Transactions on Networks and Communications - Vol. 9, No. 4
Publication Date: August, 25, 2021
DOI: 10.14738/tnc.94.10539.
Pödör, Z. (2021). Sensors and Wrong Values. Transactions on Networks and Communicaitons, 9(4). 23-28.

SOCIETY FOR SCIENCE AND EDUCATION

Sensors and Wrong Values

Zoltán Pödör
Department of Numerical Analysis, ELTE, Budapest, HUN

ABSTRACT

In the world of IoT and BigData, sensor based data collection is a really important domain. Using these tools it is possible to stow large amounts of data collection sensors in a factory or in nature in harsh environments. However, in order to obtain valuable information from these tools, it is important that potentially wrong data is discovered and handled. Automated exploration of wrong data is not a trivial task, even if similar measurements are performed in parallel with spatial differences. We present the difficulties of revealing defected data and suggest easy-to-implement procedures for detecting and handing them. We also draw attention to the potential disadvantages of these methods based on the given results.

Keywords: sensors, sensor data, missing data.

INTRODUCTION

While in 2015, the number of IoT (Internet of Things) sensor devices, which are connected to the network was 15.41 billion, it is now almost 26 billion [5]. Estimated growth is higher than ever before, and can reach 75 billion by year 2025. All these facts show that sensor-based data collection is becoming more and more important in industrial environment [1] and our analyzes, our different decisions will be increasingly based on data, which are collected by such tools, sensors.

In order for these results and decisions reliability [4], it is important that the basic data would be as complete and without errors as it is possible. In the case of missing data, the place of replacement is clear, but the methodology of replacement may be the question, for which there are numerous well-known techniques that we do not want to deal with in this paper. However, in the case of failure data, exploring their location is a serious task because the error can not only clearly indicate by wrong data - such as the 89 ° C temperature data in the summer - but theoretically good, but it is not the right measurement.

From the point of view of detecting measurement errors, it is relevant whether we are talking about a completely independent set of measurement data, or that there are time-parallel measurements that are comparable. The other measurements are able to compare individual measurements not only with the measurements of the same time series, but also with the time series of parallel measurements. The latter also provides an opportunity to detect errors that are not trivially inferred from a single time series.

In this article there are presented simple methods that are capable of detecting potentially data failure and draw attention to the disadvantages of these methods, which may mean that certain errors are not detected, or that elements that are not faulty are assumed to be wrong. The

methods were developed and tested on temperature data sets collected by sensors, and the results are also presented in connection with these data. Measurements were carried out in parallel by 10 sensors in nearly identical conditions outside of nature in a well-defined area, so we could assume that the measurement conditions were theoretically very similar.

However, the results show that this is not the case, there are several sub-data sets that certainly contain incorrect data, but the reasons require other separate tests.

DATA AND METHODOLOGY

Data

The temperature measurements were carried out by 10 sensors, parallel to a well-defined small area in the outside, with 24 measurements per day, one measurement in every hour. The data series are available for 2 years, 2017-18, and we have searched them for the wrong data and show the results.

Methodologies

In the available data series, we were looking for missing data in both horizontal and vertical ways, it means, that we searched for outliers, data failure in 10-10 measurements for a given measurement time, as well as searched for incorrect measurements in a single data series.

The vertical approach basically allows one set of data to be analyzed at the same time, and the methods basically search for data that in some respects differ significantly from other elements in their environment. This can be interpreted as smoothing the data set and calculating the distance of each measurement from the connected smoothed data. Where this distance exceeds a certain threshold, it is marked as potentially incorrect data. In this work, we used two methods [3] to smooth the data series and we wanted to choose simple procedures:

Moving average: moving averaging is one of the most common approaches for data smoothing techniques. The essence of the method is to move a window of width k above the data row, continuously moving one element forward with a fixed window width and calculating the average of the elements in the actual window. Then there are replaced the elements of the original data row in the center of the windows if the window width is odd. If the window width is even, it is necessary to have a truing step before the substitution so that the calculated average value falls into a real position. Truing means that in the case of even k, the calculated moving averages are averaged in pairs, that is, we calculate another series of moving averages this time.

The width of the window is called the order of the moving average, in practice, because of the above, it is chosen as an odd value, typically from the range 3-15 depending on the data series. The higher the order, the more smooth it is, but the greater the loss of information. At the same time, it should be noted that due to the window technique, data loss occurs at the beginning and at the end of the smoothed data series, the length of which is a whole part of k / 2.

Loess Regression: The essence of regression techniques is to fitting a one- or more variable function to the original points of the data series in order to match the points as best as possible (e. g. the least squares method). In this case, the match result will be a closed-loop function with the parameters calculated during the regression. The basic approach is looking for a global solution for fitting. In contrast, Local Regression (Loess) is about not adapting globally to the

entire data set, but taking into account only points in a given environment of that point in the current alignment step and replacing the original value with the fitted values based on regression.

The method has a smoothing parameter (n) which has to be specified by the user, thereby determining the size of the environment and thereby the degree of smoothing. The parameter is a value from the (0,1] interval, which percentage value expresses how many percent of the original measurement are involved in each local step of local regression. If the length of the whole time series is l, then m=ceiling(n*h) defines the number of points involved in regression in one step.

By default, the fitted function can fit a straight, but higher order polynomial, and in practice, a first or second degree function is fitted locally. The procedure may also weight the points in the environment using a decreasing weight based on their distance from the current point. The method performs a similar local smoothing as the moving average, with no single average training, but with local alignment and the matched data thus obtained.

For the horizontal approach, i.e. when 10-10 measurements for the same time stamp were tested, the following procedures [1] were selected in this study:
Quantile-based data search: this is the best known and most traditional method that is used in practice to find outlier data. It is not depend on the size of the data series and, and according to the literature, quite reliable. In the method, data that is at least one and a half times the size of the interquartile range (IQR = Q3-Q1) at the third quartile (Q3) or at least one and a half times the IQR at the first quartile (Q1) are considered to be outlier data. The boxplot diagrams also use this approach to determine the outliers in data mining. At the same time, it is important to note that the method is basically used for searching outbound data, it can reveal significantly different data, it is not suitable for detecting minor errors with the above mentioned parameterization.

Average and standard deviation based approach: it is practically equivalent to the previous methodology, where we deviate from the mean by means of standard deviation. The advantage of this method is that the value of constant multipliers can be defined more freely, depending on the data.

RESULTS

In the individual analysis of the data series, the moving average and loess regression give us similar results as shown in Figures 1 and 2.

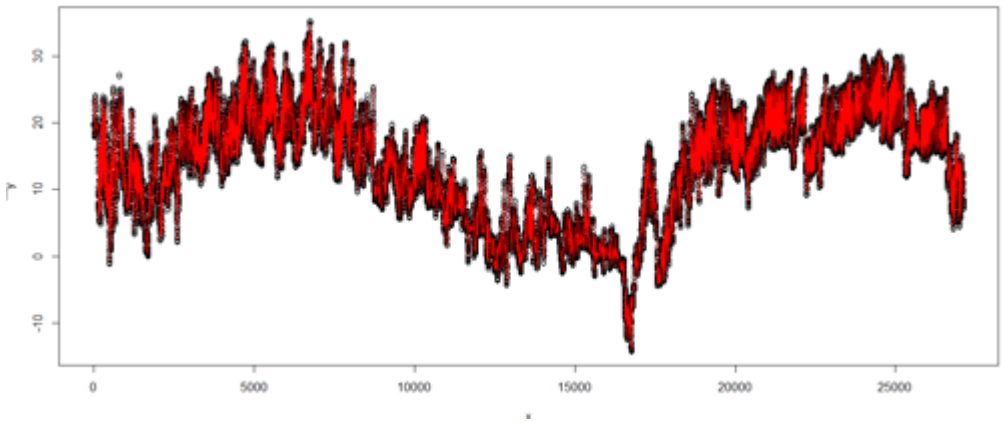


Fig. 1. Smoothed data by 7th moving average

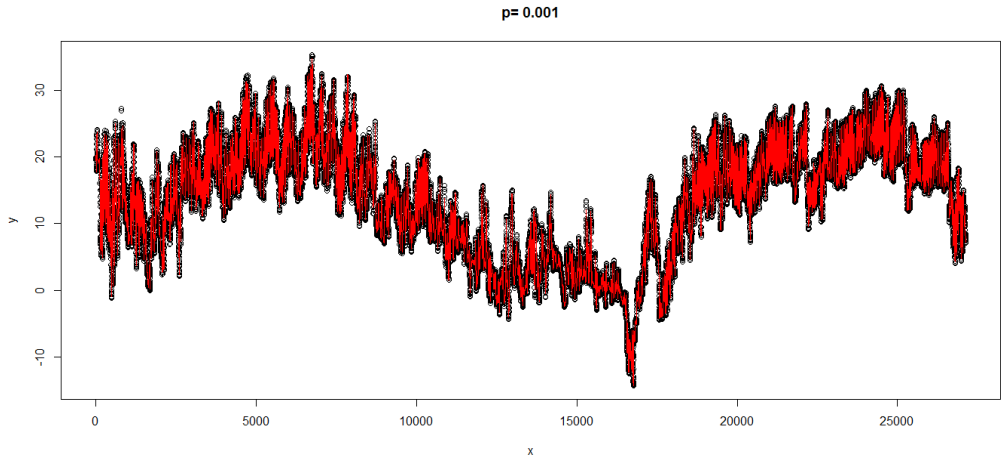


Fig. 2. Smoothed data by p=000.1 loess regression

After finishing, the distance of the individual original measurements from the smoothed values assigned to them was examined, and where this distance exceeded a threshold k, this data was marked as potential faulty data. Results are significantly influenced by the determination of k. The tests were performed for k=3, k=4 and k=5. The total number of data considered to be defective was also below 1% for k=3, which is a manageable quantity. Table 1 shows some of the typical discovered errors.

Table 1. data failures (examples)

row number	sz1	sz2	sz3	sz4	sz5	sz6	sz7	sz8	sz9	sz10
741	19.89	19.44	19.42	19.77	19.19	21.43	20.91	20.62	20.17	20.51
742	19.79	20.45	20.07	20.39	20.44	27.03	24.07	23.98	22.25	21.43
743	18.79	20.87	20.13	19.94	21.25	28.32	24.81	25.69	23.48	21.63
744	17.7	20.6	19.51	18.29	19.95	15.56	23.81	14.66	15.67	17.49
745	16.04	19.32	18.1	16.61	18.09	13.39	20.38	13.12	13.93	14.15
746	15.05	17.81	16.79	15.45	16.62	12.97	17.93	12.74	13.46	13.51
747	14.31	16.46	15.65	14.58	15.55	12.15	16.26	11.66	12.7	12.63
748	13.48	15.19	14.6	13.66	14.48	11.49	14.67	11.3	11.92	11.93
…										
3452	14.48	14.29	14.33	14.23	14.37	14.14	14.44	14.38	14.48	22.22
3453	13.2	12.96	12.97	12.87	13.09	12.87	13.16	13	13.11	22.15
3454	12.87	12.78	12.95	12.96	12.8	12.76	12.76	12.77	13.26	22.08
3455	12.05	11.65	11.81	11.73	11.75	11.54	11.82	11.87	11.94	22.06
3456	11.32	10.95	11.1	11.01	11.06	10.83	11.1	11.15	11.17	22.12
…										
3678	23.46	30.33	30.33	23.75	23.81	23.26	23.68	23.68	23.69	22.3
3679	23.83	26.34	27.49	23.54	23.49	23.68	23.81	23.8	23.52	22.16
3680	23.96	24.2	25.64	22.96	22.88	23.85	23.74	23.69	23.07	20.81
…										
8482	15.48	15.85	15.65	15.74	15.85	22.38	22.47	22.71	22.94	20.81
8483	15.07	15.37	15.2	15.1	15.33	22.3	22.34	22.64	22.86	20.64
8484	14.61	14.71	14.81	14.64	14.74	22.26	22.29	22.55	22.77	20.56
8485	14.02	14.2	14.09	14.07	14.16	22.19	22.22	22.42	22.66	20.44

The table 1. shows examples of received, supposedly incorrect measurements, but also highlights the disadvantages of the method. As shown in line 745, the measurements are compared to the mean value determined by the loess regression, and the measurements of the 6,8,9 instruments are different, but the error is not clear. Similarly, in lines 8482-85, where we see that all the measurements were considered to be wrong, one half of the devices measured values of 15, and the other half measured values around 23 (~ 19) defined by loess regression. In such cases, manual verification is required or other methodologies may be used to allow additional filtering.

SUMMARY

Reliable results can only be obtained from reliable data. More and more of the collected data comes from sensor measurements, and these sensors can malfunction and malfunction. Identifying and managing these errors is important, but not a trivial task. Automating it raises a number of questions.

References

[1] Gludovátz A., Bacsárdi L., Industry 4.0 projects' background: Experiences at the wood industrial manufactories, SEFBIS JOURNAL XI/2018, 34-41.

[2] Bencsik G., Random Correlation Methodology as a Tool for Analyzing Random Factor of Scientific Results, SEFBIS JOURNAL XII./2018, 11-20.

[3] Jeffrey S. S. (1998), Smoothing Methods in Statistics (Springer Series in Statistics), Springer

[4] Rice A. J. (2010), Mathematical Statistics and Data Analysis, Third edition, Kindle Edition

[5] https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/