Time series outlier detection and imputation

Hermine N. Akouemo and Richard J. Povinelli
Department of Electrical and Computer Engineering
Marquette University
Milwaukee, Wisconsin 53233

***Abstract*—This paper proposed the combination of two statistical techniques for the detection and imputation of outliers in time series data. An autoregressive integrated moving average with exogenous inputs (ARIMAX) model is used to extract the characteristics of the time series and to find the residuals. The outliers are detected by performing hypothesis testing on the extrema of the residuals and the anomalous data are imputed using another ARIMAX model. The process is performed in an iterative way because at the beginning the process, the residuals are contaminated by the anomalies and therefore, the ARIMAX model needs to be re-learned on "cleaner" data at every step. We test the algorithm using both synthetic and real data sets and we present the analysis and comments on those results.**

***Index Terms*—outlier, hypothesis testing, time series, ARIMAX, imputation.**

I. Introduction

In the energy domain, good forecasting results are achieved by learning the models on data sets that accurately represent the particularities of the problem. The time series data often contain anomalies which can be due to various causes ranging from human error (e.g. mistyping) to system error (e.g. erroneous measurement). The data is cleaned for the purpose of being used to train a forecasting model. Training a forecasting model on anomalous data will result in erroneous estimates. Hence, outlier detection is one of the most important and critical problems in the forecasting domain. A data point is classified anomalous depending on the context, and this is the reason why we limit our analysis to the energy domain.

This paper presents a novel approach that combines ARIMAX model and hypothesis testing to find and impute outliers in time series data sets. The contribution of the proposed algorithm is its ability to extract the time-series characteristic of the data set and focus on the residuals for outlier detection. The residuals form a distribution in which the algorithm is able to distinguish between data points in the tails of a distribution and outliers by taking into account the statistics of the residuals and the number of samples in the data set.

The next section of the paper presents some background of the outlier detection problem. Section 3 presents the ARIMAX model, the hypothesis-driven algorithm and show how both techniques are combined to form the time series detection and imputation algorithm. Section 4 presents the results and analysis.

This work was supported by the GasDay Laboratory at Marquette University.

II. Background

Outliers, in this paper, refer to data points that are considerably dissimilar to the remaining points in the data set [1]. In the energy domain, "clean" data is required to train and develop accurate models for forecasting. It has been shown that there are two types of outliers in time series: additive outliers (AO) that are isolated events and innovative outliers (IO) that are errors propagated through time in the system [2]. The author in [2] also showed how linear models can be used for the detection of AO and IO in stationary time series. The impact of anomalous data in the parameter estimation of ARIMA models have been studied by [3]–[7]. The modification of time series structure (variance changes and level shifts) by additive and innovative outliers was studied by [8]. Outlier detection using clustering techniques by considering a multi-dimensional space composed of different inputs was studied by [9]–[11]. In [12], the clustering techniques are applied in both the time and delay spaces to detect anomalies. The idea is that the delay space shows characteristics of the anomalies that are invisible or not easily extractable in the time domain. Outlier detection using neural networks, have also been studied by [13]. Another technique that uses a Kalman filter to detect and "clean" outliers was proposed by [14].

Our proposed technique uses an ARIMAX model to estimate the parameters of the time series. The parameters are skewed because of the presence of outliers in the data set as demonstrated by [3]–[5], [7]. The residuals will portray the largest anomalies only as a starting point. The outliers are then detected using hypothesis testing on the residuals, but we will not classify them as additive or innovative. Hypothesis testing efficiently avoids false positive by considering the residuals as a set of data points drawn from the same distribution and by considering the number of samples. The hypothesis testing identifies only data points that are dissimilar or inconsistent with the time series data set. As the outliers are removed, the parameter estimation will yield more valid results. The next section of this paper will present both techniques and show the time series outlier detection and imputation algorithm.

III. Techniques

This section presents the techniques used for outlier detection and imputation in this paper. The ARIMAX model and the hypothesis testing are both statistical models. The next sections give an overview of the techniques.

A. ARIMAX model

An ARMAX or autoregressive-moving average model with exogenous inputs is a class of models that describes a sta-

978-1-4799-6415-4/14/$31.00 ©2014 IEEE

tionary and invertible time series process [15]. An exogenous input is one that comes from an external system. The output of an ARMAX model is written as a linear combination of a sequence of uncorrelated random variables:

$$\mathrm{y}_t = c + \sum_{i=1}^{p} \phi_i \mathrm{y}_{t-i} + \sum_{i=0}^{q} \theta_i \epsilon_{t-i} + \sum_{i=0}^{n_x} \eta_i b_{t-i}. \quad (1)$$

where, ϵ_t is white noise. ϕ_i, θ_i and η_i are respectively the coefficients of the autoregressive, moving average and exogenous inputs. Also, p, q and n_x are respectively the orders of the autoregressive, moving average and exogenous inputs. The ARMAX model is therefore noted ARMAX(p, q, n_x). An ARMAX($p, 0, 0$) is simply an autoregressive AR(p) model.

In the case where the random process is not exactly white noise, the difference series $\bigtriangledown^n \mathrm{y}_t$ is an ARMA(p, q) process, with $\bigtriangledown \mathrm{y}_t$ = y_t - y_{t-1} [15]. The non-stationary model is referred to as the autoregressive-integrated-moving average (ARIMA) model and has a differentiation degree D as additional input. An ARIMA model with D = 0 is simply an ARMA model. In conclusion, an ARIMAX model has four parameters, an autoregressive order p, a moving-average order q, an exogenous inputs number of terms n_x and a differentiation degree D and it is noted ARIMAX(p, D, q, n_x). Given model orders and a particular time series signal, the model estimation process estimates the coefficients c, ϕ_i, θ_i, and η_i.

B. Hypothesis-driven outlier detection algorithm

Hypothesis testing is used in this paper to statistically test the likelihood of the residuals. A statistical hypothesis is a statement about the values of the parameters of a probability distribution [16]. For this paper, it is a statement about the extrema of the residuals compared to the parameters of the probability distribution. The null hypothesis (H_0) is that the extremum is not an outlier. While, the alternative hypotheses (H_{alt}) is that the extremum is an outlier. The null is rejected in favor of the alternative with a level of significance α, where α is the probability of committing a type I. A type I occurs if the null hypothesis is rejected when true and type II occurs if the null hypothesis is not rejected when it is false. We choose the probability of committing a type I error with a probability α = 0.1.

Let us define the experiment Y = {Classifying an extremum}. The possible outcomes of the experiment Y are "outlier" or "not outlier". If the probability of "outlier" in the experiment Y is p, the probability of "not outlier" is $(1 - p)$. The number of samples in the data set is n. Each classification of an extremum is an independent experiment. Therefore, the experiment Y is a Bernoulli trial. The problem is reduced to finding the number of Y Bernoulli trials needed to get an "outlier" in at least n trials and supported by the set of samples n. It corresponds to the cumulative distribution function of a geometric distribution. The number of Bernoulli trials should be less than the level of significance α for the algorithm to have exhausted all possibilities. By taking into account the number of samples and the probability of the data points, the hypothesis-driven outlier detection algorithm also sets an effective bound on how many potential outliers there might be in a data set.

Algorithm 1 HYPOTHESIS-OUTLIER-DETECTION

Require : X, α, assumed distribution Dist(X, β).

$X_{min} \leftarrow X \setminus \{\underline{x}\}$
$X_{max} \leftarrow X \setminus \{\overline{x}\}$
Dist$_{min}$ $\leftarrow$ estimate β's of X_{min}
Dist$_{max}$ $\leftarrow$ estimate β's of X_{max}
$p_{min} \leftarrow$ cdf(Dist$_{min}$, x)
$p_{max} \leftarrow$ cdf(Dist$_{max}$, x)
$g_{min} \leftarrow 1 - (1 - p_{min})^n$
$g_{max} \leftarrow 1 - (1 - p_{max})^n$
if $(g_{max} < \alpha) \vee (g_{min} < \alpha)$ **then**
  if $(g_{min} < g_{max})$ **then**
    outlierIndex $\leftarrow$ index($\underline{x}$)
  else
    outlierIndex $\leftarrow$ index($\overline{x}$)
  end if
else
  outlierIndex $\leftarrow$ nil
end if
return outlierIndex

The probability p depends on the data point values and the underlying distribution from which the point was taken. The advantage of our hypothesis-driven outlier detection algorithm is that it takes into account the number of samples and the assumed distribution from which the samples are drawn. Therefore, the technique considers the residuals as an ensemble of data points drawn from a distribution and focuses on the anomalous ones. Most importantly, the algorithm detects points that are most unlikely to be drawn from the assumed underlying distribution. The next section presents examples using both synthetic and real data sets. The synthetic data sets allow us to test the algorithms. The real data set shows the performance and feasibility of the algorithm, and the results are interpreted using energy domain knowledge.

C. Time series outlier detection and imputation algorithm

A time series data is a set of statistics, collected at regular intervals [15]. A time series can be decomposed into four elements: trend, seasonal effects, cycles and residuals. Therefore the idea behind our reasoning is that the ARIMAX model, used to estimate the parameters of the model, will extract the trend, seasonal effects and cycles characteristics of the data set. The residuals found after estimation with the ARIMAX model, form a distribution of points where outliers are detected using hypothesis testing. The ARIMAX model is re-trained on cleaner data and the new model is used to forecast the outliers.

The algorithm supposes that the order of the ARIMAX model is known *a priori*. The parameter estimates are erroneous in the beginning of the process because the data set contains anomalies, but the trend is extracted such that the residuals can depict those anomalies. After an outlier is removed, the parameter of the ARIMAX model are recalculated

with the outlier replaced by a naive impute of the mean of past and previous timesteps. This step ensures that the point is not a false positive and also removes some contamination from the imputation model. The model parameters estimation is improved after each outlier is removed. The model trained on cleaner data is used to forecast an estimate value for the data point and the estimates are replaced in the time series signal and the process continue until no more outliers are identified. The time series outlier detection and imputation algorithm is presented here.

Algorithm 2 TS-OUTLIER-DETECTION-IMPUTATION

Require : time series S, α, ARIMA model order (p, D, q), exogenous inputs b

potentialOutliers ← true
I ← ∅
while (potentialOutliers) **do**
  m ← ARIMAX(S, b)
  r ← CALCULATE-RESIDUALS(m, S, b)
  i ← HYPOTHESIS-OUTLIER-DETECTION(r, α)
  if i == nil **then**
    potentialOutliers ← false
  else
    S[i] ← $0.5(S[i-1] + S[i+1])$
    m ← ARIMAX(S, b)
    I ← I ∪ {i}
    for j = 1 : I.length **do**
      S[I[j]] ← FORECAST(m, S, I[j])
    end for
  end if
end while
return S, I

The algorithm illustrates each step of the outlier detection and imputation. It also shows the iterative nature of the process. The next section of this paper will present the data sets that we used as examples and the results obtained.

IV. RESULTS

This section gives a description of the data sets before presenting the results obtained by the algorithm using those data sets.

A. Data

We will present two data sets, a synthetic data set and a real data set. The synthetic data set is generated using an ARIMAX(3, 1, 1, 3) and with exogenous inputs drawn from a Gaussian distribution $N(0, 0.01)$. Four outliers, at the timesteps {100, 366, 394 and 395} are introduced in the data set by two different ARIMAX processes. The data set is first perturbed with outliers' samples introduced by an ARIMAX(2, 0, 1, 3) process: {(100, -4.4), (366, -2.6), (394, 2.1), (395, 2.2)}and the graph is presented in Figure 1. Then, the same synthetic data set is perturbed with outliers' samples drawn from an ARIMAX(4, 0, 1, 3) process: {(100, 7.5), (366, 10), (394, 4), (395, 5)} and the data set is presented in Figure 2.

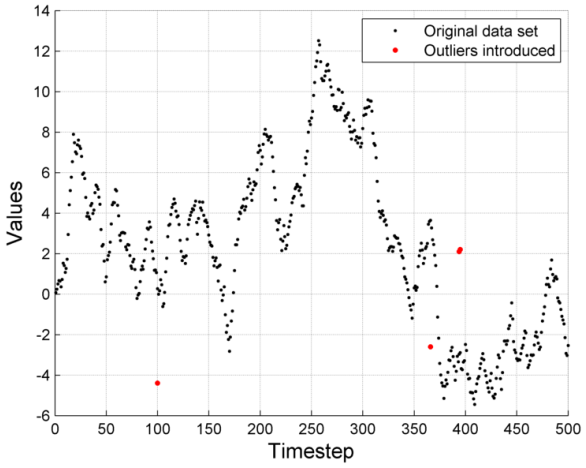

Figure 1: Synthetic data set simulated using an ARIMAX(3, 1, 1, 3) with the four outliers drawn from ARIMAX(2, 0, 1, 3) depicted on the graph

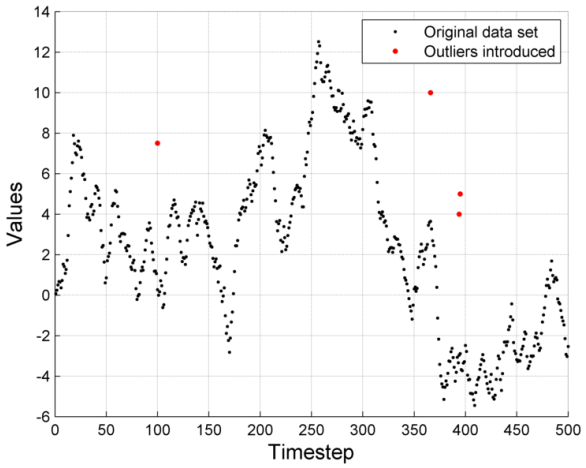

Figure 2: Synthetic data set simulated using an ARIMAX(3, 1, 1, 3) with the four outliers drawn from ARIMAX(4, 0, 1, 3) depicted on the graph

Figure 3 shows 3620 observations of aggregated daily electric load for an operating area in the United States. For this data set, temperature is used as exogenous input, the data is normalized for confidentiality purposes. We test the synthetic data sets to demonstrate that the algorithm identifies the outliers. For the real data set, we determine whether or not the real data sets contains outliers. The results for all data sets are presented and commented below.

B. Results

The results obtained using the three data sets are presented in Figure 4, Figure 5 and Figure 6. Figure 4 and Figure 5 depicts the original synthetic data set, the outliers and the imputed values of the outliers. The outliers are introduced at the same positions to make a comparison between imputed values for both cases (see Table I). For the synthetic data sets, the outliers included are all found by the algorithm. After the last outlier is found in both cases, the hypothesis-driven outlier detection algorithm returns the value NIL, which is the

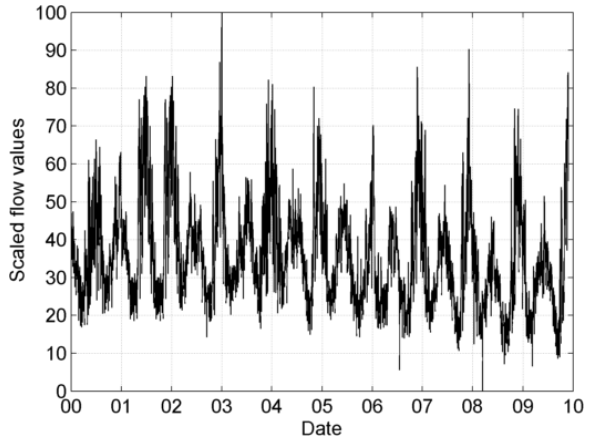

Figure 3: Daily aggregated electric load for an operating area in the United States

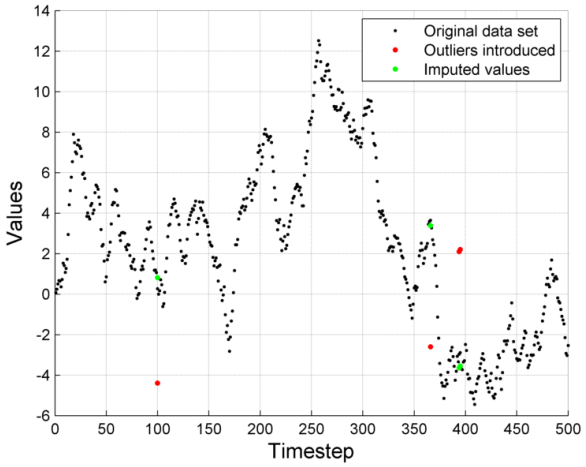

Figure 4: Algorithm results on the synthetic data set perturbed with four outliers drawn from an ARIMAX(2, 0, 1, 3) model

	Position	Actuals	Outliers	Imputed	lerrorl
Figure 4	100	0.26	-4.4	0.82	0.56
	366	3.64	-2.6	3.38	0.26
	394	-2.89	2.1	-3.64	0.75
	395	-2.96	2.2	-3.55	0.59
Figure 5	100	0.26	7.5	0.83	0.57
	366	3.64	10	3.39	0.25
	394	-2.89	4	-3.53	0.64
	395	-2.96	5	-3.65	0.69

Table I: Synthetic data set results

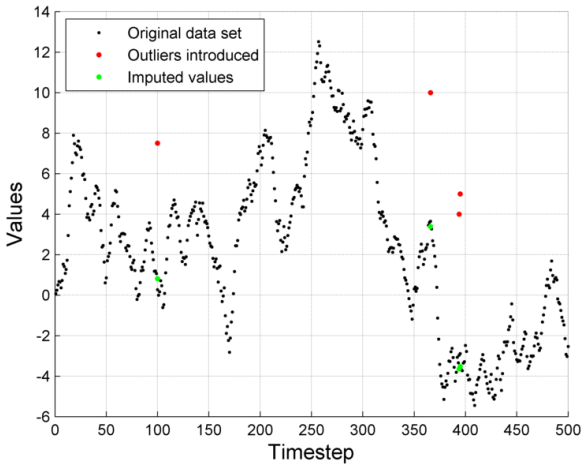

Figure 5: Algorithm results on the synthetic data set perturbed with four outliers drawn from an ARIMAX(4, 0, 1, 3) model

sentinel for the completion of the algorithm. Table I presents the results for the synthetic data sets. The table gives the position, the actual and outliers values and the imputed values found by the algorithm; it also gives a comparison (in absolute value) between actuals and imputed value because the model parameters are modified in presence of outliers.

The model used to simulate the data set is

$$\mathrm{y}_t = 0.5\mathrm{y}_{t-1} - 0.3\mathrm{y}_{t-2} + 0.2\mathrm{y}_{t-3} + 0.2\epsilon_t + 1.5m_t + 2.6m_{t-1} - 0.3m_{t-2}. \quad (2)$$

The estimated model (see Figure 4) after all outliers are removed and imputed in the case of outliers introduced by an ARIMAX(2, 0, 1, 3) process is

$$\mathrm{y}_t = 0.55\mathrm{y}_{t-1} - 0.30\mathrm{y}_{t-2} + 0.26\mathrm{y}_{t-3} + 0.10\epsilon_t + 1.61m_t + 2.5m_{t-1} - 0.38m_{t-2}. \quad (3)$$

The estimated model (see Figure 5) after all outliers are removed and imputed in the case of outliers introduced by an ARIMAX(4, 0, 1, 3) process is

$$\mathrm{y}_t = 0.55\mathrm{y}_{t-1} - 0.3\mathrm{y}_{t-2} + 0.26\mathrm{y}_{t-3} + 0.09\epsilon_t + 1.59m_t + 2.54m_{t-1} - 0.39m_{t-2}. \quad (4)$$

In general, the absolute error is low (compare to the maximum value in the data set which is 12.5). When the data point to forecast is small, the imputed values are sensitive to the model error. The results show that in general, outliers impact the estimation of the parameters of a time series data set and that removing those provide improvement to the forecasting model.

For the electric consumption data set (see Figure 6, the time series outlier detection and imputation algorithm found 11 outliers using an ARIMAX(5, 1, 3, 3) model. The residuals are assumed to be normally distributed in the outlier detection part of the algorithm. The red dots in Figure 6 depict the outliers found while the green dots are their corresponding imputed values. We should note that because the data is scaled, the point correspond to a zero is actually not zero, it is just the minimum point in the data set. Looking at Figure 6, we can notice that year 01 was much warmer than year 00, which is why the first two outliers points are expected to be much higher values than what their original values. Also, outliers in year 06, 07, 08 and 09 are correlated with temperature in those years. The temperature of the minimum point in the data set is 36°F on average for that particular day, so the electric demand value is expected to be around the baseload. The major constraint of our approach is that the assumed orders of the ARIMAX process must be close to the true orders of

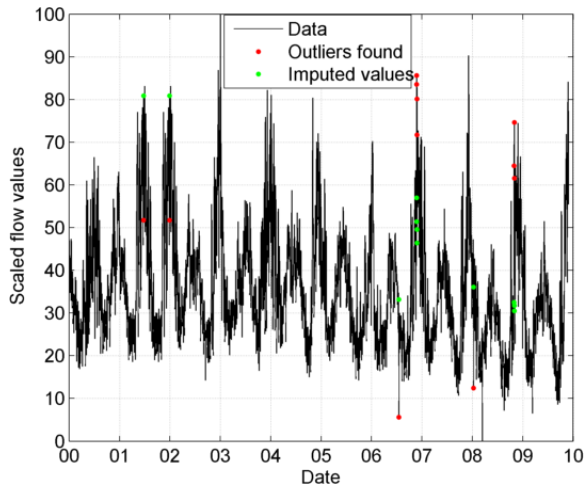

Figure 6: Algorithm result on the daily aggregated electric load data set using an ARIMAX (5, 1, 3, 3) and temperature as exogenous inputs

the system. The order depend on the number of observations and type and number of exogenous inputs. Various techniques, such as the Bayesian Information Criteria (BIC) [17] or the Box-Jenkins method [18], have been developed to estimate the orders of a time series model.

V. Conclusion

Many techniques have been developed for outlier detection. This paper presents a novel approach for outlier detection and imputation based on statistical methods. The algorithm ensures that corrupted parameters are not used for imputation by doing a naive imputation that consists of the average of the neighboring time samples before re-learning the model. The re-learned model is then used for imputation. This extra step decontaminates the model from the outlier previously found. The main contribution of this technique is the development of outlier detection algorithm based on hypothesis testing and using the number of samples in the data set, and the combination of ARIMAX and hypothesis testing to efficiently detect outliers in time series data.

References

[1] D. M. Hawkins, *Identification of outliers*. England, United Kingdom: Chapman and Hall, 1980.

[2] K. Choy, "Outlier detection for stationary time series," *Journal of Statistical Planning and Inference*, vol. 99, pp. 111–127, 2001.

[3] I. Chang, G. C. Tiao, and C. Chen, "Estimation of time series parameters in the presence of outliers," *Journal of Technometrics*, vol. 30, pp. 193–204, 1988.

[4] D. R. Martin and V. J. Yohai, "Influence functionals for time series," *The Annals of Statistics*, vol. 14, pp. 781–855, 1986.

[5] A. M. Bianco, M. García Ben, E. J. Martínez, and V. J. Yohai, "Outlier detection in regression models with arima errors using robust estimates," *Journal of Forecasting*, vol. 20, pp. 565–579, 2001.

[6] A. Zaharim, R. Rajali, R. M. Atok, I. Mohamed, and K. Jafar, "A simulation study of additive outlier in arma(1,1) model," *International Journal of Mathematical Models and Methods in Applied Science*, vol. 3, 2009.

[7] L. Denby and D. R. Martin, "Robust estimation of the first-order autoregressive parameter," *Journal of the American Statistical Association*, vol. 74, pp. 140–146, 1979.

[8] R. S. Tsay, "Outliers, level shifts, and variance changes in time series," *Journal of Forecasting*, vol. 7, pp. 1–20, 1988.

[9] A. Koufakou and M. Georgiopoulos, "A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes," *Journal of Data Mining and Knowledge Discovery*, vol. 20, pp. 259–289, 2010.

[10] F. Angiulli and C. Pizzuti, "Outlier mining in large high-dimensional data sets," *IEEE transactions on Knowledge and Data engineering*, vol. 17, pp. 203–215, 2005.

[11] E. N. Knorr and R. T. Ng, "Algorithms for mining distance-based outliers in large datasets," in *International Conference on Very Large DataBases*, 1998.

[12] A. R. Weekley, R. K. Goodrich, and L. B. Cornman, "An algorithm for classification and outlier detection of time-series data," *Journal of Atmospheric and Oceanic Technology*, vol. 27, pp. 94–107, 2010.

[13] S. Hawkins, H. He, G. Williams, and R. Baxter, "Outlier detection using replicator neural networks," *Data Warehousing and Knowledge Discovery - Lecture Notes in Computer Science*, vol. 2454, pp. 170–180, 2002.

[14] H. Liu, S. Shah, and W. Jiang, "On-line outlier detection and data cleaning," *Journal of Computer and Chemical Engineering*, vol. 28, pp. 1635–1647, 2004.

[15] W. W. S. Wei, *Time series analysis: univariate and multivariate methods*, 2nd ed. Boston, MA: Pearson Addison Wesley, 2006.

[16] D. C. Mongtgomery, *Introduction to statistical quality control*, 5th ed. Hoboken, NJ: John Wiley & Sons, 2005.

[17] G. E. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, p. 461–464, 1978.

[18] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time series analysis: Forecasting and control*, 4th ed. Hoboken, NJ: John Wiley & Sons, 2008.

[19] A. Papoulis and U. S. Pillai, *Probability, random variables and stochastic processes*, 4th ed. Boston, MA: McGraw-Hill Europe, 2002.

[20] G. Buzzi-Ferraris and F. Manenti, "Outlier detection in large data sets," *Journal of Computers and Chemical Engineering*, vol. 35, pp. 388–390, 2010.

[21] R. R. Jones, R. S. Vaught, and M. Weinrott, "Time-series analysis in operant research," *Journal of Applied Behavior Analysis*, vol. 10, pp. 151–166, 1977.

[22] I. A. McLeod and Y. Zhang, "Faster arma maximum likelihood estimation," *Journal of Computational Statistics and Data Analysis*, vol. 52, pp. 2166–2176, 2008.

[23] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, pp. 85–126, 2004.

[24] C. Fauconnier and G. Haesbroeck, "Outliers detection with the minimum covariance determinant estimator in practice," *Journal of Statistical Methodology*, vol. 6, pp. 363–379, 2009.