Unique identifier: 10.1088/1755-1315/929/1/012016

Geometric estimates of variations in space geodetic data

S I Kuzikov
Research station of Russian Academy of Science in Bishkek city, 720049, Bishkek-49, Kyrgyzstan
E-mail: ksi-63@yandex.ru

Abstract. Well-known statistical parameters have some disadvantages when analyzing space geodetic data. Geometric parameters are proposed here for estimating the variation properties of samples for various discrete datasets. The proposed parameters are logically related to each other and are based on the simplest well-known statistical parameters; they do not depend on the type of distribution of the sample under study. "Variation asymmetry" shows the shift of the arithmetic mean relative to the center of the variation interval in the units of the studied sample. "Density of variation" characterizes the level of average variability in sample units. This parameter has several times greater discriminatory sensitivity to extremely different types of variations than linear and standard deviations. The relative parameter "proportion of maximum density" shows the closeness of variation to a uniform distribution in the ranked sample and complements the indicator of variation density. An algorithm for separating different structural levels of the useful signal from emissions (noise) is proposed here based on the calculation of geometric characteristics. The iterations of dividing the sample into structurally homogeneous segments can be stopped at the level of the proportion of maximum density ≥0.9 when analyzing real GPS coordinates.

1. Introduction

The problem of separating the "useful signal" from the so-called "noise" arises when analyzing almost all digital data obtained when measuring any quantities. The concept of a useful signal is intuitively clear, but we have not found its general formal definition or description. The term noise is characterized in the encyclopedic dictionary as interference that distorts signals in the process of transmitting information over a communication channel [1].

Discrete data sets on the position or speed of movement of points are subject to analysis in the study of modern movements of the earth's crust using the methods of space and ground geodesy. In most cases, such data is obtained from signals of Global Navigation Satellite Systems (GNSS), GPS data is the main source of high-precision positioning for scientific purposes. We can get the series of intraday geocentric Cartesian coordinates XYZ when processing field GPS measurements in the GAMIT/GLOBK [2] or Bernese GNSS Software [3] programs. In this case, the time series of coordinates will be 2880 positions per day with the discreteness of recording information from the visible constellation of satellites in 30 seconds. A redundant set of XYZ coordinates gives a statistically more accurate 1-day average position than each individual of the 2880 positions. In the case of a stationary site, we a priori assume that on that day it was at rest and all its variations of coordinates were realized due to limitations on the accuracy of the measurement and calculation method.

Special methods for analyzing the daily average positions of several stations located at a distance of the first kilometers from each other make it possible to identify real relative displacements per day
of the active station by ≥ 2 mm relative to immovable sites [4]. It is possible to increase the level of accuracy in the position of the rectilinear horizontal trend and the average daily position of the station if significant deviations from this trend are excluded from the intraday series of coordinates. In the case of averaging the intraday series of coordinates, deviations of the values from the horizontal trend can be qualified as noise that degrades the accuracy of the useful signal.

We also use another type of time series of coordinates, where the trend in station displacement is fixed on the basis of average daily positions over several years of observation. Such data are presented in local topocentric NEU coordinates and the average rate of the station's annual displacement is determined from the inclined rectilinear trend. For the purpose of determining the average annual speed, the coordinates that slightly deviate from the inclined straight line will be considered a useful signal.

In this context the term "variation" should be recalled as small changes of independent variable [1]. This raises a fundamental question about the level of "smallness", how much is it? In the case of studying discrete sets of coordinates and their derivatives, by the "variation" of the parameter under study, we mean the entire set of measured values, without taking into account the "smallness". Depending on the task at hand, the "useful signal" can be at different levels of coordinate values and change places with the concepts of "interference or noise". Consider the generally accepted statistical parameters that can be used to analyze variations in samples of space geodetic data.

2. Features of traditional statistical parameters for the analysis of variations in samples

In statistical analysis, a set of parameters is often used for samples with a fixed number of its members (N), a description of this can be found in various textbooks, for example [5]. Among them are the minimum (\(MIN_X\)), the maximum (\(MAX_X\)), the variation interval (\(I_X\)) and the arithmetic mean of the sample:

\[
I_X = MAX_X - MIN_X, \quad X = \frac{\sum_{i=1}^{N} x_i}{N}
\]

(1)

In addition to the elementary characteristics of the sample, the standard deviation is usually given, less often the mean linear deviation:

\[
\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{X})^2}{N-1}}, \quad d = \frac{\sum_{i=1}^{N} |x_i - \bar{X}|}{N}
\]

(2)

These characteristics give an idea of the dispersion of the sample values relative to its mean. However, when analyzing GNSS coordinates, the study of variations within the variation interval is of greater interest. Coefficients of variation (\(V_v\)) and oscillations (\(V_l\)), linear coefficient of variation (\(V_d\)) are used to compare the level of variation of the sample relative to the mean:

\[
V_v = \frac{\sigma}{\bar{X}} \times 100\%, \quad V_l = \frac{I_X}{\bar{X}} \times 100\%, \quad V_d = \frac{d}{\bar{X}} \times 100\%
\]

(3)

These characteristics carry information about the level of variability of the parameter in units of the sample mean, but they do not contain information about the internal variation structure. GNSS XYZ coordinates have a dimension of about \(10^{-16}\) mm and a level of variation up to \(10^{-7}\) mm. Therefore, the coefficients (3) will have a dimension of the order of \(10^{-7}\) or less, which is of no practical interest. And for the NEU coordinates, the average position of the station has zero coordinates, which generally excludes the possibility of calculating the coefficients (3).

The last group of traditional statistical parameters includes Pearson skewness (\(A_p\)), cubic skewness (\(A\)) and kurtosis (\(E\)):

\[
A_p = \frac{\bar{X} - M}{\sigma}, \quad A = \frac{\sum_{i=1}^{N} (x_i - \bar{X})^3}{N\sigma^3}, \quad E = \frac{\sum_{i=1}^{N} (x_i - \bar{X})^4}{N\sigma^4} - 3
\]

(4)

The coefficient of Pearson skewness shows the level of deviation of the mean from the mode (\(M\)) in fractions of a standard deviation. Another skewness coefficient shows the level of displacement of
the modal part of the frequency plot relative to the average. The kurtosis shows at how much modal value have a higher frequency of occurrence relative to other members of the sample, or how steep the frequency graph will be for a given sample. All parameters (2-4) depend on the arithmetic mean; the part \((x_i - \bar{x})^k\) raised to different powers is especially common.

It is proposed to consider additional geometric variation characteristics, which are based on the parameters: minimum, maximum, variation interval and arithmetic mean of the sample.

3. Geometric characteristics of the variation

The sample we are studying is denoted by \(Y = \{y_1, y_2, \ldots, y_N\}\). When analyzing coordinates, the members of such a sample usually depend on the measurement time, but for simplicity, we will not take this feature into account. The proposed methodology requires ranking the members of the sample in ascending order of values \((y_1 \leq y_2 \leq \ldots \leq y_N)\). As a result, a variation series will be formed, now the indices of the sample members will reflect the order of increasing sample values. Artificial datasets will help us in understanding the essence of geometric parameters, samples \(Y_1\) and \(Y_2\) show changes in some coordinate in millimeters (figure 1).

Figure 1. Synthetic samples \(Y_1\) (a) and \(Y_2\) (b) have a symmetrical arrangement of values relative to the center of the variation rectangle ABCD. The mean values of the samples \(\bar{Y}_1\) and \(\bar{Y}_2\) are shifted by the same values with different signs relative to the middle of the variation interval (F). \(S_1\) and \(S_2\) are the areas between the plots of the samples and the nearest legs of the variation triangles.

Any series of data ranked in ascending order will have common geometrical features if we arrange the ordinal numbers of the sample members from 1 to \(N\) along the horizontal axis, and the values of the feature along the vertical axis. Variational series plots (figure 1, blue broken lines) will always start at the lower left corner A of rectangle ABCD and end at its upper right corner C. All values of the sample members will always be located within the variation rectangle ABCD, but depending on the sample size \(N\) and the variation interval \(I\), the size of such a rectangle will change. Samples \(Y_1\) and \(Y_2\) (figure 1) are symmetrical about the center of ABCD, which shows possible extremely opposite variants of the distribution of values. These samples have different means, but they have common features. Their means are located symmetrically about the middle of the CD segment (variation interval), and their standard deviations are equal. In these cases, the areas of the figures are equal \((S_1=S_2)\), enclosed between the graphs of values (blue broken lines) and the legs of the upper or lower variation triangles. The concept of "area" here is somewhat arbitrary, since the vertical axis can be represented by any units of measurement (not only in units of length), and dimensionless ordinal numbers are located along the horizontal axis. Note that the lengths of horizontal segments from ABCD are equal to the variation interval for all indices of the sample.
members \((\text{AD} = \text{BC} = N - 1 = I_N)\). If the area of variation is normalized to \((N - 1 = I_N)\), then we get a new geometric characteristic for the studied sample "variation density":

\[
D = \frac{S}{N-1} = S/I_N
\]  

(5)

The \(D\) value in figure 1 will represent the average area per distance (=1) between the closest indices of the 2 members of the sample. The normalization operation in (5) eliminates the uncertainty in the concept of "area", and \(D\) will be measured in units of the studied sample parameter \(Y\).

Let's consider the extreme possible variants of the distribution of sample values (figure 2).

![Figure 2](image)

**Figure 2.** Synthetic samples Y3 (a) and Y4 (b), showing a static and uniform distribution of the values of the studied parameter. For sample Y3, the variation area is zero. The variation area \(S_4\) coincides with the maximum possible area of the variation triangle \(S_4T\).

Sample Y3 has all 12 dimensions with the same value =7, so there is no parameter variation and the distribution of sample values will be static. In this case, the variational rectangle and triangles will degenerate into the segment AD=BC at the level of the mean value of the sample coinciding with F. In this case, the sought-for area of variation will be equal to zero \((S_3=0)\), hence the density of variation will be the lowest possible \((D_3=0)\).

Sample Y4 demonstrates the maximum possible value of the density of variation for a given interval of variation and a constant step of increasing each subsequent member of the sample. In this case, under the graph will be the maximum possible area, which is equal to the area of the variation triangle \((S_4)\). The slightest deviation of such a graph in any direction from the straight-line trend will decrease the density value, provided that the variation interval is preserved.

Note that for any uniform increment in the values of the members of the ranked sample (like Y4), the density of variation will always be equal to half of the variation interval \(D=CD/2=F\). A change in the number of members of a uniformly distributed sample with a constant variation interval will not affect the \(D\) value, which will change only when the variation interval changes.

Y3 and Y4 can be viewed as samples with a uniform increment of the values of each subsequent member of the sample. But for Y3 this increment is zero, and for Y4 it is greater than zero. The steeper the uniformly incremented sample plot, the larger the density of variation will be.

The complexity of calculating the area of variation \((S)\) can arise when creating software algorithms, which, according to formal criteria, will have to determine the required area of variation when calculating the density \((D)\). For example, for sample Y1 (figure 1), the area is taken within the lower triangle, and for sample Y2, the area is taken within the upper triangle. But theoretically, there are also more complex variants of the distributions of the investigated parameter with the intersection of the graph of the main diagonal AC (figure 3).
Figure 3. Synthetic samples Y5 (a) and Y6 (b), the graphs of which have one or more intersections with the main diagonal AC. S5Δ and S6Δ are the total areas between the plots of the samples and the diagonal AC.

Theoretically, the area of variation S for any graph of a ranged sample cannot be larger than the area of the variation triangle ST (figure 2). Therefore, the total area between the sample plot and the main diagonal SΔ (figure 3) will be included in the following expression:

\[ S = S_T - S_\Delta = \frac{IY_1N}{2} - S_\Delta \]  

(6)

Calculation of SΔ is a complex procedure, but in its process one more new variation characteristic of the sample is determined. It is convenient to represent the sequence of calculating the statistical characteristics and SΔ using the sample Y1 (figure 1) in tabular form (table 1).

Table 1. The results of calculating the statistical characteristics and variation areas using the sample Y1 as an example.

| Statistical characteristics | No. | Y1   | Yh1  | ΔY1   | ΔS1   |
|-----------------------------|-----|------|------|-------|-------|
|                             | 1   | 3.00 | 3.00 | 0.00  | 0.00  |
|                             | 2   | 3.00 | 3.73 | -0.73 | 0.36  |
|                             | 3   | 4.00 | 4.45 | -0.45 | 0.59  |
|                             | ... | ...  | ...  | ...   | ...   |
|                             | 10  | 6.00 | 9.55 | -3.55 | 3.18  |
|                             | 11  | 6.00 | 10.27| -4.27 | 3.91  |
|                             | 12  | 11.00| 11.00| 0.00  | 2.14  |

Minimum 1.00  3.00  3.00  -4.27
Average (Asymmetry) 6.50  5.08  7.00  -1.92
Maximum 12.00  11.00  11.00  0.00
Variation interval 11.00  8.00  8.00  4.27
Variation area 21.00  44.00  23.00

Columns No. and Y1, the ordinal and non-decreasing values of the observed variable are entered. Column Yh1 contains the values of the graph of the maximum density of variation, which lie on the diagonal AC, and based on the parameters for Y1, the following will be calculated:

\[ y_{h1} = \frac{(n-1)Y_1}{I_N} + MIN_{Y_1} \]  

(7)

Column ΔY1 contains the differences between the values of the rows of columns Y1 and Yh1. The average for this column ΔY1 = -1.92 shows how much the arithmetic mean for the sample Ȳ1 =
5.08 is displaced relative to the middle of the variation interval $F=7$ in units of the observed parameter (figure 1). A negative value $\Delta Y1$ indicates the magnitude of the displacement of the sample mean downward from the middle of the variation interval, and a positive value $\Delta Y2$ indicates the opposite situation. The parameter $\Delta Y$ is called "variational asymmetry"; it has the same units as the studied sample $Y$.

The maximum and minimum values of the column $\Delta Y1$ can be used to separate heterogeneous groups of sample members. Next, column $\Delta Y1$ is used to calculate the sum of the area segments between the sample plot and the main diagonal AC (figure 3). For this, calculations and rules apply:

$$\Delta S_i = \begin{cases} \frac{|\Delta y_i| + |\Delta y_{i-1}|}{2|\Delta y_{i}^2 + \Delta y_{i-1}^2|}, & \Delta y_i \Delta y_{i-1} < 0, \\
\frac{|\Delta y_i| + |\Delta y_{i-1}|}{2}, & \Delta y_i \Delta y_{i-1} \geq 0, \\
i \geq 2 
\end{cases} \quad (8)$$

Since area segments can be calculated between two adjacent lines, the numbering in (8) will start with $i=2$. The product of the differences $\Delta y_{i-1}$ and $\Delta y_i$ can serve as a criterion for applying the formula for calculating the area segment. If $\Delta y_i \Delta y_{i-1} < 0$, then these differences have opposite signs and nonzero values, which is a sign for calculating the area of symmetrically opposite similar triangles. In other cases, the formula for calculating the area of a trapezoid with a height equal to one is applicable.

The sum of all the $\Delta S_i$ in the $\Delta S1$ column will give the total area between the sample plot and the main diagonal of the variation rectangle $S_3=23$. The bottom line of table 1 in column $Y_{h1}$ also gives the area of the variational triangle according to (6) $S_{1T}=44$. In column $Y1$, according to (6), the sought-for area of variation $S1=21$ for the graph of the sample $Y1$ is calculated. According to (5), the density of variation for this sample will be $D1=S1/I_{S1}=1.91$ in units of the measured parameter.

If the members of the sample $Y$ do not have a uniform distribution, then there is such a sample $Y'$ with a uniform distribution of its members and the equalities $D=D'$ and $I'=2D$ will be satisfied (figure 4).

$$P = \frac{S}{S_t} \quad (9)$$

Figure 4. Samples Y1 (a) and Y7 (b) have equal variational areas and densities, but their proportions of maximum density differ significantly.

Samples Y1 and Y7 have different graph structures; they differ insignificantly in arithmetic mean and standard deviations, and have equal variation densities. In such cases, the differentiation of the samples can be provided by the "proportion of the maximum density".

This indicator, even with the same values of the density of variation of the sample, will differ in the proximity of the magnitude of the area of variation to the area of the variation triangle.

Geometric parameters have an advantage over traditional statistical characteristics at some positions when they are compared on the basis of a variety of synthetic samples. So the variational
asymmetry most accurately reflects the position of the mean relative to the middle of the variation interval in units of the observed parameter, as opposed to cubic asymmetry. When discriminating samples, the density of variation is several times more sensitive than the linear and standard deviations. Thus, samples with single outliers more than 10 times have a lower variation density than samples with a uniform distribution. The fraction of the maximum density approximately also qualitatively reacts to the variational differences in the samples and complements other geometric characteristics.

When studying the structure of the sample, an algorithm is proposed for removing outliers or noise, separating heterogeneous groups of members of this sample. In table 1, we find the maximum modulus of the value $|\Delta Y|_{\text{max}}$ at the stage of calculating the difference between the observed values of the sample and the values of the maximum density of variation. Then the minimum part of the sample is cut off, leaving the term with the value $|\Delta Y|_{\text{max}}$ in the main part of the sample (figure 5).

![Figure 5](image)

**Figure 5.** The principle of sequential cutting off of heterogeneous groups from members of the sample Y5 based on the criterion $|\Delta Y|_{\text{max}}$.

As a result of the first separation of the sample Y5 at the value $|\Delta Y|_{\text{max}} \approx -1.82$, a new sample Y5a with a truncated volume is formed. Further, the cutoff iterations are repeated until the result becomes satisfactory in terms of the level of geometric characteristics, or the sample becomes uniform or stationary, as in Y5b. The application of such an algorithm to the data of real intraday coordinates with 2 levels of known positions (N = 2880), the qualitative separation of positions and end emissions occurred at the level of the proportion of maximum density $P \geq 0.9$.

4. Conclusion

To assess the properties of variation of samples of geodetic and other discrete data, geometrical characteristics that are logically connected with each other and with the simplest traditional statistical parameters are proposed: variation asymmetry, variation density, and the proportion of maximum density. These characteristics do not depend on the distribution parameters and are tied to the center of the variation interval, and not to the sample mean as linear and standard deviations, and others.

Variational asymmetry reflects the position of the mean relative to the center of the variation interval in units of the observed parameter. Density of variation characterizes the level of average variability in sample units. This parameter has 10 times greater discriminatory sensitivity to extremely different types of variations than linear and standard deviations. Proportion of maximum density is relative parameter shows the closeness of variation to a uniform distribution in the ranked sample and complements the indicator of variation density.

On the basis of calculating the geometric characteristics, an algorithm has been developed for cutting off emissions (noise) from the useful signal when studying the structural levels of the sample. For the observed samples of GPS coordinates, the level of the proportion of maximum density $P \geq 0.9$ is sufficient to stop iterations to cut off the sample values that are identified as noise.
Acknowledgments
The research was carried out within the framework of the state task of the Russian Federation for Research station of the Russian Academy of Sciences in Bishkek city, under the program of basic research IX.128, topic: "Study of modern crustal movements in Central Asia using space geodesy" (AAAA-A19-11902019066-3).

References
[1] 1988 Mathematical Encyclopedic Dictionary ed Yu V Prokhorov et al (Moscow: Soviet encyclopedia) p 847
[2] Herring T A, King R W, Floyd M A and McClusky S C 2018 Introduction to GAMIT/GLOBK: Release 10.7 (Cambridge: MIT) p 54
[3] Dach R, Lutz S, Walser P and Frídez P 2018 Bernese GNSS Software, Version 5.2 (Bern: AIUB) p 826
[4] Kuzikov S I 2014 Methodical Questions and Accuracy Problems of GPS Observations by the Example of the Geodynamic Proving Ground in Bishkek Izvestiya, Physics of the Solid Earth 50 6 pp 770–84
[5] Shmoylova R A, Minashkin V G, Sadovnikova N A and Shuvalova E B 2014 Statistical Theory: Textbook (Moscow: Finance and statistics) p 656