ISSN 1330-3651
UDC/UDK 658.56 : 681.5

ONLINE DATA PREPROCESSING IN THE ADAPTIVE PROCESS MODEL BUILDING BASED ON PLANT DATA

Dražen Slišković, Ratko Grbić, Željko Hocenski

Original scientific paper

Process variables which are concerned with the quality of final product cannot often be measured by a sensor. The alternative procedure is the estimation of these difficult-to-measure process variables for which it is necessary to have an appropriate process model. Process model building, based on plant data taken from the process database, is usually the most cost-effective way to obtain a process model. Since the quality of the built model depends heavily on the modelling data informativity, preprocessing of the available measured data is an important step in such process modelling. Processes are usually time-varying and non-stationary, so that the precision of the estimation based on process model with constant parameters degrades over time. Because of that, model parameters have to be updated online. However, in order to successfully keep the precision of the estimation, it is important to use the samples which do not contain errors in the parameter updating procedure which requires a quality online data preprocessing. The online data preprocessing and online model parameter updating are discussed and presented on two examples and the influence of data preprocessing on adaptive process model quality is analyzed.

***Key words**: difficult-to-measure variable estimation, online data preprocessing, online model parameter updating, plant data*

On-line predobradba podataka u izgradnji adaptivnog modela procesa na temelju pogonskih podataka

Izvorni znanstveni članak

Vrlo često važne procesne veličine koje su povezane s kvalitetom izlaznog proizvoda nije moguće mjeriti senzorom. Alternativni postupak je procjenjivanje iznosa ovih teško-mjerljivih veličina, za što je potreban odgovarajući matematički model procesa. Izgradnja modela procesa na pogonskim podacima preuzetim iz procesne baze podataka potencijalno je najjeftiniji način iznalaženja modela. Budući da kakvoća izgrađenog modela uvelike ovisi o informativnosti raspoloživih mjernih podataka, predobrada mjernih podataka je važan korak u izgradnji modela procesa na temelju pogonskih podataka. Budući da su procesi najčešće vremenski promjenjivi i nestacionarni, točnost procjene teško-mjerljive veličine modelom procesa s konstantnim parametrima opada s vremenom. Zbog toga je potrebno prepodešavati parametre modela "online". Prilikom prepodešavanja parametara modela, kako bi se uspješno održavala točnost procjene, potrebno je koristiti uzorke koji su bez grešaka što zahtijeva kvalitetnu online predobradbu ovih uzoraka. Predobradba podataka na online način kao i online prepodešavanje parametara modela prikazani su na dva primjera te je provedena analiza utjecaja predobradbe podataka na svojstva adaptivnog modela procesa.

***Ključne riječi**: online predobradba podataka, online prepodešavanje parametara modela, pogonski podaci, procjenjivanje teško-mjerljive veličine*

1 Introduction
Uvod

Accurate and efficient online measurements of process variables which give information about final product quality are necessary for process control and optimization. However, these process variables cannot often be measured by a sensor or the measurements are too expensive and/or not reliable enough and therefore are not used. The value of these difficult-to-measure variables is usually determined by laboratory analysis based on the samples taken from the process. This kind of measurement is performed periodically, with a long time delay in obtaining information, and it does not provide continuous monitoring of the final product quality and introduction of automatic control. To provide this, the estimation of the difficult-to-measure process variables can be performed based on the process variables that are measured by sensors in the plant (so called easy-to-measure variables) and which correlate with difficult-to-measure variables [1]. For that it is necessary to have an appropriate mathematical model. In practice, the model is usually not available. Because industrial processes are generally quite complex to model, a rigorous theoretical modelling approach is often impractical, requiring a great amount of effort, or even impossible. Thus, obtaining the process model is based on the measured data [2].

In modern industrial plants there are hundreds of process variables which are measured and stored in the process database, so it is logical to use these data for process model building. However, the measured data taken from the process database are plant data and they contain plenty of different random and gross errors. Since quality of the built model depends heavily on the modelling data informativity, a preparatory part of modelling in which preprocessing of available measured data is performed is a very important step in a plant data based process modelling [3]. The preparatory part of modelling often requires more time and effort than immediate model building on the selected data set and, due to its complexity, it is also necessary to involve experts in the field of the process [4].

Most industrial processes are time-varying and/or non-stationary, so the estimation quality of the model with constant parameters often degrades with time [5-6]. To obtain a good estimation of the difficult-to-measure process variable in online application, process model parameters have to be estimated in online manner. Also, online gross error detection and online data preprocessing have to be incorporated in such applications in order to achieve successful model updating and accurate estimation of the difficult-to-measure variable of interest.

The paper is organized as follows. The data preprocessing techniques are presented in Section 2. Section 3 gives a short description of selected methods for process modelling and methods for online updating of model parameters. In Section 4 modelling result for the two case studies are presented, with accompanying discussion. Finally, conclusions are drawn in Section 5.

2
Data preprocessing
Predobradba podataka

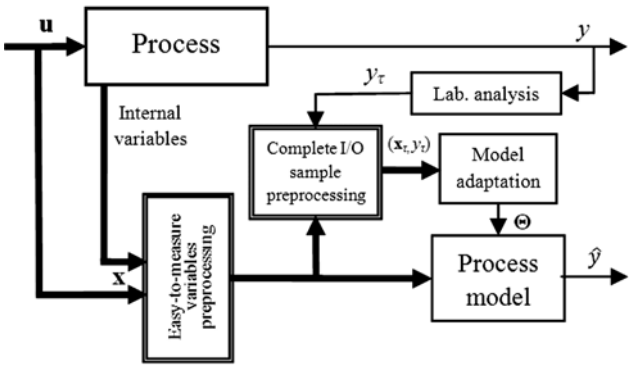
Two main types of data preprocessing can be distinguished. The first one is dealing with initial data set for offline model building while the second is regarded to the online model application.

The offline modelling, i.e. initial process model building has to be performed on informative data in order to achieve a satisfactory process model. Plant data possess a lot of impurities as a result of different disturbances, malfunctions, degradation and errors in sensors and data acquisition system. The impurities in the data for process modelling can cause incorrect model parameter estimation, especially when using least square (LS) methods, which results in an estimator with a low accuracy [7].

Therefore, data preprocessing, which includes easy-to-measure variables selection, outlier and missing values detection and replacement and data denoising, is an important step in the offline process model building [3]. As the difficult-to-measure variables are sampled with very low and variable frequency, it is not possible to detect which sample has an error and to estimate the amount of noise in the available data. Thus, difficult-to-measure variables cannot be preprocessed in the same way that easy-to-measure variables can.

In online applications newly acquired values of easy-to-measure variables (together with the built model) are used for estimation of the difficult-to-measure variable (Fig. 1). So, in order to achieve accurate estimation, current sample of easy-to-measure variables ($\mathbf{x}$) has to be preprocessed in online manner. Modifications of offline preprocessing techniques can be used in this case, but with limitation that only past samples are available for this procedure.

To track the process time-varying behaviour adaptive model has to be used in order to keep estimation precision of the difficult-to-measure process variable. In the adaptation procedure only model parameters are updated while model structure remains unchanged. Therefore, it is important to select appropriate model structure in initial model building. Model adaptation is based on complete input-output sample ($\mathbf{x}_\tau$, y_τ). To achieve successful online model parameter updating, error free samples must be provided. Due to the long delay (τ) which is introduced by laboratory analysis, this preprocessing can be done in efficient way since "future" samples are also available (samples available from the moment of sampling till the moment of getting information about difficult-to-measure variable). Thus, the model adaptation is delayed for the period τ in respect to the moment of sample taking.



u – input process variables, **x** – easy-to-measure variables, y – actual (unmeasurable) value of difficult-to-measure variable, y_τ – value of difficult-to-measure variable obtained by laboratory analysis, $\hat{y}$ – estimated value of difficult-to-measure variable, $(\mathbf{x}_\tau, y_\tau)$ – preprocessed input-output sample, τ – duration of laboratory analysis, **Θ** – model parameters.

Figure 1 *Principal schema of adaptive estimator with data preprocessing*
Slika 1. *Načelna shema rada prilagodljivog procjenjivača s pripadnim predobradbama podataka*

2.1
Selection of easy-to-measure variables
Odabir lako-mjerljivih veličina

From all available variables it is necessary to choose variables which are relevant for prediction of difficult-to-measure variable. Only the selected ones are used for offline process model building since irrelevant variables (variables that do not correlate with output variable or are in very low correlation with output variable) can potentially deteriorate modelling results [4]. Backward variable selection method and sensitivity analysis [8] are a commonly used method for variables selection in regression problems. However, it is not recommended to delete variables which are highly correlated because redundancy in the input space can potentially increase model robustness. Number of variables is usually not changed during online application.

2.2
Gross error treatment
Postupanje s grubim pogreškama

Samples with gross errors, i.e. erroneous samples, are caused by abnormal situations that happened in the past while the data were collected. Gross errors are manifested as either obvious outliers or missing values. Outliers can be defined as samples that are not consistent with the majority of the data. There are two groups of outliers: outliers which are beyond the min-max boundaries of a particular process variable (so called obvious outliers) and outliers that are within these boundaries (so called non-obvious outliers). Outliers occur as a result of a sensor malfunction or measurement error (so they are suspected to be generated by a different mechanism). The presence of these errors in the data set greatly affects model parameters estimation, especially when regression techniques are used, so it is important to detect and replace/delete such samples before offline modeling procedure and in online applications as well (see section 2.5).

Univariate offline outlier detection is often based on a visual inspection of the data. Only obvious outliers can be detected with this method. Other popular univariate approaches are 3σ edit rule and Hampel filter [9]. After the outliers identification, their values can be replaced by interpolation or some other technique or they can be removed from the data set (but only in the case of steady state model building). Slew-rate limiter can be used for non-obvious outlier detection. For multivariate outlier detection PCA combined with Q and T^2 statistics, resampling by half-means (RHM), smallest half volume (SHV) or ellipsoidal multivariate trimming (MVT) can be used [7, 10, 11].

Missing values in the data set are commonly caused by hardware failure of sensor, its maintenance or removal,

failure of communication system and by different errors in plant database. Missing values are relatively easy to find. The way to deal with missing values is to treat them as obvious outliers.

2.3
Data denoising
Filtriranje podataka

Data denoising means eliminating random contributions called noise which usually provides easier extraction of important information from some data. For this purpose linear low-pass filters are usually used which suppress signals with frequencies higher than some cut-off frequency (e.g. moving average filter). Wavelet analysis has proven to be a very efficient tool for offline and for online signal denoising (see section 2.4) and in different fields of application because it does not smooth important features of the original signal [3].

Basis of practical wavelet applications is discrete wavelet transformation (DWT). The mathematical formulation of the DWT will be omitted here and can be found elsewhere [12]. The selection of a subset of scales and positions based on powers of two (dyadic scales and positions) results in a more efficient and accurate analysis. Mallat [13] introduced multiresolution analysis where any signal $x(t) \in L^2(R)$ can be successively projected onto scaling functions and wavelet functions, respectively:

$$\begin{aligned} \Phi_{m,n}(t) &= 2^{-m/2}\Phi(2^{-m}t-n) \\ \Psi_{m,n}(t) &= 2^{-m/2}\Psi(2^{-m}t-n) \end{aligned} \quad m,n \in Z. \tag{1}$$

The inner products of the signal with $\Phi_{m,n}(t)$ are called approximation coefficients $\{a_{m,n}\}$ which represent a smoother version of the original signal and the inner products of signal with $\Phi_{m,n}(t)$ are known as the detail coefficients $\{d_{m,n}\}$.

An efficient algorithm for wavelet decomposition and the reconstruction of a discrete signal of dyadic length was developed in [13]. It consists of a repetitive application of high pass and low pass filters to calculate the wavelet decomposition of a given sequence of discrete numbers. By passing the input sequence through this pair of filters, the projection of signal onto the scaling and wavelet functions is performed as depicted in Fig. 2, where d_i and a_i (i=1,…,K) are detail and approximation coefficients of the i-th scale (level). Decomposition is carried out to a desired number of scales K by recursively applying the high and low pass filters to the approximation coefficients at the previous level. The original signal can be reconstructed from the detail coefficients of all scales and approximation coefficients of the last scale K by inverse discrete wavelet transformation. The procedure for wavelet denoising consists of three main steps:

1. Discrete wavelet decomposition of the signal.
2. Thresholding of non-significant detail coefficients at every scale of decomposition.
3. Reconstruction of signal by inverse wavelet transformation from thresholded details coefficients and approximation coefficients at scale K.

The true signal (especially in chemical process) tends to dominate the low frequency area (at high scales) while noise dominates at higher frequency area (at low scales) but

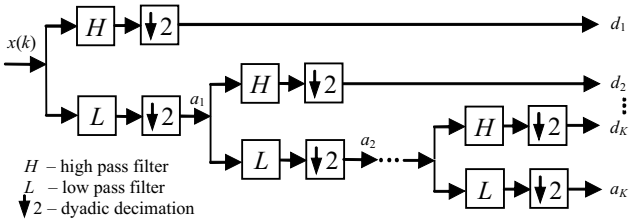


Figure 2 *Discrete wavelet decomposition*
Slika 2. *Diskretna wavelet dekompozicija*

affects all frequencies. In wavelet denoising, it is assumed that the signal contributes high amplitude coefficients which should be retained and the noise contributes low amplitude coefficients which should be removed. In contrast to smoothing techniques, where high frequency components above a certain frequency are suppressed, thresholding removes non-significant detail coefficients at every scale so all frequencies with the exception of the lowest frequency band which correspond to approximation coefficients at scale K are affected by the procedure. Therefore, wavelet denoising is very efficient in noise reduction while maintaining essential features of the original signal.

Wavelet denoising is based on either a hard or soft thresholding approach. In hard thresholding detail coefficients are set to zero if their value is above a certain threshold value (λ):

$$d_{i,j} = \begin{cases} d_{i,j} & \text{if } |d_{i,j}| > \lambda \\ 0 & \text{if } |d_{i,j}| < \lambda \end{cases}, \tag{2}$$

and in soft thresholding their values are shrunk toward zero [14]:

$$d_{i,j} = \begin{cases} d_{i,j} - \lambda & \text{if } d_{i,j} > \lambda \\ 0 & \text{if } d_{i,j} \le \lambda \\ d_{i,j} + \lambda & \text{if } d_{i,j} < -\lambda \end{cases}. \tag{3}$$

When applying thresholding, there are two main approaches: global thresholding in which a single threshold value λ is applied to all levels of decomposition, and level-dependent thresholding in which different value of threshold λ is used for every level of decomposition. The latter approach is preferred when dealing with non-stationary and/or correlated data [15]. There are several methods for selecting an appropriate threshold value for wavelet denoising [16-18].

For the sake of comparison simple unweighted moving average (MA) filter is used which behaves like low-pass filter:

$$x_s(i) = \frac{1}{2W+1}\left[x(i+W) + x(i+W-1) + \ldots + x(i-W)\right], \tag{4}$$

where $x_s(i)$ is smoothed sample, $x(i)$ original sample and W is window size. When data are smoothed with this kind of filter it is important to choose proper window size because it affects degree of the smoothing procedure [19], i.e. filter cut-off frequency.

2.4
Online data denoising
Online filtriranje podataka

Easy-to-measure variables are used as process model inputs in online estimation of the difficult-to-measure variable. Therefore, incoming samples have to be immediately preprocessed. Since the wavelet filters are generally noncausal in nature they require future samples of the measured signal. Also, only signal of dyadic length can be decomposed. Therefore multiscale filters cannot be simply used in online applications. Nounou and Bakshi [19] proposed online method for multiscale rectification (OLMS) where the signal is denoised as in section 2.3 but in a moving window of dyadic length (see Fig. 3). Since noncausal wavelet filters introduce distortion at the end of the signal, causal wavelet filters (for example 'haar') or boundary corrected wavelet filters are recommended in online denoising [19]. If not so, the last point (which is used for difficult-to-measure variable estimation) is the least accurate. Another way to deal with border distortion introduced by noncausal wavelet filter is to use window extension such as smooth window extension or symmetric window extension [20].

These problems do not appear when complete input output sample is needed for model parameter update due to the long duration of laboratory analysis.

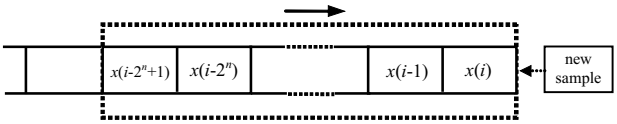


Figure 3 *Moving window in OLMS method*
Slika 3. *Pomični prozor u OLMS metodi*

Simple linear filters such as mean filter and exponentially moving average filter are very easy to implement online:

$$x_s(i) = \frac{1}{W}\left[x(i) + x(i-1) + \ldots + x(i-W+1)\right], \quad (5)$$

$$x_s(i) = \alpha x(i) + (1-\alpha)x_s(i-1), \quad (6)$$

where α is adjustable smoothing parameter. However, their drawback is their single-scale nature so simultaneous noise removal and accurate feature representation cannot be effectively achieved [19].

2.5
Online outlier and abnormal situation detection
Online otkrivanje stršećih vrijednosti i stanja kvara

While the missing values and obvious outliers are relatively easy to identify (e.g. by comparing with min/max process variable values), non-obvious outliers detection is much harder procedure. In online applications there is a need to check whether the newly collected sample of easy-to-measure variables is correct or it is an outlying observation. Univariate outlier detection methods, such as Hampel filter, can be easily implemented like a moving window filters [9]. But outliers that are multivariate in nature cannot be simply detected with these techniques.

Multivariate statistical process control (MSPC) methods are known to be effective for detecting and diagnosing abnormal process conditions such as sensor faults and process faults [21]. A frequently used method for process monitoring is Principal Component Analysis (PCA) which extracts a few independent components from highly correlated input variable space and use these components to monitor process operation [7]. The extraction is based upon decomposition of the input variables data matrix $\mathbf{X}^{n\times m}$ which is scaled to zero mean and unit variance:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \widetilde{\mathbf{T}}\widetilde{\mathbf{P}}^{\mathrm{T}} = \left[\mathbf{T}\ \ \widetilde{\mathbf{T}}\right]\left[\mathbf{P}\ \ \widetilde{\mathbf{P}}\right]^{\mathrm{T}} = \overline{\mathbf{T}}\,\overline{\mathbf{P}}^{\mathrm{T}}, \quad (7)$$

where $\mathbf{T}^{n\times l}$ and $\mathbf{P}^{m\times l}$ are the principal component scores and loadings, $\mathbf{E}^{n\times m}$ is residual, n is a number of samples, m is a number of input variables and l is a number of principal components. Correlation matrix of the input variables can be approximated as:

$$\mathbf{R} \approx \frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \overline{\mathbf{P}}\,\overline{\mathbf{\Lambda}}\,\overline{\mathbf{P}}, \quad (8)$$

where $\overline{\mathbf{\Lambda}}$ is diagonal matrix with all eigenvalues of matrix $\mathbf{R}$ in descending order. Since columns of $\mathbf{P}$ are eigenvectors of $\mathbf{R}$ assosicated with l largest eigenvalues, and $\widetilde{\mathbf{P}}$ are the remaining vectors, calculation of the $\mathbf{P}$ (PCA model) is reduced to eigenvector problem.

After developing PCA model, process monitoring is based on monitoring of the two statistics. Q statistics (or SPE index) measures the variability in the residual subspace:

$$\mathrm{Q} = \left\|\widetilde{\mathbf{x}}\right\|^2 = \left\|(\mathbf{I} - \mathbf{P}\mathbf{P}^{\mathrm{T}})\mathbf{x}\right\|, \quad (9)$$

and Hotelling's T^2 statistics measures variations in the principal component subspace:

$$T^2 = \mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{\Lambda}^{-1}\mathbf{P}^{\mathrm{T}}\mathbf{x}. \quad (10)$$

The process is considered normal if:

$$\mathrm{Q} \le \delta_\alpha^2 \ \text{ and } \ T_\alpha^2 \le \chi_{l;\alpha}^2. \quad (11)$$

where α is given significance level for which selection exists several recommendations [21]. An abnormal situation will cause at least one of the two indices to exceed the limit. However, it must be pointed out that if the sample exceeds T^2 limit but does not violate Q limit, it can be a fault, but it can be a result of the process operation region change, too.

Most industrial processes are time-varying and non-stationary. To accommodate varying behaviour and to still be able to successfully detect abnormal situations in online implementions, adaptive PCA model has to be used. A complete adaptive monitoring scheme was proposed in [22]. A similar scheme is used in this paper for erroneous samples detection:

0. Offline preprocess initial data set – select easy-to-measure variables, check if there are gross errors in the initial data set, reconstruct such samples or delete them from the data set and denoise measured data.

1. For initial data set calculate PCA model and upper control limits for statistics.
2. If there is a new sample **x**, calculate Q and T^2. If one of the confidence limits is violated, then outlying observation or abnormal situation occurred in the process and an alarm is raised to the process operator. Adaption of the model which is used for difficult-to-measure variable estimation is skipped. If abnormal situation is not detected then PCA model is recursively updated and control limits are recursively calculated.

Recursive PCA encompasses several steps [22]:

- recursive calculation of means and standard deviations of variables,
- recursive calculation of the correlation matrix,
- recursive determination of the number of principal components,
- recursive determination of the confidence limits for Q and T^2 statistics since limits can vary with time.

3
Methods for process model building
Metode za izgradnju modela procesa

Measured data taken from the process database are used for initial model developing by some offline method for process modeling (see section 3.1). Such model then serves as a basis of an estimator, which can be adaptive, i.e. model parameters are updated in online fashion during estimator use (see section 3.2).

3.1
Process model structuring
Strukturiranje modela procesa

Generally, data based process model building is in fact a search for an approximating function $\boldsymbol{f}_m(\cdot)$ which approximates the unknown natural functional dependence of the output process variable on the selected input variables. A function with a finite-dimension vector of parameters is commonly used for the approximating function, so that the model can be represented by:

$$\hat{y} = \boldsymbol{f}_m(\mathbf{x}, \boldsymbol{\Theta}), \tag{12}$$

where **x** stands for input variables vector, **Θ** vector of parameters of the approximating function $\boldsymbol{f}_m(\cdot)$, i.e. process model, $\hat{y}$ model output (estimated value of the difficult-to-measure process variable) [23].

For the chosen model structure its parameters **Θ** have to be determined based on initial data set. When designing a difficult-to-measure process variable estimator, the process model must have good prediction properties. In prediction model building the parameters are usually determined by regression, in which all model parameters are estimated based on minimization of output error of approximation:

$$\Im(\boldsymbol{\Theta}) = \frac{1}{2}\sum_{i=1}^{n}\left[y_i - \boldsymbol{f}_m(\mathbf{x}_i, \boldsymbol{\Theta})\right]^2 = \frac{1}{2}\sum_{i=1}^{n} e_i^2(\boldsymbol{\Theta}), \tag{13}$$

as in MLR (Multiple Linear Regression) and MLP (Multilayer Perceptron) models. Since the plant data are typically poorly informative, the estimation of the difficult-to-measure variable is usually performed on the great number of process variables which are in any correlation with the difficult-to-measure variable (see section 2.1). This results in highly correlated input space, so the reliability of parameter estimation by regression according to criterion (13) is usually very low.

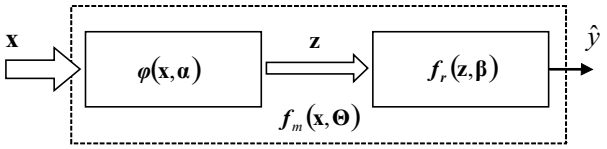


Figure 4 *Principle schema of the process model divided into two levels of projection (so-called two-level model)*
Slika 4. *Načelna shema modela procesa podijeljenog na dvije razine preslikavanja (tzv. dvo-razinski model)*

One way to improve regression modeling based on plant data is to use methods based on the input space projection into a latent subspace [23]. Hereby, a high-dimensional, correlated input space is at first projected into an adequate low-dimensional subspace, and the regression is performed on these new (latent) variables which are obtained from the projection. The projection to the latent space can be also treated as a preprocessing step in a process model building because the goal is to project important features from the input space into latent space and to discard unimportant phenomena. The model structure which is based on the input space projection into a latent subspace is shown in Fig. 4, where $\boldsymbol{\varphi}(\cdot)$ is input space projection function, $\boldsymbol{f}_r(\cdot)$ regression function, **z** latent variables, α and β are model parameters. The related model is a composition of two functions and can be presented with:

$$\hat{y} = \boldsymbol{f}_m(\mathbf{x}, \boldsymbol{\Theta}) = \boldsymbol{f}_r\left(\boldsymbol{\varphi}_j(\mathbf{x}; \boldsymbol{\alpha}_j); \boldsymbol{\beta}\right), \quad j = 1, 2, ..., l, \tag{14}$$

where l is the number of latent variables. Apart from the selection of the appropriate functions, this model structure requires definition of two separate criteria for parameter estimation.

For the input space projection different linear or nonlinear methods can be used. The most important linear methods are already mentioned PCA (see section 2.5), PLS (Partial Least Squares) [7] and CR (Continuum Regression) [23, 24]. CR method has the best properties but parameter estimation according to this method is complex and therefore not suitable for adaptive model building. PLS method results in a model with good prediction capabilities (although slightly worse than CR method) and parameter estimation is much simpler in regard to CR method so it is often used in adaptive model building.

In the second level of the process model linear or nonlinear regression can be performed. Linear regression function is usually used in adaptive models. Example of such model with two levels of projection is PLSR model where linear regression is performed on latent variables obtained by PLS method.

3.2
Adaptive process model
Adaptivni model procesa

Most industrial processes are time-varying and non-stationary and thus require adaptive rather than model with constant parameters. Apart from that, if initial data set contains lots of different errors in the measured data, then a

model with constant parameters (that are determined offline) will (probably) perform very poorly in online working whereas adaptive model can achieve good prediction capabilities after some time.

If both functions $\boldsymbol{\varphi}(\cdot)$ and $\boldsymbol{f}_t(\cdot)$ of the model (14) are linear, the resulting model can be presented as a simple one-level linear model in a way that parameters α and β are combined into a single parameter b.

First recursive partial least square (RPLS) algorithm was proposed by Helland [25]. Improvements of original RPLS were reported in [5]. PLS is based on decomposition on predictor matrix $\mathbf{X}^{n\times m}$ (easy-to-measure variables) and response matrix $\mathbf{Y}^{n\times p}$ (difficult-to-measure variables) into sums of rank one component matrices [26]:

$$\mathbf{X} = \sum_{i=1}^{l} \mathbf{t}_i \mathbf{p}_i^{\mathrm{T}} + \mathbf{E} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}, \tag{15}$$

$$\mathbf{Y} = \sum_{i=1}^{l} \mathbf{u}_i \mathbf{q}_i^{\mathrm{T}} + \mathbf{F} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F}, \tag{16}$$

where **t** and **u** are latent score vectors, **p** and **q** are corresponding loading vectors, **E** and **F** are the input and output residual matrices and l is the number of latent variables. These two equations are called outer PLS model. The latent score vectors are related by linear inner model:

$$\mathbf{U} = \mathbf{T}\mathbf{B}, \tag{17}$$

where **B** is diagonal matrix containing regression coefficients of the score model determined by PLS algorithm. Model prediction for the new samples $\mathbf{X}_t$ is given by:

$$\hat{\mathbf{Y}}_t = \mathbf{X}_t \mathbf{B}_{\mathrm{PLS}} = \mathbf{X}_t \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{B}\mathbf{Q}^{\mathrm{T}}, \tag{18}$$

where weights **W** are determined by PLS algorithm. In the case of one output variable the PLS model can be seen as a model divided into two levels of projection where the outer PLS model corresponds to the input space projection and the inner PLS model to the regression part of the model (14).

Recursive PLS algorithm consists of several steps:

0. Offline preprocess initial data set.
1. Scale initial data matrices $\mathbf{X}_0$ and $\mathbf{Y}_0$ to zero mean and unit variance. Derive initial PLS model: $\{\mathbf{X},\mathbf{Y}\}\{\mathbf{T},\mathbf{W},\mathbf{P},\mathbf{B},\mathbf{Q}\}$.
2. When new pair of data $\{\mathbf{x},\mathbf{y}\}$ is available check if sample is an outlier or if abnormal situation occurred in the process. In that case the model is not updated. If sample is correct, apply online denoising technique, normalize it and formulate new **X** and **Y**:

$$\mathbf{X} = \begin{bmatrix} \mathbf{P}^{\mathrm{T}} \\ \mathbf{x} \end{bmatrix}, \qquad \mathbf{Y} = \begin{bmatrix} \mathbf{B}\mathbf{Q}^{\mathrm{T}} \\ \mathbf{y} \end{bmatrix}. \tag{19}$$

Derive PLS model $\{\mathbf{X},\mathbf{Y}\}\{\mathbf{T},\mathbf{W},\mathbf{P},\mathbf{B},\mathbf{Q}\}$, recalculate normalizing parameters and wait for a new complete data pair.

This recursive scheme can be simply extended to block-wise RPLS, moving window PLS or RPLS with forgetting factor [5].

4
Estimation of difficult-to-measure process variable
Procjena teško-mjerljive procesne veličine

Process model serves as the basis of a difficult-to-measure variable estimator. Data preprocessing techniques and methods for process model building discussed in Section 2 and 3 are applied to the following examples: modeling of the distillation process based on data from process database and modeling of the fluid storage process based on data obtained by simulation. The influence of denoising techniques on the quality of the built models is compared in the first example while in the second example the importance of the data preprocessing in the adaptive model building is pointed out. Besides that, benefits of an adaptive model are shown with respect to a model with constant parameters.

4.1
Example I – oil viscosity estimation
Primjer I – procjena viskoznosti ulja

Viscosity is the most important quality indicator of the final product of the distillation column but it belongs to the difficult-to-measure process variable category. To estimate the viscosity of the final products, it is necessary to have a mathematical model of the process which can be built from the measured data by methods from previous section.

Measured data were taken from the database of the vacuum distillation column. Available data set has measurements of 45 easy-to-measure variables (25 temperatures, 16 flows and 4 pressures) obtained by sensors and measurements of the oil viscosity at four different heights of the distillation column obtained by laboratory analysis. These easy-to-measure process variables are potential process model inputs for estimation of viscosity as shown in Figure 5.

The selection of easy-to-measure variables for process model building was made according to the qualitative evaluation of variable significance made by process engineers and operators. According to the given evaluation, sets with 12, 25 and 45 variables were made. The first set contains 12 variables with the highest influence on the oil viscosity. The second set is equal to the first set with additional 13 variables which have a modest influence on the oil viscosity and the third set contains the second set and additional 20 variables with minimal affect on the oil viscosity.

Figure 5 *Principle schema of a viscosity estimator*
Slika 5. *Načelna shema procjenjivača viskoznosti*

The easy-to-measure variables were measured every 5 minutes during a period of 100 days, whereas the viscosity of the products was measured by laboratory analysis approximately every 3 hours in the same time period. So, 29472 samples of easy-to-measure variables and 833 samples of oil viscosity are available. Graphical inspection of the data revealed presence of noise and outliers in the easy-to-measure variables. Samples that were suspected to

be outliers were removed. Totally, 207 samples were removed, so 29265 samples of easy-to-measure variables remained and the number of complete samples dropped to 804.

The viscosity of two products with process designations UD1 and UD3 were chosen for process model outputs. The first 655 samples were used in building the process model (training data) and the other 149 samples were used to test the models.

The formed data sets were used to build linear process models with continuum regression method (CR model). PLS models in this case obtain similar results, but slightly worse. Model evaluation was performed by output estimation on the test data. For quality indicators of the process models, mean squared error (MSE) and coefficient of determination (R^2) on test data, and maximal percentage error (MPErr) as an additional indicator, were chosen.

To compare effectiveness of denoising techniques, data was denoised with wavelet filter in one case and with simple moving average filter in another case (see section 2.3).

Table 1 *Best results obtained by CR models*
Tablica 1. *Najbolji rezultat postignuti CR modelima*

data set for modeling	Model quality indicators				dim. of proj.	contin. param. γ
	test MSE	test R^2	train R^2	MPErr$_t$, %		
MOVING AVERAGE FILTER						
UD1-12	0,00427	0,6769	0,8852	7,11	8	7
UD1-25	0,00209	0,8415	0,9283	6,17	11	2,64
UD1-45	0,00215	0,8372	0,9458	6,14	18	2,08
UD3-12	0,08019	0,5196	0,3250	8,09	5	1,22
UD3-25	0,07732	0,5368	0,3911	7,73	11	2,33
UD3-45	0,10052	0,3978	0,4619	7,33	4	0,6
WAVELET FILTER						
UD1-12	0,00298	0,7744	0,8915	5,82	7	3
UD1-25	0,00164	0,8755	0,9365	6,17	11	2,33
UD1-45	0,00185	0,8599	0,9564	5,71	16	1,5
UD3-12	0,06887	0,5874	0,3604	7,82	5	1,22
UD3-25	0,06709	0,5981	0,4280	6,55	7	1,67
UD3-45	0,09138	0,4189	0,4525	7,18	17	5,68

Results of different data denoising in CR model building are presented in Tab. 1. Differences between prediction capabilities of the two models can be also seen in Fig. 6 where models output for the test data set is shown together with the values obtained by laboratory analysis. The models that were built on the wavelet filtered data sets have considerably better prediction capabilities than models that were built on the data sets smoothed with moving average (MA) filter. This means that noise was present in a wide frequency range in the available data and the multiresolution analysis had efficiently removed it, without smoothing out significant features of the original data. It can be concluded that wavelet filter is probably more promising than single scale filters in online applications also.

According to the calculated indicators, it is obvious that the estimation of the UD3 variable is poor with respect to the UD1 variable. One possible reason is high nonlinearity of the UD3 process which cannot be properly described with the linear model. Another reason might be a low correlation of the input variables and the UD3 variable. With an increase of input variables from 12 to 25, the quality of the models is enhanced, especially in the case of the UD1 process, so it can be concluded that extra knowledge about the UD1 process is incorporated in the added variables. However, models that were built on the data set containing 45 variables had no better prediction capabilities, even much worse in the case of the UD3 variable estimation. This supports the process engineers' and operators' statements

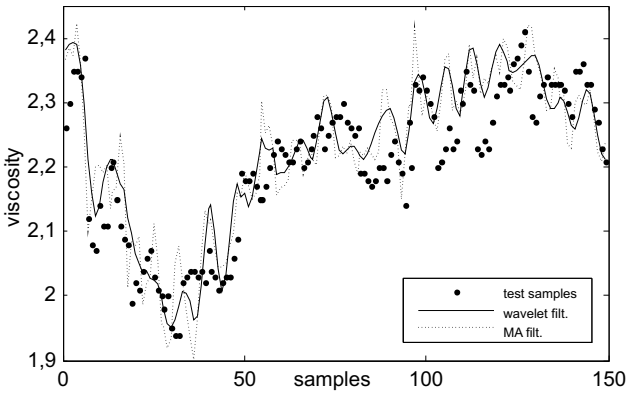


Figure 6 *UD1 viscosity estimated by CR model with 12 input variables, with differently denoised easy-to-measure variable data in training set*
Slika 6. *Procjena viskoznosti UD1 pomoću CR modela s 12 ulaznih veličina, uz različito filtrirane podatke o lako-mjerljivim veličinama u skupu za učenje*

about very low influence of the 20 variables that were added to the data set '25', to the output variables and they act just like noise.

4.2
Example II – modelling of fluid storage process
Primjer II – modeliranje procesa uskladištenja fluida

Principal schema of the complex process of fluid storage, which is a basis for the simulation model, is depicted in Fig. 7. It is supposed that difficult-to-measure process variable is level of the fluid in the third tank (h_2) and easy-to-measure variables are flows q. Also values of the positions x of the controlled valves are supposed to be known, so a total of 13 easy-to-measure variables is available. During simulation, which covered 25 hours, easy-to-measure variables were sampled every 6 seconds and level h_2 was sampled every 5 minutes. Totally 15000 samples of easy-to-measure variables and only 300 samples of difficult-to-measure variable were generated. In order to achieve a more realistic situation, noise and outliers were added to the flow measurements.

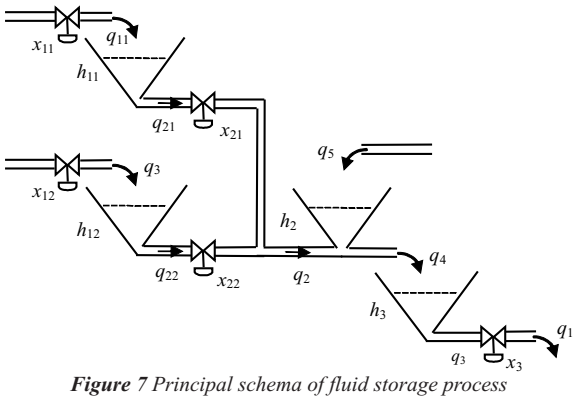


Figure 7 *Principal schema of fluid storage process*
Slika 7. *Načelna shema procesa pohrane tekućine*

The first 2500 samples, which are used for initial process model building, were preprocessed offline. Due to the lower sampling frequency of level h_2, only 50 complete input-output samples are available for offline process model building. The rest of the data (12500 samples) are used for testing non-adaptive and adaptive process models.

In order to simulate abnormal situation, bias of q_{22} and q_4 sensor was introduced between 2000^{th} - 3000^{th} sample and

drift of q_{21} sensor between 3500th - 4000th in the test data. To simulate non-stationary and time-varying process behaviour, operating point change was introduced at 5000th and the gain of one of the valves was slightly changed (after 9500th sample).

PLS model with constant parameters, built on first 2500 samples, was used for level h_2 estimation. Its prediction capabilities were tested on the test data in "online manner". Moving average filter was implemented for online data denoising. From Fig. 8 it is obvious that after the operating point change estimation precision of the PLS model is decreased. Since outlier detection and replacement algorithms were not implemented, sensor faults (bias of q_{22} and q_4 sensor between 2000th - 2500th sample) caused great error in h_2 estimation. Therefore, it is important to detect such situations and raise alarm to the process operator. If algorithm for reconstruction is available, it has to be triggered before estimation procedure.

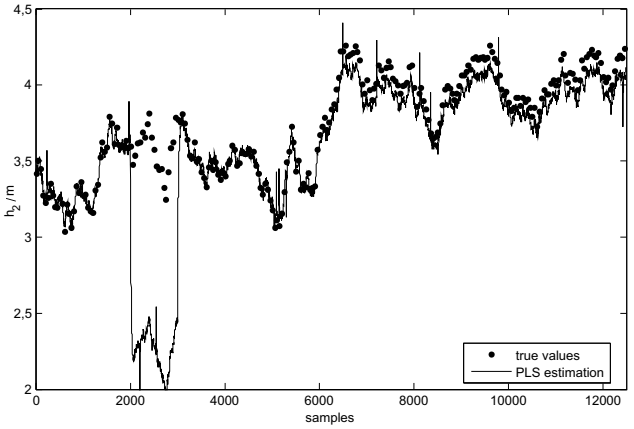


Figure 8 *Level h_2 estimation by PLS model*
Slika 8. *Procjena razine h_2 pomoću PLS modela*

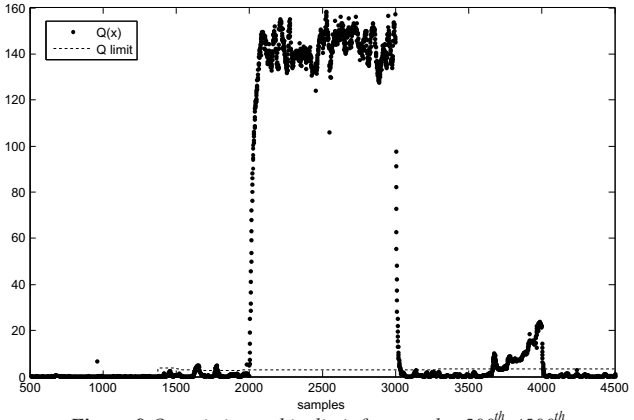


Figure 9 *Q statistics and its limit for samples 500th-4500th*
Slika 9. *Q statistika i njen limit u rasponu uzoraka od 500.-4500.*

For the purpose of online outlier detection recursive PCA model is applied (see section 2.5). Outliers in the data set and sensor bias and drifts were successfully detected with PCA model and Q statistics (see Fig. 9). Samples associated with bias and drift of sensors clearly violated Q statistics limit.

Fig. 10 shows level estimation with the PLS model that was updated in a moving average fashion (see section 3.2). Only the samples that are considered error free according to the recursive PCA model (see Fig. 9) were used for model updating. Simple moving average filter was implemented

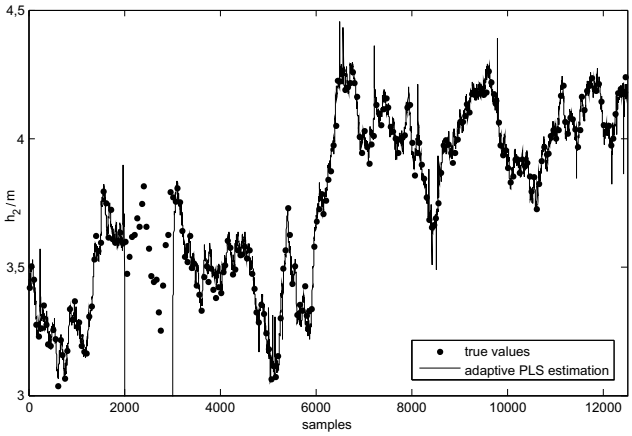


Figure 10 *Level h_2 estimation by adaptive PLS model*
Slika 10. *Procjena razine h_2 pomoću prilagodljivog PLS modela*

for online denoising. By looking at Figures 8 and 10, it can be concluded that the adaptive model tracks process changes successfully after the operating point change. Similar situation occurred as in constant PLS model testing between 2000th and 2500th sample because algorithm for outlier reconstruction was not implemented.

Fig. 11 shows comparison of constant PLS model and adaptive PLS model in the terms of absolute error of estimation for 3000th-8000th sample interval. It can be noticed that the PLS model with constant parameters looses its estimation precision after operating point change.

However, precision of an adaptive model can be decreased if erroneous samples are used for model adaptation. Estimation of the level h_2 by two differently updated PLS models is shown in Fig. 12. Parameters of the first adaptive model are updated on every available complete input-output sample while for the adaptation of the second one only error free samples are used (see Fig. 9 and 10). It can be seen that the first adaptive PLS model has lower estimation precision in 3050th-3500th sample interval compared to the second adaptive model because it has been updated on the erroneous samples (due to the bias of q_{22} and q_4 sensor in 2000th-3000th sample interval). Precision degradation is clearly visible in Tab. 2 in terms of mean squared error. Constant PLS model even shows better prediction capabilities in this sample interval. It has to be pointed out that the model precision can be recovered after a while if error free samples are used for model updating.

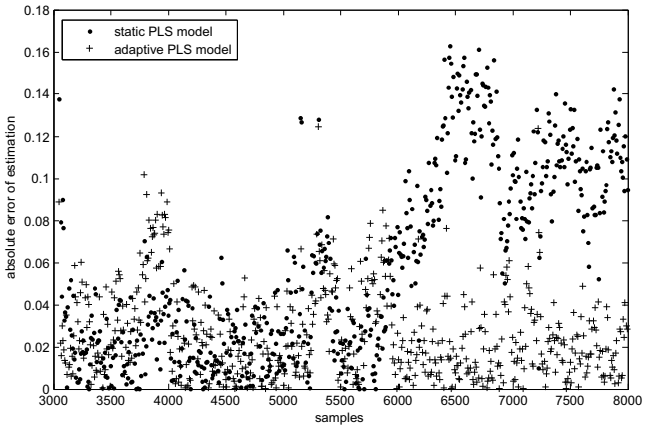


Figure 11 *Comparison of constant PLS and adaptive PLS model in terms of absolute error of estimation for 3000th-8000th sample interval*
Slika 11. *Usporedba procjene konstantnog PLS i prilagodljivog PLS modela pomoću apsolutne pogreške estimacije za raspon uzoraka od 3000.-8000.*

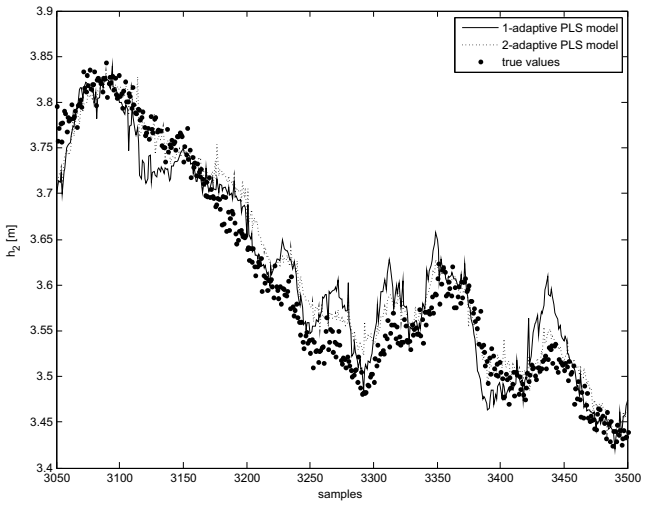


Figure 12 *Level h_2 estimation by two differently updated PLS models for 3050^{th}-3500^{th} sample interval*
Slika 12. *Procjena razine h_2 pomoću dva različito prepodešavana PLS modela za interval mjernih uzoraka od 3050.-3500.*

Table 2 *Mean squared error of PLS models for 3050^{th}-3500^{th} sample interval*
Tablica 2. *Srednja kvadratna pogreška PLS modela za raspon mjernih uzoraka od 3050.-3500.*

Model	Mean squared error
1-adaptive PLS model	0,001183
2-adaptive PLS model	0,000868
constant PLS model	0,001176

5
Conclusion
Zaključak

An important step in data based process modeling is preprocessing of easy-to-measure variables since it has a great influence on the model prediction capabilities. Data preprocessing includes three substeps: selection of easy-to-measure variables, outliers detection and easy-to-measure variable denoising. Apart from offline preprocessing of initial data set (training set), it is also necessary to preprocess samples of easy-to-measure variables that continuously arrive in online application of an estimator.

From the obtained results of the distillation process modeling it is clear that some of the available process variables are not relevant for process variable estimation of interest and they can negatively affect parameter estimation procedure. According to the presented results, models that are built on the wavelet denoised data have better prediction capabilities than models built on the MA filtered data which means that the data denoised by wavelet analysis are more informative than the data filtered with moving average filter. Wavelet analysis efficiently removes noise from the broader frequency region without smoothing out important signal features. It is reasonable to expect that these filter properties will be kept in online data denoising.

Results of the modeling of simulated fluid storage process show that adaptive model is needed when dealing with time-varying or non-stationary process. However, model parameter updating has to be carried out on error free samples. Additionally, erroneous samples of easy-to-measure variables cause great error in difficult-to-measure process variable estimation. Therefore, quality and efficient online data denoising and outlier detection are of great importance in online implementations.

6
References
Literatura

[1] McAvoy, T. J. Intelligent “Control” applications in the Process Industries. // Annual Reviews in Control. 26, (2002), str. 75-86.
[2] Kadlec, P.; Gabrys, B.; Strandt, S. Data-driven Soft Sensors in the process industry. // Computers and Chemical Engineering, 33, 4(2009), str. 795-814.
[3] Slišković, D.; Grbić, R.; Nyarko, E. K.; Data Preprocessing in Data Based Process Modeling. // in 2nd IFAC International Conference on Intelligent Control Systems and Signal Processing, 2009.
[4] Qin, S. J. Neural Networks for Intelligent Sensor and Control – Practical Issues and Some Solutions. Neural Networks for Control, Chapter 8, D. Elliott, Ed. Academic Press, 1996.
[5] Qin, S. J. Recursive PLS algorithms for adaptive data modeling. // Computers and Chemical Engineering, 22, 4(1998), str. 503-514.
[6] Kadlec, P.; Grbić, R.; Gabrys, B. Review of adaptation mechanisms for data-driven soft sensors. // Computers & Chemical Engineering, 35, 1(2011), str. 1-24.
[7] Martens, H.; Naes, T. Multivariate Calibration, 2nd edition. John Wiley & Sons, New York, 1991.
[8] Qin, S. J.; McAvoy, T. J. A data-based process modeling approach and its applications. // In reprints of the 3rd IFAC Dycord Symposium, 1992., str. 321–326.
[9] Pearson, R. K. Outliers in Process Modeling and Identification. // IEEE transactions on control system technology, 10, 1(2002), str. 55-63.
[10] Rousseeuw, P. J.; Leroy, A. M. Robust regression and outlier detection, Wiley, New York, 2003.
[11] Chiang, L. H.; Pell, R. J.; Seasholtz, M. B. Exploring process data with the use of robust outlier detection algorithms. // Journal of Process Control, 13, 5(2003), str. 437–449.
[12] Vetterli, M.; Herley, C. Wavelets and filter banks: theory and design. // IEEE Trans. Signal Process., 40, 9(1992), str. 2207-2232.
[13] Mallat, S. G. A Theory for Multiresolution Signal Decomposition: The Wavelet representation. // IEEE Transactions on pattern analysis and machine intelligence, 11, 7(1989), str. 674-693.
[14] Donoho, D. L. De-noising by soft-thresholding. // IEEE Trans. Information Theory, 41, 3(1995), str. 613–627.
[15] Shao, R.; Jia, F.; Martin, E. B.; Morris A. J. Wavelets and non-linear principal component analysis for process monitoring. // Control Engineering Practice, 7, 7(1999), str. 865 – 879.
[16] Donoho, D. L.; Johnstone, I. M. Ideal spatial adaptation via wavelet shrinkage. // Biometrika, 81, 3(1994), str. 425–455.
[17] Donoho, D. L.; Johnstone, I. M. Minimax estimation via wavelet shrinkage. // Ann. Statist., 26, (1998), str. 879–921.
[18] Stein, C. M. Estimation of the mean of a multivariate normal distribution. // Ann. Statist., 9, 6(1981), str. 1135–1151.
[19] Nounou, N. M.; Bakshi, B. R. On-Line Multiscale Filtering of Random and Gross Errors without Process Models. // AIChE Journal, 45, (1999), str. 1041-1058.
[20] Xia, R.; Meng, K.; Qian, F.; Wang, Z. Online Wavelet Denoising via a Moving Window. // Acta Automatica Sinica, 33, 9(2007), str. 897-901.
[21] Qin, S. J. Statistical Process monitoring: basics and beyond. // Journal of chemometrics, 17, 8-9(2003), str. 480-502.
[22] Li, W. H.; Yue, H.; Valle-Cervantes, S.; Qin, S. J. Recursive PCA for adaptive process monitoring. // Journal of Process Control, 10, 5(2000), str. 471-486.
[23] Slišković, D.; Perić, N.; Petrović, I. Continuum Regression in Process Modeling Based on Plant Data, // Automatika, 46, 3-4(2005), str. 173-184.
[24] DeJong, S.; Wise, B. M.; Ricker, N. L. Canonical partial least squares and continuum power regression. // Journal of Chemometrics, 15, (2001), str. 85–137.

[25] Helland, K.; Berntsen, E. H.; Borgen, O. S.; Martens, H. Recursive algorithm for partial least squares regression. // Chemometrics and Intelligent Laboratory Systems, 14, (1991), str. 129–137.
[26] Geladi, P.; Kowalski, B. R. Partial least-squares regression: A tutorial. // Anal. Chim. Acta, 185, 1(1986), str. 1-17.

Authors' addresses
Adrese autora

Doc. dr. sc. Dražen Slišković
Faculty of Electrical Engineering
J. J. Strossmayer University of Osijek
Kneza Trpimira 2B
HR-31000 Osijek, Croatia
email: drazen.sliskovic@etfos.hr

Ratko Grbić, dipl.ing.
Faculty of Electrical Engineering
J. J. Strossmayer University of Osijek
Kneza Trpimira 2B
HR-31000 Osijek, Croatia
email: ratko.grbic@etfos.hr

Prof. dr. sc. Željko Hocenski
Faculty of Electrical Engineering
J. J. Strossmayer University of Osijek
Kneza Trpimira 2B
HR-31000 Osijek, Croatia
email: zeljko.hocenski@etfos.hr