Distillation End Point Estimation in Diesel Fuel Production

I. Mohler, Ž. Ujević Andrijić, N. Bolf, and G. Galinec
Faculty of Chemical Engineering and Technology, University of Zagreb,
Department of Measurements and Process Control, Savska c. 16/5a,
HR-10 000 Zagreb, Croatia

Original scientific paper
Received: May 25, 2012
Accepted: January 3, 2013

Soft sensors for the on-line estimation of kerosene 95 % distillation end point (D95) in crude distillation unit (CDU) are developed. Experimental data are acquired from the refinery distributed control system (DCS) and include on-line available continuously measured variables and laboratory data which are consistently sampled four times a day. Additional laboratory data of kerosene D95 for the model identification are generated by Multivariate Adaptive Regression Splines (MARSplines).

Soft sensors are developed using different linear and nonlinear identification methods. Among the variety of dynamic models, the best results are achieved with Box Jenkins (BJ), Output Error (OE) and Hammerstein–Wiener (HW) model. Developed models were evaluated based on the Final Prediction Error (FPE), Root Mean Square Error (RMSE), mean Absolute Error (AE) and FIT coefficients. The best results for diagnostic purposes show BJ model. For continuous estimation of D95, OE and HW models can be used.

Key words:
Crude distillation unit, distillation end point, soft sensor, identification

Introduction

Strict product quality requirements and pollutant emission standards impose the need for effective measurement and process control in industrial plants. Therefore, a large number of process variables need to be monitored using appropriate measuring devices. The main problems are expensive analyzers and unreliability of on-line instrumentation.

Soft sensors are focused on assessing system state variables and product quality, thus replacing physical sensors and laboratory analysis. Application of soft sensors for estimating non-available or hard-to-measure process variables is very interesting in the process industry. Usually, there are a large number of continuously measured values, and these may serve as input signals for the soft sensor.[1] They can work in parallel with real sensors, analyzers, measuring devices, allowing fault detection schemes devoted to the sensor's status analysis to be implemented.[2–3] Furthermore, they can take the place of sensors which are down for maintenance, in order to keep control loops working properly and guarantee product specification without undertaking conservative production policies, which are usually very expensive.

Different model structures can be used to model real systems. In the field of industrial applications, the focus is on parametric (polynomial) structures in both linear and nonlinear versions.[4] In the last decade, soft sensor applications for the distillation unit product properties have been studied extensively.[5–13] In most industrial applications, the soft sensor design procedure based on data-driven approaches follows the sequence of the stages shown in Fig. 1.

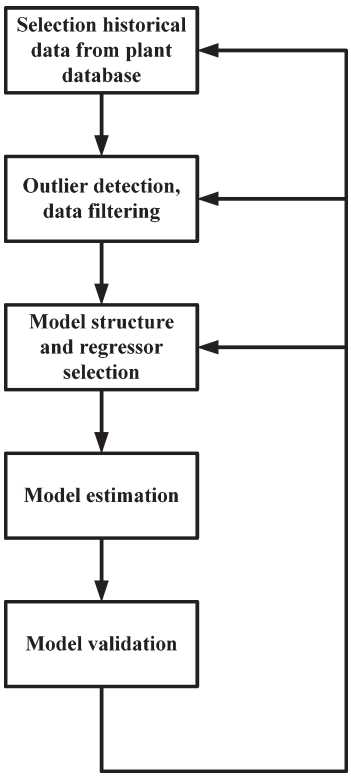


Fig. 1 – *Block scheme of the soft sensor identification procedure*

*Corresponding author: phone: +385 1 4597 148; e-mails: imohler@fkit.hr, zujevic@fkit.hr, bolf@fkit.hr, ggalin@fkit.hr.

Model development

Since the development of dynamic models demands an equal number of input and output data, additional output data were generated by Multivariate Adaptive Regression Splines algorithm (MARSpline). MARSplines algorithm operates as multiple piecewise linear regression, where each breakpoint estimated from the data defines the "region of application" for a particular (very simple) linear equation.[14] The MARSplines algorithm builds models from two-sided truncated functions (basis functions) of the predictors (x) with the following form:

$$(x-k)_+ = \begin{cases} x-k & x>k \\ 0 & else \end{cases} \quad (1)$$

The MARSplines for a dependent variable y, and M terms, can be summarized as:

$$H_{km}(x_{v(k_0m)}) = \prod_{k=1}^{k} h_{km} \quad (2)$$

where x is the predictor in the $k'th$ of the $m'th$ product. For order of interactions K=1 the model is additive, and for K=2 the model pairwise interactive.

Linear model identification

A frequently used model for the on-line estimation is the OE model:

$$\hat{y}(k) = \sum_{i=1}^{i} \{\mathbf{B}_i(q)u(k-nk) + [\mathbf{I} - \mathbf{F}_i(q)]\hat{y}_i(k)\} \quad (3)$$

q is time-shift operator; $\hat{y}(k)$ is a model output at time k, $u(k)$ is an input at time k and i is the number of model inputs.

where:

$$\mathbf{B}_i(q) = \mathbf{1} + b_1 q^{-1} + b_2 q^{-2} + \ldots + b_{nb} q^{-nb-nk+1} \quad (4)$$

is polynomial matrix over q^{-1}, $\mathbf{B}_i$ is the matrix of dimensions $n(\hat{y})$*$n(\hat{y})$, b are the polynomial coefficients of polynomial matrix $\mathbf{B}_i(q)$, nb is the number of past input samples, and nk is input delay expressed by the number of samples,

$$\mathbf{F}_i(q) = \mathbf{1} + f_1 q^{-1} + f_2 q^{-2} + \ldots + f_{nf} q^{-nf} \quad (5)$$

$\mathbf{F}_i$ is the matrix of dimensions $n(\hat{y}) \cdot n(u)$, f are the polynomial coefficients of polynomial matrix $\mathbf{B}_i(q)$, nf is the number of past model output samples.

In order to obtain a BJ model that can describe the disturbance properties, the OE model can be expanded with a parametric disturbance matrix:

$$y(k) = \frac{\mathbf{B}_i(q)}{\mathbf{F}_i(q)} u(k) + \frac{\mathbf{C}_i(q)}{\mathbf{D}_i(q)} \xi(k) \quad (6)$$

where:

$$C_i(k) = 1 + c_1 q^{-1} + c_2 q^{-2} + \ldots + c_{nc} q^{-nc} \quad (7)$$

C_i is matrix with dimensions $n(y)$*$n(\xi)$, c are the polynomial coefficients of polynomial matrix $C_i(q)$, nc is the number of past prediction error,

$$\mathbf{D}_i(q) = \mathbf{1} + d_1 q^{-1} + d_2 q^{-2} + \ldots + d_{nd} q^{-nd} \quad (8)$$

is polynomial matrix over q^{-1};

$\mathbf{D}_i$ is matrix with dimensions $n(\hat{y})$*$n(e_s)$, d are the polynomial coefficients of polynomial matrix $\mathbf{D}_i(q)$, nd is the number of past simulated prediction errors.

Nonlinear model identification

While the linear model structure is fully defined by the chosen regressors, the nonlinear model structure additionally depends on nonlinear function characteristics.

It is quite a common situation that, while the dynamics itself can be well described by a linear system, there are static nonlinearities at the input and/or output. A model with a static nonlinearity at the input is called a Hammerstein model, whereas a model with output nonlinearity is a Wiener model.[13] The block diagram in Fig. 2 represents the structure of a Hammerstein-Wiener model.

Fig. 2 – *Structure of the Hammerstein-Wiener model*

$w(k) = f(u(k))$, is a nonlinear function transforming input data $u(k)$. $w(k)$ has the same dimension as $u(k)$.

$x(k) = (\mathbf{B}_i(q)/\mathbf{F}_i(q))w(k)$ is a linear transfer function, where $\mathbf{B}_i$ and $\mathbf{F}_i$ are polynomial matrices of the linear Output-Error model. $x(k)$ has the same dimension as $y(k)$.

$y(k) = h(x(k))$ is a nonlinear function that maps the output data $x(k)$ of the linear block to the system output. $w(k)$ i $x(k)$ are internal variables that define the input and output of the linear block, respectively.

Nonlinearity of the HW model is described by static neural network:

$$q(k) = \sum_{k=1}^{n} \alpha_k \kappa(\beta_k(x - y_k)) \quad (9)$$

The network is described with sigmoid function:

$$\kappa(s) = \frac{1}{1 + e^{-s}} \tag{10}$$

where α is scalar, β is a raw vector such that $\beta(x - \gamma)$ is a scalar, and n is the number of nonlinear units.[15]

The model structure selection step is strongly influenced by the purpose of the soft sensor design. Optimal model structure parameters were determined by optimization methods: Gauss-Newton, adaptive Gauss Newton, Levenberg Marquardt, gradient Search and partial least squares method.

OE and HW models do not require past samples of measured output (variable inferred by the soft sensor) when using validation data i.e. they depend only on previous measured inputs and previous model output.

The models were evaluated based on RMSE, AE, FIT and FPE values[15] defined by:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}} \tag{11}$$

$$\mathrm{AE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i) \tag{12}$$

$$\mathrm{FIT} = 1 - \frac{\sqrt{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}} \cdot 100, \tag{13}$$

where y is the measured output, $\hat{y}$ is the simulated or predicted model output, and $\bar{y}$ is the mean of y. 100 % corresponds to a perfect fit, and 0 % indicates that the fit is no better than guessing the output to be a constant ($\hat{y} = \bar{y}$).

Akaike's Final Prediction Error (FPE) criterion provides a measure of model quality by simulating the situation where the model is tested on an estimation set and independent validation set. According to Akaike's theory, the most accurate model has the smallest FPE.[13,15]

Akaike's Final Prediction Error (FPE) is defined by the following equation:

$$\mathrm{FPE} = V\left(\frac{1 + 2d}{N}\right) \tag{14}$$

where V is the loss function, d is the number of estimated parameters, and N is the number of values in the estimation data set.

The loss function V is defined by the following equation:

$$V = \det\left(\frac{1}{N} \sum_{i=1}^{N} \varepsilon(t, \theta_N)(\varepsilon(t, \theta_N))\right)^T \tag{15}$$

where θ_N represents the estimated parameters and ε is output model error

Process description

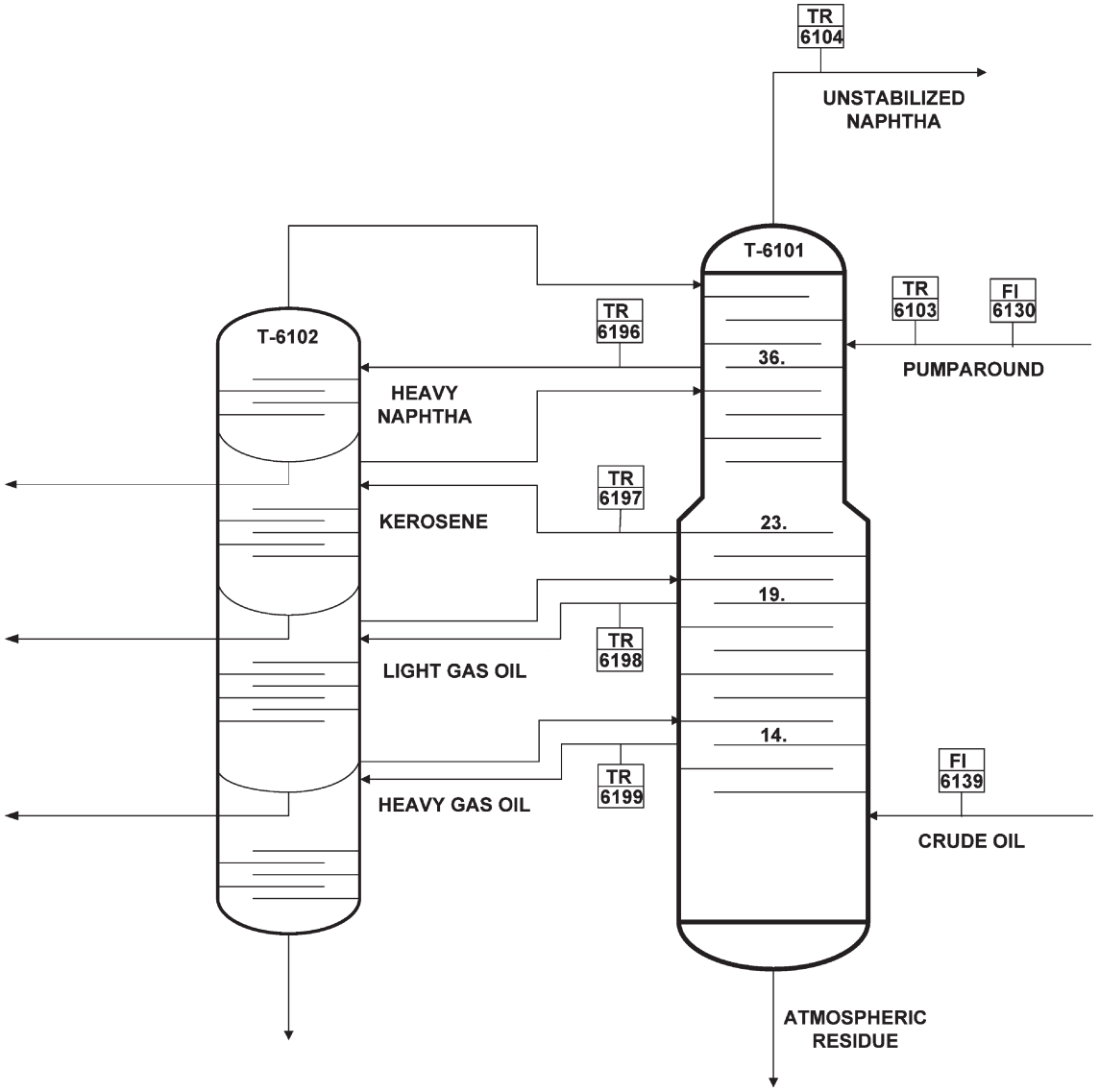
Since the CDU is the first unit in the sequence of refinery processing, it is essential that the quality of fractionation products (unstabilized naphtha, heavy naphtha, kerosene, light gas oil, heavy gas oil, atmospheric residue), be monitored and controlled. This requires that many properties should be measured online so that the unit can be effectively controlled through a feedback mechanism.[12] Heavy naphtha, petroleum, and light gas oil fractions are further used for blending of diesel fuel. These are being drained away as side fractions of the crude distillation column. Thereby, a very important product property to continuously measure and maintain is kerosene 95 % distillation point (D95).[16] Naphtha distillation properties are determined in the course of laboratory assays. Laboratory analyses are obtained by an automated distillation analyzer, which determines boiling range characteristics of various petroleum products at atmospheric pressure under appropriate conditions, based on the EN ISO 3405 standard (Petroleum products – Determination of distillation characteristics at atmospheric pressure). The end distillation point or final boiling point is defined as the maximum thermometer reading obtained during the test. However, because a fuel's end point is difficult to measure with good repeatability, the fuel's 95 % distillation point (D95) is commonly used. D95 must be maintained because very low values for D95 imply a shift to kerosene-oriented diesel fuels. This can decrease engine efficiency as well as increase maintenance requirements. Higher values for D95 can indicate very sloppy distillation operations and/or spiking with inappropriate components. Higher values can also increase soot going to the emissions control systems or into the atmosphere, and can increase maintenance requirements.

The section of the column with diesel fuel product and variables used for the estimation are given in Fig. 3.

Based on Pearson's correlation coefficients (R), PLS analysis and process expert experiences, the following variables have been chosen as the influence variables on distillation end point:

– column top temperature – (T_{TOP}), TR-6104;

– kerosene temperature – 23rd tray (T_K), TR-6197;



F i g . 3 – *Crude distillation column section*

– light gas oil temperature – 19th tray (T_{LGO}), TR-6198;

– heavy gas oil temperature –14th tray (T_{HGO}),TR-6199;

– pumparound temperature – (T_{PA}), TR-6103 and

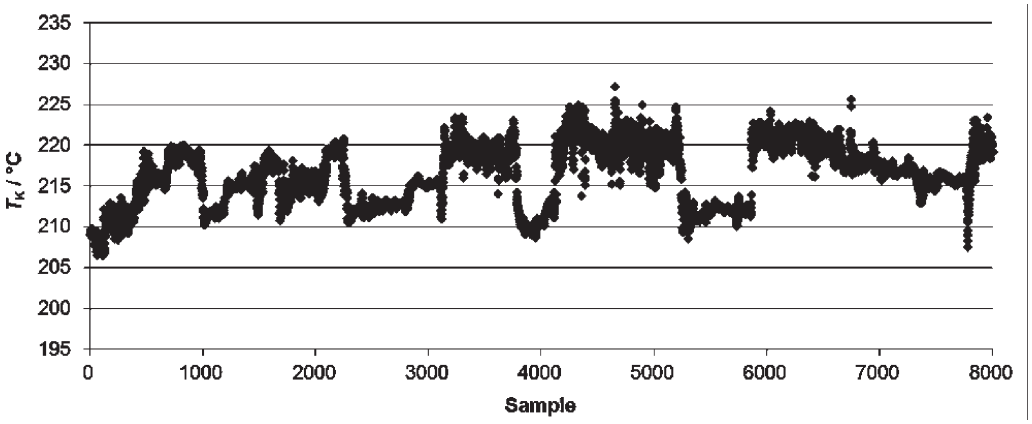
– pumparound flow rate – (F_{PA}), FI-6130.

Results and discussion

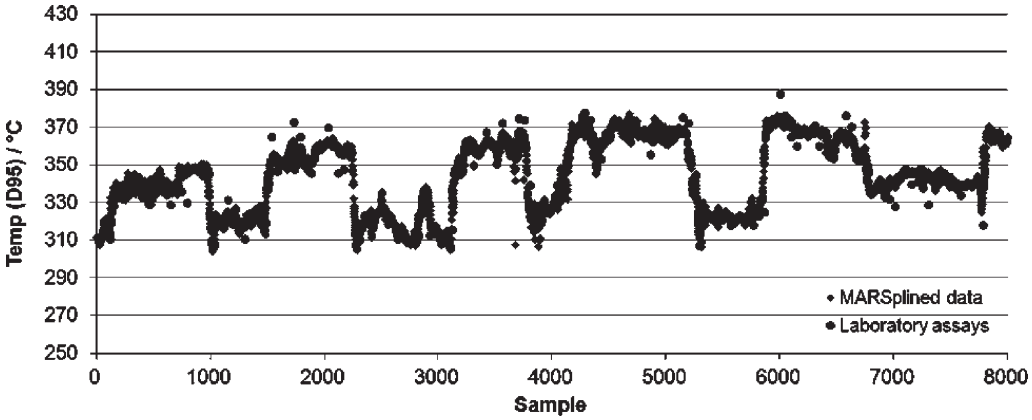
Data from the real plant were obtained in the period of ten months in a way to involve different process regimes, and therefore different quality requirements for diesel fuel. Diesel fuel samples were collected equidistantly four times a day from the plant and kerosene 95 % distillation point (D95) were determined.

The number of each input data (sampled every 5 minutes) must correspond to the number of output data, thus requiring additional output data. This was generated by the multivariate adaptive regression splines algorithm (MARSpline). Figs. 4 and 5 show trends of input kerosene temperature variable T_K, measured and splined output variables. It could be observed that changes in input obviously impacted the MARSpline output response, which approves using the MARSpline technique for generating the additional output data.

Data preprocessing was performed prior to model development. According to Shannon's sampling theorem, the chosen sampling time was 5 minutes. The extreme values were removed from the data using the "three sigma" rule. [17] Also, mean values and trends were removed from input data. Data filtering were also performed.[14] From the chosen model inputs (T_{TOP} , T_K , T_{LGO}, T_{HGO} , T_{PA} and



F i g . 4 – *The plot of kerosene temperature*



F i g . 5 – *Comparison of laboratory assays and splined output data*

F_{PA}), the variables with the greatest influence on D95 are the temperature of light gas oil, heavy gas oil, kerosene temperature, the pumparound temperature and flow.

Further, the important inputs were chosen based on PLS and Pearson correlation analysis as shown in Table 1. If the variables are mutually independent, then the correlation coefficient equals zero or close to zero, and if they are dependent, the correlation coefficient ranges from –1 to 1. A positive correlation between variables indicates that the variables are directly proportioned, and vice-versa. For the input selection, the threshold value for the correlation coefficients was set to 0.2.

The PLS analysis is shown in Table 2. The measure of importance of variables is given by its modeling power. A variable with a modeling power equal to one is completely relevant for building the

T a b l e 1 – *Correlation analysis*

	D95
F_{UK}	–0,1432
T_{VRH}	–0,4128
T_{TB}	–0,0279
T_K	0,6282
T_{LPU}	0,5115
T_{TPU}	0,2641
T_{GMR}	0,3409
F_{GMR}	–0,4979
D95	1,0000

T a b l e 2 – *PLS analysis*

Number of components is 2 in PLS model		
variables	power	importance
T_{TPU}	0,778448	1
T_K	0,590544	2
T_{LPU}	0,445702	3
T_{GMR}	0,400394	4
F_{GMR}	0,385990	5
T_{VRH}	0,334384	6
F_{UK}	0,113363	7
T_{TB}	0,052389	8

PLS model. Variables with modeling power close to the "number of components" divided by the "number of variables" are regarded to be less or non-significant, like F_{UK} and T_{TB}.

Real plant data were divided into two sets. The first 70 % of data were chosen for modeling (estimation), while the remaining 30 % of independent data were chosen for validation purposes. Initially, more than ten types of models have been developed. The linear models developed in the preliminary investigation included: Finite Impulse Response models (FIR), AutoRegressive model with eXogenous inputs (ARX), AutoRegressive Moving Average with eXogenous inputs (ARMAX), BJ, OE, state space models etc. Nonlinear models included: nonlinear FIR, nonlinear ARX and HW models with piecewise linear, sigmoid, wavenet, and other types of networks.

From a variety of developed models, the best results achieved are shown in Table 2. The optimal structure of the BJ model comprises 2 past samples of each 6 inputs, 2 past samples of model prediction errors, 2 past samples of simulated model prediction errors, 1 past sample of model prediction output and input time delay of 5 minutes, as shown in Table 1. The model shows very high FIT coefficient and very low RMS, FPE and AE from the experimental data. The average absolute deviation of the laboratory-determined distillation end point temperature is around 1.5°C, which is approximately in the range of measurement uncertainty. This kind of model can be used for advanced process control and fault detection when the measured output values are available. Fig. 4 shows good correspondence between measured and predicted outputs on the validation set.

The optimal structure of the OE model consists of 2 past samples for each of the 6 inputs with 5 minutes time delay each, 6 past samples of model prediction as shown in Table 1. The OE model shows high FIT coefficient and low RMS, FPE and absolute deviation from experimental data. Dynamic response, shown in Fig. 5, exhibits very good correspondence of measured and prediction data on the validation set.

The Hammerstein-Wiener model consists of linear dynamic block and two nonlinear static blocks, i.e. input and output static nonlinearities. Parameters of nonlinear HW model *nb*, *nf*, *nk* and the number of nonlinear units (*n*) are shown in Table 1.

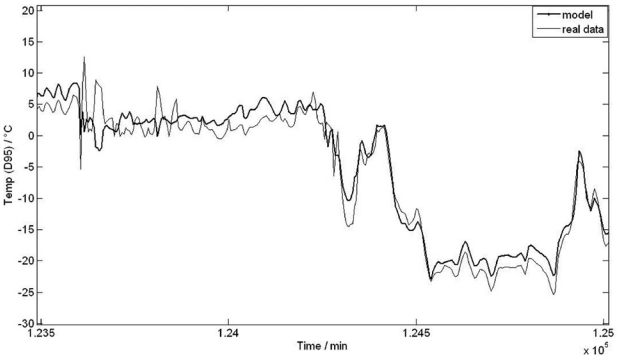
Linear block in the model is a matrix of the transfer functions containing 2 past samples for each of the 6 inputs with 5 minutes time delay each, and 3 past samples of model prediction. Static nonlinearities of all 6 inputs are presented with sigmoid network containing 10 units. The HW model

T a b l e 3 – *Model description*

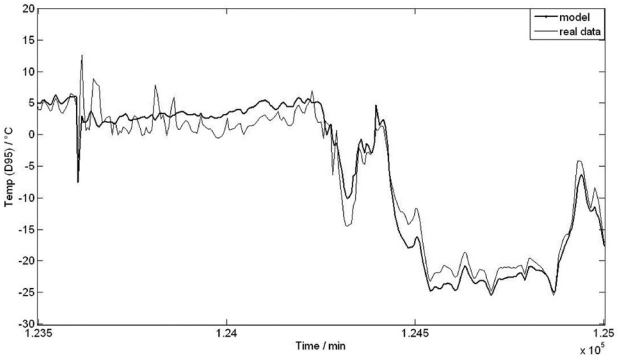
Parameters	BJ	OE	HW
na	–	–	–
nb	2	2	2
nk, min	5	5	5
nc	2	–	–
nd	2	–	–
nf	1	6	3
n	–	–	10

T a b l e 4 – *Model comparison*

Parameters	BJ	OE	HW
V	1,59	6,06	4,85
FPE	1,59	6,06	4,97
FIT, %	93,51	83,50	81,07
RMS, °C	1,48	2,33	2,74
AE, °C	0,89	1,81	2,15



F i g . 6 – *Comparison between measured and predicted outputs on validation set for BJ model*



F i g . 7 – *Comparison between measured and predicted outputs on validation set for OE model*

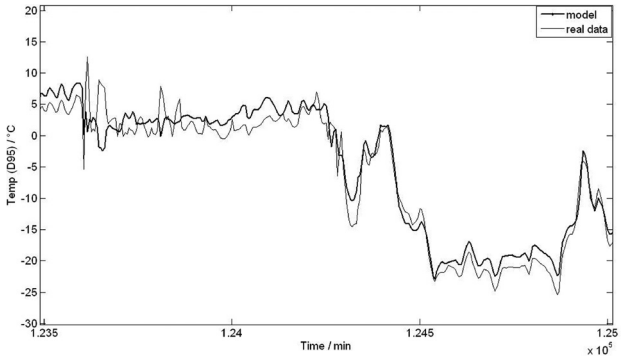


Fig. 8 – *Comparison between measured and predicted outputs on validation set for HW model*

shows satisfactory FIT coefficient and acceptable RMS, FPE and AE. The dynamic response in Fig. 6 shows a satisfactory match between measured and prediction data on the validation set. The OE and HW models do not use past outputs for prediction, so they can be used for D95 continuous estimation.

Conclusion

Based on continuous temperature and flow measurements of adequate process streams, the dynamic soft sensor models for the estimation of 95 % distillation end point were developed. Data was collected from DCS and laboratory assays. Because of the rare laboratory output data, the MARSpline generation method for additional output data was provided. Different dynamic linear and nonlinear models were developed using several identification methods.

The average absolute deviations of all model results lie within acceptable tolerance limits for on-line implementation. The BJ model shows the best performance and can be employed as the soft sensor within advance process control or for diagnostic purposes. The OE and HW models show somewhat inferior performances but still can be successfully used for the prediction of the kerosene 95 % distillation end point and process control purposes.

By real plant application of the developed soft sensors, considerable savings could be expected, as well as compliance with strict regulations for product quality specifications.

ACKNOWLEDGMENT

This paper is a result of the scientific project ZP-125-1963-1964 Soft Sensors and Analyzers for Process Monitoring and Control supported by the Croatian Ministry of Science, Education and Sports.

The authors wish to thank Mr. Damir Lončar, Mr. Ivan Radić and Mr. Ivica Jerbić with INA-Refinery Sisak, Croatia, who contributed to the work through advice and counsel.

Abbreviations

PLS – Partial Least Squares method
TOP – column top
K – kerosene
LGO – light gas oil
HGO – heavy gas oil
PA – pumparound
FIT – coefficient for the goodness of the fitted model, %

List of symbols:

$\mathbf{B}_i(q)$ – polynomial matrix
$\mathbf{F}_i(q)$ – polynomial matrix
$\mathbf{C}_i(q)$ – polynomial matrix
$\mathbf{D}_i(q)$ – polynomial matrix
$\kappa(s)$ – logsig function
$g(k)$ – sigmoid network
T – temperature point, °C
F – flow variable, kg m^{-3}
K – order of interactions
e_s – past simulated model errors

References

1. *Bolf, N., Ivandić, M. Galinec, G.* Soft sensors for Crude Distillation Unit Product Properties Estimation and Control, 16th Mediterranean Conference on Control and Automation, 1804–1809., Ajaccio, 2008
2. *Kadlec, B., Gabrys, B., Strandt, S.*, Comput. Chem. Eng. **33** (2009) 795–814.
3. *Buceti, G., Fortuna, L., Rizzo, A., Xibilia, M. G.*, Fusion Eng. Des. **60** (2002) 381.
4. *Fortuna, L., Graziani, S., Rizzo, A., Xibilia, M.G.*, Soft Sensors for Monitoring and Control of Industrial Processes (Advances in Industrial Control), Springer, London, 2007
5. *Chen, C., Mo, S., Chen, X.,* Dynamic Soft-sensor Based on Finite Impulse Response Model for Dual-rate system. Control and Decision Conference, 2173–2178., 2009
6. *Rogina, A., Šiško, I., Mohler, I., Ujević, Ž., Bolf, N.*, Chem. Eng. Res. Des. **89** (2011) 2070–2077.
7. *Napoli, G., Xibilia, M. G.*, Comput. Chem. Eng. **35** (2011) 2447–2456
8. *Yan Xuefeng*, Comput. Chem. Eng. **32** (2008) 608–621.

9. *Badhe Y., Lonari J., Sridevi U, Rao B. S., Tambe S. S., Kulkarni B. D.*, Chem. Eng. J. **97** (2004) 115–129.
10. *Schladt, M., Hu, B.*, Chem. Eng. Process. **46** (2007) 1107–1115.
11. *Dam, M., Saraf, D. N.*, Comput. Chem. Eng. 30 (2006) 722–729.
12. *Chatterjee, T., Saraf, D. N.*, J. Process Control **14** (2003) 61–77.
13. *Lin, B., Recke,B., Knudsen, J.K.H., Jrgensen, S.B.*, Comput. Chem. Eng. **31** (2003) 419–425.
14. *Hastie, T., Tibshirani, R., Friedman, J.H.*, The elements of statistical learning: Data mining, inference, and prediction, Springer, New York, 2001
15. Matlab The Language of Technical Computing, www.mathworks.com (2009)
16. *Cerić, E.*, Petroleum – Processes and products, INA and Kigen, Zagreb, 2006 (in Croatian).
17. *Ljung, L.*, System Identification: Theory for the User, 2nd ed., Prentice Hall, New Jersey, 1999