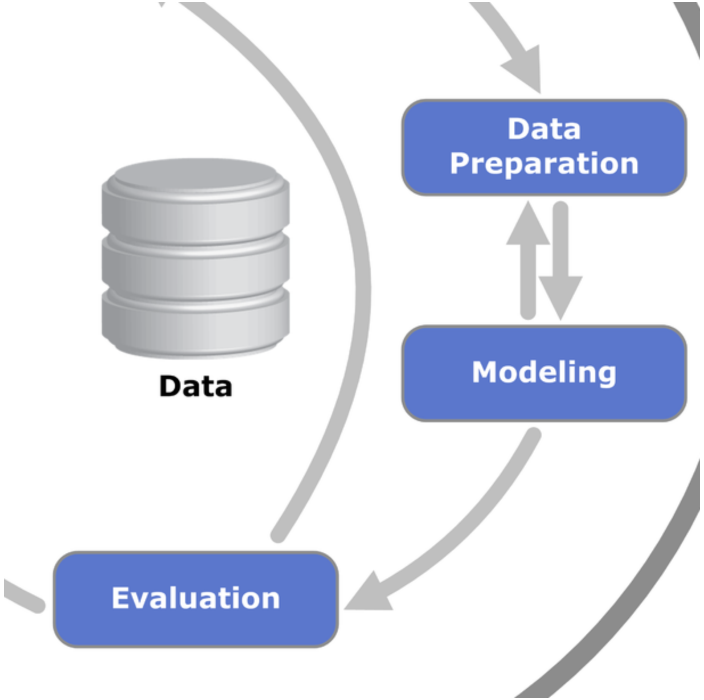
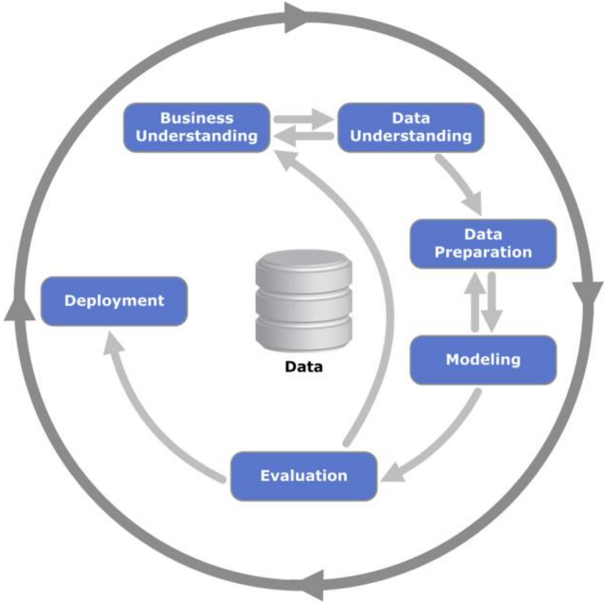
NDBI048 – Data Science
Dimenzionality reduction, clustering

Jan Hučín

20. 11. 2024

Where we are now





Outline

1. appendix to the Modelling lectures
2. multi-dimensional data – blessing and curse
3. how to reduce
4. how to exploit

Appendix to modelling 1&2

Empirical modelling

PROFINIT

- we can describe a „positive behavior“
- but data (almost) not labeled
- fraud, cheat, anomaly detection

Strategy:

- patterns of fraud (domain knowledge) – try to find in data
- what is impossible in fair case
  - playing 24 hours nonstop
  - winning 1st prize five times
  - transporting 100 km in 10 minutes
- what‘s behind weird values in data

Empirical modelling

PROFINIT

Model:

- score every anomality (e. g. X_i-$EX_i > 2\sigma \rightarrow 1$ point)
- add scores
- look at the highest scored cases and consider:
- **Are such cases suspicious?**
- if so:
  - sort cases by the score
  - get labels for cases with top scores (or, even better, different scores)

Good model = detecting well, reasonable/plausible, sustainable

Blessing and curse
of multi-dimensional data

Multi-dimensional data

ClId	APARTMENTS_AVG	BASEMENTAREA_AVG	YEARS_BEGINEXPLUATATION_AVG	YEARS_BUILD_AVG	COMMONAREA_AVG	ELEVATORS_AVG	ENTRANCES_AVG	FLOORSMAX_AVG	FLOORSMIN_AVG	LANDAREA_AVG	LIVINGAPARTMENTS_AVG	LIVINGAREA_AVG	NONLIVINGAPARTMENTS_AVG	NONLIVINGAREA_AVG
1	0,17	0,09	0,99	0,84	NA	0,24	0,10	0,54	0,58		0,14	0,19	0,00	0,00
2	0,06		0,99	0,82	0,01	0,00	0,10	0,17	0,00	0,01	0,05	0,07	0,00	0,00
3	0,06	0,00	0,99	0,80	0,00	0,00	0,14	0,17	0,04	0,00	0,05	0,06	0,00	0,11
4	0,06	0,07	0,99	0,84	0,03	0,00	0,10	0,17	0,21	0,05	0,05	0,07	0,00	0,00
5	0,08	0,04	0,98	0,75	0,02	0,00	0,03	0,17	0,21	0,01	0,06	0,04	0,01	0,06
6	0,06	0,07	0,99	0,86	0,01	0,00	0,14	0,17	0,21	0,02	0,05	0,05	0,00	0,01
7	0,13	0,09	0,98	0,76	0,00	0,00	0,03	0,17	0,21	0,10	0,10	0,05	0,00	0,01
8	0,02	0,02	0,99	0,81	0,00	0,00	0,07	0,17	0,21	0,02	0,02	0,02	0,00	0,00
9	0,02	0,00	0,97	0,63	0,00	0,00	0,07	0,04	0,08	0,01	0,01	0,01	0,00	0,00
10	0,33	0,23	0,98	0,79	0,03	0,36	0,31	0,33	0,04	0,29	0,27	0,33	0,00	0,00

Id	Q1	Q2	Q3	Q4	Q5	Q6
14056	agree	agree	agree	always	mostly	always
14057	disagree	not sure	disagree	mostly	never	mostly
14058	not sure	disagree	not sure	always	sometime	rarely
14059	agree	not sure	agree	always	never	mostly
14060	agree	agree	agree	mostly	always	mostly
14061	not sure	disagree	not sure	rarely	always	always
14062	disagree	not sure	disagree	mostly	mostly	always
14063	not sure	agree	not sure	never	always	always
14064	agree	not sure	agree	sometime	always	never
14065	disagree	disagree	disagree	never	mostly	sometime
14066	not sure	not sure	not sure	always	always	never

Multi-dimensional data

PROFINIT

many information in great detail, but…

- too much to use all
- highly correlated (multicollinearity)
- hard to find the real structure
- hard to find errors
- in E^K, data is usually sparse

Id	Q1	Q2	Q3	Q4	Q5	Q6
14056	agree	agree	agree	always	mostly	always
14057	disagree	not sure	disagree	mostly	never	mostly
14058	not sure	disagree	not sure	always	sometime	rarely
14059	agree	not sure	agree	always	never	mostly
14060	agree	agree	agree	mostly	always	mostly
14061	not sure	disagree	not sure	rarely	always	always
14062	disagree	not sure	disagree	mostly	mostly	always
14063	not sure	agree	not sure	never	always	always
14064	agree	not sure	agree	sometime	always	never
14065	disagree	disagree	disagree	never	mostly	sometime
14066	not sure	not sure	not sure	always	always	never

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

Reducing the dimensionality

compression of information:

- to discover main factors
- to make features uncorrelated
- to understand the world structure
- to find anomalities

Id	Q1	Q2	Q3	Q4	Q5	Q6
14056	agree	agree	agree	always	mostly	always
14057	disagree	not sure	disagree	mostly	never	mostly
14058	not sure	disagree	not sure	always	sometime	rarely
14059	agree	not sure	agree	always	never	mostly
14060	agree	agree	agree	mostly	always	mostly
14061	not sure	disagree	not sure	rarely	always	always
14062	disagree	not sure	disagree	mostly	mostly	always
14063	not sure	agree	not sure	never	always	always
14064	agree	not sure	agree	sometime	always	never
14065	disagree	disagree	disagree	never	mostly	sometime
14066	not sure	not sure	not sure	always	always	never

Methods
of dimensionality reduction

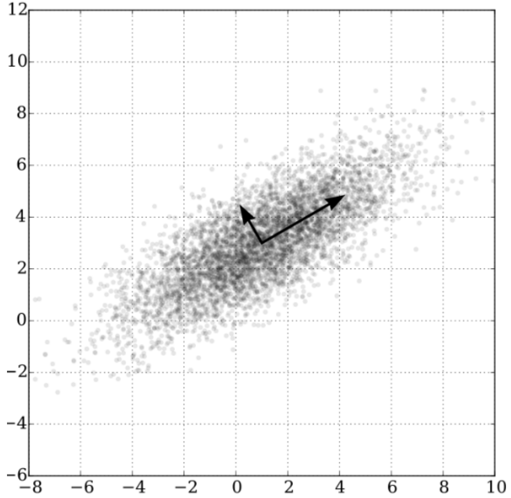
Dimensionality reduction – scoring

PROFINIT

information compression

linear:

- PCA (principal component analysis)
  - $\boldsymbol{U} = \boldsymbol{XP}$
  - $\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots \boldsymbol{u_K}$ uncorrelated, variance of $\boldsymbol{u_1}$ maximized
- factor analysis
  - linear projection of $\boldsymbol{X}$ to lower dim space (*latent factors*)
  - projection matrix = explanatory



nonlinear:

- IRT (item-response theory)
  - columns of $\boldsymbol{X}$ binary or categorical
  - estimation of „ability score“
- submodel (supervised) → score

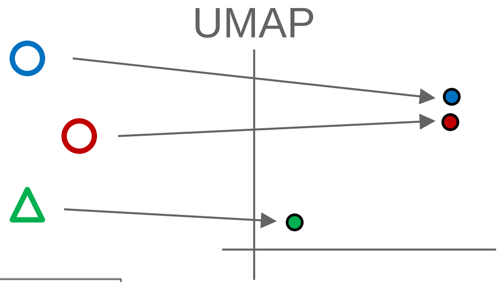
Id	It 1	It 2	It 3	It 4	It 5	It 6	It 7	It 8	Score
1	✓	✗	✓	✓	✗	✓	✗	✓	0,7
2	✓	✗	✗	✗	✓	✓	✗	✗	-0,8
3	✓	✓	✓	✓	✓	✓	✗	✓	1,8
4	✓	✗	✗	✗	✓	✗	✓	✗	-0,4
5	✗	✗	✓	✓	✗	✓	✗	✓	-0,1

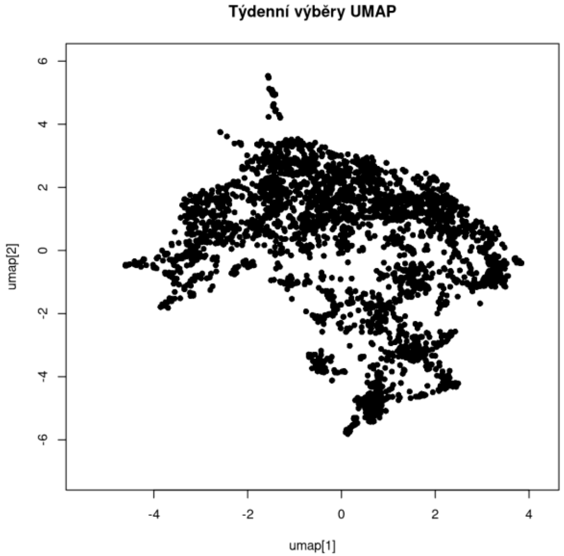
Dimensionality reduction – similarity

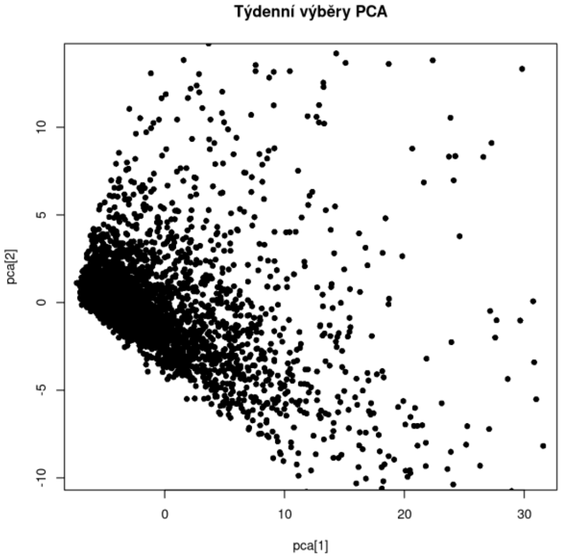
PROFINIT

redistribution of points

- linear methods (PCA) possible but often useless
- nonlinear: long distances are unimportant
  - t-sne
  - UMAP

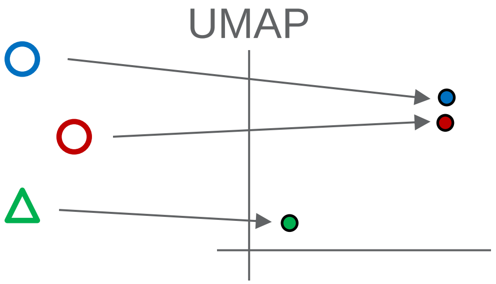


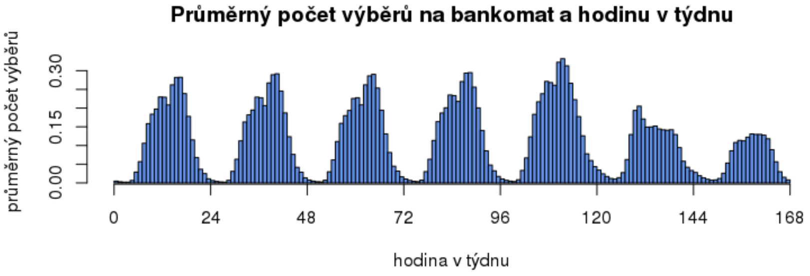


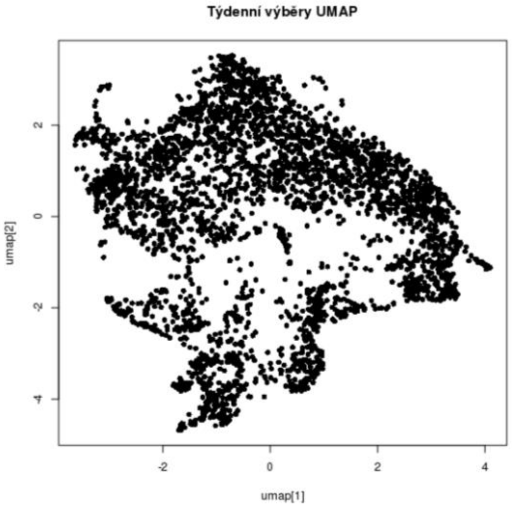


Bankomaty

- UMAP*
  - Projekce vektoru na varietu při zachování lokálního měřítka
  - Něco jako PCA na steroidech



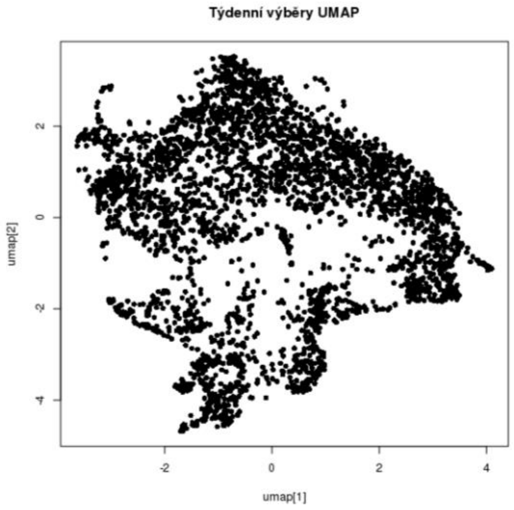


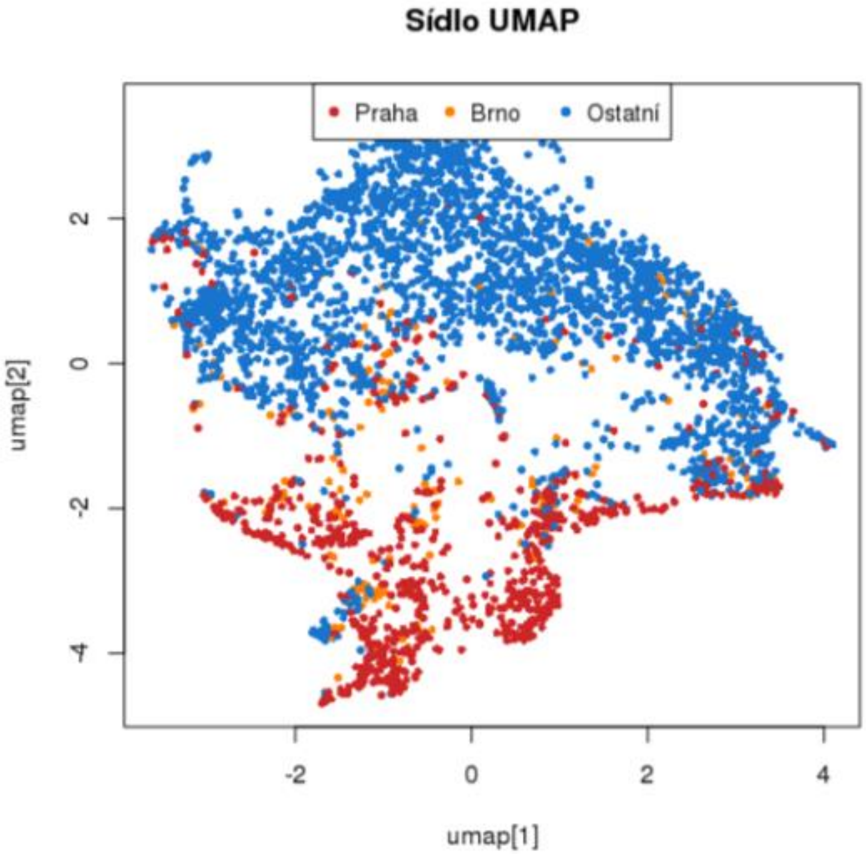


*) McInnes, L, Healy, J, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, 2018

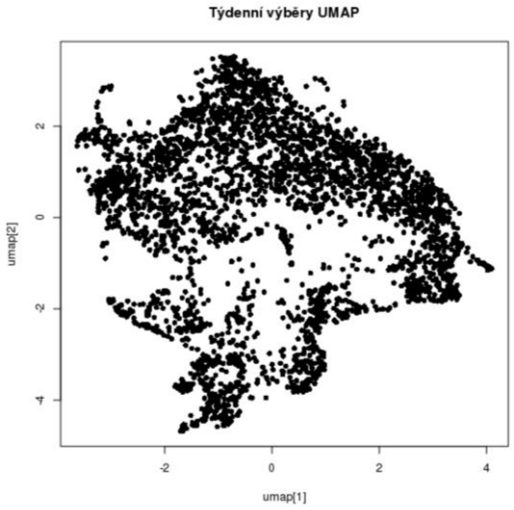
PROFINIT

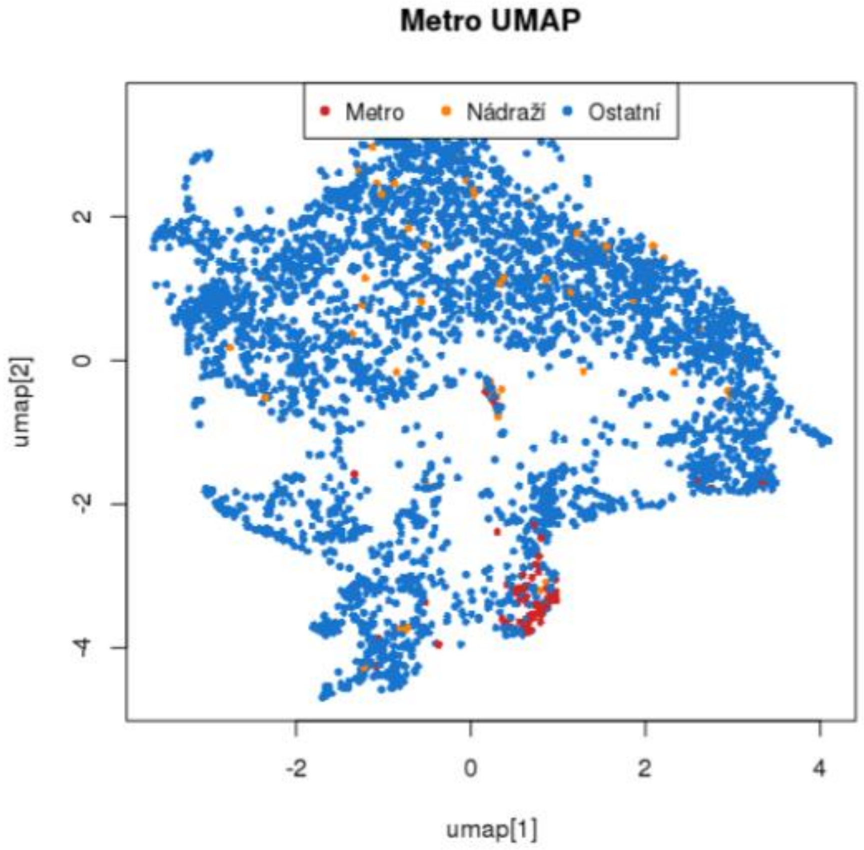
Transakční data – bankomaty



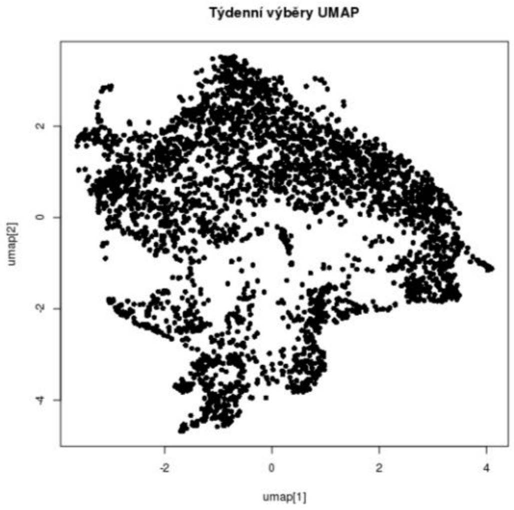


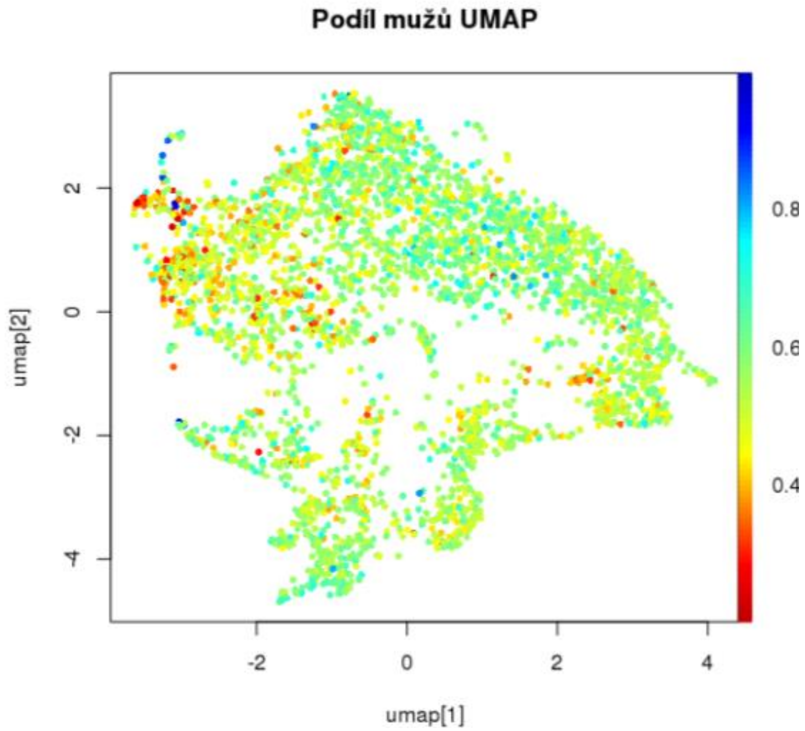
Transakční data – bankomaty



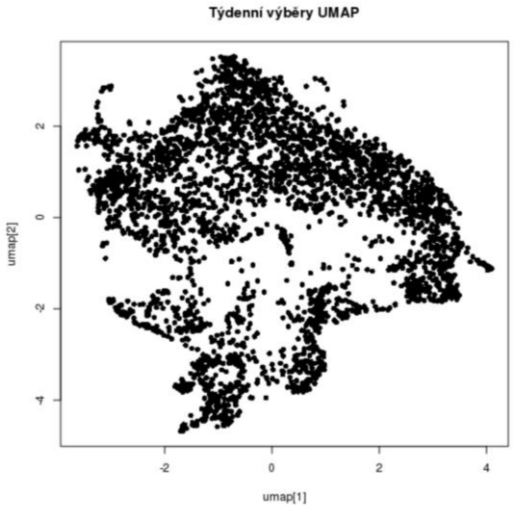


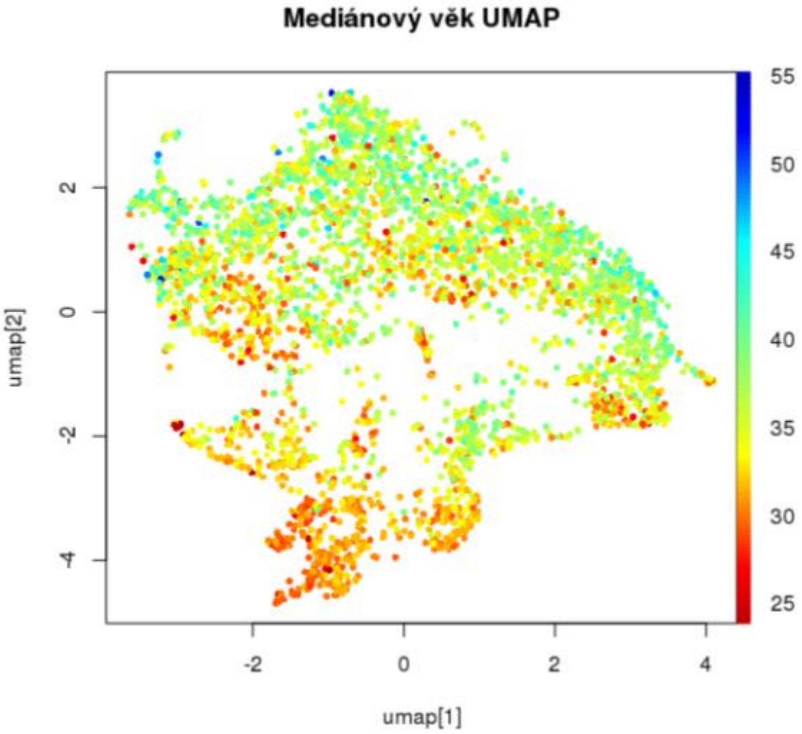
Transakční data – bankomaty





Transakční data – bankomaty





Pitfalls

PROFINIT

- numerical vs. categorical variables
- incomparable scales
- skewed distributions
- bad distance metrics

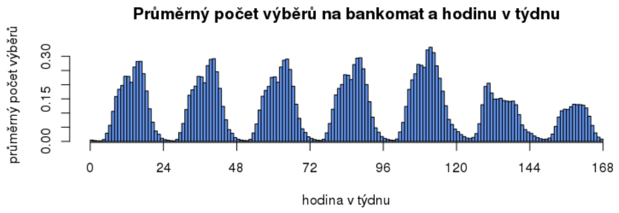
Use
of dimensionality reduction

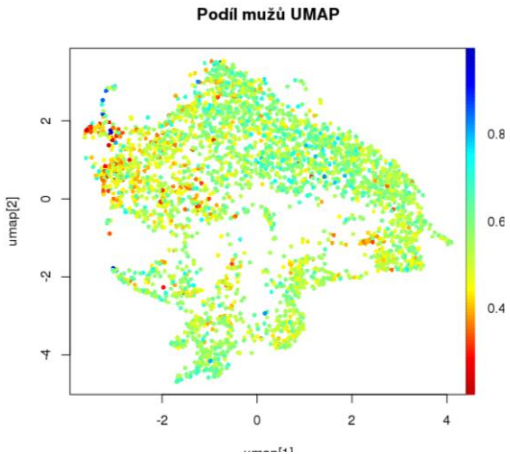
How to use the result of dim reduction

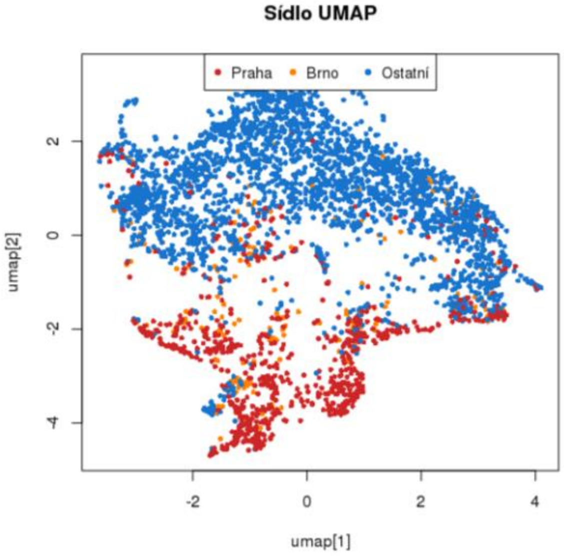
PROFINIT

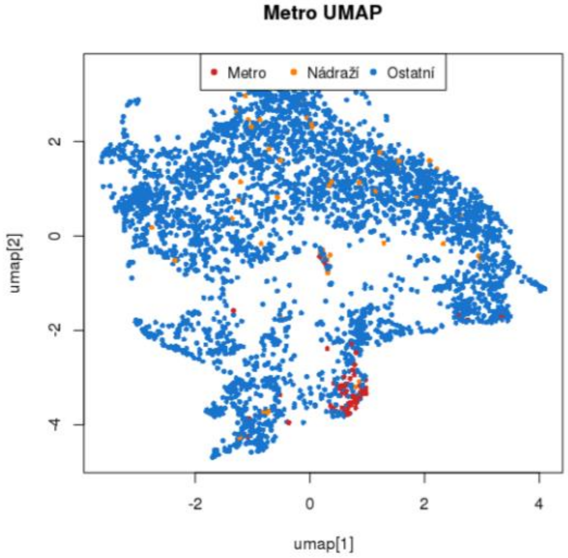
- score itself
- feature in model (risk, ability, IQ etc.)
- „world“ description (see next example)
- clustering

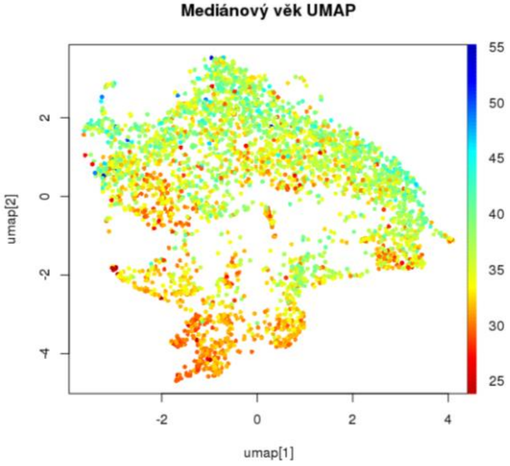
Transakční data – bankomaty



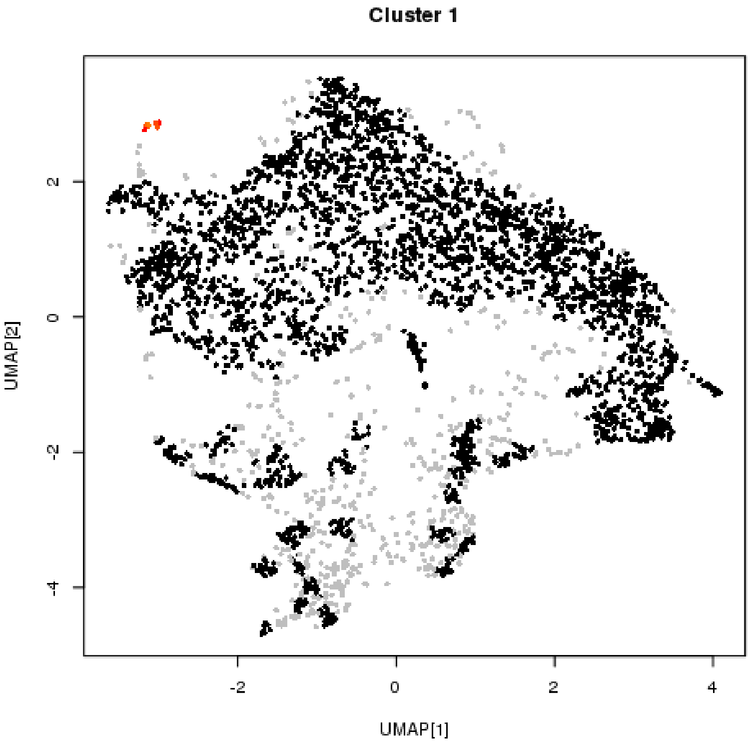


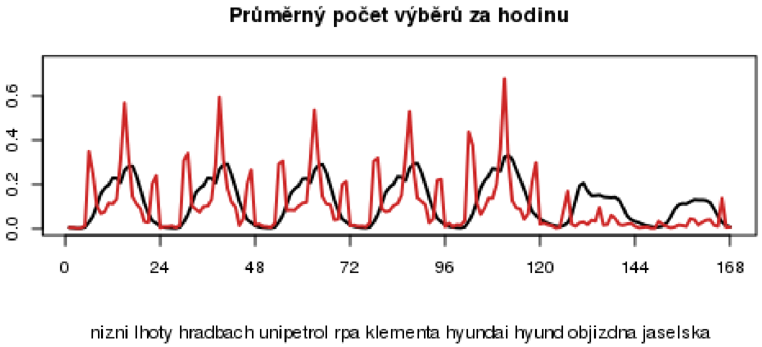






bankomaty – fabriky s 3 směnným provozem

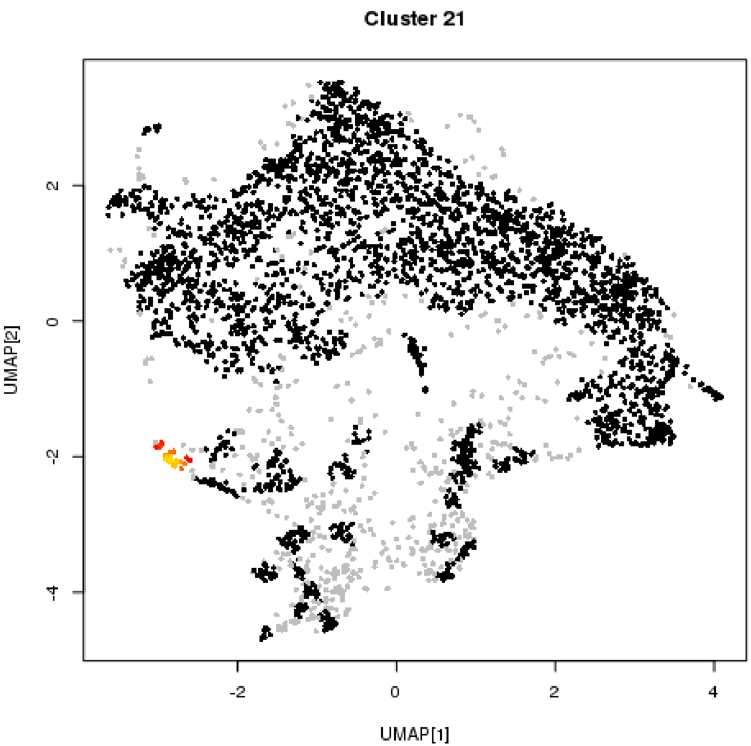


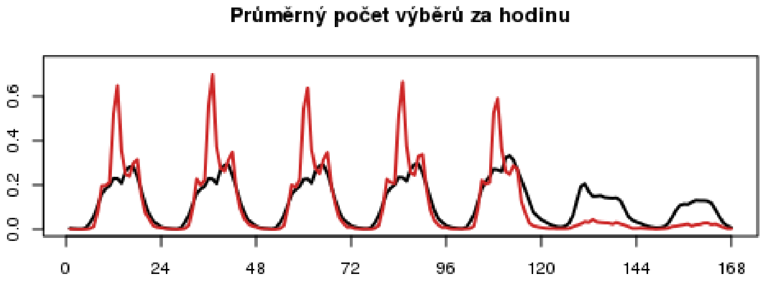


CZ::Plzen::CSOB 2052 PLZEN
CZ::KOLIN::KB ATM NA HRADBACH
CZ::NIZNI LHOTY::CS, NIZNI LHOTY, HYUND
CZ::PRELOUC::CS, JASELSKA 593
CZ::NIZNI LHOTY::CS, HYUNDAI 700/1
CZ::MLADA BOLESLA::CS, VACLAVA KLEMENTA
CZ::LITVINOV::CS, UNIPETROL RPA
CZ::OTROKOVICE::KB ATM OBJIZDNA 1628
CZ::MLADA BOLESLA::CS,VACL.KLEMENTA,SKODA
CZ::Mlada Bolesla::CSOB 1210 ML BOLESLAV

PROFINIT

bankomaty – kravaťáci jdou na oběd



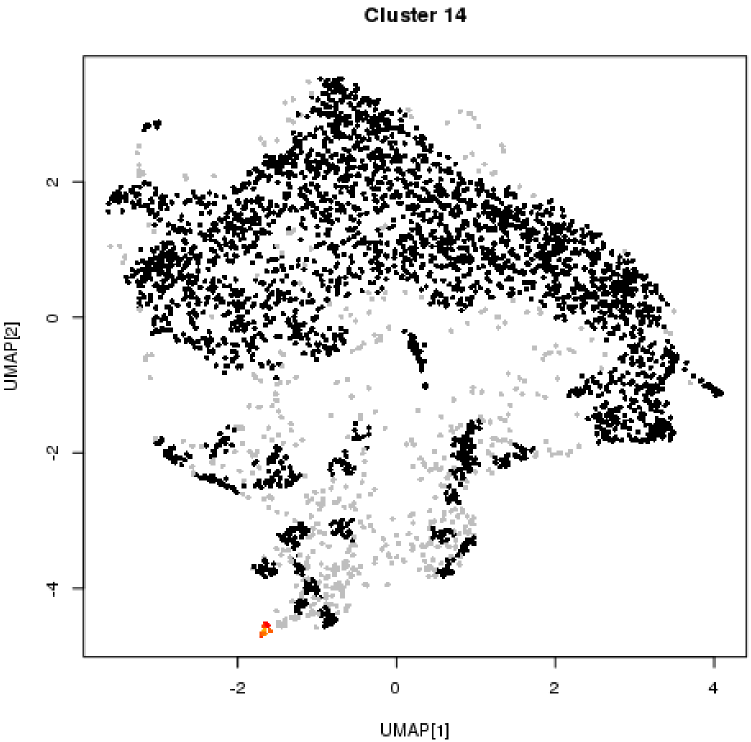


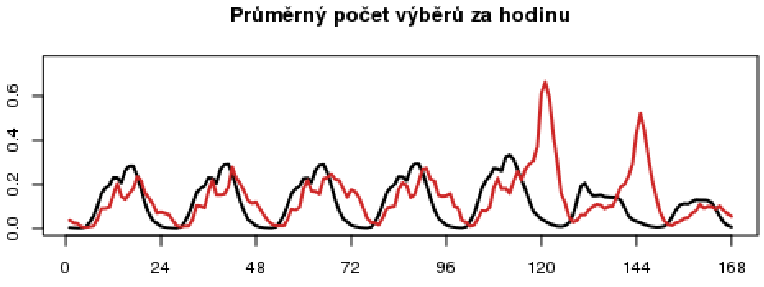
ɔankraci holandska kolejni her sberbanks1as1306 mps churchilla hrebenech parku kavc

CZ::PRAHA 4::CS. NA HREBENECH II 11
CZ::PRAHA 8::CS. POBREZNI 665/21
CZ::Ceske Budejov::CSOB 2118 C BUDEJOVIC
CZ::PRAHA 4::CS. NA PANKRACI 1683/1
CZ::Praha::RB Hvezdova
CZ::PRAHA 2 - MPS::UNICREDIT - PRAHA 2
CZ::Praha 1::CSOB 1820 PRAHA 1
CZ::Praha 4::MMB NA PANKRACI 1724/1
CZ::Brno-Kralovo::CSOB 0882 BRNO
CZ::Praha 4::RB CT Kavci hory
CZ::Praha::RB Prosek Point
CZ::Brno::SBERBANK 7013 HER.
CZ::Brno::CSOB 4029 BRNO
CZ::Praha 8::CSOB 1989 PRAHA 8
CZ::PRAHA 6::CS. NAR.TECH.KNIHOVNA

PROFINIT

Bankomaty – páteční noc





ceb prague cross club s6cy0010 vin koblizna oberbank dlouha sberbank

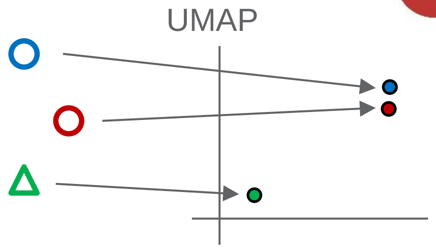
CZ::Brno::CZE00421
CZ::Ostrava::CZE00236
CZ::PRAHA 1::CZE00487
CZ::PRAHA 1::CS. DLOUHA 743/9
CZ::Praha 2::SBERBANK 7020 VIN.
CZ::Praha 1::CSOB 1216 PRAHA 1
CZ::Praha 2::CZE00183
CZ::PRAHA 1::KB ATM DLOUHA 34
CZ::Ceske Budejov::OBERBANK 7507 CEB
CZ::BRNO::KB ATM KOBLIZNA 3
CZ::Praha 1::SBERBANK 7007 P1-VN
CZ::PRAHA 1::ATM S6CY0010
CZ::Prague::ATM Cross club

Redukce dimenze

PROFINIT

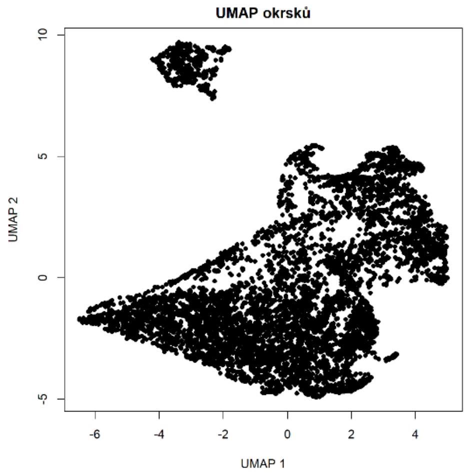
› UMAP*

– Projekce vektoru na varietu při zachování lokálního měřítka
– Něco jako PCA na steroidech



Volební okrsek

Velikost sídla, kraj
Průměrný věk, podíl mužů
Nejvyšší dosažené vzdělání
Stáří obyvatel, religiozita
Nezaměstnanost, rozvodovost, …

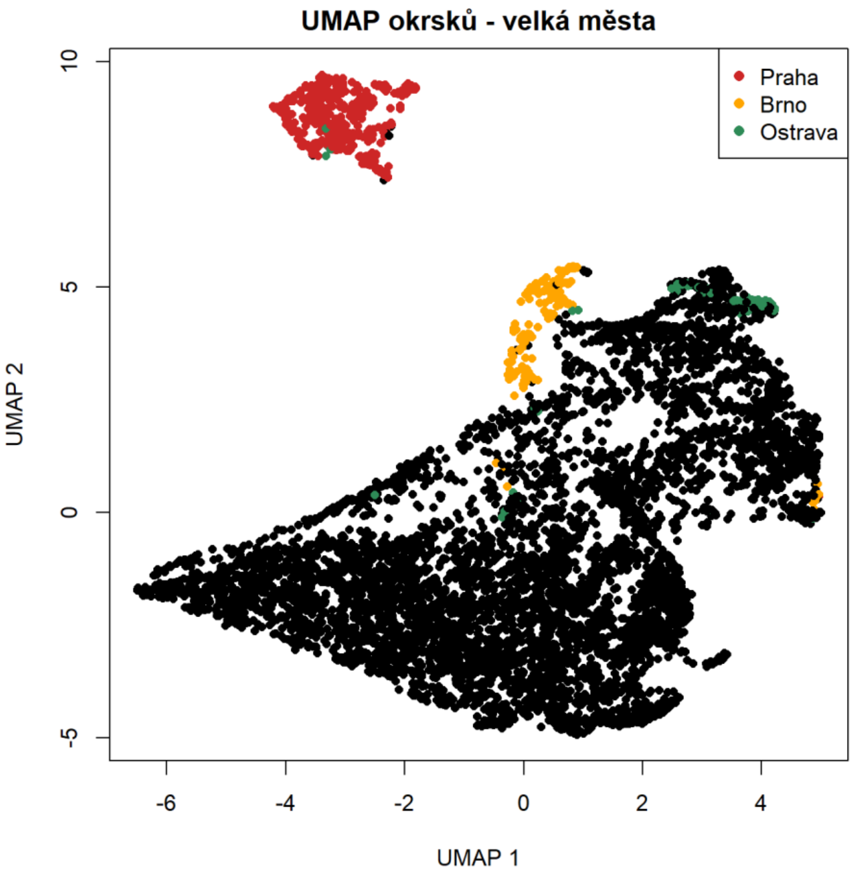


*) McInnes, L, Healy, J, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, 2018

Velká města

PROFINIT

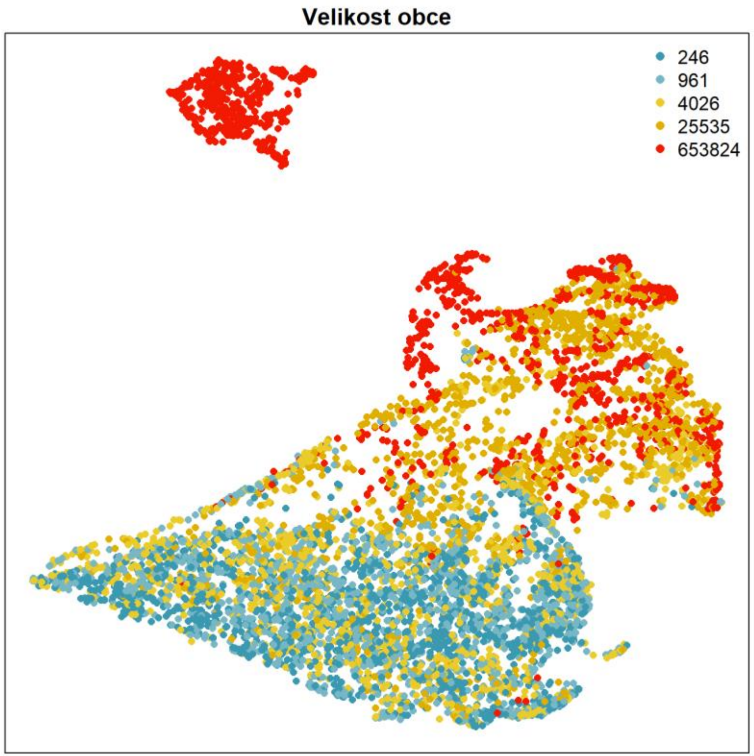
- Praha je hodně jiná
- Brno je skoro Praha
  - ale vlastně ne



Velikost obce

PROFINIT

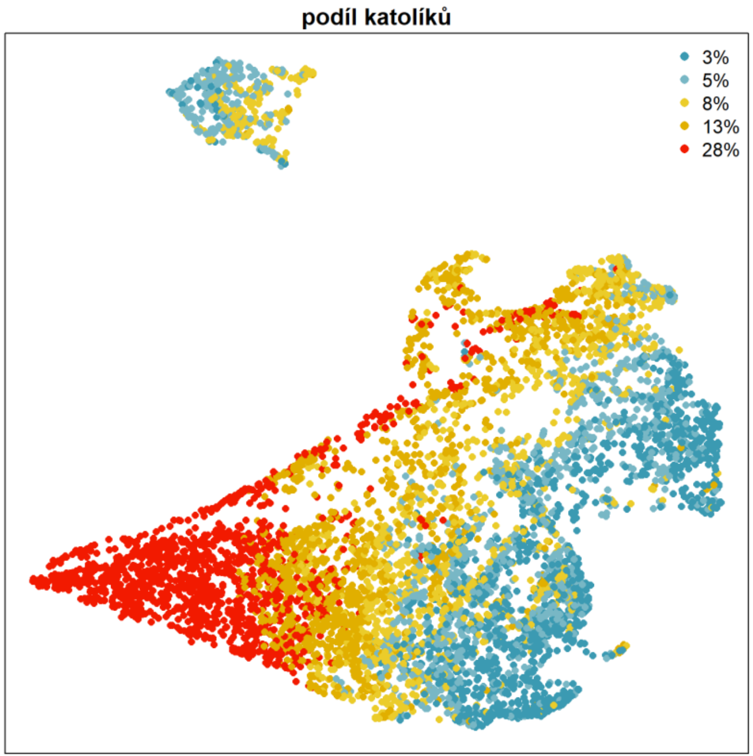
- Hlavní dělící linie
  - víceméně osa Y



Věřící

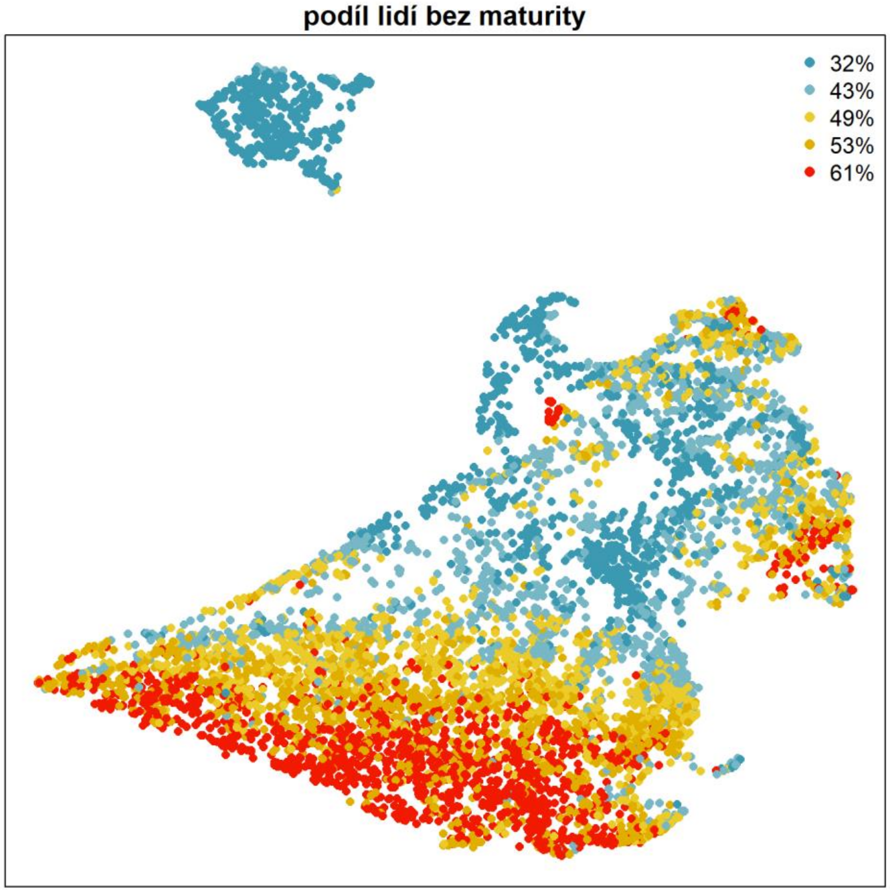
PROFINIT

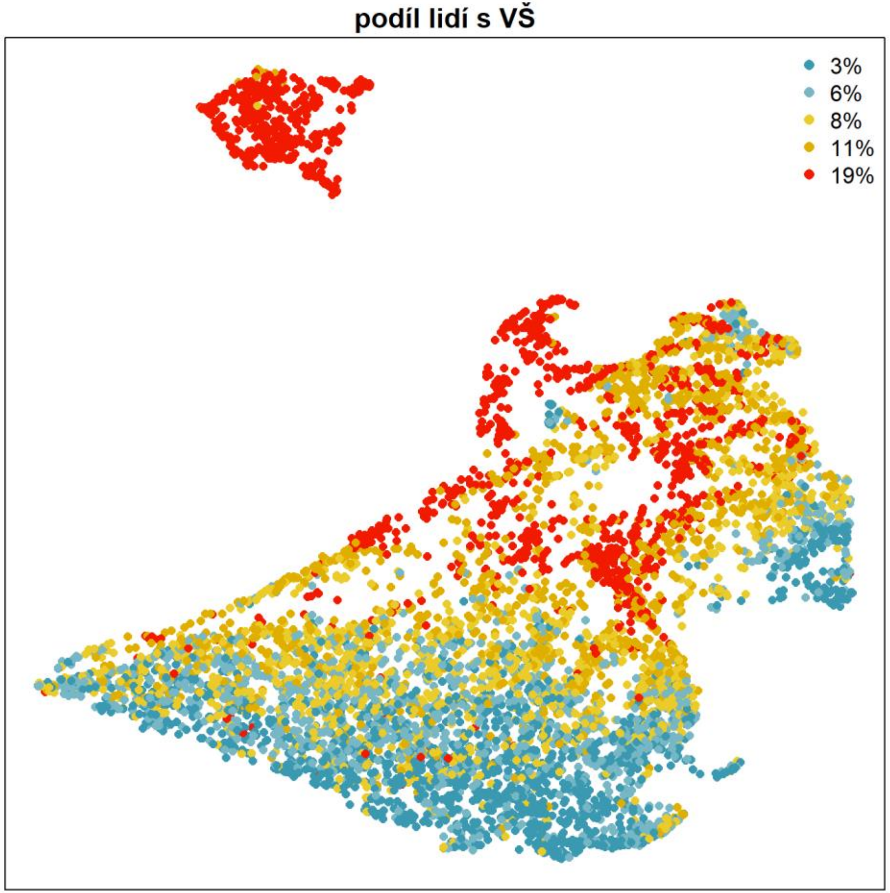
- Hlavní dělící linie
  - víceméně osa X



Vzdělání

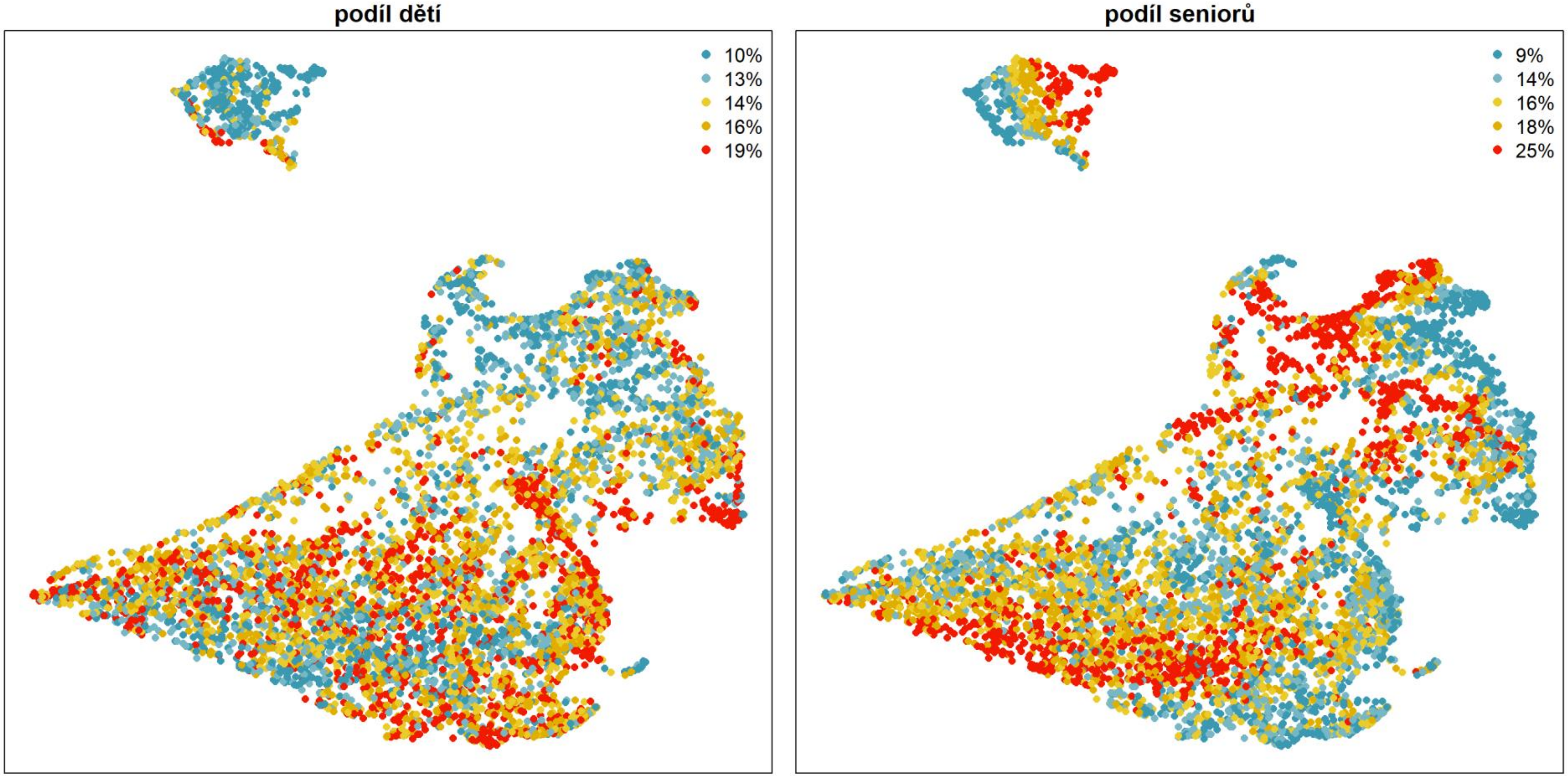
PROFINIT





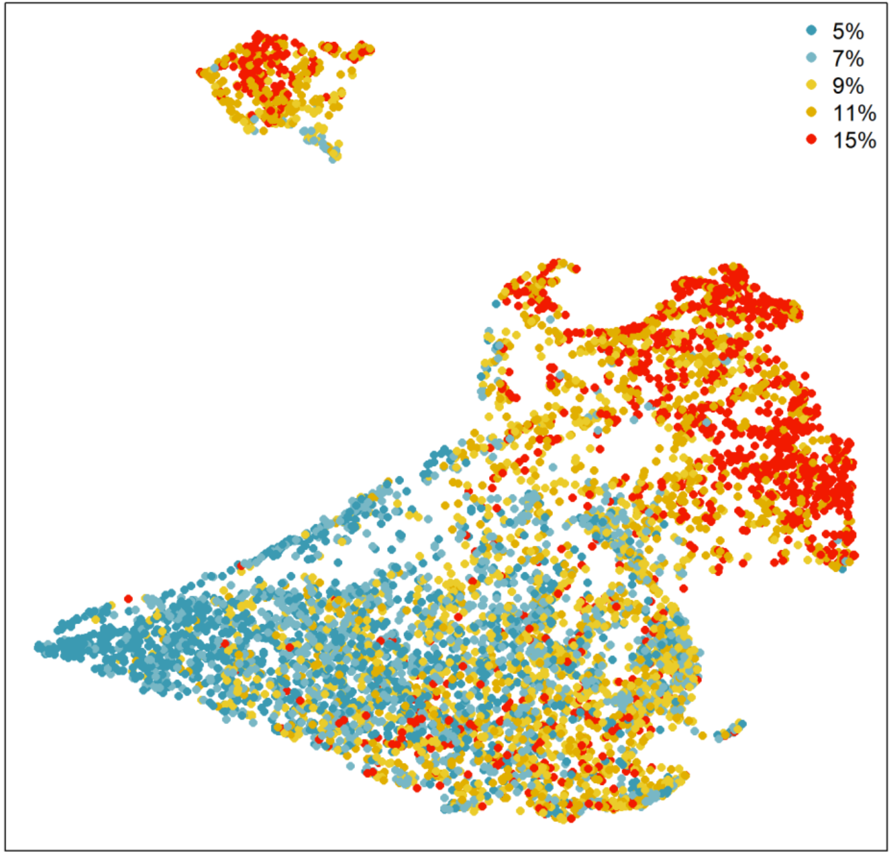
Děti a senioři

PROFINIT

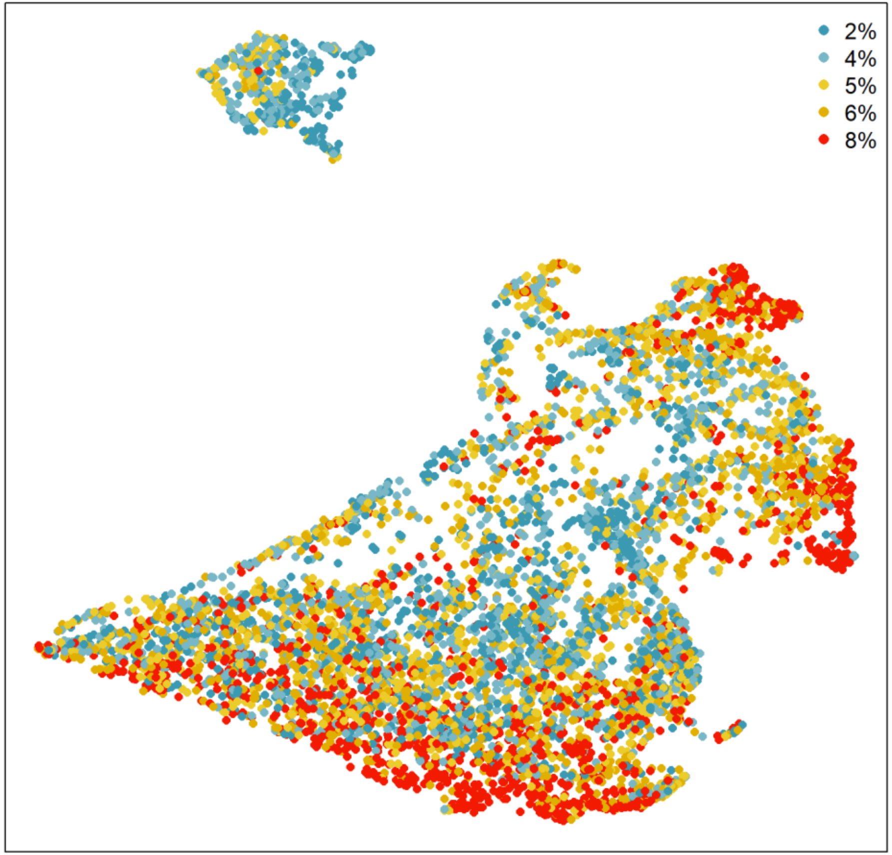


Rozvedení a nezaměstnaní

podíl rozvedených

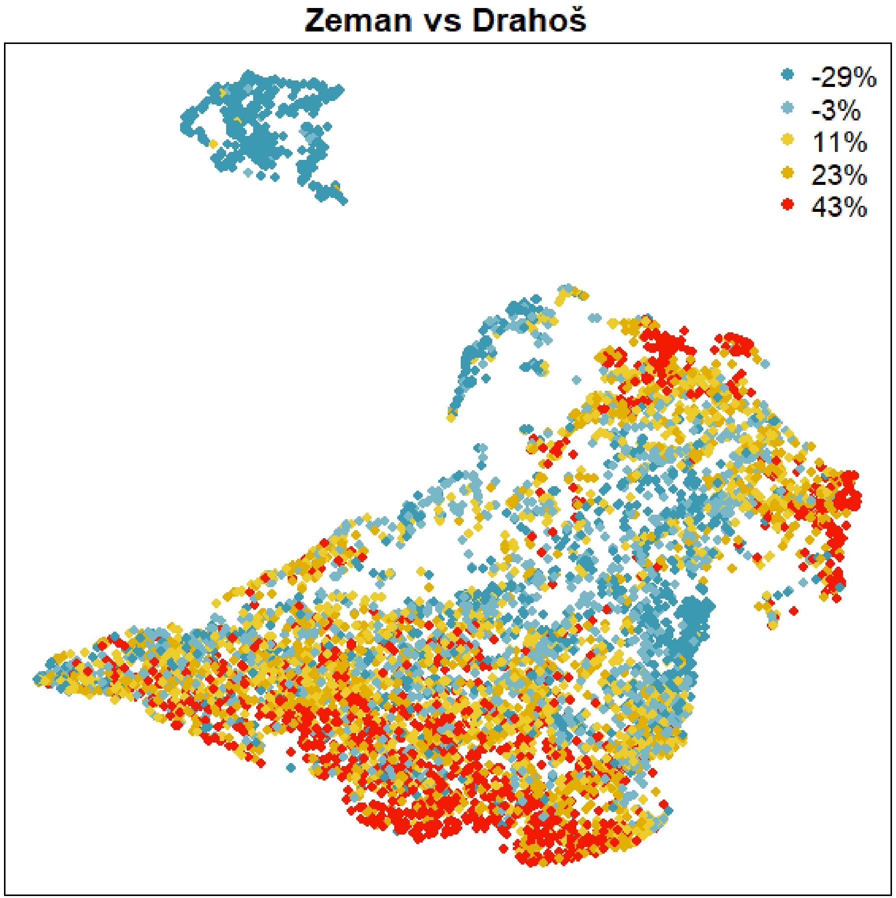


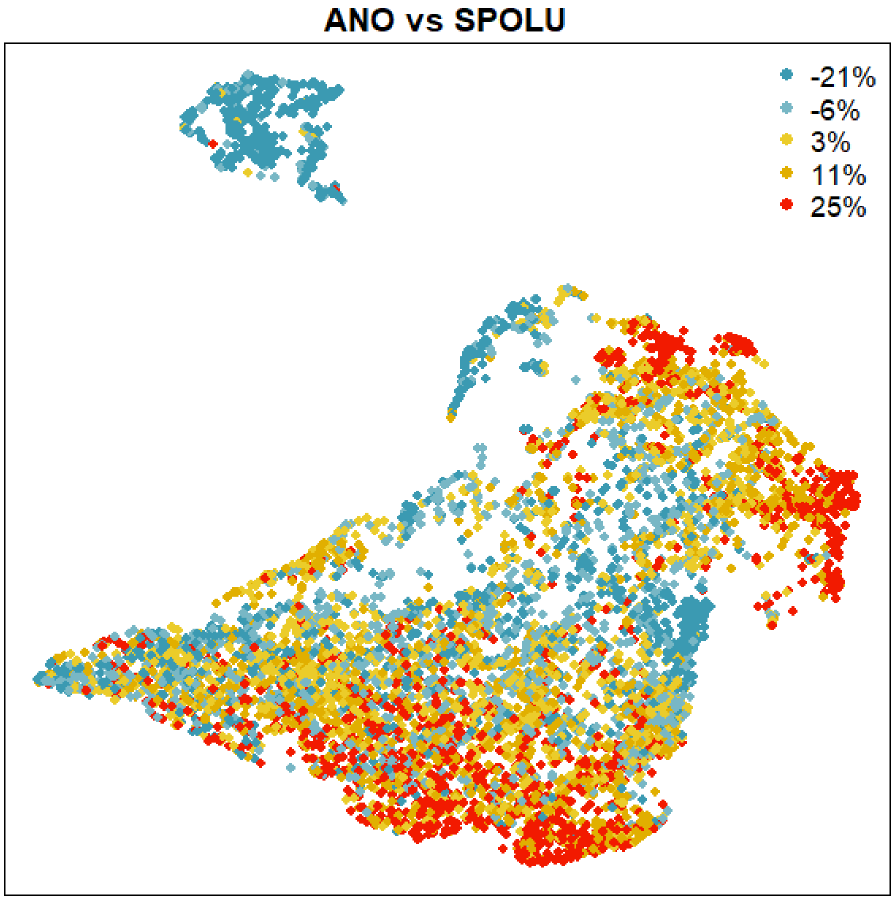
podíl nezaměstnaných



Jak dopadají volby ČR

PROFINIT





Clustering methods

PROFINIT

- grouping (global)
  - k-means
  - Gaussian (kernel)
- hierarchical (local)
  - hclust
  - DBSCAN

Questions?