From Data to Alarms: Data-driven Anomaly Detection Techniques in Industrial Settings

Rastislav Fáber[a,*], Martin Mojto[a], Karol Ľubušký[b] and Radoslav Paulen[a]

[a]*Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Bratislava, Slovakia*
[b]*Slovnaft, a.s., Bratislava, Slovakia*
rastislav.faber@stuba.sk

Abstract

This paper introduces a data-driven methodology for anomaly detection in industrial processes. Our focus is on minimizing misclassifications of normal operations and enhancing anomaly and outlier detection. This optimization is based on presumed ground truth (GT) labels associated with a dependent variable (isobutane concentration). Utilizing a moving-horizon approach on an extensive industrial dataset, we perform a comprehensive evaluation of filtering algorithms, and present a representative outlier classification. Secondly, effective anomaly detection, distinct from outlier detection, is achieved by integrating a regression model trained on measurements from independent process variables to fit the dependent variable. Trained regression models consistently achieve effective prediction, staying within an approved process tolerance.

Keywords: Streaming data, Anomaly detection, Outlier detection, Regression

1. Introduction

Ensuring reliable control in any industrial plant necessitates the validation of real-time measurements. Anomaly detection proves effective in identifying subtle signs of malfunctions that could escalate into serious issues. While operators traditionally rely on periodic laboratory samples, there is a growing interest in integrating advanced algorithms to streamline this process and minimize operational burdens. Our focus encompasses the detection of outliers, representing sudden, sharp changes in the monitored signal, as well as distinguishing between the true dynamics of the process and anomalous dynamics caused by measurement disturbances. When anomalous measurements are detected, the operation room should be alarmed.

We employ real-time data analysis to validate incoming online measurements, prioritizing simplicity — an essential factor for implementation on industrial hardware. Various methods, such as threshold or standard deviation filters (Afanasyev and Fedorova, 2019; Blázquez-García et al., 2021), have been explored for detecting outliers in univariate and multivariate time series data. Some papers consider calculating the local mean of a time series using exponentially decreasing weight factors for each prior measurement (Carter and Streilein, 2012; Roberts, 1959). An alternative involves a regression approach that utilizes data-based modeling to identify outliers in a multivariate context, leveraging the autoregressive nature of the model for a nuanced understanding of system dynamics (Yoon et al., 2022). Mathematical models, grounded in fundamental

Acknowledgements: This work is funded by the Slovak Research and Development Agency (project no. APVV-21-0019), by the Scientific Grant Agency of the Slovak Republic (grant no. 1/0691/21), and by the European Commission (grant no. 101079342, Fostering Opportunities Towards Slovak Excellence in Advanced Control for Smart Industries).

laws of nature, offer a deeper grasp of complex process dynamics. For instance, a Kalman filter can be coupled with existing models to estimate the system state (Jin et al., 2022). Another alternative involves deep autoencoders (long short-term memory or convolutional neural network-based models), for enhancing productivity in complex time-series data from the industry (Tziolas et al., 2022).

2. Methodology

Our goal is to generate reliable outcomes of outlier detection in a one-dimensional vector, denoted as $y(t)$. This vector exhibits occasional unexplained behavior and is part of a broader time-series dataset encompassing all process variables over an extended time period, denoted as $\boldsymbol{X}$. Initially, we enhance the signal clarity by filtering out random fluctuations and noise. As a dynamic model (essential for a Kalman filter) is unavailable, regression techniques offer a viable alternative. We conduct an analysis of the relationships among process variables within $\boldsymbol{X}$ to detect existing trends in the dataset.

2.1. Data Treatment

Visual inspection effectively spots systematic errors, but not all are easily caught. This section explores multivariate data methods to address errors beyond visual detection.

Three-standard-deviations rule. This widely used method assumes a normal data distribution. The 3-sigma interval *T* is defined based on the sample mean $\hat{y}$ and the standard deviation σ in Eq. (1). Observations outside this interval are considered outliers, with approximately 99.7% of data expected within 3-sigma from the mean. We use:

$$T = \hat{y} \pm 3\sigma, \tag{1}$$

$$T_{t_i} = \hat{x} \pm \chi^2_{n,0.997} \boldsymbol{S}^{1/2} e. \tag{2}$$

Eq. (2) uses the matrix square root $\boldsymbol{S}^{1/2}$ and unit vector *e* and $\chi^2_{n,0.997}$ is the quantile of the χ^2 distribution with *n* degrees of freedom and a probability level of 99.7%.

Minimum Covariance Determinant (MCD). This robust method (Rousseeuw and Driessen, 1999) detects outliers in multivariate data using the Mahalanobis distance:

$$d_{MCD,t_i} = \sqrt{(\boldsymbol{x}(t_i) - \hat{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x}(t_i) - \hat{\boldsymbol{x}})}, \tag{3}$$

which shows the dissimilarity between a measurement $\boldsymbol{x}(t_i)$ and the underlying probability distribution using the d_{MCD}. It achieves robustness by iteratively identifying data subsets with the minimum determinant of the sample covariance matrix $\boldsymbol{S}$, mitigating outlier influence. The process continues until the determinant of $\boldsymbol{S}$ stabilizes. The methods discussed focus on global anomalies and may not effectively capture local deviations. We address this in the following text by taking the temporal dimension into account.

2.2. Outlier Detection using Data Averaging

This method targets unusual values in the local signal behavior within a window of size *N*, allowing flexibility in focusing on local, temporary, or global deviations based on the chosen *N*. The confidence interval in this method is computed as:

$$T_{t_i} = \hat{y}(t_i) \pm t_{N,0.997} \sqrt{\sigma^2 / N}, \tag{4}$$

where $t_{N,0.997}$ represents the inverse of Student's *t* distribution (Student, 1908) with *N* degrees of freedom, and σ^2 corresponds to the variance within the monitored window.

Statistical Mean. This method involves calculating the absolute mean of $y(t)$ over extended periods (N ranges from months to years). The detection criterion is based on (1).

Temporal Mean. By evaluating the mean of consecutive data point differences $\Delta y(t_i) = y(t_i) - y(t_{i-1})$, we detect immediate variations in measurements. Outliers are identified when deviating from the interval in (1).

Simple Moving Average (SIMA). A dynamic average, adapting to dataset changes, uses (5) with a fixed window of past measurements (Oppenheim, 1999). Observations outside the interval (4) are identified as outliers, highlighting inconsistencies in recent history.

$$\hat{y}_{t_i} = \frac{1}{N}\sum_{j=0}^{N-1} y(t_{i-j}). \tag{5}$$

Symmetric Moving Average (SYMA). When dealing with time-series data, our knowledge of future measurements is uncertain. Thus, we use this approach only for evaluating past detection outcomes. We compute the average as:

$$\hat{y}_{t_i} = \frac{1}{N}\sum_{j=-\lfloor (N-1)/2 \rfloor}^{\lfloor (N-1)/2 \rfloor} y(t_{i-j}). \tag{6}$$

Predictive Moving Average (PMA). We enhance SIMA with additional information from a prediction model to dynamically adjust its value based on other process variables. The predicted difference is added to the filtered value obtained from past measurements:

$$\hat{y}_{t_i} = \frac{1}{N}\sum_{j=0}^{N-1} y(t_{i-j}) + \Delta\hat{y}(\boldsymbol{x}(t_i)). \tag{7}$$

2.3. Anomaly Detection using Regression Methods

We leverage predictive models to identify outliers of the dependent variable based on the positions of measurements relative to the model predictions.

Ordinary Least Squares (OLS). A standard linear regression finds model parameters $\boldsymbol{\beta}$ by minimizing the squared 2-norm of differences of observed and predicted values:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{i=1}^{\mathrm{N}} (y(t_i) - \boldsymbol{\beta}^T \boldsymbol{x}(t_i))^2. \tag{8}$$

LASSO. This method extends the regression by incorporating a penalty term based on 1-norm, encouraging model sparsity (Santosa and Symes, 1986). It effectively identifies and reduces the impact of less relevant variables by solving (9), where λ balances model accuracy and overfitting. The ℓ_1-penalization element leads some parameters to become zero, resulting in a less complex, more robust, and interpretable model.

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{i=1}^{\mathrm{N}} (y(t_i) - \boldsymbol{\beta}^T \boldsymbol{x}(t_i))^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{9}$$

Principal Component Regression (PCR). Principal Component Analysis (PCA) proves valuable in enhancing the interpretability of large, multi-dimensional datasets (Pearson, 1901). By generating new uncorrelated variables, PCA maximizes variance, reducing data dimensionality while minimizing information loss. Subsequently, OLS or LASSO can be employed to learn the model in the latent space. The synergistic application of PCA and LASSO, denoted as PCA+LASSO, harnesses the strengths of both methods.

3. Case Study

The current industrial norm relies heavily on manual processes and lab sampling for anomaly detection. Incorporating an automated algorithm would ease this burden,

notifying operators only when an outlier is detected. Our investigation involves a comprehensive industrial dataset (Fáber et al., 2023) with >500 process variables. After preprocessing, we selected 377 entities with 15,907 measurements ($\boldsymbol{X} \in \mathbb{R}^{377\times15,907}$). The monitored variable, $y(t)$, represents isobutane concentration, measured every 15 minutes. Within the dataset, we identify three outlier types: level shifts, slow drifts, and additive outliers. Level shifts cause an instant change. The dataset is affected significantly with values potentially returning to the previous level. Drifts gradually deviate, forming challenging-to-detect outliers. Additive anomalies result in unusual values for a single observation, with subsequent points unaffected. The difference between an outlier and anomaly lies in time duration, deviation on short (hours) time scale stands for an outlier; if the deviation lasts longer and requires manual calibration, we classify it as an anomaly.

4. Results

Accurately evaluating anomaly detection methods relies on the presence of anomalies in the dataset. However, even plant operators may struggle to identify anomalies reliably in historical data. To address this, we construct ground truth (GT) labels using a seventh-order SYMA, as shown in (6). This method identified 306 outliers among 15,907 measurements. It is important to note that some outliers in the dataset might not be captured in the GT labels due to missing information from other process variables $\boldsymbol{X}(t)$.

In our evaluation of filter-based approaches, we tested various filter orders and determined that order 7 was the most effective. This choice was validated on a training set of 1,881 measurements, aligning closely with laboratory samples and resulting in the detection of approximately 7.07% of outliers. Subsequent testing on a 750-measurement dataset identified around 7.87% outliers. The inclusion of a higher number of past data points, compared to the SYMA (6), highlights the substantial noise present in the data.

In regression analysis, models were trained to predict both isobutane concentration $\hat{y}(t) = \boldsymbol{\beta}^T\boldsymbol{x}(t)$ and the backward time difference of isobutane concentration $\hat{y}(t) = \boldsymbol{\beta}_\Delta^T\boldsymbol{x}(t)$ using methodology from Section 2.3. Before applying the algorithms, a preprocessing step utilized the MCD method to remove outliers from $\boldsymbol{X}$ (Section 2.1). The dataset was randomly split into training and testing sets (80/20 ratio) for model learning and evaluation. Efficacy was assessed using the Root Mean Square Error (RMSE), with deviations indicating potential outliers:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y(t_i) - \boldsymbol{\beta}^T\boldsymbol{x}(t_i))^2}, \tag{10}$$

where N is the number of training points. Specifically, we considered the ±2×RMSE confidence interval over ±3×RMSE due to challenges in capturing dataset variability, especially in the presence of outliers among independent variables.

We chose the first seven principal components for PCA-trained models, explaining approximately 62% of the overall variance using the elbow method. Additional components made minimal contributions, indicating saturation in capturing dataset variability. Similarly, for LASSO and the PMA, we applied thresholds of 0.08 and 0.8, respectively, to select relevant coefficients. This ensures that only impactful coefficients are retained, allowing for a more interpretable and effective regression outcome. The prediction model achieved RMSE values ranging from 0.4031 (LASSO) to 0.6655 (PCA+LASSO), all within ±5% of isobutane concentration, meeting the industry-standard confidence region. Variable selection by LASSO, OLS, and PCA provided insights into crucial features. LASSO identified n-butane concentration, olefin feed

Table 1: Confusion matrix entries for implemented outlier detection methods.

Method	TP	TN	FP	FN
Statistical Mean (SM)	15601	0	306	0
Temporal Mean (TM)	15431	61	245	170
Simple Moving Average (SIMA)	14553	122	184	1048
Predictive Moving Average (PMA)	14590	125	181	1011
Regression model	15175	15	291	426

concentration, and pressures of olefin and recycle streams. OLS also demonstrated reasonable variable selection, focusing on pressures at the deisobutanizer accumulator inlet/outlet. However, PCA exhibited less favorable outcomes, selecting compressor discharge, vibrations, and ventilation. Discussions with our industrial partner deemed this variable selection as unsuitable. When analyzing the prediction models, we observe a notable alignment between the fit by LASSO and laboratory measurements, prompting consideration of further investigation in future research.

In predicting output differences, LASSO yielded the lowest RMSE (0.037) among regression methods, closely followed by OLS and PCA+LASSO (RMSE = 0.038). PCR, on the other hand, yielded a higher RMSE = 0.178. Selected variables coincide with those identified by regression models, substituting some with pressure/temperature differences.

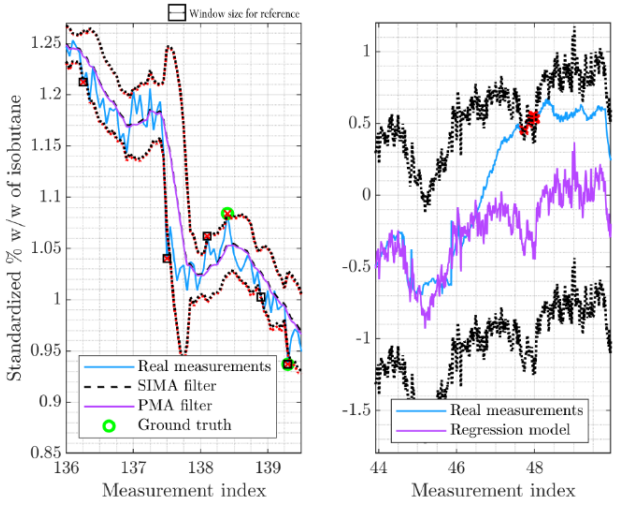

Figure 1: Comparison of the outlier detection (left) and anomaly detection (right).

We analyze the performance of each method using confusion matrices derived from the predefined GT. True positives (TP) and false positives (FP) represent correctly and incorrectly classified non-anomalous measurements, respectively. True negatives (TN) and false negatives (FN) denote correctly and incorrectly identified outliers, respectively. Results are shown in Table 1. The SM approach yielded poor results, detecting no outliers (TN = 0), which was expected given the varying isobutane concentration range across different operation points. TM performed better, correctly identifying 97% of normal operation points (TP = 15,431) yet struggled with over 80% of outliers (TN = 61). SIMA detected 40% of overall outliers (TN = 122), with no foresight into upcoming measurements. PMA improved outlier detection compared to SIMA, reducing false predictions (FP from 1,048 to 1,011; FN from 184 to 181) and enhancing overall accuracy. The TP and TN rates improved as well (TP from 14,553 to 14,590; TN from 122 to 125). These adjustments yielded the best distribution of correctly classified data and reflect the improved use of latent information.

We assessed the regression model (LASSO) using the ±2 × RMSE metric, and identified 15 outliers (TN), showcasing its unique perspective on anomalies beyond the expected range of the dependent variable. These outliers (TN+FN) signify slow, gradual drifts

requiring calibration. We illustrate the performance of the SIMA and PMA in Fig. 2 (left graph) using a selected period from testing measurements over multiple days. The fit in the latter approach enhances the mean value to better capture changes in the plant. The comparison reveals differences in outlier detection (red crosses vs. black squares), with the PMA offering additional information for more accurate predictions. While the approach may lead to a higher count of FN (identifying normal instances as outliers) concerning the GT (green circles), it simultaneously emphasizes accurately predicted instances. This method holds the potential to identify outliers that escaped detection by the SYMA, which lacked information about other independent variables $\boldsymbol{X}$.

5. Conclusions

We studied outlier detection in the process variables. We employ moving-horizon filters and integrate regression-based prediction into our approach. Notably, this method successfully identifies 40% of the outliers while reducing false detections. Conversely, regression models, despite exhibiting lower outlier detection efficacy, provide a means to detect long-term anomalies. The overall fit of the models to the dependent variable, assessed through the computed RMSE criterion, falls within the approved process tolerance. In our future research, we aim to investigate the efficacy of regression approaches for detecting anomalous measurements, mainly slow gradual shifts. Our focus will extend to exploring non-linear transformations and dynamic sensor characteristics and to the development of a comprehensive and robust anomaly detection framework.

References

D. O. Afanasyev, E. A. Fedorova, 2019. On the impact of outlier filtering on the electricity price forecasting accuracy. Applied Energy 236, 196–210.

A. Blázquez-García, A. Conde, U. Mori, J. A. Lozano, 2021. A review on outlier/anomaly detection in time series data. ACM Comput. Surv. 54 (3).

K. M. Carter, W. W. Streilein, 2012. Probabilistic reasoning for streaming anomaly detection. In: 2012 IEEE Statistical Signal Processing Workshop (SSP). pp. 377–380.

A. V. Oppenheim, 1999. Discrete-time signal processing. Pearson Education India.

IBM Corporation, 2021. SPSS Modeler 18.1.0, Outliers. URL https://www.ibm.com/docs/en/s-pss-modeler/18.1.0?topic=series-outliers

Z. Jin, J. Zhao, L. Ding, S. Chakrabarti, E. Gryazina, V. Terzija, 2022. Power system anomaly detection using innovation reduction properties of iterated extended Kalman filter. International Journal of Electrical Power & Energy Systems 136, 107613.

A. V. Oppenheim, 1999. Discrete-time signal processing. Pearson Education India.

F. Santosa, W. W. Symes, 1986. Linear inversion of band-limited reflection seismograms. SIAM Journal on Scientific and Statistical Computing 7 (4), 1307–1330.

T. Tziolas, K. Papageorgiou, T. Theodosiou, E. Papageorgiou, T. Mastos, A. Papadopoulos, 2022. Autoencoders for anomaly detection in an industrial multivariate time series dataset. Engineering Proceedings 18 (1).

Y. S. Yoon, W. Jeong, J. Kim, M. Seok, J. Park, J. Bae, K. Lee, J. H. Lee, 2022. Development of inferential sensor and real-time optimizer for a vacuum distillation unit by recurrent neural network modeling of time series data. Computers & Chemical Engineering 168, 108039

S. W. Roberts, Aug. 1959. Control chart tests based on geometric moving averages. Technometrics 1 (3), 239–250.

K. Pearson, 1901. Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Phil. Magazine and Journal of Science 2 (11), 559–572.

P. Rousseeuw, K. Driessen, 08 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.

R. Fáber, K. Ľubušký, R. Paulen, 2023. Machine Learning-based Classification of Online Industrial Datasets. 2023 24th Int. Conf. on Process Control, 132–137.