MODEL MIXING FOR LONG-TERM EXTRAPOLATION

Pavel Ettler[1], **Miroslav Kárný**[2], **Petr Nedoma**[2]

[1]COMPUREG Plzeň, s.r.o.
306 34 Plzeň, Czech Republic

[2]Institute of Information Theory and Automation (UTIA)
182 02 Praha, Czech Republic

ettler@compureg.cz (Pavel Ettler)

Abstract

Reliable extrapolation – simulation or prediction – of system output is an invaluable departure point for the control system design. For application of model-based techniques, the knowledge of the model structure is essential. It can be based purely on the physical point of view or estimated from process data while the system is considered as a *black box*. Mixing of both methods results in *grey box* modelling. Often, modelled systems are governed by several known physical laws and each of these laws implies a model, which should match the data. Nevertheless inevitable uncertainties often make simulated outputs of respective models unreliable. The problem is especially pronounced for systems with a significant time delay. This motivates search for methods, which utilize all available models at once and mix their outputs with the aim to get better results. In the paper, four variants of mixing are considered, discussed and their performance compared on industrial data. Seeming alternative – a simple complex model is discussed as well. Data for experiments came from a cold rolling mill.

Keywords: Simulation, modelling, estimation, multiple models.

Presenting Author's Biography

Pavel Ettler received the doctor degree in cybernetics from the University of West Bohemia in Plzeň. He worked as researcher at Škoda (Rolling Mills branch) for eleven years. In 1993 he joined COMPUREG, a company oriented to industrial control systems. His interests include identification and control of systems subject to uncertainties with application to metal processing and machine control. His involvement with research includes participation in several international and national research projects, mainly in co-operation with UTIA.

1 Introduction

Having data from a real system at disposal, construction of a linear model that extrapolates measured data seems to be simple. The construction relies on the following key steps: i) determine the sampling period; ii) choose the maximal model order respecting the dynamic properties of the system; iii) create the regression vector containing all available explanatory data (regressors); and iv) estimate model parameters, typically by least squares [1, 2]. Unfortunately, this purely "black box" modelling often fails in practice due to unknown correlation of data channels, imperfect measurements, unmeasurable disturbances, etc. Moreover, the gained models are usually over-parameterized and unsuitable for simulation or multi-step prediction. Thus, at least rough respecting of known physical relations seems unavoidable in model building.

Many real systems behave according to several known physical laws. A simple model based on a particular law could be sufficient providing the available data are uncorrupted and informative enough. This case is, however, rare. The systematic grey-box model building [3] is to be used whenever possible to counteract this problem. Often, however, the complexity of the resulting model is too high. Under this situation, addressed here, it is necessary to utilize all available simple models, each representing a particular anticipated relation among data. The particular model outputs are then combined into the overall model output. The paper inspects promising combination possibilities and tests them on real data. They are related to rolling mills [4, 5]. The tested case is exceptional by an inherent significant transport delay and high demands on extrapolation quality.

All considered models ${}^{\iota}\mathsf{M}$, $\iota \in \iota^* \equiv \{1, 2, \ldots, \mathring{\iota}\}$, of a system S have the form:

$$ {}^{\iota}\mathsf{M}: \;\; y_t = {}^{\iota}\theta \, {}^{\iota}\psi_t' + {}^{\iota}\xi_t, \tag{1} $$

where ι labels the model; t stands for the discrete time; y_t denotes system output; ${}^{\iota}\theta$ is a vector of unknown regression coefficients; ${}^{\iota}\psi_t$ is the regression vector; and ${}^{\iota}\xi_t$ denotes the zero mean white noise. Often, the noise can be assumed to be normal with a constant variance r_ι. Then, the output y_t is equivalently described by the normal $\mathcal{N}_{y_t}(\cdot,\cdot)$ probability density function (pdf)

$$ f\left(y_t \middle| \, {}^{\iota}\psi_t, \, {}^{\iota}\Theta\right) = \mathcal{N}_{y_t}({}^{\iota}\theta \, {}^{\iota}\psi_t', r_\iota), \;\; {}^{\iota}\Theta \equiv ({}^{\iota}\theta, r_\iota). \tag{2} $$

Parameters ${}^{\iota}\Theta$ are estimated recursively using data measured on the system. Estimates ${}^{\iota}\widehat{\theta}$ of parameters ${}^{\iota}\theta$ serve for extrapolation of the output course. They provide output predictions ${}^{\iota}\widehat{y}_t$

$$ {}^{\iota}\widehat{y}_t = {}^{\iota}\widehat{\theta}_{t-k} \, {}^{\iota}\psi_t'. \tag{3} $$

The subscript $_{t-k}$ at ${}^{\iota}\widehat{\theta}_{t-k}$ stresses that the estimates are based on data records $d^{1:t-k} \equiv (d_1, \ldots . d_{t-k})$, where $k \geq 1$ is a known estimation delay. The data record d_t contains all measurements made at time t.

2 Mixing principles

Use of a single model with regressors obtained as a union of regressors of all models is the most straightforward combination way. This approach is often applicable. Generally, it has a tendency to overparametrization and consequently to unreliable extrapolations. This property is expected to be fatal for the considered systems with a significant estimation delay. Moreover, the computational complexity increases sharply as the estimation complexity increases quadratically with the size of regression vector. This makes us avoid this option completely and focus on mixing of simple models. Considered mixing principles are discussed below.

2.1 Bayesian averaging (BA)

Bayesian paradigm [1] interprets all unknown quantities as random variables. Estimation then evaluates posterior pdf on them. For linear normal models of the type (2), the Gauss-inverse-Wishart pdf [6] of the unknown $\Theta = (\theta, r)$ reproduces during updating. The updating algorithmically coincides with recursive least squares (RLS) initiated via prior pdf. The predictive pdf $f\left(y_t \middle| \, {}^{\iota}\psi_t, d^{1:t-k}, \iota\right)$ is then Student pdf with expected value ${}^{\iota}\widehat{y}_t$ (3). Under uncertainty about the adequate model structure, the pointer ι to respective models ${}^{\iota}\mathsf{M}$ is to be taken as random variable. Then, the proper combined predictor becomes

$$ {}^{\mathring{\iota}+1}\mathsf{M}: \; f\left(y_t \middle| \, \{{}^{\iota}\psi_t\}_{\iota\in\iota^*}, d^{1:t-k}\right) = \sum_{\iota\in\iota^*} f\left(\iota \middle| \, d^{1:t-k}\right) f\left(y_t \middle| \, {}^{\iota}\psi_t, d^{1:t-k}, \iota\right), \tag{4} $$

where the probabilistic weights $f\left(\iota \middle| \, d^{1:t-k}\right)$ evolve according to the Bayes rule

$$ f\left(\iota \middle| \, d^{1:t-k}\right) = \frac{f\left(y_t \middle| \, {}^{\iota}\psi_t, d^{1:t-k}, \iota\right)}{\sum_{\iota\in\iota^*} f\left(y_t \middle| \, {}^{\iota}\psi_t, d^{1:t-k}, \iota\right) f\left(\iota \middle| \, d^{1:t-k-1}\right)} \tag{5} $$

starting from some, say uniform, prior pdf. The corresponding point prediction is

$$ \widehat{y}_t = \sum_{\iota\in\iota^*} f\left(\iota \middle| \, d^{1:t-k}\right) \, {}^{\iota}\widehat{y}_t. \tag{6} $$

This Bayesian processing [7] was given the name Bayesian averaging [8].

The computational overhead is small compared to parallel estimation and prediction with $\mathring{\iota}$ simple models.

2.2 Mixture model (MM)

Bayesian averaging weights individual predictions well. At the same time, it does not respect the possibility that for some data configurations some simple models should not be updated as the data do not belong to the model-validity domain. Noticing that the overall predictor is a mixture of predictors, it is reasonable to

consider the mixture model as the basic one

$$^{\imath+1}\mathsf{M}:\ f\left(y_t|\{\,^{\iota}\psi_t,\,^{\iota}\Theta,\alpha_\iota\}_{\iota\in\iota^*}\right)=\sum_{\iota\in\iota^*}\alpha_\iota\mathcal{N}_{y_t}(\,^{\iota}\theta\,^{\iota}\psi'_t,r_\iota) \tag{7}$$

in which probabilistic weights $\alpha_\iota, \iota \in \iota^*$ extend the set of unknown parameters. The choice is supported by the universal approximation property [9] of the popular mixture models [10].

The mixture (7) of a fixed structure can be effectively estimated in recursive mode using so called projection-based Bayesian estimation [11]. Algorithmically, it runs $\imath$ *weighted* RLS. The weights reflect a degree with which the processed data are in harmony with the updated model. In this way, drawback of the Bayesian averaging is suppressed. It makes us expect a better performance of the obtained predictor. At the same time, the computational overhead is still relatively small.

2.3 Predictions as regressors (PR)

Both Bayesian averaging BA and prediction by mixture model MM provide the overall prediction as a convex combination of individual predictions. This causes troubles if the predicted output is outside the convex hull of individual model outputs. This problem can be simply overcome if we take individual predictions as regressors in the overall model $^{\imath+1}\mathsf{M}$ combining them. It has the form (1) with

$$^{\imath+1}\psi_t=[\,^{1}\widehat{y}_t,\dots,\,^{\imath}\widehat{y}_t,1]. \tag{8}$$

The $(\imath+1)^{\rm st}$ model provides the combined prediction

$$^{\imath+1}\widehat{y}_t=\,^{\imath+1}\hat{\theta}_{t-k}\,^{\imath+1}\psi'_t \tag{9}$$

with parameter estimate $^{\imath+1}\hat{\theta}'_{t-k}$ updated by ordinary RLS. The weights of the respective predictions are generally real numbers. This, together with estimation of the offset, can provide predictions outside the convex hull. This overcomes the drawback of previous methods. Thus, we expect that it outperforms Bayesian averaging. At the same time, it suffers from the drawback justifying use of mixtures: the estimated parameters of the model $^{\imath+1}\mathsf{M}$ are assumed to be good for all data configurations.

The combination costs just a single RLS of size $\imath+1$.

2.4 Predictions as regressors in mixture (PM)

The remaining drawback implies the favorite model for combining predictors, namely, mixture with its components "sitting" on scaled individual predictions

$$^{\imath+1}\mathsf{M}:\ f\left(y_t|\{\,^{\iota}\widehat{y}_t,a_\iota,b_\iota,r_\iota,\alpha_\iota\}_{\iota\in\iota^*}\right)=\sum_{\iota\in\iota^*}\alpha_\iota\mathcal{N}_{y_t}\left(a_\iota\,^{\iota}\widehat{y}_t+b_\iota,r_\iota\right). \tag{10}$$

The combination reduces to estimation of this simple mixture. Projection-based estimation runs on components parameterized by scaling parameters (a_ι,b_ι), and variance r_ι. The resulting point prediction need not be within the convex hull of individual predictions.

2.5 Expected performance of respective variants

Previous discussion implies that mixtures MM, PM are expected outperform both Bayesian averaging and use of predictions as regressors. Behavior of the last combined extrapolator should be the best from those discussed. Taking into account the form of individual predictions, we see, that it may happen that the mixture model MM will outperform PM if: i) all individual models are enriched by offset; ii) estimated regression coefficients become product of the scaling factor $^{\iota}a$ and of original, physically motivated, coefficients. On the other hand, the use of predicted values as regressors brings added advantage: noise entering the regression vector is suppressed.

Approximate nature of mixture estimation may even destroy advantageous properties of mixture models. Thus, at the current state of the knowledge, just experimental evidence and computational demands may determine preferences between the proposed methods. At the same time, we do not expect a substantial gain by considering the most general mixtures with probabilistic weights depending on regression vectors. This is due to the need to cope with the considered time-delay in estimation, see the next section. Essentially, we have to make very long-term extrapolations and thus we have to either: i) rely on a weak dependence of individual-predictors quality on data or: ii) find time-invariants of such dependence. At present, the latter case cannot be practically solved as it requires numerical integrations in high dimensions. Thus, we have to rely on assumption that i) is valid. Possible slow variations are counteracted by a version of forgetting [6, 12].

3 Time delay problem

As mentioned in Introduction, the addressed problem was motivated by the transport delay problem inherent for rolling mills. The situation is sketched on Fig. 1. Data measurement is triggered by the strip shift of Δ length units. Current data for creating the regression vector ψ are available at time τ, i.e., for the piece of the strip in the rolling gap. The system output y – output strip thickness – is measured after k steps at time t. The task consists in the extrapolation of y_τ, $\tau = t-k$, i.e., the strip thickness just leaving the rolling gap is predicted.

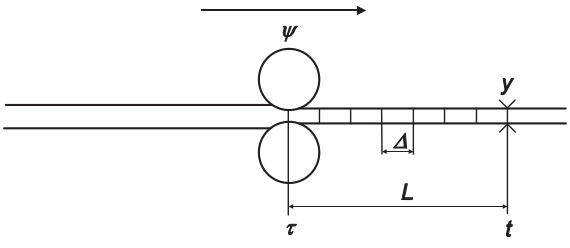

Fig. 1 Transport delay $k \approx L/\Delta$, k is a natural number.

Estimates ${}^{\iota}\hat{\theta}$ of ${}^{\iota}\theta$ parameterizing particular models ${}^{\iota}\mathsf{M}$ (1) can be updated at time t when the output y_t complements the regression vector ${}^{\iota}\psi_t$ into data vector ${}^{\iota}\Psi_t = [y_t, {}^{\iota}\psi_t]$. For prediction at time τ, just k-steps "old" parameter estimates and weights are available as indicated in (3). Utilization of the "old" parameters obviously deteriorates particular predictions. Therefore, the mixing of several predictions can be vital for counteracting this drawback.

4 Experiments

Experiments provide an insight into behavior of proposed methods and help us select the favorite one for the considered application. Moreover, they indicate whether the achieved improvement is worth increased computational demands. Comparisons are made on a typical data sample of the length 2000. In order to suppress influence of the tuning phase, characteristics of predictions are computed for the last 1600 data records.

Three underlying models ${}^{\iota}\mathsf{M}$ of the first order are used. They deal with the physical signals measured on the cold rolling mill. The signals are characterized in Table 1.

Tab. 1 Signals used in the combined models

No	Meaning	Units
1	output strip thickness	μm
2	input strip thickness	μm
3	rolling force	MN
4	input strip speed	m/s
5	output strip speed	m/s
6	ratio of speeds	
7	screwdown position	μm

All models predict signal on channel 1. Structure of regression vectors of respective simple models, combining the above channels and their delayed values, were determined from elementary physical laws, like mass conservation (mass-flow) principle, as well as from the inspection of extensive historical data. The transport delay between the rolling gap where prediction is made and the output measurements is $k = 25$. Thus predictions are calculated with utilization of k-step "old" parameters and weights.

The processing imitated recursive real-time use. The respective simple models are estimated with forgetting factor 0.999. The compound models ${}^{\iota+1}\mathsf{M}$ are estimated with forgetting factor 0.97.

The results related to respective methods are presented individually in the order corresponding to the method description. Comparison based on elementary descriptive statistics characterizing prediction errors follows.

In order to get impression about character of the predicted output, Fig. 2 shows output, the best composite prediction and prediction error of the simple model ${}^{1}\mathsf{M}$.

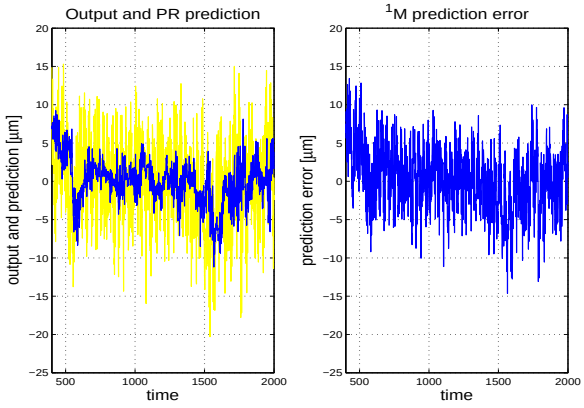

Fig. 2 Left plot: output and the best prediction; right plot: prediction error of the model ${}^{1}\mathsf{M}$

4.1 Bayesian averaging (BA)

Trajectories of the weights, i.e., posterior probabilities of respective models $f\left(\iota \middle| d^{t-k}\right)$ (5), are displayed in the top-down order in Fig. 3. Exponential forgetting is applied to them in order to influence speed of their variations. The trajectories of these weights for forgetting rates corresponding to expected slow and fast variations of simple-models validity are shown in Fig. 3.

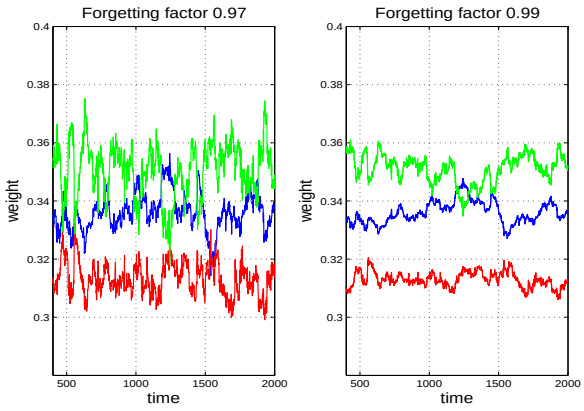

Fig. 3 Character of BA weights influenced by the value of forgetting factor. Left plot: forgetting factor 0.97, right plot: 0.99

In order to get overall picture, the trajectory of the output is displayed in the left part of Fig. 4, the output prediction in its right part. Qualitative behavior is similar in other cases and that is why predictions are not displayed in some figures.

4.2 Mixture model (MM)

Unlike a general mixture, the treated one has known structure. Its components are the combined models and this determines the number of components. The estimated component weights provide directly weights of the combined models. For the mixture, predicted and

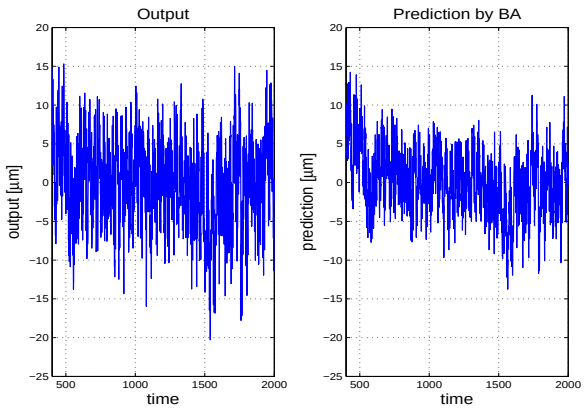

Fig. 4 System output (left plot) and its BA prediction at the rolling gap (right plot)

observed properties may differ mainly due to the non-negligible error of approximate estimation. As seen in the overall comparison, Table 2, such a deterioration occurred but still the mixing provides observable improvement comparing to individual models.

For the mixture-based output extrapolation, smooth behavior of estimates of component weights is characteristic, see left part of Fig. 5.

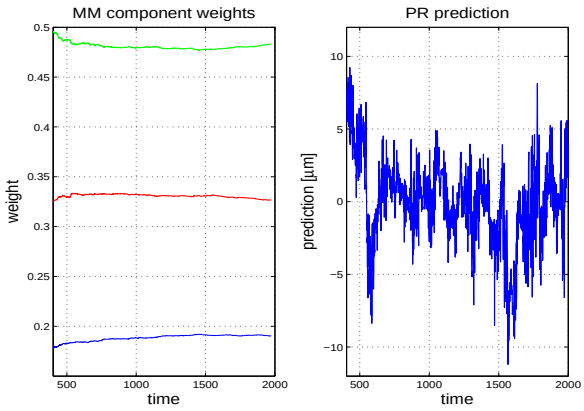

Fig. 5 MM estimates of component weights (left plot). The best prediction of the output by PR (right plot)

4.3 Predictions as regressors (PR)

Application of this model combination outperformed expectations connected with it. For the considered application, it was found as the best one. The positive shift in the quality is due to the exactly implementable Bayesian estimation. The corresponding prediction is in the right part of Fig. 5.

4.4 Predictions as regressors in mixture (PM)

This combination method was expected to provide the best results. Nevertheless they did not come up to expectation as shown in Table 2. Still the method gives good results.

4.5 Comparison of efficiency of respective methods

Qualitative comparison of considered methods is reflected on Fig. 6 where their prediction errors are displayed. Fig. 7 compares histograms of predictions errors.

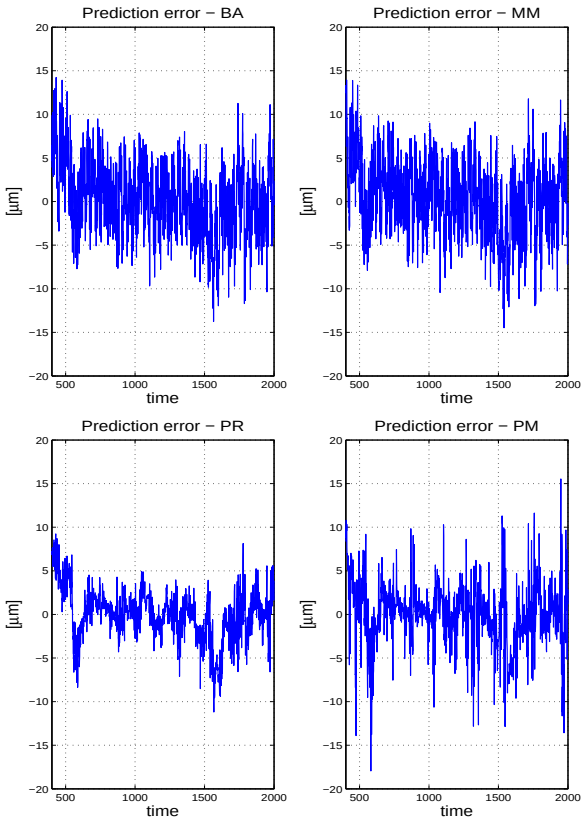

Fig. 6 Prediction errors for particular methods in the order BA, MM, PR, PM

Sample statistics evaluating prediction errors for three simple models and four combination methods respectively are summarized in Table 2.

Tab. 2 Sample statistics of prediction errors $\hat{e}^{401:2000}$

	$\mathsf{E}[\hat{e}]$	$\min[\hat{e}]$	$\max[\hat{e}]$	$\mathsf{std}[\hat{e}]$	$\mathsf{E}[\hat{e}^2]$
^{1}M	0.28	-14.63	13.43	4.13	17.15
^{2}M	0.37	-19.20	15.73	5.24	27.64
^{3}M	0.27	-13.76	14.23	4.16	17.39
BA	0.27	-13.76	14.23	4.16	17.39
MM	-0.51	-14.73	11.44	4.03	16.49
PR	**-3e-4**	**-11.19**	**9.23**	**2.94**	**8.67**
PM	-0.14	-17.95	15.54	3.57	12.83

It is worth noticing that the best combination methods from the view point of $\mathsf{E}[\hat{e}^2]$ are also the most "stable" (see $\min[\hat{e}]$ and $\max[\hat{e}]$). The improvement of the best method PR comparing even to the best simple model

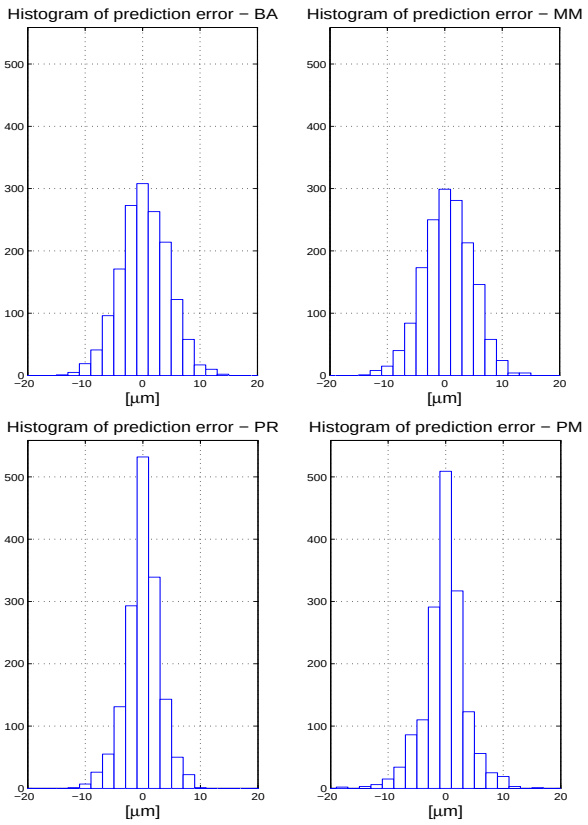

Fig. 7 Histograms of prediction errors for particular methods in the order BA, MM, PR, PM

[1]M is obviously significant. Even the worst combination method achieves the quality comparable with the best simple model. This feature is important as the observed order of methods can be case dependent.

5 Conclusions

The paper presented four promising methods of combining outputs of physically motivated extrapolation models to a single one. For the considered rolling mill application, characterized by a significant transport delay, the combination of individual predictions by a static regression model seems to be the best solution. Due to the approximations involved, the result can be case dependent. Thus, the presented general discussion should be taken into account when tailoring the methods to other types of applications.

Acknowledgements

This work was partly supported by research projects BADDYR (grant 1ET100750401 of the Academy of Sciences of the Czech Republic) and DAR (project 1M0572 of the Czech Ministry of Education).

6 References

[1] V. Peterka, Bayesian system identification, *Trends and Progress in System Identification*, P. Eykhoff, Ed., pp. 239–304. Pergamon Press, Oxford, 1981.

[2] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, London, 1987.

[3] T. Bohlin, *Interactive System Identification: Prospects and Pitfalls*, Springer-Verlag, New York, 1991.

[4] G. Rath, *Model Based Thickness Control of the Cold Strip Rolling Process*, Doctoral Thesis, University of Leoben, 2000.

[5] P. Ettler, M. Kárný and T. V. Guy, Bayes for rolling mills: From parameter estimation to decision support, *Proceedings of the 16th IFAC World Congress*, Praha, 2005.

[6] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař, *Optimized Bayesian Dynamic Advising: Theory and Algorithms*, Springer, London, 2005.

[7] M. Kárný, Algorithms for determining the model structure of a controlled system, *Kybernetika*, vol. 19, no. 2, pp. 164–178, 1983.

[8] A. E. Raftery, D. Madigan, and J. A. Hoeting, Bayesian model averaging for linear regression models, *Journal of The American Statistical Association*, vol. 97, no. 437, pp. 179–191, 1997.

[9] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Macmillan, New York, 1994.

[10] D.M. Titterington, A.F.M. Smith, and U.E. Makov, *Statistical Analysis of Finite Mixtures*, John Wiley, New York, 1985.

[11] J. Andrýsek, *Estimation of Dynamic Probabilistic Mixtures*, PhD thesis, FJFI, ČVUT, POB 18, 18208 Prague 8, Czech Republic, 2005.

[12] R. Kulhavý and M. B. Zarrop, On a general concept of forgetting, *International Journal of Control*, vol. 58, no. 4, pp. 905–924, 1993.