Translated to English from paper in Japanese, originally published in: Technical Report of the Kakioka Magnetic Observatory, 1, No.1, 1–9, 2003 with permission of the authors.

On automatic determination of criteria to detect outliers for the absolute measurement of the geomagnetic field

by

Nobukazu Ito[1] and Ikuko Fujii[2]

[1]Kanoya Magnetic Observatory
[2]Kakioka Magnetic Observatory

Received 31 January 2003; received in revised form 3 March 2003; accepted 7 March 2003

Abstract

We propose a procedure for automatic determination of criteria to distinguish spurious baseline values in the absolute measurement of the geomagnetic field based on a robust statistical method. The data used in this study are baseline values of the horizontal force (H), declination (D), and vertical component (Z) observed at Kanoya Magnetic Observatory from February 25, 2000 to December 31, 2000. The absolute measurement sessions were conducted 58, 58, and 72 times for H, Z, and D, respectively, in the period of interest. Eight baseline values were obtained in each session, giving a total of 464 baseline values each for H and Z, and 576 for D. The data set of each session has too few samples to be analyzed with a robust procedure, so the whole data set comprising 464 or 576 values was used. After the median was subtracted from the data of each session, a set of the residuals for all sessions was analyzed statistically. As a result, we found that the residuals of all sessions follow a Gaussian distribution, except for a small portion of the data set.

Quantile-Quantile (QQ) plots of the residuals normalized by the median absolute deviation (MAD) were made for each component and were used to estimate the outliers. We assumed that the outliers are normalized residuals larger than 7 for H and Z, and 5 for D, at which discontinuities were seen in the QQ plots for each component. As a result, 8, 12, and 13 outliers were detected for H, Z, and D, respectively. These criteria are slightly more robust than those estimated as three times the standard deviation that are computed from the data set without the outliers.

The results of the robust procedure used in this study indicate that the outliers can be automatically detected and that the procedure could detect outliers hidden in each session. For the baseline values in the year 2000, 5 and 7 times the MAD are good estimates of the criteria to detect the outliers. Further investigation on the robust criteria by analyzing data for other time periods will be required to improve the accuracy of the procedure.

1. Introduction

The Kakioka Magnetic Observatory observes each geomagnetic vector at a given point of time in two modes: variation observation, by which variations in the vector are measured continuously, and absolute observation, which aims at collecting baseline values. An absolute observation is conducted about once each week using a proton magnetometer and a magnetometor theodolite to measure the three components of a geomagnetic vector and compare them with the geomagnetic data collected by variation observations conducted

at a comparable point of time to derive observed baseline values.

In each session of absolute observation, eight observed baseline values are essentially averaged for each set of three components to determine the baseline value for that session. However, if any outlier is involved in the eight observed baseline values, it would hinder the estimation of the true baseline value. Although omission and repeated observation are used to detect and remove outliers, it has been pointed out that, without a quantitative standard of outlier detection available, observers are divided in their interpretation of the criteria of outlier omission.

This paper reports our attempt to quantify an outlier omission standard by stretching the concept of robust estimation (Huber, 1981) designed for data involving outliers to actual observed baseline values.

2. Robust estimation

When a set of observed values of interest involves outliers with alien characteristics, robust statistics (Huber, 1981) that probe into the characteristics of the distribution of observed values become useful with the effects of the outliers being removed. The sections that follow describe the techniques applied in this report, with regard to parameters of the distribution of observed baseline values involving outliers, and how to detect and remove the effects of outliers.

2.1 Parameters of the distribution of observed values involving outliers

Assume that a set of observed values (samples) $\{x_i\}$,$i = 1 \sim n$ has been derived after n sessions of repeated observation to estimate a given physical value U. If observation errors have averaged 0 and they are limited to random errors conforming to a normal distribution of standard deviation E, then the best estimators of physical value U and scale E of measurement dispersions are provided by sample mean μ and sample standard deviation σ, respectively, which are expressed in equations by,

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i, \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2.$$

However, if outliers that do not conform to the distribution of E are involved in the observation errors, μ and σ would be calculated with the outliers taken into account in the computation process, which could result in values remote from U and E depending on the amplitude of the outliers.

Those estimators that are scarcely vulnerable to the effects of outliers are called "robust estimators." As explained above, μ and σ are not robust estimators. The median and the median absolute deviation (MAD), on the other hand, are scarcely vulnerable to the effects of outliers and are known robust estimators of U and E, respectively. Median M is the n/2nd value in a descending order of n samples x_i, while the MAD is defined as a median in a descending order of absolute values $|x_i\text{-}M|$ of x_i and M, excluding 0.

An example demonstrates that the M and MAD are robust estimators. The sample set shown below is that of observed baseline values (unit: nT) in the H-component collected by an absolute observation on April 19, 2000 at the Kanoya Magnetic Observatory.

{33.33, 34.61, 34.54, 34.62, 34.41, 34.68, 34.79, 34.59}

Mean μ_1, standard deviation σ_1, median M_1 and median absolute deviation MAD_1 of the eight samples are 34.45, 0.46, 34.60 and 0.07 nT, respectively. Since the first baseline value of 33.33 is one that has been tested as an outlier by the analysis in this report (see Chapter 4), mean μ_2, standard deviation σ_2, median M_2 and median absolute deviation MAD_2 of the seven samples, with the first sample being removed, were calculated as 34.61, 0.12, 34.61 and 0.07 nT, respectively.

The rate of change in the mean was 0.46%, when compared with 0.03% for the median. This example indicates that the median has been less influenced by the presence of the outliers than the mean. Since the mean changed 0.16 while the standard deviation with the outlier being removed was 0.12, the change in the mean is a significant value in relation to the dispersion of the samples. So are the standard deviation and the median absolute deviation. Though about a four-fold difference existing between σ_1 and σ_2 gives them

the appearance of being distributions varying totally in the dispersion of samples, MAD_1 and MAD_2 when compared have no difference, evidencing little outlier influence.

Next, whether the median represents a true baseline value in this sample set was considered. If the sample set with the outlier removed is assumed to conform to the original distribution, then the median gets closer to the mean resulting from the removal of the outlier. Accordingly, the median is considered to play the role of an estimator of the true baseline value in such a set of samples involving outliers.

2.2 Detecting outlier

Techniques are available for using μ and σ to detect outliers involved in a population of samples. If a given sample population x_i conforms to a normal distribution, the probability of x_i expanding beyond the bounds of $\mu \pm 3\sigma$ is only 0.3 %. These techniques include defining x_i as an outlier when it meets the relation

$$|x_i - \mu| > 3\sigma \quad (1)$$

(3σ edit rule), and Thompson's critical test, in which the threshold varies with the number of samples. However, since the presence of outliers sometimes prevents μ and σ from showing a correct distribution of samples as explained in the preceding section, these methods have been pointed out to fail more often than not (Pearson, 2002).

The use of a robust M and a robust MAD in place of μ and σ would make it possible to yield an objective, quantitative estimation of the distribution of samples even if outliers are involved in their population. A modified version of the 3σ edit rule in which the components μ and σ are replaced with M and MAD, respectively, is called "Hampel's Theory for Robust Estimation" (Pearson, 2002). In this report, analyses are conducted using a similar technique, but the threshold is not confined to 3 MAD, and the value of x_i that meets the relation

$$|x_i - M| > kMAD \quad (2)$$

is defined as an outlier. In this relation, k is a parameter that determines the threshold.

2.3 Removing the effects of outliers

When outliers are detected in a population of samples in the course of the practice of the method introduced in Section 2.2, two broad routes are available to determine the statistical characteristics of the sample set with the effects of the outliers being removed. One is the hard rejection method, whereby the outliers are removed completely to allow sample x_i to conform to the original distribution of U. The current scheme of outlier omission employed in the Observatory works on this principle. The other is the soft rejection method, whereby the outliers are weighted down to bring sample x_i closer to the original distribution of U.

This report did not go as far as to debate the methodology of outlier removal.

3. Characteristics of observed baseline values

Figure 1 shows the baseline values observed at the Kanoya Magnetic Observatory from February 25, 2000 to December 31, 2000, along with their median and MAD. Because the presence of possible outliers is already known, the robust estimators of the median and MAD are used. Because the geomagnetic field is a vector that has magnitude and direction, it is represented by the three components of the horizontal force (H), vertical component (Z) and declination (D) at the Kakioka Magnetic Observatory. The observed baseline values are stated in nT for the H- and Z-components and in minutes for the D-component.

Absolute observations were conducted at about one-week intervals and were iterated until sets of eight observed baseline values each were collected. Due to repeated sessions of observations conducted occasionally, the frequency of the observations and the number of baseline values observed in each session of observation are unknown. Fifty-eight sessions of absolute observation were conducted for the H- and Z-components over the period shown in Figure 1, with 72 sessions for the D-component, yielding 464 observed baseline values for the H- and Z-components and 576 for the D-component. The more sessions and more observed baseline values for the D-component are the result of the many sessions of repeated observation involved.

Figure 1 shows that the median varies with time as it reflects secular and seasonal changes, but the population of observed baseline values derived from each session of observation distributes in about the same region around the median. Data out of a population is occasionally observed.

The objective of each session of absolute observation is to estimate the true baseline value U from x_i,i = 1,...n (where x_i represents H, D or Z). As stated earlier, the estimator μ of a baseline value is determined from the mean of n observed baseline values under the current scheme. If, in a set of n observed baseline values, any single value is found to singularly deviate from others, it is assumed as an outlier, and the baseline value is established from the mean of the observed baseline values with the outlier being removed.

To quantify the criteria of outlier detection, to which statistical distribution the observed baseline values conform, was first examined. Figure 2 shows the probability density function F (x_i) based on the assumed conformance of the observed baseline values derived from a typical day to the normal distribution.

$$F(x_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \quad (3)$$

Since only eight samples of observed baseline values are available for the day, whether outliers are involved in a widely or narrowly dispersed distribution is not clear.

Robust estimation procedures, such as omission, are an operation in which x_i is assumed to conform to a normal distribution and nonconforming data is removed as an outlier for the rest of the data to match the normal distribution. If, however, the distribution to which x_i conforms is unclear, as in Figure 2, the detection of outliers assuming a normal distribution could suffer degraded accuracy. In the distribution shown in Figure 2, for example, since M is 34.25 and MAD is 0.12, the smallest observed baseline value is 2.5 times more remote from M than from MAD, or 4.3 times more remote if the MAD of 0.07 nT (see Chapter 4) in a sample population closer to a normal distribution is used. Assuming k = 4 in Eq. 2, the same sample will be determined as an outlier depending on the choice of the value of MAD.

Figure 1 shows that the median for each session of observation changes with time but the difference between the median in a given session of observation and n observed baseline values is dispersed likewise throughout the time. Assuming that the j-th observed baseline value is x_i^j, the number of occurrences is n_j and median is x_i^j, residual $\Delta x_i^j = x_i^j - x_M^j$ was calculated with regard to every occurrence of i and j to examine the probability distribution (histogram) and probability density function of the residuals when they were viewed as one sample set. Figure 3 shows the H-component. The number of occurrences is charted at intervals of 0.1 nT in the histogram.

The 464 samples available in Figure 3 formed a distribution clearer than in Figure 2. The distribution resembled a bell centering on 0 superimposed with a few large-amplitude samples. Since the probability distribution in Figure 3 is essentially represented by the probability density function defined by Eq. 3, this distribution can be essentially regarded as a normal distribution. This hints that the estimated Δx_i^j of observation errors is homogeneous regardless of individual differences among observers and seasonable changes, except in a few exceptional cases, or it is a random variable conforming to the same distribution.

Based on Figure 3, using one year's supply of residuals, Δx_i^j might well make up for the shortage of samples, making for robust parameter estimation of a distribution of residuals.

4. Analyses

In this chapter, the robust estimation process outlined in Chapter 2 is applied to observed baseline values to detect outliers involved in the data. As mentioned in the preceding chapter, each set of observed baseline values derived from an individual session of repeated observation has its accuracy of distribution parameter estimation degraded due to the limited number of samples available, making the successful implementation of the outlier detection scheme represented by Eq. 2 unpredictable. Using one year's supply of residuals Δx_i^j, however, could make up for the shortage of samples, making for effective outlier detection. Therefore, the MAD was determined using one year's supply of residuals Δx_i^j to detect outliers in

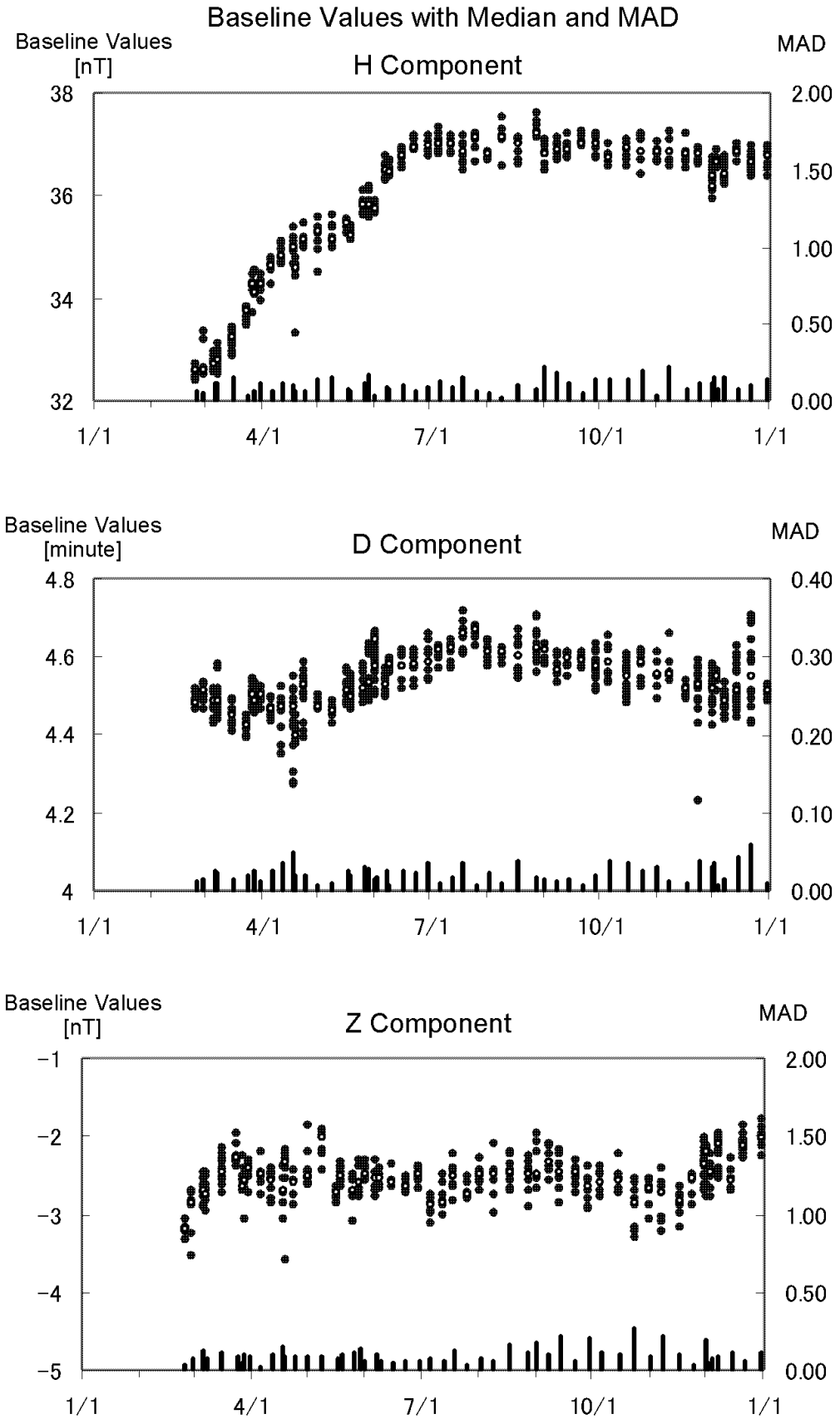

Fig. 1 Changes with Time in the Baseline Values, Medians and MADs of the H (top), D (middle), Z (bottom) Components Observed at the Kanoya Magnetic Observatory from February 25, 2000 to December 31, 2000. The observed baseline values are encircled in black, the medians in white. The MADs are charted in a bar graph at the bottom.

the residuals that meet the modified version of Eq. 2:

$$\frac{|\Delta x_i^j|}{MAD} > k \quad (4)$$

Since Δx_i^j is a value that has already been translated for a median of 0, the process of calculating the difference from the median on the upper left side can be saved.

First, the MAD is calculated. The MADs of one year's supply of residuals Δx_i^j were 0.07 nT, 0.022 minutes and 0.06 nT for the H-, D- and Z-components, respectively.

The next step was to establish the value of parameter k, which in turn determines a threshold. A Quantile-Quantile plot (QQ plot) is available as a convenient means of gaining a rough visual measure of outliers and their threshold. A QQ plot is a plot of an ascending order of residuals normalized with the MAD $\Delta x_i^j/MAD$, with the order being plotted on the horizontal axis and the normalized residual values on the vertical axis (Figure 4). When the residuals conform to a normal distribution, $\Delta x_i^j/MAD$ in the QQ plot are known to distribute in linear form (e.g., Fujii and Schultz, 1999). Outliers are plotted as values deviating significantly from the linear line.

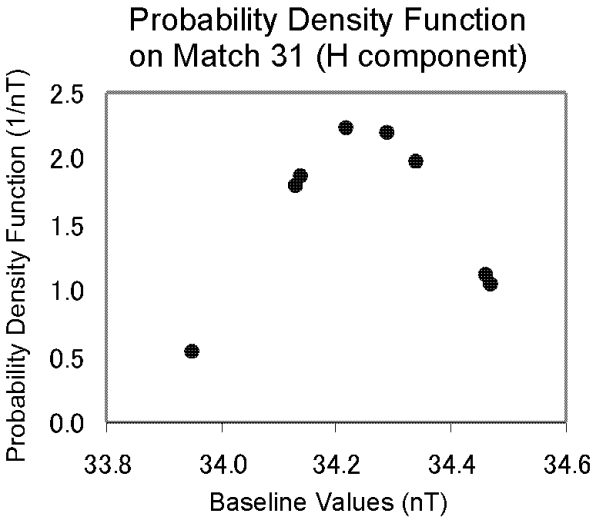

Fig. 2 Probability Density Functions of Observed Baseline Values of the H-Component Observed on March 31.

In Figure 4, all three components are linear in the middle in the plot and are either curved or intermittent in the smallest and largest parts of the order. If the purpose of outlier detection is simply to locate apparently abnormal observed baseline values on the conservative side, the candidates would be those deviating from the sequence of samples in a QQ plot. Since the H- and Z-components lost continuity with a residual/MAD size of about 7 and the D-component did so with a residual/MAD size of about 5, thresholds have been set at k = 7 for the H- and Z-components and at k = 5 for the D-component. This is equivalent to determining, as outliers, the residual amplitudes of the H-, D- and Z-components in excess of 0.49 nT, 0.11 minutes and 0.42 nT, respectively.

When outliers were detected in this way, a total of 33 residuals were determined as outliers, including eight in the H-component, 13 in the D-component and 12 in the Z-component. Figure 5 shows the observed baseline values and outliers. Obviously, data that is difficult to evaluate from

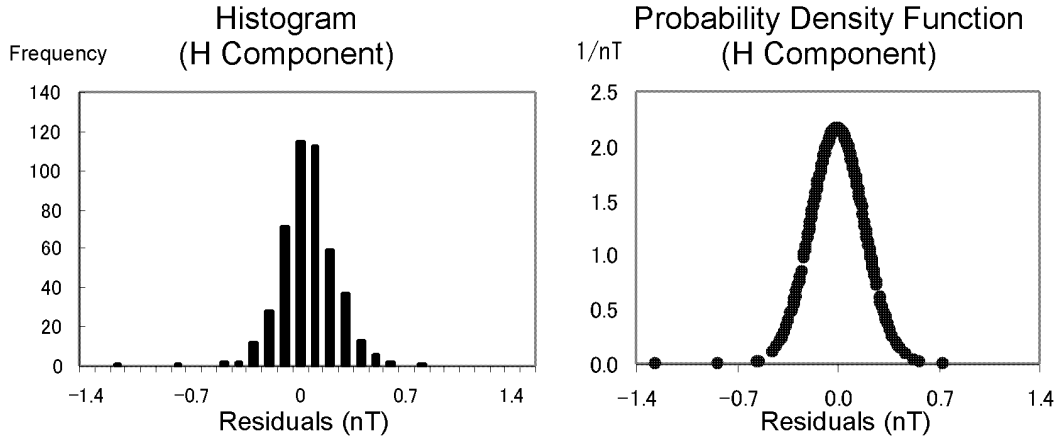

Fig. 3 Histogram (left) and Probability Density Function (right) of the Residuals of the H-Component Observed from February 25,2000 to December 31,2000.

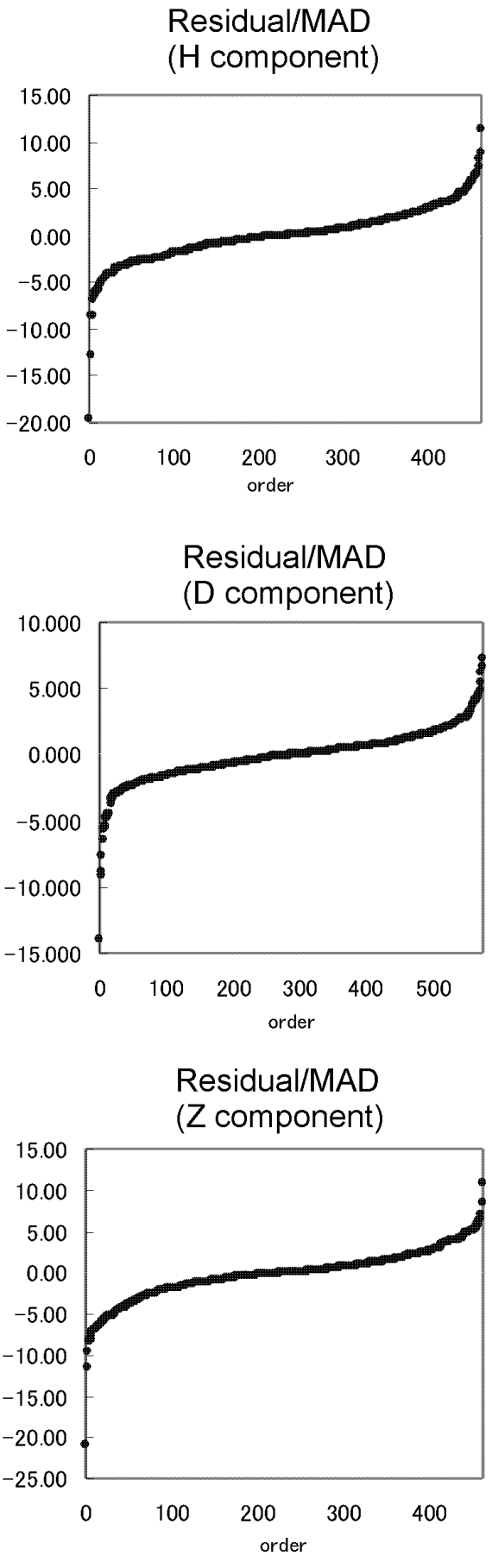

Fig. 4 QQ Plot of H (top), D (middle), Z (bottom) Components.

the status of distribution for the day, as well as apparently abnormal observed baseline data, has been detected.

The smaller the value of k, the smaller the thresholds become, allowing more values to be determined as abnormal. To determine what significance the thresholds used here have, mean of each session μ^j and standard deviation of residuals of one year's supply σ were calculated to examine $x_i^j - \mu^j$ except for the outliers detected from the residuals in the three components. Of the total of 33 outliers, 28 exceeded the range of $\pm 3\sigma$, including seven in the H-component, thirteen in the D-component and eight in the Z-component. This hints that the scheme of outlier determination used in this report was based on a standard slightly stricter than the 3σ edit rule.

5. Discussions

5.1 Comparison with the actual status of omission

The results of outlier detection in Chapter 4 and the actual omission conducted are compared to discuss the difference between the scheme of outlier detection outlined in this report and the current scheme of outlier omission.

Among the 33 abnormal observed baseline values detected in this report, one in the H-component (−19.54 in Figure 4) and one in the Z-component (−21.00 in Figure 4) were omitted without repeated observation, 11 were omitted to invoke a repeated observation and 20 were not omitted.

Among the observed baseline values that had not been tested abnormal, the observed baseline value (1.209 in Figure 4) in the D-component has been omitted singly, but the reason is unknown.

A review of the observed baseline values that had not been omitted indicated that multiple copies of observed baseline values closer to thresholds were often present on the dates of their observation. These observed baseline values might have occurred because they could not be tested in a single session of absolute observation.

Thus, the use of the method of outlier determination outlined in this report not only makes for automated, objective determination but can offer two advantages: (1) determination of outliers without requiring repeated observation,

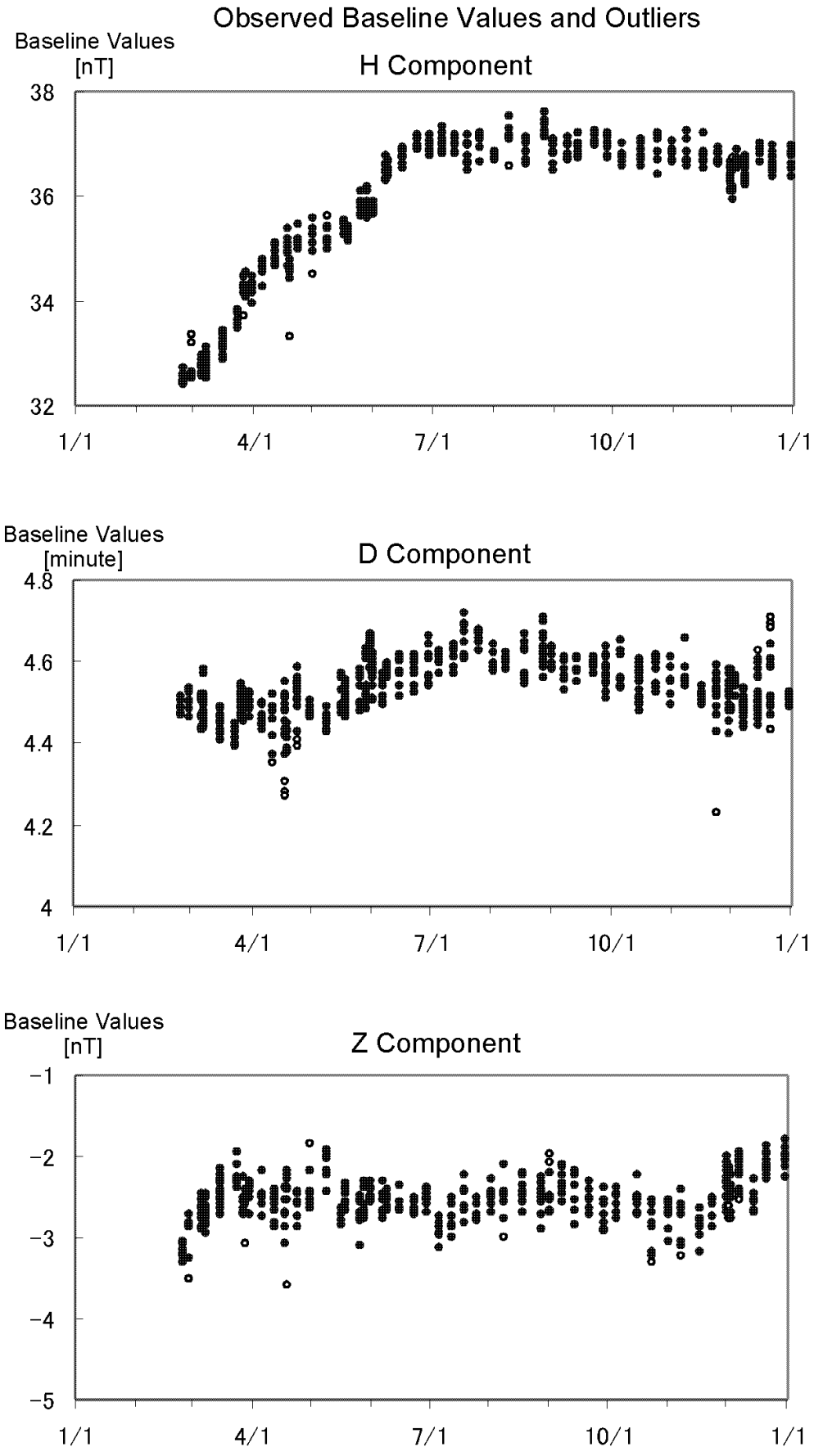

Fig. 5 Observed Baseline Values and Outliers of H (top), D (middle), Z (bottom) Components. The observed baseline values are encircled in black, the detected outliers in white.

and (2) detection of outliers that would not be located by single sessions of absolute observation.

5.2 Thresholds

In this report, thresholds have been set at k = 7 for the H- and Z-components and at k = 5 for the D-component to allow for continuity in a QQ plot. Further studies are necessary to give objectiveness to the determination of the thresholds.

Fujii and Schultz (1999) state that thresholds of 3 to 12 are empirically used, depending on the characteristics of the data of interest, whereas Xu (1989) and Xu (1998) argue that the application of such thresholds to specific cases of robust estimation would be limited because the thresholds are of no probabilistically relevant significance but are selected arbitrarily. General significance might be attached, however, to thresholds based on the distribution of residuals with the effects of outliers being removed as stated in Chapter 4.

Thresholds of k = 5-7 may offer some measure of the observed baseline value data for the year 2000 dealt with in this report. For these thresholds to be of practical value as criteria of outlier omission, data for other years should also be examined to establish the value of k.

5.3 Number of samples

The foregoing sections have demonstrated that outliers can be detected automatically by applying the concept of robust statistics. If outliers are selected as the present object of omission, baseline values closer to the true value may be obtained. But reductions in the number of samples available emerge as a concern in this situation. As reviewed in Chapter 3, each session of repeated observation comes up with too few observed baseline values to apply a normal distribution. Moreover, as the number of samples available declines further, the mean of a baseline value marked by a larger error bar would result. It would be necessary to determine how many observed baseline values at least would be needed to assure required observation accuracy, as well as formulate the criteria of outlier omission and also consider rules for initiating a repeated observation if the number of observed baseline values available falls below this level as a result of outlier omission.

The number of samples n needed to achieve the accuracy required for an absolute observation with a probability of 95% was calculated for the reader's reference. When samples of size n have standard deviation E, the maximum error $U-\mu$ between the true value U and the estimator μ becomes smaller than $1.96E/\sqrt{n}$ with a probability of 95%. If the maximum tolerable error for baseline values in the H-component is 0.10 nT and if E is substituted by sample standard deviation σ = 0.16 nT, with outliers being removed, since the value of E is unknown, then, n that meets the condition of the 95% confidence limit would be expressed by,

$$1.96 \times 0.16/\sqrt{n} \leq 0.10 \quad (5)$$
$$\therefore n \geq 9.83$$

Assuming that the H-component of the absolute observation conforms to a normal distribution, at least 10 samples would be needed to achieve the accuracy required for the absolute observation with a probability of 95%. A somewhat smaller value of n may suffice if the accuracy can be reduced.

6. Conclusions

This study attempted to detect outliers based on the concept of robust statistics proposed by Huber (1981) using one year's supply of observed baseline values in the H-, D- and Z-components recorded at the Kanoya Magnetic Observatory.

Outlier detection is conducted by determining the median and MAD from a population of samples and normalizing the difference between each sample and the median with the MAD to provide an object assessment of how remote the sample is from the center of the distribution. The number of samples in each session of absolute observation handled in this report, however, was generally as small as eight or so - too few to assure the accuracies of median and MAD calculations - that a degraded accuracy of outlier detection was feared. For this reason, one year's supply of the differences between the baseline values observed in each session of absolute observation and the median worked out in that

session were defined as a sample population for the purposes of outlier calculation.

The one year's supply of residuals were essentially distributed normally around 0 among the 464 samples in the H- and Z-components and 576 samples in the D-component, with only a few at the ends of the distribution. The H-, D- and Z-components had a MAD of 0.07 nT, 0.022 minutes and 0.06 nT, respectively. Outlier determination thresholds were set at points of discontinuity in the distribution in a QQ plot to assume as outliers the baseline values with their differences from the median at least seven times greater than the MAD for the H- and Z-components and those with their differences from the median at least five times greater than the MAD for the D-component.

There were eight, 13 and 12 outliers detected in the H-, D- and Z-components, respectively. Using the mean and standard deviation of the residuals calculated with the outliers excluded, the differences between the outliers and mean were normalized with the standard deviation to deliver an equivalent of the result attained with a threshold somewhat lower than 3. The 13 outliers were so identified under both the previous and current schemes of outlier determination. The other 20 outliers detected this time were seen to often hold observed baseline values closer to more than threshold in one session of absolute measurement.

Thus, testing outliers with a standard set using one year's supply of observed baseline values has made it possible to detect outliers that would have been hidden by day-to-day observations, thereby hinting that robust estimation could make for greater omission work simplicity. A future task facing us would be to analyze more cases to augment the objectiveness of this scheme of outlier detection from observed baseline values.

References

Fujii, I. and Schultz, A., On Data Processing Methods for the Geomagnetic Observatory Network: Part I, Proceedings of Conductivity Anomaly Symposium, 97-104, 1999

Huber, P.J., Robust Statistics, John Wiley, New York, 308pp, 1981

Pearson, R., Cleaning Data with a Nonlinear Filter, EDN Japan, 2002

Xu, P., Consequences of Constant Parameters and Robust C/R Confidence Intervals, ABSTRACTS, 1998 Japan Earth and Planetary Science Joint Meeting, 93, 1998

Xu, P., Statistical Criteria for Robust Methods, ITC Journal, No. 1, 37-40, 1989