A Deep Learning Approach for Recovering Missing Time Series Sensor Data

YiFan Zhang[1], Peter Fitch[2], Peter Thorburn[1]

AGRICULTURE & FOOD
www.csiro.au

Wireless sensor networks are in widespread use in various application areas such as agriculture, industry and environment. High frequency measurements can reveal complex temporal dynamics that may be obscured by traditional sampling methods, and offer new insights into the inner workings of a monitored system. However, the issue of missing data is relatively common in wireless sensor networks and can have a negative effect on the conclusions drawn from the data.

Sequence-to-sequence imputation model

Data Imputation Problem

Fig. 1 illustrates a common scenario for missing time series sensor data. In this case, all the sensor data during the same period of time are missing. This can be caused by a variety of factors including unstable sensor power supply, data transmission errors or regular device maintenance.

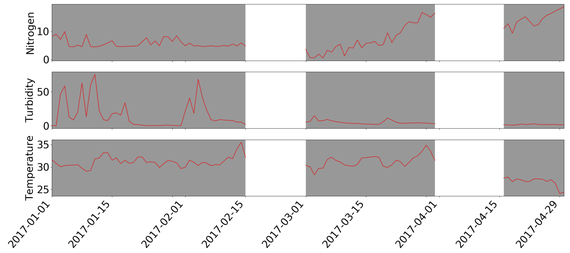

Figure 1: Time series with continuous missing data. Gray blocks highlight the available time series data (red solid lines), while the white blank spaces represent the missing data sequences for different time series.

Proposed model

We propose a new sequence-to-sequence imputation model (SSIM) for recovering missing data in wireless sensor networks. The SSIM uses the state-of-the-art sequence-to-sequence deep learning architecture, and the Long Short Term Memory Network is chosen to utilize both the past and future information for a given time. Our model is able to recover missing data sequences based on the information from both the past and future time indexes.

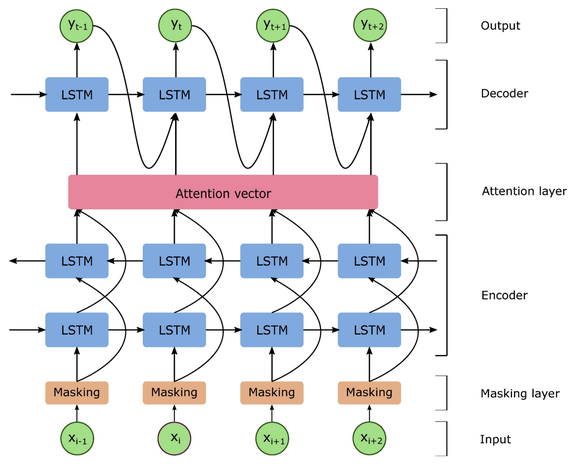

Figure 2: SSIM architecture. The encoder in the SSIM is a BiLSTM, which is comprised of a forward LSTM and a backward LSTM. The decoder in the SSIM is an unidirectional LSTM. The dropout Layers are also included in both the encoder and decoder. A masking layer is added to remove the zero-padded vectors in the input sequences.

As can be observed from Fig. 2, after taking input from the time series data, the model's masking layers filter out the zero-padded vectors in the input data samples and pass them to the BiLSTM encoder. The encoder processes data in each time index in both the forward and backward directions and generates a hidden vector as the output. The hidden vector is then processed by the LSTM decoder and the target sequence is generated recursively. During this process, the decoder calculates the attention weights in each time index so that it can focus on specific parts of the input to obtain relevant information.

Evaluation

Water quality data were collected from two monitoring sites in the Mulgrave-Russell catchment in the Great Barrier Reef, Australia. These parameters include temperature, rainfall, evaporation, radiation, vapour pressure, electrical conductivity, water discharge, water level, turbidity and nitrate.

We choose to recover the missing nitrogen sensory data in the monitoring data sets. Our model is applied to the water quality data collected from 1/1/2017 to 31/7/2017 and 1/9/2017 to 31/3/2018 as the training set. The water quality data collected in August 2017 and April 2018 were chosen as the test set. Overall, the model uses 14-month training data and 2-month testing data.

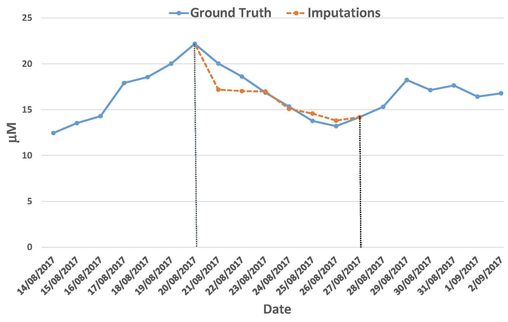

Figure 3: Recovering 6 missing nitrogen data from 21/8/2017 to 26/8/2017

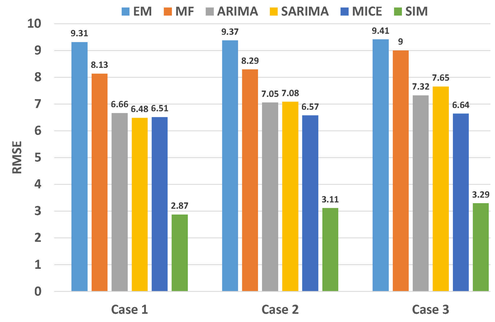

Figure 4: Evaluation of missing data imputation using the RMSE.

Fig. 3 depicts an example of recovering 6 missing nitrogen sensory data by using the SSIM. We also compared the SSIM with five different data imputation methods: ARIMA, SARIMA, Matrix Factorization (MF), Multiple Imputation by Chained Equations (MICE) and Expectation Maximization (EM). The SSIM achieves 2.87, 3.11 and 3.29 RMSE scores in cases 1, 2 and 3, respectively, as can be observed from Fig. 4. On the contrary, the best benchmark method MICE can only achieve 6.51, 6.57 and 6.64 for the three cases.

Our model is able to provide stable imputation results for most testing cases, which makes the proposed approach very promising for data recovery problems in sensor networks.

FOR FURTHER INFORMATION
YiFan Zhang
e yi-fan.zhang@csiro.au
w www.csiro.au/AF

AFFILIATION
1 CSIRO Agriculture and Food, Brisbane 2 CSIRO Land and Water, Canberra

ACKNOWLEDGEMENTS
We would like to thank the Great Barrier Reef Catchment Loads Monitoring Program for providing valuable real-time water quality monitoring data sets.