EasyChair Preprint

№ 10616

Proposal for Real-Time Pre-Processing of Anomalies on Data Collected by Meteorological Sensor Network

Salomon Mba Tene and Vivient Corneille Kamla

July 25, 2023

Rubrique

Proposal for real-time pre-processing of anomalies on data collected by meteorological sensor network

Sous-titre

Nom Prénom

Département
Université
Ville
Pays
email@email.fr

ABSTRACT. A weather station is a set of sensors that record and supply physical measurements and meteorological parameters related to climate variations in a locality. The station collects, processes and stores meteorological data for use in weather forecasting. Weather forecasting is the application of science and technology to predict atmospheric conditions for a specific location and period in the future. However, the weather forecast data produced often deviate considerably from the actual weather values. This large discrepancy can be explained by the anomaly pre-processing technique (missing values and outliers) used and sometimes by the absence of the preprocessing of anomalies in a meteorological observation system. The presence of an anomaly in one variable can compromise the prediction of the whole observation. This paper proposes an algorithm for pre-processing anomalies in a weather observation system. For the pre-processing of missing values, we have used seven imputation techniques, namely MICE, DecisionTreeRegression, BayesianRidge, LinearRegressor, ExtraTreesRegressor, KNeighborsRegressor and KNNImputer. The results showed that the MICE technique performed best, with an MSE of 0.019344.

RÉSUMÉ. Résumé

KEYWORDS : Weather Station, Data Preprocessing, Weather Forecasting, Machine Learning.

MOTS-CLÉS : Mots clefs

1. Introduction

Predicting the future is one of mankind's age-old dreams [1]. This has led mankind to technological mutation due to environmental and human change. Man's pressing need is to know weather conditions in advance in order to carry out his activities. This has given rise to a number of meteorological systems, including weather observation systems and weather forecasting systems. In addition, science aims not only to predict but also to understand the observed phenomenon, by means of an explanatory model [1, 2, 3]. Machine Learning is the science that meets this ambition [4, 5, 6, 7]. These observation systems contain anomalies in their data server which bias the forecast data [20, 14].

(HAO, MANESH, HENGXU, LEI, 2017) [25] proposed the design of a micro automatic weather station for a modern power grid and weather sensors. The designed system allows weather monitoring. Weather data is routed via GPRS. (CARLOS , JORGE, DANIEL, PABLO, 2018) [26] have developed a prototype of a solar weather station allowing the collection and storage of data. Data is transmitted via wifi. (ZAID, HADI, MOUSSA, YAQEEN , 2020) [12] presented in their article a prototype of an economical weather station with monitoring system and ZigBee communication technique. But the weather observation systems proposed by these authors do not contain the real-time anomalies preprocessing module.

However, the forecast meteorological data produced often deviate considerably from the actual meteorological values. This large discrepancy can be explained by the anomaly pre-processing technique (missing values and outliers) used and sometimes by the absence of the preprocessing of anomalies in a meteorological observation system. Furthermore, the presence of an anomaly in one variable can compromise the prediction of the whole observation [8] and can lead to poor fit and predictive modeling performance [9].

In this article, we present in section 2, a literature review on weather observation systems. The question here is whether these observation systems have a real-time anomaly pre-processing module. In Section 3, we present the methods and materials. In this section, we propose the flowchart of the real-time anomaly pre-processing module and its algorithm. We then list the different imputation techniques and metrics we'll be using in the next section. In Section 4, we present the results and analytic. In this section, we will carry out an analysis of our data, present the results of the metrics for the different imputation techniques used, and highlight the best-performing imputation technique. And finally, in section 5, we conclude by proposing some perspectives for our future scientific work.

2. Literature review on weather observation systems

Meteorology is the interdisciplinary science of atmospheric physics, which studies the state of the weather, the atmospheric environment, the phenomena produced and the laws that govern it [10]. The process of meteorological measurements is directly linked to the setting up of a weather station [10, 11, 12, 13]. Moreover, a weather station is a set of sensors that record and supply physical measurements and meteorological parameters linked to climate variations in a locality [10, 11, 13]. Thus, several authors have contributed to the proposals for meteorological observation systems and weather monitoring systems. Table 1 below lists the following: authors, meteorological parameters (MP), transmission technique (TT), detection and correction of anomalies in an observation system (DCAOS). Weather observation measurements make weather forecasting possible.

Table 1. *List of related work on the weather observation system.*

AUTHORS	MP	TT	DCAOS
(XINGANG & YU , 2010) [21]	Temperature, Air humidity, Atmospheric pressure, Wind speed, Wind direction	GSM TC35i	Absent
(YONG-HUA & SI-REN, 2011) [13]	Temperature,Relative humidity, Precipitation, Wind speed, Wind direction, Visibility	CDMA	Absent
(SUSMITHA & SOWMYABALA, 2014) [22]	Air temperature, Relative humidity, Gas	Serial communication and GSM	Absent
(GONCALO, et al., 2015) [23]	Air temperature, Relative humidity, Solar radiation	Not mentioned	Absent
(TANMAY , SHOBHIT , AKASH, & THAKARE, 2016) [24]	Temperature, Relative humidity, Precipitation, Wind speed, Wind direction, Atmospheric pressure	WIFI	Absent
(HAO, MANESH, HENGXU, & LEI, 2017) [25]	Temperature, Relative humidity, Light intensity, Wind speed, Wind direction	GPRS	Absent
(CARLOS , JORGE, DANIEL, & PABLO, 2018) [26]	Temperature, Relative humidity, Air quality, Wind speed, Atmospheric pressure	WIFI	Absent
(ANITA , ASHWINI , KAJAL, NEHA, & CHOUDHARY, 2019) [27]	Temperature, Relative humidity, Soil moisture, Precipitation	WIFI	Absent
(ZAID, HADI, MOUSSA, & YAQEEN , 2020) [12]	Temperature, Relative humidity, Dust density, Wind, Atmospheric pressure, Precipitation	ZIGBEE	Absent
(GOSAVI, BHIDE, BHOSALE, & SUTAR, 2021) [28] [41]	Temperature, Soil moisture, Light, Precipitation	WIFI	Absent
(DNYANESHWARI , PRAJAKTA , UTKRSHA , & SMITA, 2022) [29]	Temperature, Relative humidity, Precipitation		Absent

According to Table 1, we note a lack of anomaly detection and correction in a meteorological observation system. These authors did not account for real-time anomaly processing in their respective observation system. Transmission media were used in the various articles. The most widely used transmission medium is WIFI.This is due to the fact that it is more accessible. However, its coverage area is very limited, unlike GSM, which has a large coverage area. According to this Table 1 , the Temperature parameter is used in all the articles we read while parameters such as gas, visibility and soil humidity are the least used in our consulted articles. The different parameters used depend on the context of the study made by the authors.

The origins of the anomalies are mentioned in [14, 15, 16, 17, 8, 18, 19]. It should be noted that the presence of an anomaly in one variable can compromise the prediction of the whole observation [20, 14, 15, 16, 17, 8, 18, 19]. The anomaly detection and correction module will enable us to improve the accuracy and precision of weather forecast data. It is therefore important to offer a module for pre-processing weather data in an observation system.

3. Materials and Method

Any observation system needs to integrate a real-time data pre-processing module to help the forecasting system obtain good quality forecast data. Our proposed intelligent system is illustrated in Figure 1.

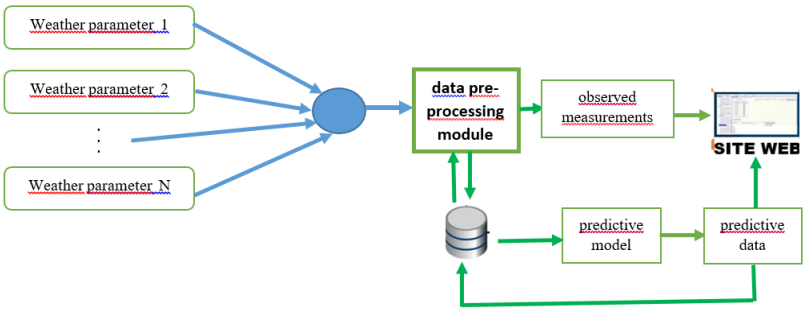
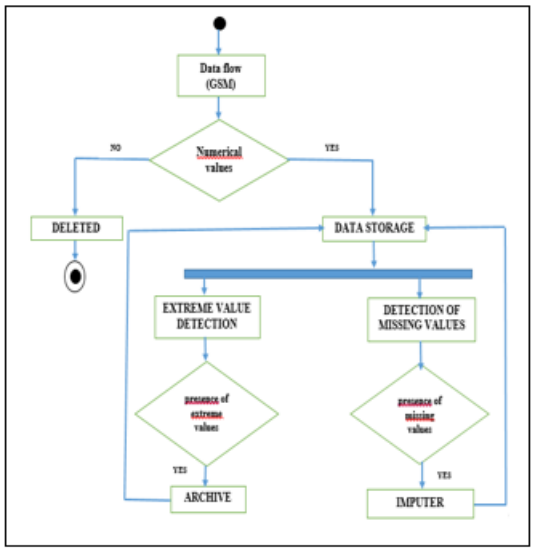


Figure 1. *Intelligent weather forecasting system based on sensor network*

Anomalies are classified into two groups namely missing values and outliers. (SALVADOR, et al., 2015) [19] mentioned three variants of outliers in his article: Data errors, Extreme values, Fraudulent entries. In our work, we will pre-process extreme values specifying extreme weather conditions and fraudulent entries indicating non-numerical values. The extreme values will be archived. The reasons for archiving these extreme values are given in [30]. Non-numerical values will be deleted. The detection of extreme values and missing values will be carried out in parallel. Missing values will be imputed using an intelligent model. The Figure 2 shows the flowchart and algorithm of the data pre-processing module in a weather observation system. The primary goal in machine learning is to enable computer programs to learn automatically by identifying anomalies in the data without human intervention and adjusting actions [31].



Data: Data flow
Result: cleaned data
while *new observation at time* T_K **do**
  detection of non-numerical values ;
  if *detection of non-numerical values is empty* **then**
    stoker this new observation ;
  else
    delete non-numeric weather parameter(s);
  end
  if *number of past observations* ≥ 10 **then**
    detection of missing values ;
    if *detection of missing values is not empty* **then**
      imputation of missing values ;
    end
  end
  detection of extreme value ;
  if *detection of extreme value is not empty* **then**
    archive extreme values ;
  end
  $T_{K-1} \leftarrow T_K$;
  $T_K \leftarrow T_{K+1}$;
end

Algorithm 1: Pre-processing

Figure 2. *Flowchart and Algorithm of data pre-processing in an observation system*

We will focus on seven Machine Learning imputation models in order to select the one with the best performance. It should be noted that the presence of missing value in

the dataset has always been a major and common problem that affects data quality and prediction results [32, 33]. Thus, missing value imputation methods must minimize the effect of incomplete data sets for the prediction model [32]. The imputation models we will use include : LinearRegressor (LR), BayesianRidge (BR), DecisionTreeRegressor (DTR), ExtraTreeRegressor (ETR), KNeighborsRegressor(KNR), KNNImputer (KNNI) and MICE. These techniques have already been used in the fields of genetics [38], real estate [34], fuzzy entropy [35], healthcare [36], energy [37] and hydrology [33]. We will evaluate these techniques using the following metrics: execution time, MSE (Mean Squared Error), RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error). The simulation will be based on meteorological data for the city of Ngaoundéré, with five meteorological parameters : temperature, precipitation, dew point, relative humidity and atmospheric pressure. For our study, we will use the 2011 data set (January 01, 2011 to December 31, 2011). The sampling time is 1 hour. We will create artificial missing values in order to evaluate the performance of each imputation model. We used a laptop with the following specifications : Windows 10 Professional 64-bit, 4GB RAM, AMD 1.80 GHz processor. The programming language used to analyze our data and implement our various imputation models is python. The IDE used is Jupiter Notebook.

4. Result and Analysic

Our dataset contains 8759 weather observations for the year 2011. When we visualize our dataset, we notice that it contains missing values. We have illustrated a single weather parameter such as the dew point symbolized by T2MDEW to visualize the missing values, as shown in Figure 3. Note that the other weather parameters also contain missing values. Interruptions in the Figure 3 symbolize the presence of missing values.

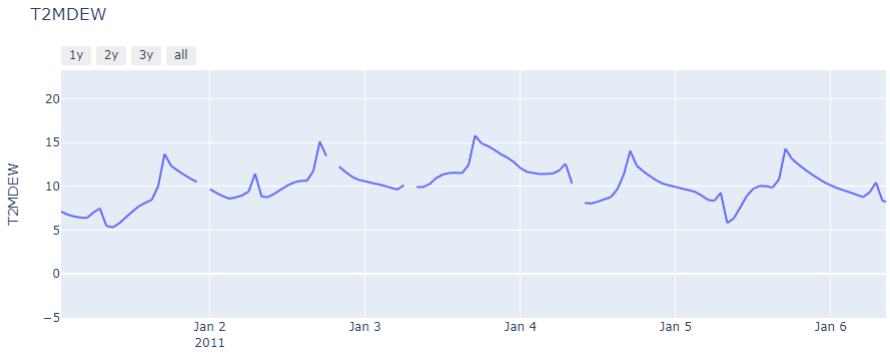


Figure 3. *presence of missing values in the dew point parameter*

We performed imputation simulations on our dataset using our implemented machine-learning imputation models. The metric results of our simulations are shown in Table 2.

According to Table 2, the **MICE model** has the best performance for missing data prediction (data imputation) with an **MSE of 0.01934 (RMSE = 0.13908 and MAE = 0.00198)** but the longest runtime for data imputation. The **LR model** has the shortest execution time for data imputation having an MSE of 0.03533. The DTR model is the

second best model on the MSE (MSE=0.2184) having a running time of 7.51 seconds. This model has a much shorter execution time than that MICE.

The choice of model will depend either on the execution time or on the MSE (RMSE, MAE) or else on the MSE having a considerable execution time.

Table 2. *Models evaluation*

METRIC	LR	MICE	KNNI	BR	DTR	ETR	KNR
EXECUTION TIME	**0.23 s**	50.09 s	0.39 s	0.56 s	7.51 s	46.25 s	2.20 s
MSE	0.03533	**0.01934**	0.08354	0.03267	0.02184	0.044778	0.05935
RMSE	0.18798	**0.13908**	0.28904	0.18074	0.14778	0.21161	0.24361
MAE	0.00323	**0.00198**	0.00493	0.00317	0.00226	0.00263	0.00304

5. Conclusion

In short, anomalies are common problems that affect data quality and the final results of forecasts. Thus, Section 1 illustrates the lack of anomaly detection and correction as presented in Table 1. To overcome this problem, we proposed a module for pre-processing time data in a weather observation system. Subsequently, we focused on the imputation of missing data. Machine Learning imputation models include LinearRegressor, BayesianRidge, DecisionTreeRegressor, ExtraTreeRegressor, KNeighborsRegressor, KNNImputer and MICE. These models were evaluated on the basis of the following metrics: EXECUTION TIME, MSE, RMSE, MAE. Our results show that the MICE model performs best, with an MSE of 0.01934 (RMSE = 0.13908 and MAE = 0.00198). But the disadvantage of this model is its execution time (50.09 seconds). We chose this imputation model in our data pre-processing module because it is more accurate. However, meteorological data pre-processing is not enough, since we observe climate variation and change. This climate variation or climate change is a factor that also biases the forecast data, as the meteorological predictive models are static. Consequently, these predictive models do not adapt to variations in weather observations. For this reason, there is also a considerable discrepancy between forecast data and weather observation data. This problem will be the subject of our next article. In this next article, we will propose a Machine Learning predictive model that automatically adjusts to meteorological variation. it should be noted that the primary goal in machine learning is to enable computer programs to learn automatically without human intervention and adjusting actions.

6. References

[1] A. CHLOÉ-AGATHE, "Introduction au Machine Learning", 2019.

[2] X. SHI, Z. CHEN, H. WANG, Y. DIT-YAN, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting", 2016.

[3] M. SAMIR TOUKOUROU, "Application de l'apprentissage artificiel à la prévision des crues éclair", 2010.

[4] S. DADHICH, V. PATHAK, R. MITTAL, R. DOSHI, "Machine learning for weather forecasting", 2021.

[5] A. MOSAVI, P. OZTURK, C. KWOK-WING, "Flood Prediction Using Machine Learning Models : Literature Review", *Water*, 2018.

[6] C. HU, Q. WU, S. JIAN, N. LI, Z. LOU, "Deep Learning with a Long Short-Term Memory Networks Approach for Rainfall-Runoff Simulation", *Water*, 2018.

[7] L. XUAN-HIEN, H. VIET HO, G. LEE, S. JUNG, "Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting", *Water*, 2019.

[8] F. ROLLON, C. BACHECHI, L. PO , "Anomaly Detection and Repairing for Improving Air Quality Monitoring", *Sensors*, 2023.

[9] B. JASON, "Automatic Outlier Detection Algorithms in Python", 2020.

[10] M. ANTONIO PONCE-JARA, C. VELÁSQUEZ-FIGUEROA, D. TONATO-PERALTA, G. PAREDES-MORILLO, "Diseño de una estación meteorológica automática para registrar las variables solar y eólica", *Revista Arbitrada Interdisciplinaria KOINONIA*, 2020.

[11] A. MUNANDAR, H. FAKHRURROJA, M. IIHAM RIZQYAWAN, R. PUTRA PRATAMA, J. WINARYO WIBOWO, I. ASFY FAKHRY ANTO, "Design of Real-time Weather Monitoring System Based on Mobile Application using Automatic Weather Station", 2017.

[12] Z. KHUDHUR HUSSEIN, H. JAMEEL HADI, M. RIYADH ABDUL-MUTALEB, Y. SABAH MEZAAL, "Low cost smart weather station using Arduino and ZigBee", *TELKOMNIKA*, 2020.

[13] C. YONG-HUA, L. SI-REN, "Design and realization of an utomatic weather station at island", 2011.

[14] L. MIN-KI, M. SEUNG-HYUN, K. YONG-HYUK, M. BYUNG-RO, " Correcting Abnormalities in Meteorological Data by Machine Learning ", *IEEE*, 2014.

[15] G. JAIN, B. MALLICK, " A Review on Weather Forecasting Techniques ", *International Journal of Advanced Research in Computer and Communication Engineering*, 2016.

[16] D. DODAKE, C. WAGHMARE, " A literature review on Improvement of Weather prediction by using Machine learning ", *International Journal of Research Publication and Reviews*, 2022.

[17] Z. MINGHU, G. JIANWEN, L. XIN, J. RUI, " Data-Driven Anomaly Detection Approach for Time-Series Streaming Data ", *Sensors*, 2020.

[18] P. VIVIANE, " Traitement des valeurs aberrantes : concepts actuels et tendances générales ", 2005.

[19] S. GARCÍA, L. JULIÁN, H. FRANCISCO, " Data Preprocessing in Data Mining ", 2015.

[20] P. SHARMA, S. PRAKASH, " Real Time Weather Monitoring System Using Iot ", *ITM Web of Conferences 40*, 2021.

[21] X. GUO, Y. SONG, " Design of Automatic Weather Station Based on GSM Module ", 2010.

[22] SUSMITHA, SOWMYABALA , "Design and Implementation of Weather Monitoring and Controlling System", *International Journal of Computer Applications*, vol. 97, num. 3, 2014.

[23] G. MESTRE, A. RUANO, H. DUARTE, S. SILVA, H. KHOSRAVANI, S. PESTEH, M. PEDRO, R. HORTA, "An Intelligent Weather Station", *Sensors*, 2015.

[24] T. PARASHAR, S. GAHLOT, A. GODBOLE, THAKARE, "Weather Monitoring System Using Wi-Fi", *International Journal of Science and Research (IJSR)*, vol. 5, 2016

[25] H. LI, M. KUMAR OCHANI, H. ZHANG, L. ZHANG, "Design of micro-automatic weather station for modern power grid based on STM32", *Journal of engineering*, 2017

[26] C. MORÓN, J. PABLO DIAZ, D. FERRÁNDEZ, P. SAIZ, "Design, Development and Implementation of a Weather Station Prototype for Renewable Energy Systems", *Energies*, 2018.

[27] A. BHAGAT, A. THAKARE, K. MOLKE, N. MUNESHWAR, CHOUDHARY, "IOT Based Weather Monitoring and Reporting System Project", *International Journal of Trend in Scientific Research and Development (IJTSRD)*, vol. 3, 2019.

[28] GOSAVI, BHIDE, BHOSALE, SUTAR, "IOT BASED WEATHER MONITORING SYSTEM USING ARDUINO-UNO", 2021.

[29] D. NAGANE, P. MORE, U. CHAVAN, S. GAWADE, "IOT Based Weather Reporting System", *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, vol. 2, 2022.

[30] R. CERVENY, "Les extrêmes météorologiques et climatiques archivés par l'OMM", 2018.

[31] S. ABGHARI, "Data Mining Approaches for Outlier Detection Analysis", 2020.

[32] R. ARMINA, A. MOHD ZAIN, N. AZIZAH ALI, R. SALLEHUDDIN, "A Review On Missing Value Estimation Using Imputation Algorithm", *Journal of Physics*, 2017.

[33] F. ORIANI, S. STISEN, M. DEMIREL, G. MARIETHOZ, "Missing Data Imputation for Multisite Rainfall Networks: A Comparison between Geostatistical Interpolation and Pattern-Based Estimation on Different Terrain Types", 2020.

[34] H. VASANI, H. GANDHI, S. PANCHAL, S. MISHRA, "House Price Prediction Using Advanced Regression Techniques", *SPRINGER*, 2022.

[35] F. KHALID KARIM, H. ELMANNAI, A. SELEEM, S. HAMAD, S. M. MOSTAFA, "Handling Missing Values Based on Similarity Classifiers and Fuzzy Entropy Measures", *ELECTRONICS*, 2022.

[36] K. SEU, K. MI-SUN, H. LEE, "An Intelligent Missing Data Imputation Techniques: A Review", 2022.

[37] A. AFZAL, S. ALSHAHRANI, A. ALROBAIAN, A. BURADI, S. AFGHAN KHAN, "Power Plant Energy Predictions Based on Thermal Factors Using Ridge and Support Vector Regressor Algorithms", *ENERGIES*, 2021.

[38] A. YADAV, A. RASOOL, A. DUBEY, N. KHARE , "A Hybrid Approach for Missing Data Imputation in Gene Expression Dataset Using Extra Tree Regressor and a Genetic Algorithm", *SPRINGER*, 2023.