proceedings

MDPI

Proceedings

Optimal Missing Value Estimation Algorithm for Groundwater Levels †

Klemen Kenda [1,2,*], Filip Koprivec [1] and Dunja Mladenić [1,2]

[1] Artificial Intelligence Laboratory, Jožef Stefan Institute, Ljubljana, 1000, Slovenia
[2] Jožef Stefan International Postgraduate School, Ljubljana, 1000, Slovenia
klemen.kenda@ijs.si, koprivec.filip@gmail.com, dunja.mladenic@ijs.si
* Correspondence: klemen.kenda@ijs.si; Tel.: +386-31-420-418
† Presented at the 3rd EWaS International Conference on "Insights on the Water-Energy-Food Nexus", Lefkada Island, Greece, 27-30 June 2018.

Published: date (leave it empty)

Abstract: .

Keywords: missing values, data cleaning, data fusion, machine learning, ensembles

1. Introduction

Groundwater is an important indicator of changes in the climate. To assess the changes in comparison to groundwater withdrawal and different land-use long term availability of the data is needed. Usually different sensors for measuring groundwater levels are active in different time periods and might be subjected to very different properties of collected data (frequency, precision, etc.). In this paper we present a missing value estimation algorithm based on available data from active near-by sources and test it on data available from Ljubljana aquifer. Groundwater level measurements for Ljubljana aquifer are quite sparse. First measurements have been conducted in 1949 by one sensor and no sensor has measured groundwater levels continuously up until the present day. There are time intervals with more active sensors (1967-1971, 2002 – 2017), however – in between only a few sensors have been recording measurements. Additionally – each sensor has random missing values during the intervals, when measurements have been taken and also frequencies of data measurements change significantly between different time periods.

There are a plethora of reasons for missing data: hardware or software malfunction, human error (in the early measurements of groundwater), intentional removal (when data is corrupted). We choose the process of data imputation to deal with missing values, which means that we want to approximate the missing data points. There are many different ways to impute missing values. We can substitute a missing value with a mean, with a substitute from another individual, from a hot deck (randomly chosen value from an individual who has similar other variables), cold-deck (like in hot-deck, but the individual is chosen systematically), with regression (based on other variables), stochastic regression (adding a random residual value to regression), interpolation and extrapolation (estimate from the other observations from the same individual).

Authors of (Higashijima, 2010) propose usage of regression tree for missing data imputation of sparsely sampled time-series. Their models rely on nearby sensors' measurements. They propose the usage of linear interpolation at a pre-processing step to improve model accuracies. We also use linear regression interpolation within the sparse datasets, but also propose usage of b-splines. In addition to regression trees we also test other methods like linear regression, random forests and gradient boosted trees. We propose the usage of the optimal algorithm and optimal feature set.

Authors of (Lopes, 2010) report on usage of simple (interpolation with last value and mean) and sophisticated methods, such as linear regression and PCA. Tensor-based methods also produce very

good results in estimating missing values (Tan, 2013). In some (extreme) cases missing values were successfully imputed with this method, where data of one or several days were completely missing.

In (Kenda, 2019) authors use short-term Kalman filter models for imputation of missing or corrupt time-series data in a streaming scenario.

In groundwater missing data imputation (Nunes, 2004) suggested groundwater nitrate monitoring network optimization with reduction of measuring nodes. They estimate errors at missing nodes with linear methods. In our work we take a similar approach, but do not care about network reduction.

Authors of (Srebotnjak, 2012) suggest the usage of hot-deck imputation method on a global scale, where they replace missing values with a value from the donor cases, which match the recipient node in a set of specified variables from the same dataset. Value is chosen randomly or from the deck with the closes similarity. In our work the target node has been observed before therefore regression models can provide better estimations.

In this work we try to estimate missing values of a particular sensor based on the data-driven regression models with features from the neighboring sensors. We model each sensor with an ensemble of data-driven models with all available combinations of available sensors. Algorithm selects the most appropriate model (with lowest estimated error) and uses it to predict missing values at a particular time. Final result of the algorithm is the full dataset of all available sensors in the system, which can be used for further climate and urban-planning studies.

2. Materials and Methods

2.1. Data and Data Acquisition

2.3.1. Linear Regression

3. Results

3.1. Exploratory data analysis

Underground water level acquisiton is quite sparese during middle period (1970-2008) and very sparsely acquired in first period (frequency of data acquisiton in years between 1960 and 1970 varies, but rarely exceeds more than 2 measurements per month), sample date frame with more frequent and reliable data points was selected for experiements. Crucially, since method used for missing data estimation relies on nearby sensors and availability of their data, time range from beginning of year 2013 until the end of 2015 was selected, consisting of 1092 different measurements.

During this period, 12 measurement stations were active, exposing a few different characteristics, as seen from their plots from Figure 2. With first glance, see twosensors that highly deviate from others: namely 85012 an 85073. Other measurements seem to follow roughly similar pattern, but with varying amplitude and some individual features. Judging from correlation matrix presented in Figure 1, measurements seem to be correlated among nearby sensors. But not so much between the groups. Additionally, last set of measurements contains a long interval of missing values, that was predicted at the end.

3.2. Modeling and evaluation

Each data set was modeled by optimal combination of other sets (excluding the last which contained too long interval of missing values), by linear regression, support vector regressor and random forrest regressor. For each set, optimal subset of predictors and prediction model was selected according to R^2 score (full dataset was split in ratio of 7:3, train to test set and each model was evaluated on the same test-set, for last measuring station with missing results, same procedure was performed but only on available data).

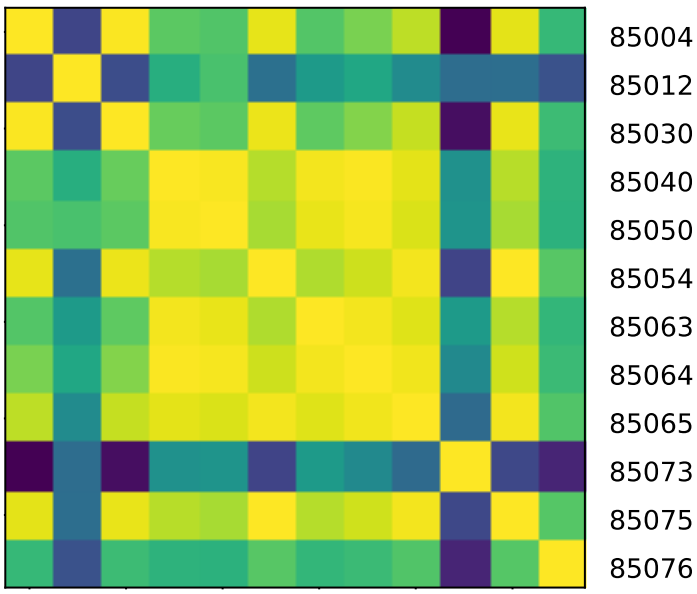

Figure 1: Correlation matrix

Results of best performing predictions are presented in Table 1. For grouped sensors, mean of R^2 and RMSE is presented with standard deviation in parenthesis, optimal predictors are not presented, as they vary from sample to sample (varying from 7 to almost all others) interestingly, measure station with id 85012 is always present as predictor, although it's coefficient is mostly somewhat small (ranging from almost 0 to 0.17). Optimal predictors are presented as last two digits of sensor id, as numbered in Figure 2.

Table 1: Modeling results

	R^2	RMSE	Optimal algorithm	Optimal predictors
grouped	0.9976 (0.0028)	0.0318 (0.0244)	Linear regression	varies
85012	0.9388	0.0993	Linear regression	04, 50, 54, 63, 64, 65, 73, 75
85073	0.8572	2.7233	Random forest regression	04, 30, 50, 54, 63

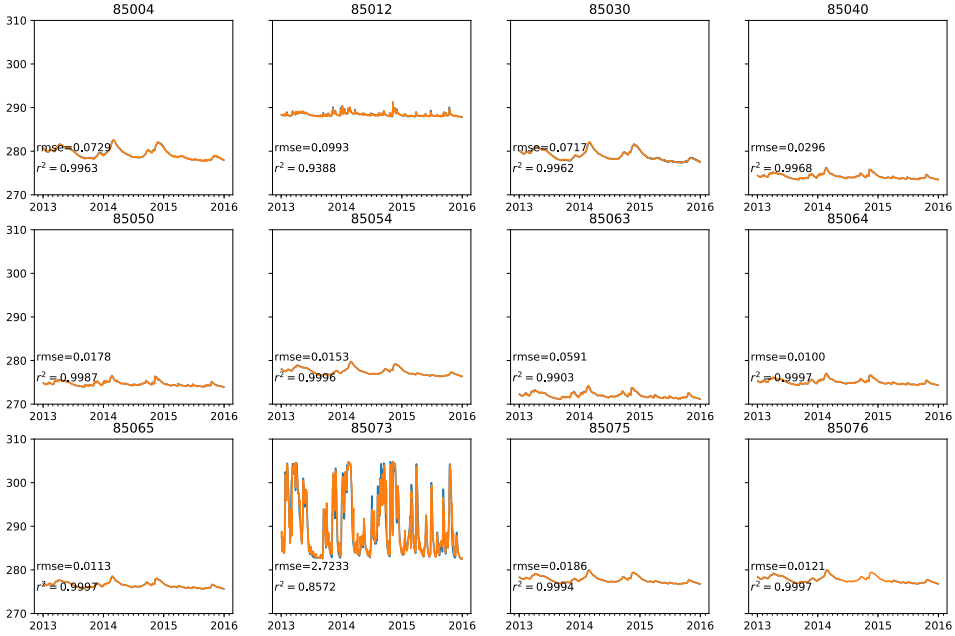

Figure 2: Best performing predicted values overlaid on real values

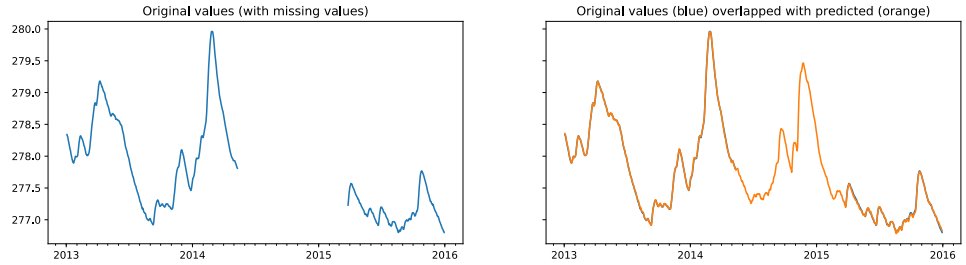

Figure 3: Prediction of missing values for sensor 85067

4. Discussion

5. Conclusions

Acknowledgments: The work described in this paper has been conducted within the projects Water4Cities and PerceptiveSentinel. Both projects have received funding from the European Union's Horizon 2020 programmes. Water4Cities is a Research and Innovation Staff Exchange project under grant agreement number 734409. PerceptiveSentinel is a Research and Innovation project under grant agreement number 776115. This paper and the content included in it do not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of its content.

Author Contributions: Klemen Kenda conceived and designed the experiments and wrote the paper; Filip Koprivec performed the experiments; Dunja Mladenić provided additional analysis of the data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lopes, J., Bento, J., Huang, E. Traffic and mobility data collection for real-time applications. (ITSC), 2010 13th …, 2010.

2. H. Tan, G. Feng, J. Feng, W. Wang, Y.-J. Zhang, and F. Li, "A tensor-based method for missing traffic data completion," Transportation Research Part C: Emerging Technologies, vol. 28, pp. 15–27, Mar. 2013.
3. Nunes, L. M., E. Paralta, M. C. Cunha, and L. Ribeiro (2004), Groundwater nitrate monitoring network optimization with missing data, Water Resour. Res., 40, W02406, doi: 10.1029/2003WR002469.
4. Srebotnjak, T., Carr, G., de Sherbinin, A., Rickwood, C., A global Water Quality Index and hot-deck imputation of missing data, Ecological Indicators 2012, Vol. 17, pp. 108-119, doi: 10.1016/j.ecolind.2011.04.023.
5. Higashijima, Y., Yamamoto, A. , Nakamura, T., Nakamura, M., Matsuo, M. Missing Data Imputation Using Regression Tree Model for Sparse Data Collected via Wide Area Ubiquitous Network, 2010 10th IEEE/IPSJ International Symposium on Applications and the Internet, Seoul, 2010, pp. 189-192. doi: 10.1109/SAINT.2010.18
6. Kenda, K., Mladenić, D. Autonomous Sensor Data Cleaning in Stream Mining Setting. *Business Systems Research Journal* **2019**, Vol.: 9 (in press).