Bayan

by Budi Irmawati

Submission date: 09-Apr-2023 02:12AM (UTC-0500)
Submission ID: 2059415690
File name: 56-78-1-PB.pdf (650.35K)
Word count: 1997
Character count: 10704

Proceedings International Conference on Science and Technology (ICST)
e-ISSN: 2722-7375 pp: 255-260, Vol. 1, Juni 2020

The Handling of Missing Data At Nurul Bayan Geomagnetic Station, Lombok

Khairunnas, [1]Budi Irmawati, [2]Teti Zubaidah, [2]Bulkis Kanata, [1]Arik Aranta, [2]Dwi Ratnasari

[1]Department of Informatic Engineering University of Mataram
[2]Department of Electrical Engineering University of Mataram
Jl. Majapahit 62, Mataram 83125, Indonesia
run.khairunnas@gmail.com, [budi-i, tetizubaidah, uqikanata, arikaranta, dwi.ratnasari]@unram.ac.id

Abstract. The geomagnetic data processing as required in earthquake prediction studies is currently being interesting discussions, particularly in Lombok island. For good quality data processing, it is very important to consider the quality of the data recorded at the geomagnetic stations, whether they were recorded continuously in the order of time sequence or not. The time sequence data are essential in the processing of geomagnetic data, so that data does not lose their histories and make them as valid data. The Nurul Bayan Geomagnetic Station has been operated since April 2018, designated for earthquake mitigation in the Lombok regions. However, the obtained recorded data at this station have gaps or emptiness and they are not written well, so the data do not match the time sequence (one-minutes or-seconds). In this study, 24 hours of one-second geomagnetic data were taken from Nurul Bayan station from August to November 2018. We tackled the missing data using the interpolation method, and then we validated them using k-means clustering. This clustering method we used to see whether the filling gaps give an improvement in data processing. The result is the filling missing data was proved to have a benefit to increase the accuracy in data processing using k- means clustering.

Keywords. Geomagnetic data, Earthquake, Interpolation method, k-means clustering, Nurul Bayan, Lombok.

1. Introduction

Earthquake occurrences are natural disaster that are common in Indonesia. In a past few years massive earthquakes occurred frequently. For example, a year ago Lombok island was hit by a massive earthquake for several months continuously and caused a lot of damage. Those occurrences encourage the researchers to find a better disaster mitigation by studying earthquake precursors **[1]**. Speaking about the earthquake precursor, the researchers have found that geomagnetic signals may be recognized as an earthquake precursor since many years ago. One example is a research that analyzed geomagnetic data using wavelet transformation to determine the quietest time range of geomagnetic signal processing in relation to earthquakes precursor in the region of Lombok **[2]**.

On the other hand, in the data processing, it is very important to consider the quality and condition of the data when they are recorded at the geomagnetic station, whether they are recorded continuously (time series) or not. The data sequence is essential in the processing of geomagnetic data to not lose their histories and guarantee that they are valid. However, the obtained recorded data sometimes have missing records, so they do not match the time sequence (one-minutes or-seconds). The missing data can be caused by device mall functions or eventually due to technical problems like electricity

outages. In Lombok island we have a new station in Nurul Bayan, North Lombok, which operated since 2018. The data from its have missing records.

Regarding all above problems, we have to handle those missing data to get a proper and valid data. This research purpose is to handle missing data recorded at Nurul Bayan Geomagnetic station to get a valid data. The second purpose is to evaluate whether the method to handle the missing data have significant benefit to improve the quality in data processing. We tackled the missing data using an interpolation method, then we validated using k-means clustering. Finally, we evaluated whether the filling missing data improved in the data processing, in this case is clustering.

2. Data Collection and Literature Study

2.1. Nurul Bayan Geomagnetic Station

Nurul Bayan (NRB) Geomagnetic Station was operated since April 2018, but only have scalar measurement tool. After the Lombok earthquake there was an addition of instrument to measure vectors in three axes (X,Y,Z). At that time, the instrument was operated temporarily from the end of August 2018 to January 2019. The limited time was because it was a portable instrument and installed for emergency reason. The NRB station was managed by BMKG (Meteorology, Climatology, and Geophysical Agency) with PUI (Center of Excellence in Science and Technology) Geomagnetic Laboratory of University of Mataram.

Fig. 1. Fluxgate magnetometer used to obtain geomagnetic data at NRB station.

This research elaborated the geomagnetic data collected from NRB station in a range time between August 28th and November 30th 2018. In this research, we only used the total geomagnetic field value (F). Table I shows an example of geomagnetic data used in this research which contain missing data. The record in the 12th seconds was not written in the table. The number of data from the 10th to 21st - seconds should contain 11 records, but there are only 6 ones.

Table I. Data Geomagnetic Which Contain Missing

Datetime	F
20181018145900	45750.79
20181018145900	45751.00
20181018145902	45750.84
20181018145903	45750.69
20181018145904	45750.57
20181018145905	45750.58

2.2. K-Means Clustering

Clustering is the process of grouping data sets into groups of objects in a cluster that have a higher resemblance, but very different from other objects in different clusters. K-means is a clustering algorithm that groups data according to the closest data to the center (centroid). The purpose of k- means is to group the data by maximizing the similarity of data in the same cluster and increasing the

dissimilarity of data between clusters. The measure of similarity of data in one cluster is a distance function. The similarity is assumed to be maximum if the data have the shortest distance to the centroid point [3].

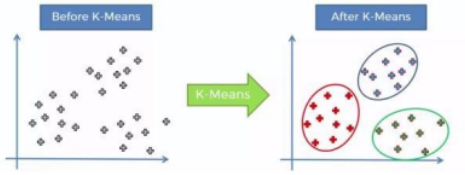


Fig. 2. K-means clustering visualization.

lowchart in Fig 3 shows how k-means clustering works.

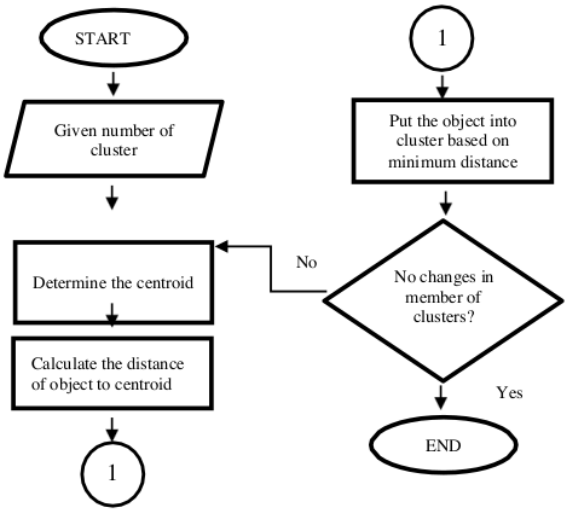


Fig. 3 Flowchart of k-means.

2.3. Euclidean Distance

Euclidean distance is the most popular distance metric. It commonly known as straight line distance as written in Equation (1).

$$D(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} \quad (1)$$

2.4. Silhouette Coefficient

Silhouette Coefficient is used to see the quality and strength of a cluster, how well an object is placed in a cluster. This method is a combination of cohesion and separation methods. Resulted cluster are said to be optimal if the distance between members in the same cluster is small/close, and the distance between members from different clusters is large/far.

2.5. Local Minima

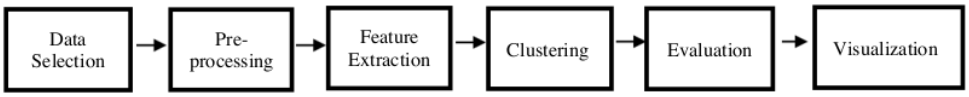
Local minima or local minimum is also known as relative minima which refers to the minimum value among other values and does not include the minimum value of the whole data (global minimum). Determination of local minima is done to obtain anomalous points in the total geomagnetic field. This is because the tendency of total geomagnetic fields is said to be an anomaly when there is a decrease in F value [4].

2.6. Interpolation Method

Interpolation is the process of "guessing" the value of data by paying attention to other existing data. Interpolation is a technique for finding the value of a missing variable in a known range of data. Sequential data usually have a pattern. The pattern can be polynomial or clustered where each pattern is handled differently [5].

3. Methodology

This section explains the proposed methodology to handle the missing data collected from Nurul Bayan Geomagnetic station.



3.1. Data Selection

In this study, the data used are only during 09.00-12.00 and 23.00-03.00 Central Indonesian Time (WITA). With the difference between UTC and WITA is 8 hours, therefore, we used the data ranged between 01.00-04.00 and 15.00-19.00 UTC. Then, we only selected the F and datetime attributes as shown in Tables I.

3.2. Preprocessing

The step that were done in preprocessing.

1. Search the missing records based on datetime column. Table II shows that there are missing records between row 1 and 2 and between row 2 and 3.

Table 2. Geomagnetic Data Which Contain Missing Records

Datetime	F
20181018145910	45750.79
20181018145911	45751.00
20181018145915	45750.84
20181018145916	45750.69

2. Insert the datetime information and set the F column of missing rows to NaN as shown in Table III.
3. Replace the NaN in F column using interpolation as shown in Table V.

Figure 4 show the comparison of data before and after the missing data was handled. The graph shows that 72.599 records data were missing.

3.3. Feature Extraction

Feature extraction is a process to obtain the best part of features that have useful information from the data. In this research, the extracted feature was F. The feature extraction process was carried out by referring to points on the local minima.

After local minima points were obtained, we took ten rows before and after local minima. Then we calculated the difference between two consecutive rows. This extraction resulted total 18 features. The extracted features were clustered using k-means. The example of it result show in figure 4.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	0.11	-0.16	0.13	-0.27	0.27	-0.17	-0.07	-0.10	0.00	-0.26	0.22	-0.13	0.06	-0.09	0.09	0.11	-0.23	0.00
1	0.13	-0.27	0.27	-0.17	-0.07	-0.10	0.00	-0.15	0.16	-0.13	0.06	-0.09	0.09	0.11	-0.23	0.00	-0.10	-0.04
2	0.27	-0.17	-0.07	-0.10	0.00	-0.15	0.16	-0.26	0.22	-0.09	0.09	0.11	-0.23	0.00	-0.10	-0.04	0.04	0.07
3	-0.07	-0.10	0.00	-0.15	0.16	-0.26	0.22	-0.13	0.06	0.11	-0.23	0.00	-0.10	-0.04	0.04	0.07	-0.16	0.09
4	0.22	-0.13	0.06	-0.09	0.09	0.11	-0.23	0.00	-0.10	0.07	-0.16	0.09	0.00	0.00	-0.09	-0.05	0.14	0.06
5	-0.09	0.09	0.11	-0.23	0.00	-0.10	-0.04	0.04	0.07	0.00	0.00	-0.09	-0.05	0.14	0.06	0.00	-0.01	0.18
6	-0.10	-0.04	0.04	0.07	-0.16	0.09	0.00	0.00	-0.09	0.06	0.00	-0.01	0.18	0.00	-0.07	0.00	0.16	-0.04
7	-0.16	0.09	0.00	0.00	-0.09	-0.05	0.14	0.06	0.00	0.00	-0.07	0.00	0.16	-0.04	0.11	-0.08	0.00	0.01
8	0.14	0.06	0.00	-0.01	0.18	0.00	-0.07	0.00	0.16	-0.08	0.00	0.01	0.16	-0.10	0.04	0.06	-0.03	0.00

Fig. 4. Feature extracted.

We extracted features both in the real data and the filled missing data. Both datasets produced different number of local minima as shown in Figure 6.

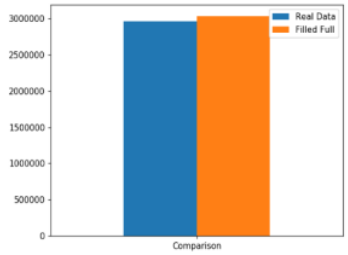


Fig. 5. Number of missing

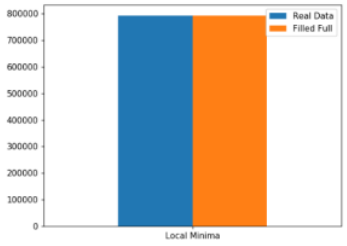


Fig. 6. The number of local minima

Table 3. The Filled Missing Data Data Of Geomagnetic Data

DATETIME	F
20181018145910	45750.79
20181018145911	NAN
20181018145912	45751.00
20181018145913	NAN
20181018145914	NAN
20181018145915	45750.84
20181018145916	45750.69

Table 4. The Filled Missing Using Interpolation

DATETIME	F
20181018145910	45750.79
20181018145911	45750.90
20181018145912	45751.00
20181018145913	45750.95
20181018145914	45750.89
20181018145915	45750.84
20181018145916	45750.69

Table 5. The Number Of Local Minima

DATA	NUMBER OF LOCAL MINIMA
REAL DATA	**793060**
FILLED FULL	792814

Table 6. Accuracy

DATA	ACCURACY
REAL DATA	99,71%
FILLED FULL	**99,72%**

3.4. Clustering

We did cluster on both real data and full data to see whether the filling missing data has impact in data processing, in this case was clustering. The two datasets resulted three optimum number of clusters. By using silhouette coefficient method, the cluster accuracy for both the two dataset shows in Table VI. Our experimental result shows an improvement in clustering accuracy, it may not significant because the data recorded many earthquakes that cause the data consistency is quite low. Therefore, we need to find other method especially in the feature extraction to produce better cluster.

4. Conclusion

The handling of missing data is something that we have to do to get a better geomagnetic data to associate with disaster mitigation. The missing data may indicate that there are unknown phenomena. However, many missing data will affect the computation. Therefore, filling the missing data was proved may have benefit in increase the computational accuracy.

References

[1] B. Kanata, T. Zubaidah, B. Irmawati and C. Ramadhani, "Comprehensive Geomagnetic Signal Processings for Sucessful Earthquake Prediction," in *International Conference on Quality in Research*, 2013.

[2] B. Kanata and T. Zubaidah, "Quiet Time Localization of Geomagnetic Data for Japanese Regions," in *International Conference on Quality in Research*, 2015.

[3] Suyanto, Data Mining untuk Klasifikasi dan Klasterisasi Data, Bandung: Penerbit INFORMATIKA, 2017.

[4] L. G. Hee and M. H. A. Jr., "Mobile Robots Navigation, Mapping, and Localization Part I," in *Encyclopedia of Artificial Intelligence*, 2009, p. 8.

[5] P. Howard. J, Computational Methods for Numerical Analysis with R, CRC Press, 2017.

Bayan

ORIGINALITY REPORT

13%	10%	9%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Mataram Student Paper	5%
2	M A Kamal, D Youlla. "Municipal solid waste vehicle routing optimization based on region clustering of Pontianak City West Kalimantan", IOP Conference Series: Earth and Environmental Science, 2019 Publication	3%
3	Eka Mala Sari Rochman, Sri Herawati, Ach. Khozaimi, Endang Indriyani, Bain Khusnul Khatimah, Aeri Rachmad. "Clustering tourism places in Madura based facilities using fuzzy C-means", AIP Publishing, 2023 Publication	2%
4	Submitted to Pennsylvania State System of Higher Education Student Paper	1%
5	Bulkis Kanata, Teti Zubaidah. "Quiet time localization of geomagnetic data for Japanese regions", 2015 International Conference on Quality in Research (QiR), 2015	1%

	Publication	
6	Submitted to Universitas Sebelas Maret Student Paper	1 %
7	www.igi-global.com Internet Source	<1 %
8	Terttu Luukkonen. "10. Quantitative Techniques in Evaluation in Western Europe", Walter de Gruyter GmbH, 1997 Publication	<1 %
9	te.unram.ac.id Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 3 words
Exclude bibliography	On		

Bayan

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6