Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Evaluation of Nonlinear Inferential Models to Estimate the Products Quality of Industrial Distillation Process

G. Digo, N. Digo, I. Mozharovskii, A. Torgashov

*Institute of automation and control processes Far Eastern Branch of RAS,
Vladivostok, Russia, 690041 (e-mail: torgashov@iacp.dvo.ru)*

Abstract: The problem of improvement of existing approaches for evaluation of inferential models used in the advanced process control systems is considered. The method of finding the identifiability index limit, based on the physically-chemical essence of the industrial distillation process, for the inferential models containing nonlinear transformations of measured inputs is given. The problem of handling the narrow variability range of key inputs in the training sample is considered. The paper contains the example of evaluating the inferential models based on the proposed method applied to the nonlinear industrial distillation plant.

Keywords: quality control, process models, identifiability, distillation columns.

1. INTRODUCTION

Operational control of the continual technology process product quality is carried using the monitoring of the industrial situations. The monitoring itself is carried by collecting and processing the data, received from the sensors, and the laboratory analysis. Nowadays, to improve the quality of the output product of the oil-chemistry and refining technological process the three ways to control the quality are used: laboratory analysis, the stream analyzer data, and the inferential models (IM) data (or soft sensors, virtual analyzer etc.) (Fortuna *et. al.,* 2007; Petr Kadlec *et. al.,* 2009). But the results of the analysis, made in the industrial laboratory are not always full and operative enough, also quite seldom and cannot be used for the real-time quality control. The stream analyzers need calibration and are extremely expensive and though sometimes not affordable. In contrast, the IM, having only a little less accuracy, are much more cheaper and reliable. The principle of the IM work is based on the continued determining of the quality by the mathematical model, describing its link to the current measured process parameters as flowrate, temperatures, pressures etc.

Also, the technological objects such as distillation column (DC) of the oil-refinery industrial (malfunction of the DC can lead to ecological disaster) must be monitored permanently.

In reality the structures of most of the plants are weakly formalized due to the lack of data about them, so before building the IM, the rating of it's identifiability should be done (the opportunity to make the model based on the given input-output data under corresponding restrictions). That is the unusual task of identifiability of the IM of the low formalized structure in general. Due to lack of the articles, describing the problem of IM identifiability (Fisher 1966; Gorskii *et. al.,* 1987), the problem of creating the numerical algorithm of analyzing the IM identifiability for the industrial needs is still important.

This article is describing the problem of determining the factors that influence the accuracy of the IM identifiability of complicated nonlinear low-formalized plants. It gives the method of determining the limit value of identifiability of the IM, based on the calculation of the physically-chemical rigorous distillation model in order to form the nonlinear function type for each IM input.

The example of determining the IM and its identifiability index of the refinery industrial DC by the isopentane and benzene components in the top product is considered.

2. FORMULATION AND ANALYSIS OF THE PROBLEM

We consider the mass transfer process of rectification, flowing in the industrial apparatus such as DC. The plant model has *m* input of the measured variables $x_1,\dots,x_m$ and output *Y*, connected by a functional relationship

$$Y = F(\mathbf{X},\mathbf{B}) + \varepsilon\,, \tag{1}$$

where $\mathbf{X} = (x_1,\dots,x_m)$ is the vector of the measured input variables, $\mathbf{B} = (\beta_1,\dots,\beta_m)$ - vector of model coefficients, ε - output error.

Supposing the true relationship (structure) *F* type in (1) is not known, but from the available physical-chemical rigorous model of the DC follows, that there are nonlinear relationships between the output and input variables.

From the description of method proposed in (Digo *et. al.,* 2013) for determining the identifiability index of nonlinear objects with unknown model structure based on the algorithm of alternating conditional expectations (ACE) proposed by Breiman and Friedman (1985) follows that it is rated by a threshold value, depending on the specific characteristics and conditions of the plant. Therefore it is necessary to set its threshold value for each sufficiently narrow class of the examined objects. As shown in previous studies (Digo *et. al.,*

2013) for the object (1) the nonlinear IM is acceptable in form of

$$\theta(Y) = \alpha + \sum_{i=1}^{q} \Phi_i(x_i) + \varepsilon \,, \qquad (2)$$

where θ - is the output variable Y function; Φ_i – nonlinear functional transformation of input $x_i, i = 1,...,q$; ε-error.

However, due to the use of ACE to the industrial data containing the error, analytic type of function Φ_i is unknown, which makes finding the threshold identifiability parameter harder. Therefore, the task is determining the analytical form of the function Φ_i , taking into account the physico-chemical characteristics of each input $x_i, i = 1,...,q$. To solve it, it is necessary to:

1. Make a physical-chemical rigorous model (system of nonlinear differential-algebraic equations of high dimension) and execute its calibration;

2. Form a sample of data to construct IM, making experiment based on a calibrated physical-chemical model;

3. Apply the ACE algorithm to a sample to find Φ_i from (2);

4. Make the approximation of nonlinear functions Φ_i to receive their analytic form.

As a result of steps 1-4, a class of nonlinear input functions, included in the IM, will be reasonably formed.

3. CALIBRATION OF NONLINEAR PHYSICOCHEMICAL RIGOROUS MODEL OF THE DISTILLATION PROCESS ON INDUSTRIAL DATA

The mass-transfer processes in industrial fractionating columns operating in constantly acting perturbation such as fluctuations of feed temperature and composition, vapor pressure and loads, are poorly formalized. Since fractionating columns are objects with time-varying statistical characteristics, it is necessary to use the data samples of the industrial DC and data sample generated based on the rigorous DC model in order to find a structure of IM. Nonlinear dependence between the output and each input may be revealed using the ACE algorithm.

A calibration should be completed with the help of industrial data to approximate the steady-state regime of a DC with an actual system, i.e. to find appropriate values of characteristics that can provide the closest to the real operating conditions result.

A calibration of the DC steady-state regime in view of physicochemical nature of the process involves the following steps:

- to use values of chemical substances concentrations (that can be known from the laboratory studies, such as mean values of several months) for the feed stream;

- to set the modes of the DC operation according to key process parameters;

- to select the tray efficiency minimizing the mismatch between industrial data and the rigorous model.

To approximate the DC steady-state regime model to industrial working conditions the following steps were done:

- the average concentration of 28 individual hydrocarbon components in the feed stream were used, as well as the values of process parameters needed to calculate the DC steady-state operating point (the first four rows of Table 1);

- Murphree tray efficiency values (Murphree, 1925) that provide the smallest deviation of industry data from design data were sorted out (last four rows of Table 1).

When selecting Murphree efficiency on industrial data for all DC trays and the average steady-state regime by isopentane fraction and concentration of benzene components in the distillate the following criteria were used:

$f_{isopen\tan e}(E) = (x^{m}_{isopen\tan e}(E) - x_{isopen\tan e})^2$,

$f_{benzene}(E) = (x^{m}_{benzene}(E) - x_{benzene})^2$, $f_{isopen\tan e}(E)$ - the residual function of isopentane fraction in the distillate, $x^{m}_{isopen\tan e}(E)$ - isopentane fraction calculated using model; $x_{isopen\tan e}$ - isopentane industrial data, $f_{benzene}(E)$ - the residual function of concentration of benzene components in the distillate; $x^{m}_{benzene}(E)$ - concentration of benzene generating components calculated using model; $x_{benzene}$ - industrial data of the concentration of benzene components.

Table 1. The results of calibration of rigorous DC model

Process parameter	Industrial data	Model data
Feed, kg/hr	51730	51730
Pressure at the top of the column, kgf/cm^2	1.93	1.93
Distillate flowrate (*D*), kg/hr	23089	23089
Reflux ratio	1.6	1.6
Top temperature, °C	85.5	80.3
Bottom temperature, °C	119.5	121.2
Isopentane in *D*, mass. frac.	0.1795	0.1790
Benzene components in *D*, mass. frac.	0.1877	0.1855

The values of the residual (error) functions and Murphree efficiency are shown at figure 1. The calibration results show that Murphree efficiency coefficient equal to $E = 0.3$ gives the smallest deviation of industrial data from the rigorous model data. The fitting of efficiency values for each individual separation stages did not provide significant improvements of calibration results.

4. OBTAINING OF NONLINEAR FUNCTIONS OF IM INPUT VARIABLES

To define a class of nonlinear functions of the inputs $\mathit{\Phi}_i$ that describes the dependence of the output function with the expression (2), a data sample based on rigorous process model should be made with the help of the calibrated model above. Basing on the model type of functions $\theta(Y)$ and $\Phi_i(x_i)$, $i=1,...,q$, should be estimated according to the algorithm ACE. As the result, we can make recognition among linear and nonlinear input functions. The nonlinear functions Φ_i should be approximated for getting their analytic form. Coefficients in all analytical expressions, including linear coefficients, should be corrected for approximating the results to the operating points of the industrial DC.

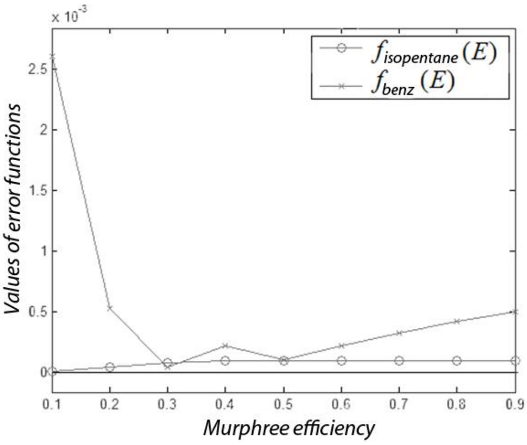


Fig. 1. The values of Murphree trays efficiency during rigorous DC model calibration with industrial data.

As it was mentioned above the IM structure was defined on the previously generated data set that contains 5 input variables in order to determine the quality of the industrial DC final product over the isopentane fraction in distillate (Fig. 2). The model includes the specific nonlinear functions of saturation on inputs, associated with pressure, reflux and bottom temperature of the column. Thus, the following variables are inputs of IM: x_1 is the flow of distillate (top product) kg/hr; x_2 is the flowrate of reflux, kg/hr; x_3 is the top pressure of the column, kgf/cm^2; x_4 is the top temperature of the column, °C; x_5 is the bottom temperature of the column, °C.

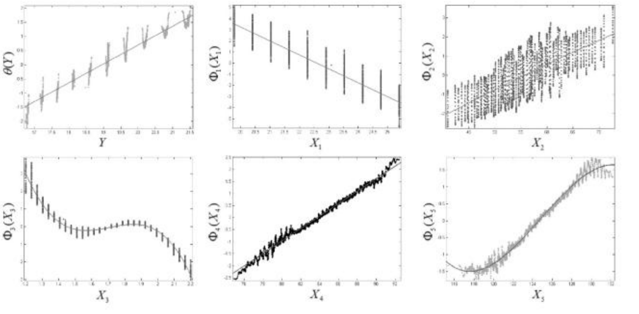


Fig. 2. The results of applying the ACE algorithm to determine the IM structure over the isopentane fraction in the DC's distillate.

It is obvious that variables x_2, x_3, x_5 effect on the output nonlinearly, and the effect of variables x_1, x_4 on the output it is close to a linear one. Approximated reconstructed variables for output $Y_{isopen\tan e}$ and for inputs x_1, x_4, x_2, x_3, x_5 are of the following form:

$$\begin{aligned}
&\theta(Y_{isopen\tan e}) = 0.665 \cdot Y_{isopen\tan e} - 12.61,\\
&\Phi_1(x_1) = -1.251 \cdot x_1 + 28.26,\\
&\Phi_4(x_4) = 0.2593 \cdot x_4 - 21.73,\\
&\Phi_2(x_2) = -0.3437 \cdot 10^{-3} \cdot x_2{}^2 + 0.1703 \cdot x_2 - 8.412,\\
&\Phi_3(x_3) = -30.55 \cdot x_3^3 + 156.6 \cdot x_3^2 - 265.7 \cdot x_3 + 149.2,\\
&\Phi_5(x_5) = -2.023 \cdot 10^{-3} \cdot x_5^3 + 0.7578 \cdot x_5^2 - 94.31 \cdot x_5 + 3899.
\end{aligned} \tag{3}$$

Likewise, the IM structure for concentration of the sum of benzene components in the distillate was obtained and also contains 5 input variables (Figure 3).

It is obvious from fig.3 that inputs x_3, x_4 effect nonlinearly on the $Y_{benzene}$, and the effect of variables x_1, x_2, x_5 on the output it is close to a linear one. Approximated reconstructed variables for output $Y_{benzene}$ and for inputs x_1, x_2, x_5, x_3, x_4 are of the following form:

$$\begin{aligned}
&\theta(Y_{benzene}) = 34.33 \cdot Y_{benzene} - 6.211,\\
&\Phi_1(x_1) = -1.531 \cdot 10^{-11} \cdot x_1^3 + 1.05 \cdot 10^{-6} \cdot x_1^2 - 0.02396 \cdot x_1 + 181.7,\\
&\Phi_2(x_2) = 8.863 \cdot 10^{-5} \cdot x_2 - 2.957 \cdot 10^{-3},\\
&\Phi_5(x_5) = 0.2574 \cdot x_5 - 32.05,\\
&\Phi_3(x_3) = -12.37 \cdot x_3^3 + 63.9 \cdot x_3^2 - 113.0 \cdot x_3 + 68.2,\\
&\Phi_4(x_4) = 3.283 \cdot 10^{-3} \cdot x_4^3 - 0.8303 \cdot x_4^2 + 70.17 \cdot x_4 - 1982.0.
\end{aligned} \tag{4}$$

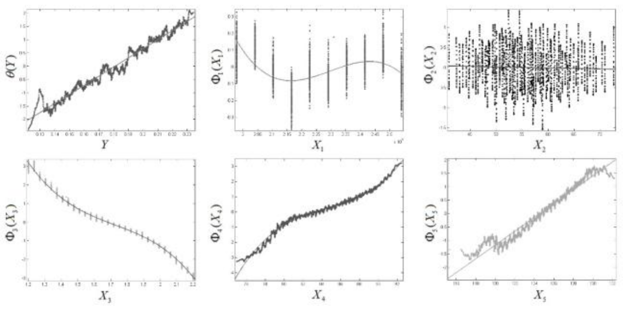


Fig. 3. The results of applying the ACE algorithm to determine the IM structure over the concentration of sum of benzene components in the DC distillate.

5. CALCULATING THE LIMIT VALUE OF THE IDENTIFIABILITY INDEX OF IM

When the model structure is unknown we proposed to use the identifiability index H_p during IM evaluation described by Digo *et. al.* (2013).

Based on the reference data sample from rigorous model is formed, the type of $\theta(Y)$ and $\Phi_i(x_i)$ functions is determined with the ACE algorithm, and the identifiability index H_Y is calculated for the input variable. It is obvious, that for the so formed data sample the value of the identifiability index is high, because there are no noise and almost no error. Let`s take into account some error in the data sets (e.g. 10% of the output data range), considering it as the unmeasured disturbance in the model. We need to mention, that the $H_Y^{10\%}$ will go lower if the impact of unmeasured factors increased. Consequently, when calculating the limit value of the model identifiability index with the reference data sample the reasonable error of the plant measurements should be always considered.

The numerical algorithm of determining the limit value of the identifiability index within the evaluation of IM for the industrial DC includes the following:

- forming the reference data sample of the given size, considering the selecting model structure (3);

- calculating the identifiability value for the input variable H_Y for the reference data sample;

- determining the parameters variation admissible range, which are used in the model (3);

- determining the identifiability value H_Y with considering different sources of uncertainties and errors in the industrial data set;

- choosing the limit value H_p of the model identifiability value.

Examples of choosing the limit value H_p. To find the threshold value $H_{p\ isopentane}$ of identifiability by approximated function (3) we should construct a reference data sample (K×5), which may be different to the size of the available industry data set. The values of the input variables $x_i, i = 1,2,3,4,5$ are generated by normally distributed random numbers in the specified ranges: $20.0 \le x_1 \le 25.0$, $37.0 \le x_2 \le 71.0$, $1.3 \le x_3 \le 2.1$, $76.0 < x_4 \le 92.0$, $116.0 < x_4 \le 132.0$, K=60.

Without taking into account the output measurement error, the parameter of identifiability for the reference data sample is $H_{Y\,isopentane} = 60.53$.

In the presence of unmeasured factors in the form of some errors and uncertainties (5% - 35% of the variation range of output) calculated identifiability values confirm that with the growing influence of unaccounted factors the $H_{Y\ isopentane}$ decreases.

For example, with a 5% uncertainty of the output $H_{Y\ isopent}^{5\%} = 27.80$, and while 30% of the change $H_{Y\ isopent}^{30\%} \approx 5.0$. Fig. 4 shows the change of $H_{Y\,isopentane}$ value depending on the unaccounted factors and uncertainty level in the industrial data set.

In reality, the error value of the output by the share of isopentane is 15%, which corresponds to the threshold parameter of identifiability $H_{p\ isopent} = H_{Y\ isopent}^{15\%} = 10.0$.

Similarly, the threshold value $H_{p\ benz}$ of the identifiability index is obtained for the sum of benzene components measurement error, which is also 15%, using a reference sample of the same size, the same input data and the approximated functions (4).

Without taking into account the output measurement error, the identifiability index for the reference sample is $H_{Y\ benz} = 35.2526$. Fig. 4 shows the change $H_{Y\ benz}$ depending on the output uncertainty range in 5% - 35%. The threshold identifiability index of the IM for the output of the sum of benzene components in the case of a 15% error of the output measurement is $H_{p\ benz} = H_{Y\ benz}^{15\%} = 17.0$.

As the result of application of the algorithm of determining the identifiability index of IM to the DC industrial data, we received $H_{Y\ isopent}^{indust} = 13.09$, $H_{Y\ benz}^{indust} = 18.0216$ by the isopentane and concentration of benzene components, respectively. It is obvious that the following inequalities $H_{Y\ isopent}^{indust} > H_{p\ isopent}$ and $H_{Y\ benz}^{indust} > H_{p\ benz}$ are holding.

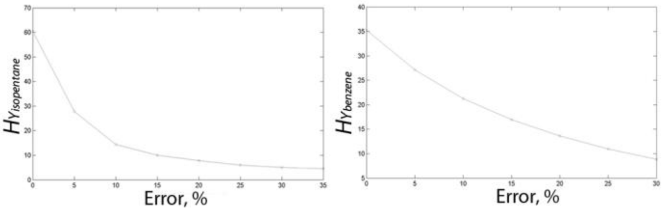


Fig. 4. The influence of measurement errors on output identifiability indexes $H_{Y\ isopent}$ and $H_{Y\ benz}$, respectively.

Therefore, IM can be identified, based on the given industrial data taking into account the 15% output uncertainty.

6. DEVELOPMENT OF IM WITH INPUTS HAVING A SMALL RANGE OF VARIATION

In practice we have to deal with a situation where a priori information is showing that the output of the plant is significantly affected by one or more inputs, but such inputs in the model are not statistically significant. This is due to the fact that during the plant operation, their variation ranges are limited by the technological reasons, for example, due to safety constraint etc. The below is an algorithm that ensures the handling of a priori data about specific input variables from the available data obtained from nonlinear plant with small range of variation of key (from physical-chemical sense) inputs.

Let us examine the plant (1), the structure F of which is unknown, but the available rigorous physical-chemical model of the DC means that there are nonlinear relation between the output and specific input variables. In order to find them, we have to apply, for example, to regression models, usually multi-dimensional, nonlinear by the input variables

$$\hat{Y} = b_0 + \sum_{i=1}^{p} b_i f_i \qquad (5)$$

and to select the structure, based on the available a priori information. In (5) $\hat{Y}$ - is the predicted value of the output variable Y; $f_i = f_i(x_1,\ldots,x_m)$, $i = 1,\ldots,p$, - some of the functions of measured input variables $x_1,\ldots,x_m$; $b_0, b_1,\ldots,b_p$ - coefficient estimates from (1).

Let us assume, firstly for simplicity, that a linear model is made, based of the available data, and equation (5) takes the form

$$\hat{Y} = b_0 + \sum_{i=1}^{m} b_i x_i \qquad (6)$$

in which the coefficient b_k is insignificant in statistical sense for input variable x_k having small range of variation (or small standard deviation). But the physical-chemical properties mean that this variable should have a significant impact on output. When making an adequate model in such circumstances, a correcting factor should be derived, that takes into account the influence of such input, based on the calibrated rigorous DC model. For input x_k an extended sample of data is formed by increasing the range of its variability as long as the adequate description is obtained in the form of polynomial function. The derived polynomial is used in the correction term d_k in the model for industrial data.

To simplify the description of the proposed algorithm we redefine input variables in (6) so that the variable x_k with insignificant coefficient b_k becomes the last, i.e.

$$\hat{Y} = a_0 + \sum_{i=1}^{m-1} a_i x_i + a_m x_m \qquad (7)$$

$a_0 = b_0$, $a_i = b_j$, $j = 1,2,\ldots,k-1,k+1,\ldots m$, $a_m = b_k$. In this notation, the algorithm reduces to the following sequence of steps.

1. An extended data set (sample) is formed by increasing the range of variability of input x_m using the rigorous calibrated DC model.

2. According to the extended sample the nonlinear model $Y_m = a_m x_m = a_m^0 + \sum_{j=1}^{q} a_m^{(j)} \cdot x_m^j$ is derived for input x_m, where q - the order of the polynomial of x_m variable.

3. The correction term is calculated, and in the original sample the output variable Y is transformed to $Y_d = Y - Y_m$.

4. According to the available sample, the IM is constructed from (m-1) variables $\hat{Y}_d = a_0 + \sum_{i=1}^{m-1}\sum_{j=1}^{p} a_i^{(j)} \cdot x_i^j$, where q – order of input polynomials (for example, using OLS method).

5. The value of the output variable $\hat{Y} = \hat{Y}_d + Y_m$ is determined.

The proposed algorithm is tested on generated samples and data obtained from an industrial distillation column.

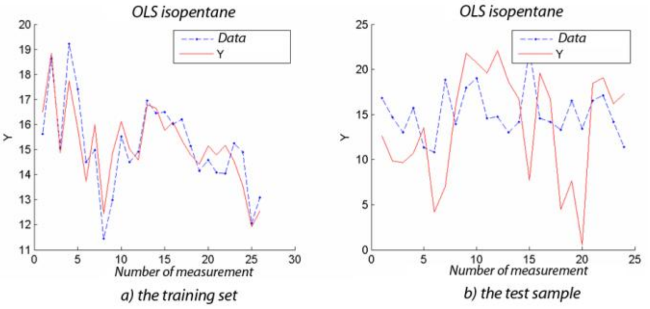
Example. Consider the problem to make the IM of the DC end-product quality by isopentane mass fraction in the distillate with the a priori information about the effect of pressure (P) at the top of the column. The pressure as input to IM was found to be insignificant because of the small range of variability in the available industrial data set.

The inputs for IM are the following process variables: x_1 - flowrate of the distillate (top product) kg/hr; x_2 - reflux flowrate, kg/hr; x_3 - pressure at the top of the column, kgf/cm^2; x_4 - temperature at the top of the column, °C; x_5 - the column bottom temperature, °C.

When building the IM for determining the quality of the output product of the concentration of isopentane without a priori information about the effect of pressure at the top of the column we get:

$$Y_{isopent} = 181.2661 + 0.3864{*}x_1 + 0.0490\ {*}x_2$$
$$-1.2504{*}x_4 + 0.1965{*}x_5 - 45.4786{*}x_3 .$$

The coefficient of determination $R^2_{isopent}$ is equal to for testing data set 0.1402 and for training sample 0.8113 (Fig. 5), respectively.



a) the training set

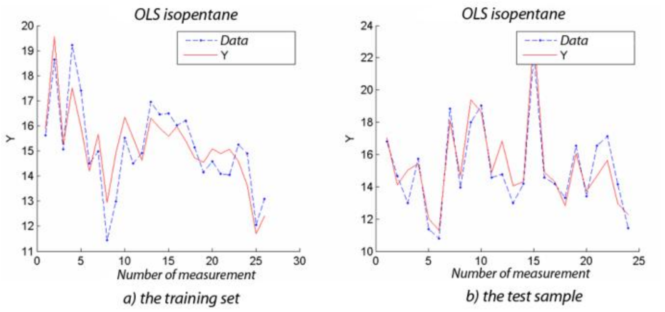
b) the test sample

Fig. 5. IM performance without taking into account the pressure correction term.

Using a priori information about the effect of pressure on the concentration of isopentane in the DC distillate and rigorous static model, an IM is made as close to the modes of operation of industrial DC as possible. The dependence of concentration of isopentane in the distillate with the pressure change from 1 kg/cm^2 to 2 kg/cm^2 on a rigorous DC model in form $Y_m = 2.863 \cdot x_m^2 + 4.432 \cdot x_m + 7.708$ is found. Taking into account the influence of pressure, the model of the output product quality by the mass fraction of isopentane was found:

$$\hat{Y}_{isopent} = 66.458 + 0.4817x_1 + 0.117x_2 - 1.035x_4 +$$
$$+ 0.0802x_5 + 2.863x_3^2 - 4.432x_3 .$$

The coefficient of determination $R^2_{isopent}$ for test sample is 0.8727, for training data set 0.7878, respectively (Fig. 6).



a) the training set

b) the test sample

Fig. 6. The performance of IM taking into account the pressure correction term.

The practical result of the usage of proposed algorithm (taking into account a priori information of statistically insignificant inputs) consists in the improvement of quality of the IM by the 63.25% of R^2 increase for the test sample.

7. CONCLUSIONS

This article reviews the improved method of making the IM for the nonlinear industrial plants. The analysis of identifiability of IM based on the industrial data was fulfilled. The method allows to choose the limit value of identifiability index with consideration of output uncertainty and to avoid the lengthy process of search and substantiation of structure of IM using conventional regression analysis.

The way to take into account important (with the physics-chemical meaning) inputs in the IM in case of their low variability range has been proposed. As the result, the R^2 is increased on the test sample, if it contains new data, not included in the training data set. The improvement of IM quality leads to successful industrial application of advanced control systems.

The results of applying of proposed method are demonstrated by the example of IM evaluation of the isopentane mass fraction in the overhead product of industrial DC.

8. ACKNOWLEDGEMENTS

This work is partially supported by grants of the Far Eastern Branch of the Russian Academy of Sciences No. 12-I-P-17-02.

REFERENCES

Breiman L., Friedman J. (1985) Estimating optional transformations for multiple regression and correlation // *Journal of the American Statistical Association*. Vol. 80. pp. 580-598.

Digo G., N. Digo, I. Mozharovskii and A. Torgashov (2013). Analysis of Identifiability of Nonlinear Plants with Weakly Formalized Structure //IFAC Proceedings Voumes (IFAC-PapersOnline): 7th IFAC Conference on Manufacturing Modeling, Management and Control. IFAC MIM'2013. June 19-21: Russian Federation, Saint Petersburg. pp. 1224-1229.

Fisher F. (1966) The identification problem in econometrics. New York: McGraw-Hill.

Fortuna, L., Graziani, S., Rizzo and A., Xibilia, M.G. (2007). *Soft sensors for monitoring and control of industrial processes*. London: Springer-Verlag.

Gorskii V. G., E. A. Katsman, F. D. Klebanova and A. A. Grigoryev (1987), Numerical study of parameter identifiability for nonlinear models // *Theoretical and Experimental Chemistry*, Volume 23, Issue 2, pp 181-186.

Murphree E. V. (1925) Rectifying Column Calculations with Particular Reference to n-component Mixtures // *Industrial & Engineering Chemistry*. 17 (7). pp. 747-750.

Petr Kadlec, Bogdan Gabrys and Sibylle Strandt (2009). Data-driven soft sensors in the process industry // *Computers and Chemical Engineering*. Vol. 33. pp. 795–814.