# Hierarchická shluková analýza

**Úloha:** Chceme provést shlukovou analýzu, aniž by byl předem dán počet shluků.

**Možné řešení 1:** Provedeme shlukovou analýzu pro různé počty shluků a vybereme „vhodný kompromis“ mezi složitostí modelu (počtem shluků) a dosaženou hodnotou kritéria.
(Existuje řada jednoduchých kritérií, ne příliš teoreticky zdůvodněných.)

**Možné řešení 2:** Vytvoříme strom, ve kterém „blízké objekty“ budou „blízko“, tzv. **dendrogram**.
Při použití *zdola nahoru* nám dovolí najít k objektu libovolný počet jeho „nejbližších sousedů“.
Při použití *shora dolů* nám dovolí rozdělit objekty na libovolný počet shluků.

Existuje řada metod na konstrukci dendrogramu *zdola nahoru* nebo *shora dolů*; zde ukážeme konstrukci založenou na (standardní) fuzzy ekvivalenci.

Vyjdeme z fuzzy relace $R$ mezi $N$ objekty, která popisuje jejich „podobnost“ a je symetrická a reflexivní.

Tento krok vyžaduje znalost uživatele, ale nemusí být obtížné navrhnout vhodné kritérium.

Chtěli bychom, aby fuzzy relace $R$ byla i tranzitivní, ale to už bývá pro uživatele těžké zajistit, ba i ověřit.
Místo toho můžeme použít **tranzitivní uzávěr** fuzzy relace $R$.
Ten v případě **standardní** tranzitivity lze získat následujícími postupy:

**Možné řešení 1:** Pro každý řez fuzzy relace $R$ najdeme tranzitivní uzávěr, čímž dostaneme systém řezů tranzitivního uzávěru fuzzy relace $R$.
Z řezů je jen konečně mnoho různých a všechny jsou ostré relace, takže je to proveditelné klasickými algoritmy.

**Možné řešení 2:** Konstruujeme posloupnost fuzzy relací $R^{(n)}$, $n \in \mathbb{N}$:

$$\begin{aligned} R^{(1)} &= R\,, \\ R^{(n+1)} &= R^{(n)} \fatsemi R\,, \qquad n \in \mathbb{N}\,. \end{aligned}$$

Hodnoty fuzzy relace $R^{(n+1)}$ jsou

$$R^{(n+1)}\_i, j = \max_k \min\{R^{(n)}_{i,k}, R_{k,j}\}\,,$$

kde indexy $i, j, k$ probíhají množinu indexující objekty, $\{1, \ldots, N\}$, což je výpočet podobný násobení matic.

Tato posloupnost matic je neklesající, takže má limitu, navíc ji dosáhne po konečném kroků, což poznáme z podmínky ukončení $R^{(n+1)} = R^{(n)}$.

Jelikož dále je posloupnost konstantní,

$$R^{(n)} \fatsemi R^{(n)} = R^{(2n)} = R^{(n)}\,,$$

takže $R^{(n)}$ je S-tranzitivní.
Je to tranzitivní uzávěr fuzzy relace $R$.

Řezy standardní ekvivalence $R^{(n)}$ jsou ostré ekvivalence, každý z nich definuje rozklad na třídy (shluky), které odpovídají větvím dendrogramu.
Pro menší hodnoty $\alpha$ určuje řez $\mathcal{R}_{R^{(n)}}(\alpha)$ více tříd ekvivalence.
Z toho už snadno sestrojíme dendrogram, ve kterém navíc každé větvení má definovanou hodnotu $\alpha$, při kterém dochází k rozdělení jeho větví. To může být další užitečná informace, kvantifikující nalezenou podobnost.

# Příklad aplikace

Stav infekce Covid-19 v evropských zemích (dle údajů Johns Hopkins University z 5. 5. 2020).

Každá země je charakterizována dvěma souřadnicemi:
1. Počet pozitivních testů na 100 000 obyvatel.
2. Počet úmrtí z pozitivně testovaných v %.

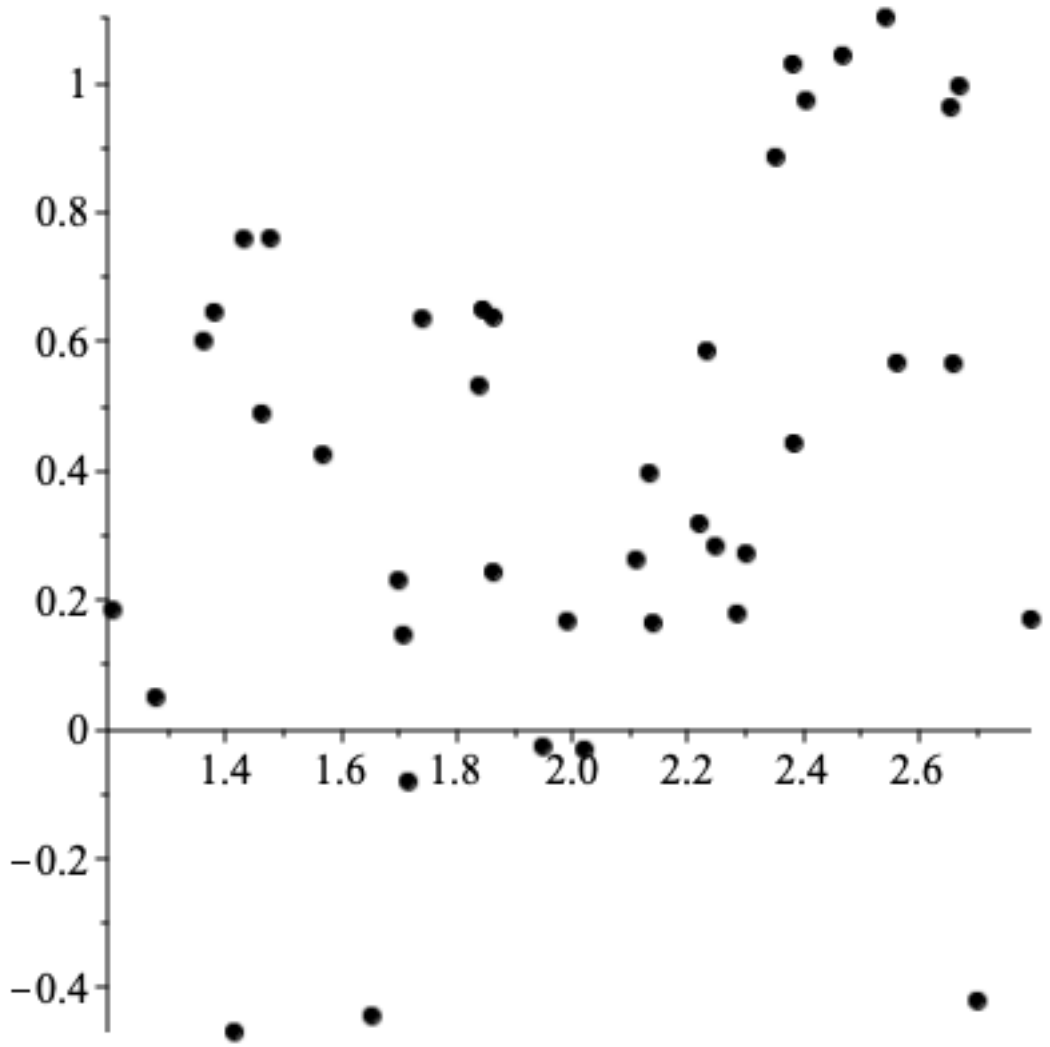
Data byla nejdříve zlogaritmována. (Vynecháno Kosovo, kde nebylo žádné úmrtí.)
Poté byla každá souřadnice normována.
Rozdílnost zemí $i, j$ posuzujeme podle euklidovské vzdálenosti $d_{i,j}$ transformovaných souřadnic.
Na hodnoty z intervalu $\langle 0, 1\rangle$ ji převádíme klesající funkcí

$$R(i,j) = \exp(-d_{i,j}^2/m)\,,$$

kde dělitel byl zvolen $m = 2$, aby typické hodnoty byly vhodně rozloženy na intervalu $\langle 0, 1\rangle$.

Při tomto postupu nevadí, že souřadnice byly v různých měřítkách, na 100 000, resp. na 100 obyvatel.
Nezáleží na základu logaritmu, korekci střední hodnoty, ani na koeficientu $m$, který neovlivní tvar dendrogramu, jen výšku jeho sloupců.
Naopak je důležité, že po normování má každá souřadnice jednotkovou výběrovou směrodatnou odchylku, takže euklidovská norma má smysl.

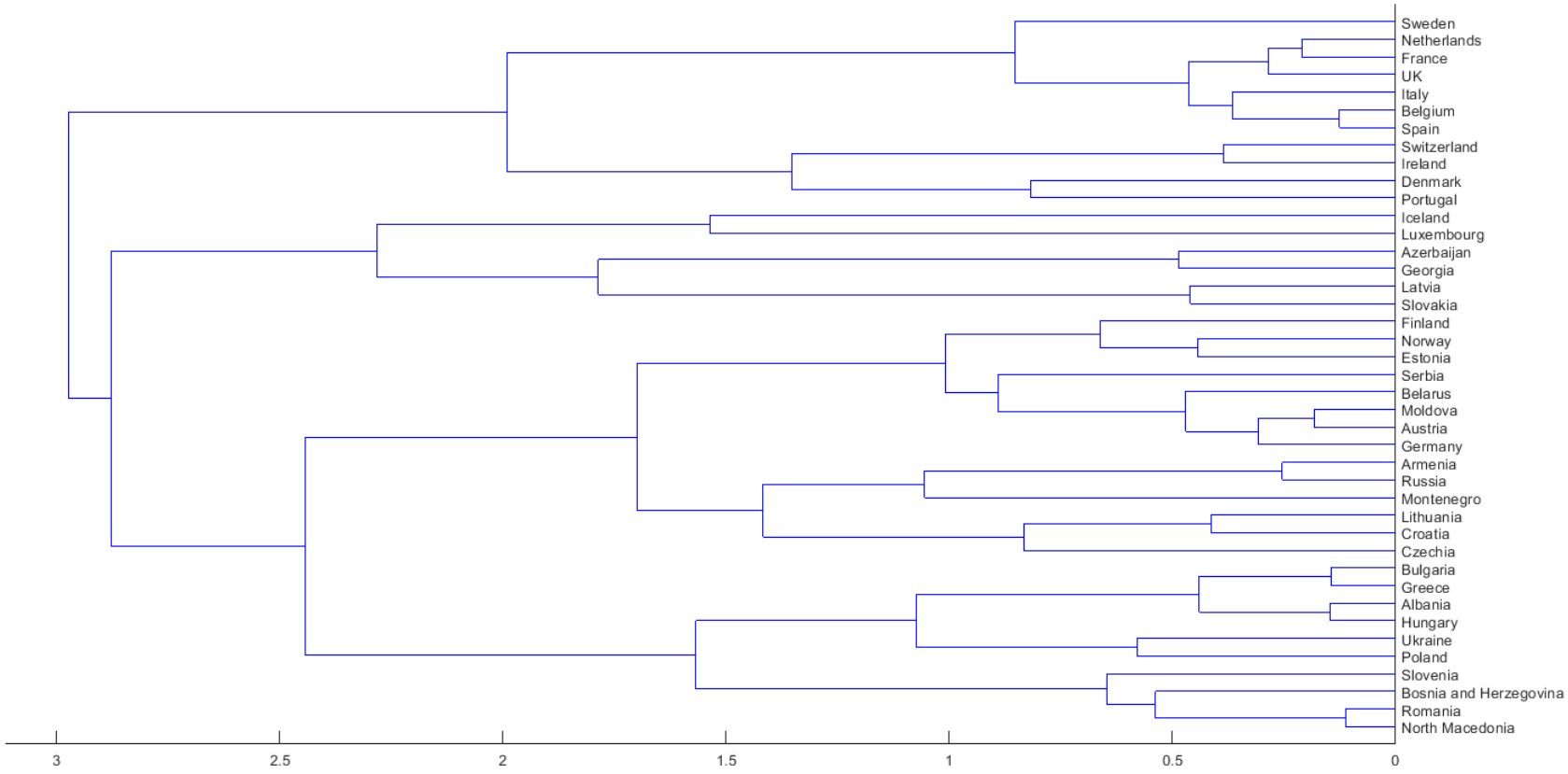
Rozložení vstupních dat (obě souřadnice v dekadických logaritmech):



Vstupní data:

| | | | |
|---|---|---|---|
| 1 | Czechia | 73 | 1.75 |
| 2 | Slovakia | 26 | 0.34 |
| 3 | Poland | 37 | 2.66 |
| 4 | Hungary | 30 | 5.75 |
| 5 | Portugal | 242 | 2.77 |
| 6 | Spain | 468 | 9.89 |
| 7 | Ireland | 457 | 3.68 |
| 8 | UK | 294 | 11.03 |
| 9 | France | 254 | 9.4 |
| 10 | Belgium | 452 | 9.17 |
| 11 | Netherlands | 241 | 10.7 |
| 12 | Luxembourg | 623 | 1.48 |
| 13 | Germany | 200 | 1.87 |
| 14 | Switzerland | 365 | 3.69 |
| 15 | Italy | 349 | 12.63 |
| 16 | Austria | 177 | 1.92 |
| 17 | Slovenia | 69 | 3.4 |
| 18 | Croatia | 51 | 1.4 |
| 19 | Serbia | 136 | 2.49 |
| 20 | Bosnia & Herzeg. | 55 | 4.32 |
| 21 | Montenegro | 52 | 0.83 |
| 22 | North Macedonia | 73 | 4.34 |
| 23 | Albania | 27 | 5.74 |
| 24 | Greece | 24 | 4.42 |
| 25 | Bulgaria | 23 | 3.99 |
| 26 | Romania | 70 | 4.46 |
| 27 | Moldova | 166 | 2.08 |
| 28 | Ukraine | 29 | 3.08 |
| 29 | Belarus | 193 | 1.51 |
| 30 | Russia | 105 | 0.93 |
| 31 | Lithuania | 50 | 1.7 |
| 32 | Latvia | 45 | 0.36 |
| 33 | Estonia | 129 | 1.83 |
| 34 | Denmark | 171 | 3.85 |
| 35 | Iceland | 503 | 0.38 |
| 36 | Norway | 138 | 1.46 |
| 37 | Sweden | 225 | 7.68 |
| 38 | Finland | 98 | 1.47 |
| 39 | Georgia | 16 | 1.53 |
| 40 | Armenia | 89 | 0.94 |
| 41 | Azerbaijan | 19 | 1.12 |

Z tranzitivního uzávěru fuzzy relace $R$ byl vytvořen následující dendrogram:
(změna svislého měřítka je artefakt prostředí Matlab)



| Krajina | Počet nakazených na 100 000 obyvateľov | Percento úmrtí zo všetkých nakazených | Zaradenie pri 7 zhlukoch |
|---|---|---|---|
| Slovakia | 26,00 | 0,34 | 1 |
| Latvia | 45,00 | 0,36 | 1 |
| Georgia | 16,00 | 1,53 | 2 |
| Azerbaijan | 19,00 | 1,12 | 2 |
| Spain | 468,00 | 9,89 | 3 |
| UK | 294,00 | 11,03 | 3 |
| France | 254,00 | 9,40 | 3 |
| Belgium | 452,00 | 9,17 | 3 |
| Netherlands | 241,00 | 10,70 | 3 |
| Italy | 349,00 | 12,63 | 3 |
| Sweden | 225,00 | 7,68 | 3 |
| Portugal | 242,00 | 2,77 | 4 |
| Ireland | 457,00 | 3,68 | 4 |
| Switzerland | 365,00 | 3,69 | 4 |
| Denmark | 171,00 | 3,85 | 4 |
| Luxembourg | 623,00 | 1,48 | 5 |
| Iceland | 503,00 | 0,38 | 5 |
| Poland | 37,00 | 2,66 | 6 |
| Hungary | 30,00 | 5,75 | 6 |
| Slovenia | 69,00 | 3,40 | 6 |
| Bosnia and Herzegovina | 55,00 | 4,32 | 6 |
| North Macedonia | 73,00 | 4,34 | 6 |
| Albania | 27,00 | 5,74 | 6 |
| Greece | 24,00 | 4,42 | 6 |
| Bulgaria | 23,00 | 3,99 | 6 |
| Romania | 70,00 | 4,46 | 6 |
| Ukraine | 29,00 | 3,08 | 6 |
| Czechia | 73,00 | 1,75 | 7 |
| Germany | 200,00 | 1,87 | 7 |
| Austria | 177,00 | 1,92 | 7 |
| Croatia | 51,00 | 1,40 | 7 |
| Serbia | 136,00 | 2,49 | 7 |
| Montenegro | 52,00 | 0,83 | 7 |
| Moldova | 166,00 | 2,08 | 7 |
| Belarus | 193,00 | 1,51 | 7 |
| Russia | 105,00 | 0,93 | 7 |
| Lithuania | 50,00 | 1,70 | 7 |
| Estonia | 129,00 | 1,83 | 7 |
| Norway | 138,00 | 1,46 | 7 |
| Finland | 98,00 | 1,47 | 7 |
| Armenia | 89,00 | 0,94 | 7 |

# Diskuse výsledků

Na zemích, v nichž situaci známe, si můžeme ověřit, že výsledky zhruba odpovídají očekávání.
Blízko sebe se nacházejí země v podobné situaci, s podobným zdravotnickým systémem, historií a často i geografickou polohou (např. země bývalé Jugoslávie).

Snadno bylo možné zahrnout další parametry, např. geografické nebo ekonomické.

Důležité je, že postup, odzkoušený na menších datech, lze posléze aplikovat i na velká data, která už bychom sami těžko zpracovali.

S odstupem času pak umožňují tyto výsledky porovnat vývoj v různých zemích, které byly v dané době v podobné situaci, ale zvolily odlišný postup boje proti šíření infekce. To dovoluje vyhodnotit účinnost přijatých opatření.

# Reference

[1] Michalíková, A.: Fuzzy množiny v informatike. UMB, Banská Bystrica, 2020.

[2] Navara, M., Olšák, P.: Základy fuzzy množin. ČVUT, Praha, 2. vyd., 2007.
http://cmp.felk.cvut.cz/ navara/Zaklady_fuzzy_mnozin/fuzzy-2ed.pdf

Text vypracovali M. Navara a A. Michalíková v květnu 2020.