Autonomous Time Series Data Processing on Historical and Real-Time Settings

Ricardo Sousa
Instituto Superior Técnico
Universidade de Lisboa
Lisboa, Portugal
ricardo.filipe.sousa@tecnico.ulisboa.pt

***Abstract*—Heterogeneous sensor networks, including water distribution systems and traffic monitoring systems, produce abundant time series data for monitoring network dynamics and detecting events of interest. Nevertheless, errors and failures in the calibration, data storage or acquisition can occur on some of the sensors installed in those systems, producing missing and/or anomalous values. This work proposes a computational system, referred as AutoMTS, for the fully autonomous cleaning of multivariate time series data using strict quality criteria assessed against ground truth extracted from the targeted series data. The proposed methodology is parameter-free as it relies on robust principles for the assessment, hyperparameterization and selection of methods. AutoMTS coherently supports an extensive set state-of-the-art methods for (multivariate) time series imputation and outlier detection-and-treatment, considering both point and segment/serial occurrences. A comprehensive evaluation of AutoMTS is accomplished using heterogeneous sensors from two water distribution systems with varying sampling rates, water consumption patterns, and inconsistencies. Results confirm the relevance of the proposed AutoMTS system.**

***Index Terms*—parameter-free learning, multivariate time series, missing values imputation, outlier detection, heterogeneous sensor networks, real-time data**

I. Introduction

The placement of heterogeneous sensors within complex systems – whether physiological, mechanical, digital, geophysical, environmental or urban – offers the possibility to acquire comprehensive views of their behavior along time. Sensorized systems produce abundant time series data, used for monitoring purposes or the detection of events of interest. However, the placed sensors are susceptible to failures and errors associated with sensor calibration and data acquisition-transmission-storage Gill et al. [1], producing time series data with missing and anomalous values. In this context, time series data are generally subjected to initial processing stages for leveraging their quality for the subsequent mining stages.

Processing time series data produced by networks of heterogeneous sensors is, nevertheless, a laborious process due to four major reasons. First, the selection and parameterization of the processing methods is highly dependent on the regularities of the target series data and challenged by the wide diversity of approaches currently available. Second, the profile of errors can be diversified, each leading to different processing choices. In this context, the type and amount of anomalies and missing values can largely affect decisions. Third, different types of sensors – such as water flow, pressure and water quality sensors in water distribution systems – may benefit from dissimilar processing methods. In fact, sensors of the same type but with singular calibrations, sampling rates, or positioning within the monitored system can as well benefit from different choices. Fourth and finally, different systems equipped with identical sensors do not necessarily benefit from the same processing options. Consider water distribution network (WDN) systems, water consumption patterns can highly vary between WDNs or along time, impacting decisions. Also, different WDNs may be susceptible to unique externalities, affecting the profile of observed errors.

In addition, time series data processing generally yields sub-optimal results. First, cross-variable relationships in multivariate time series data are commonly disregarded. For instance, flow and pressure sensors in WDNs are generally correlated, and thus co-located or nearby sensors can guide the treatment of low-quality series data. Second and understandably, optimal decisions are challenged by the wide diversity of available processing approaches, multiplicity of sensors, and profile of errors observed per sensor.

This work proposes a methodology for the fully autonomous cleaning of multivariate time series, for historical and real-time data settings, that is able to address the introduced challenges. The proposed methodology, referred as AutoMTS (**Auto**nomous **M**ultivariate **T**ime **S**eries data processing), offers three major contributions. First, AutoMTS provides strict guarantees of optimality as it places robust processing decisions against ground truth extracted from the targeted series data.

Second, AutoMTS provides a comprehensive coverage of available processing options, currently providing state-of-the-art methods for missing imputation, outlier detection and gross-error removal from time series data, some of them able to consider cross-variable dependencies in the presence of multivariate time series data. Also, we further guarantee the presence of methods able to deal with both point and segment/serial missing and outlier values.

Third, AutoMTS is parameter-free as it relies on robust principles to assess, hyperparameterize and select state-of-the-art processing methods.

AutoMTS is provided as both a graphical and programmatic

tool satisfying strict usability criteria.

The manuscript is structured as follows. Section II provides essential background and surveys recent contributions on time series data processing. Section III described the AutoMTS approach for either historical data and real-time data settings. Section IV comprehensively assesses the adequacy of AutoMTS using two real-world heterogeneous networks as study cases. Finally, concluding remarks and major implications are synthesized.

II. Background and related Work

This section offers a structured view on how to process inconsistencies in (multivariate) time series, providing essential *background*, surveying *recent contributions*, and describing the pre-processing *methods* implemented in AutoMTS.

Time series data processing. Signals produced by sensors are generally represented as *time series*, an ordered set of observations $\vec{x}_{1..T} = (\vec{x}_1, ..., \vec{x}_T)$, each $\vec{x}_t$ being recorded at a specific time point t. Time series can be *univariate*, $\vec{x}_t \in \mathbb{R}$, or *multivariate*, $\vec{x}_t \in \mathbb{R}^m$, where $m > 1$ is the order (number of variables).

Errors associated with the calibration, measurement, storage, logger communication and synchronization of sensors are associated with inconsistencies on the produced time series. As a result different types of errors can be observed, including: 1) anomalous values, 2) missing values; 3) duplicate values; 4) atypical values or gross errors (impossibilities in a given domain); and 5) incorrectly timestamped observations (arbitrarily-high sampling delays).

Low-quality data can be rectified. The task of *pre-processing time series* is the process of leveraging quality data to facilitate the subsequent extraction of useful information from the time series. In this context, cleaning the identified inconsistencies is an important step, and the one targeted in this work.

Time series can be decomposed into *trend*, *seasonal*, *cyclical*, and *irregular components* using additive or multiplicative models Jain [2]. Processing can take place on the original series or separately on each component. Classical approaches for time series analysis generally rely on statistical principles, including *auto-regression*, *differencing* and *exponential smoothing* operations to either detect deviations from expectations as well as to impute missing values Wei [3].

Time series typically have an internal structure with domain-specific meanings. In this context, normalization, resampling, piecewise aggregate approximation, symbolic aggregate approximation, and transformations (including Fourier, Wavelet and other forms of window-based feature extraction) can support the analysis of the internal structure of time series. However, finding suitable representations is highly dependent on the subsequent mining ends and therefore is not considered part of the processing pipeline proposed in our work.

Missing value imputation. Missing observations, commonly referred as missing values, can be characterized by the underlying stochastic processes that describe their occurrence: i) missing completely at random (MCAR) where there is no distribution characterizing their occurrence, generally caused by punctual problems on data transmission-storage-acquisition; ii) missing at random (MAR) where missings are independent of the value of the observation but dependent on the other non-missing observations (e.g. sensor malfunction under high temperatures); and iii) not missing at random (NMAR) where missings essentially depend on the value of the observation (e.g. sensors failing measuring high pressure). Complementary, missing values can be described by their *type* – whether point, sequential or mixed similarly to outliers – and *amount* from a given period.

There are three typical choices to deal with missing values: i) force removal, leading to gaps on the time series to be handled along the subsequent time series processing steps; ii) replace them with a dedicated value or symbol; and iii) estimate their values using imputation principles. Missing removal can be listwise (indiscriminate missing deletion) or pairwise (controlled deletion in accordance with the amount) Osman et al. [4]. Missing imputation can either produce hot-deck estimates from similar/nearby observations or from matched segments of the time series; or cold-deck estimates from external time series datasets Osman et al. [4].

Last observation carried forward (LOCF) and next observation carried backward (NOCB) are simplistic methods based on the closest available observation. Linear interpolation linearly combines last and next observations. Usually, the seasonal component is removed at the beginning and included after linear interpolation is done. Moving average (MA) can include further observations to estimate the missing value, $\hat{\vec{x}}_t = \frac{1}{m}\sum_{j=-k}^{k} \vec{x}_{t+j}$ where $[t-k, t+k]$ is a centered window of *2k+1* length (also termed order). When the sequential values are all missing observations, the window size can dynamically expand until two non-missing values occur. In this context, linear interpolation is a moving average or order 2. Average (median) imputation corresponds to a moving average (median) with unbounded order, imputing the average (median) of all non-missing occurrences. The expectation maximization algorithm (EM) has been also suggested for estimating missing observations within multivariate time series data, although in its original form disregards time dependencies. Amelia combines the EM method with bootstrapping to impute missing values in time series data using principles from multiple imputation. Classical approaches for time series modeling, including SARIMA and Holt-Winters Wei [3], are also viable imputation candidates when time series have well-established regularities.

k-nearest neighbors (kNN) can be applied to impute both point and sequential missings from (multivariate) time series. To this end, time series are subjected to segmentation, and the value estimates inferred from the closest neighbor subsequences. Particular attention should be paid to its parameterization, as kNN performance highly depends on the selected distance (e.g. ability to tolerate shift and scale misalignments on the time and amplitude axes) and number of neighbors. In the presence of multivariate time series data, MissForests Stekhoven et al. [5] uses principles from random forest approaches to deal with mixed-variables (relevant when

dealing with heterogeneous sensors) in accordance with the frequency of missing values (chained principle). Despite its role, it neglects time dependencies between observation. The time-extended version of multivariate imputation by chained equations (MICE) Buuren et al. [6] is able to addresses such drawback while still accounting for cross-variable dependencies.

Osman et al. [4] proposed an ensemble approach that selects between classical imputation techniques (such as moving average) and modern alternatives in accordance with the type (MAR or MCAR) and amount of missings. In addition to some of the surveyed methods, modern imputation techniques further include reconstruction methods based on principal component analysis Ilin et al. [7] and machine learning techniques such as Gaussian process regression, tensor-based methods Garg et al. [8], and neural networks, specially auto-associative neural networks Luo et al. [9].

Moritz et al. [10] extensively compares multiple-imputation approaches by deleting observations from time series with varying trend and seasonal characteristics. Multiple-imputation approaches rely on multiple estimates to reduce biases. For instance, Aggregated values Zeileis et al. [11] is an estimator from mean estimates collected at multiple temporal granularities (overall, yearly, monthly and daily mean). Seasonal Kalman filters and model-based approaches have been also applied within multiple-imputation settings Kowarik et al. [12] and Moritz et al. [10].

Imputation methods have been also proposed in the context of specific domains. In water-energy-gas distribution systems, the well-recognized Quevedo method Quevedo et al. [13] estimates missings from observations collected at similar periods from previous days, weeks, months and years. Barrela et al. [14] further proposed a estimator that combines both forecast and backcast missing observations values generated by TBATS and ARIMA models, accommodating multiple seasonality.

For real-time data, Fan et al. [15] proposed a model called On-line Missing Value Imputation (OLMVI), which can analyze observations of information and impute the missing observations before they are added into a database. Using a correlation matrix (updated as new data enter the system) and the attributes, imputation candidates are computed by assigning an imputation score to each of them. The candidate with the highest score is the one used as imputed value. Also Osman et al. [4] present approaches for missing values imputation in real-time data from a WDN.

Time series outlier detection. *Outliers* are observations significantly deviating from expectations as to arouse suspicion of being generated by a different mechanism Hawkins [16]. Outliers can occur in point or serial forms. *Point outliers* (also referred as punctual or singular outliers) can be detected against the whole series (*global outliers*) or against observations that occur on nearby time points or share the same context (local/*contextual outliers*). *Sequential outliers* (also referred as segment or serial outliers) are anomalous subsequences of contiguous observations. Outliers can be further characterized in accordance with their causation and impact Chan [17]: additive outliers affect the time series for a single time period; level shift outliers have preserved/continuous effects; temporary change outliers show an exponential decaying over time; and innovational outliers affect the nearest subsequent observations. *Outlier analysis* generally comprises anomaly *scoring*, *detection* and *treatment* steps. Treatment either denotes the removal (planting missing values) or re-estimation of outlier values. Approaches for outlier analysis are generally categorized according to *distribution*-based, *depth*-based, *distance*-based, *density*-based and *clustering*-based *approaches* Aggarwal [18].

Outlier analysis can be applied on the raw time series or over its irregular component once decomposed. Simple methods for point outlier detection rely on *deviation criteria* or *inter-quartile ranges* assessed on the irregular component. Generally, this class of methods fits empirical or statistical distributions and fix thresholds on what it is expected to occur. Despite their simplicity, time dependencies are disregarded. *Local outlier factor* (LOF) Breunig et al. [19] approach minimizes this drawback by computing anomaly scores based on the local density of an observation with respect to its neighbours where the neighborhood criteria can include temporal and cross-variable distances. *Isolation forests* Liu et al. [20] recursively generate partitions from multivariate series data by randomly selecting a feature and a split value for the feature. Presumably the anomalies need fewer partitions to be isolated compared to "normal" points, thus yielding smaller trees. *Parametric models* from maximum likelihood estimates are also available Chen et al. [21].

Gupta et al. [22] provide a comprehensive survey of contributions on outlier detection over temporal data structures, including (geolocalized) time series data. The approaches to detect *point outliers* are grouped into five major categories: predictive, profile-based models, information-theoretic, classification and clustering approaches. In the context of predictive models, a score is assigned to each observation as a deviation from the estimated value. Estimates can be computed using imputation techniques for univariate and multivariate time series data previously covered. Profile-based approaches trace a normal profile for the time series using classical time series models Wei [3] and more recent advances, including recurrent neural networks that act as auto-encoders Guo et al. [23]. Anomaly scores are then inferred by testing deviations against the approximated profile. The principle behind the less common information theoretic approaches is that the removal of outlier results in higher abstraction ability (time series representations with lower error bound) Jagadish et al. [24].

Approaches for sequential outlier detection traditionally compare subsequences segmented under multi-scale sliding windows to identify dissimilar subsequences. Keogh et al. [25] outlines principles to surpass the computational complexity of computing pairwise time series distances between all subsequences, including heuristics to reorder candidate subsequences, locality sensitive hashing, Haar wavelets, and joint use of symbolic aggregations with augmented tries.

These are used for an improved ordering of subsequences. An additional challenge is the fact that sequential outliers may have an arbitrary length. Chen et al. [26] proposed a new class of approaches that satisfy this premise: a pattern (subsequence of two consecutive points) is defined and outliers are composed of infrequent patterns on either the original time series or compressed time series recovered after wavelet transform.

Time series *clustering* algorithms are as well used to detect sequential outliers. Generally, these approaches segment the inputted series to identify anomalous segments, paying particular attention to distance metrics between time series (including metrics to tolerate misalignments) and barycenter criteria whenever applicable. Understandably, traditional clustering algorithms can be also applied to detect outliers from (multivariate) time series by assuming independence between observations. HOT SAX Keogh et al. [27] also offers the possibility to detect sequential outliers, referred as time series discords, from symbolic representations of the time series. HOT SAX, originally prepared to detect global sequential outliers, was later on extended towards local sequential outliers Toshniwal et al. [28].

Real-time data yields a significant difference: the non fixed length of the data and the requirement that the processing task must be efficient enough to deal with the sampling rate of the new coming observations. Therefore, the referred methods for outliers detection can substantially differ. According to Gupta et al. [22], the methods can be divided into: 1) evolving prediction models, 2) distance based outliers for sliding windows and 3) outliers in high-dimensional data streams, where multiple investigations on the field are depicted in the paper.

Other inconsistencies. In the presence of domain knowledge, *atypical values* or gross errors in time series can be detected by fixing upper and/or lower bounds on the acceptable values. *Duplicate values* are harder to detect as they may not necessarily result in anomalous values. Duplicates can have different causes: 1) accumulation of values from previous observations (generally preceded by missing occurrences), and 2) multiplicity of measurements within a single time step. Density-based outlier approaches are generally considered for the former case, while rule-based analysis of timestamps against sampling expectations are pursued for the latter case. Finally, *irregular sampling rates* observed within or between sensors or between sensors often result from faulty sensor synchronization. Diverse transforms and dedicated time series analysis algorithms have been proposed to deal with irregular measurements Fatehi et al. [29] and Xue-Bo et al. [30].

Parameter-free and autonomous processing. The literature on autonomous selection of either parametric or non-parametric methods for time series processing is scarce, generally providing series-dependent contributions and focusing on a single processing task. Rayana et al. [31] and Zimek et al. [32] proposed ensemble principles to infer anomaly scores from multiple estimates, validated in specific data domains. Similarly, ensemble principles for imputing missing observations in time series have been proposed Li et al. [33] and Oehmcke et al. [34]. Böhm et al. [35] introduced CoCo, a parameter-free method for detecting outliers in data with unknown underlying distributions. Despite the relevance of these contributions, to our knowledge there are not yet methodologies for autonomously assessing, parameterizing and selecting methods able to treat time series unsupervisedly.

III. Solution

Despite the relevance of the surveyed contributions, existing time series pre-processing methods are generally oriented towards specific data regularities and types of errors. Thorough comparisons are thus necessary to place proper decisions, a generally laborious and difficult process due to the difficulty of performing objective assessments in the absence of ground truth.

The AutoMTS is a parameter-free methodology, a composition of steps that guarantee the robust assessment, hyper-parameterization and selection of state-of-the-art processing methods in accordance with the regularities and inconsistencies observed in the inputted series data. We developed to approaches: one for the historical data setting and one for the real-time data setting.

A. Historical data

AutoMTS receives as input a pointer to a database or file with the raw time series data, and produces as output the processed data without inconsistencies in accordance with strict quality criteria.

The major idea behind AutoMTS is to generate precise ground truth for the sound and quality-driven evaluation of available processing options. To this end, AutoMTS relies on two major principles: i) detection of conserved segments within the inputted series data, and ii) modeling the type and amount of observed errors. Under these principles, the assessment can be conducted by purposefully planting inconsistencies along the conserved segments. In this way, available processing options can be objectively assessed.

AutoMTS provides a good coverage of available processing options, providing methods for missing imputation, outlier detection and gross-error removal from time series data. With the aim of handling errors of varying profile, AutoMTS incorporates processing methods able to deal with both point and serial missing and outlier values. In addition, AutoMTS is able explore the aided processing guidance provided by correlated variables within multivariate time series data. To this end, state-of-the-art processing methods able to capture cross-variable dependencies are further supported in AutoMTS.

1) Methodology: AutoMTS is a sequential approach for pre-processing time series produced from heterogeneous networks. The four major steps are: given a (multivariate) time series, the *first step* is to treat non-cumulative duplicates. After the time series is cleansed of duplicates, the *second step* is the detection of atypical values against background knowledge. For instance, in the context of water flow and pressure sensors, lower bounds are generally zero and upper bounds fixed in

accordance with pipe specifications. Atypical values are then translated into missing values to be dealt later in the process. On the *third step*, we detect outlier observations. This is a core step in our pipeline as the wide-diversity of state-of-the-art methods for outlier detection needs to be robustly assessed using the methodology proposed in section III-A2. The selected method, already hyperparameterized, is then applied to detect outliers in the target (multivariate) time series. The detected outliers, along with their anomaly scores, will be given to the user and he may opt to either discard the outliers (default option) or mark some of the detected outliers to be retained in the time series. The *fourth step* is to impute values on the missing observations, including originally missing occurrences as well as the removed outliers and atypical values. Similarly with the third step, this is another core step within the AutoMTS process. The assessment methodology for hyperparameterizing and selecting imputation methods is introduced in section III-A3. Once missing occurrences are imputed, the treated time series is returned by AutoMTS.

2) Autonomous outlier detection (step 3): The third step purposefully plants artificial outliers in the conserved segments of the inputted time series eithe point-wise and/or segment-wise. The robust planting of artificial outliers is essential to gather ground truth for the objective assessment of the methods, necessary to their hyperparameterization and comparison. AutoMTS runs by default 30 process simulations to collect performance estimates.

The outlier detection methods available in the AutoMTS are standard deviation, inter quartile range, isolation forests, local outlier factor, DBScan and HOT SAX.

Let *TP* (true positives) be the correctly detected outliers, *TN* (true negatives) be observations correctly identified as non-outliers, *FP* (false positives) be the incorrectly detected outliers, and *FN* (false negatives) be the non-detected outliers wrongly. To evaluate the behavior of outlier detection methods, we suggest as essential performance views the analysis of recall,

$$\text{recall} = \frac{TP}{TP + FN},$$

to understand the percentage of correctly identified outliers, as well as precision,

$$\text{precision} = \frac{TP}{TP + FP},$$

to understand whether the retrieved outliers were identified at the cost of retrieving non-outlier observations (false positives). To objectively guide the hyperparameterization and selection steps, these complementary views can be combined within scores, such as the F1-score,

$$\text{F1-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}},$$

which is not free of criticisms [36] due to the inherent characteristics of the harmonic mean.

3) Autonomous missing imputation (step 4): The fourth step generates missing observations within conserved segments of the inputted time series in accordance with the profile of missing data observed along the non-conserved segments.

Similarly to the generation of artificial outliers, the generation of artificial missings, is essential to gather ground truth for objective assessments required for the hyperparameterization and selection of imputation methods.

For generating the ground truth, three major steps are undertaken. First, AutoMTS verifies whether the largest conserved segment satisfies a minimum length assumption (four weeks). To generate the artificial missing values random values are selected, for point-wise and/or sequence-wise observations, and are turned into missing observations. By default, 30 process simulations are considered to collect performance estimates

The univariate imputation methods available in the AutoMTS are: random sample, interpolation, LOCF, NOCB and moving average. The supported multivariate methods are: random forests, expectation maximization, kNN, Mice and Amelia.

To evaluate the performance of imputation methods, residue-based scores are considered, including the mean absolute error (MAE),

$$\text{MAE} = \Sigma_{i=1}^{n} |\hat{\vec{x}}_{t_i} - \vec{x}_{t_i}|,$$

where $\vec{x}$ and $\hat{\vec{x}}$ are the observed and imputed time series respectively, and n is the number of missings; the root mean squared error (RMSE),

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{\vec{x}}_{t_i} - \vec{x}_{t_i})^2}{n}},$$

the symmetric mean absolute percentage error (SMAPE); and the percentage of missing values imputed since not all imputation methods may not encounter necessary conditions for imputing certain missing observations.

4) Final remarks on the behavior of historical setting: The state-of-the-art methods supported along the third and fourth steps of the AutoMTS pipeline are tested one by one. A good portion of these methods require the input of parameter values. In this context, hyperparameterization is conducted using the planted inconsistencies in order to identify the best parameters. To this end, we rely on Bayesian optimization [37] due to its inherent ability to traverse only the most promising areas of the search space, thus promoting efficiency. The hyperparameterization should be driven by one of the performance views previously introduced. By default, F1-score is selected for the hyperparameterization of outlier detection methods, while RMSE is the default criteria to guide the hyperparameterization of missing imputation methods.

Once parameterized, methods are then evaluated using the same performance views and the best performing method is selected and used to pre-process the original data.

B. Real-time data

As stated before, AutoMTS uses a sequential approach for pre-processing time series, although in the real-time setting, the approach is also conditional. Instead of having access to the totality of the time series, we receive a new observations with a specific period, so we have to deal observations-wise instead of dataset-wise. The four major steps are the following: given a new observation, the *first step* is to check if it is a duplicate. If it is then we remove the observation, if it is not we go to the next step. The *second step* is to detect if the observation is a gross-error. If it is we remove the observation and impute as a missing value, if it is not we go to the next step. The *third step* is to detect the observation as an outlier. If it is we remove the observation and impute as missing value, if it is not we got to the next step. The *fourth step*, we check if the observation is a missing value. If it is we impute it and if it is not we can store the observation because we know it is cleaned.

Even more, we have to consider that in the real-time setting we have a time limitation, i.e., we have to complete the four previous steps, before a new observation comes into the system. Because of that we need to take into account two new principles: historical window and buffer. For our real-time approach to work we require a historical window, i.e., previous stored observations. This window will help our methods to behave better since they have the window size of observations to work with. The buffer is what we call a window of saved observations before they enter the system. With this we can use methods that require observations after to the observation we are working with. The window and buffer sizes are dependable of the periodicity of the new coming observations, e.g., if an observation takes 1 hour to get into the system, we can use a bigger window and buffer size than if the periodicity of the series is 1 minute.

1) Autonomous outliers detection step 3: The third step autonomously detects if the new observation is an outlier. The available methods are the same as section III-A2, but the observations of the window are considered our ground truth and the planted outliers are planted on them as explained in section in section III-A2.

C. Missing values imputation (step 4)

The fourth step autonomously detects if the new observation is a missing value. The available methods are the same as section III-A3 and the hyperparameterization is performed in the same way as in section III-B1.

For the regular imputation the window and buffer are used and given to the method, more specifically for methods that require further observations to the current one, e.g., moving average and NOCB. Although, the addiction of the buffer is mainly used for this methods, they might not be able to perform, because of the buffer size, or simply because the buffer are composed with missing observations.

1) Final remarks on the behavior of real-time setting.: Instead of doing the hyperparameterization like in the historical data setting and run the hyperparameterization for every new coming observation, we run the hyperparameterization with a pre-defined time period, e.g., once a day. Because of that we have a pre-defined outlier detection and missing values imputation method and every time the hyperparameterization occurs, we select the best new method with all the observations stored in the system, including the new observations since the previous hyperparameterization.

D. Computational complexity

Considering the presence of k_1 pre-processing methods, each with $O(T_i)$ complexity, then the complexity of executing them is $\sum_i^{k_1} O(T_i) = O(k_1 T_{max})$. Assuming that the conducted Bayesian optimization per method converges in a bounded number of k_2 iterations for each method, then $O(k_1 k_2 T_{max})$. Finally, considering the presence of k_3 testing settings in accordance with the detected error profiles in the original series (e.g. k_3=2 for missing and outlier segments with well-defined rate and length distributions), then AutoMTS has $O(k_1 k_2 k_3 T_{max})$ complexity. k_1 and k_3 are constants. Given a window of bounded size w, the majority of pre-processing methods are linear on the window size, yielding $O(k_1 k_2 k_3 w)$.

IV. Results

To assess the significance of the proposed contributions, AutoMTS is extensively evaluated in two water distribution network systems with heterogeneous sensors, producing observations at varying sampling rates, and subjected to unique water consumption patterns and error profiles.

The gathered results confirm the relevance of the proposed AutoMTS methodology, highlighting that processing choices are highly specific to each sensor and thus guarantees of optimality can only be provided under comprehensive and robust assessments. Also, results further offer a thorough comparison of state-of-the-art imputation and outlier detection methods, assessing their ability to handle diverse error profiles in real-world series data with varying regularities, on the historical and real-time settings.

Results are organized in three major steps. First, we describe the networks of heterogeneous sensors that will be used as study cases, exploring some of the produced time series. Second, we provide a thorough comparison of state-of-the-art methods to detect outliers and impute missings, showing that their adequacy is highly dependent on the time series regularities and error profiles. Finally, we assess AutoMTS, quantifying its performance gains.

A. Study cases: Beja and Barreiro water distribution systems

A Water Distribution Network (WDN) is a system composed of pumps, pipelines, tanks and other elements for delivering water in adequate quantities, pressure and quality for the everyday needs. WDNs can be equipped with an arbitrarily-high number of heterogeneous sensors, including water flow and pressure sensors.

The results of this article were obtained in collaboration with two major water utilities: Barreiro city Council and Beja

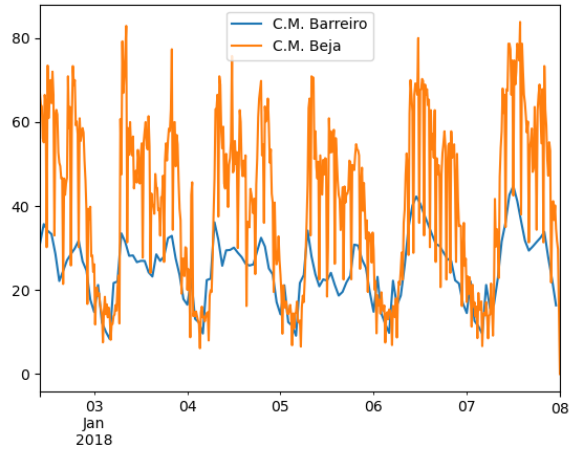

Fig. 1: Sensor measurements over 5 illustrative days for both Barreiro and Beja WDNs water flow sensors.

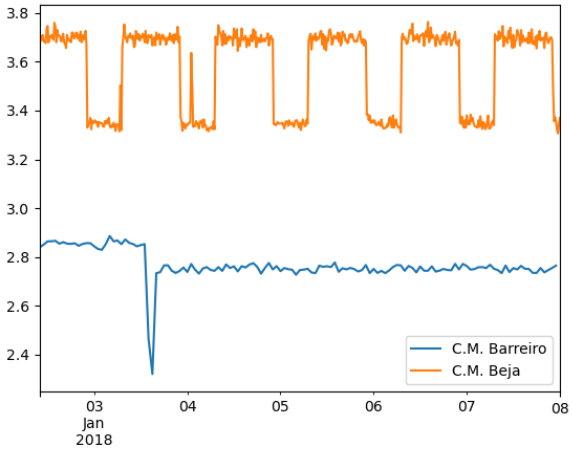

Fig. 2: Sensor measurements over 5 illustrative days for both Barreiro and Beja WDNs water pressure sensors.

city Council, which provided time series representative of their telemetry systems.

Barreiro WDN is composed by 14 sensors of water flow and pressure that provide aggregated measurements on an hourly basis along 2018. The time series has 8473 observations, an amount inferior to the total yearly hours given the presence of weekly periods without measurements – real sequential missings – and the presence of a scarce number of punctual missings. Beja WDN offers water flow and pressure measurements along a two-year period (5/2017 to 4/2019) with an approximate 5-minute sampling rate. Each time series has over 200.000 observations, a irregular sampling rate and the presence of missing values along segments of lower extension than those observed in the Barreiro WDN.

Figures 1 and 2 depicts the water flow and water pressure series from sensors located near the principal tanks in the Barreiro and Beja WDNs. As one can observe, the pressure and flow series from show highly dissimilar structure. In addition, sensors of the type show considerably different regularities for different water distribution systems. These observations motivate the need to perform processing decisions separately for each sensor from the monitored systems.

B. Experimental setting

To assess the impact of placing appropriate choices along the processing stages in accordance with the characteristics and inconsistencies observed along time series, we consider the water flow and pressure time series from Barreiro and Beja WDNs and applied the proposed AutoMTS methodology to generate ground truth. To facilitate the interpretability of results, we further varied the profile of the planted inconsistencies for some of the conducted analyzes. The major parameters controlling the experimental setting are:

- available methods for point outlier detection (e.g. IF (IF)) and sequential outlier detection (e.g. SAX), and the corresponding parameters;
- planted outlier profiles, including: i) frequency of outliers (2% and 10%); ii) type of outliers (point versus sequential); and iii) length of sequential outliers;
- available methods for missing imputation from univariate series (e.g. moving average) or multivariate series (e.g. MICE), and corresponding parameters;
- planted missing profiles, including: i) frequency of missing values (from 2% to 10%); ii) type of missings (point versus sequential); and iii) length of sequential missing observations.

The presented results provide the average performance collected from 30 simulations. A stochastic process to generate inconsistencies in accordance with the introduced parameters is used to produce each simulation. Random seeds are considered to guarantee fair comparisons between methods.

The parameters controlling the experimental setting are abreviated as the following: *WP* (water pressure sensor), *WF* (water flor sensor), *P* (point-wise), *S* (sequential/segment-wise), *2%* (2% of planted anomalies) and *10%* (10% of planted anomalies). The score for the outliers detection methods is the F1-score. On the other hand, for the missing values imputation the decisive method is the RMSE.

For the real-time setting for the experimental setting was used a buffer size of 5 and window size of 200.

C. AutoMTS performance

Table I provides a comprehensive analysis of the performance of the available outliers detection methods on multiple settings for the historical and real-time settings. In the historical setting the isolation forests (IF) method leads for the water pressure sensors while the inter quartile range (IQR) is the best performing method for the water pressure sensors. Even more for the water pressure sensors with 2% of planted anomalies, the scores obtained are considered low, even tough they are the better. For the real-time setting the dominant method is the inter quartile range for all settings except the segments in the 2% of planted anomalies, where the standard deviation and

Setting	Historical		Real-time	
	Method	Score	Method	Score
WP:P:2%	IF	0.322±0.00	IQR	0.56±0.02
WF:P:2%	IQR	0.787±0.10	IQR	0.435±0.17
WP:S:2%	IF	0.301±0.00	Std deviation	0.567±0.17
WF:S:2%	IQR	0.827±0.07	DBScan	0.298±0.02
WP:P:10%	IF	0.867±0.03	IQR	0.435±0.17
WF:P:10%	IQR	0.926±0.02	IQR	0.868±0.05
WP:S:10%	IF	0.854±0.03	IQR	0.892±0.00
WF:S:10%	IQR	0.929±0.03	IQR	0.71±0.1

TABLE I: Best performing methods for the outliers detection methods on multiple settings for the Barreiro WDN.

Setting	Historical		Real-time	
	Method	Score	Method	Score
WP:P:2%	Interp.	0.048±0.01	Interp.	0.039±0.02
WF:P:2%	Interp.	16.871±2.52	MA	11.716±3.59
WP:S:2%	MA	0.061±0.06	Interp.	0.045±0.04
WF:S:2%	MA	24.745±12.15	Rand. forest	12.093±3.5
WP:P:10%	Interp.	0.05±0.00	Interp.	0.046±0.01
WF:P:10%	Interp.	17.246±1.34	Interp.	11.815±1.58
WP:S:10%	Mean	0.169±0.01	Nocb	0.046±0.05
WF:S:10%	MA	24.396±12.08	Nocb	13.917±5.12

TABLE II: Best performing methods for the missing values imputation methods on multiple settings for the Beja WDN.

the DBScan are better. Overall the results are not as high as the historical setting, but end up having good results.

Although this are the best performing methods (best values of F1-score) for each of the settings, in the real-time setting it is important to consider the time it takes to detect the outliers. All this methods were able to detect one outlier under 1 second, so we can assure that they are performing well.

Table II provides a comprehensive analysis of the performance of the available outliers detection methods on multiple settings for the historical and real-time settings. In the historical data setting, the interpolation is the most dominant method, followed by the moving average (MA) and surprisingly the mean method. The disparity in the scores are between the water pressure and water flow sensors, mostly because of the differences of both values ranges that those two sensor have. Although the moving average is getting good results, some times might not be possible to impute the totality of the missing values (mainly on the sequential missing values), so the results here might be a little misleading. Even more, the interpolation method is particularly strong in point-wise imputation.

For the real-time data setting, the interpolation is again dominant, but now the NOCB and random forests are also part of the best available methods. The interpolation is again good for point-wise missings, except for the water flow sensor with 2% of planted anomalies, where the moving average is the best performing method. The random forest which is a multivariate method is the best performing method for the setting water flow sensor, sequential with 2% of artificial missings, further proving the relevance of multivariate methods in this work. At last the NOCB method perform better for 10% of sequential artificial generated missings, which also proves the importance of the buffer.

Complementary, Figure 3 offers a graphical description of previous results for the Barreiro WDN on the historical setting for the water flow sensor, further showing how the performance of different outlier detection methods vary with the amount of planted outliers. As we could see previously the inter quartile range method performs better than the others and improve the results with higher percentages of planted outliers. The HOT SAX method is the only which deals mostly for sequence of outliers is the best performing mthod with 5% of outliers. The standard deviation tends to get worst with high percentage of outliers.

Figure 4 offers complementary results for the real-time setting for the water pressure sensor of Beja WDN. All methods tend to get worse with higher percentages of missing values, except the mean method which performs similarly for any number of missings. The random forests also have a stable performance like the mean method, and while he is not a competitive method at lower percentages is the best perfoming methods for higher one water pressure sensors.. The LOCF and interpolation methods require dependencies of closer observations tend to get worse with higher percentages of missings but are the best performing method for lower percentages.

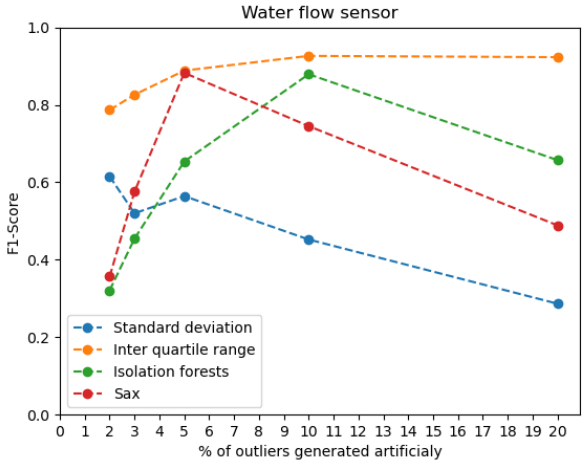

Fig. 3: Performance of outlier detection methods with varying percentage of point outliers planted in time series for water flow sensor in Barreiro WDN, for historical setting

D. AutoMTS tool

Figure 5 provides a snapshot of the AutoMTS tool. On the left panel it is possible to upload the file which contains the time series dataset. Different file formats are supported, including .xlsx and .csv, as well as different data representations. An illustrative representation of the input data is a table with timestamped rows containing the measurements and as many columns as the number of sensors (time series). If sensors have temporally misaligned measurements, each row can alternatively describe a single event, identifying the timestamp, sensor and collected measurement. To guarantee that ground truth is assessed over the provided series data, each sensor needs to have at least one period of four weeks without missing observations. Otherwise, synthetic series are generated for the

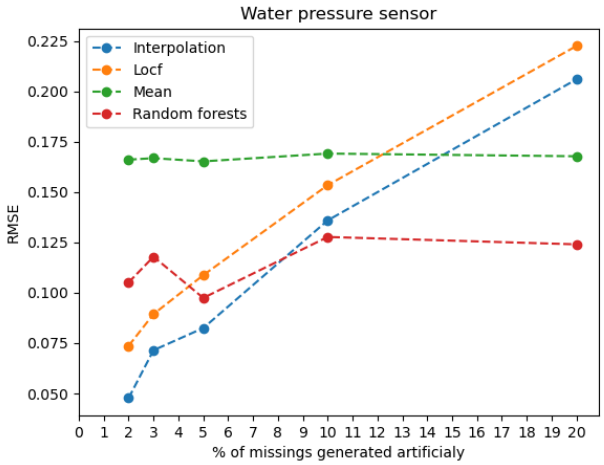

Fig. 4: Performance of missing imputation methods with varying percentage of point missings planted in time series for water pressure sensor in Beja WDN, for real-time setting

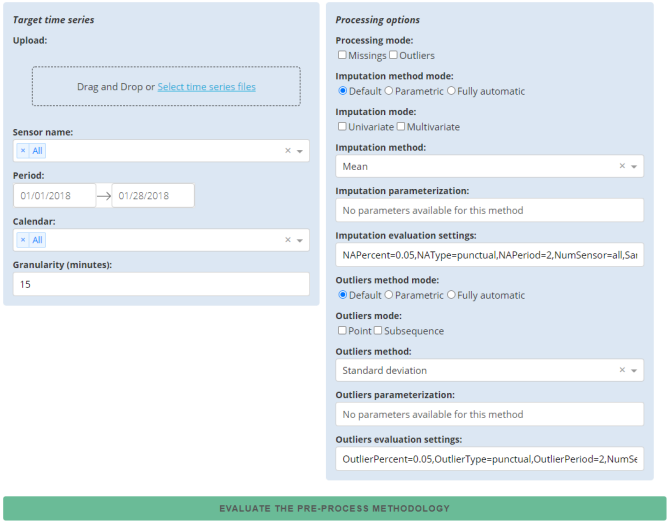

Fig. 5: Graphical user interface.

parameterization and selection of methods. Once the uploaded dataset passes the initial validation process, it is possible to filter the dataset by selecting the time series (sensors) that we want to process. This can be done using *sensor name* fields. It is possible to further filter the observations by time period on the *period* field, the days of the week on the *calendar* field (e.g. weekdays, holidays, saturdays), as well as the desirable time granularity for the target time series.

On the right panel it is possible to select the steps along the AutoMTS pipeline to be accomplished, in particular whether we want to conduct missing imputation and/or outlier detection. For both options, it is possible to select one of three distinct modes: i) the *default* mode which provides a simple rule-based decision on what is the most appropriate method given the general characteristics of the inputted series data; ii) the *parametric* mode which allows the user to select a desirable method method and its parameters; and, at last, iii) the *fully automatic* mode which runs AutoMTS (section III) to autonomously identify the best method for each one of the sensors selected in the left panel.

The user can optionally specify the profile of the artificially planted missing values and outlier values to be considered along the evaluation stage of AutoMTS (as well as to provide statistics whenever the user opts to select default and parametric modes). Here the user can select the type, percentage and duration of artificial missings and outliers. It is also possible to select the number of sensors on where we want to plant the artificial inconsistencies. Finally, the user can also specify whether the inconsistencies must occur at the same time for the inputted set of sensors or planted for each sensor individually, thus mimicking different real-world problems in heterogeneous networks.

After running the query the user can select which outliers he wants to maintain or remove and then it is possible download the time series selected already cleaned.

V. Conclusion

In this work we were able to complete our objectives and produce two approaches for the fully-autonomous and quality-driven processing of time series data produced by networks of heterogeneous sensors. Each one is parameter-free and offers guarantees of optimality. To optimize pre-processing choices, ground truth is created from conserved time series segments for possible error profiles. In addition, AutoMTS provide a coverage of state-of-the-art methods of outliers detection and missing values imputation. AutoMTS implements processing methods able to work with point-wise or sequence observations, and with cross-variable dependencies in the presence of multivariate time series data. Also, our methodology can work with varying types and amount of missing and outliers values, including both point and sequential occurrences of different duration and recurrence.

The experimental evaluation in two real world study cases of water distribution network systems with different sampling rates, water consumption patterns and error profiles confirm the significance of AutoMTS contributions and highlight that pre-processing choices are highly specific to each sensor. Thus guarantees of optimality can only be provided under a robust evaluation. Also the results further offer a comparison of the available methods, showing the strengths and limitations when handling multiple error profiles in real-world time series.

Also the time that takes to pre-process in the real-time setting and its results show that our approach is applicable for the treatment of time series in data streams.

Our approach also enables and facilitate the incorporation and/or implementation of other methods, with almost no effort.

At last, this work provides a functional tool for the WISDOM project, that would help the detection of events of interest and acquire a comprehensive view of the behaviour produced by their time series data.

A. Future work

Even though the work was completed with success, some complementary extensions could be implemented in order to improve our approach. For the imputation methods, some specific time series methods could be added, e.g., the 10-Min flow model from Quevedo, as long some variations of the moving average method which could deal better with sequential missings.

Moreover, the implementation of our real-time script in a real system could provide a better insight on the behaviour of our methodology, practical validation with the water entities would be performed and evaluation of the pre-processing in the detection leakage and on consumption patterns. Finally, further evaluation on irregular time series sampling can be supported.

VI. Scientific communication

During the development of this methodology the article Sousa et al. [38] was published and presented in the EAI Qshine 2020 - 16th EAI International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness. This article only depicted the historical data setting of the AutoMTS tool. The paper that extends the contributions for the real time setting is under submission.

References

[1] P. Gill, N. Jain, and N. Nagappan. "Understanding Network Failures in Data Centers: Measurement, Analysis, and Implications". In: *Proceedings of the ACM SIGCOMM 2011 Conference*. SIGCOMM '11. Toronto, Ontario, Canada: Association for Computing Machinery, 2011, 350–361. ISBN: 9781450307970. URL: https://doi.org/10.1145/2018436.2018477.

[2] R. K. Jain. "A state space model-based method of seasonal adjustment". In: *Monthly Lab. Rev.* 124 (2001), p. 37.

[3] W. W. Wei. "Time series analysis". In: *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*. 2006.

[4] M. S. Osman, A. M. Abu-Mahfouz, and P. R. Page. "A Survey on Data Imputation Techniques: Water Distribution System as a Use Case". In: *IEEE Access* 6 (2018), pp. 63279–63291. ISSN: 2169-3536.

[5] D. J. Stekhoven and P. Bühlmann. "MissForest—non-parametric missing value imputation for mixed-type data". In: *Bioinformatics* 28.1 (Oct. 2011), pp. 112–118. ISSN: 1367-4803. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/28/1/112/583703/btr597.pdf. URL: https://doi.org/10.1093/bioinformatics/btr597.

[6] S. van Buuren and K. Groothuis-Oudshoorn. "mice: Multivariate Imputation by Chained Equations in R". In: *Journal of Statistical Software* 45.3 (2011), pp. 1–67. ISSN: 1548-7660. URL: https://www.jstatsoft.org/v045/i03.

[7] A. Ilin and T. Raiko. "Practical Approaches to Principal Component Analysis in the Presence of Missing Values". In: *J. Mach. Learn. Res.* 11 (Aug. 2010), pp. 1957–2000. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1756006.1859917.

[8] L. Garg et al. "Tensor-Based Methods for Handling Missing Data in Quality-of-Life Questionnaires". In: *IEEE JBHI* 18.5 (2014), pp. 1571–1580. ISSN: 2168-2208.

[9] Y. Luo et al. "Multivariate time series imputation with generative adversarial networks". In: *Advances in Neural Information Processing Systems*. 2018, pp. 1596–1607.

[10] S. Moritz et al. "Comparison of different Methods for Univariate Time Series Imputation in R". In: (Oct. 2015).

[11] A. Zeileis and G. Grothendieck. "zoo: S3 Infrastructure for Regular and Irregular Time Series". In: *Journal of Statistical Software, Articles* 14.6 (2005), pp. 1–27. ISSN: 1548-7660. URL: https://www.jstatsoft.org/v014/i06.

[12] A. Kowarik and M. Templ. "Imputation with the R Package VIM". In: *Journal of Statistical Software, Articles* 74.7 (2016), pp. 1–16. ISSN: 1548-7660. URL: https://www.jstatsoft.org/v074/i07.

[13] J. Quevedo et al. "Validation and reconstruction of flow meter data in the Barcelona water distribution network". In: *Control Eng. Practice* 18.6 (2010), pp. 640 –651. ISSN: 0967-0661. URL: http://www.sciencedirect.com/science/article/pii/S0967066110000791.

[14] R. Barrela et al. "Data reconstruction of flow time series in water distribution systems – a new method that accommodates multiple seasonality". In: *Journal of Hydroinformatics* 19.2 (Dec. 2016), pp. 238–250. ISSN: 1464-7141. eprint: https://iwaponline.com/jh/article-pdf/19/2/238/390915/jh0190238.pdf. URL: https://doi.org/10.2166/hydro.2016.192.

[15] F. Fan, Z. Li, and Y. Wang. "On-line imputation for missing values". In: *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. 2017, pp. 1–5.

[16] D. Hawkins. *Identification of outliers*. Monographs on applied probability and statistics. London [u.a.]: Chapman and Hall, 1980. X, 188. ISBN: 041221900X. URL: http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+02435757X&sourceid=fbw_bibsonomy.

[17] W.-S. Chan. "Understanding the effect of time series outliers on sample autocorrelations". In: *Test* 4.1 (1995), pp. 179–186.

[18] C. C. Aggarwal. "Outlier analysis". In: *Data mining*. Springer. 2015, pp. 237–263.

[19] M. Breunig et al. "LOF: Identifying Density-Based Local Outliers." In: vol. 29. June 2000, pp. 93–104.

[20] F. T. Liu, K. M. Ting, and Z. hua Zhou. "Isolation Forest". In: *In ICDM '08: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. IEEE Computer Society*. 2008, pp. 413–422.

[21] C. Chen and L.-M. Liu. "Joint Estimation of Model Parameters and Outlier Effects in Time Series". In: *Journal of the American Statistical Association* 88.421 (1993), pp. 284–297. eprint: https://doi.org/10.1080/01621459.1993.10594321. URL: https://doi.org/10.1080/01621459.1993.10594321.

[22] M. Gupta et al. "Outlier Detection for Temporal Data: A Survey". In: *IEEE Transactions on Knowledge and Data Engineering* 26.9 (2014), pp. 2250–2267. ISSN: 2326-3865.

[23] Y. Guo et al. "Multidimensional time series anomaly detection: A gru-based gaussian mixture variational autoencoder approach". In: *Asian Conf. on Machine Learning*. 2018, pp. 97–112.

[24] H. V. Jagadish, N. Koudas, and S. Muthukrishnan. "Mining Deviants in a Time Series Database". In: *Proceedings of the 25th International Conference on Very Large Data Bases*. VLDB '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 102–113. ISBN: 1-55860-615-7. URL: http://dl.acm.org/citation.cfm?id=645925.758373.

[25] E. Keogh et al. "Finding the most unusual time series subsequence: algorithms and applications". In: *Knowledge and Information Systems* 11.1 (2007), pp. 1–27. ISSN: 0219-3116. URL: https://doi.org/10.1007/s10115-006-0034-6.

[26] X.-Y. Chen and Y.-Y. Zhan. "Multi-scale anomaly detection algorithm based on infrequent pattern of time series". In: *Journal of Computational and Applied Mathematics* 214.1 (2008), pp. 227 –237. ISSN: 0377-0427. URL: http://www.sciencedirect.com/science/article/pii/S0377042707001100.

[27] E. Keogh, J. Lin, and A. Fu. "Hot sax: Efficiently finding the most unusual time series subsequence". In: *Fifth IEEE Int. Conf. on Data Mining (ICDM'05)*. Ieee. 2005, 8–pp.

[28] D. Toshniwal and S. Yadav. "Adaptive outlier detection in streaming time series". In: *Proceedings of International Conference on Asia Agriculture and Animal, ICAAA, Hong Kong*. Vol. 13. 2011, pp. 186–192.

[29] A. Fatehi and B. Huang. "Kalman filtering approach to multi-rate information fusion in the presence of irregular sampling rate and variable measurement delay". In: *Journal of Process Control* 53 (2017), pp. 15–25.

[30] J. Xue-Bo, D. Jing-Jing, and B. Jia. "Target tracking of a linear time invariant system under irregular sampling". In: *Int. Journal of Advanced Robotic Systems* 9.5 (2012), p. 219.

[31] S. Rayana and L. Akoglu. "Less is more: Building selective anomaly ensembles". In: *Acm transactions on knowledge discovery from data (tkdd)* 10.4 (2016), pp. 1–33.

[32] A. Zimek, R. J. Campello, and J. Sander. "Ensembles for unsupervised outlier detection: challenges and research questions a position paper". In: *Acm Sigkdd Explorations Newsletter* 15.1 (2014), pp. 11–22.

[33] L. Li et al. "Missing value imputation for traffic-related time series data based on a multi-view learning method". In: *IEEE Transactions on Intelligent Transportation Systems* 20.8 (2018), pp. 2933–2943.

[34] S. Oehmcke, O. Zielinski, and O. Kramer. "kNN ensembles with penalized DTW for multivariate time series imputation". In: *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2016, pp. 2774–2781.

[35] C. Böhm et al. "CoCo: Coding Cost for Parameter-Free Outlier Detection". In: *Proceedings of the 15th ACM SIGKDD IC on Knowledge Discovery and Data Mining*. KDD '09. Paris, France: Association for Computing Machinery, 2009, 149–158. ISBN: 9781605584959. URL: https://doi.org/10.1145/1557019.1557042.

[36] J. Davis and M. Goadrich. "The relationship between Precision-Recall and ROC curves". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 233–240.

[37] J. Snoek, H. Larochelle, and R. P. Adams. "Practical Bayesian Optimization of Machine Learning Algorithms". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 2951–2959. URL: http://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms.pdf.

[38] R. Sousa, C. Amado, and R. M. C. Henriques. "AutoMTS: fully autonomous processing of multivariate time series data from heterogeneous sensor networks". In: *16th EAI Int. Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine)*. Oct. 2020.