Kvantitativní analýza internetového provozu (3)

Ladislav Lhotka
⟨*lhotka@cesnet.cz*⟩

Osnova přednášky

- Metody sběru dat – SNMP
- Časové řady
- Autokorelační funkce
- Časové řady s trendem a sezónními vlivy
- Predikce časových řad

SNMP

Simple Network Management Protocol: RFC 3416 (verze 2)

Na straně serveru (managed device) se o komunikaci stará *agent* (softwarový démon).

Zabezpečení přístupu: *SNMP communities* (read-only, read-write)

Dva režimy:

1. Dotaz-odpověď: klient/manažer požádá agenta o určitou akci, ten ji provede (nebo také ne) a pošle odpověď.
2. *Trap*: asynchronní zpráva generovaná agentem, která manažera upozorňuje na změnu stavu zařízení nebo nějaký jev.

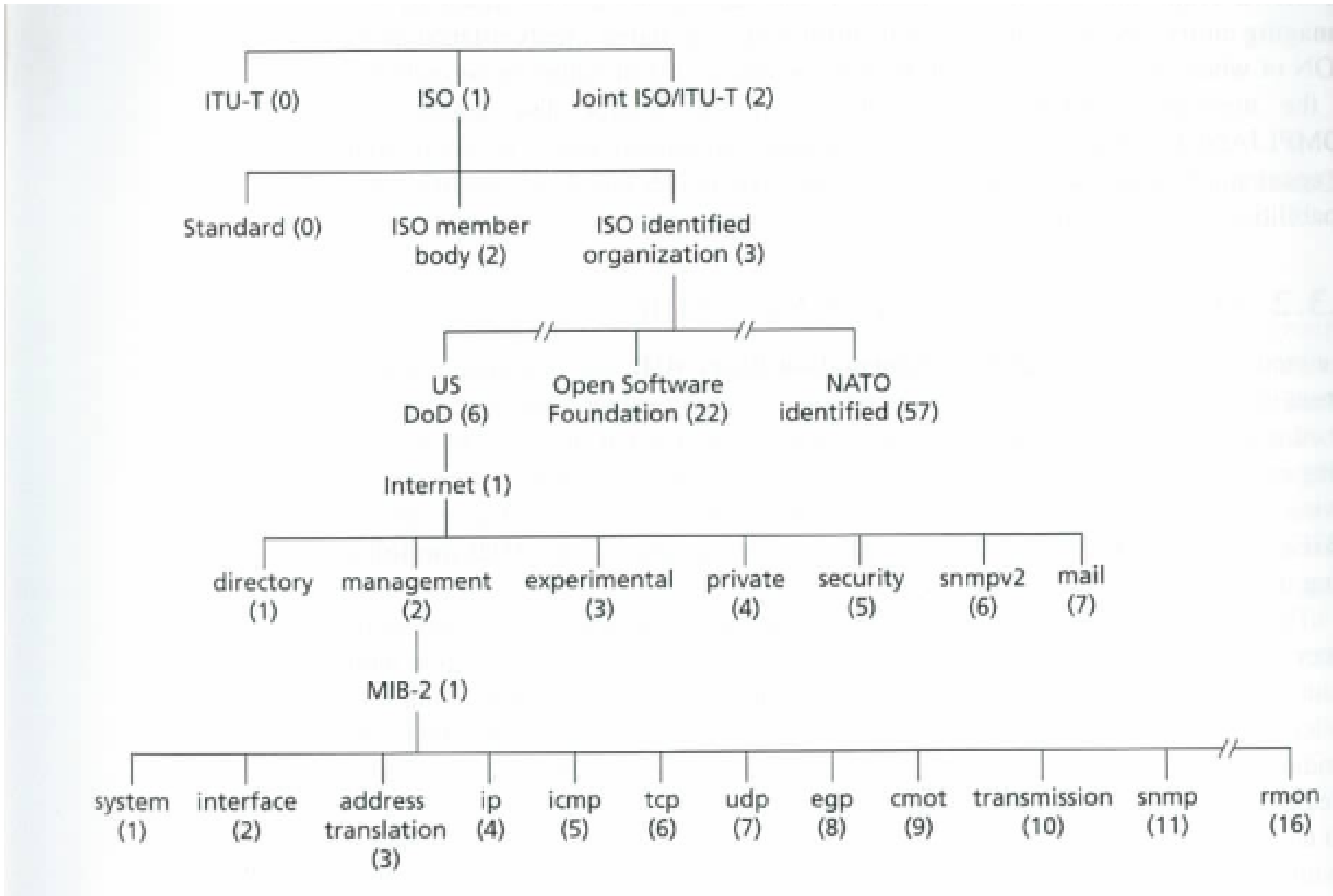
Management objects

Každý objekt reprezentuje konkrétní údaj (konfigurační parametr, statistiku) nebo jejich posloupnost.

Soubor (databáze) objektů se nazývá *Management Information Base (MIB)*.

Objekty jsou dále tématicky členěny (podle typu zařízení nebo služby apod.) do *MIB modulů*.

Struktura MIB



Dotazy na data

U všech jsou jako parametry OID objektů, jejichž hodnota se žádá.

- `GetRequest` jedna nebo více hodnot MIB objektů
- `GetNextRequest` hodnota následujícího prvku posloupnosti
- `GetBulkData` umožňuje přenést celé posloupnosti

Software pro SNMP

- Net-SNMP:
  http://www.net-snmp.org
- MRTG (Multi-Router Traffic Grapher):
  http://oss.oetiker.ch/mrtg/
- Nagios:
  http://www.nagios.org/

Časové řady

Diskrétní časová řada je posloupnost pozorování jedné nebo více náhodných veličin uspořádaná v čase:

$$\{X_{t_1}, X_{t_2}, \dots, X_{t_n}\}$$

Předpokládáme ekvidistantní časový interval – řada se pak zapisuje

$$\{X_1, X_2, \dots, X_n\}$$

Proměnná X_n obvykle závisí na předchozí historii, např. opakované hody kostkou se nepovažují za časovou řadu.

Značení

$\overline{X} = \mathrm{E}(X)$ . . . střední hodnota

$\mathrm{D}(X)$ . . . rozptyl

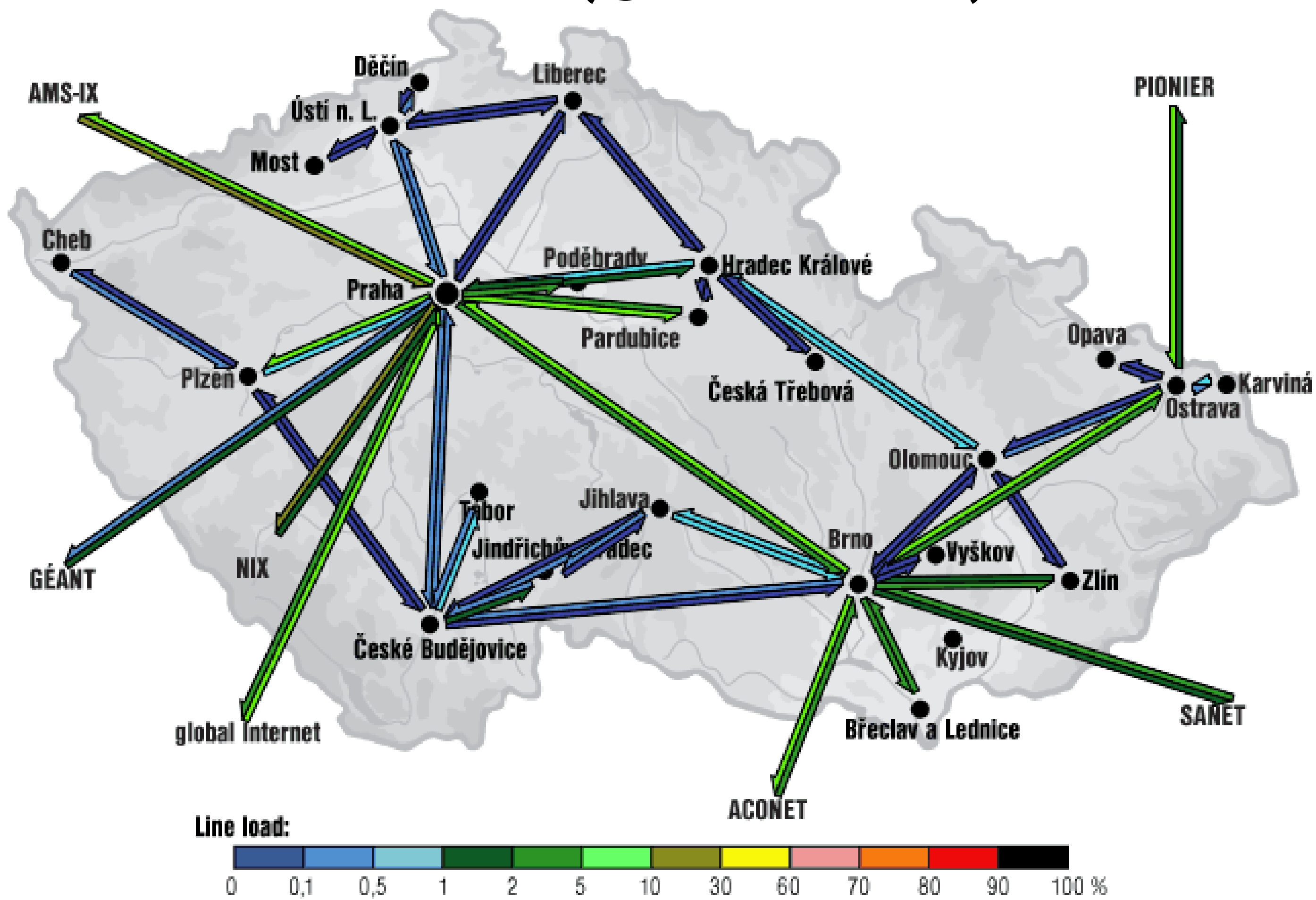
$\sigma(X)$ . . . standardní odchylka

$\mathrm{C}(X, Y)$ . . . kovariance

$\rho(X, Y)$ . . . koeficient korelace

Objemové statistiky provozu

Vzorek z mezinárodního okruhu („global Internet“)



telia-3roky.data

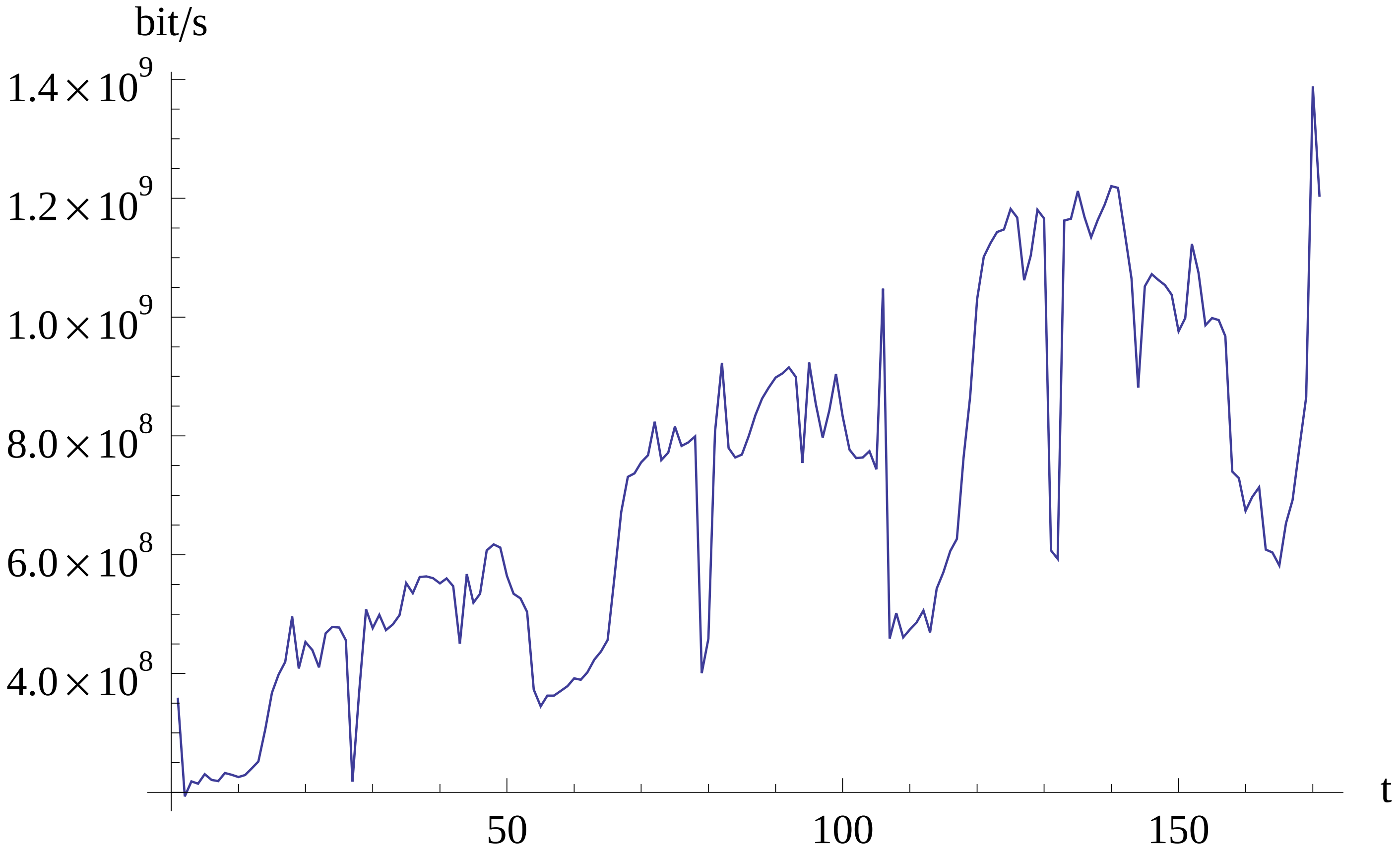
2005-06-29T20:00:00 – 2008-10-01T20:00:00, Δt = 1 týden.

```
1120089600  357446623.982743
1120694400  193276988.836032
1121299200  218319212.625317
1121904000  214722634.92688
1122508800  230670664.3727
1123113600  220978686.182301
...
```

Levý sloupec: Unix time [s]

Pravý sloupec: průměrná hodnota za $\langle t - \frac{\Delta t}{2}, t + \frac{\Delta t}{2} \rangle$ [bit/s]

Public Internet Praha – 3 years



Software pro časové řady

- R – *http://www.r-project.org*
- S-PLUS, SAS
- Mathematica, MAPLE

Stacionarita časové řady

Stacionarita 1. řádu: $E(X_t)$ nezávisí na t.

Stacionarita 2. řádu: Platí podmínka stacionarity 1. řádu a navíc $C(X_t, X_{t+s})$ závisí pouze na s.

Autokorelační funkce

Autokovarianční funkce pro řadu se stacionaritou 2. řádu:

$$\gamma_s = C(X_t, X_{t+s}) = E((X_t - \overline{X})(X_{t+s} - \overline{X}))$$

Autokorelační funkce

$$\rho_s = \frac{C(X_t, X_{t+s})}{\sigma(X_t)\sigma(X_{t+s})} = \frac{\gamma_k}{\gamma_0}$$

Odhady ACF ze vzorku

$$\hat{\rho}_s = \frac{\sum_{t=1}^{n-s}(X_t - \overline{X})(X_{t+s} - \overline{X})}{\sum_{t=1}^{n}(X_t - \overline{X})^2}$$

Složky časové řady

$$X_t = T_t + S_t + e_t$$

T_t . . . trend

S_t . . . sezónní složka

T_t . . . trend

e_t . . . zbytek

Separace složek

Buys-Ballotova tabulka

	1	2	$\dots$	T	Celkem	Průměr
	x_1	x_2	$\dots$	x_T		
	x_{T+1}	x_{T+2}	$\dots$	x_T		
	$\vdots$	$\vdots$	$\vdots$	$\vdots$		
	x_{n-T+1}	x_{n-T+2}	$\dots$	x_n		
Celkem						
Průměr						

T je perioda

Určení trendu – klouzavý průměr

$$T_t = \frac{1}{2a + 1} \sum_{i=-a}^{a} x_i$$

V *R* lze použít funkci `filter()`.

Pro naše data zkusíme klouzavý průměr přes 52 hodnot:

```
filter(bps, filter=rep(1/52,52))
```

Stroje pracují za nás

Funkce `stl()` odhaduje z časové řady X_t trend T_t pomocí regresní metody LOESS a dopočítává sezónní složku S_t a zbytek e_t.

Ze seznamu hodnot je ale potřeba vytvořit objekt „time-series“:

```
rada.6 <- ts(bps, start=0, freq=6)
plot(stl(rada.6, s.window="periodic"))
```

Predikce

Střední hodnota celé řady obvykle nefunguje.

Klouzavý průměr několika předchozích hodnot je lepší.

```
filter(bps, sides=1, filter=rep(1/3,3))
```

Exponenciální vyhlazování

Předpovídaná hodnota

$$\hat{X}_{t+1} = \hat{a}_t$$

$$\begin{aligned}\hat{a}_t &= \alpha X_t + (1-\alpha)\hat{X}_t \\ &= \alpha X_t + (1-\alpha)\hat{a}_{t-1}\end{aligned}$$

Váha předchozích pozorování α (smoothing parameter) se zmenšuje směrem do minulosti:

$$\hat{a}_t = \alpha X_t + \alpha(1-\alpha)X_{t-1} + (1-\alpha)^2\hat{a}_{t-2} = \dots$$

Počáteční hodnotu můžeme nastavit např. takto: $\hat{a}_0 = X_1$.

Holtova metoda – lineární trend

$$\hat{X}_{t+1} = \hat{a}_t + \hat{b}_t \Delta t$$

Zvolíme Δt za časovou jednotku.

$$\begin{aligned}
\hat{a}_t &= \alpha X_t + (1 - \alpha)(\hat{a}_{t-1} + \hat{b}_{t-1}) \\
\hat{b}_t &= \beta(\hat{a}_t - \hat{a}_{t-1}) + (1 - \beta)\hat{b}_{t-1}
\end{aligned}$$

Wintersova metoda – sezónní variace

Aditivní model

$$\hat{X}_{t+1} = \hat{a}_t + \hat{b}_t \Delta t + \hat{s}_t$$

$$\hat{a}_t = \alpha X_t + (1 - \alpha)(\hat{a}_{t-1} + \hat{b}_{t-1})$$
$$\hat{b}_t = \beta(\hat{a}_t - \hat{a}_{t-1}) + (1 - \beta)\hat{b}_{t-1}$$
$$\hat{s}_t = \gamma(X_t - \hat{a}_t) + (1 - \gamma)\hat{s}_{t-p}$$

Multiplikativní model

$$\hat{X}_{t+1} = (\hat{a}_t + \hat{b}_t \Delta t)\hat{s}_t$$

$$\hat{a}_t = \alpha X_t + (1 - \alpha)(\hat{a}_{t-1} + \hat{b}_{t-1})$$
$$\hat{b}_t = \beta(\hat{a}_t - \hat{a}_{t-1}) + (1 - \beta)\hat{b}_{t-1}$$
$$\hat{s}_t = \gamma(X_t/\hat{a}_t) + (1 - \gamma)\hat{s}_{t-p}$$