Přednáška 5 – Model směsi Poissonových komponent

Poissonovy komponenty $f_j(y_t|\lambda)$

$$f_j(y_t|\lambda) = \exp\{-\lambda\}\frac{\lambda^{y_t}}{y_t!}$$

$j \in \{1, \ldots, n_c\}$, n_c – počet komponent

Ukazovátko ukazuje aktivní komponentu

$c_t \in \{1, 2, \ldots, n_c\}$

Příklad:

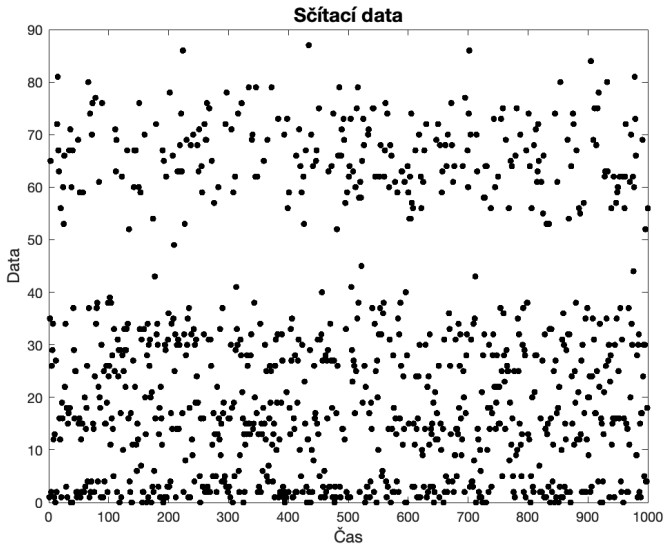
Komponenta 1: $f_1(y_t|\lambda)$, $\lambda = 2$

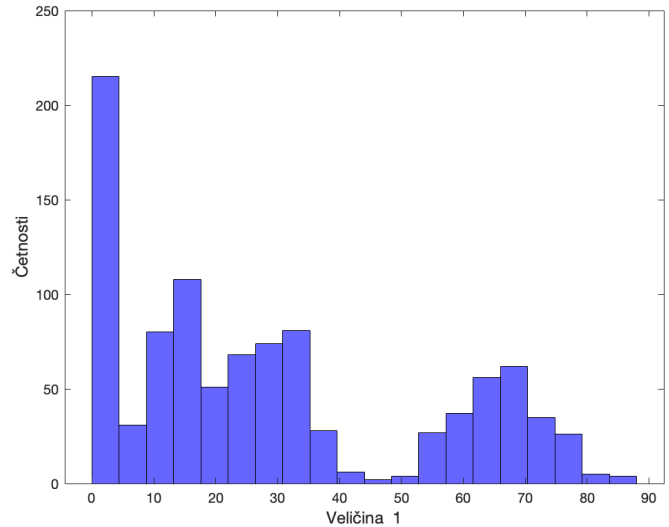
Komponenta 2: $f_2(y_t|\lambda)$, $\lambda = 15$

Komponenta 3: $f_3(y_t|\lambda)$, $\lambda = 30$

Komponenta 4: $f_4(y_t|\lambda)$, $\lambda = 60$

- Kolik hodnot má ukazovátko?
- Model ukazovátka?

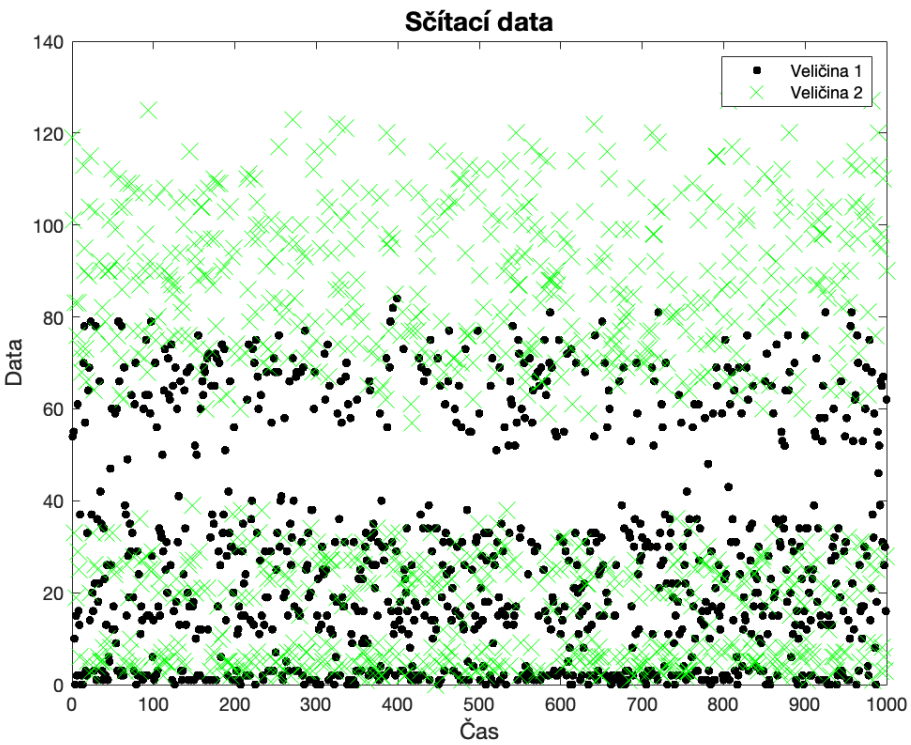


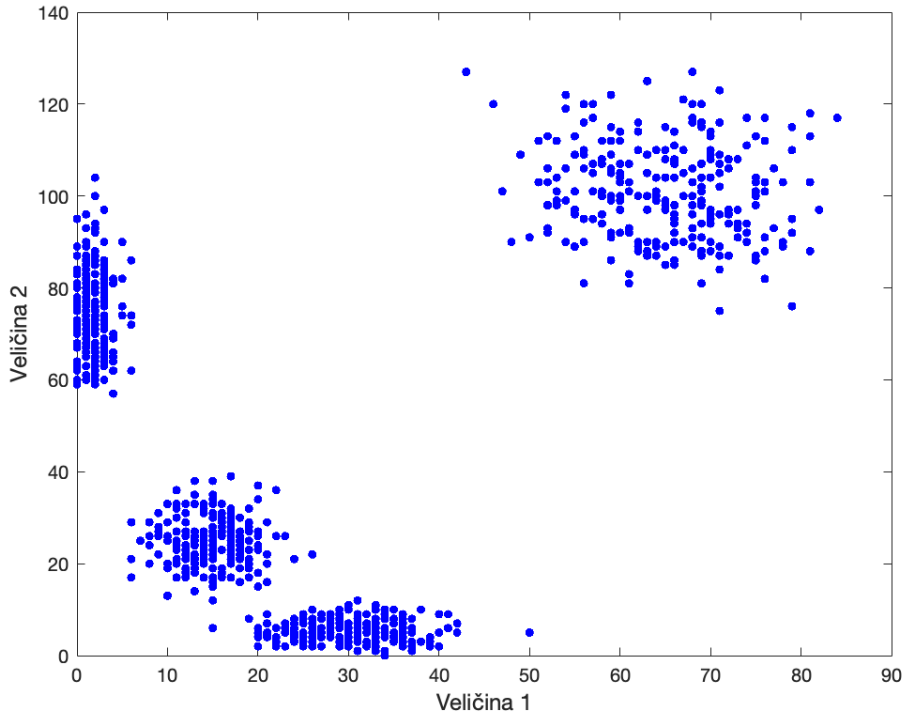


Program – generování multimodálních sčítacích dat

```
clear, clc, close
% Generování multimodálních sčítacích dat z modelu směsi
% Poissonových komponent (2 nezávislé veličiny)

nd = 1000;  % počet dat

La{1} = [2; 15; 30; 66]; % parametry komponent y1
La{2} = [75; 25; 5; 101]; % parametry komponent y2

ny = numel(La);  % počet sčítacích veličin

% Simulace
c = zeros(1, nd); % počáteční podmínky
y = zeros(ny, nd);

for t = 1:nd
    cum_prob = cumsum([0.25, 0.25, 0.25, 0.25]); % rovnoměrný model ukazovátka
    c(t) = sum(rand(1, 1) > cum_prob) + 1;  % generování ukazovátka

    for i = 1:ny
        y(i, t) = poissrnd(La{i}(c(t)));  % generování sčítacích veličin
    end
end
```

Simulovaná sčítací data





- Co ovlivní vzdálenost mezi shluky?

Rekurzivní Bayesovský odhad modelu směsi distribucí

- Shlukování – odhad parametrů komponent Θ z průběžně měřených dat
  (identifikujeme shluky)
- Klasifikace – třídíme data do odhadnutých komponent
  (odhad ukazovátka c_t v čase t)

Základní princip – obecně:

- Inicializace komponent

V cyklu:

- Měříme data
- Vzdálenosti od jednotlivých komponent – proximity
- Normalizace proximit – váhy komponent
  (pravděpodobnost, že je komponenta aktivní)
- Maximální váha – bodový odhad ukazovátka
  (klasifikace)
- Update statistik s váženými daty
- Přepočet bodových odhadů podle typu komponent
- Jdeme na krok 1

Pro čas $t = 0$

1. Nastavíme počet komponent + jejich počáteční statistiky
2. Vypočteme počáteční bodové odhady parametrů $(\hat{\lambda}_0)_j$

Pro čas $t = 1, 2, \ldots$, pro každou komponentu

1. Měříme nová data y_t
2. Určíme proximity m_j – dosadíme do komponent $(\hat{\lambda}_{t-1})_j$ a y_t
3. Určíme váhy komponent

$$w_{j;t} = \frac{m_j}{\sum_{i=1}^{n_c} m_i}$$

4. Bodový odhad ukazovátka: $\hat{c}_t = \arg\max_j w_t$
5. Update statistik s váženými daty:

$$(S_t)_j = (S_{t-1})_j + w_{j;t} y_t, \quad (\kappa_t)_j = (\kappa_{t-1})_j + w_{j;t},$$

6. Přepočet bodových odhadů $(\hat{\lambda}_t)_j = \frac{(S_t)_j}{(\kappa_{t-1})_j}$
7. Jdeme na krok 1

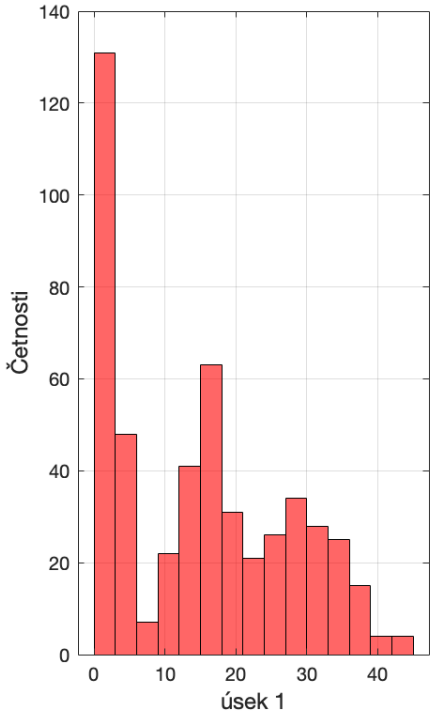
Příklad: klasifikace počtu vozidel online - dva úseky

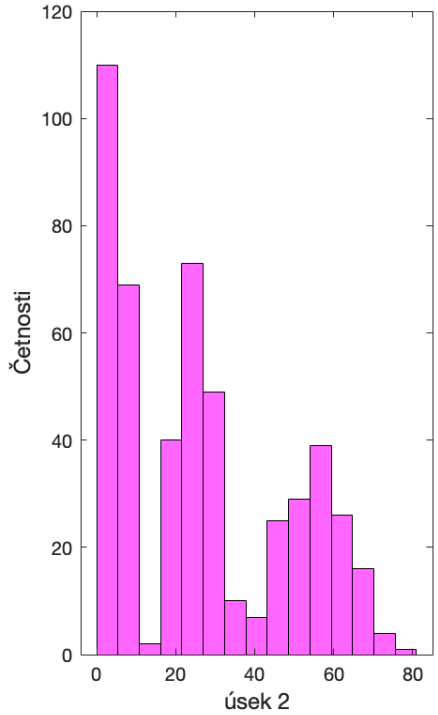
$y_t = [y_{1;t}; y_{2;t}]$ – počty vozidel
t – minuty

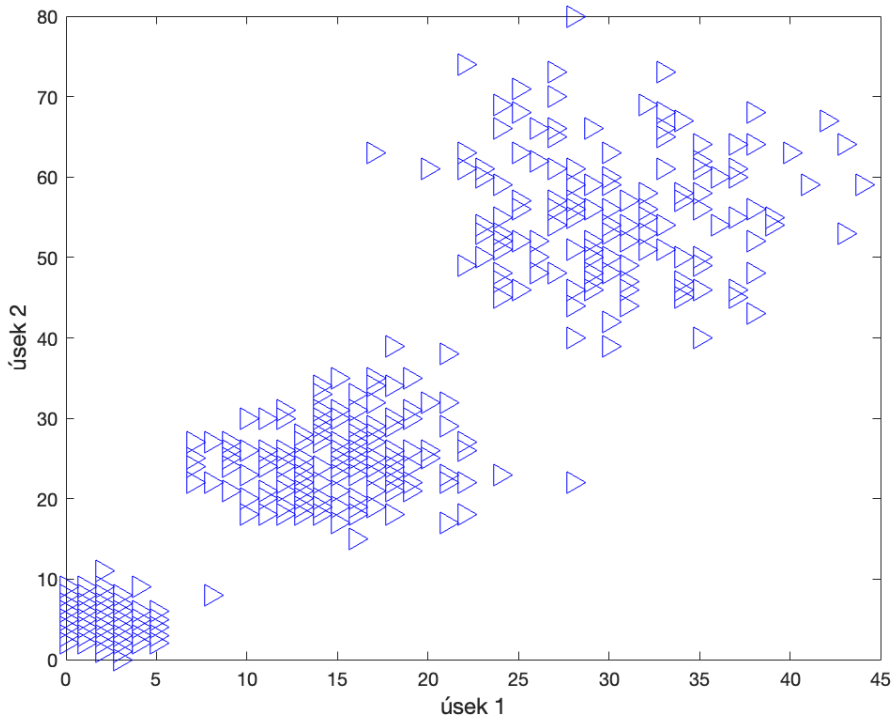
$c_t \in \{1, 2, 3\}$ – dopravní špička
(noc, dopoledne, odpoledne)

Inicializace:

- Počet komponent – podle histogramu
- Počáteční bodové odhady – expertně, histogram, x-y graf







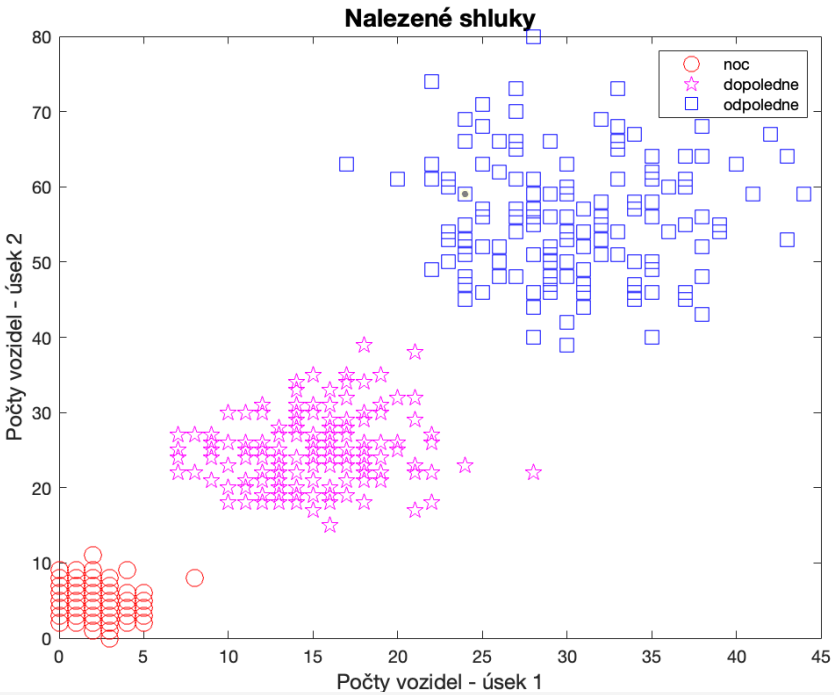
Program - inicializace

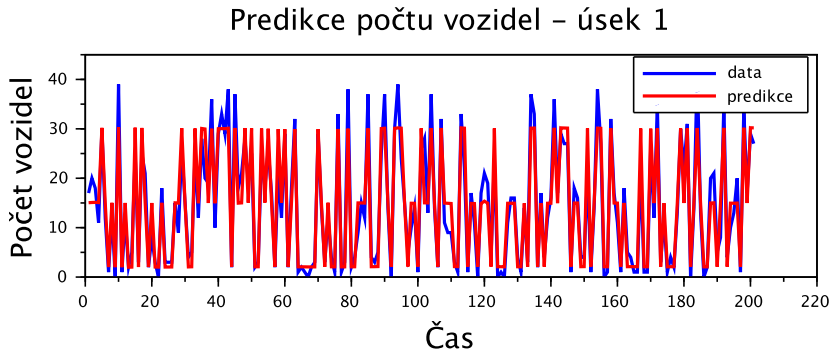
```
nc = 3; % Počet komponent - expertně nebo z histogramu (počet kopečků)

% Počáteční bodové odhady  histogramu nebo z jednotlivých kopečků
% - expertně (např. noc, dopoledne, odpoledne)
LaE{1} = [2; 15; 30]; % Pro úsek 1
LaE{2} = [5; 30; 60]; % Pro úsek 2


ka{1} = ones(nc, 1); % Počáteční počítadlo pro úsek 1
ka{2} = ka{1}; % Pro úsek 2

% Počáteční statistiky
for i = 1:ny
    S{i} = LaE{i} .* ka{i};
end
```

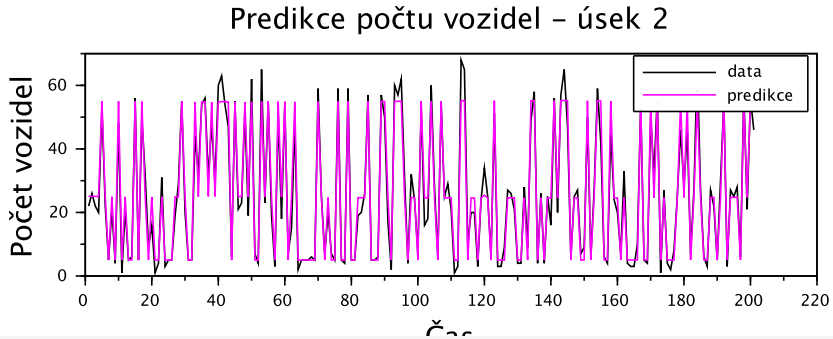
Program - shlukování a klasifikace

```
% Odhad, shlukování a klasifikace online
for t = 1:nd

    for i = 1:nc
        mm1(i) = poisspdf(y(1, t), LaE{1}(i)); % Proximity pro úsek 1
        mm2(i) = poisspdf(y(2, t), LaE{2}(i)); % Pro úsek 2
    end
    m = mm1 .* mm2;
    w = m ./ sum(m); % Váhy
    [~, cp(t)] = max(w); % Bodový odhad ukazovátka - klasifikace

    % Update statistik
    for j = 1:ny
        for i = 1:nc
            S{j}(i) = S{j}(i) + w(i) * y(j, t);
            ka{j}(i) = ka{j}(i) + w(i);
        end
        LaE{j} = S{j} ./ ka{j}; % Bodové odhady
    end
```

Výsledky a validace klasifikace








```
    % Validace - predikce z komponent s odhady
    bb = [0; 0];
    for j = 1:nc
        for i = 1:ny
            bb(i) = bb(i) + w(j) * LaE{i}(j); % Vážený průměr z komponent
        end
    end
    yp(:, t) = bb; % predikce
end

% Výsledky
RMSE=sqrt(mean((y' - yp').^2)) % Chyba predikce (pro sčítací data jako pro spojitá)
```

Poissonova regrese – predikce sčítacích dat

y_t – Poissonovo rozdělení

$$f(y_t|\lambda) = \exp\{-\lambda\}\frac{\lambda^{y_t}}{y_t!}$$

Poissonova regrese

$$\ln(\underbrace{\bar{y}_t}_{\lambda}) = x_{1;t}\theta_1 + x_{2;t}\theta_2 + \ldots + k$$

λ = průměr = rozptyl

$x_t = [x_{1;t} \quad x_{2;t} \quad \ldots \quad x_{n;t}]$ – vysvětlující data

Linearizace + nejmenší čtverce

metoda 1

$$\underbrace{\begin{bmatrix} \ln(y_1) \\ \ln(y_2) \\ \ldots \\ \ln(y_{nd}) \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} x_{1;1} & x_{2;1} & \ldots & 1 \\ x_{1;2} & x_{2;2} & \ldots & 1 \\ \ldots & & & \\ x_{1;nd} & x_{2;nd} & \ldots & 1 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \theta_1 \\ \theta_2 \\ \ldots \\ k \end{bmatrix}}_{\theta}$$

Odhad offline:

trénovací data

$$\hat{\theta} = (X'X)^{-1}X'Y$$

Predikce online:

testovací data

$$\hat{Y} = \exp\{X\hat{\theta}\}$$

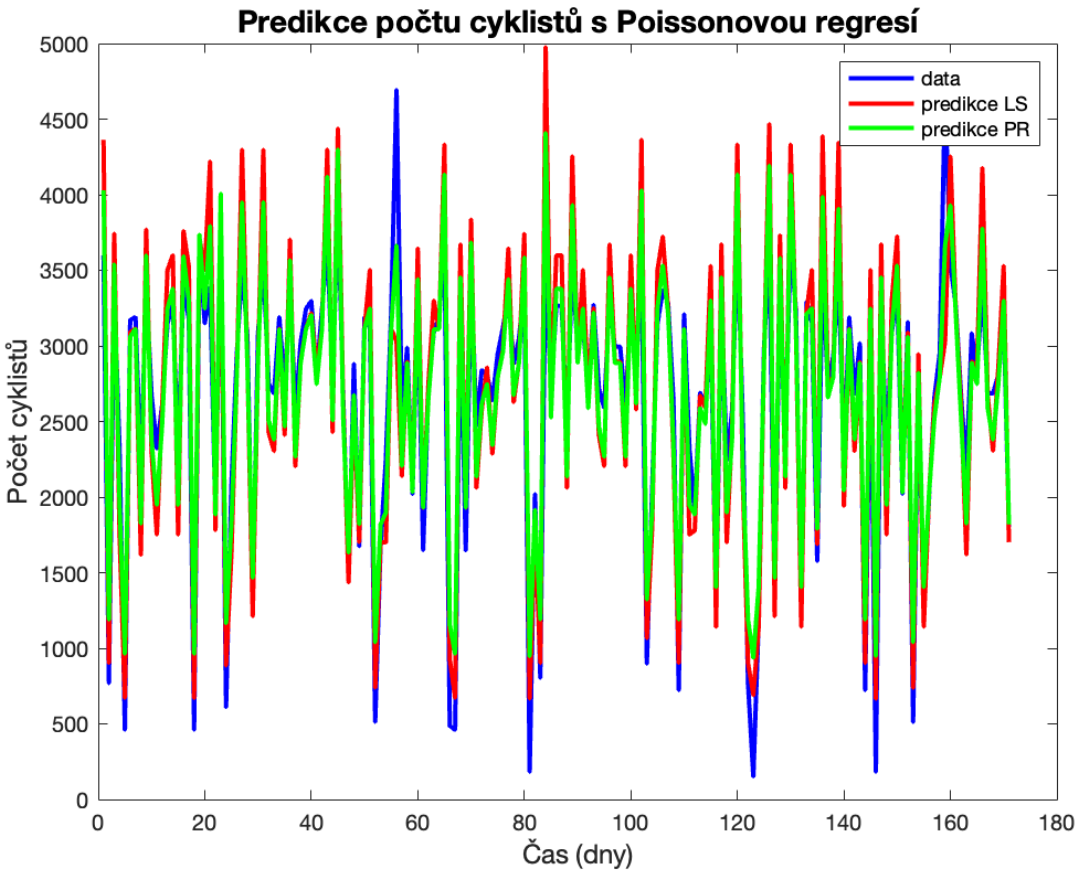
Metoda maximální věrohodnosti

metoda 2

$$\ln L_t(\theta) = \ln \prod_{\tau=1}^{t} \text{model}(\text{trénovací data}_\tau), \quad \hat{\theta} = \arg\max_{\theta} \ln L_t(\theta)$$

Příklad – predikce počtu cyklistů na Brooklynském mostě

y_t – počty cyklistů
$x_{1;t}$ – nejvyšší teploty (°F)
$x_{2;t}$ – nejnižší teploty (°F)
$x_{3;t}$ – srážky (mm)
t – dny



Odkaz na zdroj

Program – Linearizace + nejmenší čtverce

```
y=data(:,4); % Počty cyklistů
x=data(:,1:3); % Nejvyšší a nejnižší venkovní teploty (F) a srážky (mm)

nd=size(y,1); % Počet dat

ndTr=round(nd/100*80); % Počet trénovacích dat - 80%

yL=y(1:ndTr);xL=x(1:ndTr,:); % Trénovací data

yT=y((ndTr+1):end);xT=x((ndTr+1):end,:); % Testovací data


% Odhad s trénovacími daty (linearizace a nejmenší čtverce)
Y=log(yL);
X=[xL ones(size(xL,1),1)];
pom=X'*X+0.0000001*eye(4,4);
thetaLS=(pom\X')*Y; % Odhad regresních koeficientů inv(pom)*X'*Y;

% Testování s testovacími daty
XT=[xT ones(size(xT,1),1)];
ypLS=exp(XT*thetaLS);
```

Program – Metoda maximální věrohodnosti

```
% % Odhad s trénovacími daty (maximální věrohodnost)
b0=[.1 .1 0.1 1]'; % počáteční parametry pro optimalizaci
X=[xL ones(size(xL,1),1)];
Y=yL;
[nill,thetaPR]=poiReg(X,yL,b0);

% Testování s testovacími daty
XT=[xT ones(size(xT,1),1)];

yPR=exp(XT*thetaPR);
RMSE = sqrt(mean((yT - yPR).^2))
```

Poznámky

- reálná data – nadměrně/nedostatečně rozptýlená – velká chyba predikce
- Vhodné modely:
  - negativní binomická regrese
  - generalizovaný Poissonův model
  - směs Poissonových regresí