Generování umělých dat
Zpoždění vlaků

Jan Žítek

Tomáš Pokorný

Cíl

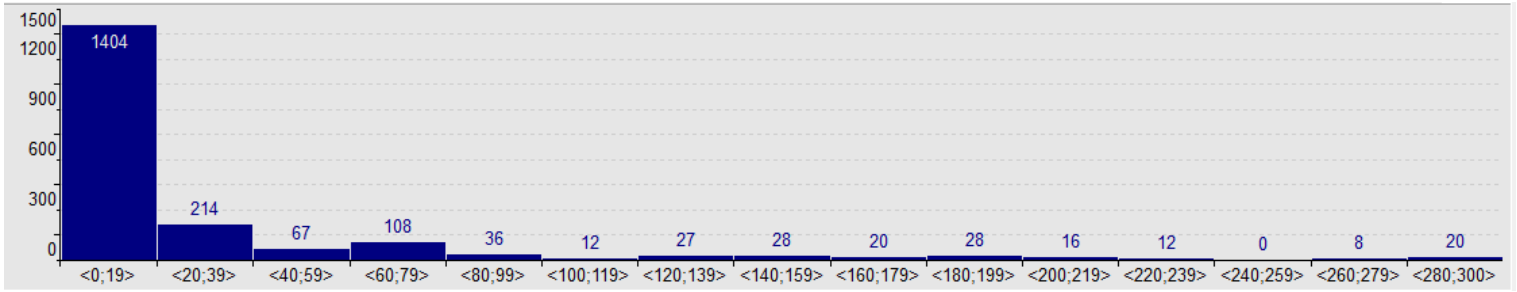
- Vygenerovat věruhodný obraz možných reálných dat
- Konkrétně data o vlakových spojích a jejich zpožděních

- Řádky představují jednotlivé záznamy o vlakových spojích
- Sloupce představují informace o daných spojích

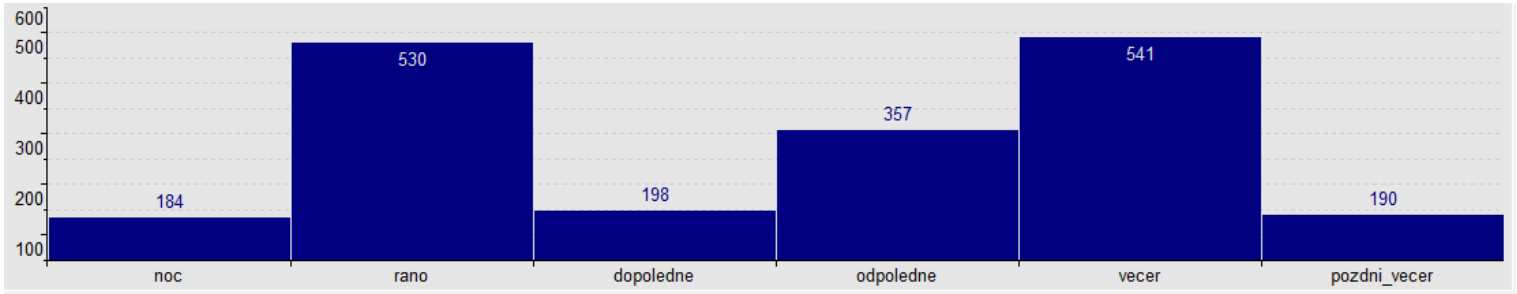
- Celkový počet záznamů: 2000
- Naše data napodobují možná reálná data z roku 2014

Struktura dat

- Zpoždění (minuty)
  - Celé číslo, hodnoty: 0 – 300, vlak je zrušen při zpoždění ≥ 240



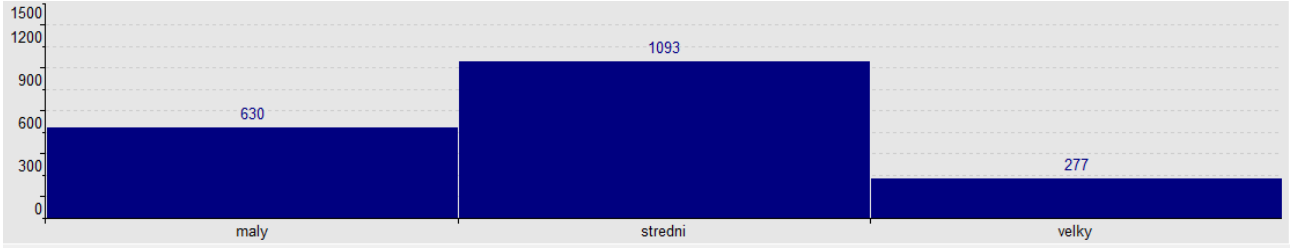
- Čas příjezdu (hodiny, decimální pojetí)
  - Decimální, hodnoty: 0 – 23,99
    - DZ: V ranních a večerních hodinách jezdí více vlaků než v ostatních denních dobách



noc <22;4), ráno <4;9), dopoledne <9;12), odpoledne <12;16), vecer <16;20), pozdni_vecer <20;22)

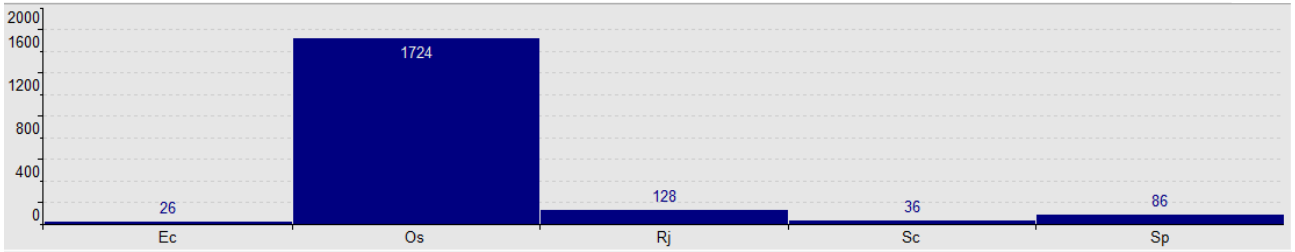
Struktura dat

- Datum
  - Hodnoty: 1.1.2014 – 31.12.2014
    - rovnoměrné rozdělení
- Mimořádná událost
  - Nominální, hodnoty: ano, ne
    - V datech: Ano – 6,8 %, Ne – 93,2 %
- Provoz na trati
  - Nominální, hodnoty: malý, střední, velký
    - V datech: Malý – 31,5 %, Střední – 54,6 %, Velký – 13,9 %



Struktura dat

- Vlak
  - Nominální, hodnoty: Os, Ec, Rj, Sc, Sp
    - DZ: Os – 89,0 %, Rj – 6,0 %, Sp – 3,5 %, Ec – 1,0 %, Sc – 0,5 %
    - V datech: Os – 86,2 %, Rj – 6,4 %, Sp – 4,3 %, Ec – 1,3 %, Sc – 1,8 %



- Výluka
  - Nominální, hodnoty: ano, ne
    - Ano – 10 %, Ne – 90 %

Struktura dat

- atributy závislé na hodnotách ve sloupci *Datum* (externí data)

- Srážky
  - Decimální
    - hodnoty: od 0 do 43,18
    - Průměr: 1,35

- Teplota
  - Decimální
    - Hodnoty od -7,56 do 27.056
    - Průměr 10,75

- Svátek
  - Nominální, hodnoty: ano, ne
    - Ano – 7,1 %, Ne – 92,9 %

Vztahy zanesené do dat

- Při mimořádné události na trati je zpoždění vlaku více než 30 minut
  - Mimořádná událost(ano) $\Rightarrow_{0.6,\ 2\ \%}$ Zpoždění(<30;300>)
    - Nalezení vztahu v datech:

Nr.	Id	Conf	Hypothesis
1	1	0.783	**TMimoradna_udalost**(*ne*) >÷< **VZpozdeni**(*<0;29>*)
2	2	0.600	**TMimoradna_udalost**(*ano*) >÷< **VZpozdeni**(*<30;300>*)

- Mimořádné událost často nastávají také v ranních hodinách, pokud průměrná denní teplota klesne pod 5 stupňů
  - Teplota(<=4) $\wedge$ Čas příjezdu(<4;9)) $\Rightarrow_{0.6,\ 2\ \%}$ Mimořádná událost (ano)
    - Nalezení vztahu v datech (ručně vybraná kategorie v sukcedentu):

Nr.	Id	Conf	Hypothesis
1	1	0.625	**PTeplota**(*pod_4_vcetne*) & **VCasPrijezdu**(*rano*) >÷< **TMimoradna_udalost**(*ano*)

Vztahy zanesené do dat

- Pokud se jedná o trať s velkým provozem, pak výluka na trati znamená zpoždění u všech vlaků.
  - Provoz(Velký) $\wedge$ Výluka(Ano) $\Rightarrow_{0.8,\ 2\ \%}$ Zpoždění(ano)
    - Nalezení vztahu v datech:

Nr.	Id	Conf	Hypothesis
1	1	0.816	**TProvoz**(*velky*) & **TVyluka**(*ano*) >÷< **VZpozdeni**(*zpozden*)

- Pokud se jedná o trať s velkým provozem (a nemusí dojít k výluce ani mimořádné události na trati), tak vznikají zpoždění u více než 30 % osobních vlaků.
  - Provoz(Velký) $\Rightarrow_{0.6,\ 2\ \%}$ Zpoždění(ano) / Vlak(Os)
    - Nalezení vztahu v datech (ručně vybraná kategorie v sukcedentu):

Nr.	Id	Conf	Hypothesis
1	1	0.602	**TProvoz**(*velky*) >÷< **VZpozdeni**(*zpozden*) / **VTypVlaku**(*Os*)

Vztahy zanesené do dat

- Pokud se jedná o trať se středním provozem, pak výluka na trati znamená zpoždění osobních a spěšných vlaků
  - Provoz(Střední) $\wedge$ Výluka(Ano) $\Rightarrow_{0.4,\ 2\ \%}$ Zpoždění(zpožděn) / Vlak(Os) $\vee$ Vlak(Sp)
    - Nalezení vztahu v datech (ručně vybraná kategorie v antecedentu i sukcedentu):

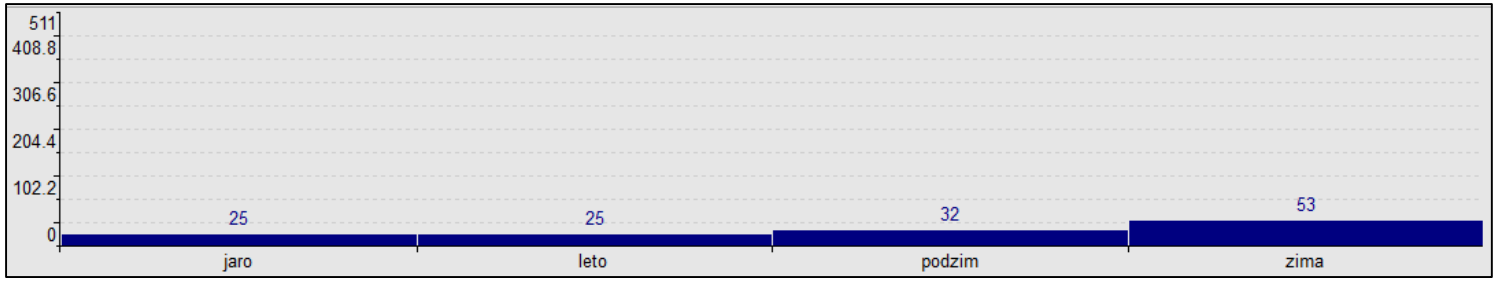
Nr.	Id	Conf	Hypothesis
1	1	0.440	**TProvoz**(*stredni*) & **TVyluka**(*ano*) >÷< **VZpozdeni**(*zpozden*) / **VTypVlaku**(*Os*)

    - Vztah zaveden pouze pro Os
- V pracovní dny dochází k více zpožděním než v nepracovní dny (víkendy a svátky)
  - Den(pracovní) $\wedge$ Svátek(ne) $\Rightarrow^{+}_{0.3,\ 2\ \%}$ Zpoždění(zpožděn)
    - Nalezení vztahu v datech (ručně vybraná kategorie v sukcedentu):

Nr.	Id	AvgDf	Hypothesis
1	1	0.301	**DDenVTydnu**(*Pracovni*) & **DSvatek**(*ne*) >÷< **VZpozdeni**(*zpozden*)

Vztahy zanesené do dat

- Mimořádné události vznikají nejčastěji v zimě
  - Vlaky ≈ $_{\text{Sum(Zima)>=55 \% C}}$ Roční období/Mimořádná událost(ano)
    - Histogram s počty výluk podle ročních období ve vygenerovaných datech



- Čím větší jsou průměrné denní srážky, tím delší je zpoždění vlaků
  - Vlaky ≈ $_{\text{TauB 0.4}}$ Srážky x Zpoždění

Komentář:

Při zadávání vztahu byla u kvantifikátoru v sekci *Parameters* zaškrtnuta volba *Absolute value of TauB for Kendall's coefficient.* Tím bylo dovoleno zavést vztah o požadované síle jak přímé závislosti, tak nepřímé. V tomto případě byl bohužel zanesen vztah nepřímý (hodnota koeficientu menší než -0,4). Do dat byl tedy zanesen opačný vztah, než bylo požadováno.

Zhodnocení

- Rozložení hodnot ve sloupcích respektují doménové znalosti.
- Do dat bylo zaneseno 7 skrytých vztahů (+ 1 částečně a 1 chybně)

- Oproti specifikaci byly vypuštěny 2 vztahy.

- Můžeme tvrdit, že data obstojným způsobem napodobují skutečné hodnoty zpoždění vlaků z roku 2014.

- Možná vylepšení:
  - Oprava vztahu vyjadřujícího přímou úměru mezi výší srážek a délkou zpoždění
  - Zesílení některých stávajících vztahů v datech
  - Zavedení dalších vztahů
  - Mírné rámcové snížení hodnot zpoždění

Použité zdroje

- LISp-Miner – Generování umělých dat
  - http://lispminer.vse.cz/wiki/doku.php?id=mrm:start
- Data o počasí z NOAA Satellite and Information Service
  - Dostupné na: http://www7.ncdc.noaa.gov/CDO/cdo
- Články na webu:
  - Aktuálně.cz: Zpoždění přes hodinu? Kdy vám České dráhy dají odškodnění. http://zpravy.aktualne.cz/finance/zpozdeni-pres-hodinu-drahy-zavedly-dobrovolne-odskodneni/r~17ef74547ae311e4840b002590604f2e/
  - České dráhy: O společnosti https://www.cd.cz/infoservis/o-spolecnosti/-3540/
  - České dráhy: Tiskové zprávy http://www.ceskedrahy.cz/tiskove-centrum/tiskove-zpravy/-14775/
  - České Dráhy: Standardy kvality společnosti České dráhy a.s. https://www.cd.cz/assets/infoservis/cim-se-ridime/standardy-kvality-spolecnosti-ceske-drahy--a-s-.pdf
  - Železničář.cz: ČD loni přepravily rekordní počet cestujících v pětileté historii https://zeleznicar.cd.cz/zeleznicar/hlavni-zpravy/cd-loni-prepravily-rekordni-pocet-cestujicich-v-petilete-historii/-6499/