

Předzpracování dat

Měření a zpracování dat (MZD)

Ondřej Příbyl

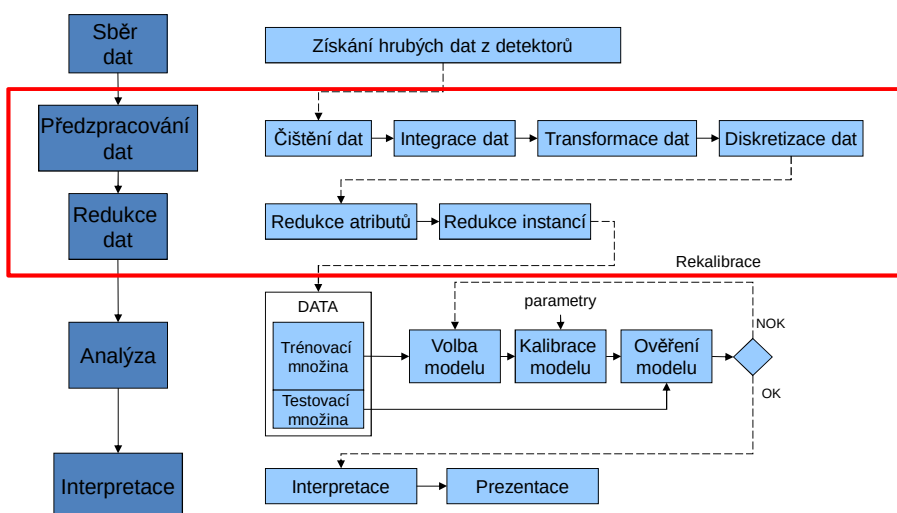
Ústav aplikované matematiky
ČVUT v Praze, Fakulta dopravní



Ondřej Příbyl K611MDS

strana 1

Hlavní kroky – obsah prezentace



Ondřej Příbyl K611MDS

Diskuse

- Co znamená - předzpracování dat?
- Proč je předzpracování dat nutné?

Ondřej Přibyl K611MDS

Proč je třeba předzpracovávat data?

Data v reálním světě jsou „špinavá“

- Nekompletní
- Obsahují šum
- Nekonzistentní
 - Data obsahují protichůdné informace
- Chybná
 - obsahují špatné údaje vzniklé chybami měřicích přístrojů i lidské obsluhy

- Nejednoznačná
 - popsána pomocí příliš mnoha atributů – není zřejmé které jsou relevantní
- Složitá
 - Data mají formu složitého relačního schématu a
 - ne jednoduché tabulky nutné pro matematické algoritmy

Motto: Pokud nejsou kvalitní data, nebudou kvalitní ani výsledky analýzy!

Ondřej Přibyl K611MDS

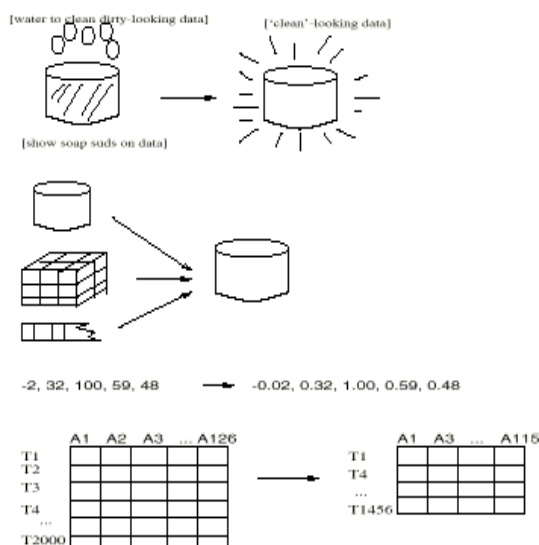
Jak určit kvalitu dat? (Metriky kvality dat)

- **Přesnost** (accuracy)
 - měřeno obvykle statistickými charakteristikami pro chybu, např. směrodatná odchylka
- **Úplnost** (completeness)
 - zda statistické charakteristiky dat nejsou ovlivněny výběrovými efekty.
- **Konzistence** (consistency)
- **Včasnost** (timeliness)
 - za jakou dobu lze data aktualizovat
- **Důvěryhodnost** (believability)
- **Přidaná hodnota** (added value),
- **Interpretabilita** (interpretability),
- **Dostupnost** (accessibility)
 - Technologické, legislativní a procesní bariéry

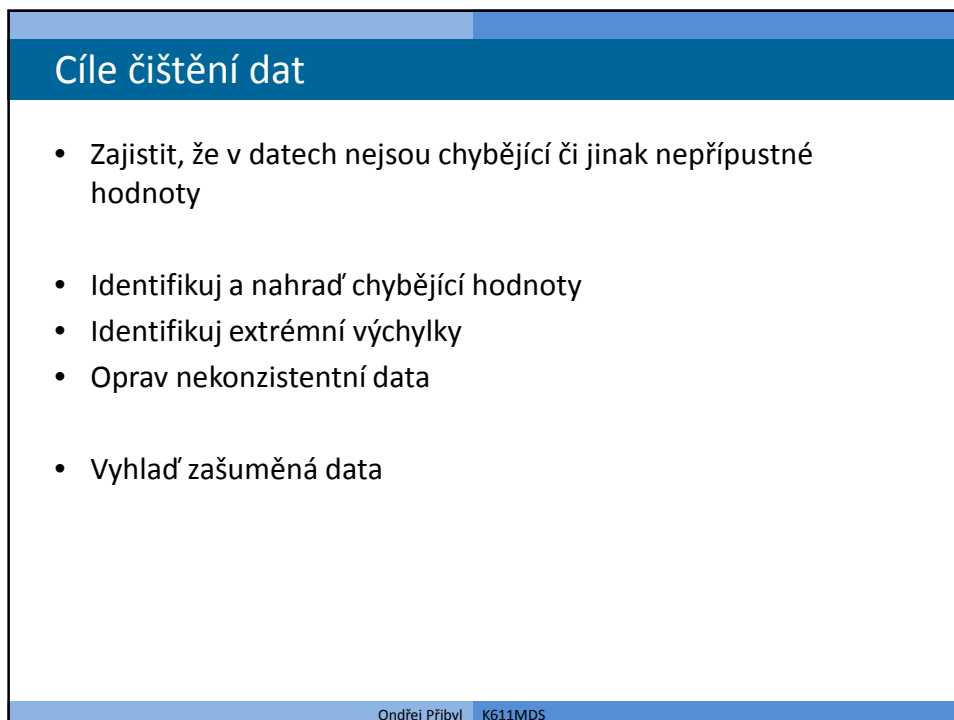
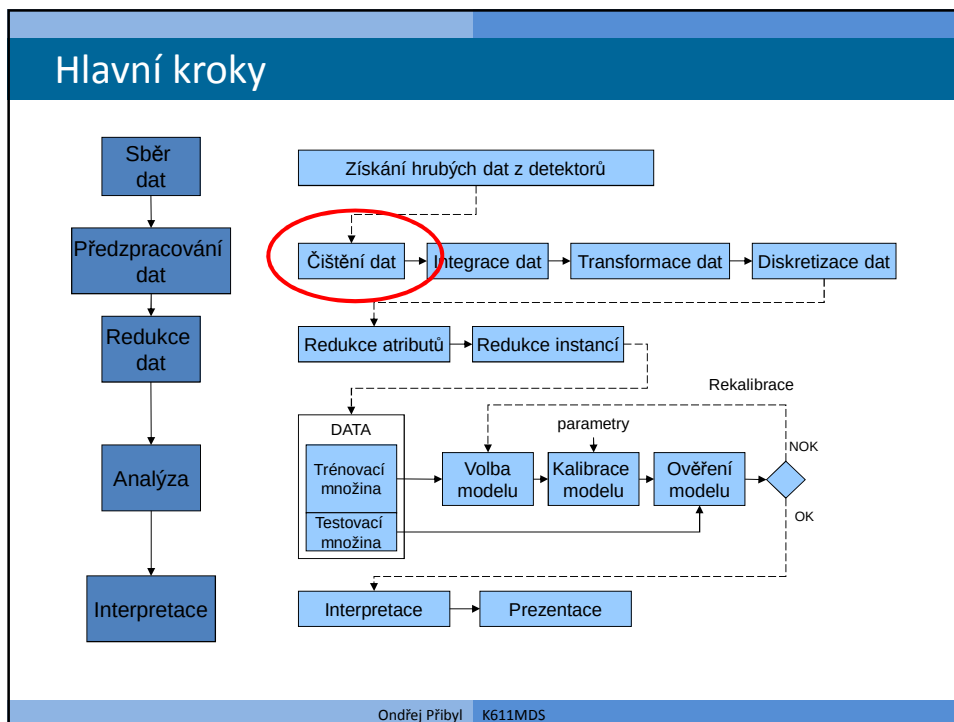
Ondřej Přibyl K611MDS

Hlavní oblasti předzpracování dat

- Čištění dat
- Integrace dat z více zdrojů
- Transformace dat
- Redukce dat
- Diskretizace dat



Ondřej Přibyl K611MDS



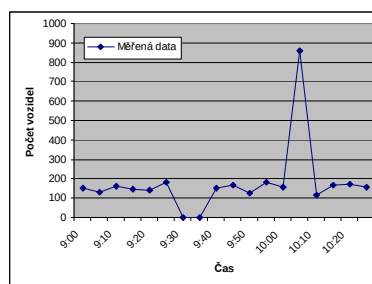
Rozdělení metod čištění dat

- Dle způsobu měření dat
- A. Data v tabulární formě
- B. Časové řady

Atributy

Objekty

Jméno	Zařazení	Roky	Počet publikací
Petr	asistent	3	5
Jan	docent	8	15
Michaela	profesor	12	54
Jiří	asistent	5	?
David	docent	7	?
Jana	asistent	1	1
Martina	asistent	3	12
Petr	docent	5	51
Michal	profesor	10	35
Martin	profesor	8	45



Ondřej Přibyl K611MDS

Chybějící data

Jak dojde ke ztrátě dat?

- nefunkčnost senzorů či problém při přenosu a zápisu dat.
- Nevyplněný dotazník
- **Jak opravit následující tabulku?**

Jméno	Zařazení	Roky	Počet publikací
Petr	asistent	3	5
Jan	docent	8	15
Michaela	profesor	12	54
Jiří	asistent	5	?
David	docent	7	?
Jana	asistent	1	1
Martina	asistent	3	12
Petr	docent	5	51
Michal	profesor	10	35
Martin	profesor	8	45

Ondřej Přibyl K611MDS

Jak opravit chybějící data - A. Tabulární forma?

- Ignorování vzorku
 - obvykle se použije pokud chybí označení třídy (při klasifikaci)
- Manuální vyplnění
 - náročné, nemožné (vytvoření stejných podmínek, stejné subjekty)?
- Vyplnění globální konstantou
 - např. „Chybějící“
- Vyplnění střední hodnotou všech vzorků
- Střední hodnota pro dané atributy
- Pravděpodobnostní modely
 - regrese, rozhodovací stromy, ...

Ondřej Přibyl K611MDS

Jak opravit chybějící data - A. Tabulární forma?

- Vyplnění střední hodnotou všech vzorků
- Střední hodnota pro dané atributy
- Pravděpodobnostní modely (regrese, rozhodovací stromy, ...)

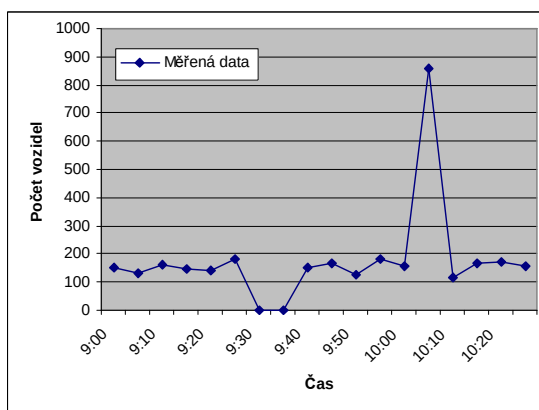
Průměr	30
Průměr asistent	6
Průměr docent	33
Průměr profesor	45
Průměrný roční počet publikací	4

Jméno	Zařazení	Roky	Počet publikací
Petr	asistent	3	5
Jan	docent	8	15
Michaela	profesor	12	54
Jiří	asistent	5	?
David	docent	7	?
Jana	asistent	1	1
Martina	asistent	3	12
Petr	docent	5	51
Michal	profesor	10	35
Martin	profesor	8	45

Ondřej Přibyl K611MDS

Jak opravit chybějící data - B. Časová řada?

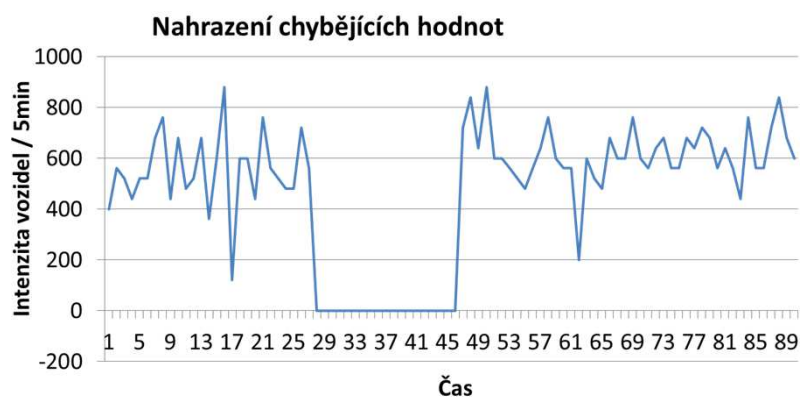
- V prvním kroku je třeba identifikovat nečistá data !
- **Diskuse** - Jak identifikovat chybějící hodnoty v případě časové řady?
- Jak zjistit zda se jedná o naměřenou hodnotu či o chybu?
 - Rozsah hodnot,
 - Statistika,
 - Kontext
 - ...



Ondřej Příbyl K611MDS

Jak opravit chybějící data - B. Časová řada?

- **Diskuse**
 - Jak nahradit chybějící hodnoty v případě časové řady?



Ondřej Příbyl K611MDS

strana 14

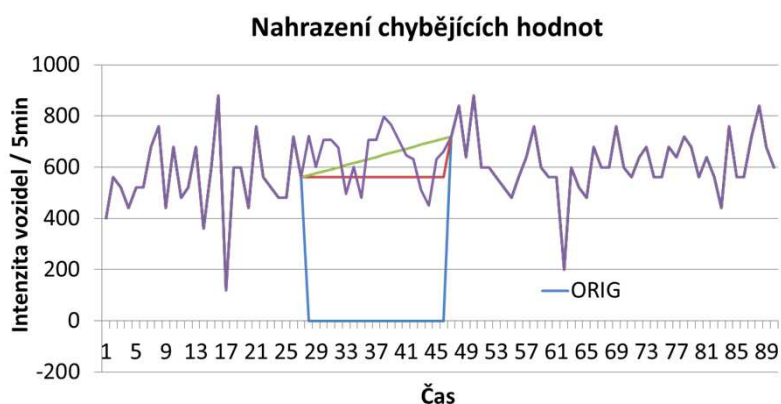
Chybějící hodnoty – časová řada

- Náhrada poslední hodnotou
 - nahradí se poslední správně naměřenou hodnotou.
- Průměr platných hodnot
 - místo chybné hodnoty se použije průměr poslední platné hodnoty před výpadkem a první po výpadku.
- Lineární spojnice platných hodnot
 - Další strana
- Nahrazení dle statického modelu
 - Další strana

Ondřej Příbyl K611MDS

Jak opravit chybějící data - Řešení

- Diskuse
 - Jak identifikovat a nahradit chybějící hodnoty v případě časové řady?

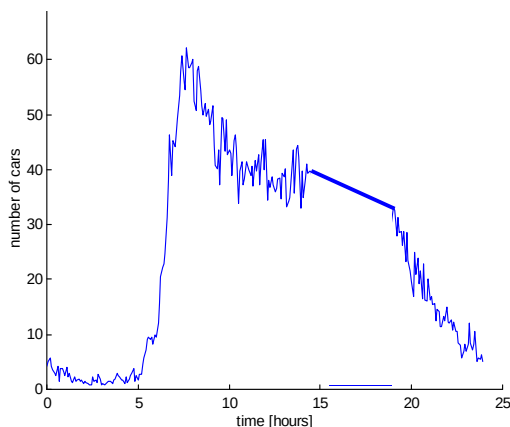


Ondřej Příbyl K611MDS

strana 16

Chybějící hodnoty – časová řada

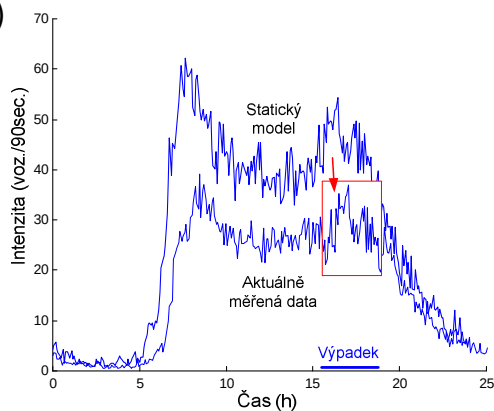
- Lineární spojnice platných hodnot
 - místo chybějící hodnoty se lineární spojnice poslední platné hodnoty před výpadkem a první po výpadku.



Ondřej Příbyl K611MDS

Chybějící hodnoty – časová řada

- Nahrazení dle statického modelu
 - „typické“ chování (takzvaný statický model) je možné použít pro odhad chybějících hodnot.
 - úprava (koeficientem) pro přizpůsobení aktuálnímu rozsahu dat



Ondřej Příbyl K611MDS

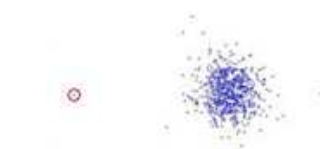
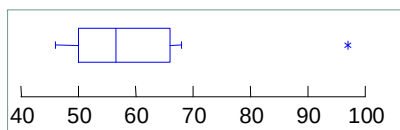
Diskuze

- Co je to outlier?
- Jak identifikovat outliery ?
- Jak odstranit outliery?

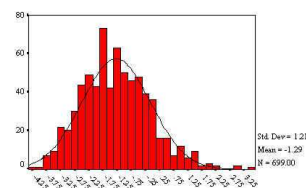
Ondřej Příbyl K611MDS

Detekce extrémních hodnot

- Outliers je třeba identifikovat, popřípadě odstranit
- Metody – **jednorozměrné**
 - Gausovské rozložení - Z score
 - Histogram
 - Boxplot



$$z_i = \frac{(x_i - \bar{x})}{s}$$



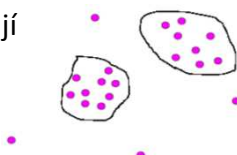
Ondřej Příbyl K611MDS

Detekce extrémních hodnot

Metody – vícerozměrné

- Shlukování

- Najdi „podobná“ data a odstraň ta která se vymykají

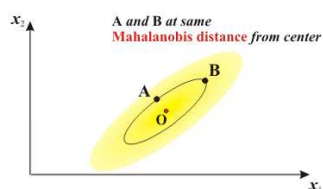


- Binning

- Seřad' hodnoty atributů a sluč je do skupin (bins)
- Vyhlaď je podle středních hodnot,

- Mahalanobisova vzdálenost

- vzdálenost od centroidu



- Regrese

- Najdi data ležící daleko od regresní funkce



Ondřej Přibyl K611MDS

FILTRACE DAT

Ondřej Přibyl K611MDS

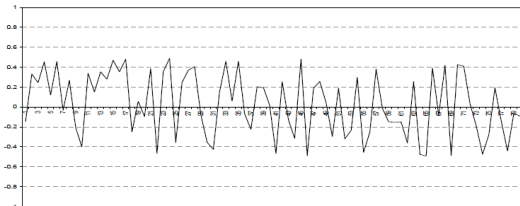
Diskuze

- Co je cílem filtrace?
- Co je to bílý šum (noise)?

Ondřej Příbyl K611MDS

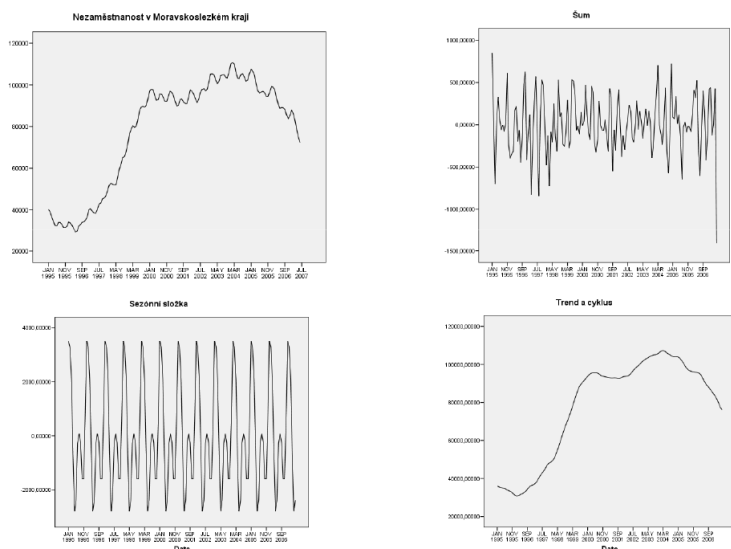
Šum analytického signálu

- náhodné zvýšení nebo snížení měřeného signálu
- Šum, jehož suma je **nulová** v časovém intervalu pozorování, se označuje jako bílý šum
- Šum, jehož suma je **nenulová** v časovém intervalu pozorování, se označuje jako náhodný šum
- Šum je významný jen z hlediska intervalu pozorování
- Intenzivní a náhlé změny signálu (spiky) nelze doslovně považovat za šum - jejich původ bývá v okamžitém porušení funkce měřicího zařízení



Ondřej Příbyl K611MDS

Dekompozice časové řady



source: http://www.spss.cz/files/ruzne/vsb/casove_rady.pdf

Ondřej Přibyl K611MDS

Filtrování dat

- **Důvody pro využití filtrování dat**

- metoda pro čištění dat
- odstranění náhodné složky z dat – odstranění šumu (náhodné složky)

A) Filtrace v **časové oblasti**

- Obvykle se jedná o okno definované velikosti ve kterém se spočítá střední hodnota a ta se použije pro odstranění šumu.

B) Filtrace ve **frekvenční oblasti**

- Data jsou transformována z časové oblasti do oblasti frekvenční.
- Odebráním vysokofrekvenčních složek dojde k odfiltrování šumu a náhodných složek.

Ondřej Přibyl K611MDS

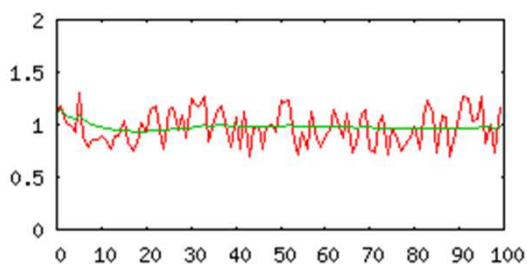
Filtrace dat v časové oblasti

- Veličiny s konstantní hodnotou
 - jednoduchým průměrováním
- Veličiny s pomalu se měnící hodnotou
 - Plovoucí průměr (moving average)
 - Vážený plovoucí průměr (weighted moving average)
 - Exponenciální vyhlazování (exponential smoothing)

Ondřej Příbyl K611MDS

Filtrace dat v časové oblasti

- **Jednoduché průměrování**



$$\hat{x}_n = \frac{\hat{x}_{n-1} \cdot (n-1) + x_n}{n}$$

Ondřej Příbyl K611MDS

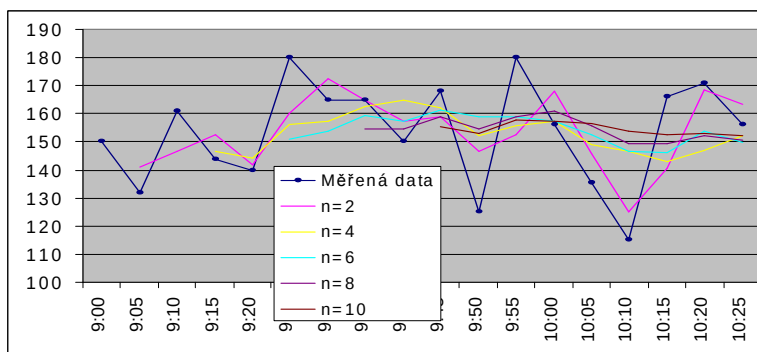
strana 28

Filtrace dat v časové oblasti

- **Plovoucí průměr** (moving average)

- Pokud neexistuje periodický cyklus
- Všechny hodnoty mají stejnou váhu
- k ... velikost ,paměti‘

$$\hat{x}_n = \frac{1}{k} \sum_{i=1}^k x_{(n-i)+1}$$



Ondřej Přibyl K611MDS

strana 29

Filtrace dat v časové oblasti

- **Vážený plovoucí průměr** (weighted moving average)

- Rozlišuje vliv jednotlivých měření

5 Day Weighted Moving Average				
Most Recent	Weight	Data	Weighted Data	
	5	* 90	= 450	
	4	* 85	= 340	
	3	* 82	= 246	
	2	* 80	= 160	
Oldest	1	* 77	= 77	
Totals	15		1273	/ 15 = 84.86

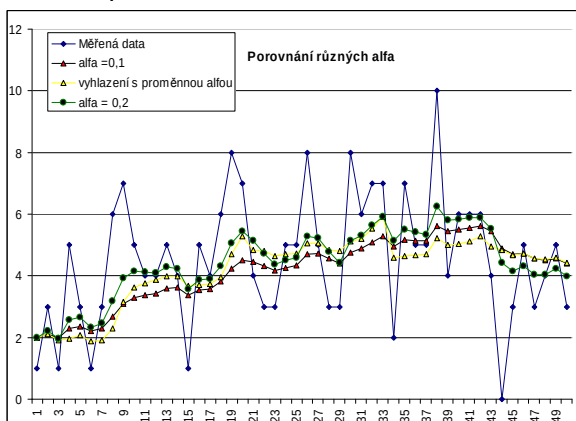
5 Day WMA

$$\hat{x}_n = \frac{\sum_{i=1}^k w_i \cdot x_{(n-i)+1}}{\sum_{i=1}^k w_i}$$

Ondřej Přibyl K611MDS

Filtrace dat v časové oblasti

- Exponenciální vyhlazování (exponential smoothing)
 - S pevnou hodnotou alfa
 - S proměnnou hodnotou alfa



$$\bar{v}_1 = v_1$$

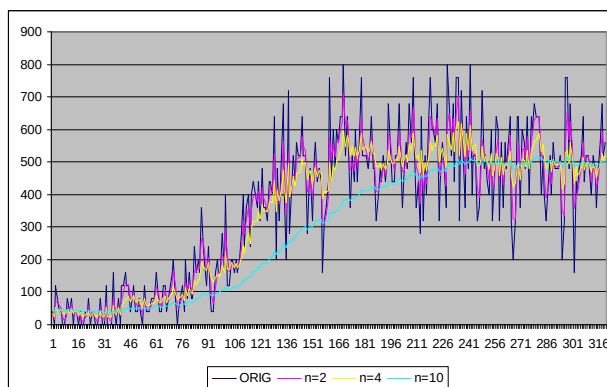
$$\bar{v}_t = \bar{v}_{t-1} + \alpha(v_t - \bar{v}_{t-1})$$

Ondřej Přibyl K611MDS

strana 31

Diskuze

- Jak volit optimální velikost okna pro plovoucí průměr? Je lepší $n=5$, $n=10$ nebo $n=20$?

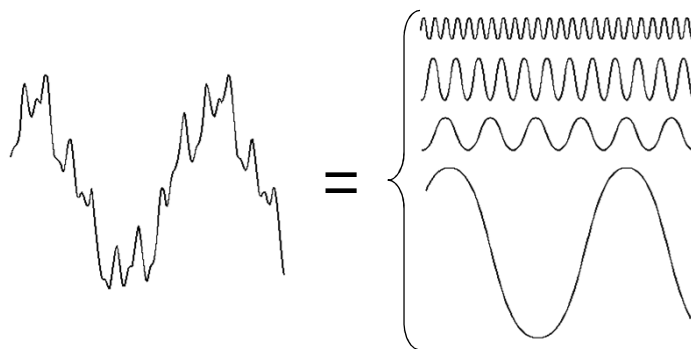


Ondřej Přibyl K611MDS

strana 32

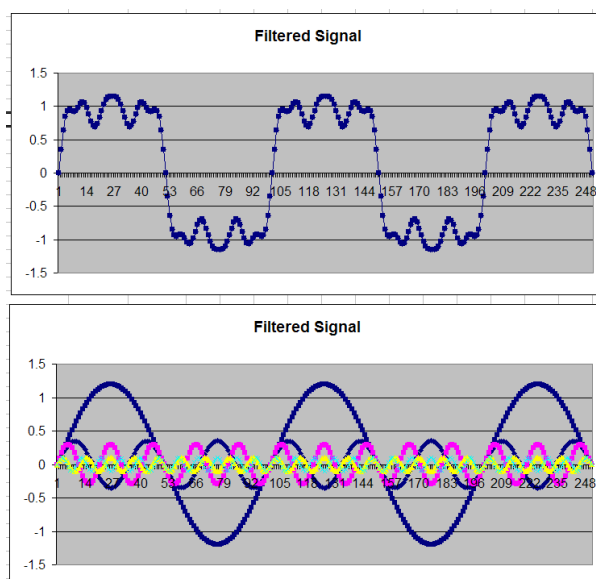
Filtrování dat ve frekvenční oblasti

- Fourierova transformace.
 - Libovolnou periodickou časovou řadu lze nahradit superpozicí sinusových a cosinusových funkcí



Ondřej Příbyl K611MDS

Další příklad...



Ondřej Příbyl K611MDS

http://www.ws.binghamton.edu/fowler/fowler%20personal%20page/EE301_files/EECE301%20Discussion%2012a%20DT%20Filter%20Applications.pdf

Základní kroky pro filtrování ve frekvenční oblasti

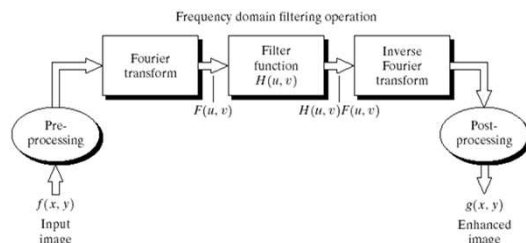
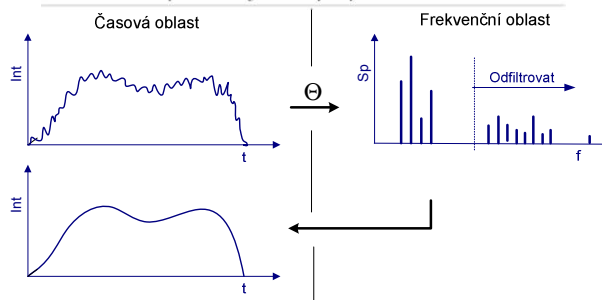


FIGURE 4.5 Basic steps for filtering in the frequency domain.



Ondřej Přibyl K611MDS

Výsledek low-pass filtru

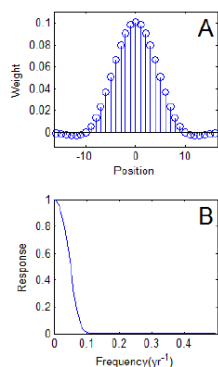


Figure 8.8. Low-pass filter by Hamming-window method. (A) filter weights. (B) Frequency response. Filter response is 0.5 at wavelength 20 yr.

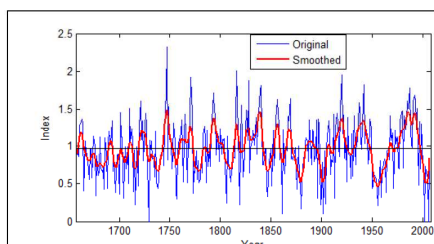


Figure 8.10. Annual index of tree growth, before and after smoothing by binomial filter. Series was extended by reflection across endpoints prior to filtering to provide filtered series with same time coverage as original series.

http://www.ltrr.arizona.edu/~dmeke/notes_8.pdf

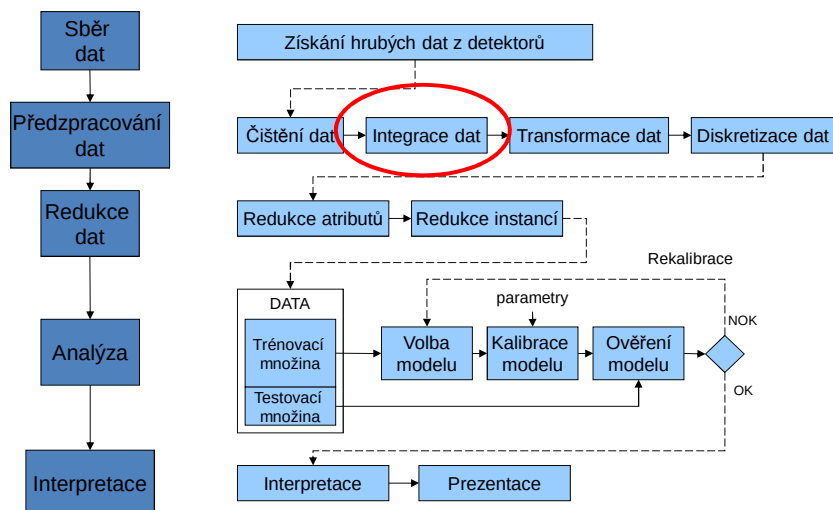
Ondřej Přibyl K611MDS

Filtrace

- Filtrace vždy **znamená ztrátu nějaké informace**
 - Cílem je odfiltrování bílého šumu (tedy informace s nulovou střední hodnotou)
- Přílišná filtrace může poškodit data
- Metoda a nastavení filtrace je třeba volit
 - dle cíle a
 - dle následné analytické metody
- V důsledku filtrace ke **zpoždění informace** (pokud potřebujeme rychle reagovat na změnu trendu dat (např. identifikace nehod))
- Bayesovské metody se o filtraci postarají, při lineární regresi je určité filtrování spíše výhodou

Ondřej Přibyl K611MDS

Hlavní kroky



Ondřej Přibyl K611MDS

Integrace dat

- Integrace dat:
 - Kombinování dat z více zdrojů
- Důvod pro integraci dat?
 - Je třeba **odstranit duplicity a konflikty** v datech, aby nedošlo k nekonzistentním stavům
- Důvody nekonzistence?
 - I / int / Int1
 - Věk / Datum narození
 - tabA_Int / tabB_Int
 - Cena (EUR) / Cena (Kč)
 - Rychlost (kmh) / Rychlost (mph)

Ondřej Přibyl K611MDS

Diskuse

- Jak odhalím redundantní data?

Ondřej Přibyl K611MDS

strana 40

Jak zjistit redundantní data při integraci?

- Redundantní data (časové řady) mohou být zjištěna pomocí korelační analýzy
 - popisuje lineární vztahy mezi veličinami (míra závislosti)
- Párový (Pearsonův) **korelační koeficient**
 - míra vyjádření “těsnosti **lineární** vazby”. (od -1 do +1)

$$R = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{(n - 1) s_x s_y} \quad R = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

- “Vypočteme [aritmētické průměry](#) souborů X a Y, vynásobíme sumy [čtverců](#) odchylek od těchto průměrů obou souborů. Tím jsme spočetli tzv. [kovarianci](#), což je však absolutní veličina, pro výpočet relativní veličiny pak kovarianci dělíme odmocninou násobku [rozptylu](#) souboru X a souboru Y. “

Ondřej Přibyl K611MDS

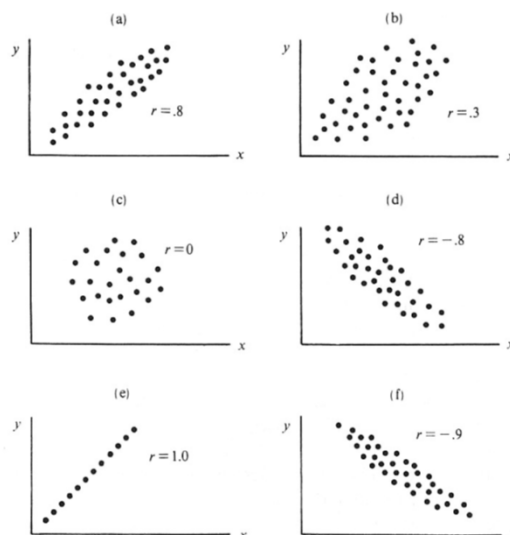
Diskuse

- Co znamenají hodnoty korelačního koeficientu rovné
 - 0?
 - 1?
 - -1?
 - 0,5?

Ondřej Přibyl K611MDS

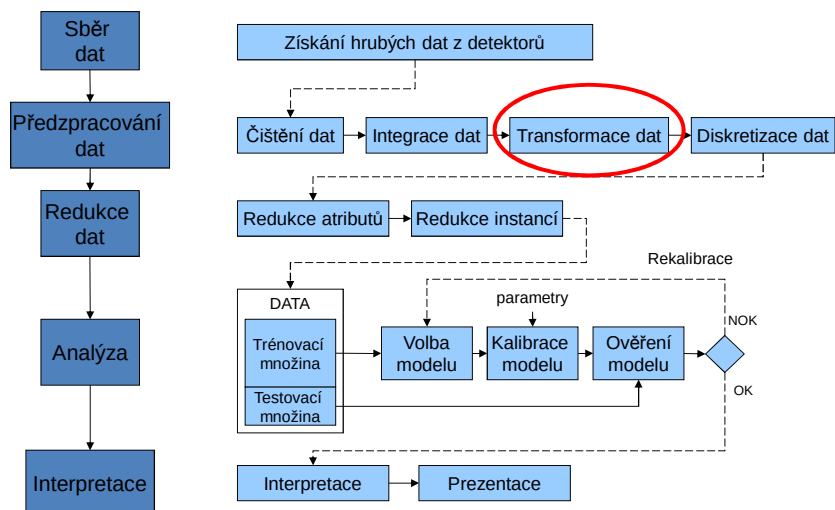
strana 42

Ukázka korelačních koeficientů



Ondřej Přibyl K611MDS

Hlavní kroky

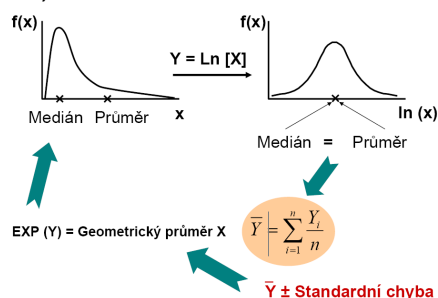


Ondřej Přibyl K611MDS

Transformace dat

Odstranění závislosti atributů na jednotkách měření

- Transformace
 - **Logaritmická**, odmocninová, ...
- Agregace
 - souhrny, vytváření globálních pravidel
- Generalizace
 - koncept hierarchického rozvrstvení
- Normalizace
 - Změna měřítka tak, aby data náležela do určitého intervalu
 - Vhodné pro porovnávání různých dat

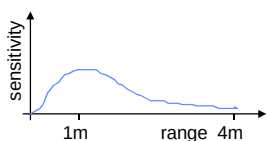


Ondřej Přibyl K611MDS

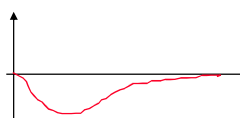
Příklad transformace -

nelinearita

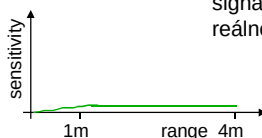
• Korekce mlhy



Skenery do venkovního prostředí- Outdoor- vykazují při opticky nečistém prostředí zvýšenou citlivost okolo 1,3m



Podle křivky je citlivost upravována opozitním signálem, který je přičítán k reálnému signálu



Výsledná křivka citlivosti je téměř plochá. To zaručuje konstantní citlivost pro celou zaručenou vzdálenost detekce.

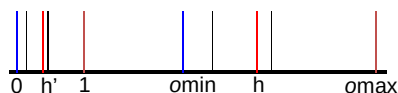
Ondřej Přibyl K611MDS 4. přednáška

46

Transformace dat - normalizace

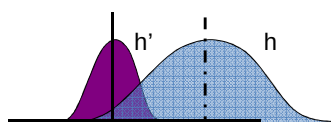
- Lineární normalizace

$$h' = \frac{(h - o \min)(new \max - new \min)}{o \max - o \min} + new \min$$

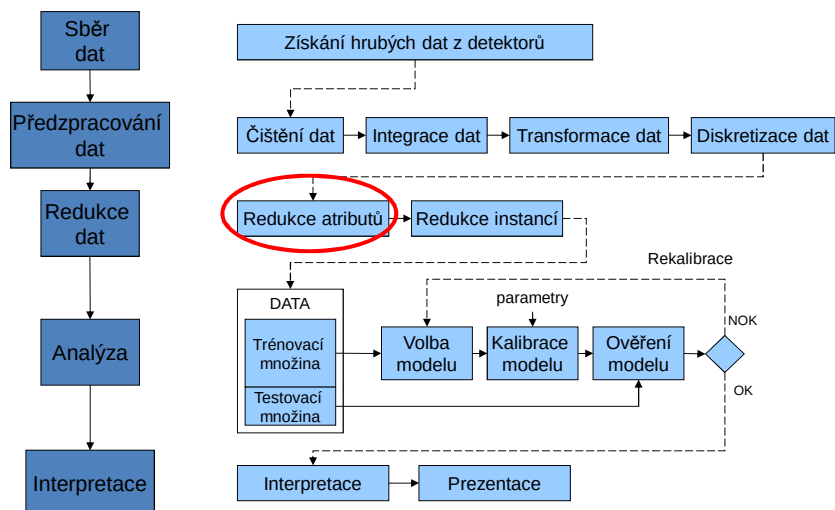


- Z-score normalizace

$$h' = \frac{h - mean_A}{std_A}$$



Hlavní kroky



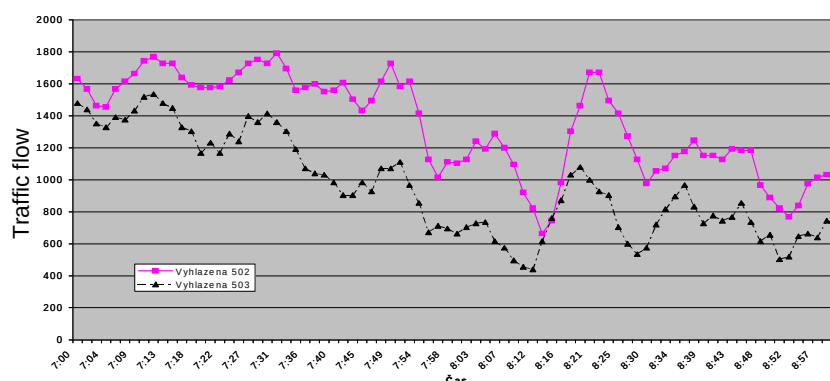
Proč je třeba redukce dat?

- V praxi se sbírá **obrovské množství dat**
 - Na většině řízených křižovatek se nachází v každém jízdním pruhu alespoň jedna indukční smyčka
 - Z každé indukční smyčky se v pravidelných intervalech 90 sekund sbírají informace o intenzitě dopravy a obsazenosti detektoru.
- Toto obrovské množství dat se přenáší na hlavní dopravní řídící ústřednu v Praze
- **Co s tím?**
 - Klasické algoritmy jsou zahlceny
 - Není možné najít „důležitá data“ – informaci o dopravní nehodě, a pod.

Ondřej Přibyl K611MDS

Redukce dat - Motivace

- Sbíraná data často nesou podobnou, či zcela stejnou informaci
 - Viz. detektory umístěné na jedné linii za sebou



Ondřej Přibyl K611MDS

Redukce dat

- Strategie redukce dat
 - Redukce dimenzionality
 - Komprese dat
 - Redukce vzorků
 - Diskretizace

Redukce dat - cíl

Získat redukovanou množinu dat, která je mnohem menší objemově, ale produkuje (téměř) stejné analytické výsledky.

- **Výběr** podmnožiny atributů
 - Vyber minimální podmnožinu všech atributů, které dostatečně reprezentují původní rozdělení dat

Ondřej Přibyl K611MDS

Redukce dimenzionality - Jak vybrat atributy?

Heuristické metody (exponenciální počet možností) jsou **greedy**

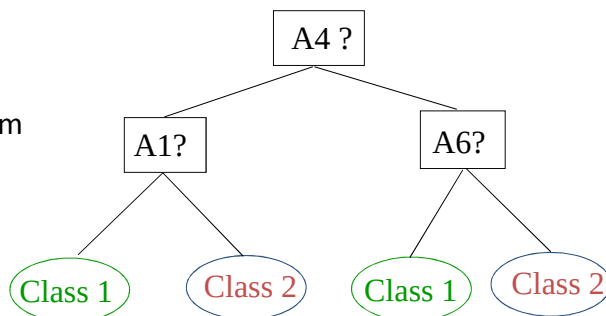
- **Vynecháním**
 - Konstantních atributů
 - Řídce obsazených atributů
 - Atributů s duplicitní informací (věk x datum narození)
 - Korelační analýza, ANOVA
- **Sloučením** atributů
- **Analyticky**
 - Rozhodovací stromy
 - Fourierova transformace, Wavelet transformace
 - Principal component analysis
 - Shlukování

Ondřej Přibyl K611MDS

Příklad rozhodovacího stromu

- Původní množina atributů **{A1, A2, A3, A4, A5, A6}**

- Výsledný strom

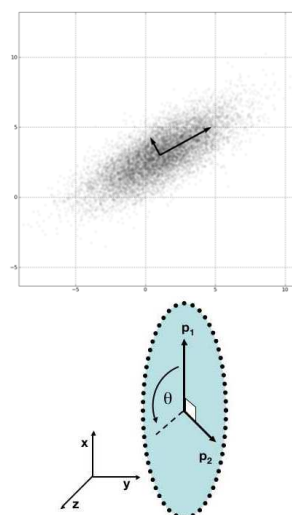


- Redukovaná množina atributů: **{A1, A4, A6}**

Ondřej Přibyl K611MDS

Metoda PCA (Principal Component Analysis)

- Analýza hlavních komponent
 - Snížení dimenze dat s co nejmenší ztrátou informace
- Nové atributy/dimenze:
 - Lineární kombinace původních
 - Nekorelované
 - Ortogonální
 - Zachycují co nejvíce původní variance v datech



Ondřej Přibyl K611MDS

Komprese dat

- Komprese řetězců
 - Mnohé existující algoritmy (mpeg4, jpg, ...)
 - Obvykle bezztrátové
- Využívá se pro kompresi audio/video dat
- Příklad - Kódování RLE (Run-Length Encoding)
 - posloupnost opakujících se bytů se nahradí jednou hodnotou s uvedením počtu opakování

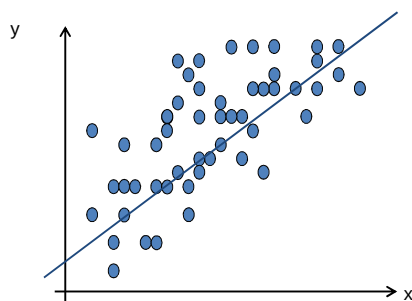
"AAAhooooj"

"<3A>h<4o>j"

Ondřej Přibyl K611MDS

Redukce počtu vzorků – parametrické metody

- Parametrické metody
 - Pokud původní data odpovídají určitému modelu, je možné je nahradit tímto modelem



$$y = \alpha + \beta \cdot x$$

$$y = 5 + 10 \cdot x$$

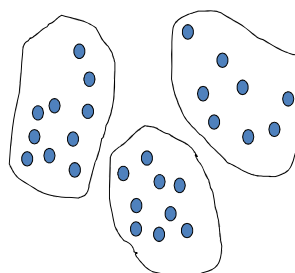
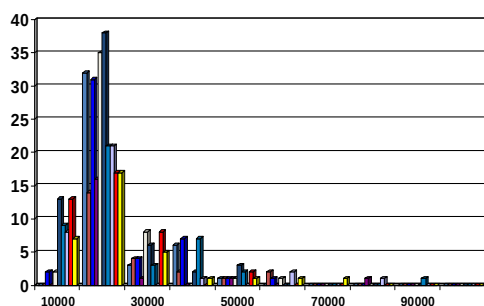
$$\{5;10\}$$

Ondřej Přibyl K611MDS

strana 56

Redukce počtu vzorků – neparametrické metody

- Neparametrické metody
 - Data není možné nahradit modelem
 - Hlavní skupiny: Histogram, shlukování, a další



Ondřej Přibyl K611MDS

strana 57

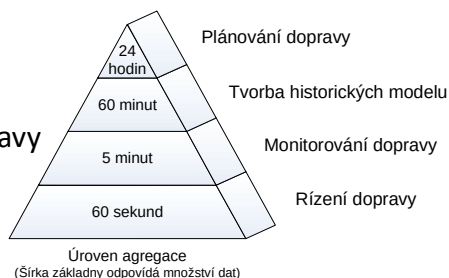
Hierarchická redukce - agregace

- Různé dopravní problémy mají různé požadavky na data

Agregace

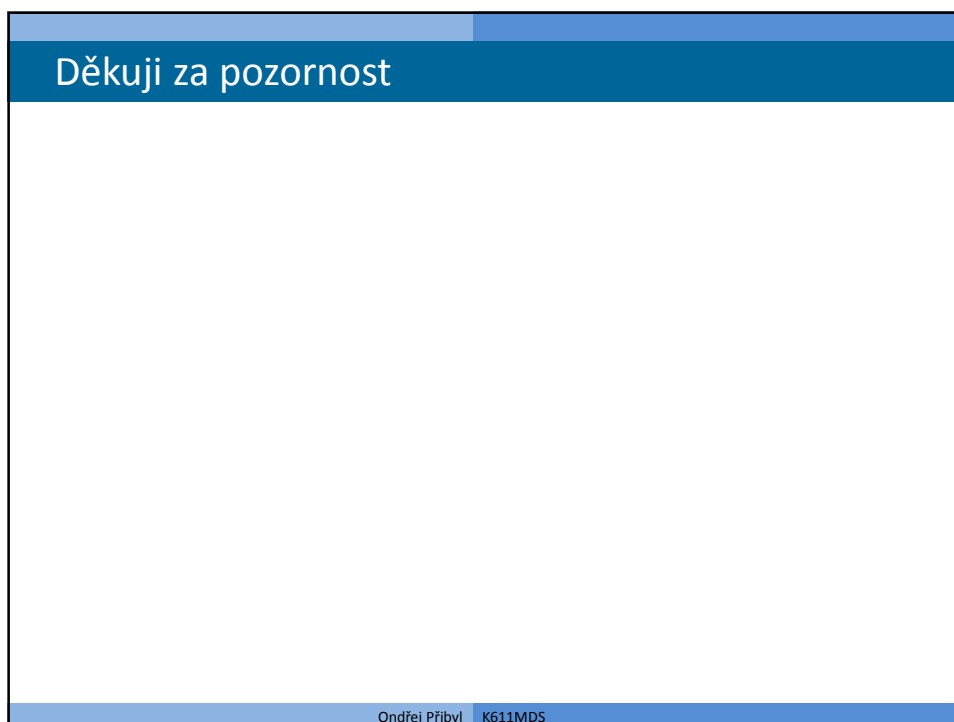
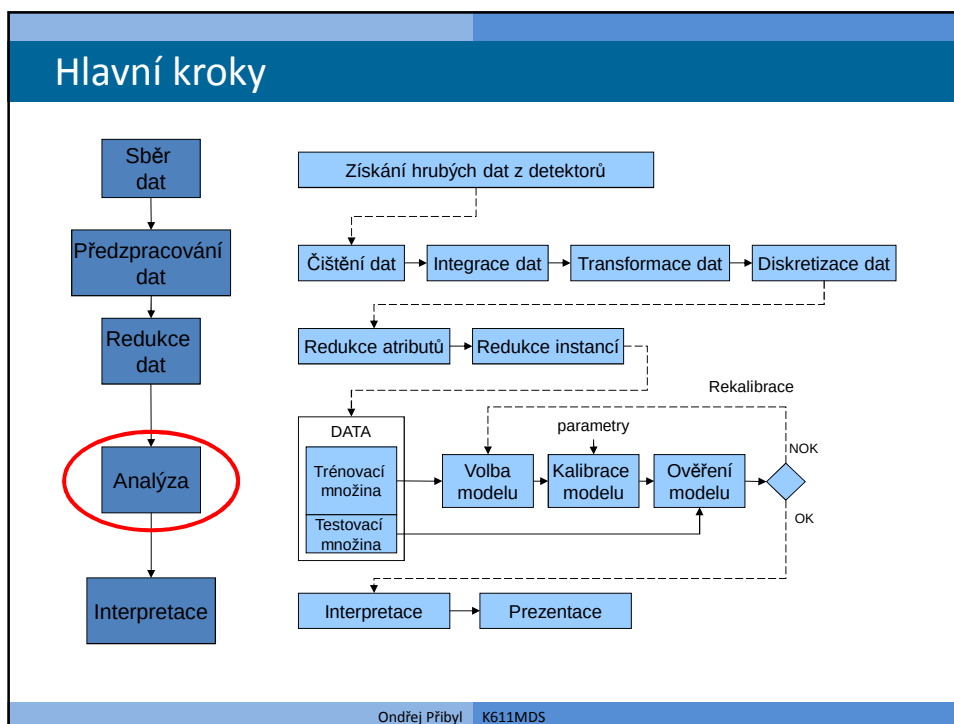
- Příklad – Měření intenzity dopravy

- Původně měřená data
individuální vozidla
- Potřeby řízení uzlu
[počet vozidel/60 sec]
- Potřeby monitorování dopravy
[počet vozidel/5min]



Ondřej Přibyl K611MDS

strana 58



Témata ke zkoušce

- Důvody a motivace pro předzpracování dat
- Metody a aplikace čištění dat
- Náhrada chybějících hodnot
- Lineární / z-score normalizace
- Filtrace dat (časová / frekvenční oblast)
- Metody redukce dat
- Korelační analýza
- Princip PCA

Zdroje:

- <http://isddc.dot.gov/OLPFiles/FHWA/010393.pdf>
- http://www.kxcad.net/cae_MATLAB/techdoc/learn_matlab/bq45sar.html
- <http://botany.upol.cz/prezentace/duch/statistika4.pdf>
- <http://botanika.bf.jcu.cz/suspa/vyuka/materialy/Statistika13.ppt#256,1,Korelace>
- <http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm>
- <http://www.public.iastate.edu/~alicia/stat328/Time%20Series.pdf>