Development of a Surrogate-model Based Energy Efficiency Estimator for a Multi-step Chemical Process

Markku Ohenoja Tero Vuolio Teemu Pätsi Petri Österberg Mika Ruusunen

Environmental and Chemical Engineering Research Unit, Control Engineering, University of Oulu, Finland,
{forename.surname}@oulu.fi

Abstract

Energy efficiency is increasingly being considered as a critical measure of process performance due to its importance both in production costs and in environmental footprint. In this work, an indirect energy efficiency estimator was developed for the Tennessee Eastman (TE) benchmark process for the first time. The TE model was first modified to provide the reference values of energy efficiency. A sophisticated model selection scheme was then applied to build the surrogate-model. The results indicate reasonable model performance with mean absolute prediction error around 1.7%. The results also highlight the limitations present in the training set, which are, together with other practical implementation issues, discussed in this work.

Keywords: Chemical Process Engineering, Tennessee Eastman, Energy Efficiency, PLSR models, Model adaptation

1 Introduction

Energy efficiency is an important factor to consider in modern chemical process engineering; the efficiencies often are concentrated to reducing energy costs per product. Higher energy prices strongly contribute to increasing operating and manufacturing costs. Additionally, inefficiencies in energy usage also contribute to higher greenhouse gas emissions and environmental footprint. It has been concluded that the improvements in energy efficiency require pragmatic and holistic approaches (Drumm *et al.*, 2012).

The increased computational resources have enabled energy efficiency estimation and monitoring using large data sets collected from process plants. The dynamic losses (difference between the current energy consumption and the historical or theoretical energy consumption) can be estimated from the process data and visualized to the plant operators (Drumm *et al.*, 2012). The predictive soft sensors could also assist in the selection of process paths at least with a suboptimal energy efficiency (Nikula *et al.*, 2016). However, with large and complex data sets, the development of soft sensors is not straightforward. It should be also mentioned that regardless of the suboptimality, typically large energy savings can be realized in the chemical industry because of the high production volumes (Saygin *et al.*, 2011).

This study demonstrates the development of a real-time data-driven energy efficiency estimator using an artificial data set. For this aim, a multivariate simulation study with the Tennessee Eastman (TE) process benchmark is carried out. The TE process is a multi-step chemical process with relatively slow dynamics and consequently large delays. After introducing a step change, the settling time is approximately 24–48 hours, severely complicating the analysis (Downs and Fogel, 1993). The TE process has five main unit comprising an exothermic reactor, a condenser, a compressor, a separator, and a stripper. The operating cost of the TE process is related to the loss of product and reactants (in purge and product streams), steam utilization and the compressor work (Konge *et al.*, 2020).

Being an open-loop unstable process, the TE process has been extensively used to develop and test plant-wide control strategies (e.g. Larsson *et al.*, 2001; Jämsä, 2018). The scenarios embedded to the benchmark model have also resulted as numerous studied aimed for fault detection and diagnosis (e.g. Kulkarni *et al.*, 2005; Xie and Bai, 2015; Zou *et al.*, 2018). In addition, the plant-wide, nonlinear nature of the TE process has gained attention for developing surrogate models; For example, Tran and Georgakis (2018) used Net-elastic regularization and D-optimal designs to reach steady-state surrogate models with reduced complexity. Sheta *et al.* (2019) developed dynamic NNARX models with interpretable structures for four TE outputs. Recently, Konge *et al.* (2020) proposed several machine learning based regression modeling techniques for building lower dimensional subsystems and performing process operability analysis to the TE process.

However, the energy efficiency estimation of the TE process is still an unexplored topic. The energy balances for the reactor, the product separator, the stripper and the mixing zone were introduced by Jockenlhövel *et al.* (2003).

2 Material and methods

2.1 Energy efficiency

Energy efficiency is here defined as the energy consumed by the process divided by the amount of

DOI: 10.3384/ecp211851

product produced. Hence, the value should be minimized in order to minimize the energy usage per produced tons. Both terms should also involve possible losses related to production and energy utilization, having negative effect to the energy efficiency, namely increasing the value. In TE process model, the product losses are negligible (less than 0.7%) and the model does not account for the energy losses. Therefore, the simplified definition of the instantaneous energy efficiency for the product component *n* at time instant *k* is calculated as in Eq. (1):

$$\eta_n(k) = \frac{P(k)}{m_n(k)} \tag{1}$$

Where η_n is the energy efficiency with respect to component *n* at time instance *k*, *P* [MJ/h] is the energy consumed per hour by the compressor and reboiler, and m_n [ton/h] is the amount of produced component *n* per hour. In TE process, the components of interest are the liquid products *G* and *H*.

In order to extract the instantaneous, real value of the energy efficiency from the TE model, a set of modifications to the simulation were required:

1. The average liquid densities of the product streams were calculated based on the measured molar fractions, and component liquid densities given in Downs and Vogel (1993),
2. The product mass flows were calculated from the average liquid densities and the measured product volumetric flows,
3. The reboiler energy was calculated from the measured steam mass flow according to Jockenhövel *et al.* (2003).

The product stream's molar fractions were the delay and disturbance free model outputs, while the other measurements consisted of the default delays and noise levels of TE benchmark. The energy efficiency described above represents the reference (target) signal for the surrogate model.

2.2 Simulation scenario

In TE process, gaseous reactants *A, C, D* and *E* are converted into liquid products *G* and *H*, and byproduct *F* (Downs and Vogel, 1993). TE model by Balthelt *et al.* (2015) is used in this study to generate the simulated process data. In the simulation, the base case operational mode of the TE process is considered, where the target product mass ratio of *G* and *H* is set to 50/50. The simulation was run with disturbance flags disabled and using the decentralized control strategy included in the TE simulator.

First, a subset of manipulated variables was selected using first order finite difference-based sensitivity analysis of the inputs with respect to η_G and η_H (energy efficiency of components *G* and *H*). The ranges for the selected variables were determined based on simulations and earlier findings from the literature. It is well known that the ranges of inputs need to be reduced as the number of inputs is increased (Konge *et al.,* 2020; Tran and Georgakis, 2018). Table 1 lists the selected variables and their feasible ranges applied to this study.

Next, a Monte Carlo type simulation scenario is formulated. There, the TE process is simulated for two months (60 days, 1440 h) to mimic a typical set of routine process data. The set points of the manipulated and operational variables are changed pseudo randomly to illustrate the effect of sudden changes in the production and on the energy consumption.

The simulation was performed in a following way; Firstly, a random number generator was initialized. Secondly, a random time instant between 24 and 48 hours was selected from an even distribution. Then, one to four manipulated variables are randomly selected to the adjusted time step. Finally, their values are randomly chosen from an even distribution and previously adjusted variables are changed back to nominal values to keep the process within control range.

The time spans for the set points changes were chosen to occur between 24 and 48 hours after the previous change in order to ensure robust process behavior and following the recommendations in original TE model (Downs and Vogel, 1993). Using a step size of five seconds, the resulting data matrix consist of 1,036,801 rows (time instants) and 43 columns (simulated process variables).

Table 1. Setpoints of the manipulated variables and their range in TE simulation.

Manipulated variable	*Nominal value*	*Lower bound*	*Upper bound*
Production rate [m^3/h]	22.9	20.5	24.0
Stripper level (%)	50.0	40.0	60.0
Component G in product (mole-%)	53.7	51.0	57.0
Component A in reactor feed (mole-%)	55.0	49.5	65.9
Components A&C in reactor feed (mole-%)	58.6	52.7	64.4
Reactor temperature [°C]	120.0	118.0	125

2.3 Data preprocessing

The simulated data had a substantial start-up transient. Hence, the first 1000 data points (1.39 h) were excluded from the training set prior to modeling.

The data matrix was then down-sampled to reduce the effect of delays and measurement noise present in the simulated process measurements. The down-sampling was performed with 6-minute averaging, resulting as a data matrix with 14,400 x 43 (41 inputs, 2 reference outputs). Then, each of the input variables were delay-compensated using a discrete time shift with a maximum lag of 30 minutes (*i.e.*, 5 different time shifts with a 6-minute sampling time). Consequently, before filtering there were 41 x 5 = 205 input variable candidates.

Prior to model selection the down-sampled data set was divided into training and testing sets. The division was simply made with respect to simulation time, reserving the latter 30% for testing.

2.4 Model selection and validation

In the second phase, a dynamic surrogate model for the operational mode one is constructed based on the training data (70% of the whole set) to estimate the energy efficiency. Prior to model selection, the input data space was normalized such that $X = \{0,1\}$ using the min-max-scaling.

The model structure here is based on the partial least-squares (PLS) regression. The estimator is selected based on a sufficiently representative training set of 40 days and tested with an independent time series of 20 days. The delay analysis and input variable selection are carried out using out signal correlation-based filtering, using linear correlation with the desired output and the time compensated signals as the filtering metric. In filter-based variable selection, the rule for variable inclusion or exclusion is given as

$$i = \begin{cases} 1, & R \geq T \\ 0, & R < T \end{cases}, \tag{2}$$

where *i* is the logical iterator, *R* is the linear Pearson product-moment correlation coefficient and *T* is the manually selected threshold. In this study, the threshold was set heuristically to T = 0.25. Consequently, the filtered estimator is

$$\hat{y} = \mathrm{X} b_{\mathrm{PLS}} + \varepsilon, \tag{3}$$

where b_{PLS} is the estimated parameter vector with the PLS algorithm, X is the input data matrix, $\hat{y}$ is the estimated output and the ε is the residual term with $\mathrm{N}(0, \sigma^2)$. The PLS parameter estimation is performed for the filtered matrix, i.e. $X[i=1]$ with the algorithm presented in de Jong (1993). The number of PLS components was selected using a grid search with cross-validation, consequently resulting as 4 and 3 selected components for the models of η_G and η_H, respectively. The objective function in selection was based on k-fold sequential cross-validation. After testing different values of the k-fold, a 3-fold cross-validation was selected.

The model performance was evaluated with the following figures of merit including *R*, RMSE (root mean squared error) and MAPE (mean absolute prediction error).

3 Results and discussion

3.1 Model Selection

Using the presented model selection procedure, a feasible model was identified. The figures of merit for the model training and testing results for the two energy efficiencies are shown in Table 2. For the η_G, the figures of merit for the out of sample data set (test set) can be considered sufficient for process control purposes. The predictions can be considered to be within ±0.0028 MJ/ton (2.8 kJ/ton) with 95.4% confidence. Similarly, for the final product component *H*, the model performance is comparable to the previous model with slightly higher correlation coefficient. The 95.4% confidence interval for energy efficiency model for component *H* was ±0.0032 MJ/ton (3.2 kJ/ton). In addition, it can be seen that the model's testing set performance metrics are quite optimistic for the component *H*, which can be seen as a higher correlation coefficient and lower error values compared to the training set.

Table 2. Figures of merit for the identified PLSR models.

Criteria	*Training*		*Testing*	
Product	*G*	*H*	*G*	*H*
R	0.86	0.88	0.85	0.89
RMSE, kJ/ton	1.6	1.6	1.6	1.6
MAPE, %	1.7	1.7	1.7	1.6

3.2 Model applicability

The test set estimations using the selected models for η_G and η_H are presented in Figure 1 and Figure 2, respectively, with corresponding confidence intervals of the selected estimators.

According to Figures 1 and 2, the testing set shows decreased performance, and for some regions the output value seems to interpolate poorly. In data-driven modeling, this often could indicate overfitting the model during the training phase, which means that the model parameters are biased because of estimating the noise in the system rather than the true dependencies. Utilization of an overly complex model as the estimator is a common cause of this behavior. (Hastie *et al.*, 2009)

In the modeling case of this study, the lack of fit in the test set seems to be at least partially explained with the non-similar distributions of the training and testing input data sets, often referred as the covariate shift (Moreno-Torres *et al.*, 2012). This issue is discussed in the following.

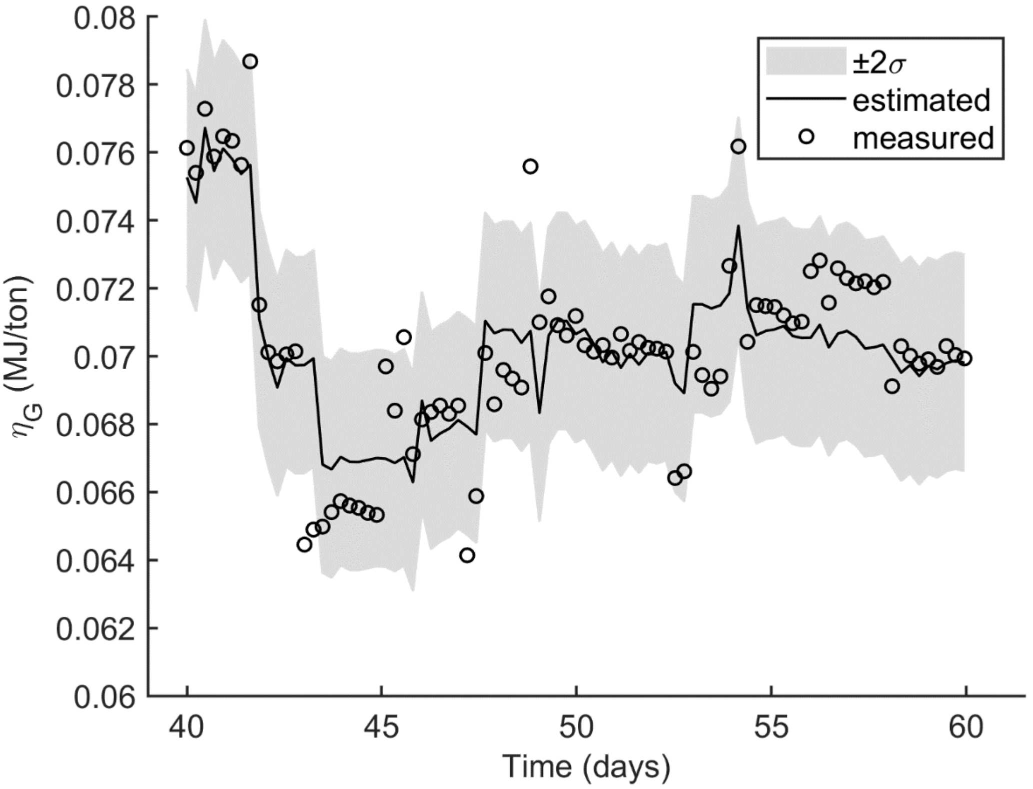

Figure 1. Measured and estimated energy efficiency for component *G* with corresponding confidence intervals. Only every 50 sample is plotted for the sake of clarity.

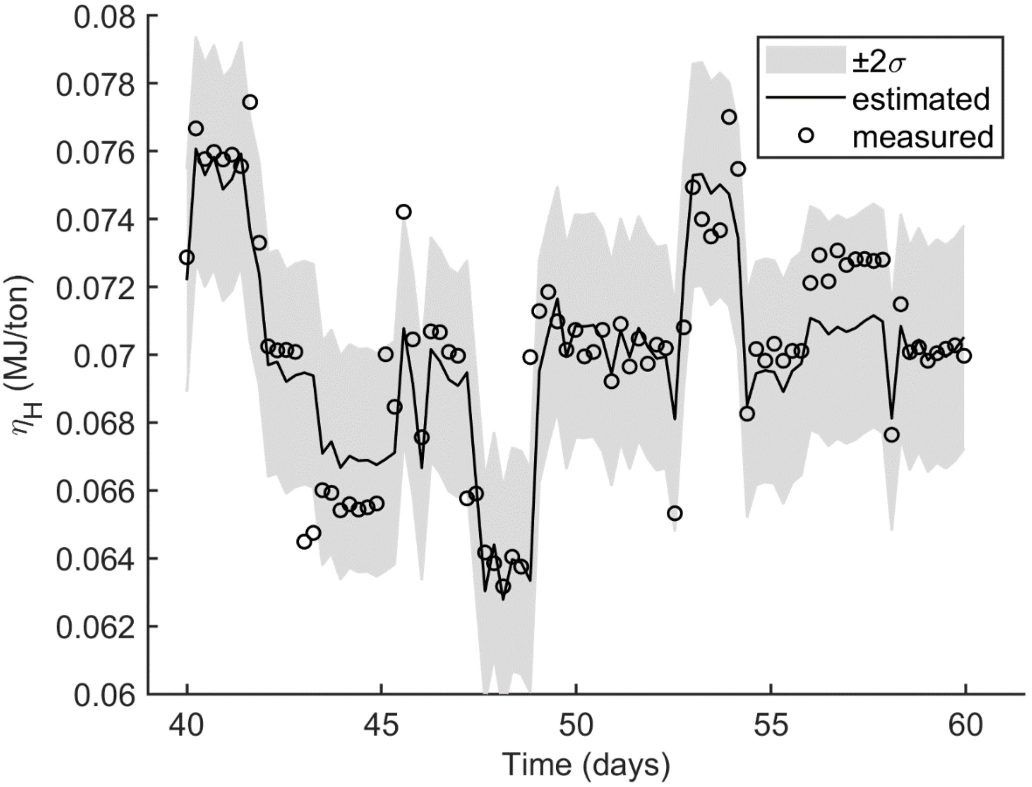

Figure 2. Measured and estimated energy efficiency for component *H* with corresponding confidence intervals. Only every 50 sample is plotted for the sake of clarity.

The difference between the training and test sets was further analyzed using Euclidean histogram distance (Ma *et al.*, 2010) and histogram intersection similarity (Swain and Ballard, 1991) and Kullback-Leibler (KL) divergence (Mathiassen *et al.*, 2002) in sliding windows for each *m* inputs used in the PLSR model. Using these, a novel metric is presented and denoted as the I_α. The I_α is given here as

$$I_\alpha = \sum_{j=1}^{m} b_{\mathrm{PLS},j}\left(\alpha D_j - s_j\right), \qquad (4)$$

where $D_j = D_j(X_{train}, X_{test})$ is the Euclidean distance between the training and test set histograms for input j and $s_j = s_j(X_{train}, X_{test})$ is the similarity between the train and test set histograms for input *j*. It was found that the KL divergence provided practically the same information as the presented distance metric, thus it was intentionally left out from the definition of the index. However, the KL divergence was found also to give a qualitative indication of the data drift. The α_j for an input variable is defined as the fraction of samples out of range in a test set window. The global parameter α is the maximum of all α_j's. It can be seen from the Eq. (4) that the proposed metric highlights the variables with more significant effect on the input. In addition, the α thresholds the Euclidean histogram distance, and the s_j acts as a penalty if the train and the test set have non-similar histograms. It should be noted that the higher index values I_α indicate a higher covariate shift, and thus higher histogram similarity needs to decrease the value of the proposed index.

The applied metrics are illustrated together with the RMSE for component *G* in Figure 3. The visual inspection in the Figure 3 shows that in fact that the training set might not be representative, as some of the model inputs diverge from the training data set. It can be seen from the Figure 3 that it is apparent that the covariate shift correlates well with the observed modelling error with testing data. Thus, monitoring the input space could be at least partially used to aid in the decision-making concerning the need of soft sensor maintenance. De facto, in actual use this issue would have to be fixed with model adaptation (or model re-training) to a more comprehensive training set. However, the recognition and tackling not only the covariate shift, but also the other type of dataset shifts such as the prior probability shift and the contextual shift (Moreno-Torres *et al.*, 2012) in real-time demands further studies on model adaptation. From these, the prior probability shift is the most of obvious to be dealt with, especially in the case of models with single output.

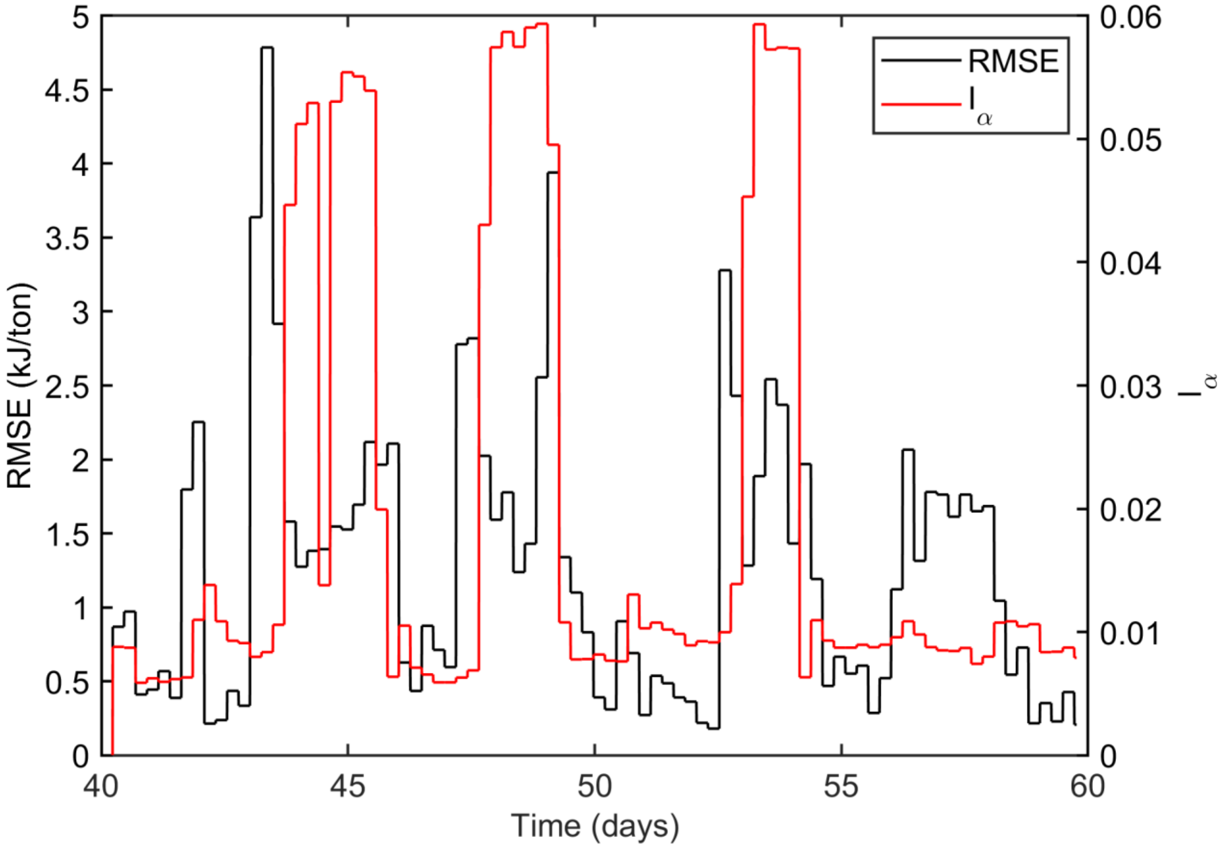

Figure 3. Illustration on the possibility of monitoring the soft sensor's covariate shift (I_α, red line) and its effect on the RMSE with testing data (black line). The illustration is computed with the sliding window size of 50 samples for the component *G*.

3.3 Practical Implications

The results of the study give guidelines to soft sensor selection in monitoring of energy intensive production processes with large input delays. Monitoring of the energy efficiency provide basis for real-time optimization of the processes also with respect to energy consumption. Hence, the study contributes to life-cycle analysis theme of the multi-step chemical processes, and by that demonstrates how the soft sensors could be utilized to lower the carbon footprint of an industrial process. In order to comprehend the analysis from the process engineering point of view, some further considerations related to practical implementation to TE process are required.

First of all, the surrogate model developed here utilizes a rather simple approach. Although a more sophisticated variable construction, delay estimation and variable selection methods may enhance the estimator performance, in industrial applications it is often beneficial to have a model structure in a representable format. In case of PLSR model, and with limited number of projections, this requirement can be met.

Additionally, it is important to give insight on the explanatory variables used in the model. In the case presented, after the correlation-based filtering the subset of 18 and 14 variables for the two surrogate models were used in PLS model estimation. From these, the most important ones were found to be:

- Component *F, G* and *H* mol-% in product stream,
- Component *D* mol-% in purge stream,
- Reactor temperature (°C),
- Product separator underflow (m^3/h),
- Stripper underflow (m^3/h),
- Compressor power (W),
- Condenser cooling water outlet temperature (°C).

Based on the presented list, the liquid molar fraction measurements of the final product components (*G, H*) in the product stream, together with the by-product *F* are needed. In TE model, these are sampled with relative high frequency of 0.25 h and 0.25 h delay. The soft sensor approach utilized in this paper considered a maximum lag of 0.5 h, suggesting that the indirect energy efficiency estimation is strongly based to recent analysis results from the product composition. Similarly, the purge stream molar fraction of reactant *D* is assumed to be measured with interval of 0.1 h and delay of 0.1 h in TE model. These assumptions set high requirements for the online gas and liquid analyzers.

As indicated by Konge *et al.* (2020), the steam cost in overall cost-efficiency of the TE process is relatively small. On the other hand, it might have more important effect to the energy efficiency. This cannot be directly seen from the most important variables selected to the energy efficiency model. However, the product separator underflow is used as an explanatory variable and this liquid stream is directed to the stripper, having impact to required steam consumption. Utilization of temperature measurements from several process points and the compressor work as explanatory variables also have natural connection to process energy efficiency. Finally, the surrogate model also uses the production rate (stripper underflow) as an input. Thus, the model incorporates most of the variables affecting to the energy efficiency by definition given in Eq. (1).

Finally, it was highlighted in this work, and also in previous studies related to surrogate modeling of TE process (e.g. Sheta *et al.*, 2019), that the selection of the training data deserves attention. Sheta *et al.* (2019) suggest approaches such as peak shaving and smoothing of intensive changes as pre-processing methods to avoid overfitting problems. However, as indicated in Section 3.2, the implementations in real systems typically need to include also efficient model adaptation as all the process points are seldom available in the training data.

Development of ensemble models can also help to reduce the estimator uncertainties and to overcome the challenges related to unseen process points (Hastie *et al.*, 2009). In addition, gradual changes due to fouling and wear of equipment, or even process design changes (which could be expected if the training set is extended over very long time period) set challenges to any surrogate models. Hence, maintenance of the soft sensor to ensure its performance over time is in fact a very interesting and important topic to study.

4 Conclusions

In this work, an indirect energy efficiency estimator was developed for the Tennessee Eastman (TE) benchmark. For this aim, the TE benchmark was modified to be suited for generating the necessary data with a realistic simulation scheme. Based on the simulated data, a surrogate-model was selected using a sophisticated model selection scheme. The final model structure was the Partial Least-Squares (PLS) regression. With these, a reasonable model performance was obtained. By monitoring the histogram similarity metrics along with the test set estimation error, it was found that the applicability of the estimator could be partially limited because of the covariate shift. All and all, the data drift was identified to be an important factor that plausibly could complicate the use of soft sensors in industrial applications. In this simulation study, this was attributed to multivariable nature of the process and motivate the future research towards selection and maintenance of soft sensors.

Acknowledgements

This work was partially funded by Business Finland via project ‘Operational eXcellence by Integrating Learned information into AcTionable Expertise (OXILATE)’, under the ITEA3 Programme and project ‘Autonomous

Processes Facilitated by Artificial Sensing Intelligence (APASSI)’.

References

A. Bakdi and A. Kouadri. An improved plant-wide fault detection scheme based on PCA and adaptive threshold for reliable process monitoring: Application on the new revised model of Tennessee Eastman process. *Journal of Chemometrics*, 32(5), 2018. doi: 10.1002/cem.2978

A. Bathelt, N. Ricker, and M. Jelali. Revision of the Tennessee Eastman Process Model. *IFAC-PapersOnLine*, 48(8):309– 314, 2015. doi: 10.1016/j.ifacol.2015.08.199S.

S. de Jong. SIMPLS: An Alternative Approach to Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263, 1993. doi: 10.1016/0169-7439(93)85002-X

J. Downs and E. Vogel. A plant-wide industrial process control problem. *Computers & Chemical Engineering*, 17(3):245-255, 1993. doi: 10.1016/0098-1354(93)80018-I

C. Drumm, J. Busch, W. Dietrich, J. Eickmans, and A. Jupke. STRUCTese® – Energy efficiency management for the process industry. *Chemical Engineering and Processing – Process Intensification*, 67:99–110. doi: 10.1016/j.cep.2012.09.009

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. 2009. Springer Science & Business Media.

T. Jockenhövel, L. Biegler, and A. Wächter. Dynamic optimization of the Tennessee Eastman process using the OptControlCentre. *Computers & Chemical Engineering*. 27:1513–1531, 2003. doi: 10.1016/S0098-1354(03)00113-3

N. Jämsä. *Model predictive control for the Tennessee Eastman process*, M.Sc. Thesis, Aalto University, 2018.

U. Konge, A. Baikadi, J. Mondi, and S. Subramanian. Data-Driven Model Based Computation and Analysis of Operability Sets Using High-Dimensional Continuation: A Plant-Wide Case Study. *Industrial & Engineering Chemistry Research*, 59(21):10043—10060, 2020. doi: 10.1021/ acs.iecr.9b07087

A. Kulkarni, V. Jayaraman, and B. Kulkarni. Knowledge incorporated support vector machines to detect faults in Tennessee Eastman Process. *Computers & Chemical Engineering*, 29(10):2128-2133, 2005. doi: 10.1016/j.compchemeng.2005.06.006

T. Larsson, K. Hestetun, E. Hovland, and S. Skogestad. Self-Optimizing Control of a Large-Scale Plant: The Tennessee Eastman Process. *Industrial & Engineering Chemistry Research*, 40:4889–4901, 2001. doi: 10.1021/ ie000586y

Y. Ma, X. Gu, and Y. Wang. Histogram similarity measure using variable bin size distance. *Computer Vision and Image Understanding*, 114(8):981–989, 2010. doi: 10.1016/j.cviu.2010.03.006

J. Mathiassen, A. Skavhaug, and K. Bø. Texture similarity measure using Kullback-Leibler divergence between gamma distributions. In *Proceedings – 7th European Conference on Computer Vision, ECCV 2002, 28-31 May, 2002, Copenhagen, Denmark*, pages 133–147, 2002. doi: 10.1007/3-540-47977-5_9

J. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. Chawla and F. Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521-530, 2012. doi: 10.1016/j.patcog.2011.06.019

R. Nikula, M. Resonant, and K. Leiviskä. Data-driven framework for boiler performance monitoring. *Applied Energy*, 183:1374-1388, 2016. doi: 10.1016/j.apenergy.2016.09.072

N. Ricker. Tennessee Eastman Challenge Archive, 2015, https://depts.washington.edu/control/LARRY/TE/download.html#Basic_TE_Code.

D. Saygin, M. Patel, E. Worrell, C. Tam, and D. Gielen. Potential of best practice technology to improve energy efficiency in the global chemical and petrochemical sector. *Energy*, 36(9):5779-5790, 2011. doi: 10.1016/j.energy.2011.05.019

A. Sheta, M. Braik and H. Al-Hiary. Modeling the Tennessee Eastman chemical process reactor using bio-inspired feedforward neural network (BI-FF-NN). *The International Journal of Advanced Manufacturing Technology*, 103:1359–1380, 2019. doi: 10.1007/s00170-019-03621-5

M. Swain and H. Ballard. Color indexing. *International Journal of Computer Vision*, *7*(1):11-32, 1991. doi: 10.1007/BF00130487

A. Tran and C. Georgakis. On the estimation of high-dimensional surrogate models of steady-state of plant-wide processes characteristics. *Computers & Chemical Engineering*, 116:56–68, 2018. doi: 10.1016/j.compchemeng.2018.02.014

D. Xie and L. Bai. A hierarchical deep neural network for fault diagnosis on Tennessee-Eastman process. In *Proceedings - 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 9-11 December, 2015, Miami, FL, USA*, pages 745–748, 2015. doi: 10.1109/ICMLA.2015.208

W. Zou, Y. Xia, and H. Li. Fault diagnosis of Tennessee-Eastman process using orthogonal incremental extreme learning machine based on driving amount. *IEEE Transactions on Cybernetics*, 48(12):3403–3410, 2018. doi: 10.1109/TCYB.2018.28