International Congress on Environmental Modelling and Software

9th International Congress on Environmental Modelling and Software - Ft. Collins, Colorado, USA - June 2018

Jun 26th, 9:00 AM - 10:20 AM

Neural Network for Tuning-Friendly Automatic Outlier Detection in Water Quality Time Series

Cyril Garneau
Université Laval, cyril.garneau.1@ulaval.ca

Peter A. Vanrolleghem
Université Laval, Peter.Vanrolleghem@gci.ulaval.ca

iEMSs International Congress on Environmental Modelling and Software

9th International Congress on Environmental Modelling and Software
Fort Collins, Colorado, USA, Mazdak Arabi, Olaf David, Jack Carlson, Daniel P. Ames (Eds.)
https://scholarsarchive.byu.edu/iemssconference/2018/

Neural Network for Tuning-Friendly Automatic Outlier Detection in Water Quality Time Series

Cyril Garneau[a] and Peter A. Vanrolleghem[a]
[a]modelEAU, Université Laval, Québec, QC, Canada (cyril.garneau.1@ulaval.ca)

Abstract: An Auto-Encoder Neural Network for Time Series (AENNTS) was developed to detect outliers in environmental time series that are subject to difficult measurement conditions leading to data quality problems. The neural network aims at replicating a time series as closely as possible, but with an internal structure which enforces some data compression, and thus imprecise replication. To identify outliers an exponentially weighted moving average (EWMA) and a mean absolute deviation (MAD) are then calculated on the basis of the residuals between the original time series and its AENNTS replication. The method requires only seven parameters, which could all be fixed to their default value in this application. The ratio of outliers to accepted data ranged from 1% to 5% across the various time series tested, which was deemed acceptable and similar to other, more difficult to tune methods.

***Keywords**: Data quality evaluation; fault detection; automated monitoring systems; time series analysis; data validation environmental time series.*

1 INTRODUCTION

With the increasing presence of sensors to monitor environmental systems, that are often problematic in terms of fouling and other data quality issues, outlier detection procedures are necessary to tackle the huge amount of data generated. Furthermore, due to the increasingly large number of large time series to validate, manual validation of time series is no longer possible and automatic methods must be developed. Cleaned time series can afterwards be filtered, screened for fault detection and then used for process control, modelling or decision support.

Outliers from time series exhibiting stable signals can easily be removed by tracking the local mean and standard deviation σ and accepting all data within, e.g., 3σ from the mean (Montgomery, 2009). However, environmental time series often exhibit large discontinuities or highly unstable behaviour, which requires more advanced outlier detection methods. An earlier advanced method that involves forecasting an incoming data point and comparing the actual measure to the forecast has been used to accept data or reject it as an outlier (Alferes and Vanrolleghem, 2016). This method proved valuable on dynamic time series but required a relatively complex method tuning procedure that had to be repeated for each sensor and each measurement location.

Automatic outlier detection is a challenge in all data-rich environments and innovative approaches have been pursued. Hawkins et al. (2002) applied a Replicator Neural Network to network intrusion detection and breast cancer diagnosis. In both cases, the objective was to have the RNN reproduce the original dataset as well as possible. By tracking the replication error, it was then possible to identify outliers. Chen et al. (2017) applied a similar methodology to various datasets collected from biology, network intrusion and geography. Furthermore, problems such as denoising or anomaly detection are similar to outlier detection and were solved by autoencoder (Martinelli et al., 2004; Sakurada and Yairi, 2014; Xiong and Zuo, 2016). However, it is worth noting that all these applications were performed on multivariate datasets and no example was found on a univariate time series.

In this work, an automatic outlier detection based on neural networks is presented for filtering univariate time series. This method is fast to compute, tuning-friendly and is applicable to a wide range of water quality sensors.

2 MATERIAL AND METHODS

The Auto-Encoder Neural Network for Time Series (AENNTS) builds a constrained neural network to reproduce a set of original data. The error of replication is then tracked to identify poorly reconstructed points and flag them as outliers. Outliers are flagged on the basis of a method that combines an exponentially weighted average (EWMA) that tracks the average residual and a mean absolute deviation (MAD) that tracks the variability of the residual. The EWMA + MAD method was taken from Montgomery (2009).

2.1 The AENNTS method

The autoencoder neural network for time series aims to reproduce its own inputs. A record, i.e. a vector of data points provided to the neural network for replication, consists of n consecutives data points in time. A time series of N data points therefore allows for $N - n$ individual records for training the neural network. In this work, a value of $n = 5$ was used, although other values could be used. Therefore, for a large number of datapoints (typically over 100 000), the number of records is approximately the number of individual datapoints.

The neural network used is a feed forward NN containing three hidden layers and one output layer (Figure 1). In order to enforce the compression of information, the middle hidden layer should have less than n neurons. Having 5 – 3 – 5 neurons in each hidden layer gave the results presented in this work. The output layer must have n neurons.

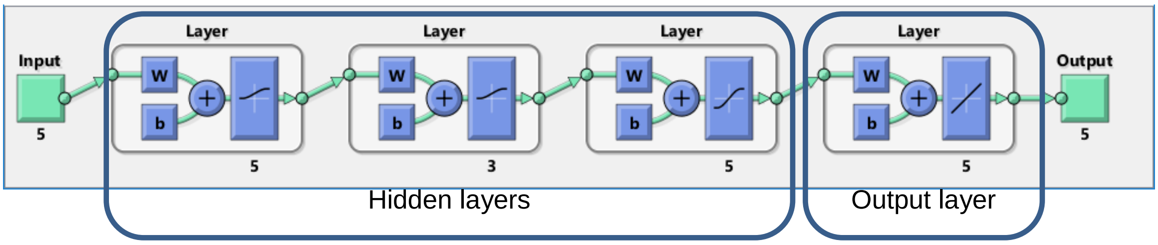

Figure 1. Layout of the neural network for outlier detection.

The training set for the neural network was fixed at 95% of the available records while 3% were kept for validation of the network and 2% for testing. This choice is justified by the fact that the number of records is high (thousands of records) and that the only objective of the neural network is to reproduce its training set.

A trained neural network provides n estimations of each data point. The average of these estimations was taken as the replicated time series. Outliers are then identified by filtering the residual (observed value minus AENNTS-replicated value) through an exponentially weighted moving average (EWMA, equation 1).

$$EWMA_1 = x_1$$
$$EWMA_i = (1 - \alpha) \times EWMA_{i-1} + \alpha \times x_i \quad (1)$$

where $EWMA$ is the exponentially weighted moving average at time i, α is a parameter between 0 and 1 and x_i is the residual at time i. A large value of α gives the most weight on the last residual while a small value of α gives the most weight to the moving average. This will be a tuning parameter.

The mean absolute deviation (MAD) is computed with a similar equation (Eq. 2), but it tracks the mean absolute error rather than the last observed residual.

$$MAD_1 = 0$$
$$MAD_i = (1-\beta) \times MAD_{i-1} + \beta \times (|EWMA_{i-1} - x_{i-1}|) \quad (2)$$

where MAD_i is the mean absolute deviation at time i and β is a parameter between 0 and 1 where a large value gives the most weight to the error between the mean and the last residual observed while a small value gives the most weight to the last computed MAD. This will be a tuning parameter.

Based on the EWMA and the MAD, a data point is accepted if it complies with equation 3.

$$EWMA_i - \sigma \times MAD_i < x_{i+1} < EWMA_i + \sigma \times MAD_i \quad (3)$$

where the σ parameter is the tolerance to the variation of noise level.

2.2 Sample datasets

Two datasets were used to assess the performances of the filter. The first dataset consists of conductivity (Hach D3725E2T) and turbidity measurements (Hach Solitax) at the inlet of a small water resource recovery facility (WRRF). The filtering of the dataset is challenging due to a fast pumping sequence that brings water from two distinct sectors with very different conductivities.

The second dataset comes from the deployment of an ammonium probe (ammo::lyser from S::CAN) in a pilot WRRF. This dataset was also subject to regular cleaning with compressed air and maintenance by the operators.

3 RESULTS

3.1 Parameters of the AENNTS method

The full procedure requires fixing seven parameters, as shown on Table 1. These values were found from experience and from manual tuning. Important to note is that all computations presented in this paper used the default values for all parameters in Table 1.

Table 1. Parameters of the AENNTS method and proposed default values. All computations presented in the current paper used these default values.

PARAMETER NAME	PROPOSED DEFAULT VALUE	DEFINITION
$\boldsymbol{n}$	5	Number of datapoints per record
$\boldsymbol{L1 - L2 - L3}$	5 – 3 – 5	Number of neurons in the three hidden layers
$\boldsymbol{\alpha}$	0.2	Weight of the EWMA
$\boldsymbol{\beta}$	0.2	Weight of the MAD
$\boldsymbol{\sigma}$	3	Accepted variance of the residuals

Experience further suggested that the ratio of outliers was most sensitive to σ. Changing its value from 3 to 2 would identify up to twice as many outliers. In addition, it must be remembered that the parameter $L2$ must be lower than n to ensure the signal to be compressed. If $L2 \geq n$, the neural network can, in theory, replicate the original signal with arbitrary precision.

3.2 Performance of the AENNTS method

The first WRRF dataset is particularly challenging due to a pumping sequence that changes the origin of the wastewater on a regular basis (around 20 minutes in the data series shown in Figure 2). The two sources of wastewater that arrive at the measurement location have different conductivities. The outlier detection must therefore not flag too many outliers when the water quality changes. Furthermore, starting around noon, on Figure 2, an automatic cleaning with compressed air happened during measurement, pushing the conductivity down to 20-30 µS/cm. It can be observed that the outlier detection filter was able to remove most of these artifacts from the time series. In summary, the method identified 5.4% of the values in the time series as outlier, which is similar to what other, less practical methods would have detected. Therefore, despite an apparent large number of outliers on Figure 2, the cleaned signal did not loose significant information.

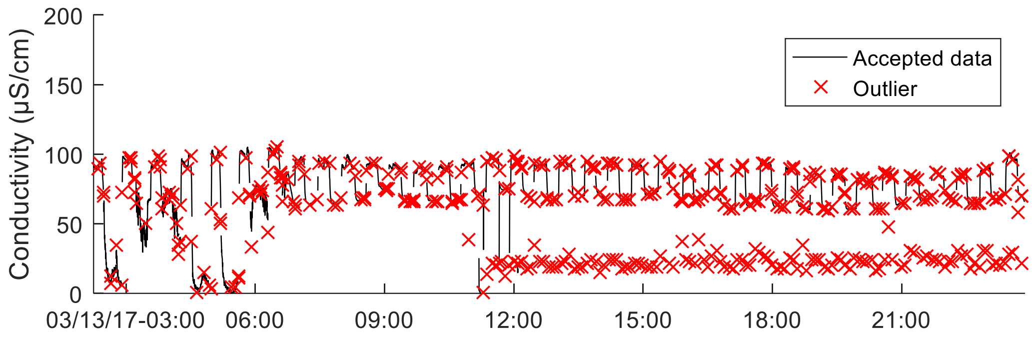

Figure 2. Result of the outlier detection procedure on the conductivity of the wastewater influent of a WRRF. Datapoints that were identified as outliers are replaced by red "x" and break the continuous line of accepted data.

When applied to the ammonium sensor of the pilot WRRF, the filter detected most of the outliers that are caused by the automatic air-based cleaning (from April 11th to April 13th, for example) and a manual maintenance operation on April 10th. Applied to a full year of data, 3.5% of the points were identified as outliers.

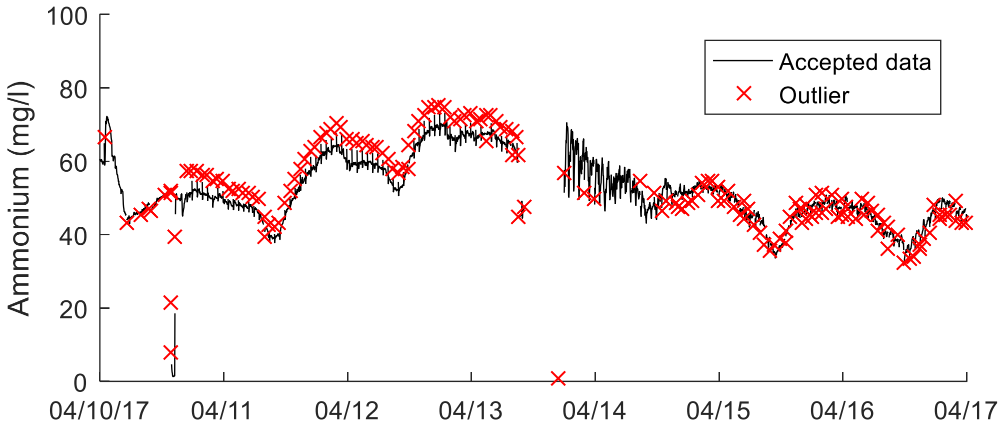

Figure 3. Result of the outlier detection procedure on the ammonium concentration in a pilot WRRF. Datapoints that were detected as outliers are replaced by red "x" and break the continuous line of accepted data.

The outlier detection method was also applied to time series not presented here (conductivity, temperature, chemical oxygen demand, turbidity). The method consistently identified between 1% and 4% of individual outliers. A rapid rate of change in the signal was often identified as an outlier, but as soon as the signal stabilizes, no more outliers were identified.

4 DISCUSSION AND CONCLUSIONS

Applying an autoencoder neural network strategy to identify outliers in a time series proved to be an efficient and tuning-friendly method for a variety of signals. In our experience, methods such as the one proposed by Alferes and Vanrolleghem (2016) are efficient, but require significant parameter tuning for each new measurement task. For example, Alferes' method is prone to missing a large variation in the signal (such as a pumping sequence), which leads to a loss of data until the auto-restart technique embedded in the method is triggered. The AENNTS method is much less sensitive to such variation since only the replication error is used to identify outliers. Therefore, since that replication error average is expected to stay close to zero, chances that the signal will be completely out of control are significantly reduced. On the other hand, it must be remembered that only punctual outliers are identified with the method. If a water quality sensor is removed from water for a long time (i.e. for cleaning or as the result of an external problem), the AENNTS will probably identify the first few points as outliers, but will not detect the continued sensor fault. The outlier detection is therefore a pre-treatment to fault detection.

Going forward, the AENNTS shows promises for various developments: First, the version presented here only works off-line, i.e. it cannot process a continuously incoming data stream. However, neural network training functions exist that can be used to continuously update the weights of the network to efficiently keep track of the signal and to perform reasonably frequent outlier detection. Second, the method proposed here is strictly univariate (a single signal at a time). When multiple sensors record different signals, neural networks are perfectly adapted to merge these different forms of information and extract non-linear correlations between them to improve their outlier detection capabilities. Third, due to their learning ability and their non-linear nature, neural networks have excellent potential for various data treatment objectives. Examples may include gap filling, fault detection, short-term forecasting, etc.

REFERENCES

Alferes, J., Vanrolleghem, P.A., 2016. Efficient automated quality assessment: Dealing with faulty on-line water quality sensors. AI Commun. 29, 701–709. https://doi.org/10.3233/AIC-160713

Chen, J., Sathe, S., Aggarwal, C., Turaga, D., 2017. Outlier Detection with Autoencoder Ensembles, in: Proceedings of the 2017 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, Houston, Texas, USA, pp. 90–98. https://doi.org/10.1137/1.9781611974973.11

Hawkins, S., He, H., Williams, G., Baxter, R., 2002. Outlier Detection Using Replicator Neural Networks, in: Kambayashi, Y., Winiwarter, W., Arikawa, M. (Eds.), Data Warehousing and Knowledge Discovery. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 170–180.

Martinelli, M., Tronci, E., Dipoppa, G., Balducelli, C., 2004. Electric Power System Anomaly Detection Using Neural Networks, in: Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science. Presented at the International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Springer, Berlin, Heidelberg, pp. 1242–1248. https://doi.org/10.1007/978-3-540-30132-5_168

Montgomery, D.C., 2009. Introduction to statistical quality control, 6th ed. ed. Wiley, Hoboken, N.J.

Sakurada, M., Yairi, T., 2014. Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction, in: Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, MLSDA'14. ACM, New York, NY, USA, pp. 4:4–4:11. https://doi.org/10.1145/2689746.2689747

Xiong, Y., Zuo, R., 2016. Recognition of geochemical anomalies using a deep autoencoder network. Comput. Geosci. 86, 75–82. https://doi.org/10.1016/j.cageo.2015.10.006