Extended isolation forest – application to outlier detection in geomagnetic data

S A Imashev

Research station of Russian Academy of Sciences in Bishkek city, 720049, Bishkek-49, Kyrgyzstan
E-mail: sanzar.imashev@gmail.com

Abstract. The aim of this study is to present a method for detection of outliers in the time series of total intensity of geomagnetic field using Extended Isolation Forest algorithm. The method consists of three steps: 1) generation of additional features that take into account the regular daily variation and smooth behaviour of normal data, 2) detection of potential outliers based on ensemble of extended isolating trees and 3) subsequent refinement based on difference between the outlier and its replacement with interpolated value. Application of the method for detection of outliers in yearly time series of the total geomagnetic field at Ak-Suu and Kegety stations showed that the algorithm identifies both global and contextual outliers. Average classification metrics for the method are characterized as high and have the following values: precision 94.3%, recall 93.9% and F-score 94.5%, and probabilities of errors of the first and second kind are comparable to similar algorithms used for detection of outliers in magnetograms of different sampling rate.

1. Introduction
The time series of the total geomagnetic field contains trend and periodic components as well as the irregular part in the form of various anomalies [1]. The source of these anomalies can be both human activity, as well as natural phenomena, such as magnetic storms and thunderstorm activity [2]. These outliers represent abnormal points that are not similar to neighboring parts and are deviate from the expected "normal" behavior [3]. The presence of these outliers in the time series complicates further data processing and requires a preliminary stage of removing them or replacing with appropriate values.

This paper presents a technique for detection of outliers in the time series of the total intensity of geomagnetic field based on Extended Isolation Forest (EIF) algorithm.

2. Data
Monitoring of the total intensity of the Earth’s magnetic field is carried out at the Research Station of the Russian Academy of Sciences (RS RAS) via the network of observation sites (figure 1).
Figure 1. Stationary observation sites of geomagnetic monitoring network of RS RAS: 1 – Ak-Suu, 2 – Shavai, 3 – Chunkurchak, 4 – Tash-Bashat, 5 – Issyk-Ata, 6 – Kegety, 7 – Karagai-Bulak.

Ak-Suu stationary observation site is the base station of the network with the lowest impact of the industrial noises. Figure 2 shows the time series of the total geomagnetic field at the Ak-Suu station in 2020. The time series contains both global and contextual anomalies [3].

Figure 2. Time series of the total geomagnetic field at Ak-Suu station in 2020 (red - secular trend).

While the global outliers usually have extremely high or low values and thus can be easily distinguished from the normal data, the contextual ones [3] are characterized with the values within the typical (normal) range, but deviate from the typical diurnal variation by their location (figure 3).

Figure 3. Example of global and contextual outliers (Ak-Suu station, 04-09 May 2020).

The dominating part of outliers in the form of spikes was registered in the warm seasons of the year (May-September), which is primarily associated with thunderstorm activity, when the electric currents induced in the Earth causes sudden changes in the normal variations of the geomagnetic field [4].
It is known that the geomagnetic field contains a regular diurnal component, which is caused by external current sources associated with the processes of magnetosphere-ionosphere interaction [4, 5]. Grouping the data by hour with the subsequent calculation of the median value gives an estimation of the average daily profile of the Sq-variations (figure 4). The presence of this periodic component can help in identifying contextual outliers as the deviation from this normal daily profile.

**Figure 4.** Average daily profile of the total geomagnetic field at Ak-Suu station in 2020 (solid blue - median).

### 3. Method

The peculiarities of the total geomagnetic field considered above make it difficult to use the standard methods for outlier detection, such as the z-score [6], interquartile range method [7], Chauvinet test [8], median absolute deviation (MAD) [9] and etc. As a result, more sophisticated methods of anomaly detection should be applied. For example, there are algorithms based on the theory of discrete mathematical analysis and developed for outlier detection in magnetometer data of various sampling rates: SP [2] and SPs [1] algorithms. However, it should be noted that the features of the time series discussed above, as well as the difficulty of manual selection, complicate the labeling of normal and outlier objects for the training set, which turn the problem of outlier detection, in the general case, to unsupervised learning tasks [10].

The isolation tree algorithm is based on binary decision tree. The root of the tree is the entire set of features and at each next node, a random feature and partitioning threshold are selected in the interval from the minimum to the maximum value of the selected feature. In this case, the tree is built until each object appears in a separate leaf. In such way of the tree growth, outliers will most often appear in leaves with a low depth (closer to the root), while to isolate an object from a cluster of normal data, additional levels of partitioning will be required. Thus, the depth of the leaf in the constructed decision tree will be the final answer. The estimate of the average leaf depth, meaning the anomaly score is calculated as:

$$S(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$  \(1\)

where, \(h(x)\) – number of branches in the tree for object \(x\), \(c(n)\) – normalizing constant for a dataset of size \(n\), \(E(h(x))\) – average value of \(h(x)\) for the ensemble of isolation trees. Based on the values of the anomaly score \(S\) for each object in the set, one can classify it as the outlier or normal data.

Unlike the classical isolation forest algorithm [11], in its extended version, the feature space is split not by vertical and horizontal lines, but by lines of random slopes [12].

At the first stage, two additional features, taking into account the peculiarities of the time series, such as slow change in field values and a repeating diurnal variation are generated. The first feature – difference in absolute value between the current and subsequent (in time) field value and the second one – the average value of the at a given time of day, calculated as the median of the field values recorded at the Nth second from the beginning of the day for all days of the time series. The addition of these features helped to increase the quality of the classification. Firstly, the average daily variation acts as a model, helping to identify points with an abnormal field value for a given time of day; and secondly, the
difference applied to normal data (slow changes) gives relatively small values, while for outliers, it takes extreme values and thus, corresponding the derivative. The next step in the proposed methodology is to determine the level of significance for detected outliers, i.e. for practical purposes, it is advisable to identify not all outliers, but only those that exceed a certain threshold, just as an expert in manual processing misses insignificant outliers of small values. For this purpose, the algorithm has a procedure for removing an outlier and filling the missing value with interpolated one with the subsequent calculation of difference between original outlier value and its substitute. This step makes it possible to significantly reduce the number of false positive classifications and the influence of uncertainty in the value of the anomaly score $S$, which defines the boundary between anomaly and normal data. It should be noted that even if we assume that all outliers below this threshold value (for example, 3 nT) are significant, then this step will not affect the final result due to its small amplitudes.

4. Test
The correctness of outlier detection procedure was tested on time series of magnetic field variations at Ak-Suu and Kegety stations for 2020 and 2021 (the first half of the year). For the construction of the confusion matrix, the outliers in the time-series were firstly detected manually by an expert. Classification metrics for automatic detection of outliers in the time-series are presented in table 1.

| Metrics  | Ak-Suu 2020 | Ak-Suu 2021 | Kegety 2020 | Kegety 2021 |
|----------|-------------|-------------|-------------|-------------|
| TP       | 256         | 74          | 2361        | 746         |
| FN       | 6           | 4           | 157         | 89          |
| FP       | 13          | 10          | 52          | 31          |
| TN       | 1580845     | 652232      | 1578550     | 651454      |
| Precision| 0.9517      | 0.8809      | 0.9785      | 0.9601      |
| Recall   | 0.9771      | 0.9487      | 0.9376      | 0.8934      |
| F-score  | 0.9642      | 0.9136      | 0.9576      | 0.9256      |
| P1       | 0.0229      | 0.0513      | 0.0624      | 0.1066      |
| P2       | 0.0483      | 0.1191      | 0.0215      | 0.0399      |

Based on the values of precision, recall and F-score, the classification results can be characterized as quite good. An additional metrics such as probability of errors of I and II types ($P_1$ and $P_2$) show the values close to the levels, corresponding to independent examination of the SP algorithm ($P_1 = 0.01$, $P_2 = 0.09$) [2] and SPs ($P_1 = 0.14$, $P_2 = 0.089$) [1].

For independent control of classification quality one can use lag-plots, where the normal data points are spread along the diagonal due to strong autocorrelation and the outliers are distributed randomly with the specular reflection with respect to the diagonal. Figure 5 shows that for both of the stations all the global and vast majority of contextual outliers are detected.
5. Conclusion

The presented method of detection of outliers in time series of the total intensity of geomagnetic field is based on Extended Isolation Forest algorithm. Generation of additional features that take into account the slow variations and Sq-periodicity of normal data, as well as replacement of detected outlier with interpolated value, significantly improved the quality of outlier detection compared to the results obtained using isolation forest for the original test dataset with only one feature.

Application of the methodology for detection of outliers in time-series of the total geomagnetic field at Ak-Suu and Kegety stations showed that it makes it possible to classify not only global outliers, but also contextual ones. The average classification metrics of outlier detection for these time series can characterize the result as quite good: accuracy 94.3%, completeness 93.9% and F-score 94.5%. Probabilities of errors of the first and second kind are comparable with the corresponding values of the special-purpose algorithms SP and SPs used for detection of outliers in magnetograms.

6. Reference

[1] Soloviev A A, Agayan S M, Gvishiani A D, Bogoutdinov S R and Chulliat A 2012 Recognition of disturbances with specified morphology in time series: Part 2. Spikes on 1-s magnetograms Izvestiya, Physics of the Solid Earth 48 pp 395–409
[2] Bogoutdinov S R, Gvishiani A D, Agayan S M, Solovyev A A and Kihn E 2010 Recognition of disturbances with specified morphology in time series. Part 1: Spikes on magnetograms of the worldwide INTERMAGNET network Izvestiya, Physics of the Solid Earth 46 1004–16
[3] Chandola V, Banerjee A and Kumar V 2009 Anomaly detection ACM Comput. Surv. 41 1–58
[4] Campbell W H 2012 Introduction to Geomagnetic Fields (Cambridge University Press) p 337
[5] Manda M and Korte M 2011 Geomagnetic observations and models vol 5 (Dordrecht: Springer) p 344
[6] Kreyszig E, Kreyszig H and Norminton E J 2011 Advanced engineering mathematics (Hoboken
[7] Williamson D F, Parker R A and Kendrick J S 1989 The box plot: a simple visual method to interpret data *Annals of internal medicine* **110** 916–21

[8] Milosan I 2017 Application of the Chauvenet criterion to detection of the aberrant data obtained in the industrial processes *AFASES 2017* **19** 389392

[9] Leys C, Ley C, Klein O, Bernard P and Licata L 2013 Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median *Journal of Experimental Social Psychology* **49** 764–6

[10] Mohri M, Rostamizadeh A and Talwalkar A 2018 *Foundations of machine learning (Adaptive computation and machine learning)* (Cambridge, Massachusetts: The MIT Press) p 504

[11] Liu F T, Ting K M and Zhou Z 2008 Isolation Forest *2008 Eighth IEEE Int. Conf. on Data Mining (Pisa)* 413-22

[12] Hariri S, Kind M C and Brunner R J 2021 Extended Isolation Forest *IEEE Trans. Knowl. Data Eng.* **33** 1479–89

**Acknowledgements**

The work was performed within the framework of the state assignment for Federal State Budgetary Institution of Science Research Station of the Russian Academy of Sciences in Bishkek city (No. AAAA-A19-119020190064-9).