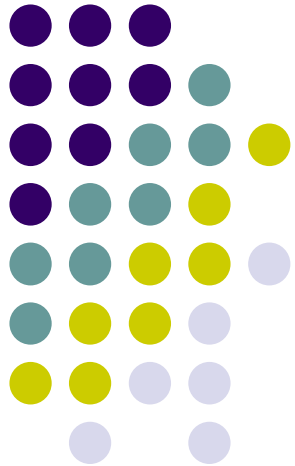
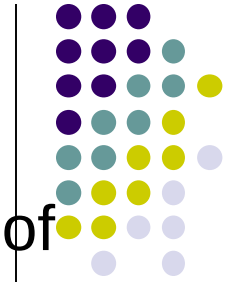
Statistical and Soft Computing Modelling Methods Applied to Economic Data

DUŠAN MARČEK
Faculty of Economic, Department of Applied Informatics, VSB-TU Ostrava

Dusan.Marcek@vsb.cz

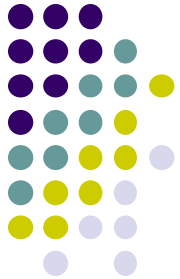


Outlines



- Basic principles of identifying input-output functions of systems and forecasting
- A classification table of statistical and NN methods for time series modelling and forecasting
- Explanatory (Statistical, causal) Modelling
- Flow chart of building an appropriate time series model
- Model based on B-J and H-W methodology - estimation
- SV Regression Model
- Automated modelling approach
- Conclusions

Basic principles of identifying input-output functions of systems and forecasting



There are three major approaches to forecasting – explanatory, time series, machine learning

- **Explanatory modelling and forecasting assumes a cause and effect relationship between the inputs into the system and its output. In this system any change in inputs will affect the output of the system in a predicable way.**
- **Time series**
- **In a special case machine learning such as SV regression or ANN offers relatively new ways for improving forecast method.**

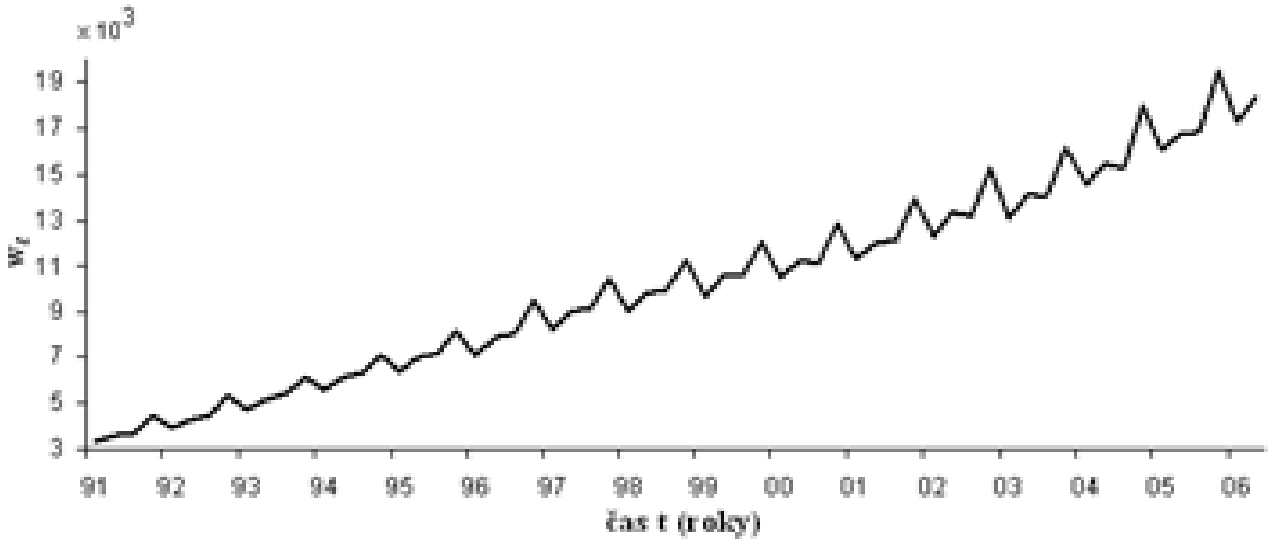
A classification table of statistical and NN methods for time series modelling and forecasting

E/Statistical-Probab. M. SC

Standard regression/ econometric models	Without or with Seasonal comp.	Integration of 3 IT:	ANN
			Fuzzy logic systems
Latest models: State-Space Models (K. filtration)	Structural models. EC, VEC models		Machine learning
Transfer Function M.		ANN:	
Models: ARIMA,		Perceptron-type NN	
ARCH-GARCH		**RBF NN**	classic
			soft (fuzzy logic)
Special models:	Bayes M.		**granular**
- ARIMA/ARCH-GARCH models require more costs of development, - there is not a convenient way to modify or update the estimates of the model parameters as each new observation becomes available and one has to periodically completely develop and refit the model.		Believe	
		Recurrent	
		Computational NN offer exciting adwantages such as learning, adaptation, fault-tolerace, parallelizm, generalization.	

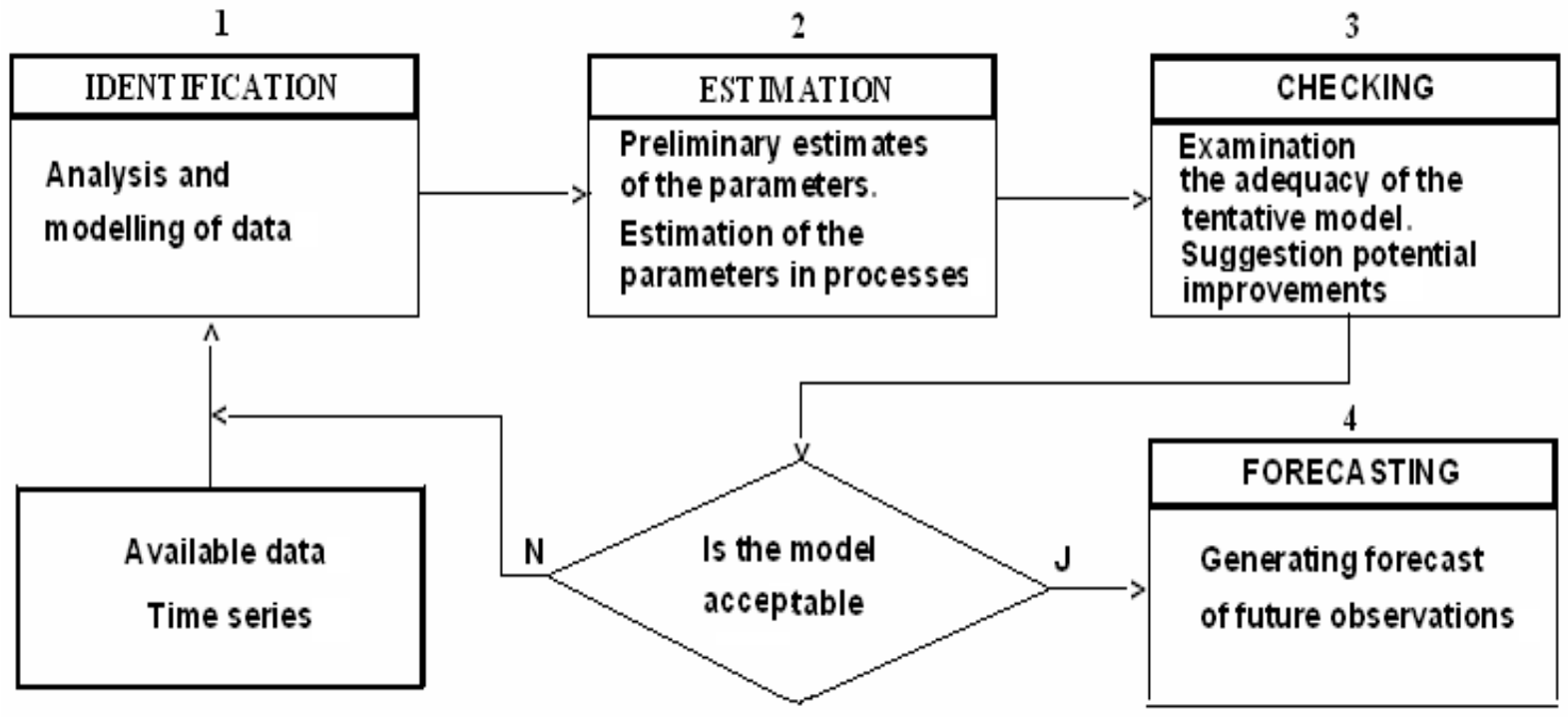
Explanatory (Statistical: ARIMA and Holt-Winter´s Methodology) Modelling

To study the modelling problem of wages quantitatively the quarterly data from 1991Q1 to 2005Q4 was collected



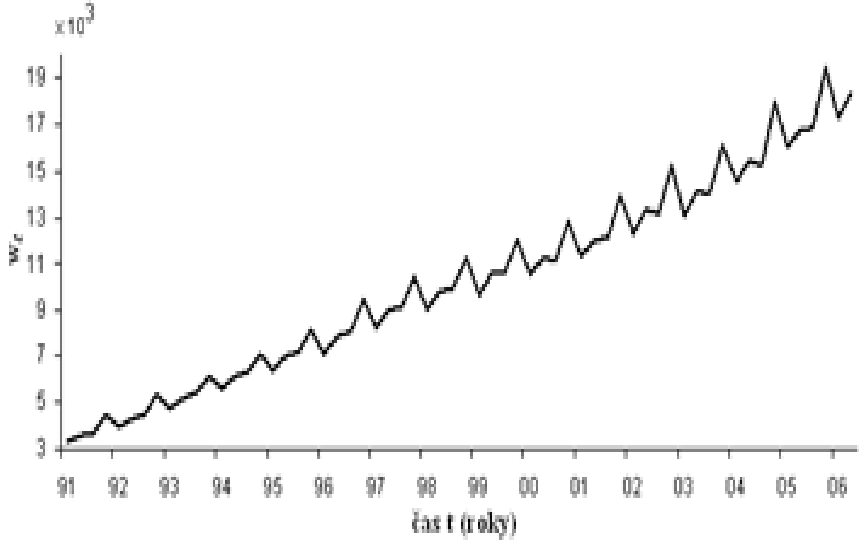
A chart of wages development

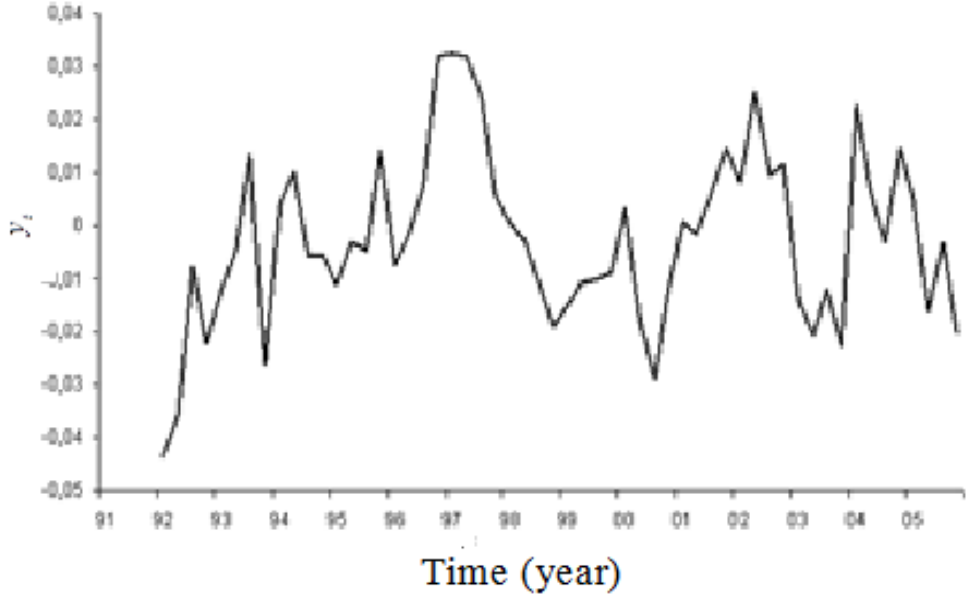
Flow chart of building an appropriate time series model



Model based on B-J methodology - estimation

=





Time plot of the transformed (stationary) time series of wages

Model based on B-J methodology - estimation

To study the estimation problem, we looked to determine the maximum lag for which the PACF coefficient was statistically significant and the lag given the minimum the form

$$\hat{y}_t = -0.001656 - 0.4567 y_{t-1} + 0.9052 \varepsilon_{t-1} + 0.588 \varepsilon_{t-2} + 0.365 \varepsilon_{t-3}$$

which is the ARIMA(1,2,3) series. The approximate and predictive accuracy of the model is documented in *Table 1* and visually shown in *Figure 4* on the left.

We also developed the exponential time series model according to Holt-Winter´s methodology with additive seasonal component. This model was developed using the methodology available on the free online resource www.it.iitb.ac.in/~praj/acads/seminar/04329008_ExponentialSmoothing.pdf.

Table 1

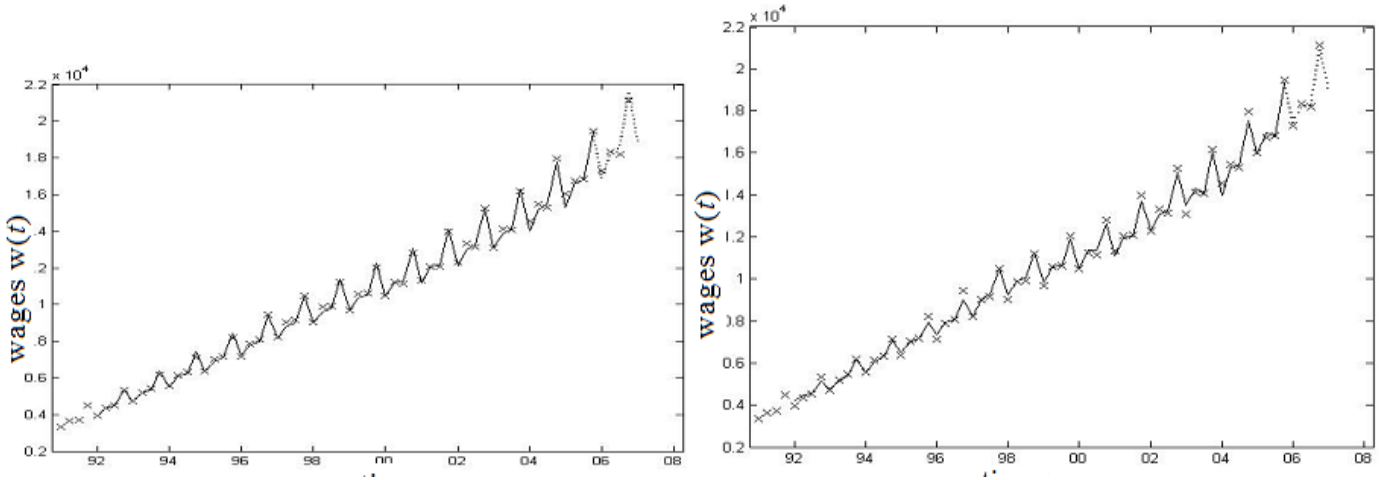
The summary approximation and predictive characteristics of the ARIMA and Holt-Winter´s model. MSE_A reports the precision for coaching the set A (1991-2006). MSE_E expresses the ex-ante predictions (MSE_E (3) or MSE_E (4) for 3 or 4 quarters of 2006

Model	model/Fig.no.	p	q	MSE_A	MSE_E (3)	MSE_E (4)
ARIMA	Time series/11.8	1	3	25289	104830	111880
Holt-Winter´s	Time series/11.9			34959	11216	18899

Model based on B-J and H-W methodology- estimation

$$\hat{y}_t = -0.001656 - 0.4567 y_{t-1} + 0.9052\varepsilon_{t-1} + 0.588\varepsilon_{t-2} + 0.365\varepsilon_{t-3}$$

which is the ARIMA(1.2,3) series. The approximate and predictive accuracy of the model (11.63) is documented in next table 11.3 and visually shown in next Fig. left



On the left, a chart of wages developments and their forecasts based on the ARIMA model.

The crosses represent actual values the wages time series. On the right, a wage chart and their forecasts based on the Holt-Winter´s model.

The full line represents estimates of the time series The dashed line represents an estimate of ex ante forecasts for year 2007

Model	model/Fig.no.	p	q	MSE_A	MSE_E (3)	MSE_E (4)
ARIMA	Time series/11.8	1	3	25289	104830	111880
Holt-Winter´s	Time series/11.9			34959	11216	18899

The summary approximation and predictive characteristics of the ARIMA and Holt-Winter´s model. MSE_A reports the precision for coaching the set A (1991-2006). MSE_E expresses the ex-ante predictions (MSE_E (3) or MSE_E (4) for 3 or 4 quarters of 2006)

SV Regression Model

$$w_t = \phi\, w_{t-4} + \varepsilon_t$$

$$w_t = b_0 + b_1 t + \varepsilon_t, \qquad t = 1, 2, \ldots, 60$$

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{x}^T \mathbf{w} + b \qquad f(\mathbf{x}) = \sum_{\substack{i=1 \\ j=1}}^{n} (\alpha_i - \alpha_i^*)\psi(\mathbf{x}_i \mathbf{x}_j) + b \qquad f(\mathbf{x}, \mathbf{w}, b) = \psi(\mathbf{x}_i, \mathbf{x}_j)\mathbf{w} + b$$

$$\mathbf{w} = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)\mathbf{x}_i \qquad b = \frac{1}{n}\left(\sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \mathbf{w}) \right)$$

$$\max_{\alpha, \alpha_i^*} -\frac{1}{2} \sum_{i,j=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\psi(\mathbf{x}_i^T \mathbf{x}_j) - \varepsilon \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i (\alpha_i - \alpha_i^*)$$

subject to constrains

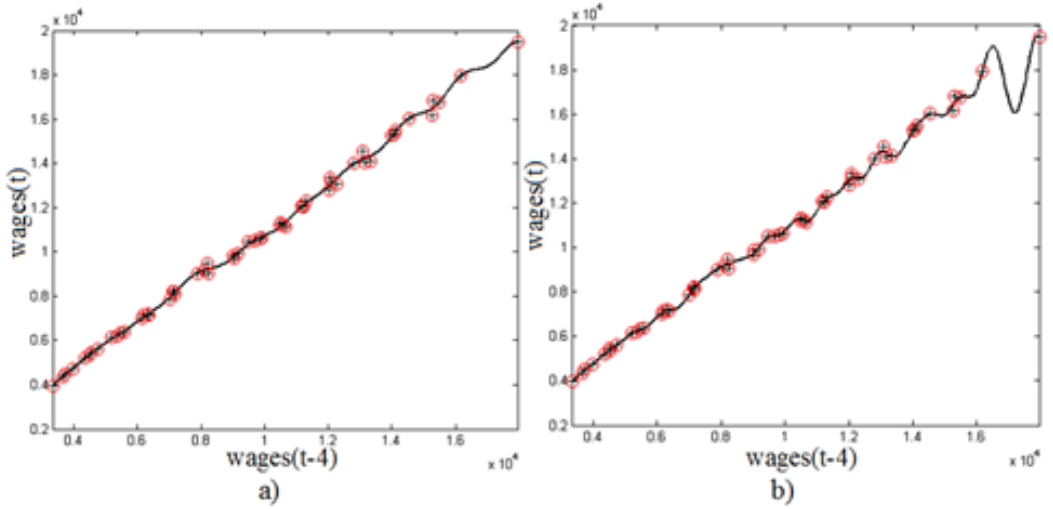
$$\begin{cases} \dfrac{\partial L_p}{\partial w} = 0 & \rightarrow \mathbf{w} = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i)\psi(\mathbf{x}_i), \\ \dfrac{\partial L_p}{\partial b} = 0 & \rightarrow \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) = 0, \\ \dfrac{\partial L_p}{\partial \xi_i} = 0, \dfrac{\partial L_p}{\partial \beta_i} = 0 & \rightarrow 0 \le \alpha_i \le C, \quad i = 1, \ldots, n, \\ \dfrac{\partial L_p}{\partial \xi_i^*} = 0, \dfrac{\partial L_p}{\partial \beta_i^*} = 0 & \rightarrow 0 \le \alpha_i^* \le C, \quad i = 1, \ldots, n \end{cases}$$

where L_p is the Lagrangian with Lagrange multipliers given by $\alpha_i, \alpha_i^* \ge 0; \beta_i, \beta_i^* \ge 0$

ξ_i, ξ_i^*

SV Regression Model

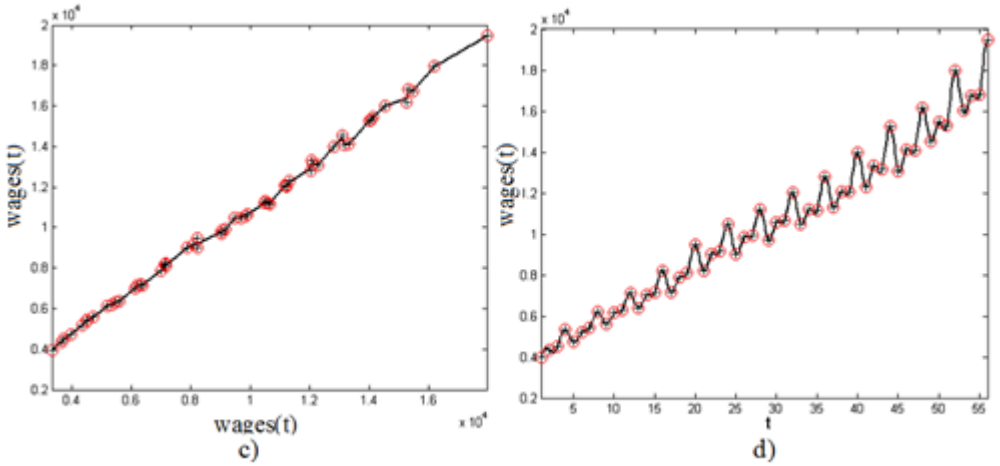
model/fig.	kernel function	σ	degree -d	loss function	MSE
Causal (a)	RBF	11		**insensitive**	15590
Causal (b)	RBF	6		**insensitive**	10251



Results of the SV regression for various kernel functions, loss functions and standard deviations σ according to *Table*

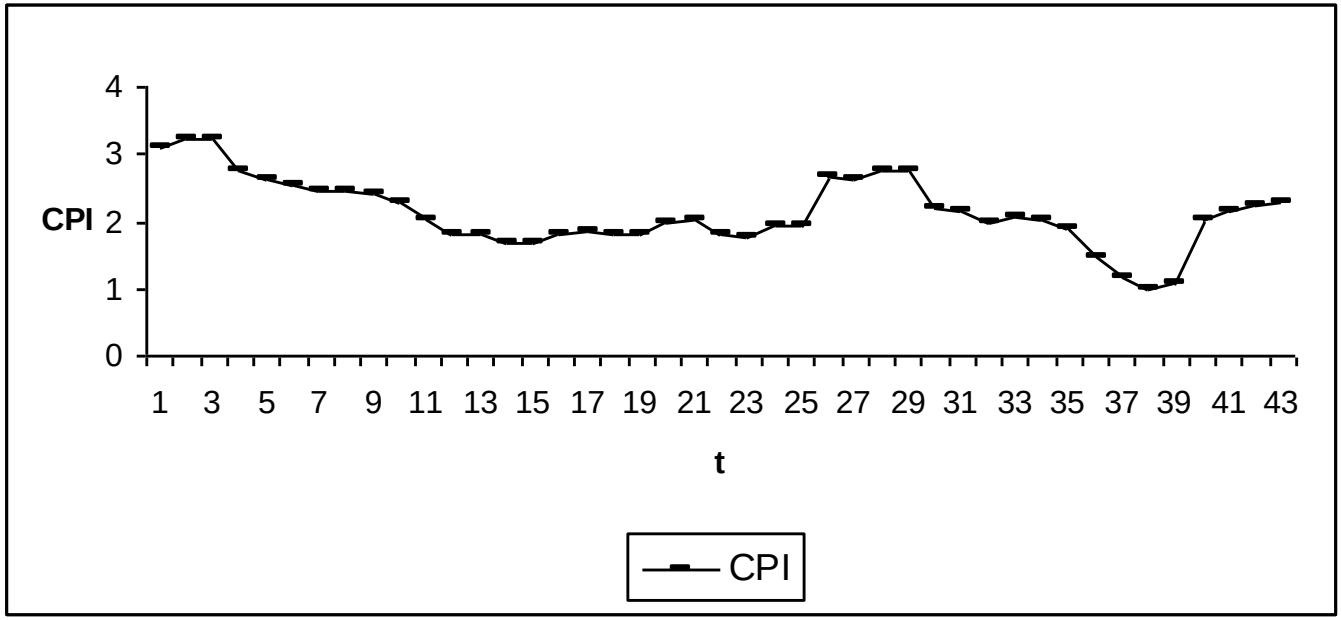
SV Regression Model

model/fig.	kernel function	σ	degree -d	loss function	MSE
Causal (c)	ERBF	6		**insensitive**	3315,7
Causal (d)	RBF	1		**insensitive**	0,421



Explanatory (Econometric, causal) Modelling

To study the modelling problem of inflation quantitatively the quarterly data from 1993Q1 to 2003Q4 was collected concerning the consumption price index *CPI*, aggregate wages *W* and unemployment *U*.



Natural logarithm of quarterly inflation from January 1993 to December 2003

Automated modelling

The strategy for selecting an appropriate model is based on so called a "specific to general" methodology (Dynamic Modelling in Economics - DME)

The DME methodology leads to two stage modelling procedure

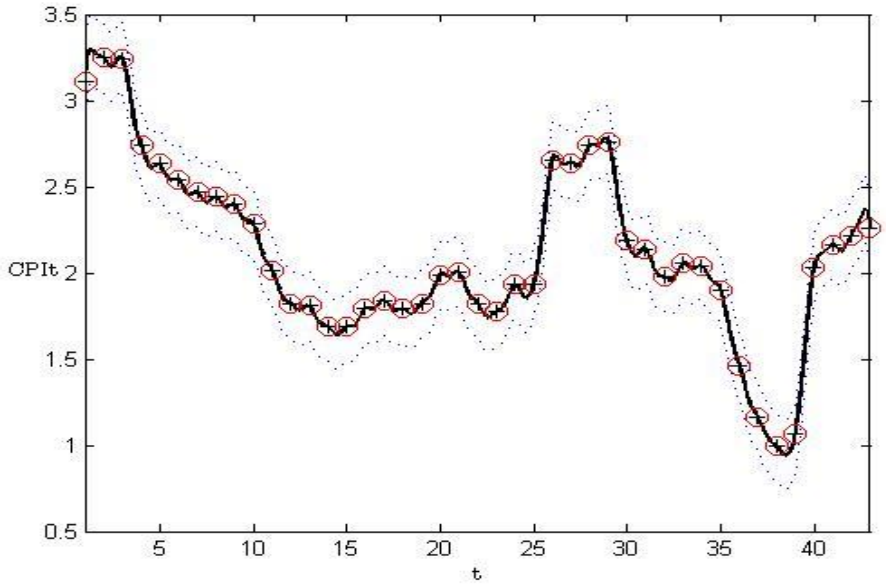
$$\hat{CPI}_t = 11.3302 - 1.355W_t + 1.168U_t = 11.3302 - 1.355W_t + 1.168U_t \quad (2)$$

$R^2 = 0.374, \quad DW = 0.511$

$$\hat{CPI}_t = \underset{(0.229)}{0.5941} - \underset{(0.3387)}{0.0295W_{t-1}} - \underset{(0.1035)}{0.00359U_{t-1}} + 0.84524CPI_{t-1}, \qquad R^2 = 0.7762 \quad (3)$$

$$\hat{CPI}_t = \beta_0 + \beta_1 CPI_{t-1} = \underset{(0.072)}{0.292} + 0.856CPI_{t-1} \quad (4)$$

Fig.	MODEL	KERNEL		DEGREE-d	LOSS FUNCTION	R^2	RMSE
2f	time series (6)	RBF	0.52		insensitive	0.9999	1.1132

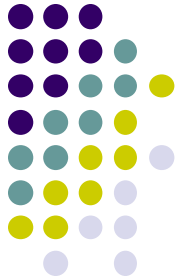


Empirical Comparisons and Conclusion

The use of an SV machine is a powerful tool to the solution many economic problems. The estimated parameters, in contrast with statistical estimators, have any economic interpretation to structural model, all parameters in the model are fixed, and there is no possibility to test the stability of the parameters.

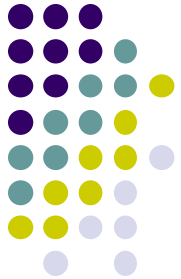
Unfortunately, the SVM´s method does not explicitly define how the forecast is determined, the point estimates of the fitted model are simple values without any degree of confidence for the results.

The prediction results are given in tab 1. The model in Fig. b gives best predictions outside the estimation period and clearly dominates the other models. We have shown that too many model parameters results in overfitting, i.e. a curve fitted with too many parameters follows all the small fluctuations, but is poor for generalisation.

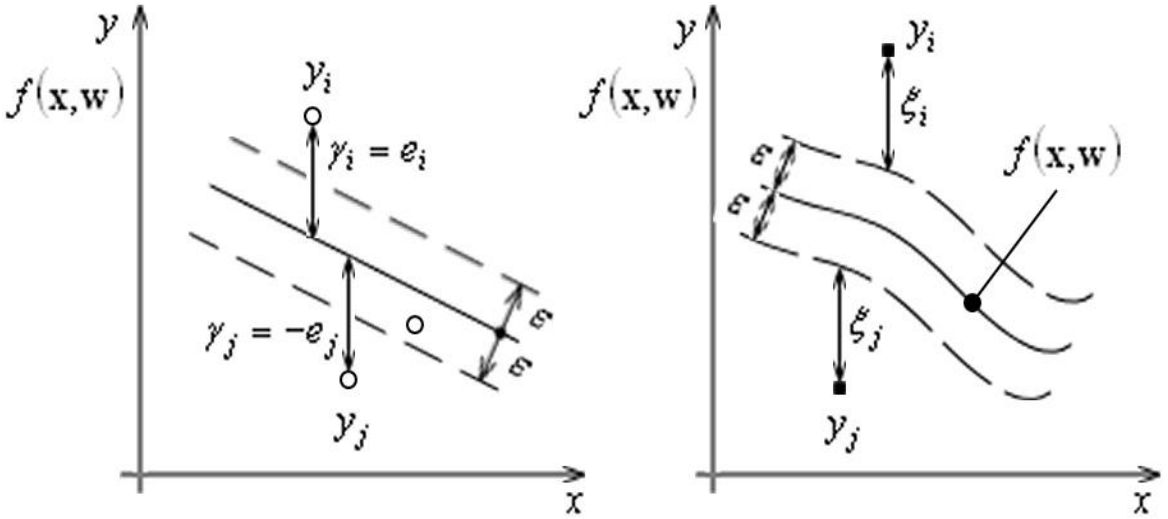


Thank you

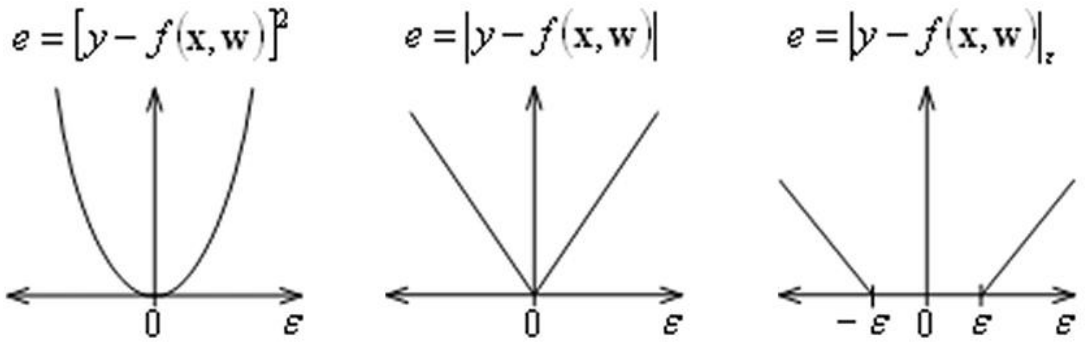
Some kernel functions



$\psi(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$	lineárna pre lineárne SVM
$[(\mathbf{x}_i^T \mathbf{x}_j + 1)]^d$	polynomická stupňa *d* pre polynomické SVM
$\exp\ [-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2]$	Gaussova pre RBF SVM
$\tanh[(\mathbf{x}_i^T \mathbf{x}_j + b)]$	sieť perceptrónového typu s jednou skrytou vrstvou



SVM linear and nonlelinear regression



Loos functions: quadratic L_1, absolute L_2 and ε-ignorujúca norma L_3

$$|y - f(\mathbf{x}, \mathbf{w})|_{\varepsilon} = \begin{cases} 0 & \text{if } |y - f(\mathbf{x}, \mathbf{w})| \le \varepsilon, \\ |y - f(\mathbf{x}, \mathbf{w})| - \varepsilon & \text{otherwise,} \end{cases}$$

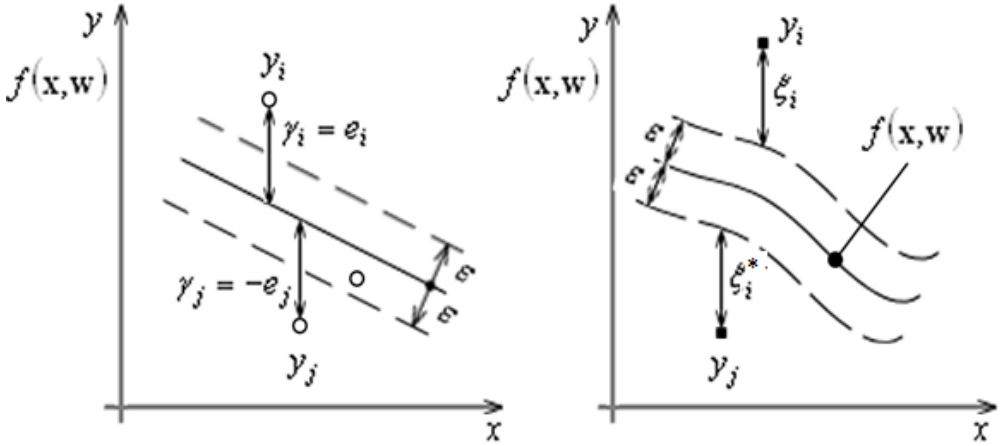
Causal Models, Experimenting with Non-linear SV Regression

Given a training set of *n* data points $\{y_i, x_i\}_{i=1}^{n}$ where $x_i \in R^n$ denotes *i*th input, and $y_i \in R$ is the *i*th output, the support vector appoach aims at constructing a function of the form

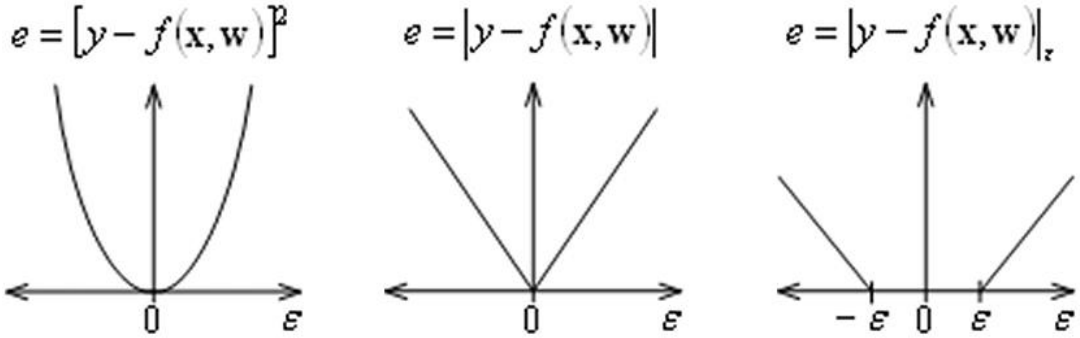
$$f(\mathbf{x}) = \sum_{\substack{i=1 \\ j=1}}^{n} (\alpha_i - \alpha_i^*)\psi(\mathbf{x}_i \mathbf{x}_j) + b$$

where $\alpha_i - \alpha_i^*$ are positive real constants and *b* is real constant, $\psi(./.)$ is so called kernel function.

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{x}^T \mathbf{w} + b \qquad f(\mathbf{x}) = \sum_{\substack{i=1 \\ j=1}}^{n} (\alpha_i - \alpha_i^*)\psi(\mathbf{x}_i \mathbf{x}_j) + b$$



SVR linear and nonlelinear regression



$$|y - f(\mathbf{x}, \mathbf{w})|_\varepsilon = \begin{cases} 0 & \text{if } |y - f(\mathbf{x}, \mathbf{w})| \le \varepsilon \\ |y - f(\mathbf{x}, \mathbf{w})| - \varepsilon & \text{otherwise} \end{cases}$$