Model Generation and its Applications in Financial Sector

Vadim STRIJOV
Computing Center of RAS

October 5th, 2009 at the Institute of Applied Mathematics, METU

Russian Academy of Sciences

joins the national academy of Russia and a network of scientific research institutes from across the Russian Federation as well as scientific and social units.

- **Founded in 1724 by decree of Emperior Peter I the Great**
  - 470 institutions
  - 55,000 researchers
  - 16 Nobel laureates

- Section of Applied Mathematics and Informatics,
  - **Computing Center**

Computing Center of RAS

- **Founded in 1955**
- **Fields of the scientific research**
  - computational methods
  - mathematical modeling
  - mathematical methods of pattern recognition
- **276 researchers**
  - **8** academicians and corresponded members of RAS
  - **75** researchers have DSc degree
  - **136** researchers have PhD degree

The scientific principal is acad. Yuri I. ZHURAVLEV

Data mining

is a collection of methods for extracting

- unexplored,
- nontrivial,
- useful,
- and interpretable

patterns, models and facts from the data.

Data mining is important to support decisions in various fields of science, economics and finance.

Supervised learning

- Regression (forecasting)
- Classification
- Parameter estimation

Non-supervised learning

- Clustering
- Association rule learning
- Visualizing

Main Themes

- Regression analysis, introduction
- European options from the data mining point-of-view
- Model selection principles

Regression analysis

$$\mathrm{E}(y \mid \mathbf{x}) = f(\mathbf{w}, \mathbf{x})$$

$$y = f(\mathbf{w}, \mathbf{x}) + \varepsilon$$

Data generation hypothesis, example

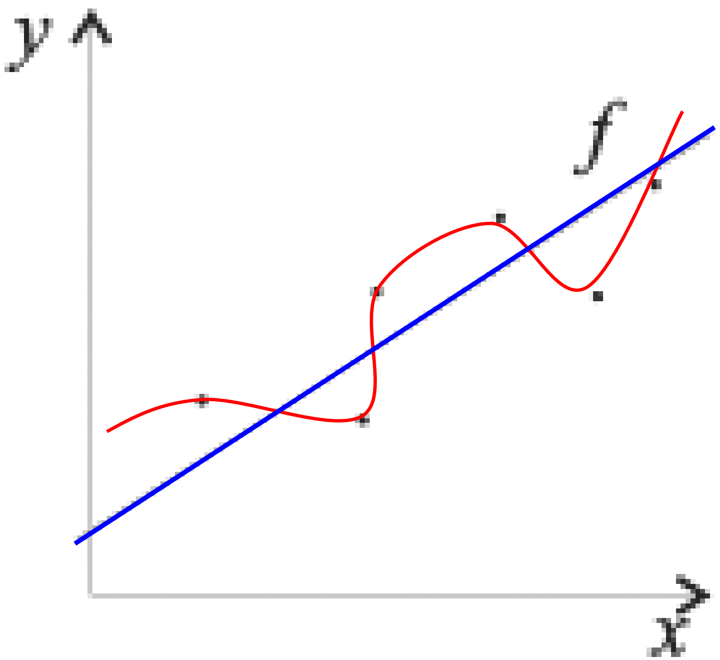
$$\varepsilon \sim N(0, \sigma_\varepsilon^2)$$

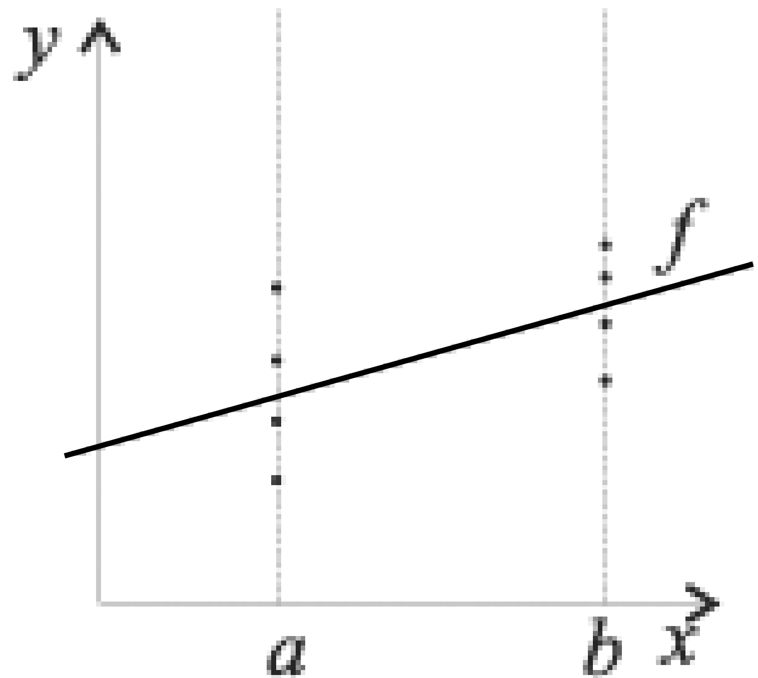
$$p(\varepsilon \mid \eta) = h(\varepsilon) g(\eta) \exp\{\eta u(\varepsilon)\}$$

Two data sets for regression

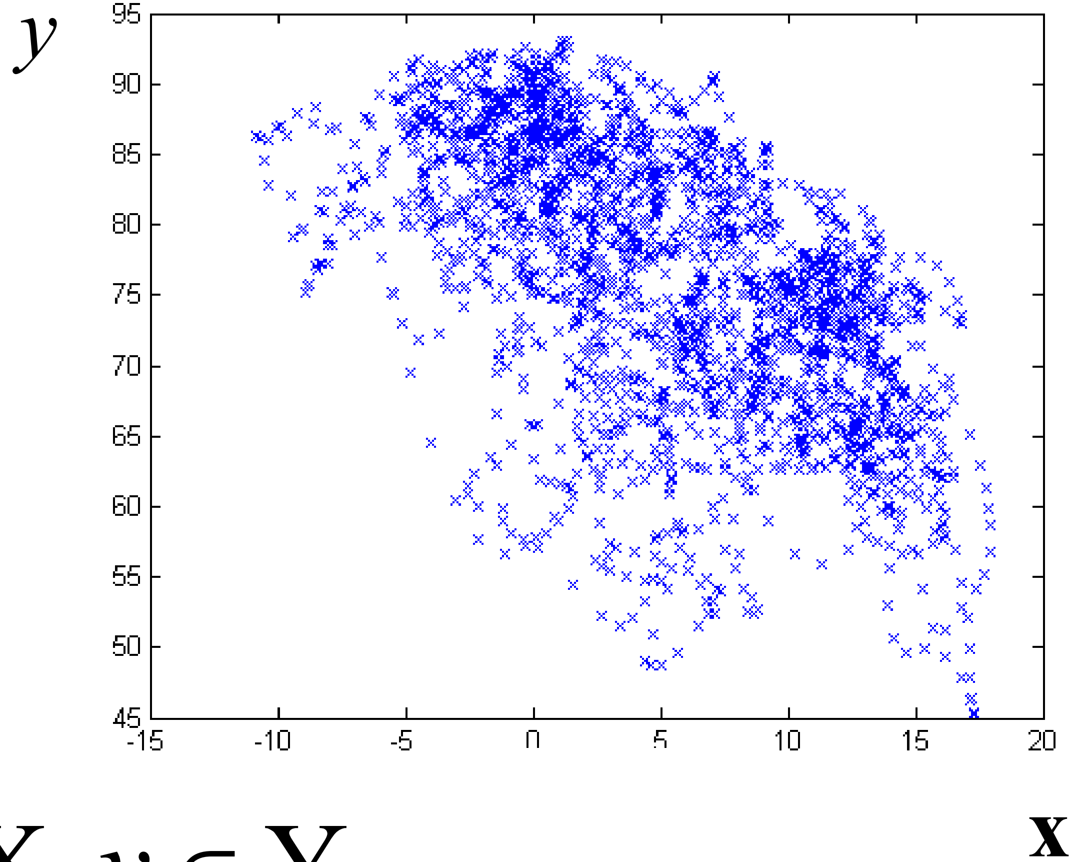
- Interpolation
- Approximation

- Regression





“Typical” data for regression modeling



$$\mathbf{x} \in \mathrm{X}, y \in \mathrm{Y}$$

Regression model

is a parametric family of functions.

$$f : \mathrm{W} \times \mathrm{X} \to \mathrm{Y}$$

$$f : \Re^{W} \times \Re^{X} \to \Re^{1}$$

$$y = f(\mathbf{w}, \mathbf{x}) + \varepsilon$$

$$f \,|_{\mathbf{w}=\mathbf{w}'} : \mathrm{X} \to \mathrm{Y}$$

Linear regression

$$y = f(\mathbf{w},\mathbf{x}) + \varepsilon = \sum_{j=1}^{W} w_j g_j(x^{(j)}) + \varepsilon = \langle \mathbf{w}, \mathbf{g}(\mathbf{x}) \rangle + \varepsilon$$

$$D = \{(\mathbf{x}_i, y_i) \mid i = 1,\dots,m\}$$

$$G = \{g_1, g_2\} = \{x^0, \mathrm{id}\}$$

$$y_i = w_1 + w_2 x_i$$

$$g_1(x_i) = x_i^0 \mapsto 1$$

$$g_2(x_i) = x_i^1 \mapsto x_i$$

$$X = \begin{pmatrix} g_1(x_1) & g_2(x_1) \\ \dots & \dots \\ g_1(x_m) & g_2(x_m) \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_m \end{pmatrix}$$

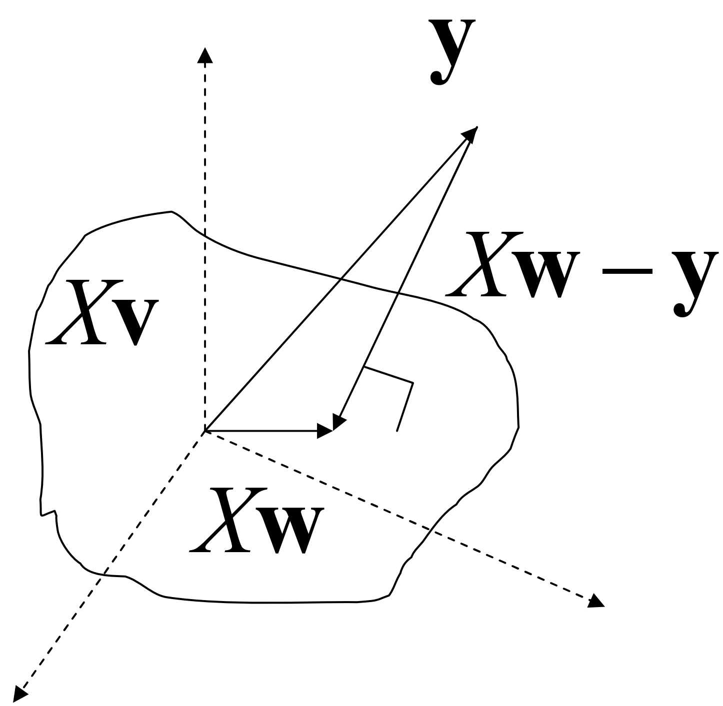
Normal equation

$$SSE = \|X\mathbf{w} - \mathbf{y}\|_2^2 \to \min$$

$$(X\mathbf{v})^T (X\mathbf{w} - \mathbf{y}) = 0$$

$$X^T X\mathbf{w} - X^T\mathbf{y} = 0$$

$$\mathbf{w} = (X^T X)^{-1} X^T \mathbf{y}$$

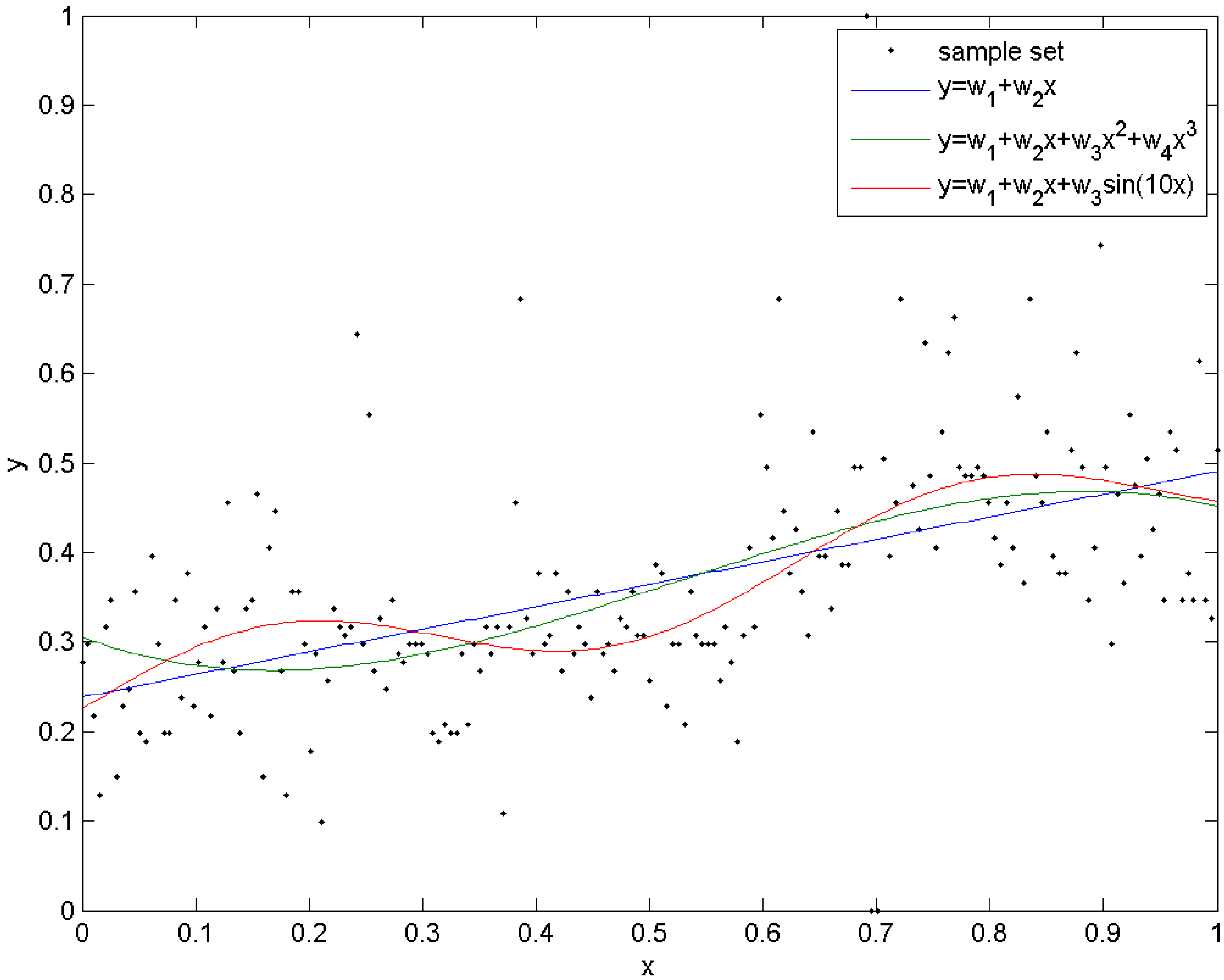


Three models

```
X = [x.^0, x]; % matrix of substitutions

f = inline('[x.^0, x,]','x');                % f1
f = inline('[x.^0, x, x.^2, x.^3]','x');% f2
f = inline('[x.^0, x, sin(10*x)]','x'); % f3

X = f(x);          % matrix of substitutions
w = (X'*X)\(X'*y); % solve normal equation
yr = X*w;          % recover dependent variable
r = y-y3;          % residual vector
SSE = r'*r;        % sum squared errors
```


sample set
y=w1+w2x
y=w1+w2x+w3x2+w4x3
y=w1+w2x+w3sin(10x)
y
x
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Questions of regression analysis

- How to choose a family of models?
- How to select a model from the family?
- What is the data generation hypothesis?
- How to set the target function?
- How to tune the model parameters?

European option

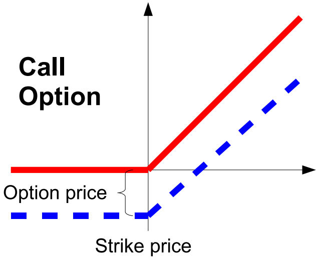
- The option is an instrument that conveys the right, but not the obligation, to engage in a future transaction on some underlying security.

- European option is an option that may only be exercised on expiration.

European option

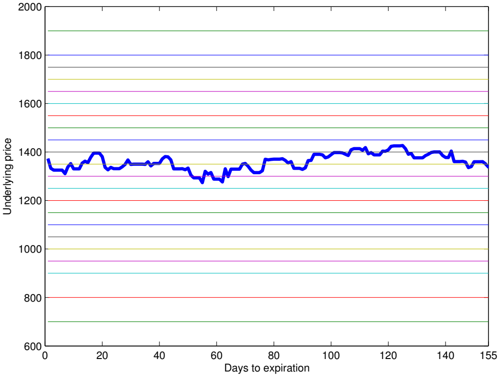
$$C_t = F(\sigma, P, B, K, t),$$

C_t — option price,
σ — volatility,
P — price of security,
B — risk-free rate,
K — strike price,
t — time to expiration.



$$C_t = \mathcal{N}\left(\frac{ln(\frac{P}{K}) + t(B + \frac{\sigma^2}{2})}{\sigma\sqrt{t}}\right) - Ke^{-Bt}\mathcal{N}\left(\frac{ln(\frac{P}{K}) + t(B - \frac{\sigma^2}{2})}{\sigma\sqrt{t}}\right)$$

Historical price of security

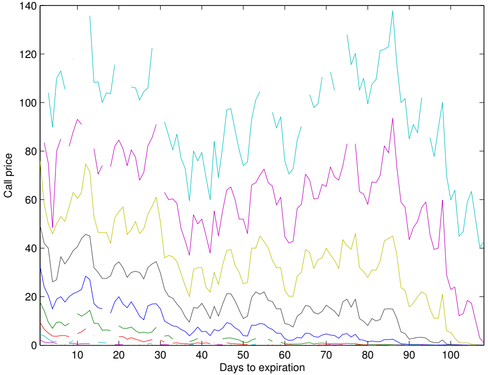


t – time to expiration, years,

P — security price.

Horizontal lines correspond to strike prices K.

Historical prices of options K



t — time to expiration, years,
C — option price.

How to calculate the volatility?

Volatility most frequently refers to the standard deviation of the returns of a financial instrument. It is often used to quantify the risk of the instrument over a time period.

Implied volatility of an option is the volatility implied by the market price of the option based on an option pricing model.

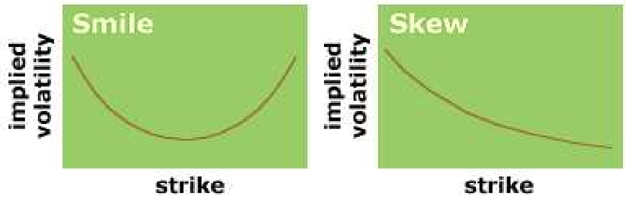
$$\sigma^{\text{imp}} = \arg\min_{\sigma}(C_{\text{hist}} - C(\sigma, P, B, K, t)).$$

We consider implied volatility as the dependent variable of the regression model.

Our knowledge about volatility helps us to estimate the risk of capital investments.

Implied volatility

The implied volatility depends on the time t and strike price K.



Volatility model, given by experts

A model for traders at the Russian trade system

$$\sigma = \sigma(\mathbf{w}) = w_1 + w_2(1 - \exp(-w_3 x^2)) + \frac{w_4 \arctan(w_5 x)}{w_5},$$

$$\text{где} \quad x = \frac{\log(K) - \log(C(t))}{\sqrt{t}}.$$

Model assumptions [Daglish, 2006]

- The volatility depends on the option price.
- The volatility proportional to inverse square root of the maturity.

Historical prices

K= 13.50, 13.00, 12.50, 12.00, 11.50, 11.00

Maturity	K1	K2	K3	K4	K5	K6	Price
-91	0.105	0.16	0.24	0.36	0.56	0.725	11.27
-90	0.105	0.16	0.24	0.35	0.56	0.725	11.29
-87	0.105	0.16	0.21	0.36	0.56	0.725	11.34
-86	0.105	0.16	0.21	0.32	0.56	0.725	11.2
-85	0.105	0.16	0.21	0.32	0.48	0.725	11.18
-84	0.105	0.16	0.215	0.33	0.625	0.725	11.5
-83	0.105	0.16	0.22	0.41	0.625	0.725	11.41
-80	0.105	0.16	0.25	0.42	0.69	0.885	11.48
…	…	…	…	…	…	…	…

Given data

$t \in \{t_1, \ldots, t_\tau, \ldots, t_{64}\} = \mathrm{T}$ set of time ticks

$K \in \{K_1, \ldots, K_k, \ldots, K_6\} = \mathbf{K}$ set of strike prices

$C = C(t, K)$ historical option prices

$P = P(t)$ historical security prices

The desired model

$$\sigma = f(t, K)$$

Index mapping

Implied volatility

$$\sigma_{t,K} = \arg\min_{\sigma}(C_{t,K}^{hist} - C(\sigma, P_t, B, K, t))$$

Sample set for regression analysis

$$\sigma_{t,K} \mapsto \sigma_i, i = \tau + k(|\mathbf{T}|-1)$$

$$(t_i, K_i) \in \mathbf{T} \times \mathrm{K}$$

The regression model

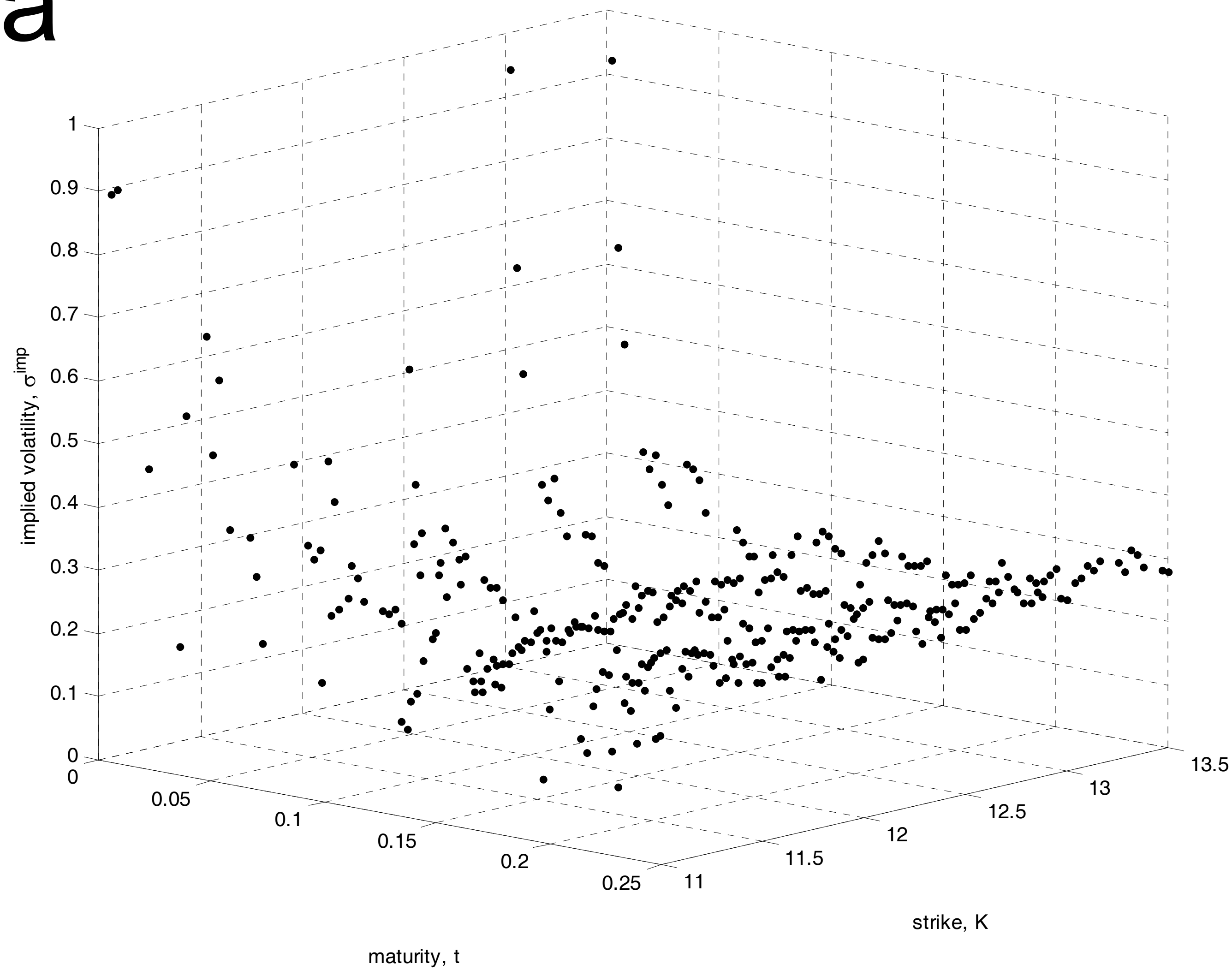
$$\sigma_i = f(t_i, K_i)$$

Volatility models, toy version

$$f_1 = w_0 + w_1 t^2 + w_2 tK + w_3 K^2$$

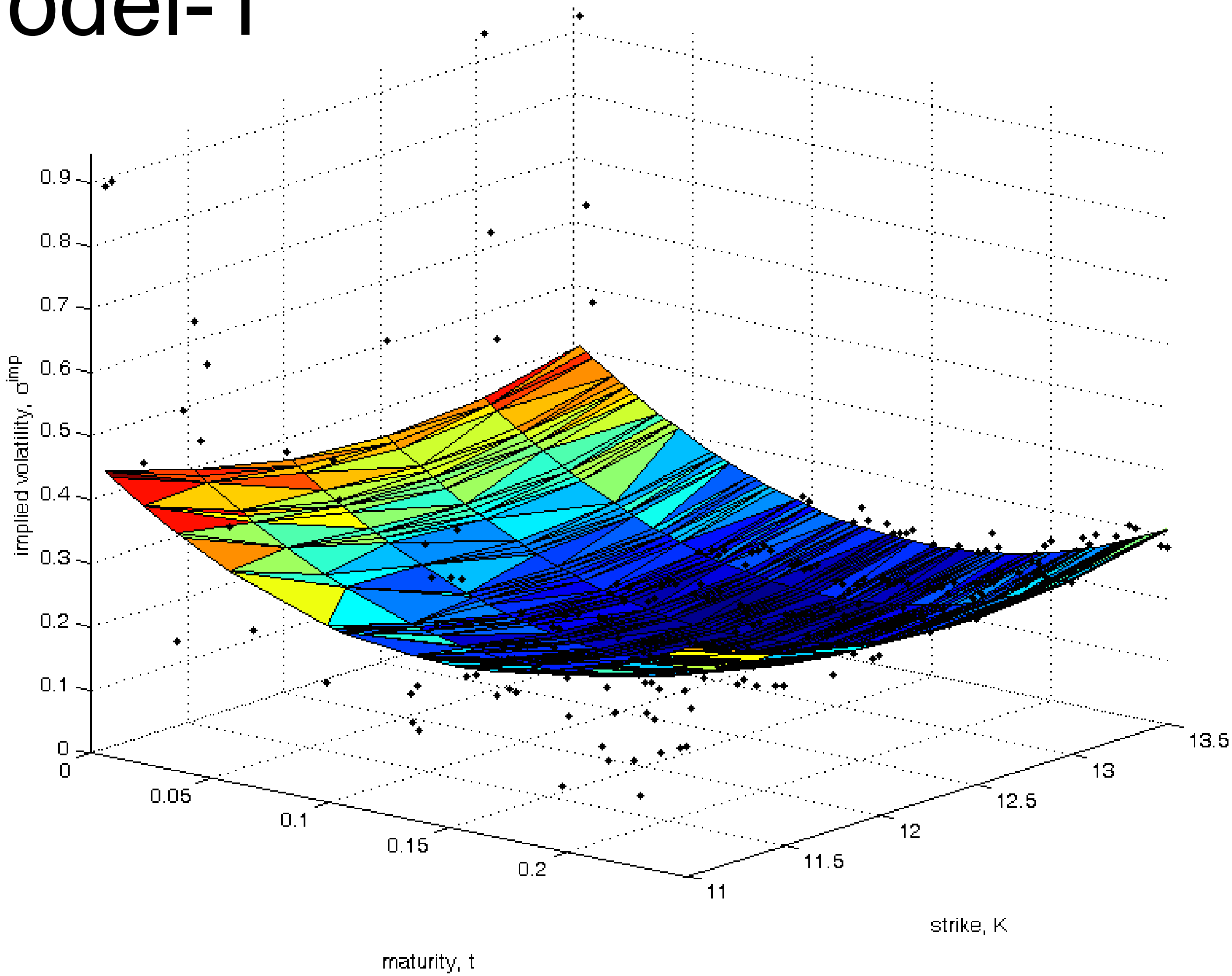
$$f_2 = w_0 + w_1 t^2 + w_2 K^2 + w_3 \frac{\sqrt{K}}{1+\exp(t)} + w_4 \frac{(\exp(t)\sqrt{t})\sqrt{K}}{K}$$

Data



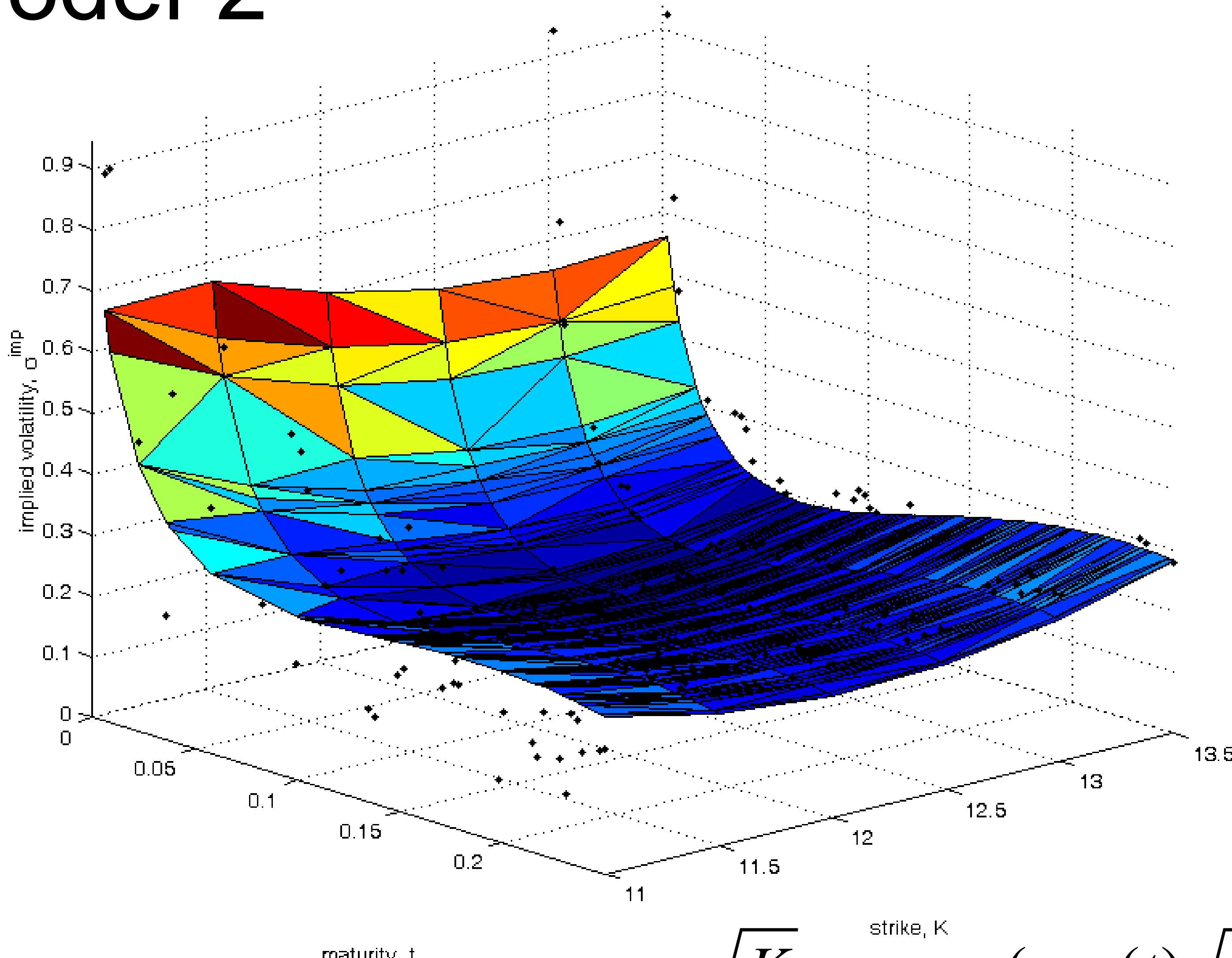
implied volatility, σ^{imp}
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.05
0.1
0.15
0.2
0.25
maturity, t
11
11.5
12
12.5
13
13.5
strike, K

Model-1



$$f_1 = w_0 + w_1 t^2 + w_2 tK + w_3 K^2$$

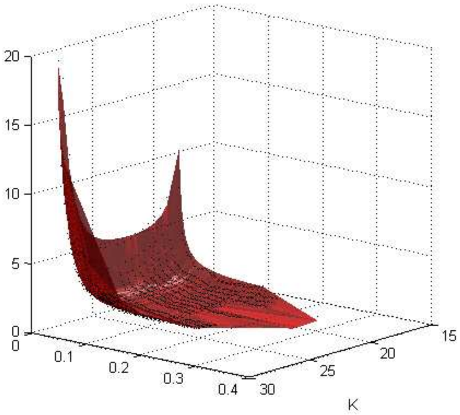
Model-2



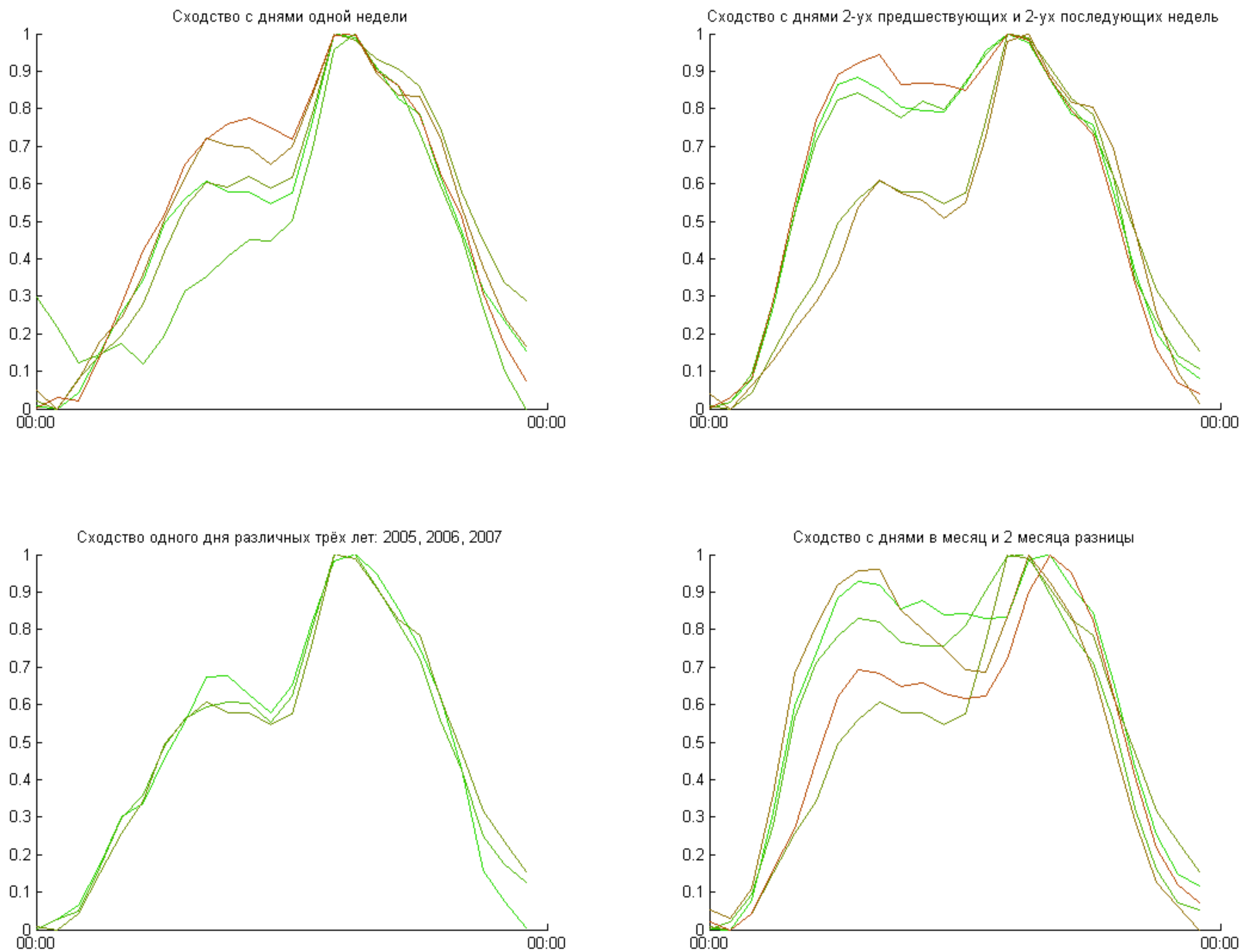
$$f_2 = w_0 + w_1 t^2 + w_2 K^2 + w_3 \frac{\sqrt{K}}{1+\exp(t)} + w_4 \frac{(\exp(t)\sqrt{t})\sqrt{K}}{K}$$

Non-linear model

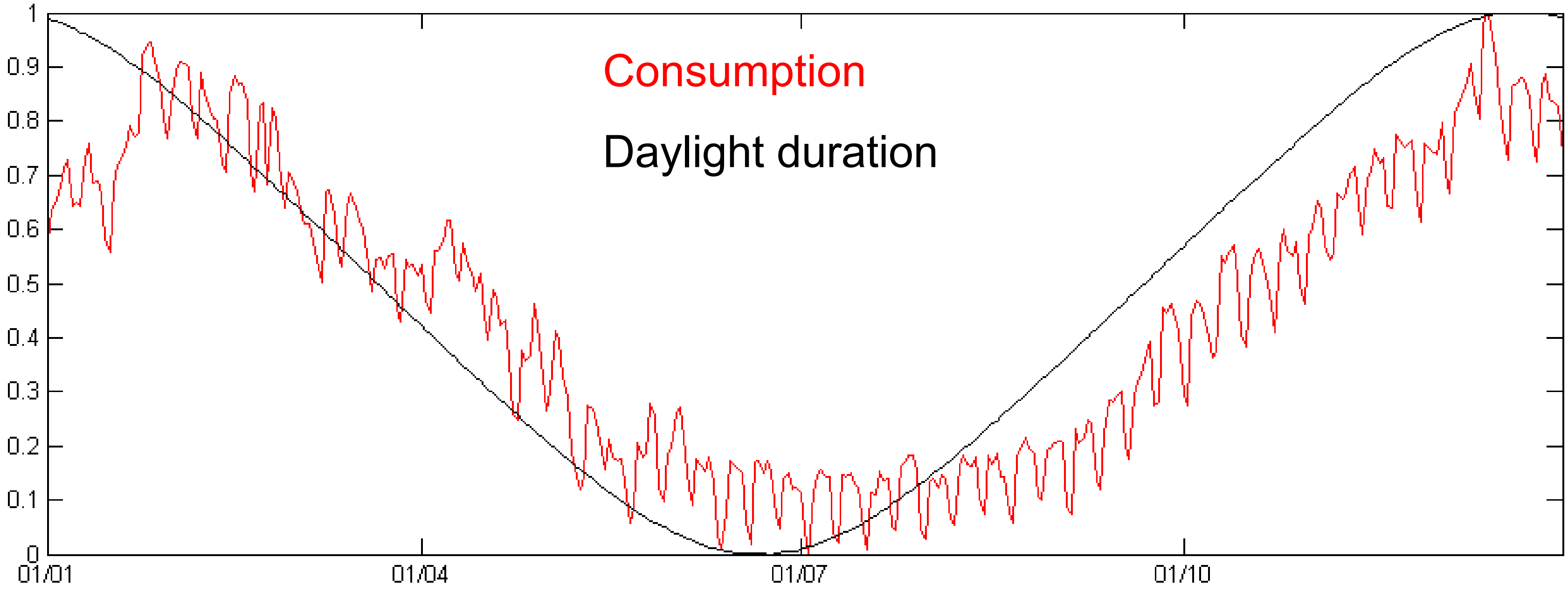
$$\sigma = \frac{(w_1 K^2 + w_2 K + w_3)}{\sqrt{t}}$$



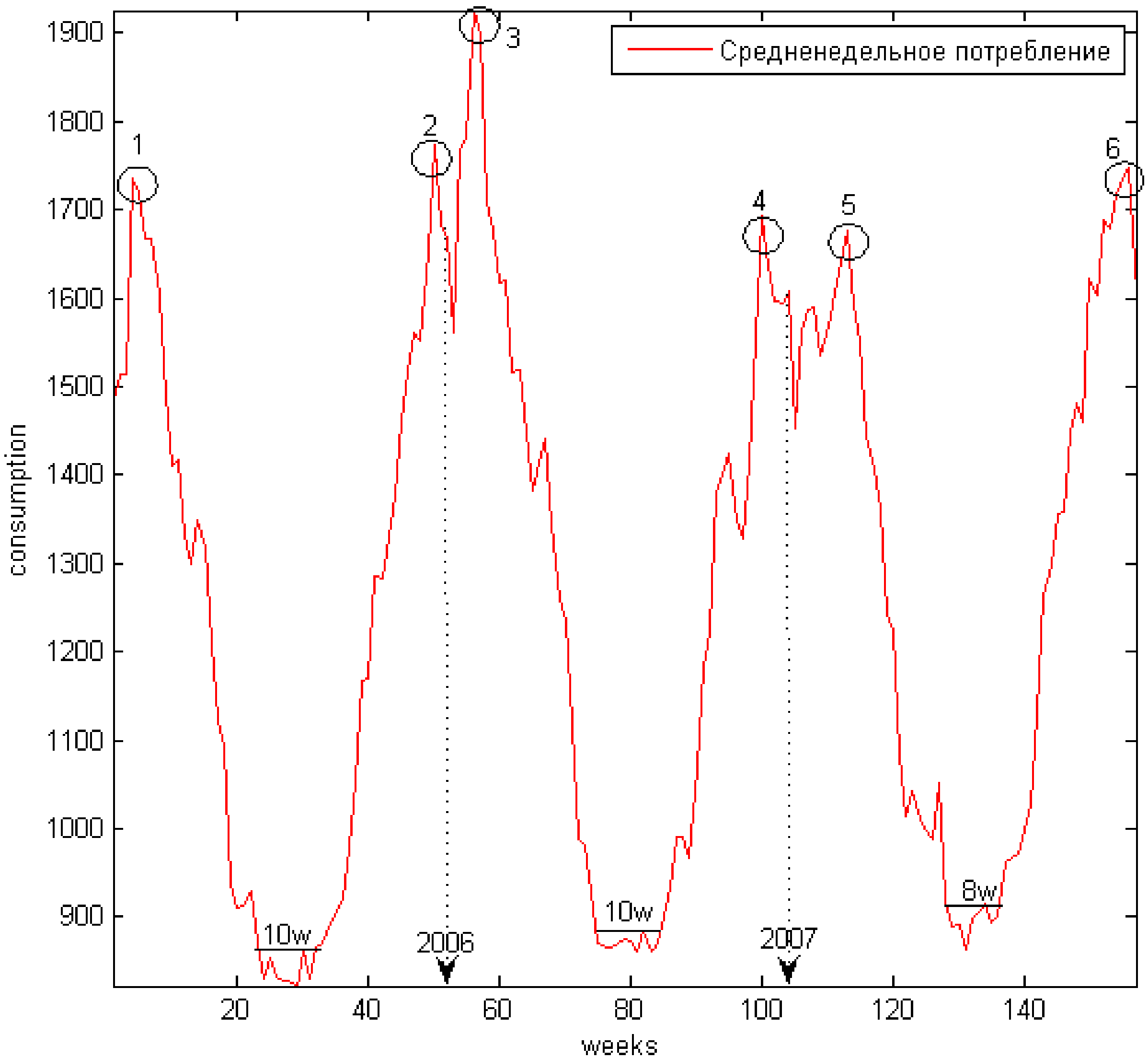
Hourly energy consumption



Daily energy consumption



Weekly energy consumption



Problem statement

Let there be given:
$\mathbf{x} = [x_1, \dots, x_{T-1}]^T$, $x \in \mathbb{R}^1$ — time series,
$t_{\tau+1} - t_\tau = \text{const}$,
k is a period and $T = mk$.
One must:
to forecast the next value x_T.
The reshaped time series is $(m \times k)$-matrix

$$X^{\text{combined}} = \begin{pmatrix} x_T & x_{T-1} & \dots & x_{T-k+1} \\ x_{(m-1)k} & x_{(m-1)k-1} & \dots & x_{(m-2)k+1} \\ \dots & \dots & \dots & \dots \\ x_{nk} & x_{nk-1} & \dots & x_{n(k-1)+1} \\ \dots & \dots & \dots & \dots \\ x_k & x_{k-1} & \dots & x_1 \end{pmatrix}.$$

The regression problem

$$X^{\text{combined}} = \left(\begin{array}{c|ccc} x_T & x_{T-1} & \dots & x_{T-k+1} \\ \hline \\ x_{(m-1)k} & x_{(m-1)k-1} & \dots & x_{(m-2)k+1} \\ \dots & \dots & \dots & \dots \\ x_{nk} & x_{nk-1} & \dots & x_{n(k-1)+1} \\ \dots & \dots & \dots & \dots \\ x_k & x_{k-1} & \dots & x_1 \end{array}\right).$$

In a nutshell,

$$\left(\begin{array}{c|c} x_T & \mathbf{x}_{\text{test}}^T \\ \hline \\ \mathbf{y} & X \end{array}\right).$$

In terms of linear regression:

$$\mathbf{y} = X\mathbf{w},$$

$$y^* = x_T = \langle \mathbf{x}_{\text{test}}^T, \mathbf{w} \rangle.$$

Further model generation

Let there be given:
a set of the functions $G = \{g_1, \dots, g_r\}$, for example
$g_1 = 1$, $g_2 = \sqrt{x}$, $g_3 = x$, $g_4 = x\sqrt{x}$.

The generated regression model $X =$

$$\left(\begin{array}{ccc|c|ccc}
g_1 \circ x_{T-1} & \dots & g_r \circ x_{T-1} & \dots & g_1 \circ x_{T-k+1} & \dots & g_r \circ x_{T-k+1} \\
\hline
g_1 \circ x_{(m-1)k-1} & \dots & g_r \circ x_{(m-1)k-1} & \dots & g_1 \circ x_{(m-2)k+1} & \dots & g_r \circ x_{(m-2)k+1} \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
g_1 \circ x_{nk-1} & \dots & g_r \circ x_{nk-1} & \dots & g_1 \circ x_{n(k-1)+1} & \dots & g_r \circ x_{n(k-1)+1} \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
g_1 \circ x_{k-1} & \dots & g_r \circ x_{k-1} & \dots & g_1 \circ x_1 & \dots & g_r \circ x_1
\end{array}\right).$$

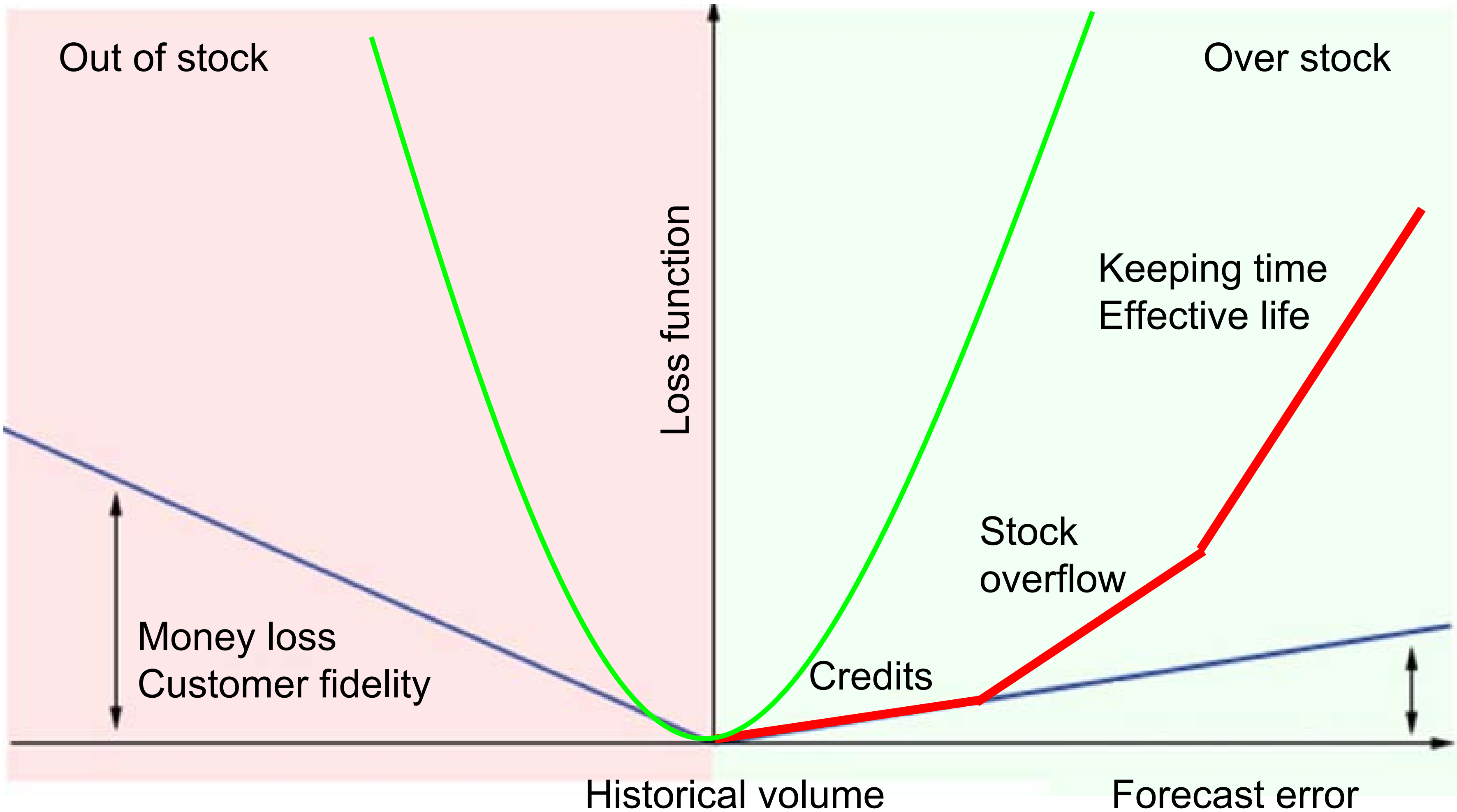
Time series forecasting

1. There is a historical time series of the volume off takes (i.e. foodstuff).

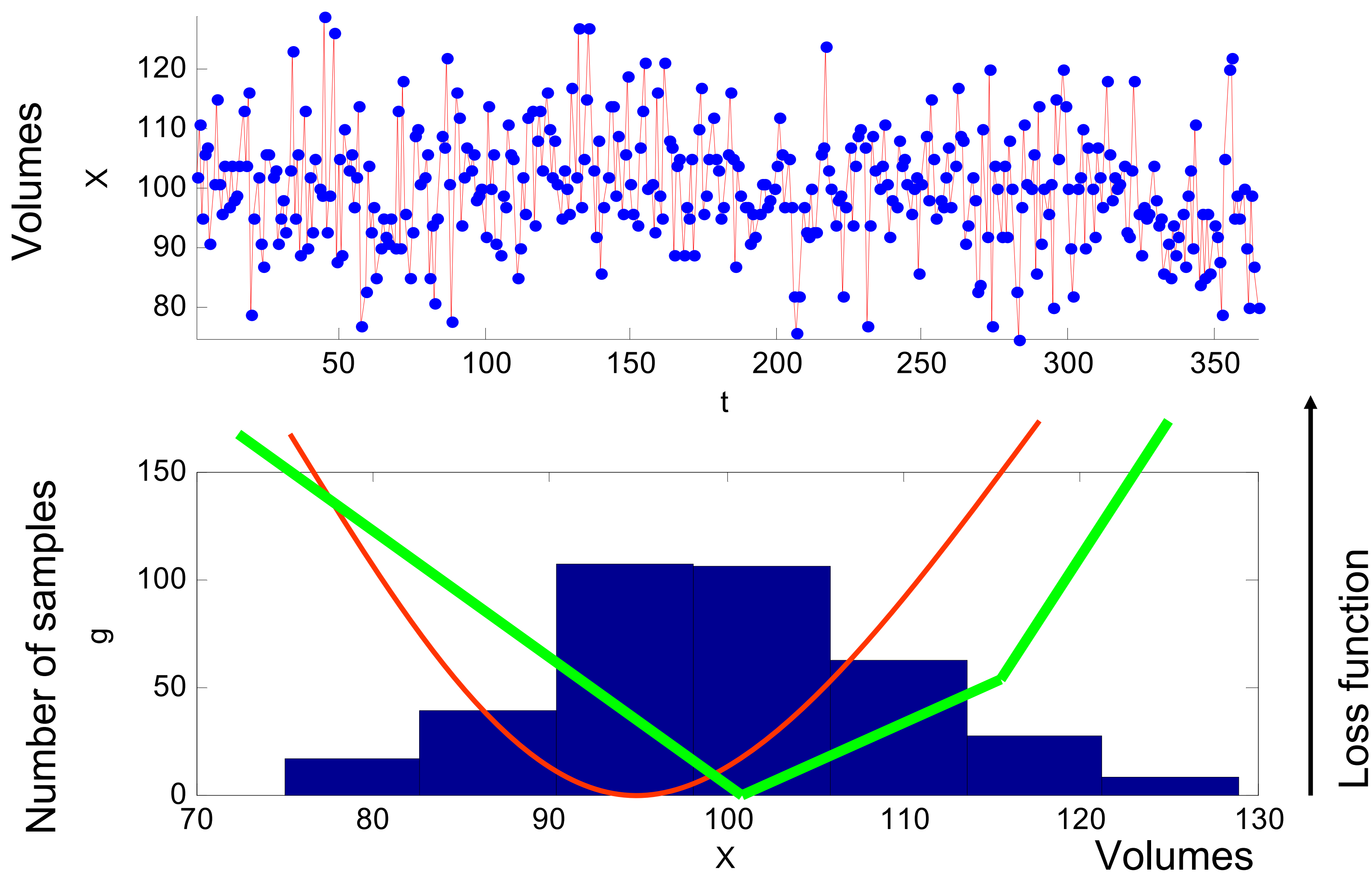
2. Let the time series be homoscedastic.

3. Using the loss function one must forecast the next sample.

Asymmetrical loss function



The time series and the histogram



Let there be given:

$\Gamma = \{ (X_i, g_i) \}$, $i=1, \ldots, N$ – histogram of the time series samples empirical distribution,

$L(Z, X)$ – loss function.

Problem:

For Γ and L, one must find the optimal forecast value X^*.

Solution:

$$X^* = \underset{Z \in \{X_1, \ldots, X_N\}}{\arg\min} \sum_{i=1}^{N} g_i L(Z, X_i).$$

Result:

X^* – the optimal forecast.

Ventia non sunt multiplicanda praeter necessitatem

William of Ockham
1285-1349

Occam's rasor: entities (model elements) must not be multiplied beyond necessity