanding-dynamic-time-warping.html>

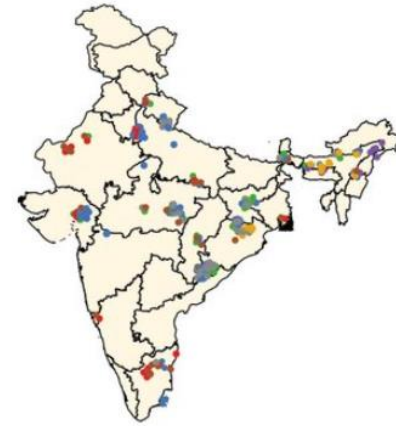


Dynamic Time warping

- Charakterizácia a zhlukovanie časových radov s podobnými časovými vlastnosťami



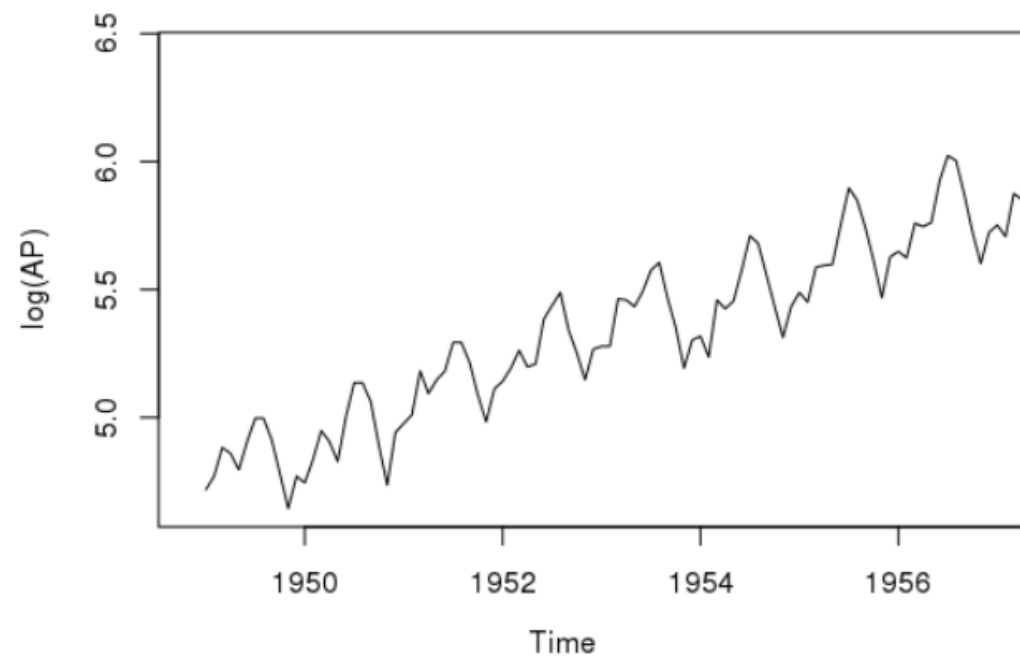
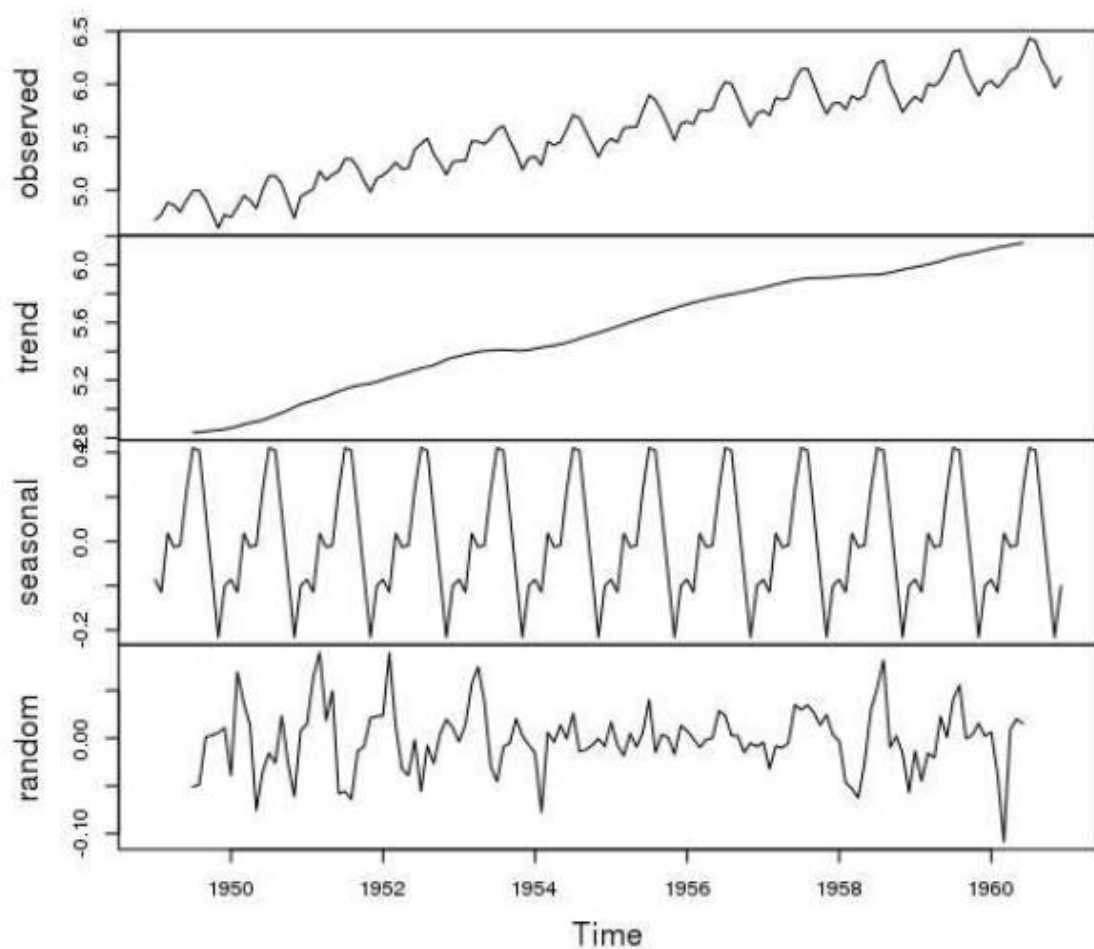
Skúmanie druhovo špecifických vzorcov populačnej dynamiky komárov .



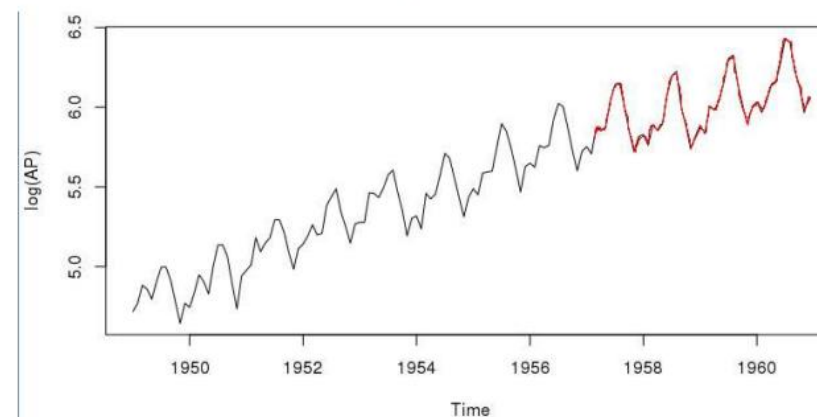
https://www.researchgate.net/publication/348494644_The_Ecological_Structure_of_Mosquito_Population_Dynamics_Insights_from_India/figures?lo=1



Časové rady



Intuitívne: ak nenastane nejaký šok, zachová sa rastúci trend a sezónnosť, teda nejako takto:





Časové rady

- **Kointegrácia** - V kointegrácii je zahrnutá informácia o dlhodobých vzťahoch pôvodných nestacionárnych radov, teda dlhodobá rovnováha medzi premennými.
- **ARDL modely**

