AutoMTS: Fully Autonomous Processing of Multivariate Time Series Data from Heterogeneous Sensor Networks

Ricardo Sousa[1,2], Conceição Amado[1], and Rui Henriques[2](✉)

[1] CEMAT and Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
{ricardo.filipe.sousa,conceicao.amado}@tecnico.ulisboa.pt

[2] INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
rmch@tecnico.ulisboa.pt

Abstract. Heterogeneous sensor networks, including water distribution systems and traffic monitoring systems, produce abundant time series data with an arbitrarily-high multivariate order 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. AutoMTS is provided as an open-source tool available at https://github.com/RicardoFLNSousa/AutoMTS/tree/master.

Keywords: Parameter-free learning · Multivariate time series · Missing values imputation · Outlier detection · Heterogeneous sensor networks

1 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

X. Wu et al. (Eds.): QShine 2020, LNICST 381, pp. 154–178, 2021.
https://doi.org/10.1007/978-3-030-77569-8_12

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 [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 suboptimal 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 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. To this end, series data are automatically explored in order to detect conserved segments and identify the profile of observed errors, which are then planted in the conserved segments for the sound comparison of available processing choices.

Second, AutoMTS provides a comprehensive coverage of available processing options, currently providing over twenty state-of-the-art methods for missing imputation, outlier detection and gross-error removal from time series data. Particular attention was placed to guarantee the presence of state-of-the-art methods 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.

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.

AutoMTS is provided as both a graphical and programmatic tool satisfying strict usability criteria.

The manuscript is structured as follows. Section 2 provides essential background and surveys recent contributions on time series data processing. Section 3 described the AutoMTS approach. Section 4 comprehensively assesses the adequacy of AutoMTS using two real-world heterogeneous networks as study cases. Finally, concluding remarks and major implications are synthesized.

2 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 preprocessing *methods* implemented in AutoMTS.

Time Series Data Processing. Signals produced by sensors are generally represented as *time series*, an ordered set of observations $\mathbf{x}_{1..T} = (\mathbf{x}_1, ..., \mathbf{x}_T)$, each $\mathbf{x}_t$ being recorded at a specific time point t. Time series can be *univariate*, $\mathbf{x}_t \in \mathbb{R}$, or *multivariate*, $\mathbf{x}_t \in \mathbb{R}^m$, 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 *preprocessing 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 [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 *autoregression*, *differencing* and *exponential smoothing* operations to either detect deviations from expectations as well as to impute missing values [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) [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 [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, $\hat{\mathbf{x}}_t = \frac{1}{m}\sum_{j=-k}^{k} \mathbf{x}_{t+j}$ 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 [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 [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) [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 [7] and machine learning techniques such as Gaussian process regression, tensor-based methods [8], and neural networks, specially auto-associative neural networks [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 [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 [10,12].

Imputation methods have been also proposed in the context of specific domains. In water-energy-gas distribution systems, the well-recognized Quevedo method [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.

Time Series Outlier Detection. *Outliers* are observations significantly deviating from expectations as to arouse suspicion of being generated by a different mechanism [15]. 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 [16]: 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* [17].

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) [18] 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* [19] 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 [20].

Gupta et al. [21] 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 [3] and more recent advances, including recurrent neural networks that act as auto-encoders [22]. 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) [23].

Approaches for sequential outlier detection traditionally compare subsequences segmented under multi-scale sliding windows to identify dissimilar subsequences. Keogh et al. [24] 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. [25] 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 [26] 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 [27].

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 [28,29].

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. [30] and Zimek et al. [31] 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 [32,33]. Böhm et al. [34] 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.

3 Solution: Autonomous Time Series Data Processing

Despite the relevance of the surveyed contributions, existing time series preprocessing 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. In this context, we propose a novel approach for the fully Autonomous processing of Multivariate Time Series data, referred as AutoMTS. 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. Annotations, including bounds associated with the estimated anomaly scores, and performance statistics can be optionally outputted.

The AutoMTS is a parameter-free methodology, a composition of steps that guarantee the robust assessment, hyperparameterization and selection of state-of-the-art processing methods in accordance with the regularities and inconsistencies observed in the inputted series data. 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 and, depending on their length, on synthetically generated series using the approximated component-wise regularities. In this way, available processing options can be objectively assessed.

AutoMTS provides a good coverage of available processing options, providing over twenty state-of-the-art 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.

3.1 Methodology

AutoMTS is a sequential approach for preprocessing time series produced from heterogeneous networks. The four major steps are depicted in Fig. 1. Given a (multivariate) time series, the *first step* is to treat non-cumulative duplicates through a rule-based inspection of sampling irregularities (see Sect. 2). 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 Sect. 3.2. 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 outputted 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 Sect. 3.3. Once missing occurrences are imputed, the treated time series is returned by AutoMTS.

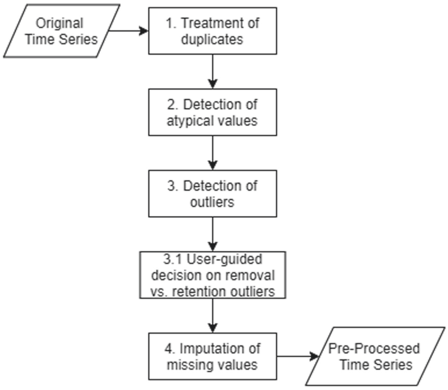


Fig. 1. Time series preprocessing methodology.

3.2 Autonomous Outlier Detection *(Step 3)*

The third step purposefully plants artificial outliers in the conserved segments of the inputted time series in accordance with the signal regularities observed along those segments. The regularities reveal information related with the point-wise and segment-wise distribution of values to guide the planting of point and segment outliers. 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.
For generating the ground truth, five major steps are undertaken:

1. the time series is decomposed into trend, seasonal, cyclical and noise/irregular components;
2. the distribution of values observed along the irregular component is dynamically fitted into a well-known probability distribution using both the Kolmogorov-Smirnov and χ^2 statistical tests;
3. the tails of the approximated distributions are used to plant point outlier values randomly distributed along the irregular component;
4. sequential outliers are further planted by guaranteeing a residual joint probability of the observed values along the artificial subsequence;
5. the irregular component with the planted point and sequential outliers is added to the original trend, seasonal and cyclical components.

The statistical properties of this five-step process guarantee the presence of non-trivial outliers resembling the characteristics of real-world anomalies. AutoMTS runs by default 30 process simulations to collect performance estimates.

Some of the outlier detection methods available in the AutoMTS are standard deviation, inter quartile range, isolation forests, LOF, 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 [35] due to the inherent characteristics of the harmonic mean. Complementary integrative scores able to reconcile recall and precision views at alternative anomaly score thresholds, including the area under the ROC curve (AUC), can be alternatively selected [35].

3.3 Autonomous Missing Imputation *(Step 4)*

The fourth step wittingly 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. The profile of missing observations essentially discloses information on their temporal distribution, nature (point versus sequential), length, amount and periodicity (well-defined versus random). Similarly to the generation of artificial outliers, the removal of observations 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 times the seasonal factor as default). If the largest conserved segment does not satisfies the assumption, the segment is replaced by an artificial time series. To generate the artificial time series, the original time series should be decomposed in order to approximate its core components. The irregular component is regenerated in accordance with the underlying distribution and added to the remaining components to produce a synthetic time series without missing occurrences. Second, the approximated percentage amount and temporal distribution of punctual missings in the original time series is used to remove observations from the conserved segment or synthesized time series. Third, and finally, sequential missings are planted in accordance with the distribution of their extension and recurrence on the original time series.

The statistical properties of this three-step process guarantee the presence of missing observations resembling real-world characteristics. By default, 30 process simulations are considered to collect performance estimates.

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

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

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

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

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

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.

3.4 Computational Complexity

Considering the presence of k_1 preprocessing 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 preprocessing methods are linear on the window size, yielding $O(k_1 k_2 k_3 w)$.

3.5 Final Remarks on the Behavior of AutoMTS

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 [36] 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. If the length of the largest conserved segment (or synthesized time series) permits, the segment is further segmented into two subsequences, one for hyperparameterization and other for the final method evaluation. In this way, we prevent the overfitting of the selected parameters.

4 Results

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.

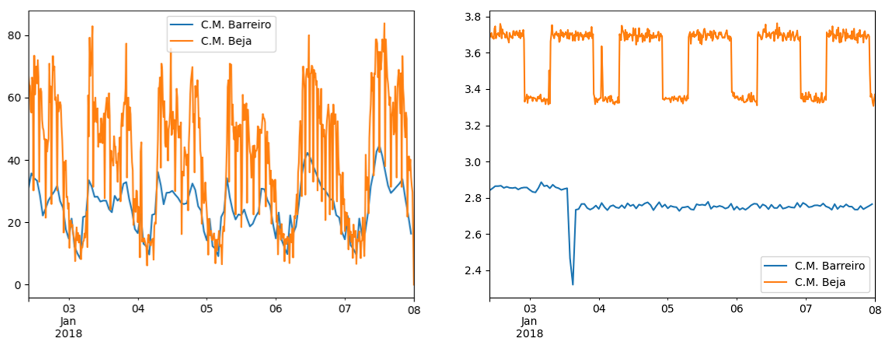
Study Cases: Beja and Barreiro Water Distribution Systems

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 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.

Figure 2a depicts the water flow series from sensors located near the principal tanks in the Barreiro and Beja WDNs, while Fig. 2b depicts the time series



(a) water flow (b) water pressure mean

Fig. 2. Sensor measurements over 5 illustrative days for both Barreiro and Beja WDNs.

produced by the approximately co-located water pressure sensors. 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.

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:

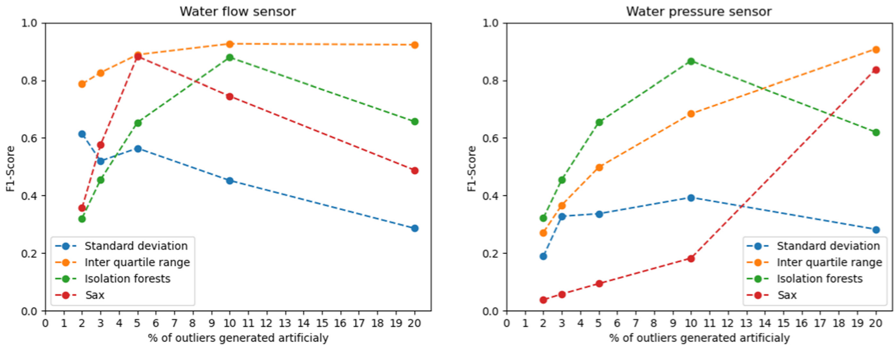
– available methods for point outlier detection (e.g. isolation forests) and sequential outlier detection (e.g. SAX), and the corresponding parameters;
– planted outlier profiles, including: i) frequency of outliers (1% to 10%); ii) type of outliers (point versus sequential); and iii) length of sequential 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 1% to 20%); 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.

4.1 AutoMTS Performance

Table 1 provides a comprehensive analysis of the performance of multiple outlier detection methods on time series data produced from different sensors installed within the Barreiro and Beja WDNs. We can observe that different settings – different sensors, water distribution systems, outlier types – propel different choices. Considering F1-score and recall, while isolation forests appears to be the most promising option for water pressure sensors, inter-quartile range performance is particularly good on water flow sensors. The recall of the most surveyed methods significantly differs between WDNs. Understandably, as AutoMTS selects the best choice available, it shows optimal performance across major performance views.

Figures 3a and 3b offer a complementary graphical description of previous results, further showing how the performance of different outlier detection methods vary with the amount of planted outliers. Illustrating, HOT SAX is not competitive when considering a low amount of outliers (offers a good recall yet low precision due its focus on outlier segments), yet performance improves with a medium-to-high amount of outliers. The analysis of these figures further highlights that there are significant changes in performance associated with changes on the amount of outlier values. These variations can affect processing decisions (e.g. isolation forests versus inter-quartile range in pressure sensors), further supporting the relevance of the proposed AutoMTS methodology.



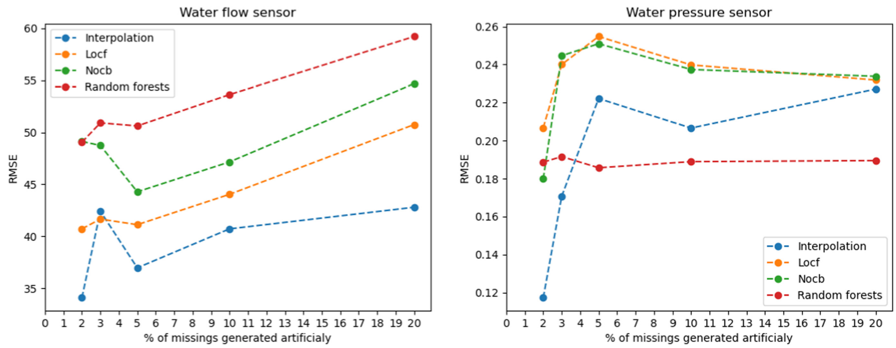
(a) water flow sensor at Barreiro WDN (b) water pressure mean at Barreiro WDN

Fig. 3. Performance of outlier detection methods with varying percentage of planted point outliers in time series produced from heterogeneous sensors.

Similarly to Table 1, Table 2 gathers results on the performance of missing imputation methods on time series data produced from different sensors placed within the Barreiro and Beja WDNs. Decisions are similarly dependent on the target sensor, network and missing profile (type and amount). For instance, while interpolation shows generally good performance on water flow sensors is not competitive on water pressure sensors. The characteristics of the Beja WDN, where measurements are collected under a smaller sampling rate, presents compelling evidence towards the use moving average imputation technique. Finally, we can observe a decreased performance of single-value estimators such as LOCF and NOCB for imputing missing segments and an increased performance of multi-point estimators such as moving average estimators for this sequential type of missings. These remarks underline the role of AutoMTS.

Table 1. Performance of outlier detection methods for water pressure and flow sensors from Barreiro and Beja WDNs with planted point and sequential outliers on up to 10% of observations.

		Barreiro WDN				Beja WDN			
		F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall
pressure: point	Standard deviation	0.393 ± 0.06	0.921 ± 0.01	0.259 ± 0.05	0.832 ± 0.05	0.414 ± 0.02	0.926 ± 0.00	0.261 ± 0.02	$\mathbf{1.0 \pm 0.00}$
	Inter quartile range	0.684 ± 0.01	0.908 ± 0.00	1.0 ± 0.00	0.52 ± 0.01	0.686 ± 0.02	0.952 ± 0.00	0.522 ± 0.03	$\mathbf{1.0 \pm 0.00}$
	Isolation forests	$\mathbf{0.867 \pm 0.03}$	0.973 ± 0.01	$\mathbf{0.873 \pm 0.03}$	$\mathbf{0.861 \pm 0.03}$	$\mathbf{0.75 \pm 0.03}$	0.95 ± 0.01	$\mathbf{0.751 \pm 0.03}$	0.75 ± 0.03
	Local outlier factor	0.179 ± 0.07	0.835 ± 0.01	0.18 ± 0.07	0.178 ± 0.07	0.0 ± 0.00	0.801 ± 0.00	0.0 ± 0.00	0.0 ± 0.00
	Dbscan	0.0 ± 0.00	0.899 ± 0.00	0.0 ± 0.00	0.0 ± 0.00	0.0 ± 0.00	0.9 ± 0.00	0.0 ± 0.00	0.0 ± 0.00
	SAX	0.182 ± 0.00	0.106 ± 0.00	1.0 ± 0.00	0.1 ± 0.00	0.61 ± 0.11	0.945 ± 0.01	0.448 ± 0.11	$\mathbf{1.0 \pm 0.00}$
	AutoMTS	0.867 ± 0.03	0.973 ± 0.01	0.873 ± 0.03	0.861 ± 0.03	0.75 ± 0.03	0.95 ± 0.01	0.751 ± 0.03	0.75 ± 0.03
flow: point	Standard deviation	0.452 ± 0.04	0.93 ± 0.00	0.293 ± 0.03	$\mathbf{1.0 \pm 0.00}$	0.477 ± 0.02	0.932 ± 0.00	0.313 ± 0.02	$\mathbf{1.0 \pm 0.00}$
	Inter quartile range	$\mathbf{0.926 \pm 0.02}$	0.986 ± 0.00	0.864 ± 0.04	$\mathbf{1.0 \pm 0.00}$	$\mathbf{0.964 \pm 0.01}$	0.993 ± 0.00	$\mathbf{0.981 \pm 0.01}$	0.949 ± 0.00
	Isolation forests	0.879 ± 0.05	0.976 ± 0.01	$\mathbf{0.886 \pm 0.05}$	0.873 ± 0.05	0.855 ± 0.02	0.971 ± 0.00	0.856 ± 0.02	0.854 ± 0.02
	Local outlier factor	0.15 ± 0.05	0.829 ± 0.01	0.151 ± 0.05	0.149 ± 0.05	0.084 ± 0.03	0.817 ± 0.01	0.084 ± 0.03	0.083 ± 0.03
	Dbscan	0.837 ± 0.06	0.966 ± 0.01	0.881 ± 0.09	0.801 ± 0.03	0.774 ± 0.01	0.944 ± 0.00	0.954 ± 0.02	0.651 ± 0.01
	SAX	0.745 ± 0.10	0.96 ± 0.01	0.603 ± 0.12	$\mathbf{1.0 \pm 0.00}$	0.598 ± 0.11	0.944 ± 0.01	0.435 ± 0.10	$\mathbf{1.0 \pm 0.00}$
	AutoMTS	0.926 ± 0.02	0.986 ± 0.00	0.864 ± 0.04	1.0 ± 0.00	0.964 ± 0.01	0.993 ± 0.00	0.981 ± 0.01	0.949 ± 0.00
pressure: segment	Standard deviation	0.373 ± 0.05	0.918 ± 0.00	0.243 ± 0.04	0.807 ± 0.04	0.415 ± 0.03	0.926 ± 0.00	0.262 ± 0.02	$\mathbf{1.0 \pm 0.00}$
	Inter quartile range	0.665 ± 0.00	0.898 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.498 ± 0.00	0.692 ± 0.02	0.953 ± 0.00	0.529 ± 0.03	$\mathbf{1.0 \pm 0.00}$
	Isolation forests	$\mathbf{0.854 \pm 0.03}$	0.97 ± 0.01	0.853 ± 0.03	$\mathbf{0.855 \pm 0.03}$	$\mathbf{0.75 \pm 0.04}$	0.95 ± 0.01	$\mathbf{0.751 \pm 0.04}$	0.749 ± 0.04
	Local outlier factor	0.151 ± 0.07	0.828 ± 0.01	0.15 ± 0.07	0.151 ± 0.07	0.0 ± 0.00	0.801 ± 0.00	0.0 ± 0.00	0.0 ± 0.00
	Dbscan	0.0 ± 0.00	0.897 ± 0.00	0.0 ± 0.00	0.0 ± 0.00	0.0 ± 0.00	0.9 ± 0.00	0.0 ± 0.00	0.0 ± 0.00
	SAX	0.185 ± 0.00	0.108 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.102 ± 0.00	0.617 ± 0.11	0.946 ± 0.01	0.455 ± 0.11	$\mathbf{1.0 \pm 0.00}$
	AutoMTS	0.854 ± 0.03	0.97 ± 0.01	0.853 ± 0.03	0.855 ± 0.03	0.75 ± 0.04	0.95 ± 0.01	0.751 ± 0.04	0.749 ± 0.04
flow: segment	Standard deviation	0.45 ± 0.05	0.928 ± 0.00	0.292 ± 0.04	$\mathbf{1.0 \pm 0.00}$	0.486 ± 0.02	0.932 ± 0.00	0.322 ± 0.01	$\mathbf{1.0 \pm 0.00}$
	Inter quartile range	$\mathbf{0.929 \pm 0.03}$	0.987 ± 0.01	0.869 ± 0.05	$\mathbf{1.0 \pm 0.00}$	$\mathbf{0.956 \pm 0.01}$	0.991 ± 0.00	$\mathbf{0.982 \pm 0.03}$	0.934 ± 0.03
	Isolation forests	0.889 ± 0.05	0.978 ± 0.01	$\mathbf{0.889 \pm 0.05}$	0.889 ± 0.05	0.855 ± 0.02	0.971 ± 0.00	0.856 ± 0.02	0.854 ± 0.02
	Local outlier factor	0.157 ± 0.05	0.829 ± 0.01	0.157 ± 0.05	0.157 ± 0.05	0.091 ± 0.04	0.819 ± 0.01	0.091 ± 0.04	0.091 ± 0.04
	Dbscan	0.816 ± 0.10	0.962 ± 0.02	0.864 ± 0.14	0.779 ± 0.07	0.773 ± 0.03	0.944 ± 0.01	0.953 ± 0.05	0.651 ± 0.02
	SAX	0.795 ± 0.08	0.966 ± 0.01	0.667 ± 0.11	$\mathbf{1.0 \pm 0.00}$	0.61 ± 0.11	0.945 ± 0.01	0.448 ± 0.11	$\mathbf{1.0 \pm 0.00}$
	AutoMTS	0.929 ± 0.03	0.987 ± 0.01	0.869 ± 0.05	1.0 ± 0.00	0.956 ± 0.01	0.991 ± 0.00	0.982 ± 0.03	0.934 ± 0.03



(a) water flow sensor at Beja WDN (b) water pressure mean at Beja WDN

Fig. 4. Performance of missing imputation methods with varying percentage of point missings planted in time series from heterogeneous sensors.

Figures 4a and 4b extend some of the presented settings, offering a complementary graphical description sensitive to the amount of planted missing values. Generally, the higher the amount of missing observations, the higher the imputation difficulty. These figures highlight the presence of significant performance differences related with the amount of missing observations, further suggesting the relevance of understanding the missing profiles when placing preprocessing decisions. For instance, while random forests is generally a non-competitive method for a small amount of missings, it is the suggested option to impute high amounts of missing observations in water pressure series. This last remark further pinpoints the relevance of considering cross-variable dependencies.

Tables 3 and 4 in appendix provide complementary results on the behavior of both outlier detection and missing imputation methods to handle point and sequential inconsistencies.

4.2 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 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*

Table 2. Performance of missing imputation methods for water pressure and flow sensors from Barreiro and Beja WDNs with planted point and sequential missing values on 2% of observations.

		Barreiro WDN				Beja WDN			
		RMSE	MAE	SMAPE	%	RMSE	MAE	SMAPE	%
pressure: point	Mean	0.051 ± 0.07	0.025 ± 0.02	0.941 ± 0.89	1.00	0.166 ± 0.01	0.154 ± 0.01	4.325 ± 0.21	1.00
	Median	0.051 ± 0.07	0.024 ± 0.02	0.927 ± 0.89	1.00	0.19 ± 0.02	0.124 ± 0.02	3.495 ± 0.52	1.00
	Random sample	0.058 ± 0.07	0.034 ± 0.04	1.268 ± 1.43	1.00	0.23 ± 0.05	0.169 ± 0.06	4.759 ± 1.64	1.00
	Interpolation	0.052 ± 0.08	0.023 ± 0.02	0.897 ± 1.04	1.00	**0.048 ± 0.01**	**0.03 ± 0.01**	0.842 ± 0.17	1.00
	Locf	**0.039 ± 0.07**	**0.019 ± 0.02**	0.747 ± 0.93	1.00	0.067 ± 0.02	0.036 ± 0.01	1.017 ± 0.22	1.00
	Nocb	0.073 ± 0.11	0.03 ± 0.03	1.247 ± 1.53	1.00	0.057 ± 0.02	0.033 ± 0.01	0.923 ± 0.22	1.00
	Moving average	0.047 ± 0.07	0.021 ± 0.02	0.822 ± 0.88	1.00	0.08 ± 0.02	0.042 ± 0.01	1.168 ± 0.27	1.00
	Random forests	0.074 ± 0.07	0.037 ± 0.02	1.425 ± 1.00	1.00	0.173 ± 0.01	0.132 ± 0.01	3.709 ± 0.39	1.00
	EM	0.092 ± 0.06	0.065 ± 0.02	2.422 ± 0.87	1.00	0.231 ± 0.02	0.19 ± 0.02	5.343 ± 0.50	1.00
	Knn	0.059 ± 0.05	0.032 ± 0.02	1.216 ± 0.82	1.00	0.164 ± 0.01	0.132 ± 0.01	3.707 ± 0.41	1.00
	Mice	0.083 ± 0.06	0.046 ± 0.02	1.722 ± 0.95	1.00	0.22 ± 0.02	0.152 ± 0.02	4.304 ± 0.50	1.00
	Amelia	0.095 ± 0.06	0.069 ± 0.02	2.547 ± 0.95	0.98	0.229 ± 0.01	0.187 ± 0.01	5.247 ± 0.37	0.97
	AutoMTS	0.039 ± 0.07	0.019 ± 0.02	0.747 ± 0.93	1.00	0.048 ± 0.01	0.03 ± 0.01	0.842 ± 0.17	1.00
flow: point	Mean	8.89 ± 1.25	7.461 ± 1.31	34.015 ± 6.62	1.00	47.609 ± 5.38	36.619 ± 3.97	48.716 ± 4.41	1.00
	Median	9.061 ± 1.27	7.47 ± 1.35	33.91 ± 7.02	1.00	49.51 ± 6.59	34.352 ± 4.72	45.846 ± 4.99	1.00
	Random sample	11.894 ± 4.25	9.956 ± 4.17	41.997 ± 12.05	1.00	58.912 ± 19.86	46.49 ± 20.79	60.353 ± 29.31	1.00
	Interpolation	**2.801 ± 0.86**	**2.0 ± 0.66**	10.056 ± 4.14	1.00	**16.871 ± 2.52**	**11.96 ± 1.36**	9.655 ± 2.54	1.00
	Locf	4.221 ± 1.22	3.264 ± 0.90	15.714 ± 4.51	1.00	20.677 ± 3.41	14.189 ± 1.94	23.204 ± 3.36	1.00
	Nocb	4.52 ± 1.17	3.484 ± 0.91	16.934 ± 4.57	0.99	20.526 ± 2.83	14.425 ± 1.73	23.604 ± 3.12	1.00
	Moving average	6.68 ± 2.04	5.314 ± 1.63	25.299 ± 6.89	1.00	20.534 ± 3.08	14.683 ± 1.64	23.764 ± 3.29	1.00
	Random forests	10.115 ± 1.86	8.315 ± 1.73	37.123 ± 8.07	1.00	46.03 ± 5.63	34.514 ± 3.76	46.435 ± 3.94	1.00
	EM	12.125 ± 2.88	9.982 ± 2.50	47.819 ± 11.22	1.00	64.264 ± 6.33	49.785 ± 4.78	77.618 ± 6.87	1.00
	Knn	9.771 ± 1.60	8.051 ± 1.46	36.148 ± 7.26	1.00	48.345 ± 5.78	36.213 ± 3.97	48.019 ± 4.41	1.00
	Mice	12.69 ± 2.57	10.433 ± 2.42	48.719 ± 11.95	1.00	72.407 ± 6.63	54.85 ± 5.43	67.096 ± 5.80	1.00
	Amelia	12.548 ± 2.03	10.346 ± 1.76	46.246 ± 8.02	0.98	67.114 ± 5.31	53.254 ± 5.21	75.529 ± 6.72	0.97
	AutoMTS	2.801 ± 0.86	2.0 ± 0.66	10.056 ± 4.14	1.00	16.871 ± 2.52	11.96 ± 1.36	19.655 ± 2.54	1.00
pressure: sequential	Mean	0.025 ± 0.06	0.014 ± 0.02	0.535 ± 0.80	1.00	0.166 ± 0.03	0.156 ± 0.02	4.399 ± 0.66	1.00
	Median	**0.024 ± 0.06**	**0.013 ± 0.02**	0.518 ± 0.80	1.00	0.185 ± 0.07	0.129 ± 0.06	3.653 ± 1.69	1.00
	Random sample	0.035 ± 0.06	0.024 ± 0.03	0.908 ± 1.20	1.00	0.227 ± 0.07	0.174 ± 0.08	4.924 ± 2.23	1.00
	Interpolation	0.03 ± 0.06	0.021 ± 0.04	0.834 ± 1.82	1.00	0.117 ± 0.03	0.09 ± 0.03	2.538 ± 0.92	1.00
	Locf	0.039 ± 0.10	0.029 ± 0.07	1.176 ± 3.10	1.00	0.207 ± 0.08	0.15 ± 0.08	4.228 ± 2.31	1.00
	Nocb	0.029 ± 0.06	0.018 ± 0.02	0.67 ± 0.87	1.00	0.18 ± 0.07	0.123 ± 0.08	3.489 ± 2.17	1.00
	Moving average	0.03 ± 0.07	0.022 ± 0.05	0.875 ± 2.12	0.67	**0.061 ± 0.06**	**0.044 ± 0.05**	1.224 ± 1.42	0.13
	Random forests	0.074 ± 0.09	0.034 ± 0.03	1.35 ± 1.30	1.00	0.189 ± 0.03	0.149 ± 0.03	4.187 ± 0.77	1.00
	EM	0.086 ± 0.05	0.064 ± 0.02	2.371 ± 0.75	1.00	0.227 ± 0.03	0.188 ± 0.03	5.286 ± 0.71	1.00
	Knn	0.049 ± 0.06	0.026 ± 0.02	1.013 ± 0.83	1.00	0.177 ± 0.03	0.145 ± 0.03	4.073 ± 0.79	1.00
	Mice	0.047 ± 0.06	0.027 ± 0.02	0.994 ± 0.81	1.00	0.229 ± 0.03	0.164 ± 0.03	4.633 ± 0.84	1.00
	Amelia	0.082 ± 0.05	0.063 ± 0.02	2.302 ± 0.88	1.00	0.236 ± 0.03	0.194 ± 0.03	5.441 ± 0.74	1.00
	AutoMTS	0.024 ± 0.06	0.013 ± 0.02	0.518 ± 0.80	1.00	0.117 ± 0.03	0.09 ± 0.03	2.538 ± 0.92	1.00
flow: sequential	Mean	8.922 ± 3.13	7.691 ± 3.31	33.079 ± 12.44	1.00	49.262 ± 15.83	38.472 ± 11.40	44.656 ± 7.20	1.00
	Median	8.956 ± 3.00	7.574 ± 3.26	32.503 ± 12.70	1.00	52.412 ± 20.38	38.268 ± 14.57	44.13 ± 10.56	1.00
	Random sample	10.477 ± 4.33	8.774 ± 3.75	37.703 ± 16.27	1.00	61.061 ± 22.51	49.609 ± 20.98	59.152 ± 29.97	1.00
	Interpolation	**5.889 ± 2.26**	**4.855 ± 2.15**	21.426 ± 8.18	1.00	34.086 ± 11.53	25.531 ± 7.72	31.003 ± 8.26	1.00
	Locf	9.482 ± 3.27	7.23 ± 2.73	32.741 ± 11.95	1.00	40.719 ± 13.46	31.17 ± 9.71	37.961 ± 13.23	1.00
	Nocb	8.971 ± 3.37	7.244 ± 3.08	31.694 ± 12.54	1.00	49.136 ± 19.13	37.041 ± 13.07	43.945 ± 12.47	1.00
	Moving average	9.006 ± 3.95	7.262 ± 3.38	33.103 ± 13.17	0.67	**24.745 ± 12.15**	**19.515 ± 9.65**	23.337 ± 12.06	0.13
	Random forests	10.211 ± 2.87	8.412 ± 3.00	35.549 ± 11.01	1.00	49.06 ± 14.73	37.128 ± 10.54	43.422 ± 6.87	1.00
	EM	12.111 ± 3.38	10.346 ± 3.01	47.673 ± 11.92	1.00	69.439 ± 16.02	54.84 ± 12.28	78.177 ± 9.67	1.00
	Knn	10.035 ± 2.98	8.375 ± 3.01	35.499 ± 11.47	1.00	51.228 ± 14.47	39.216 ± 10.56	46.073 ± 6.67	1.00
	Mice	12.611 ± 4.13	10.549 ± 3.47	47.185 ± 14.90	1.00	72.33 ± 13.95	56.269 ± 11.41	64.379 ± 8.45	1.00
	Amelia	12.232 ± 3.02	10.263 ± 2.76	45.0 ± 11.82	1.00	68.419 ± 13.26	55.239 ± 10.25	72.444 ± 7.02	1.00
	AutoMTS	5.889 ± 2.26	4.855 ± 2.15	21.426 ± 8.18	1.00	34.086 ± 11.53	25.531 ± 7.72	31.003 ± 8.26	1.00

type and *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.

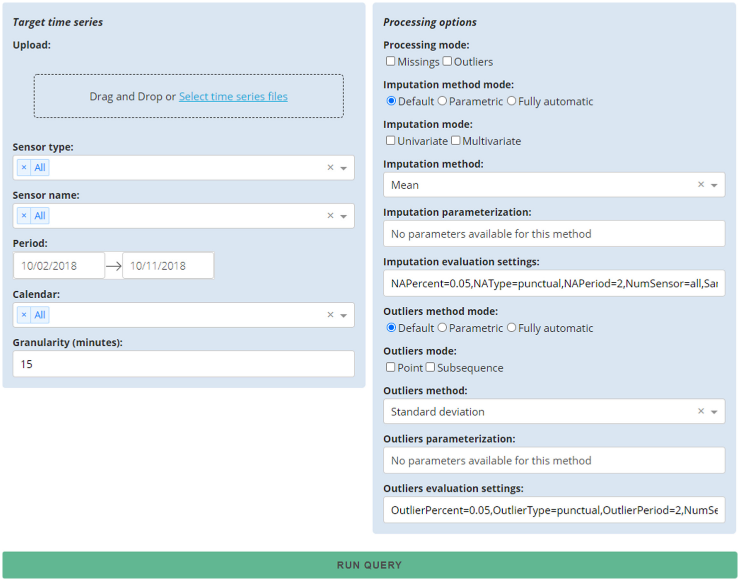


Fig. 5. AutoMTS tool: graphical user interface.

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 (Sect. 3) 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 application will return the original series with the missing values imputed and the outliers detected, together with performance statistics whenever the user opted for generating ground truth by planting artificial inconsistencies. Figure 6 provides a summarized view of the outputs. The user can use interactive zooming and filtering facilities on the displayed series, and access a generated report with the results of the assessment with a similar format as the ones presented along the previous section.

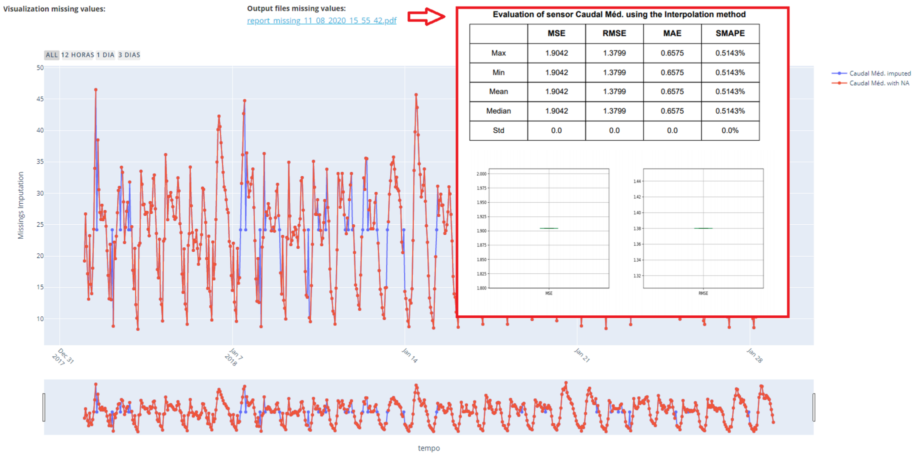


Fig. 6. AutoMTS tool: output overview.

5 Conclusion

This work proposed a methodology, AutoMTS, for the fully-autonomous and quality-driven processing of time series data produced by networks of heterogeneous sensors. AutoMTS is parameter-free and offers strict guarantees of optimality as it places robust principles to assess, hyperparameterize and select state-of-the-art processing methods. To this end, ground truth is produced from conserved series segments in accordance with the eligible error profiles. AutoMTS further provides a comprehensive coverage of state-of-the-art methods for missing imputation, outlier detection and gross-error removal from time series data. AutoMTS implements processing methods able to explore the aided guidance from cross-variable dependencies in the presence of multivariate time series data. In addition, we guarantee the presence of methods able to deal with varying types and amount of missing and outlier values, including both point and serial occurrences of varying duration and recurrence.

The experimental assessment of AutoMTS over two real-world study cases – water distribution network systems with different sampling rates, water consumption patterns and error profiles – confirm the significance of the above contributions. 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, evidencing inherent strengths and limitations to handle diverse error profiles in real-world series data with varying regularities.

This work opens up possibilities for the processing of networks of sensors, particularly those networks that are large in size, heterogeneous in nature, or whose regularities are subjected to significant changes along time. AutoMTS surpasses the need for laborious processing decisions in these contexts, autonomously leveraging time series data quality for subsequent analytics.

As future work, we aim to extend the proposed methodology to guarantee the online processing of time series data streams for real-time monitoring tasks.

Acknowledgements. The authors thank the support of *Câmara Municipal do Barreiro*, *Câmara Municipal de Beja*, *Infraquinta* and *Câmara Municipal de Lisboa*. This work is further supported by national funds through *Fundação para a Ciência e Tecnologia* under projects WISDOM (DSAIPA/DS/0089/2018), ILU (DSAIPA/DS/0111/ 2018), FARO (PTDC/EGE-ECO/30535/2017), and the INESC-ID (UIDB/50021/2020) and CeMAT/IST-ID (UID/Multi/04621/2019) pluriannuals.

A Supplementary Material

Table 3. Performance of outlier detection methods for water pressure and flow sensors from Barreiro and Beja WDNs with planted point and sequential outliers on 2% of observations.

		Barreiro WDN				Beja WDN			
		F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall
pressure: point	Standard deviation	0.189 ± 0.11	0.976 ± 0.00	0.149 ± 0.09	$\mathbf{0.262 \pm 0.13}$	0.132 ± 0.06	0.982 ± 0.00	0.072 ± 0.03	$\mathbf{1.0 \pm 0.00}$
	Inter quartile range	0.272 ± 0.00	0.896 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.157 ± 0.00	0.058 ± 0.04	0.981 ± 0.00	0.031 ± 0.02	0.767 ± 0.42
	Isolation forests	$\mathbf{0.322 \pm 0.00}$	0.918 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.192 ± 0.00	0.337 ± 0.01	0.922 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.202 ± 0.00
	Local outlier factor	$\mathbf{0.322 \pm 0.00}$	0.918 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.192 ± 0.00	0.0 ± 0.00	0.881 ± 0.00	0.0 ± 0.00	0.0 ± 0.00
	Dbscan	0.0 ± 0.00	0.979 ± 0.00	0.0 ± 0.00	0.0 ± 0.00	0.0 ± 0.00	0.98 ± 0.00	0.0 ± 0.00	0.0 ± 0.00
	SAX	0.038 ± 0.00	0.023 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.019 ± 0.00	$\mathbf{0.684 \pm 0.29}$	0.941 ± 0.11	$\mathbf{1.0 \pm 0.00}$	0.582 ± 0.29
	AutoMTS	0.322 ± 0.00	0.918 ± 0.00	1.0 ± 0.00	0.192 ± 0.00	0.684 ± 0.29	0.941 ± 0.11	1.0 ± 0.00	0.582 ± 0.29
flow: point	Standard deviation	0.615 ± 0.11	0.989 ± 0.00	0.454 ± 0.12	$\mathbf{1.0 \pm 0.00}$	$\mathbf{0.779 \pm 0.03}$	0.991 ± 0.00	0.772 ± 0.04	$\mathbf{0.788 \pm 0.02}$
	Inter quartile range	$\mathbf{0.787 \pm 0.10}$	0.993 ± 0.00	0.659 ± 0.13	$\mathbf{1.0 \pm 0.00}$	0.445 ± 0.01	0.95 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.287 ± 0.01
	Isolation forests	0.321 ± 0.00	0.918 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.191 ± 0.00	0.332 ± 0.00	0.92 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.199 ± 0.00
	Local outlier factor	0.309 ± 0.02	0.917 ± 0.00	0.964 ± 0.07	0.184 ± 0.01	0.089 ± 0.02	0.891 ± 0.00	0.269 ± 0.07	0.054 ± 0.01
	Dbscan	0.602 ± 0.06	0.976 ± 0.00	0.959 ± 0.12	0.439 ± 0.04	0.415 ± 0.02	0.949 ± 0.00	0.912 ± 0.06	0.269 ± 0.01
	SAX	0.357 ± 0.12	0.916 ± 0.05	$\mathbf{1.0 \pm 0.00}$	0.224 ± 0.09	0.504 ± 0.08	0.959 ± 0.01	$\mathbf{1.0 \pm 0.00}$	0.341 ± 0.07
	AutoMTS	0.787 ± 0.10	0.993 ± 0.00	0.659 ± 0.13	1.0 ± 0.00	0.779 ± 0.03	0.991 ± 0.00	0.772 ± 0.04	0.788 ± 0.02
pressure: segment	Standard deviation	0.141 ± 0.12	0.977 ± 0.00	0.114 ± 0.10	$\mathbf{0.189 \pm 0.15}$	0.136 ± 0.06	$\mathbf{0.981 \pm 0.00}$	0.074 ± 0.03	$\mathbf{0.967 \pm 0.18}$
	Inter quartile range	0.253 ± 0.00	0.895 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.145 ± 0.00	0.054 ± 0.05	0.98 ± 0.00	0.028 ± 0.03	0.733 ± 0.44
	Isolation forests	$\mathbf{0.301 \pm 0.00}$	0.917 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.177 ± 0.00	$\mathbf{0.341 \pm 0.01}$	0.922 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.205 ± 0.00
	Local outlier factor	0.3 ± 0.00	0.917 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.177 ± 0.00	0.0 ± 0.00	0.881 ± 0.00	0.0 ± 0.00	0.0 ± 0.00
	Dbscan	0.0 ± 0.00	0.981 ± 0.00	0.0 ± 0.00	0.0 ± 0.00	0.0 ± 0.00	0.98 ± 0.00	0.0 ± 0.00	0.0 ± 0.00
	SAX	0.035 ± 0.00	0.021 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.018 ± 0.00	0.64 ± 0.31	0.925 ± 0.13	0.976 ± 0.13	0.561 ± 0.31
	AutoMTS	0.301 ± 0.00	0.917 ± 0.00	1.0 ± 0.00	0.177 ± 0.00	0.341 ± 0.01	0.922 ± 0.00	1.0 ± 0.00	0.205 ± 0.00
flow: segment	Standard deviation	0.653 ± 0.11	0.991 ± 0.00	0.494 ± 0.12	$\mathbf{1.0 \pm 0.00}$	$\mathbf{0.787 \pm 0.03}$	0.992 ± 0.00	0.778 ± 0.05	$\mathbf{0.799 \pm 0.02}$
	Inter quartile range	$\mathbf{0.827 \pm 0.07}$	0.995 ± 0.00	0.711 ± 0.10	$\mathbf{1.0 \pm 0.00}$	0.441 ± 0.03	0.948 ± 0.01	$\mathbf{1.0 \pm 0.00}$	0.283 ± 0.02
	Isolation forests	0.3 ± 0.00	0.917 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.177 ± 0.00	0.336 ± 0.00	0.92 ± 0.00	$\mathbf{1.0 \pm 0.00}$	0.202 ± 0.00
	Local outlier factor	0.294 ± 0.01	0.916 ± 0.00	0.981 ± 0.04	0.173 ± 0.01	0.093 ± 0.03	0.891 ± 0.00	0.277 ± 0.10	0.056 ± 0.02
	Dbscan	0.597 ± 0.06	0.976 ± 0.00	0.983 ± 0.09	0.429 ± 0.04	0.416 ± 0.03	0.948 ± 0.00	0.91 ± 0.07	0.27 ± 0.01
	SAX	0.366 ± 0.12	0.928 ± 0.04	$\mathbf{1.0 \pm 0.00}$	0.231 ± 0.09	0.508 ± 0.09	0.958 ± 0.01	$\mathbf{1.0 \pm 0.00}$	0.346 ± 0.08
	AutoMTS	0.827 ± 0.07	0.995 ± 0.00	0.711 ± 0.10	1.0 ± 0.00	0.787 ± 0.03	0.992 ± 0.00	0.778 ± 0.05	0.799 ± 0.02

Table 4. Performance of imputation methods for water pressure and flow sensors from Barreiro and Beja WDNs with planted point and sequential missing values on 10% of observations.

		Barreiro WDN				Beja WDN			
		RMSE	MAE	SMAPE	%	RMSE	MAE	SMAPE	%
pressure: point	Mean	0.064 ± 0.05	0.025 ± 0.01	0.933 ± 0.43	1.00	0.166 ± 0.00	0.154 ± 0.00	4.335 ± 0.08	1.00
	Median	0.065 ± 0.05	0.024 ± 0.01	0.921 ± 0.42	1.00	0.192 ± 0.01	0.125 ± 0.01	3.517 ± 0.18	1.00
	Random sample	0.07 ± 0.05	0.033 ± 0.03	1.247 ± 0.91	1.00	0.23 ± 0.04	0.168 ± 0.05	4.737 ± 1.45	1.00
	Interpolation	0.058 ± 0.05	0.019 ± 0.01	0.731 ± 0.37	0.99	**0.05 ± 0.00**	**0.031 ± 0.00**	0.857 ± 0.06	1.00
	Locf	0.062 ± 0.05	0.021 ± 0.01	0.796 ± 0.36	0.99	0.068 ± 0.01	0.036 ± 0.00	0.999 ± 0.10	1.00
	Nocb	0.062 ± 0.06	0.021 ± 0.01	0.808 ± 0.43	0.99	0.065 ± 0.01	0.035 ± 0.00	0.973 ± 0.09	1.00
	Moving average	**0.055 ± 0.04**	**0.018 ± 0.01**	0.712 ± 0.33	0.99	0.078 ± 0.01	0.04 ± 0.00	1.113 ± 0.10	1.00
	Random forests	0.089 ± 0.04	0.037 ± 0.01	1.414 ± 0.41	1.00	0.175 ± 0.01	0.133 ± 0.01	3.752 ± 0.18	1.00
	EM	0.098 ± 0.03	0.065 ± 0.01	2.404 ± 0.28	1.00	0.228 ± 0.01	0.188 ± 0.01	5.275 ± 0.23	1.00
	Knn	0.069 ± 0.04	0.032 ± 0.01	1.19 ± 0.38	1.00	0.166 ± 0.01	0.135 ± 0.01	3.806 ± 0.15	1.00
	Mice	0.084 ± 0.04	0.043 ± 0.01	1.574 ± 0.49	1.00	0.221 ± 0.01	0.154 ± 0.01	4.337 ± 0.24	1.00
	Amelia	0.104 ± 0.04	0.068 ± 0.01	2.508 ± 0.54	0.90	0.234 ± 0.01	0.19 ± 0.01	5.342 ± 0.21	0.90
	AutoMTS	0.055 ± 0.04	0.018 ± 0.01	0.712 ± 0.33	0.99	0.05 ± 0.00	0.031 ± 0.00	0.857 ± 0.06	1.00
flow: point	Mean	8.878 ± 0.58	7.331 ± 0.57	32.584 ± 2.57	1.00	47.573 ± 2.71	36.25 ± 1.68	48.104 ± 1.96	1.00
	Median	8.976 ± 0.57	7.312 ± 0.58	32.392 ± 2.70	1.00	49.678 ± 3.18	34.009 ± 2.14	45.249 ± 2.21	1.00
	Random sample	12.186 ± 4.31	10.31 ± 4.14	42.831 ± 12.28	1.00	59.523 ± 17.85	46.715 ± 19.59	60.721 ± 28.42	1.00
	Interpolation	**2.927 ± 0.32**	**2.055 ± 0.25**	9.821 ± 1.23	0.99	**17.246 ± 1.34**	**12.137 ± 0.79**	19.849 ± 1.08	1.00
	Locf	4.602 ± 0.47	3.49 ± 0.34	16.146 ± 1.65	0.99	21.231 ± 1.60	14.692 ± 0.85	23.877 ± 1.62	1.00
	Nocb	4.878 ± 0.55	3.649 ± 0.35	16.924 ± 1.43	0.99	21.133 ± 1.57	14.565 ± 1.01	24.047 ± 1.59	1.00
	Moving average	7.021 ± 0.79	5.48 ± 0.59	25.245 ± 2.28	0.99	21.595 ± 1.46	15.242 ± 0.83	24.077 ± 1.57	1.00
	Random forests	10.086 ± 0.83	8.085 ± 0.78	35.154 ± 3.40	1.00	47.097 ± 3.20	34.74 ± 2.29	46.244 ± 2.00	1.00
	EM	12.08 ± 0.85	9.806 ± 0.72	46.242 ± 3.74	1.00	66.136 ± 3.22	50.566 ± 2.36	79.039 ± 3.73	1.00
	Knn	9.639 ± 0.73	7.857 ± 0.66	34.557 ± 2.72	1.00	47.803 ± 2.11	35.747 ± 1.66	47.494 ± 2.00	1.00
	Mice	13.291 ± 1.05	10.893 ± 1.03	49.257 ± 5.83	1.00	72.356 ± 5.04	54.984 ± 4.83	66.829 ± 5.40	1.00
	Amelia	12.11 ± 0.94	9.78 ± 0.82	42.972 ± 4.07	0.90	65.755 ± 2.65	51.807 ± 2.21	71.872 ± 3.35	0.90
	AutoMTS	2.927 ± 0.32	2.055 ± 0.25	9.821 ± 1.23	0.99	17.246 ± 1.34	12.137 ± 0.79	19.849 ± 1.08	1.00
pressure: sequential	Mean	0.021 ± 0.02	0.013 ± 0.00	0.464 ± 0.13	1.00	**0.169 ± 0.01**	0.157 ± 0.00	4.416 ± 0.13	1.00
	Median	**0.02 ± 0.02**	**0.012 ± 0.00**	0.427 ± 0.13	1.00	0.197 ± 0.01	**0.13 ± 0.01**	3.684 ± 0.34	1.00
	Random sample	0.03 ± 0.03	0.022 ± 0.03	0.816 ± 0.95	1.00	0.23 ± 0.04	0.168 ± 0.05	4.729 ± 1.31	1.00
	Interpolation	0.03 ± 0.06	0.021 ± 0.04	0.817 ± 1.58	1.00	0.207 ± 0.01	0.164 ± 0.01	4.618 ± 0.39	1.00
	Locf	0.043 ± 0.11	0.032 ± 0.08	1.265 ± 3.34	1.00	0.24 ± 0.03	0.177 ± 0.04	4.998 ± 1.15	1.00
	Nocb	0.022 ± 0.02	0.014 ± 0.01	0.533 ± 0.20	1.00	0.237 ± 0.03	0.175 ± 0.03	4.955 ± 0.91	1.00
	Moving average	0.031 ± 0.07	0.022 ± 0.05	0.887 ± 2.12	0.12	0.061 ± 0.07	0.044 ± 0.05	1.225 ± 1.42	0.03
	Random forests	0.064 ± 0.04	0.027 ± 0.01	1.013 ± 0.33	1.00	0.189 ± 0.01	0.147 ± 0.01	4.155 ± 0.24	1.00
	EM	0.078 ± 0.01	0.061 ± 0.01	2.21 ± 0.19	1.00	0.23 ± 0.01	0.189 ± 0.01	5.326 ± 0.22	1.00
	Knn	0.045 ± 0.02	0.024 ± 0.00	0.871 ± 0.16	1.00	0.178 ± 0.01	0.146 ± 0.01	4.112 ± 0.19	1.00
	Mice	0.056 ± 0.02	0.032 ± 0.01	1.167 ± 0.31	1.00	0.229 ± 0.01	0.162 ± 0.01	4.587 ± 0.30	1.00
	Amelia	0.073 ± 0.02	0.056 ± 0.01	2.026 ± 0.41	0.89	0.238 ± 0.01	0.194 ± 0.01	5.433 ± 0.33	0.89
	AutoMTS	0.02 ± 0.02	0.012 ± 0.00	0.427 ± 0.13	1.00	0.169 ± 0.01	0.157 ± 0.00	4.416 ± 0.13	1.00
flow: sequential	Mean	8.839 ± 1.09	7.415 ± 1.25	32.556 ± 4.45	1.00	53.752 ± 13.36	40.987 ± 9.06	46.883 ± 3.85	1.00
	Median	8.884 ± 0.98	7.321 ± 1.20	32.073 ± 4.38	1.00	56.946 ± 18.43	41.172 ± 12.99	46.95 ± 8.74	1.00
	Random sample	11.589 ± 3.59	9.644 ± 3.29	41.054 ± 11.91	1.00	65.0 ± 21.24	51.602 ± 19.29	60.692 ± 29.12	1.00
	Interpolation	10.273 ± 1.21	8.493 ± 1.13	37.44 ± 4.24	1.00	40.72 ± 10.25	31.124 ± 7.20	37.789 ± 6.21	1.00
	Locf	12.074 ± 2.19	9.705 ± 1.90	43.785 ± 8.47	1.00	44.045 ± 11.33	33.286 ± 8.19	40.576 ± 11.90	1.00
	Nocb	11.092 ± 2.30	9.055 ± 1.90	39.8 ± 8.14	1.00	47.163 ± 15.20	37.099 ± 12.08	44.922 ± 12.72	1.00
	Moving average	**8.803 ± 3.96**	**7.065 ± 3.39**	31.923 ± 13.69	0.12	**24.396 ± 12.08**	**19.166 ± 9.47**	23.783 ± 12.51	0.03
	Random forests	9.913 ± 1.07	7.961 ± 1.05	34.648 ± 3.79	1.00	53.638 ± 13.67	39.78 ± 9.41	45.504 ± 4.33	1.00
	EM	12.193 ± 1.44	9.957 ± 1.25	46.583 ± 5.41	1.00	73.736 ± 13.64	57.226 ± 10.29	79.993 ± 5.61	1.00
	Knn	9.431 ± 1.13	7.696 ± 1.17	33.592 ± 4.34	1.00	54.104 ± 12.18	40.735 ± 8.19	47.224 ± 3.77	1.00
	Mice	12.148 ± 1.79	9.931 ± 1.55	44.663 ± 7.75	1.00	75.86 ± 6.55	58.838 ± 5.76	67.721 ± 4.48	1.00
	Amelia	12.125 ± 1.38	9.771 ± 1.29	43.056 ± 5.90	0.89	70.062 ± 8.17	55.106 ± 6.29	70.747 ± 3.56	0.89
	AutoMTS	8.839 ± 1.09	7.415 ± 1.25	32.556 ± 4.45	1.00	40.72 ± 10.25	31.124 ± 7.20	37.789 ± 6.21	1.00

References

1. Gill, P., Jain, N., Nagappan, N.: Understanding network failures in data centers: measurement, analysis, and implications. In: Proceedings of the ACM SIGCOMM 2011 Conference, SIGCOMM 2011, pp. 350–361. Association for Computing Machinery, New York (2011). ISBN 9781450307970. https://doi.org/10.1145/2018436.2018477
2. Jain, R.K.: A state space model-based method of seasonal adjustment. Monthly Lab. Rev. **124**, 37 (2001)
3. We, W.W.S.: Time series analysis. In: The Oxford Handbook of Quantitative Methods in Psychology, vol. 2 (2006)
4. Osman, M.S., Abu-Mahfouz, A.M., Page, P.R.: A survey on data imputation techniques: water distribution system as a use case. IEEE Access **6**, 63279–63291 (2018). ISSN 2169–3536. https://doi.org/10.1109/ACCESS.2018.2877269
5. Stekhoven, D.J., Bühlmann, P.: MissForest–non-parametric missing value imputation for mixed-type data. Bioinformatics **28**(1), 112–118 (2011). ISSN 1367–4803. https://doi.org/10.1093/bioinformatics/btr597
6. van Buuren, S., Groothuis-Oudshoorn, K.: MICE: multivariate imputation by chained equations in R. J. Stat. Softw. **45**(3), 1–67 (2011). ISSN 1548–7660. https://doi.org/10.18637/jss.v045.i03
7. Ilin, A., Raiko, T.: Practical approaches to principal component analysis in the presence of missing values. J. Mach. Learn. Res. **11**, 1957–2000 (2010). ISSN 1532–4435. http://dl.acm.org/citation.cfm?id=1756006.1859917
8. Garg, L., Dauwels, J., Earnest, A., Leong, K.P.: Tensor-based methods for handling missing data in quality-of-life questionnaires. IEEE JBHI **18**(5), 1571–1580 (2014). ISSN 2168–2208. https://doi.org/10.1109/JBHI.2013.2288803
9. Luo, Y., Cai, X., Zhang, Y., Xu, J., et al.: Multivariate time series imputation with generative adversarial networks. In: Advances in Neural Information Processing Systems, pp. 1596–1607 (2018)
10. Moritz, S., et al.: Comparison of different Methods for Univariate Time Series Imputation in R, October 2015
11. Zeileis, A., Grothendieck, G.: zoo: S3 infrastructure for regular and irregular time series. J. Stat. Softw. Articles **14**(6), 1–27 (2005). ISSN 1548–7660. https://doi.org/10.18637/jss.v014.i06. https://www.jstatsoft.org/v014/i06
12. Kowarik, A., Templ, M.: Imputation with the R package VIM. J. Stat. Softw. Articles **74**(7), 1–16 (2016). ISSN 1548–7660. https://doi.org/10.18637/jss.v074.i07. https://www.jstatsoft.org/v074/i07
13. Quevedo, J., et al.: Validation and reconstruction of flow meter data in the barcelona water distribution network. Control Eng. Practice **18**(6), 640–651 (2010). ISSN 0967–0661. https://doi.org/10.1016/j.conengprac.2010.03.003. http://www.sciencedirect.com/science/article/pii/S0967066110000791
14. Barrela, R., et al.: Data reconstruction of ow time series in water distribution systems - a new method that accommodates multiple seasonality. J. Hydroinformatics **19**(2), 238–250 (2016). ISSN 1464–7141
15. Hawkins, D.M.: Identification of outliers. Monographs on applied probability and statistics. Chapman and Hall, London [u.a.] (1980). ISBN 041221900X. http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+02435757X&sourceid=fbw_bibsonomy
16. Chan, W.-S.: Understanding the effect of time series outliers on sample autocorrelations. TEST **4**(1), 179–186 (1995)

17. Aggarwal, C.C.: Outlier analysis. In: Data Mining, pp. 237–263. Springer, Heidelberg (2015)
18. Breunig, M., Kriegel, H.-P., Ng, R., Sander, J.: LOF: Identifying density-based local outliers. **29**, 93–104 (2000). https://doi.org/10.1145/342009.335388
19. Liu, F.T., Ting, K.M., Zhou, Z.: Isolation forest. In: ICDM 2008: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422. IEEE Computer Society (2008)
20. Chen, C., Liu, L.-M.: Joint estimation of model parameters and outlier effects in time series. J. Am. Stat. Assoc. **88**(421), 284–297 (1993). https://doi.org/10.1080/01621459.1993.10594321
21. Gupta, M., Gao, J., Aggarwal, C.C., Han, J.: Outlier detection for temporal data: a survey. IEEE Trans. Knowl. Data Eng. **26**(9), 2250–2267 (2014). ISSN 2326–3865. https://doi.org/10.1109/TKDE.2013.184
22. Guo, Y., et al.: Multidimensional time series anomaly detection: a GRU-based Gaussian mixture variational autoencoder approach. In: Asian Conference on Machine Learning, pp. 97–112 (2018)
23. Jagadish, H.V., Koudas, N., Muthukrishnan, S.: Mining deviants in a time series database. In: Proceedings of the 25th International Conference on Very Large Data Bases, VLDB 1999, pp. 102–113. Morgan Kaufmann Publishers Inc., San Francisco (1999). ISBN 1-55860-615-7. URL http://dl.acm.org/citation.cfm?id=645925.758373
24. Keogh, E., Lin, J., Lee, S.-H., Van Herle, H.: Finding the most unusual time series subsequence: algorithms and applications. Knowl. Inf. Syst. **11**(1), 1–27 (2007.) ISSN 0219–3116. https://doi.org/10.1007/s10115-006-0034-6
25. Chen, X.-Y., Zhan, Y.-Y.: Multi-scale anomaly detection algorithm based on infrequent pattern of time series. J. Comput. Appl. Math. **214**(1), 227–237 (2008). ISSN 0377–0427. https://doi.org/10.1016/j.cam.2007.02.027. http://www.sciencedirect.com/science/article/pii/S0377042707001100
26. Keogh, E., Lin, J., Fu, A.: Hot sax: efficiently finding the most unusual time series subsequence. In: Fifth IEEE International Conference on Data Mining (ICDM 2005), p. 8. IEEE (2005)
27. Toshniwal, D., Yadav, S.: Adaptive outlier detection in streaming time series. Proc. Int. Conf. Asia Agric. Animal ICAAA Hong Kong **13**, 186–192 (2011)
28. Fatehi, A., Huang, B.: Kalman filtering approach to multi-rate information fusion in the presence of irregular sampling rate and variable measurement delay. J. Process Control **53**, 15–25 (2017)
29. Jin, X.-B., Du, J.-J., Bao, J.: Target tracking of a linear time invariant system under irregular sampling. Int. J. Adv. Robot. Syst. **9**(5), 219 (2012)
30. Rayana, S., Akoglu, L.: Less is more: building selective anomaly ensembles. ACM Trans. Knowl. Discov. Data (TKDD) **10**(4), 1–33 (2016)
31. Zimek, A., Campello, R.J.G.B., Sander, J.: Ensembles for unsupervised outlier detection: challenges and research questions a position paper. ACM Sigkdd Explor. Newslett. **15**(1), 11–22 (2014)
32. Li, L., Zhang, J., Wang, Y., Ran, B.: Missing value imputation for traffic-related time series data based on a multi-view learning method. IEEE Trans. Intell. Transp. Syst. **20**(8), 2933–2943 (2018)
33. Oehmcke, S., Zielinski, O., Kramer, O.: kNN ensembles with penalized DTW for multivariate time series imputation. In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 2774–2781. IEEE (2016)

34. Böhm, C., Haegler, K., Müller, N.S., Plant, C.: Coco: coding cost for parameter-free outlier detection. In: Proceedings of the 15th ACM SIGKDD IC on Knowledge Discovery and Data Mining, KDD 2009, pp. 149–158. Association for Computing Machinery, New York (2009). ISBN 9781605584959. https://doi.org/10.1145/1557019.1557042
35. Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240 (2006)
36. Snoek, J., Larochelle, H., Adams, R.P.: Practical Bayesian optimization of machine learning algorithms. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 2951–2959. Curran Associates Inc. (2012). http://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms.pdf