

Digital humanities

Analýza sociálních sítí a korespondenční analýza

Jindřich Marek

Analýza sociálních sítí a koresp. analýza

- porovnání výsledků v prezentaci
- naše zdrojová data pro oba typy analýz:
 - přehled o barokní tiskové produkci v Praze
 - bibliografické záznamy
 - odvozená tabulka s cca 13 tis. záznamy: přiřazení jazyka a tématu
 - dvě skupiny kategorií
 - jazyk (filtr na jen tři nejvíce zastoupené: latina, němčina a čeština)
 - téma (náboženství, právo, ...)

Analýza sociálních sítí

- struktura sociálních a kulturních vazeb mezi aktéry (lidé, instituce, postavy v románech, ale i např. kategorie)
 - sociogram (grafické vyjádření vztahů osoby): uzly (nodes) a hrany (edges)
 - hledá klíčové postavy, silné a slabé vazby, komunity, periferie

Analýza sítě: podkladové výpočty

- je to v podstatě vážená spektrální (nebo singulární) dekompozice příslušné matice (matice sousednosti, Laplaceovy matice nebo matice incidence)
 - získané faktory představují základní struktury – centrální uzly, komunity a latentní dimenze, které určují dynamiku a uspořádání sítě
- postup
 - původní data
 - kontingenční tabulka (příklad: řádky = jazyky, sloupce = témata, hodnoty = počty)
 - spektrální/singulární dekompozice: výsledky
 - vlastní vektory (např. Fiedlerův vektor): reprezentují uzly (z řádků i sloupců)
 - centrální míry (např. PageRank, centrality): měří důležitost jednotlivých uzlů
 - zpravidla ještě embedding (snížení počtu rozměrů v datech)

Síťové grafy

- výpočet (matrix)
- vizualizace (layout)
- popis (interpretace)

a. Image speaks for itself



"El resultado es una imagen real de lo que son hoy en día las Humanidades Digitales a nivel global"

Pino-Díaz J. and Fiorimonte D. (2018). "La Geopolítica De Las Humanidades Digitales: Un Caso De Estudio De DH2017 Montreal", DH2018, Mexico City.

b. Helps to visualise the argument



Figure 4. Largest components of social networks from problem and data-matrix and from deconstructed datasets (center and right-most), with $\gamma = 0.5$ and $\gamma = 1.0$.

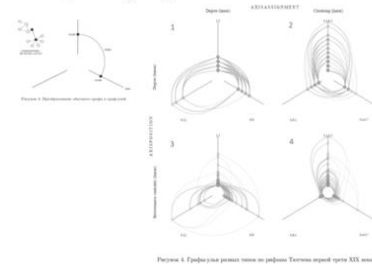
Colavizza G. et al. (2016). "A Method for Record Linkage with Sparse Historical Data", DH2016, Krakow.

c. Highlights a specific node



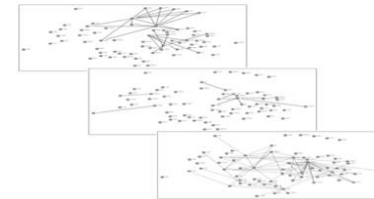
Moretti G. et al. (2016). "Building Large Persons' Networks to Explore Digital Corpora", DH2016, Krakow.

d. Comparing layouts



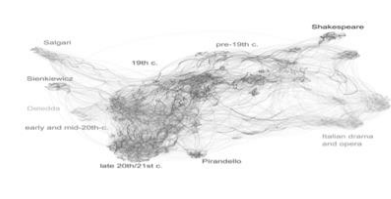
Sozinova O. (2016). "Complex Networks-Based Approach to Russian Rhyme History Description: Linguistics and Database", DH2016, Krakow.

e. Same network over time



Wright C. (2016). "The Formation of Australia's Economic History Community, 1950-1970: A Multi-dimensional Network Analysis", DH2016, Krakow.

f. Visualising communities



Rybicki J. et al. (2018). "Polysystem Theory And Macroanalysis: A Case Study Of Senkiewicz In Italian", DH2018, Mexico City.

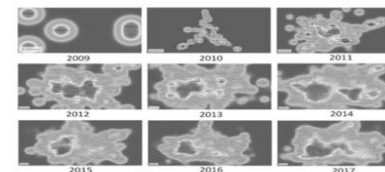
g. Image presented as a map



Figure 2. The provisional ACA network map in DH, data from journals *CHum*, *LLC/DSH*, and *DHJ*, 1966-2016, created using VOSviewer

Gao J. et al. (2017). "The Intellectual Structure of Digital Humanities: An Author Co-Citation Analysis", DH2017, Montreal.

h. Comparing densities



Gao J. et al. (2018). "Visualising The Digital Humanities Community: A Comparison Study Between Citation Network And Social Network", DH2018, Mexico.

i. Visualising modularity clustering



Figure 1. The four Great Awakenings in traditional classification (top); divided by modularity into 3 (center) and 4 (bottom) groups.

Choinski M. and Rybicki J. (2017). "Networks of the Great Awakenings: Classification of Puritan Sermons by Word Usage Statistics", DH2017, Montreal.

Visual analysis

Overall organisation

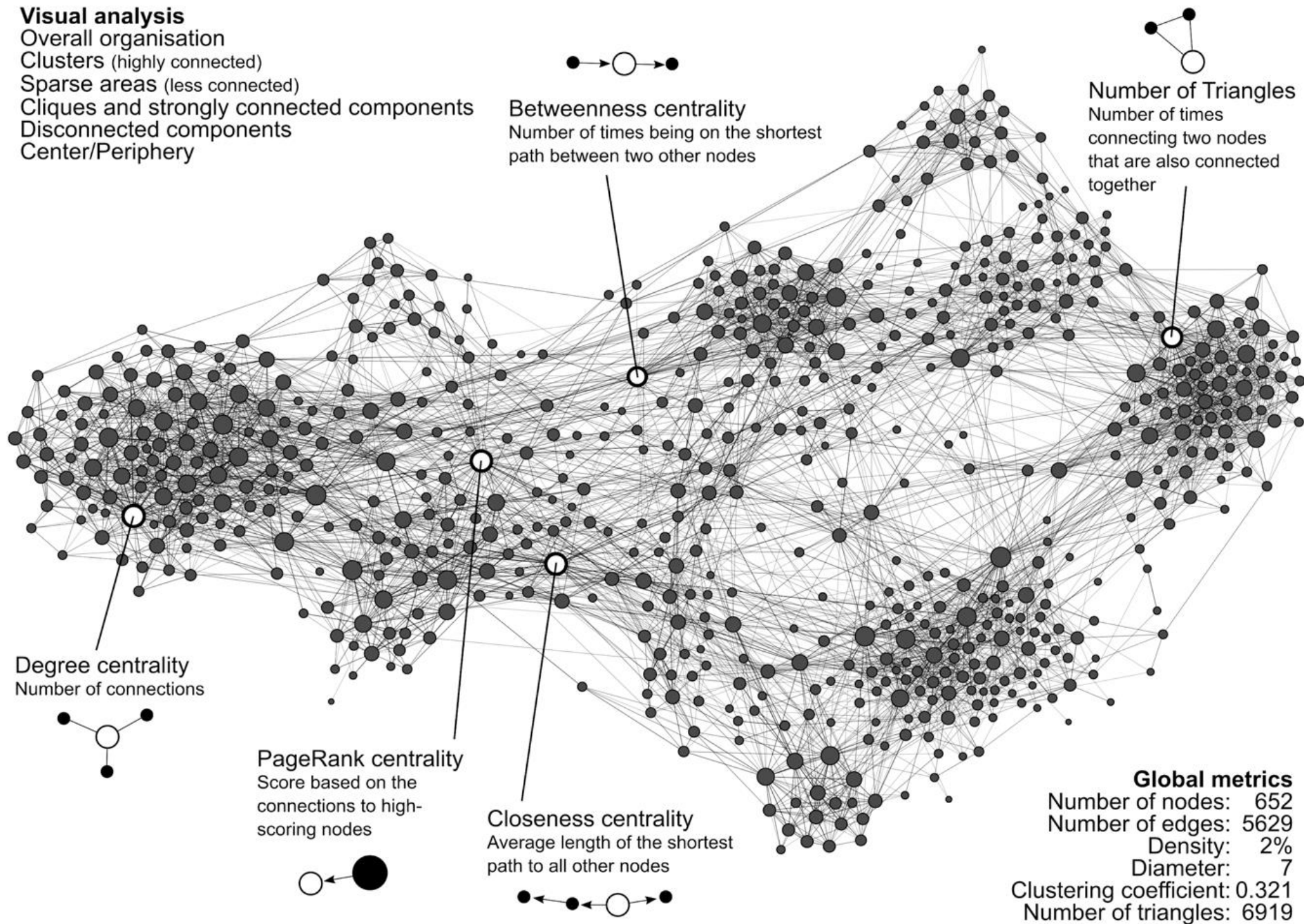
Clusters (highly connected)

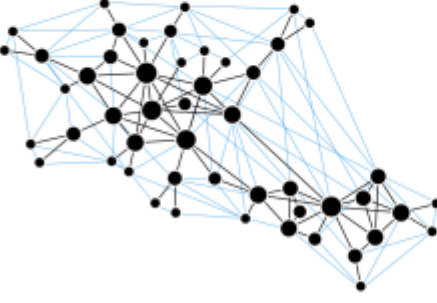
Sparse areas (less connected)

Cliques and strongly connected components

Disconnected components

Center/Periphery



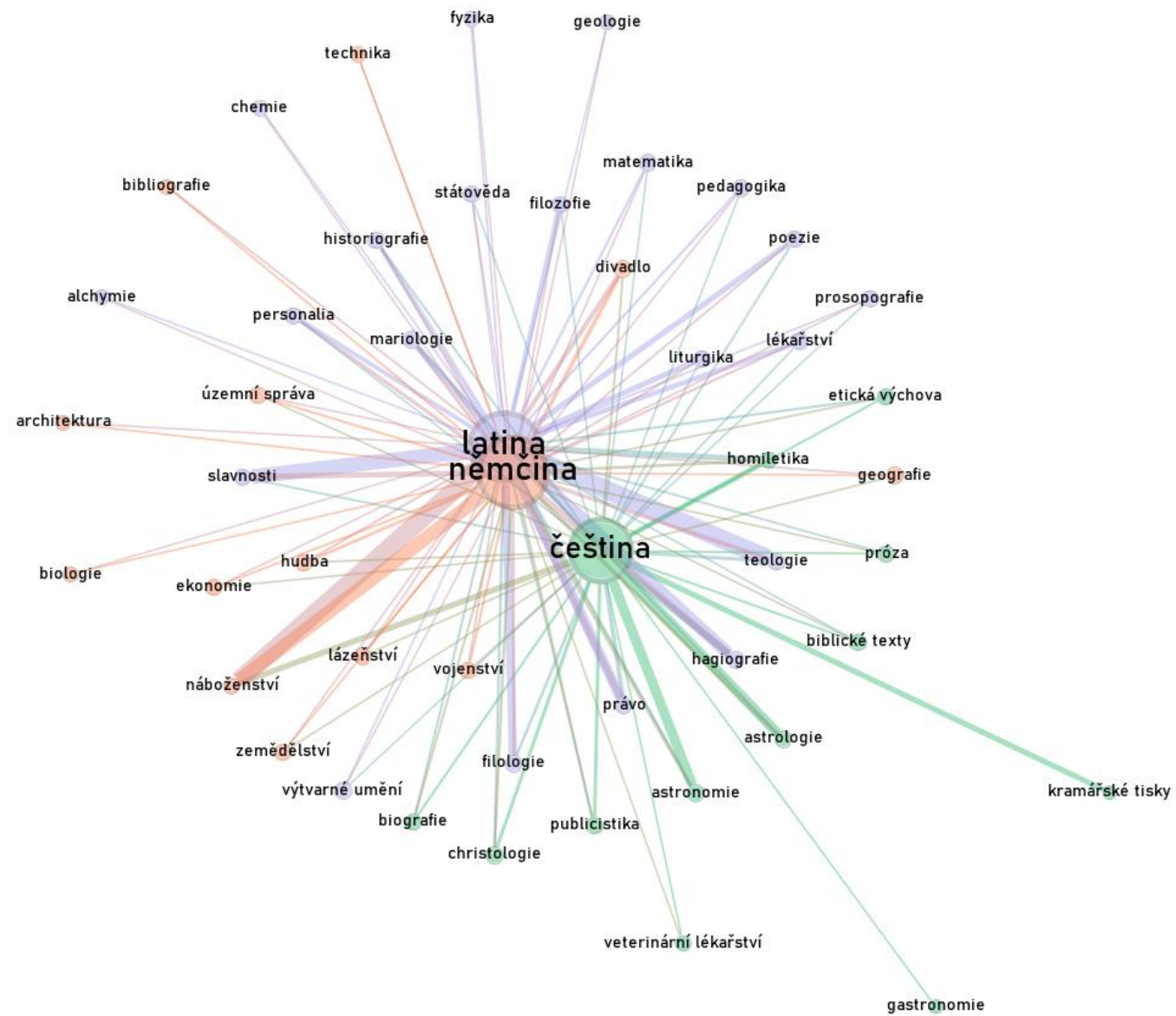
	NOTION	VISUAL ANALYSIS		COMPUTATIONAL ANALYSIS		INTERPRETATIVE POTENTIAL
Global	 Graph size (nodes) SIMPLE DEFINITION How many nodes are there in the graph?	VISUAL PATTERN There are few or many nodes.	ISSUES Though it is easy to have an estimation of the total number of nodes, visualization decisions (for example, setting node sizes on a large scale) can make nodes with few connections difficult to see.	TOPOLOGICAL PATTERN Nodes count.	ISSUES None. Note that in graph theory the count of nodes is referred to as "graph order" while the "graph size" refers to the count of edges.	USEFULNESS Very basic information but useful when comparing networks.
		HERMENEUTIC CRITERIA Counting the nodes.		COMPUTATIONAL CRITERIA Counting the nodes.		"TRANSLATION" The intuition of the size of a network is appropriate.
Global	 Graph size (edges) SIMPLE DEFINITION How many edges are there in the graph?	VISUAL PATTERN There are few or many edges.	ISSUES The total number is hard to estimate as soon as the graph is not a simple diagram anymore. The distribution of edges and their weight has an influence on the visual estimation. The difficulty to count edges visually is a known issue, and probably impossible to overcome.	TOPOLOGICAL PATTERN Edges count.	ISSUES None. Note that in graph theory the count of edges is referred to as "graph size" while the count of nodes is "graph order".	USEFULNESS Very basic information but useful when comparing networks.
		HERMENEUTIC CRITERIA Counting the edges.		COMPUTATIONAL CRITERIA Counting the edges.		"TRANSLATION" Sometimes translated as size in natural language, but the number of edges is usually expressed in comparison to the number of nodes to indicate density or complexity, not for itself.
Global	 Density SIMPLE DEFINITION How connected are the nodes overall?	VISUAL PATTERN The graph is more or less compact. If only certain parts are more compact, see "clusters" below.	ISSUES The less edges there are, the easier to estimate. High densities are difficult to distinguish because the overall appearance of a graph with 60% edges will look close to a graph with 100%. The visual aspect also depends on the layout algorithm used: some are more efficient at representing clusters.	TOPOLOGICAL PATTERN Network density.	ISSUES The formula of density slightly changes depending on the type of networks (directed or not, self-loops allowed or not...).	USEFULNESS A very important notion that allows to compare networks of different sizes if they are produced in the same way or on the basis of comparable data sets.
		HERMENEUTIC CRITERIA Accumulation of edges, cluttered groups of nodes, "hairball". Easier to estimate in situation of comparison.		COMPUTATIONAL CRITERIA Divide the actual number of edges by the number of all potential edges.		"TRANSLATION" Density, complexity, completeness.
		VISUAL PATTERN	ISSUES	TOPOLOGICAL PATTERN	ISSUES	USEFULNESS

Analýza sociálních sítí: další pojmy

česky	anglicky
řízený / neřízený graf	(un)directed layout
modularita	modularity
centralita vlastního vektoru	eigenvector-centrality

Nástroje pro analýzu sociálních sítí

- Gephi >>
 - také webová aplikace Gephi Lite <https://gephi.org/gephi-lite/>
- VOSviewer
 - odkazy k oběma nástrojům – viz prezentace ke cvičení
- knihovny a balíčky pro R a Python
 - knihovny pro R: statnet, igraph, tidygraph, ggraph, ...
 - balíčky pro Python: networkx (+ matplotlib / nxviz / plotly), igraph, graph-tool, pygraphistry, networkit



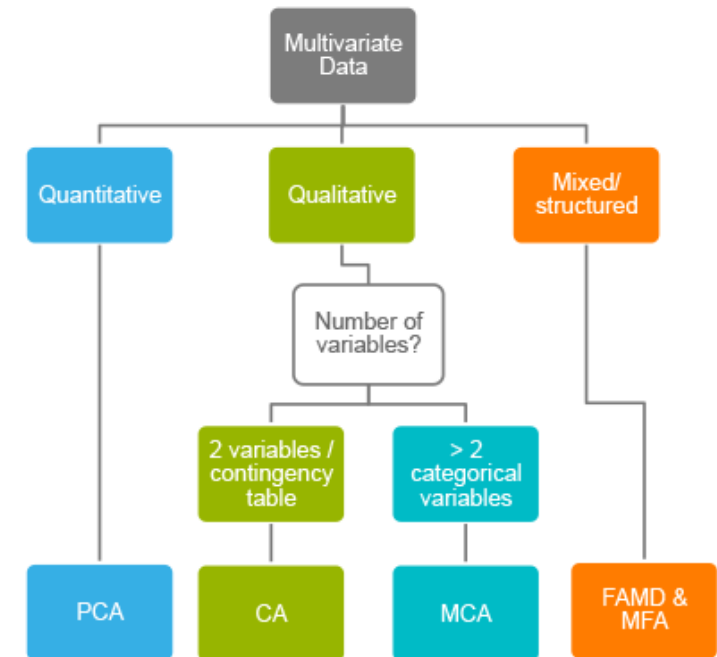
Korespondenční analýza

- statistická metoda pro vizualizaci vztahů mezi kategoriemi v kontingenčních tabulkách
 - složitá vícerozměrná data ve 2D (případně 3D) mapě
- cíl: odhalit latentní struktury v datech

Kontext: metody hlavních komponent

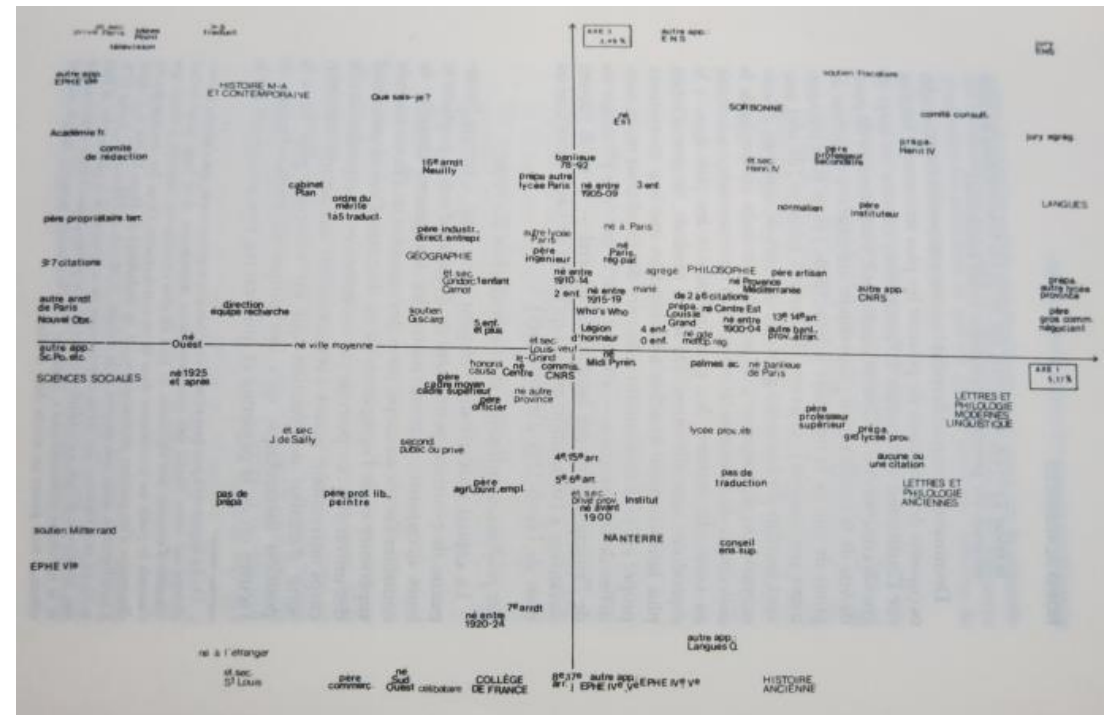
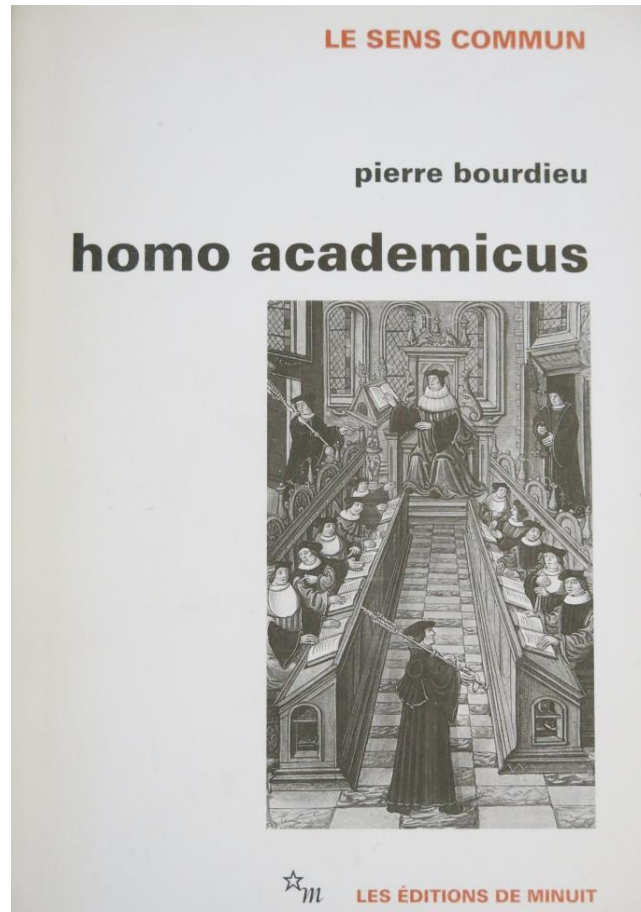
- multivariační data
 - kvantitativní
 - PCA (analýza hlavních komponent)
 - kvalitativní
 - 2 proměnné / kontingenční tabulka → CA (korespondenční analýza)
 - více než 2 kategorické proměnné → MCA (mnohonásobná korespondenční analýza)
 - smíšená / strukturovaná → FAMD & MFA
 - FAMD: faktorová analýza smíšených dat
 - MFA: mnohonásobná faktorová analýza

Principal Component Methods *Methods to Summarize & Visualize Multivariate Data*



- PCA: Principal Component Analysis
- (M) CA: (Multiple) Correspondence Analysis
- FAMD: Factor Analysis of Mixed Data
- MFA: Multiple Factor Analysis

Korespondenční analýza: Pierre Bourdieu

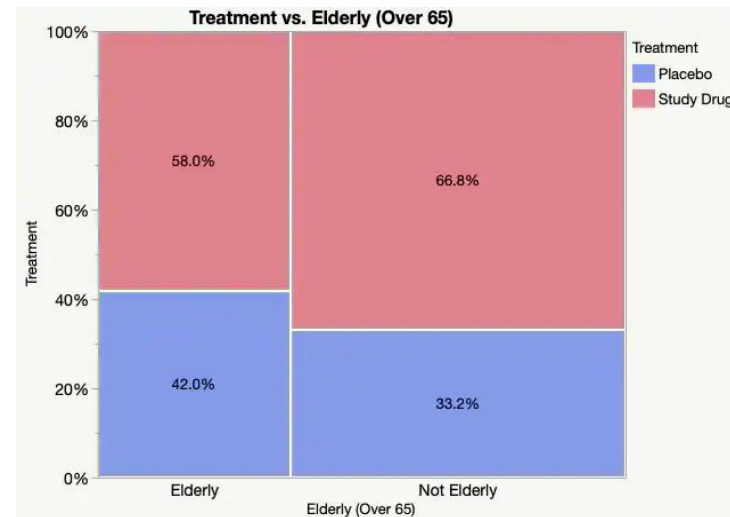
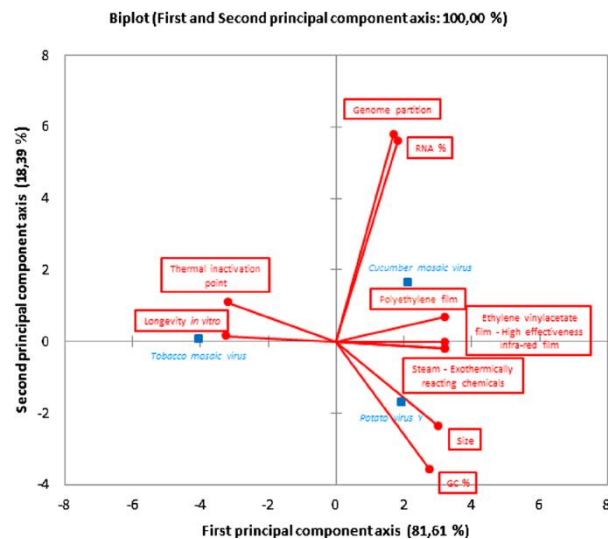


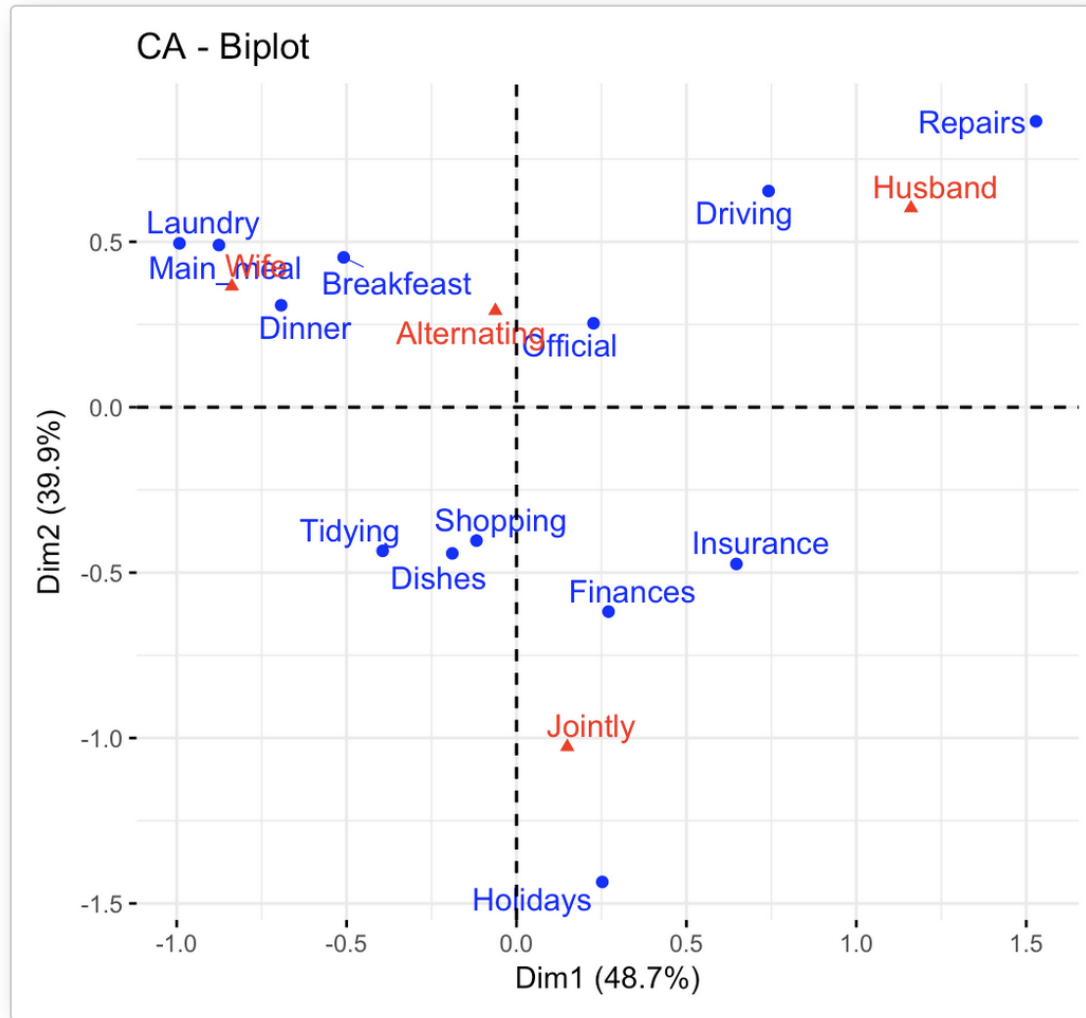
Koresp. analýza: podkladové výpočty

- je to v podstatě vážená singulární dekompozice (SVD) kontingenční tabulky
 - v případě MCA singulární dekompozice Burtovy matice
- postup
 - původní data
 - kontingenční tabulka (příklad: řádky = jazyky, sloupce = témata, hodnoty = počty)
 - výsledky (podklady pro graf)
 - levé singulární vektory (reprezentace řádků – např. jazyků)
 - pravé singulární vektory (reprezentace sloupců – např. témat)
 - singulární hodnoty (síla jednotlivých dimenzí)
 - zpravidla ještě [další] embedding (snížení počtu rozměrů v datech)

Korespondenční analýza: vizualizace

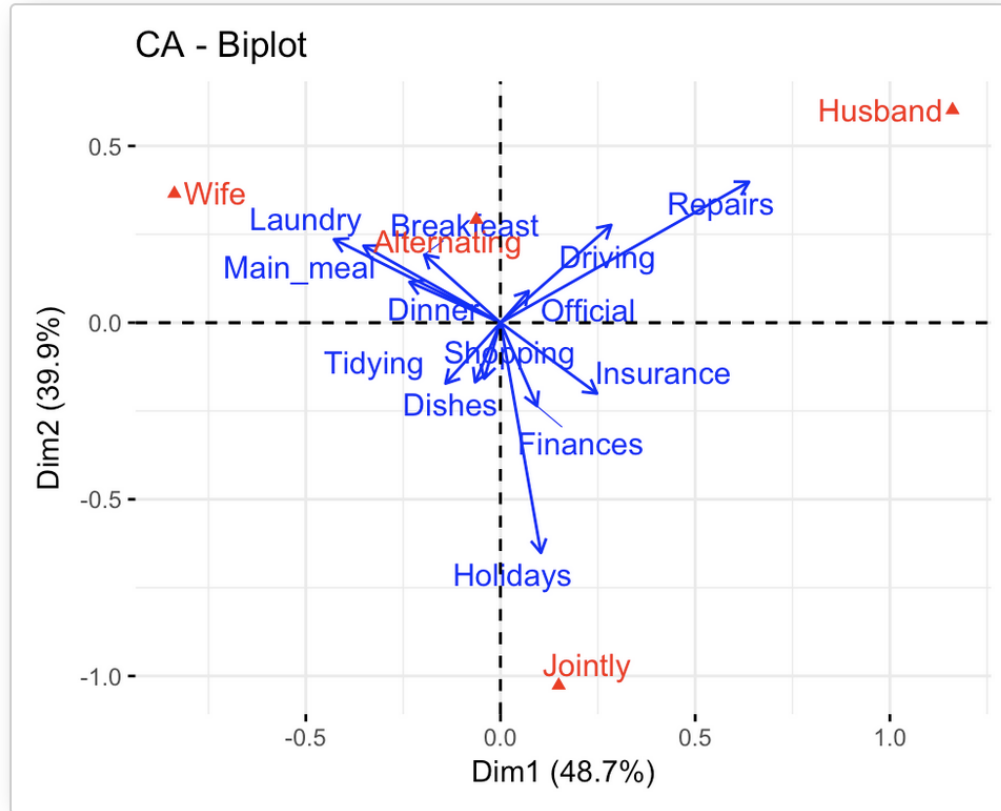
- biplot
 - dvojí koordináty
 - způsoby výpočtu
 - symetrický, asymetrický
- mozaikové zobrazení





The graph above is called *symetric plot* and shows a global pattern within the data. Rows are represented by blue points and columns by red triangles.

The distance between any row points or column points gives a measure of their similarity (or dissimilarity). Row points with similar profile are closed on the factor map. The same holds true for column points.



In the graph above, the position of the column profile points is unchanged relative to that in the conventional biplot. However, the distances of the row points from the plot origin are related to their contributions to the two-dimensional factor map.

The closer an arrow is (in terms of angular distance) to an axis the greater is the contribution of the row category on that axis relative to the other axis. If the arrow is halfway between the two, its row category contributes to the two axes to the same extent.

Mnohonásobná korespondenční analýza: MCA

jméno	automobil	oblíbená barva	mobil
František	Škoda	červená	Samsung
Josef	Škoda	červená	Apple
Karel	Volvo	zelená	Apple
Ludvík	Peugeot	bílá	Motorola

- metoda zpracování:
 - vícecestné tabulky
 - zdvojení (nausea_y, nausea_n)
 - indikátorová / Burtova matice

MCA

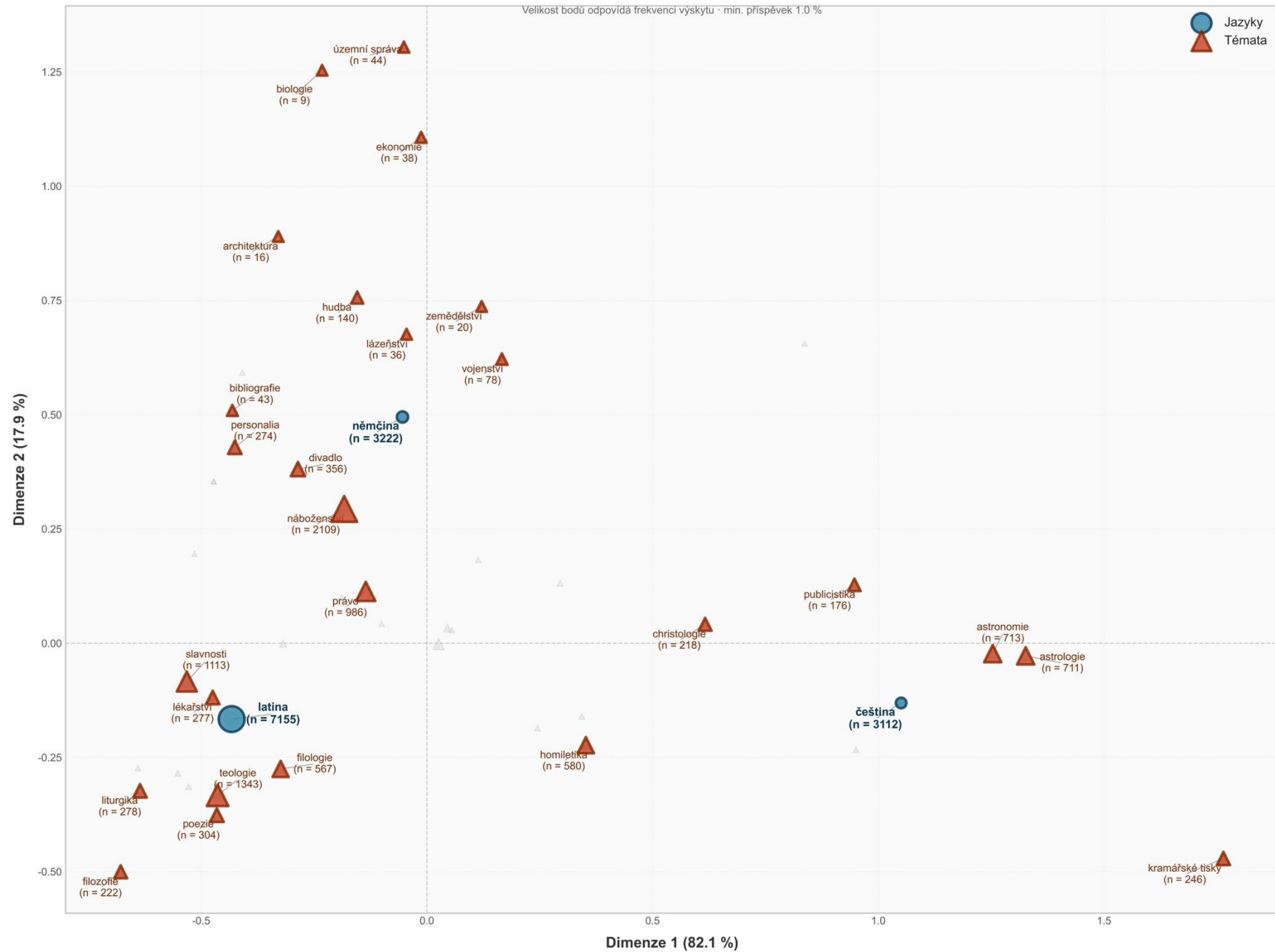
- explorativní vícerozměrná analýza
 - shluková analýza
 - analýza založená na redukci rozměrů na menší počet
- skupiny jedinců s podobným profilem
 - aktivní jedinci
- vztahy mezi kategoriemi proměnných
 - aktivní proměnné
 - doplňkové proměnné: kvantitativní
 - doplňkové proměnné: kvalitativní
- $\cos^2$: kvalita reprezentace
- příspěvek k rozměrům v %

Nástroje pro korespondenční analýzu

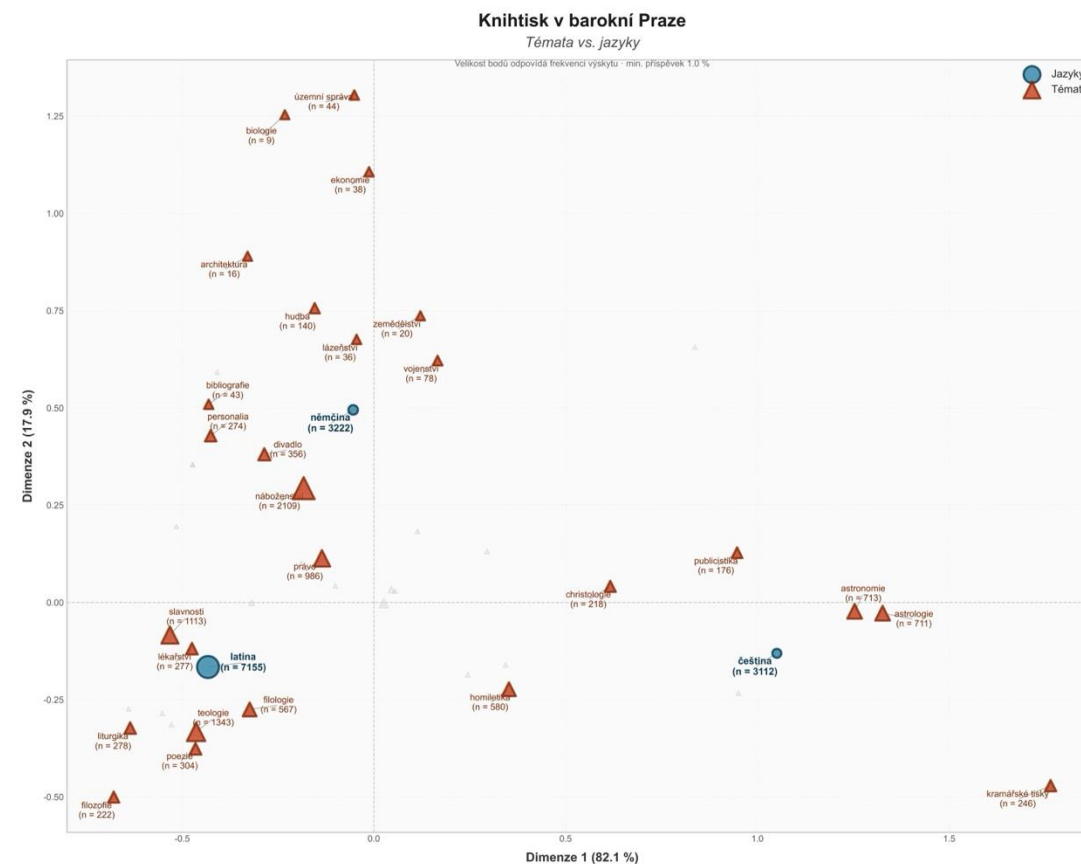
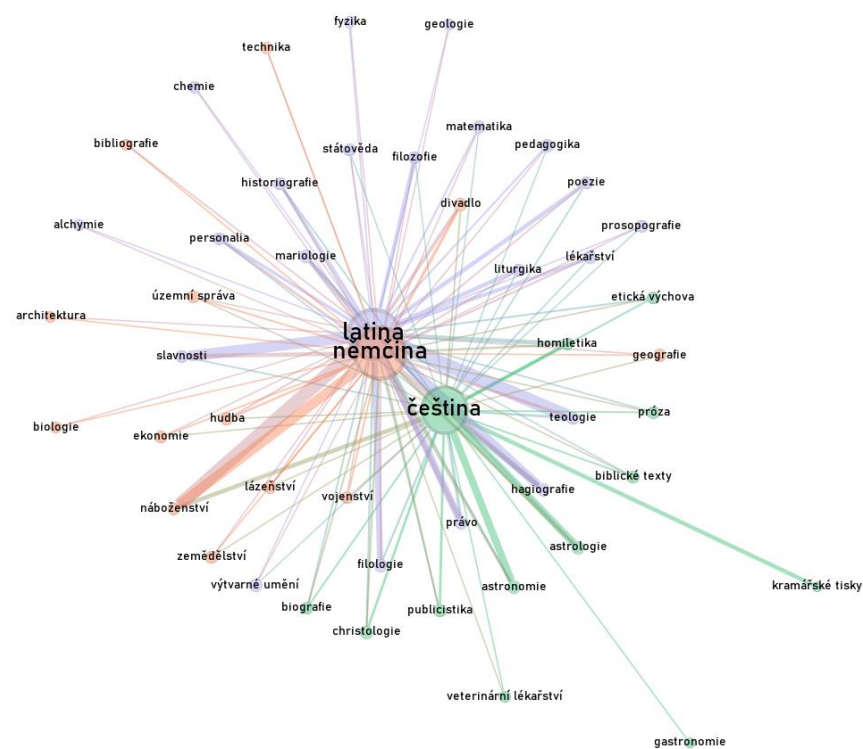
- knihovny a balíčky pro R a Python
 - *vždy jeden nebo uvedená kombinace*
 - knihovny pro R: ca; MASS; FactoMiner + factoextra; GDATools; ...
 - balíčky pro Python: prince >>; scikit-learn; mca; ...
- SPSS, SAS, Stata, MATLAB
 - a další komerční softwary

Knihtisk v barokní Praze

Témata vs. jazyky



Analýza sociálních sítí vs. korespondenční



Analýza sociálních sítí vs. korespondenční

- co mají společného?

„sociogram je založen na stejných datech a byla použita optimalizační technika, která je založena na víceméně stejných principech jako [singulární rozklad](#) v korespondenční analýze“

de Nooy (2003)