

Dobývání znalostí

Analýza zpoždění spojů ČD typu EC

Michal Hampl, Otto Gold

17.5.2008

Osnova

- **Využití metodiky SEMMA**
 - **Podrobný popis použitého systému a úprava dat**
 - **Prezentace výsledků**
 - **Aplikace jako výsledek dobývání a uvedení získaných znalostí do rutinní praxe**
-

Realizace (S)emma - Sample

- o z dostupných dat o průjezdu vlaků kontrolními body ČD
 - o Dráhy nejsou ochotna zveřejnit aktuálnější informace, naše data pocházejí z roku 2004
 - o Protože dráhy provozují množství spojů, vybrali jsme si pouze vlaky EC (nejméně spojů na hlavních koridorech) za jeden měsíc
-

Realizace s(E)mma - Explore

- o analýza úplnosti informací**
 - n Problém 1: výrazný rozdíl mezi počty záznamů o jednotlivých spojích**
 - n Problém 2: výrazný rozdíl mezi četností informací z různých kontrolních bodů**
- o spolehlivost pro jednotlivé části trati**
 - n kontrola histogramy**
 - n definice nových podmínek na data (minimální # pro spolehlivost přes zastávky)**

Realizace se(M)ma - Modify

- o **využití SQL databáze (velký objem dat)**
 - n **Usnadnění transformace dat (př. Funkce pro získání explicitního dne v týdnu)**
 - n **Usnadnění získání popisných statistik**
 - n **Příprava pro vývoj aplikace (dále...)**
-

Realizace sem(M)a - Model

- **Modelování popisných statistik**
 - **Vytvoření modelu pro naivní Bayesovský klasifikátor**
 - **Aplikace, která bude na základě starých dat modelovat nové informace o předpokládaném zpoždění**
 - **Motto: „zadej kdy a čím jedeš a já ti řeknu kdy dorazíš“**
-

Realizace semm(A) – Assess

- **Od počátku byla účelem aplikace výsledků v praxi („všichni jezdíme vlakem“)**
 - **PHP aplikace provádějící bayesovskou klasifikaci dle starých dat**
 - **Předpoklad: situace na koridorech se nezměnila od roku 2004 :D**
-

Úprava dat

- o **Problém 1: některé vlaky máji výrazně více záznamů než jiné**
 - o **Problém 2: v délce trasy u vybraného úseku max. Dat a u zbytku min.**
 - n (redukce na úsek)
 - o **Problém 3: použitelnost by byla pro koridor, ale linky se příliš nepřekrývají**
 - n (výběr „spolehlivých“ koridorů)
-

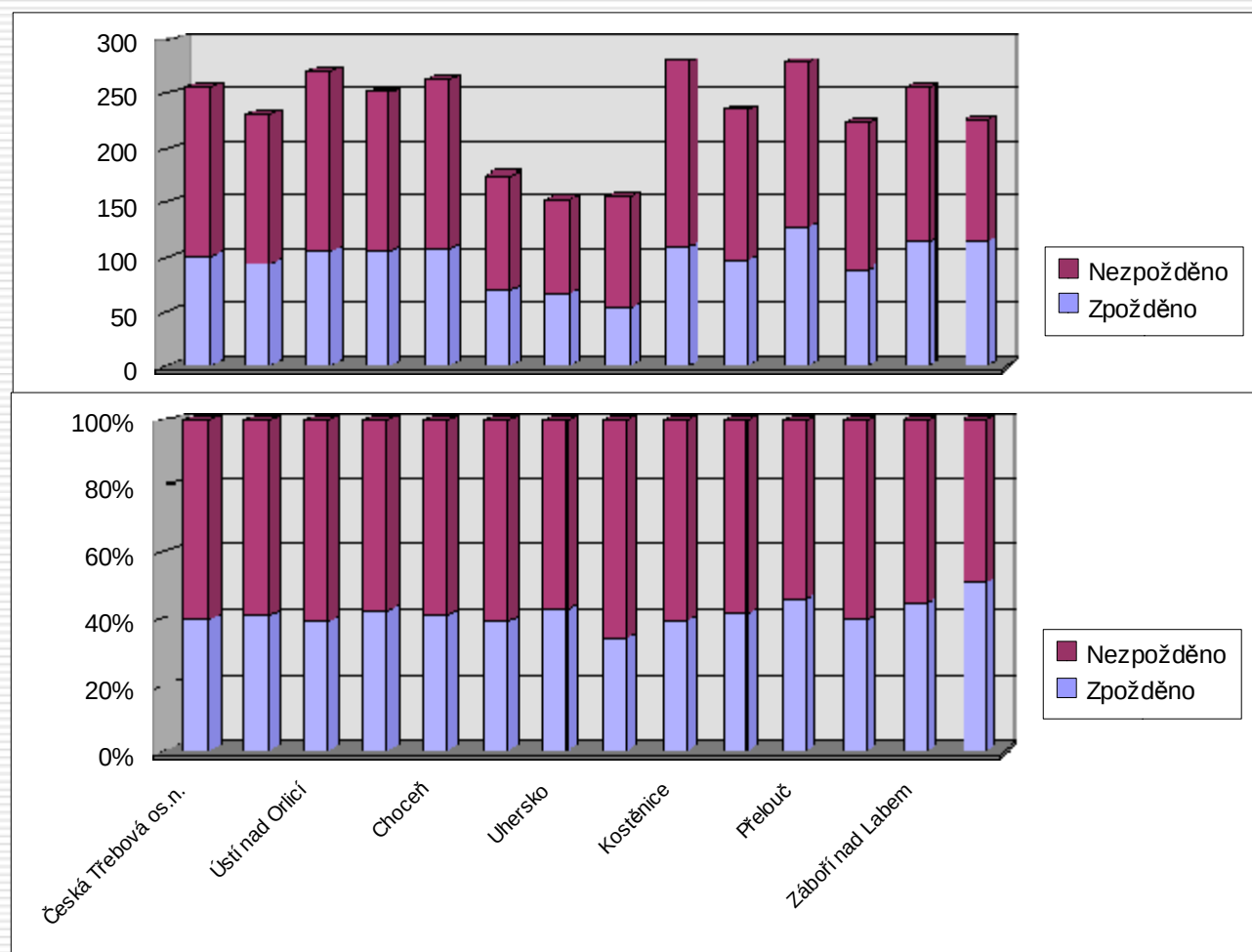
Úprava dat ilustrace (1)



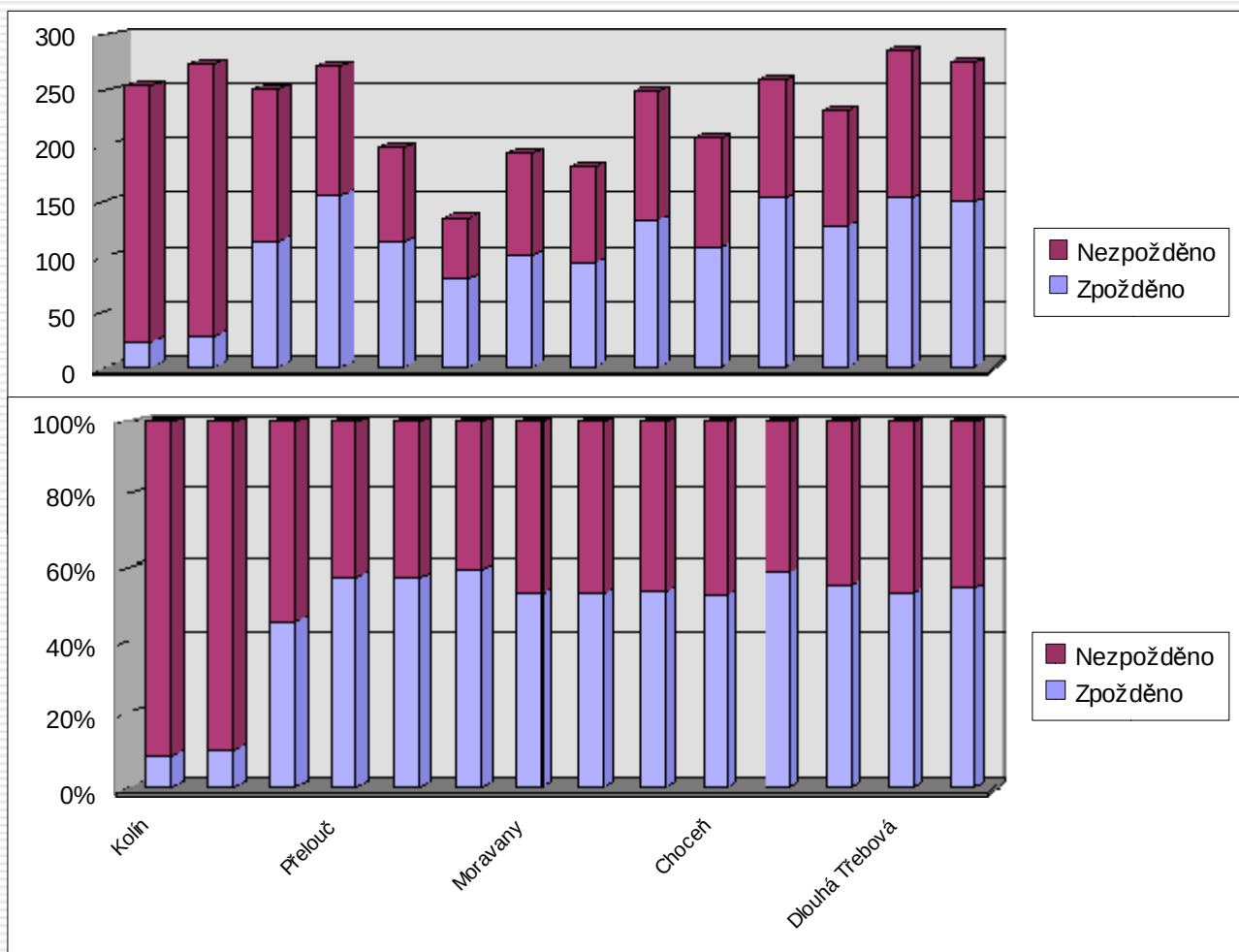
Úprava dat ilustrace (2)

- o **ceska trebova - kolin - 106, 120, 132, 140, 142**
 - o **kolin - ceska trebova - 107, 121, 133, 141, 143**
 - o **praha - usti - 170, 172, 174, 178**
 - o **usti - praha - 171, 173, 177**
-

Výsledky – KO / ČT (1)



Výsledky – KO / ČT (2)



Aplikace

http://info.gootik.net/dm.php?run=1&Vlak=142&Den=3 - Windows Internet Explorer

http://info.gootik.net/dm.php?run=1&Vlak=142&Den=3

Go Bookmarks 182 blocked AutoLink AutoFill Settings

http://info.gootik.net/dm.php?run=1&Vlak=142&Den=3

Home RSS Print Page Tools

Vlak: 142
Den v tydnu: Streda

Interval	P(interval)	P(vlak interval)	P(den v tydnu interval)	P(interval) * P(vlak interval) * P(den v tydnu interval)
0-1	373 / 794 = 0.470	3 / 373 = 0.000	55 / 373 = 0.147	0.010
1-5	123 / 794 = 0.155	7 / 123 = 0.447	20 / 123 = 0.163	0.004
5-10	95 / 794 = 0.120	6 / 95 = 0.211	12 / 95 = 0.126	0.002
10-20	124 / 794 = 0.156	6 / 124 = 0.097	20 / 124 = 0.161	0.004
20-30	41 / 794 = 0.052	8 / 41 = 0.488	6 / 41 = 0.146	0.001
30-60	32 / 794 = 0.040	5 / 32 = 0.188	5 / 32 = 0.156	0.001
60-	6 / 794 = 0.008	1 / 6 = 0.833	0 / 6 = 0.000	0.000

Done

Internet 100%