Machine Learning Supports Robust Operation of Thermosiphon Reboilers

David Appelhaus[1],*, Yan Lu[1], René Schenkendorf[2], Stephan Scholl[1], Katharina Jasch[1]

DOI: 10.1002/cite.202100063

The analysis of process and equipment operational data in chemical engineering regularly requires a high level of expert knowledge. This work presents a Machine Learning-based approach to evaluate and interpret process data to support robust operation of a thermosiphon reboiler. By applying an outlier detection, potentially interesting and unstable operating conditions can be identified quickly. A multidimensional regression allows to forecast the circulating mass flow. The results obtained fit well into the current state of research and manual evaluation of thermosiphon reboilers.

Keywords: Forecasting, Instability detection, Machine learning, Outlier detection, Thermosiphon reboiler

Received: March 12, 2021; *revised:* July 26, 2021; *accepted:* October 05, 2021

1 Introduction

Experimental data contain substantial information to be used for the development and optimization of processes and apparatuses. The analysis of experimental data involves knowledge of the physical principles and the processes under consideration. Additionally, a high amount of empirical knowledge is relied upon. In the case of high-dimensional process data, it can be challenging to identify cause-effect relationships, which results in labor-intensive data evaluation. Correlations are often determined using spreadsheet programs or numerical calculation programs, identifying the most relevant parameters. In many cases, the evaluation and interpretation of data can be accelerated if artificial intelligence and data science methods are used [1, 2]. This work shows how data science and machine learning (ML) can assist in the evaluation of data and how these results compare to existing process knowledge.

For this purpose, historical measurement data of a thermosiphon reboiler (TR) were examined. TRs are a type of evaporator often used to evaporate thermally insensitive, low-viscosity mixtures [3]. TRs are characterized by low-maintenance and low-cost operation with simultaneously high heat transfer, low fouling tendency, and short residence times [4]. However, various types of instability can occur [5]. These instabilities are manifested by fluctuations in the induced circulating mass flow and can arise due to the complex coupling of fluid dynamics and heat transfer. In terms of mathematical modelling, this is reflected by the physical equations for fluid dynamics and heat transfer which can have multiple solutions or by dynamic feedback effects [6–9]. The instabilities reduce the efficiency of TR and can lead to safety-critical situations and operational failure. They can be counteracted by selecting suitable operating conditions. However, this limits the permissible operating window.

The detection, identification, and prediction of these unstable areas are therefore essential for the design and operation of TR and can enlarge the operating window. Various authors have investigated the occurrence of unstable areas and identified stable operating conditions. The results are usually recorded in the form of stability maps. Therefore, operational parameters or dimensionless numbers are plotted against each other and stable operating ranges are defined [10, 11].

Operational instabilities depend on process parameters, such as pressure, subcooling, processed fluids, equipment design, etc. By using stability maps, the number of parameters considered is reduced to a handful. While this increases clarity, there is a risk that interdependencies may be misidentified or not recognized at all. ML algorithms allow a multidimensional analysis of TR and might improve the identification of instabilities and process control. They are likely to help avoid instabilities that occur, extending the operation area, and providing new process insights.

ML refers to data-driven, automated learning methods, which have been the subject of research since the 1950s but have received increased attention in recent years due to better algorithms and higher computing power [12, 13].

[1] David Appelhaus, Yan Lu, Prof. Dr.-Ing. Stephan Scholl, Dr.-Ing. Katharina Jasch
d.appelhaus@tu-braunschweig.de
TU Braunschweig, Institute for Chemical and Thermal Process Engineering, Langer Kamp 7, 38106 Braunschweig, Germany.

[2] Prof. Dr.-Ing. René Schenkendorf
Harz University of Applied Sciences, Automation & Computer Sciences Dep., Friedrichstrasse 57–59, 38855 Wernigerode, Germany.

Especially for the analysis and prediction of high-dimensional data with unknown correlations, ML can be applied to analyze these data more efficiently. It can support evaluating operational data, improving the understanding of the underlying processes, making predictions of future operating conditions, and revealing process states via so-called soft sensors, which calculate a target variable on the basis of other measured variables. The latter case has already been successfully implemented for TRs [14, 15].

Herein, the historical data set of a TR is analyzed using standard ML algorithms to show the advantages in data evaluation, identifying possible unstable areas, and predicting future system states.

Sect. 2 introduces the experimental setup of the TR. The used mathematical methods for analyzing the recorded data are shown, and the advantages and drawbacks for chemical engineering processes are illustrated.

Sect. 3 presents and discusses the results of a principal component analysis (PCA). An outlier detection as an unsupervised learning procedure is used to identify data of interest for further analysis while the found results are compared with existing knowledge about TR. The prediction of the circulation mass flow demonstrates the beneficial use of ML methods to identify critical values beforehand. This knowledge can be used to improve process design as well as equipment operation in the future.

2 Material and Methods

Within this section, the experimental setup of the investigated TR (Sect. 2.1) is illustrated. The used methods for the ML-based data analysis are described in Sect. 2.2 and Sect. 2.3. The data were analyzed using Python 3.7.4. The packages used for the evaluation were SciPy, scikitlearn, Numpy, and Pandas [16–19]. To extract relevant features and efficiently roll the data, tsfresh was used [20].

2.1 Experimental Setup

The experiments were carried out in a TR with a vertical shell and tube heat exchanger made of stainless steel SS1.4571 (Fig. 1). The reboiler E1 consists of three reboiler tubes with a geometry of $d_0 \times s \times L = 20 \times 2 \times 1500$ mm. The system is heated up by condensing steam at the shell side. Then, the bulk flow reaches its local boiling temperature. The two-phase flow leaves the reboiler into a vapor-liquid separator D1. The condensed vapor is recirculated to the phase separator. A detailed description of the experimental setup is given by Lu et al. [21].

The operation pressure was controlled using vacuum pump P3. The static liquid head h_s* specifies the liquid level as the ratio of the distance h between the liquid head in D1 and the inlet of E1 to the length of the reboiler tubes L_{RT}:

$$h_s^* = \frac{h}{L_{RT}} \times 100\% \tag{1}$$

Water as well as a water-glycerol mixture with $x_{Gly} = 0.33\,mol_{Gly}mol_{ges}{}^{-1}$ were used as process fluids. The operating range (Tab. 1) was chosen to work at the lower operating limit. Operating in vacuum with a small driving temperature difference and a low static liquid head is challenging for TR as unstable operating behaviors with fluctuating recirculation flows can occur [4]. Thus, a wide range of operating conditions and characteristic behaviors of TR were used for the ML analysis. Unstable operation also exists at high thermal loads, where burnout and dynamic fluctuations can occur, defining the upper operating limit. However, within the data examined, this case has not been the primary subject of investigation.

Table 1. Investigated operating conditions.

Condition	
Operating pressure [bar abs]	0.1–0.4
Driving temperature differences between reboiler inlet and heat side [K]	5–10
Static liquid head [%]	75–120

2.2 Machine Learning Based Data Analysis

2.2.1 Data Preparation

A total of 140 individual experiments were investigated. As the condensing mass flow measurement was based on weighing cycles, the average mass flow was determined. A mass flow averaged over three measured values was calculated for the periods with a continuous increase in the measured condensate mass. The condensate in container D2 was pumped back into container D1 at regular intervals. No mass flow could be determined for this period, so the overall mean condensate flow was assumed for these data points. Furthermore, measured data which have no influence on the system or represent calculated variables were not considered.

Measurement data from process engineering plants are generally available as series of discrete measurement points. A discrete-time series of an experiment i and a sensor j consisting of m measurements is forming a data series of the form:

$$S_{i,j} = \left(s_{i,j}(t_1), s_{i,j}(t_2), \ldots, s_{i,j}(t_m)\right)^T \tag{2}$$

It is possible to consider each of the m measuring points as a feature. However, this leads to a changing number of features for every experiment and large feature matrices. Furthermore, by reducing the measuring intervals, the data volume can theoretically become infinite. For this reason, it

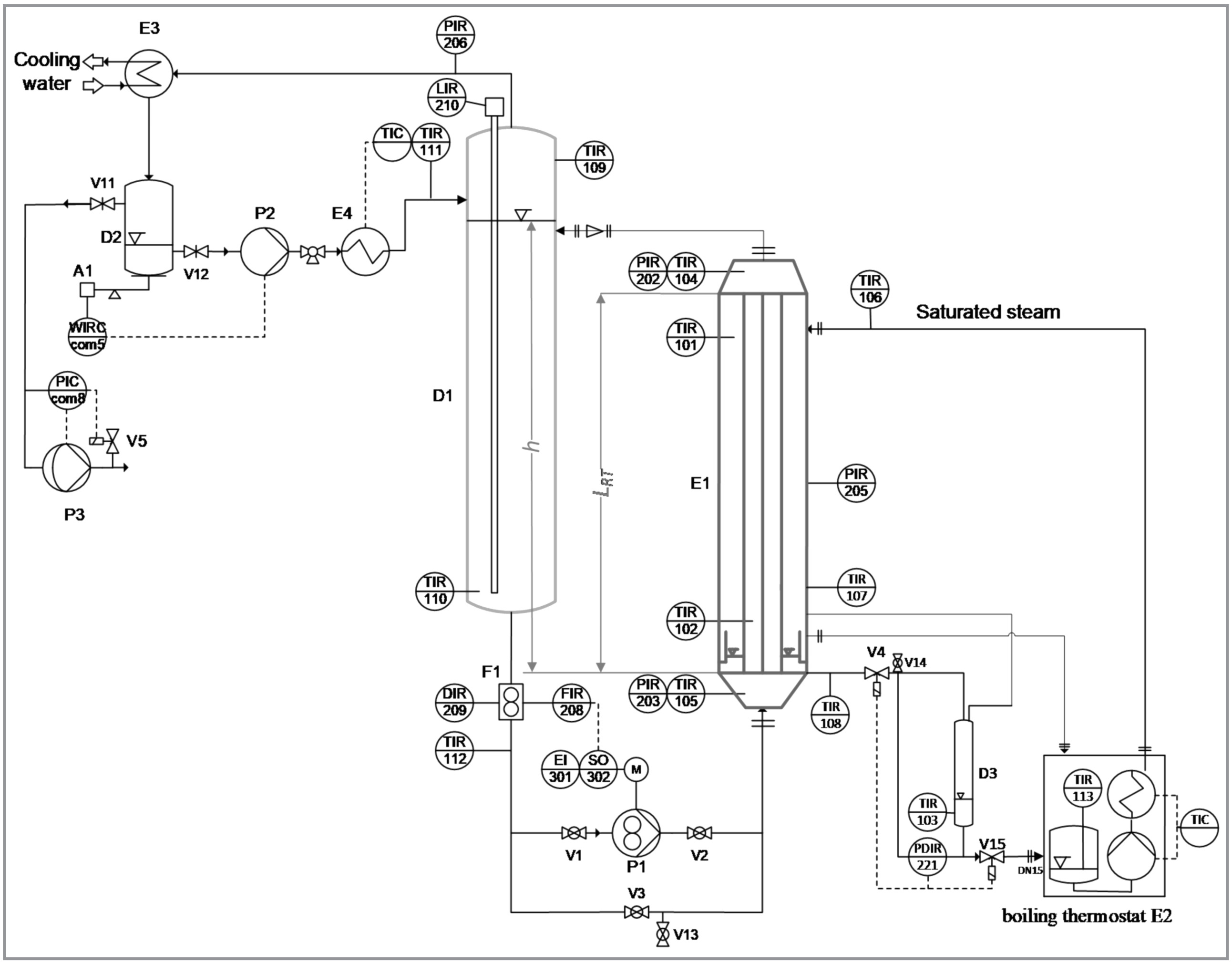


Figure 1. Experimental setup with a vertical shell and tube heat exchanger (E1), vapour-liquid separator (D1), Coriolis flowmeter (F1) phase separator (D2) including weighing (A1) and preheating (E4), and vacuum pump (P3).

is necessary to extract the relevant information. A detailed list of the extracted features can be found in Tab. 2. The set of features was chosen because it covers typical statistical features and at the same time provides partial temporal information. It was also beneficial compared to extracting only the means and standard deviations. Furthermore, no better clustering and outlier detection results were found for a significantly expanded feature set. Depending on the application, the set of extracted features can be further optimized based on the individual problem.

Table 2. List of extracted features and parameters.

Feature	Parameters
sum of values	None
Median	None
Mean	None
Skewness	None
Kurtosis	None
Quantiles	$q \in \{10\,\%,\ 30\,\%,\ 70\,\%,\ 90\,\%\}$
Variance	None
Maximum	None
Minimum	None
Autocorrelation	$\text{lag} \in \{1,\ 2,\ 3\}$

In Tab. 2, the respective parameters denote variations of the extracted features. "None" means that no further specifications were made. Extracted features whose values were not quantifiable (not a number or infinite) were removed. The features were scaled using min-max scaling, as this scaling method proved advantageous for data visualization and outlier detection. However, the scaling method might be adapted for other use-cases.

In process engineering, similar process information is measured at several points (e.g., temperature) whose values correlate. The number of relevant features can therefore be reduced by principal component analysis (PCA). Furthermore, many statistical values, e.g., median and mean, are likely to be correlated [20].

As stated in Eq. 3, when using PCA, the original data matrix $X \in \mathbb{R}^{n \times m}$ is projected into a new principal component space $P \in \mathbb{R}^{n \times m}$ through a matrix $\Gamma \in \mathbb{R}^{m \times m}$, whereby the principal components are arranged according to their share of the total variance. n is the number of samples while m is the number of features within the original data.

$$P = \Gamma^{\mathrm{T}} X \tag{3}$$

By considering only a portion of the total variance, the number of dimensions $p \leq m$ to be considered can be reduced. The higher the total variance included, the higher the number of dimensions considered. Therefore, it is necessary to find a compromise between simplification and loss of information, for which different empirical methods have been developed over time. Depending on the application, required accuracy and interpretability, the proportion of variance taken into account can vary and is usually between 90 and 99 %. In this work, 95 % of the total variance of the original characteristics was used, preserving most of the variance while reducing the number of dimensions significantly. Alternatively, dedicated methods could be used to identify an appropriate number of principal components. For example, by plotting the eigenvalues over the number of components, a point, called the elbow, can be identified at which the slope of the curve suddenly changes, or all components with eigenvalues less than the mean of all eigenvalues are discarded [22, 23].

2.2.2 Outlier Detection

Data sets often contain individual points that differ significantly from the true or expected distribution. These points are called outliers. In process engineering, this can indicate defective measuring technology or unusual operating conditions so that the identification of outliers can be helpful for process analysis. Herein, it is dealt with the extent to which an outlier analysis can be used to detect unstable operating areas in TR.

For outlier analysis, different methods of unsupervised learning can be used [24]. As standard in learning the input values are available. Still, no information about a classification or a target vector, density-based and distance-based algorithms, among others, can be applied for outlier detection. In these algorithms, the relative density or the distance to the population is estimated, whereby a data point is classified as an outlier if a limit value is exceeded. We used Mahalanobis distance as a distance-based approach because density-based approaches would likely have resulted in false-positive outliers due to the small number of measurements relative to dimensionality. One distance-based approach is the Mahalanobis distance [25]. In multidimensional space, the Mahalanobis distance describes the distance of a measurement to the center of the population, considering the covariance matrix, i.e., the distribution of the measurements.

It thus describes the scaled distance of a data point i with the coordinate vector x_{i} to a hyperspherical distributed data set with the covariance matrix S and the mean value $\bar{\mu}$ [25]. Thus, in contrast to simpler distance measures like the Euclidean distance, the number of false-positive outliers can be reduced [26].

The squared Mahalanobis distance $d_{\mathrm{m}}{}^2$ is calculated as:

$$d_{\mathrm{m}}^2 = (x_{\mathrm{i}} - \bar{\mu})^{\mathrm{T}} S^{-1} (x_{\mathrm{i}} - \bar{\mu}) \tag{4}$$

The Mahalanobis distance is closely related to Hotelling's T-squared distribution used in multivariate data analysis and fault detection [26]:

$$T^2 = \frac{n}{n-1} (x_{\mathrm{i}} - \bar{\mu})^{\mathrm{T}} S^{-1} (x_{\mathrm{i}} - \bar{\mu}) = \frac{n}{n-1} d_{\mathrm{m}}^2 \tag{5}$$

Assuming that the process data corresponds to a multivariate normal distribution, the T^2 statistic can be used. Therefore, the proportionality to the F-distribution with p and $n - 1$ degrees of freedom is applied. Here p corresponds to the number of features and n to the number of data points [26].

$$T^2 \sim \frac{p(n-1)}{n-p} F_{\mathrm{p,n-p,1-}\alpha} \tag{6}$$

A criterion for the value of the Mahalanobis distance from which a data point is classified as an outlier with the significance level α is [26]:

$$d_{\mathrm{m}}^2 > \frac{p(n^2-1)}{n(n-p)} F_{\mathrm{p,n-p;}\alpha} \tag{7}$$

Because the Mahalanobis distance is calculated in the Principal Component Space (PCS), a prediction error may occur in the principal components that are not considered.

To consider deviations in the residual subspace the squared projection error (SPE) can be used. The SPE measures the distance of the measuring point to its projection in the principal component space (PCS)

$$SPE = \left|\left|(I - \Gamma\Gamma^{\mathrm{T}})x\right|\right|^2 \tag{8}$$

with I as the identity matrix and Γ as the projection Matrix according to Eq. (3). As an upper control limit, different criteria can be used. In the context of this work the criterion δ_{h} from Jackson and Mudholkar is used [28]:

$$SPE > \delta_{\mathrm{h}}^2, \tag{9}$$

$$\delta_{\mathrm{h}}^2 = \theta_1 \left(\frac{c_\alpha \sqrt{2\theta_2 \mathrm{h}_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right)^{\frac{1}{\mathrm{h}_0}} \tag{10}$$

with

$$\Theta_{\mathrm{i}} = \sum_{\mathrm{j=p+1}}^{\mathrm{m}} \lambda_{\mathrm{j}}^{\mathrm{i}} \quad i = 1, 2, 3 \qquad (11)$$

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2} \qquad (12)$$

Here, m is the number of total features, p is the number of components in the main component space, and λ_{i} is the i-th eigenvalue of the covariance matrix. c_α describes the value for the uppermost α % of the standard normal distribution. A data point was assumed to be an outlier if either the critical value for the SPE or the Mahalanobis distance was exceeded.

2.3 Forecasting

ML allows regressions to be developed purely from patterns within the data. This enables forecasting even for complex systems [29, 30].

For forecasting the circulating mass flow of a TR, the past measured values including all measured values were defined as input variables. The procedure is illustrated in Fig. 2. In this work, exemplary 3 and 7 past measured values were considered as input variables. Thus, the dimension of X_{i} is 3 × m and 7 × m, respectively, while Y_{i} is a scalar.

For a first overview, simple linear regression methods with regularization as well as nonlinear regressors were investigated in this work.

The linear regression methods used were Lasso and Ridge with cross-validation (CV) and BayesianRidge from the package Scikit Learn.

The predicted value $\hat{y}$is calculated by a linear combination of the used features scaled by a set of weight w:

$$\hat{y}(W, X) = w_0 + w_1 x_1 + \ldots + w_{\mathrm{m}} x_{\mathrm{m}} \qquad (13)$$

For Lasso and Ridge, the individual weights are regularized in value by the ℓ_1 and ℓ_2-norm, respectively. Although Lasso and Ridge regression are both conceptually similar, the results may differ. In addition, Lasso prefers to set the individual coefficients to exactly 0, which can make it slightly more interpretable than Ridge. In Bayesian Ridge, the prefactor of the regularization term is not constant, but instead the distribution of the weights is fitted to a Gaussian distribution. The regularization parameters are adjusted based on the given data by maximizing the log marginal likelihood [31]. The weights obtained are thus similar to Ridge, with BayesianRidge being slightly more stable for ill-posed problems.

As nonlinear regressors, a decision-tree-regressor and a k-nearest-neighbor (kNN) regressor were used. The Decision-Tree-Regressor is well suited for understanding the underlying algorithm while achieving high accuracies, whereas kNN is particularly interesting from a process engineering point of view, as it can provide an insight into whether similar process conditions lead to similar changes in mass flow, provided that an acceptable accuracy is achieved.

The parameters of the individual regressors were adjusted with the Gridsearch-algorithm of scikit-learn. Deviating from the standard parameters of the regression methods, the values of Tab. 3 were used.

The coefficient of determination R^2 was used as a measure of quality for the evaluation of the regression methods:

$$R^2 = 1 - \frac{\sum_{\mathrm{i}} (y_{\mathrm{i}} - f_{\mathrm{i}})^2}{\sum_{\mathrm{i}} (y_{\mathrm{i}} - \bar{y})^2} \qquad (14)$$

The term $y_{\mathrm{i}} - f_{\mathrm{i}}$ is the difference between actual and predicted value, while $y_{\mathrm{i}} - \bar{y}$ represents the difference between the actual value and the mean value.

The direct interpretation of R^2 values can be misleading if stationary processes are investigated because the average value of the respective experiment already represents a good approximation of the following measured value in terms of the R^2 value. Because of that, in addition to a direct prediction of the circulating mass flow, the derivation of it was used as a target value for the regression.

For the case of a random walk process, no correlation should be producible with R^2 > 0, whereas R^2 > 0 indicates a correlation between input and output variables [32].

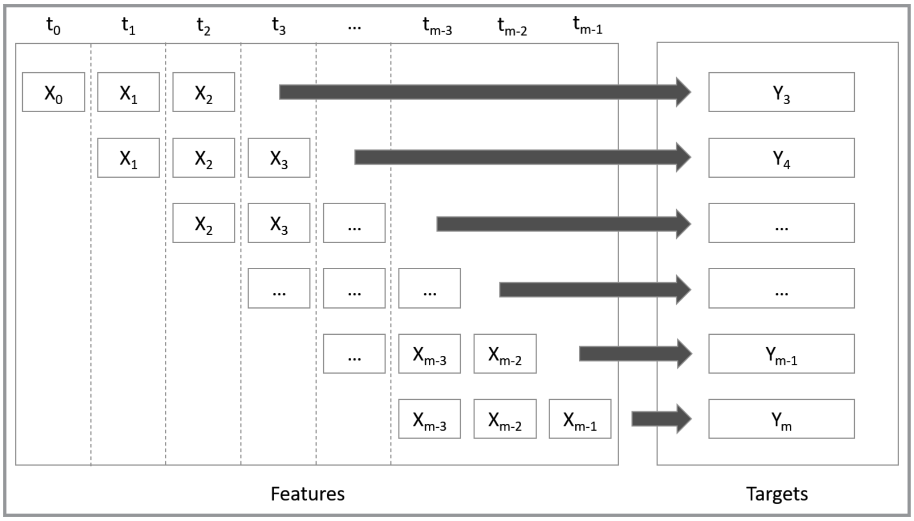


Figure 2. Schematic representation of the forecasting based on three previous measurements.

Table 3. List of the used hyperparameters within the regression. Not listed parameters were left at default.

Regression function	Parameter	Values
Lasso	α	100 logarithmically spaced values in the interval $[10^{-1}, 10^{4}]$
	Cross validations	5
Ridge	α	100 logarithmically spaced values in the interval $[10^{-1}, 10^{4}]$
	Cross validations	5
Decision-Tree-Regressor	maximum depth	6
K-Neighbors-Regressor	number of neighbors	5

3 Results and Discussion

After PCA (Sect. 3.1), an outlier detection was carried out, and the detected outliers were evaluated considering the prevailing operating conditions in Sect. 3.2. In Sect. 3.3, the data were analyzed regarding their suitability for forecasting.

3.1 Principal Component Analysis and Associated Operating States

By applying PCA, the number of features was reduced from 327 extracted features to 25 principal components. Therefore, the number of features could be reduced by 92 %, preserving 95 % of the total variance. Although each principal component greater than 10 contributes less than 1 % to the total variance, their total number was left at 25 because possible anomalies in the higher principal components may be relevant, especially in light of the outlier analysis in Sect. 3.2. Nevertheless, this shows that in a more defined problem, e.g., a classification problem, the number of principal components could be further reduced. The individual experimental data points in the first two principal components were plotted in the PCS, representing 64.5 % of the total variance (Fig. 3).

The individual tests were assigned to the respective test medium (a), operating pressure (b), static liquid height (c), and circulating flow (d). Higher circulating flows were achieved at higher pressures, liquid levels, and at low viscosities. These findings are typical for stable TR conditions. Thus, the results can be an indicator for occurring instabilities, although no unambiguous cluster for unstable operating conditions could be identified here. Furthermore, these results agree with literature values and operational experience. It is unanimously reported that low process pressures

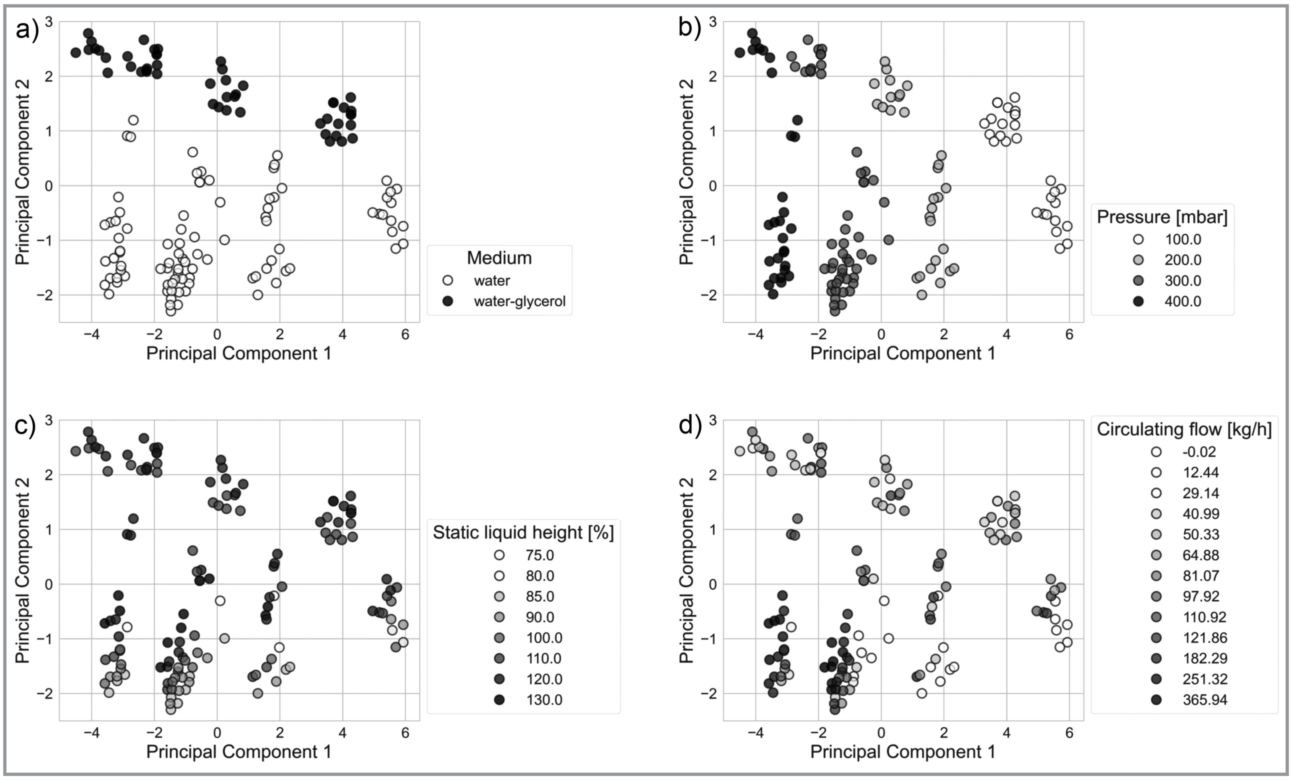


Figure 3. Illustration of the experiments within the first two principal components and the associated operating states.

and high viscosities can lead to a drying up of the circulation flow, whereby higher static liquid heights have a stabilizing effect in the area of low process pressures [33, 34]. The associated operating conditions are not the only relevant factors for occurring instabilities, as there is a scattering of the mean circulating flow within the different groups. Since only the first two principal components are plotted against each other, further correlations could be recognizable by considering additional principal components or different extracted features. Nevertheless, process data from TRs can be evaluated quickly and the effects of operating parameters described in the literature can be understood much better. Possible questions can be formulated and analyzed more precisely based on already existing data and ready-made methods. This form of visualization can improve interpretability and prevents false conclusions.

3.2 Outlier Detection and Instability Analysis

Especially in large data sets, the question of which experimental parameters lead to changes in the operating conditions is often difficult to answer. For the TR investigated here, such "points of interest" could be the occurring instabilities and, therefore, the operating limits. Outlier detection is one possibility to identify these unstable operating points through unsupervised learning, assuming that most of the examined data represent stable operating conditions. The full dataset was classified in Inliers and Outliers by means of the Mahalanobis distance identifying possible instabilities.

The number of outliers was plotted in Fig. 4 over the process parameters medium, process pressure, static liquid head, and circulating flow. The majority of 73 % of the experiments were classified as inlier, while 27 % of the data points were classified as outliers. For an outlier analysis, the value of 27 % is relatively high due to the experimental nature of the data. The data result from investigations for characterization of the unstable operating areas. They, therefore, include significantly more critical operating conditions than can be expected within an industrial plant. Thus, although the 27 % are relatively high, they appear realistic in this context. However, the assumption that the data follows a Gaussian distribution is only conditionally valid.

Due to the definition of the Mahalanobis distance as a distance-based algorithm, the question arises whether the detected outliers represent unstable operating points or whether they are boundary points of the examined state space, which, however, show a stable circulation.

First, the number of outliers for water and water-glycerol is examined. Thereby no accumulation of outliers could be found. This is remarkable as the fluid properties of the medium have a significant influence on the process. However, they only have a limited significance about whether the investigated operating conditions are stable or unstable,

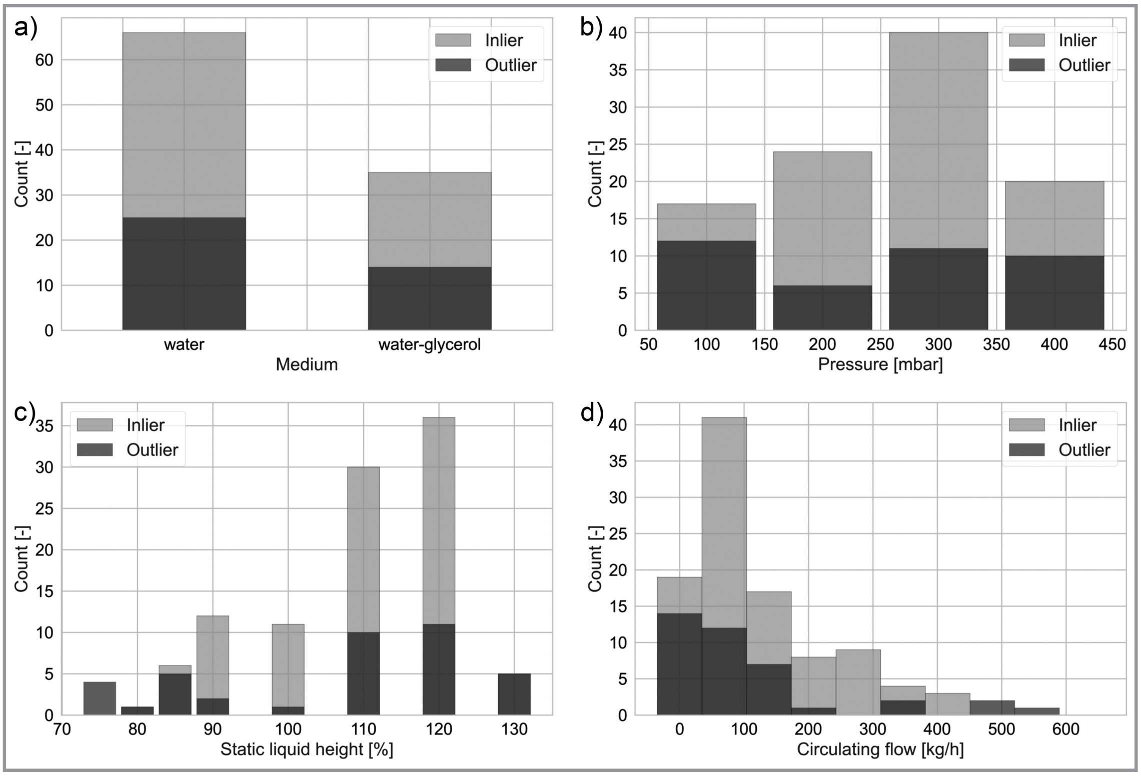


Figure 4. Illustration of the number of outliers depending on the associated operating state.

which indicates that the outliers detected might represent instabilities. Nevertheless, it must be carefully checked whether apparent accumulations can be attributed to process variations.

For the pressure and static liquid height, an above-average number of outliers can be determined for the marginal areas of the parameter space. This can be an indicator of deviating operating conditions and possible instabilities. The latter is particularly supported because an above-average number of outliers was found for circulating flows near 0 kg h^{-1} and for high circulating flows, whereby instabilities are typically to be expected here. For circulating flows between 300 and 400 kg h^{-1}, it could also be shown for the investigated TR that fluctuations exist between laminar and turbulent flow and that associated instabilities occur [21].

There is also a disproportionately high number of outliers for low and high static liquid heights of 75–85 and 130 %. This corresponds with the results of Liu et al., according to which an optimum exists for the static liquid height, and instabilities can occur for differing values [11]. The reasons for this are that for very high static liquid levels, phase change is suppressed due to hydrostatic pressure, while for very low levels, burnout can occur in the apparatus. In the experiments shown here, the optimum static liquid height was found to be in the range of 90–120 %, which can also be observed by outlier detection.

At an operating pressure of 100 mbar, most outliers were classified proportionately. This corresponds with other investigations, according to which low operating pressures and correspondingly lower boiling temperatures have a destabilizing effect on the circulation flow [34]. The main reason for this is that at lower operating pressure, the volume of the vapor component increases significantly. As a result, the circulating mass flow reacts much more sensitively to fluctuations in heat transfer and instabilities are favored.

In contrast, there is a high number of outliers for an operating pressure of 400 mbar, which contradicts this statement. Here, however, a more detailed look showed that they were detected mainly for very low and high circulating flows and high static liquid heights. Especially for the water-glycerol mixture, only low temperature differences and thus low circulation flows could be achieved due to the limited heating capacity of the used thermostat. The outliers are therefore found for operating conditions that tend to be regarded as unstable, although this would not be expected when considering the pressure solely.

The outlier detection used were suitable for identifying unstable areas described in the literature without prior knowledge. For TR, this can be an excellent possibility to assist the plant operator of a plant in critical operating conditions, which is already used in a similar form in statistical process control [35, 36]. Furthermore, this methodology can also be applied to unknown process conditions, and thus interesting experimental results can be identified. This can accelerate process evaluation and prevent false assumptions especially if the process under investigation is depending on a high number of parameters.

3.3 Prediction of the Circulating Mass Flow

As an important criterion for efficiency and stability of TR, the circulating mass flow was predicted. 50 out of the full dataset of 140 performed experiments were randomly selected, and their features were determined according to Sect. 2.3. The training was performed with 70 %/30 % training/test data.

The results of the investigated regression methods are shown in Fig. 5. The standard deviations were calculated, performing a 3-fold validation. For this purpose, the

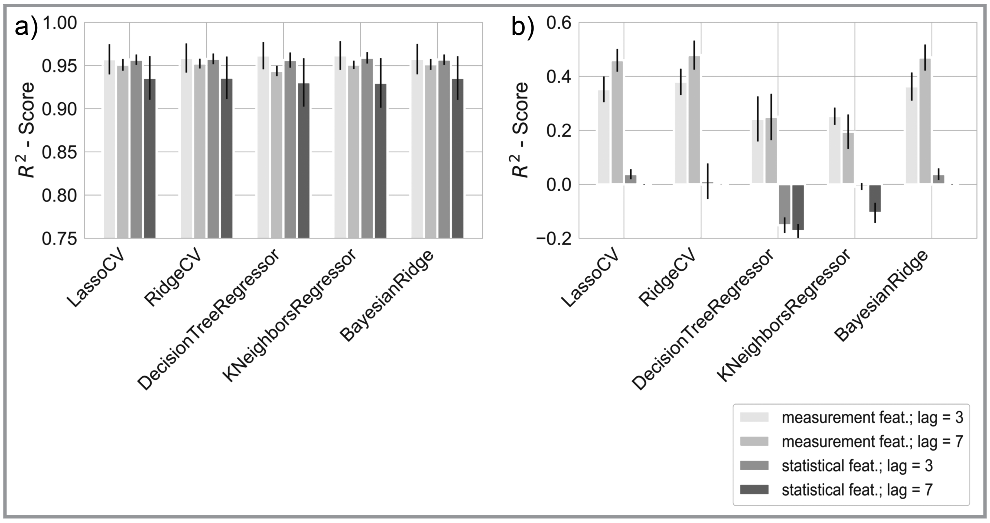


Figure 5. a) R^2 Score for the prediction of the circulating mass flow. b) R^2 Score for the prediction of the deviation of the circulating mass flow.

algorithm used was trained and tested three times, each time with a new randomly selected data set consisting of 50 experiments. For each of the regression methods used, a distinction is made about whether statistical features are extracted from the measured values used (statistical feat.) or whether they are used directly (measurement feat.). The past three (lag = 3) or seven (lag = 7) measuring points are used. Using the circulating mass flow directly as target value (Fig. 5a), values for R^2 near 1 are achieved for all features and regression methods. The difference between the different methods lies within the standard deviation. Considering a higher number of past measurements seems to worsen the regression, which could be due to the loss of temporal information. Using the actual value of the circulating mass flow as target, there is the possibility that a fault occurs, assuming that the mean value is a good assumption for the next value. Although this results in high values for R^2 there is no information about the future system behavior included.

Therefore, the derivative of the circulating flow was also considered as a target value. The results are shown in Fig. 5b. The values obtained for R^2 are significantly lower than before.

For this case, feature extraction is shown to be disadvantageous for all regression methods. It can be concluded that there is a significant loss of relevant information due to feature extraction, which could be counteracted by the extraction of other features. Using the extracted features here, the values of R^2 were around 0, which means that no information gain is possible.

For the measurement features, the linear regression methods could achieve values of $R^2 > 0.3$. This means that, in principle, there is a deterministic correlation between measured system values and the fluctuation of the circulating mass flow. Considering the individual measured quantities in the form of linear regressions can reflect this correlation. However, it cannot be determined which variables significantly influence the fluctuation because some of the used measurements like the various temperature sensor values correlate with each other.

The generally low values of R^2 could result from the generally poor predictability of the temporal change due to random fluctuations. However, it is more likely that the information contained in the measured data is not sufficient for a better prediction. The increase in the number of experiments considered did not lead to an increase in the prediction accuracy. It is conceivable that system variables influencing the instability were not measured, whereas the known influencing variables were recorded.

However, it is most likely that due to the low measurement frequency of about 0.3 Hz, temporal information is not sufficiently captured. Therefore, it is expected that increasing the measurement frequency will also increase the prediction accuracy.

The linear regression methods showed better predictive accuracy than the nonlinear regressors kNN and Decision Tree. These were chosen in particular to allow interpretability, which does not seem reasonable given the low R^2 values. One reason for the poorer performance of kNN could be that the data points are not dense in the parameter space studied. For the Decision Tree only discrete output values are considered. This could worsen the generalizability of this algorithm. However, it is possible that other nonlinear regression methods may allow a significantly better prediction accuracy. The approach of using procedurally known relationships to map nonlinear relationships while keeping the computational cost low seems promising. This could be done, for example, by multiplying or exponentiating the original measured values based on known physical principles.

The linear regression methods used all have similar predictive accuracy. The selection of a specific algorithm therefore depends on the application. Bayesian Ridge, for example, does not require any pre-selection of the regression parameters, but on the other hand also has the highest training effort. Ridge is a widely used algorithm, which seems to be particularly suitable for transfer to other applications. With Lasso, individual weights are preferably set to 0 due to the ℓ_1-norm, which is particularly advantageous for feature selection. This is also of interest with regard to a good interpretability of the algorithms used.

It is shown that the use of up to 7 past measurement data improves the prediction accuracy compared to the usage of just 3 past measurements. The measurements were taken every 3 to 4 seconds. So, the process conditions prevailing more than 9 seconds ago still seem to contribute to the changes in mass flow. However, it is also possible that despite the regularization used, the regression algorithms are over-fitted. By varying the considered measurements used at a higher measuring frequency and the regularization parameter, the optimum number of past measured values can be determined.

As can be seen from the lower value of R^2, the temporal information is of considerable importance for the prediction accuracy. As time progresses, the individual input variables have less influence on future values. A further improvement of the prediction accuracy can thus be expected with an increase in the measurement frequency. This method can be used to determine optimal measurement frequencies for a process under consideration. Above a particular cut-off frequency, the additional information should have only a marginal effect on the prediction accuracy.

Knowledge about the future process behavior enables the development of improved control strategies. The realization and implementation of such control strategies are subject of extensive research and require an accurate model to predict the state of the plant, which usually requires numerous assumptions in addition to a good understanding of the system [37].

As shown here, ML can also be used for predicting the TR system behavior. With $0.5 > R^2 > 0.3$, the prediction accuracy still offers much potential for improvement, but fluctuations can be predicted within certain limits. The necessary system knowledge is low, so this method can be applied to complex systems.

For further optimization, the incorporation of process knowledge for the development of task-specific features can be beneficial. Systematically taking into account and not taking into account the uncorrelated input variables can offer significant gains in knowledge about how the target variable depends on the input variables under investigation. This can reduce the investigation effort and can help to identify relevant operational parameters.

4 Conclusions

Using the analysis of operational data of a thermosiphon reboiler as a showcase, the great potential of ML and systematic data analysis for an improved understanding of equipment performance has been demonstrated. PCA has proven to be advantageous in the evaluation and analysis of complex data sets. Applying PCA, trends in data can be identified quickly and effectively, especially in highly dimensional datasets.

For the TR under investigation, the operating variables pressure and static liquid height, as well as the medium viscosity, can be compared qualitatively well with the prevailing average circulation flow by the principal component analysis, and correlations between process parameters can be identified.

An outlier detection revealed potentially unstable operating areas without the need for extensive process knowledge. Thus, for the parameters examined, the experimental data can be efficiently pre-selected, and areas of interest can be identified. For a more detailed evaluation, the outliers and test results can be classified more precisely and evaluated based on existing process knowledge.

Furthermore, it was investigated whether there is a deterministic correlation between the measured values and the change in the circulation flow. It was shown that R^2 values of over 0.3 can be achieved by directly using the previously measured values and linear regression methods. In contrast, statistical features or direct predictions of the circulation flow did not provide any additional information about the process.

Using nonlinear regressions or increasing the measuring frequency, the prediction accuracy could be further increased so that a control system for preventing unstable operating states could be established. By changing the input variables, deterministic relationships could be identified quickly, even without process knowledge or validated in the case of calculated process variables.

ML offers a valuable set of methods for analyzing chemical engineering equipment in general and TR in special. The time required for data evaluation can be significantly reduced, system understanding can be improved, and unknown relationships can be identified effectively.

Acknowledgement

Yan Lu would like to thank the German Federal Environmental Foundation (Deutsche Bundesstiftung Umwelt – DBU) for the Doctoral Scholarships. Open access funding enabled and organized by Projekt DEAL.

Symbols used

c	[–]	sum of standard normal distribution
d_m	[–]	Mahalanobis distance
d_0	[m]	outer diameter
f	[–]	predicted value
h	[m]	height
h_s*	[%]	static liquid height
I	[–]	identity matrix
L	[m]	length
m	[–]	total number of measurements
n	[–]	number of samples
p	[–]	number of principle components
P	[–]	principal component space matrix
R^2	[–]	coefficient of determination
s	[m]	wall strength
$s_{\mathrm{i,j}}(t_1)$	[–]	data point of sensor j and experiment *i*
$S_{\mathrm{i,j}}$	[–]	time data series of sensor j and experiment *i*
S	[–]	covariance matrix
w	[–]	linear regression weights of features
x	[mol mol^{-1}]	mole fraction
x	[–]	coordinate vector
X	[–]	original feature matrix
y	[–]	true value

Greek letters

α	[–]	significance level
Γ	[–]	Projection Matrix
δ_h	[–]	delta criterion
λ	[–]	eigenvalue
$\bar{\mu}$	[–]	mean vector

Sub- and Superscripts

Gly	glycerin
m	Mahalanobis
RT	reboiler tube

Abbreviations

CV	cross validation
kNN	k-nearest-neighbor

ML	machine learning
PCA	principal component analysis
PCS	principal component space
SPE	squared prediction error
TR	thermosiphon reboiler

References

[1] D. A. Beck, J. M. Carothers, V. R. Subramanian, J. Pfaendtner, *AIChE J.* **2016**, *62 (5)*, 1402–1416.

[2] V. Venkatasubramanian, *AIChE J.* **2019**, *65 (2)*, 466–478. DOI: https://doi.org/10.1002/aic.16489

[3] A. Mersmann, M. Kind, J. Stichlmair, *Thermische Verfahrenstechnik: Grundlagen Und Methoden*, Springer, Berlin **2005**.

[4] S. Arneth, J. Stichlmair, *Int. J. Therm. Sci.* **2001**, *40 (4)*, 385–391. DOI: https://doi.org/10.1016/S1290-0729(01)01231-5

[5] A. W. Sloley, *Chem. Eng. Prog.* **1997**, *93 (3)*.

[6] L. E. O'Neill, I. Mudawar, *Int. J. Heat Mass Transfer* **2018**, *123*, 143–171.

[7] V. Pandey, S. Singh, *Chem. Eng. Sci.* **2017**, *168*, 204–224. DOI: https://doi.org/10.1016/j.ces.2017.04.041

[8] M. Kessler, S. Kabelac, *Heat Transfer XIII: Simulation and Experiments in Heat and Mass Transfer* (Eds: B. Sunden, C. A. Brebbia), WIT Press, Southampton, UK **2014**, 325–336.

[9] L. C. Ruspini, C. P. Marcel, A. Clausse, *Int. J. Heat Mass Transfer.* **2014**, *71*, 521–548. DOI: https://doi.org/10.1016/j.ijheatmasstransfer.2013.12.047

[10] J. A. Boure, A. E. Bergles, L. S. Tong, *Nucl. Eng. Des.* **1973**, *25 (2)*, 165–192. DOI: https://doi.org/10.1016/0029-5493(73)90043-5

[11] Y. Liu, Z. Li, Y. Li, Y. Jiang, D. Tang, *Appl. Therm. Eng.* **2019**, *151*, 262–271. DOI: https://doi.org/10.1016/j.applthermaleng.2019.02.031

[12] A. M. Turing, *Mind* **1950**, *59 (236)*, 433–460. DOI: https://doi.org/10.1093/mind/LIX.236.433

[13] F. Corea, *An Introduction to Data: Everything You Need to Know About AI, Big Data and Data Science*, Springer International, Cham **2019**.

[14] S. Zaidi, *Chem. Eng. Res. Des.* **2015**, *98*, 44–58. DOI: https://doi.org/10.1016/j.cherd.2015.04.002

[15] S. Zaidi, *Chem. Eng. Sci.* **2012**, *69 (1)*, 514–521. DOI: https://doi.org/10.1016/j.ces.2011.11.005

[16] P. Virtanen et al., *Nat. Methods.* **2020**, *17 (3)*, 261–272. DOI: https://doi.org/10.1038/s41592-019-0686-2

[17] F. Pedregosa et al., *J. Mach. Learn. Res.* **2011**, *12 (85)*, 2825–2830.

[18] C. R. Harris et al., *Nature* **2020**, *585 (7825)*, 357–362. DOI: https://doi.org/10.1038/s41586-020-2649-2

[19] W. Mckinney, *pandas: a Foundational Python Library for Data Analysis and Statistics. Python High Performance Science Computer*, **2011**. www.researchgate.net/publication/265194455_pandas_a_Foundational_Python_Library_for_Data_Analysis_and_Statistics/citations

[20] M. Christ, N. Braun, J. Neuffer, A. W. Kempa-Liehr, *Neurocomputing* **2018**, *307*, 72–77. DOI: https://doi.org/10.1016/j.neucom.2018.03.067

[21] Y. Lu, K. Jasch, S. Scholl, *Chem. Ing. Tech.* **2021**, *93 (7)*, 1142–1151. DOI: https://doi.org/10.1002/cite.202000236

[22] H. Abdi, L. J. Williams, *WIREs Comput. Stat.* **2010**, *2 (4)*, 433–459. DOI: https://doi.org/10.1002/wics.101

[23] J. Josse, F. Husson, *Comput. Stat. Data Anal.* **2012**, *56 (6)*, 1869–1879. DOI: https://doi.org/10.1016/j.csda.2011.11.012

[24] V. J. Hodge, J. Austin, *Artif. Intell. Rev.* **2004**, *22 (2)*, 85–126. DOI: https://doi.org/10.1007/s10462-004-4304-y

[25] P. C. Mahalanobis, *Natl. Inst. Sci. India* **1936**, *2 (1)*, 49–55.

[26] S. J. Qin, *J. Chemom.* **2003**, *17 (8–9)*, 480–502. DOI: https://doi.org/10.1002/cem.800

[27] R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing*, 4th ed., Academic Press, San Diego, CA **2017**.

[28] J. E. Jackson, G. S. Mudholkar, *Technometrics* **1979**, *21 (3)*, 341–349.

[29] Z. Wu, D. Rincon, P. D. Christofides, *Ind. Eng. Chem. Res.* **2020**, *59 (6)*, 2275–2290. DOI: https://doi.org/10.1021/acs.iecr.9b03055

[30] J. B. Rawlings, C. T. Maravelias, *AIChE J.* **2019**, *65 (6)*, e16615. DOI: https://doi.org/10.1002/aic.16615

[31] M. E. Tipping, *J. Mach. Learn. Res.* **2001**, *1 (6)*, 211–244.

[32] C. Chatfield, *Time-Series Forecasting*, Chapman and Hall, London **2000**.

[33] C. Tompkins, M. Corradini, *Nucl. Eng. Des.* **2018**, *332*, 267–278. DOI: https://doi.org/10.1016/j.nucengdes.2018.03.018

[34] P. J. Heggs, A. Alane, in *Proc of the 2010 14th International Heat Transfer Conference*, ASMEDC, Washington, DC **2010**.

[35] I. Jaffel, O. Taouali, M. F. Harkat, H. Messaoud, *IFAC-PapersOnLine.* **2015**, *48 (21)*, 1397–1401. DOI: https://doi.org/10.1016/j.ifacol.2015.09.720

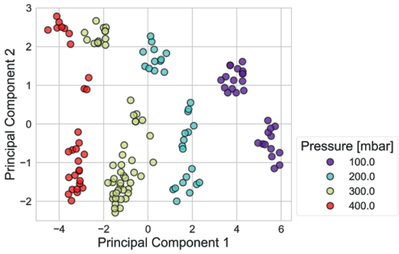
[36] G. Verdier, A. Ferreira, *IEEE Trans. Semicond. Manuf.* **2011**, *24 (1)*, 59–68. DOI: https://doi.org/10.1109/TSM.2010.2065531

[37] A. S. Kumar, Z. Ahmad, *Chem. Eng. Commun.* **2012**, *199 (4)*, 472–511. DOI: https://doi.org/10.1080/00986445.2011.592446

DOI: 10.1002/cite.202100063

Machine Learning Supports Robust Operation of Thermosiphon Reboilers

David Appelhaus, Yan Lu, René Schenkendorf, Stephan Scholl, Katharina Jasch*



Research Article: Thermosiphon reboilers have complex flow patterns and unstable areas. In this work historical experimental data were analyzed using machine learning. Potentially unstable test points were identified using outlier detection. Different regression methods were investigated regarding their prediction accuracy of the circulating mass flow. .... ■