Earth Sciences
2022; 11(2): 29-34
http://www.sciencepublishinggroup.com/j/earth
doi: 10.11648/j.earth.20221102.11
ISSN: 2328-5974 (Print); ISSN: 2328-5982 (Online)

Unequally Interval Data Processing Across Fault Deformation Measurement

Peizhi Wu[1], Tianhai Liu[2], Mingyong Lu[2, *], Yan Xiong[1], Leyin Hu[1], Pingfa Zhang[2], Jiannong Wen[1], Hong Ji[1], Gang Feng[1]

[1]Beijing Earthquake Agency, Beijing, China

[2]National Earthquake Response Support Service, Beijing, China

Email address:

quakewu@163.com (Peizhi Wu), lmy9988@163.com (Mingyong Lu)

[*]Corresponding author

To cite this article:

Peizhi Wu, Tianhai Liu, Mingyong Lu, Yan Xiong, Leyin Hu, Pingfa Zhang, Jiannong Wen, Hong Ji, Gang Feng. Unequally Interval Data Processing Across Fault Deformation Measurement. *Earth Sciences*. Vol. 11, No. 2, 2022, pp. 29-34. doi: 10.11648/j.earth.20221102.11

Received: March 10, 2022; **Accepted**: April 1, 2022; **Published**: April 14, 2022

Abstract: According to different regions, conditions and requirements, the cross-fault measurement specifications is allowed to measure at different resurvey periods, and resulted in unequal interval observation data. The unequal interval observation data is a common phenomenon data, the difference on both sides of the fault is observed by geological investigation, historical record, artificial observation, simulated record, digital sampling, encrypted observation before and after the event, change of observation equipment, change of observation environment, human factors, etc, and the unequal interval observation data is obtained. The characteristics of the unequal interval observation data is not only shown in time, but also in space. The unequal interval observation data is usually preprocessed into equal interval data by some kind of algorithm chosen before the subsequent complex calculation. In the data processing of cross-fault measurement, the unequal interval observation data is usually preprocessed into equal interval data, and then calculated, which leads to a series of new problems, such as time calculation, synchronization, master-slave relationship, comparability and so on. In view of unequal interval observation data in cross-fault measurement, some new problems are tried to solve in unequal interval data matching calculation by using conventional methods combined with some algorithm requirements, data characteristics and practical experience, and their adaptability in various algorithms is investigated in this paper. These works contribute to the improvement and development of cross-fault survey data processing methods, and enhance the role of cross-fault survey data in earthquake protection and disaster reduction.

Keywords: Cross-Fault Deformation Measurement, Retest Period, Unequal Interval Data, Synchronization Domain, Comparability

1. Introduction

According to the cross-fault deformation measurement standard in China [1-3], the retest period can be 1 month, 2 months, 3 months, 4 months, 6 months, 12 months and so on. The retest period should also be kept at equal intervals and be processed in the same month. The retest date can fluctuate within 1/6 of the retest period. Moreover, it also can be adjusted according to the earthquake situation. For the metropolitan cross-fault deformation measurement, the probability that the re-measurement period is 1 month is 7/8 and the probability that the re-measurement period is 2 months is 1/8 [4]. Consider a case that the retest period of item A is 1 month and that of item B is 2 months. When comparing the relative changes of the two items, the measurement data does not match due to the different retest period. This difference stems from spatial factors. If the retest period of the two items is the same, but they are tested in different months, the data is asynchronous if it is calculated in months.

Considering another cases that several test values of an item has been observed and the test intervals of these value are 10a, but intervals of 10a for an item has been observed, many events may cause that the data is still at unequally interval in time. These events can be the change of retest period, supplementary test, additional test before or after

earthquakes, data impairment of outdoor test stations, etc. It is worth noticing that most of the test data is unequally interval according to the following requirements in the standard: 1) The retest period multiplied by (1 ± 0.17) is the detection limit. 2) There must be test baselines and checking levels more than 30a. 3) There must be less than 20% data within the detection limit. As a result, whether in the space or time domain, it is more reasonable to regard the data of cross-fault deformation measurement as unequally interval data from a holistic perspective. Equally interval data can be processed directly using existing software [5-11], while for the unequally interval data, it first needs to be converted to equally interval data using specific preprocessing algorithms, then conducts complex subsequent processing [12].

In the data processing of cross-fault deformation measurements, except converting the original into equally interval data for processing, the more general way is using unequally interval data directly for calculations as much as possible. However, the calculation results are also unequally interval, causing a series of problems.

For example, before calculating the correlation coefficient of two time series, data matching must be performed first. For two time series with the same start and end times but different time interval, we need to check whether the data is time synchronized, and then determine the check standard in advance. But for two groups of data with the same time interval, the results can be calculated as long as they are paired in order, without checking whether each pair of data is synchronized in advance. Assuming the observation time is represented as year-month-day, and the retest period of item A and B is 1 month and 4 months, respectively. If we set the year-month-day to be the same as the synchronization condition, a lot of data may be discarded since the standard allows the retest date can be changed within 1/6 of the retest period, which is apparently unreasonable. Therefore, it is necessary to set an appropriate time synchronization domain for data matching. Therefore, the software for data processing of cross-fault deformation was developed [13]. In this software, the time synchronization domain can be 1 month, 2 months, 3 months, and 4 months. For selecting the most suitable in a data matching task, we need to analyze the characters of the data, and users' confidence in the credibility of data to ensure that the distribution of data pairs in time is relatively uniform, the data discarded as little as possible, and the calculated correlation coefficient is the closest to the reality. The unequal interval cross fault data processing method of this software is introduced in this paper. In the data processing method, we make necessary matching processing for unequal interval data through reasonable assumptions to meet the requirements of various algorithms.

2. Several Necessary Conventions

In the data processing of cross-fault deformation measurement, different algorithms have different requirements for data. Some commonly used conventions are described below.

2.1. Time Algorithm Convention

In the earthquake deformation monitoring, the sliding window algorithm is a common algorithm. The observation date of cross-fault deformation measurement is represented by year-month-day, and the window length and step length are represented by month. Since the number of days in a year or a month is not fixed, there are various algorithms for time calculation, such as month as unit, day as unit, retest as unit, calendar year-month-day representation, uniform, and so on. These are all reasonable, but they are applied to different scenarios. Different algorithms have differences in calculation results. For example, if we use a uniform algorithm that sets the number of days in a year to 365.24d, the number of days in a month, window length, step length, etc. are all decimals. Thus, there is a rounding error when rounding the date, making the calculated period lengths not exactly the same. This phenomenon also exists in various algorithms using months as the unit, i.e., the actual calculated years, months, window length, and step length are not strictly equal. The uniform algorithm has the advantages of simplicity, a small error, and being scientific. Although it is contrary to the general definition, the performance is better. The time algorithm is a convention. Without a special statement, one software system should only use one time algorithm. Mixed-use will reduce the comparability of results.

2.2. Synchronization Domain Convention

The synchronization domain is a time domain, which is artificially set based on data characteristics, algorithm requirements, user experience, etc. We define that all measurements within the synchronous domain are considered synchronous. For example, if item A is measured twice in a month and the synchronization domain is one month, the two measurements are considered to be synchronous. Data matching is required when removing data annual variation and calculating correlation coefficient, difference, rate, strain, fault activity, etc. Time synchronization is often a basic requirement for data matching. For example, let t_{ai} and a_i represent the observation date and the measurement value of item A, respectively, where i is the data index. Similarly, the ith data of item B is (t_{bi},b_i). After traversing all the data of item B to get $|t_{a1} - t_{bi}|$, if $|t_{a1} - t_{bi}|$ is less than the synchronization domain, a_1 and b_i are synchronized. Otherwise, data synchronization failed. Obviously, the choice of the synchronization domain is related to the value of the retest period. For item A, if the proportion of the observation interval is 1 month is 17%, and the proportion of the observation interval is 2 months is 80%, the retest period of item A is likely to be 2 months. Similarly, the retest period of item B is likely to be 3 months. Under this assumption, the synchronization domain can be 2 months, 3 months, or other values. After selecting a synchronization domain, there are still one-to-one matching, one-to-many matching, and matching failures in the data matching process. As a result, we need to derive an optimal synchronization domain selection

method based on the chosen algorithm and confidence for realizing the best match.

2.3. Master-Slave Convention

Subtraction is necessary for the calculation of rate, strain, fault activity, and removing data annual variation. We define the minuend as the master and the subtrahend as the slave. The synchronization domain where the minuend is located is called the master synchronization domain, and the synchronization domain where the subtrahend is located is called the slave synchronization domain. Theoretically, when calculating the correlation coefficient, there is no master-slave relationship between item A and item B, in other words, item A and item B have the same status. We assume that two pieces of data of item A are located in the synchronization domain, and six pieces of data of item B are located in the synchronization domain. If the data matching is based on the measured time, only one pair of data can be matched and six pieces of data are discarded. If the data matching is based on the data serial number, two pairs of data can be matched and four pieces of data are discarded. If we consider item A as the master and item B as the slave, two pairs of data can be matched and four pieces of data are discarded. If the data matching is based on the mean of the data in the synchronization domain, only one pair of data can be matched. If we consider item B as the master and item A as the slave, all data can be matched. The last data matching method works best if the premise is to use as many observations as possible. As a result, even though using an algorithm without the master-slave relationship, we can realize the best data matching in the unequally interval data processing by artificially setting a master-slave relationship.

2.4. Comparability Convention

The calculation results are meaningless if there is no comparability between the data. The data from across-fault deformation measurement can be simplified into two columns for date and measurement. We can get the time series by performing the difference on the measured value. Imagine an extreme case, the difference between two measurement values of the same item is 10mm. If the difference between the date of the two measurement values is 1d, the result mentioned above means an earthquake may have occurred. If the difference between the date corresponding to the two measurement values is 10a, the result mentioned above is just a normal value. Therefore, if the difference between the measured values is considered and the time difference between the measured values is ignored, the comparison is theoretically meaningless, since it does not meet the comparability requirement. The difference of the measured value is often used in cross-fault deformation measurement calculation. Some calculation formulas also contain difference factors. When solving the time series about difference value, it is necessary to consider whether their time difference is the same, that is, whether the comparability condition is satisfied. The denominator of the rate calculation is the difference of the time and the numerator is the difference of the measured value, so the time series of rates are comparable. Therefore, it is better to use rate instead of difference. If we cannot replace difference with rate, we need to assess the impact caused by unequally interval data. When the time dispersion of unequally interval data is small, difference calculation can be performed, otherwise, we need to conduct the necessary preprocessing. Obviously, we can reduce the time dispersion by appropriately reducing the synchronization domain, but this can also result in more discarded data.

3. Data Matching for Unequally Interval Data

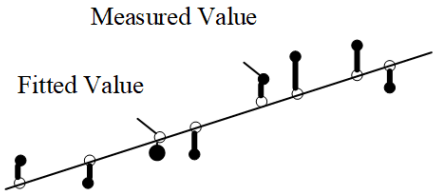
The data matching in this example targets the most reasonable matching or best matching, and it also aims at directly using unequally interval data as much as possible. Moreover, additional data processing procedures is described in the paper.

3.1. Trend Removal

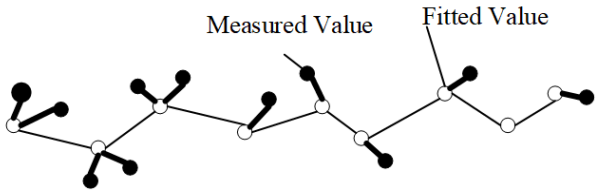
When the changing trend of the time series is linear, we can use the fitted value and residual value of the time series obtained by linear regression to finish the trend removal task. When the trend can be fitted by other functions, the calculation process is similar to the process that uses a liner function. In the algorithms based on these unequally interval data, the data is matched using the measured value and the time corresponding to the values. All the input data can be applied, and the output is also unequally interval. Figure 1(a) shows the trend removal under a liner function. In this figure, the solid circles represent the master (minuend or input), the hollow circles mean the slave (subtraction or fitted value), and the polyline passing through the hollow circles is the fitting line. In order to highlight the characteristics of data matching, the X and Y axes are not shown in the figure.

3.2. Annual Variation Removal

The process of annual variation removal is similar to that of trend removal. We first segment the year according to the synchronization domain, and then use the anomaly method to remove the annual variation. The output is also unequally interval data. In the data used in this example, the data with a retest period of 3 months accounts for 70%, so we consider the retest period is 3 months. Then, we calculated the mean of all measurements in each quarter of the time series, and use the mean of each quarter to represent the average annual variation. In the anomaly method, the measurement value is the master, and the average annual change is the slave. We subtract the slave from the master in the synchronization domain to realize annual change removal. All data is used in this algorithm. Figure 1(b) shows the annual variation removal algorithm. In Figure 1(b), the solid circles represent the measured value, the hollow circles represent the fitted value, and four consecutive fitted values are 1a. Cycle according to the annual, the broken line of the hollow circles is the average annual variation curve.



(a) Data matching based on trend removal

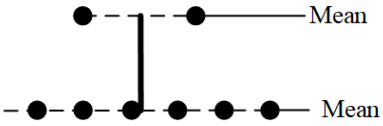


(b) Data matching based on annual varation removal (partial).

(c1) Data matching based on closest measurement time

(c2) Data matching based on data serial number

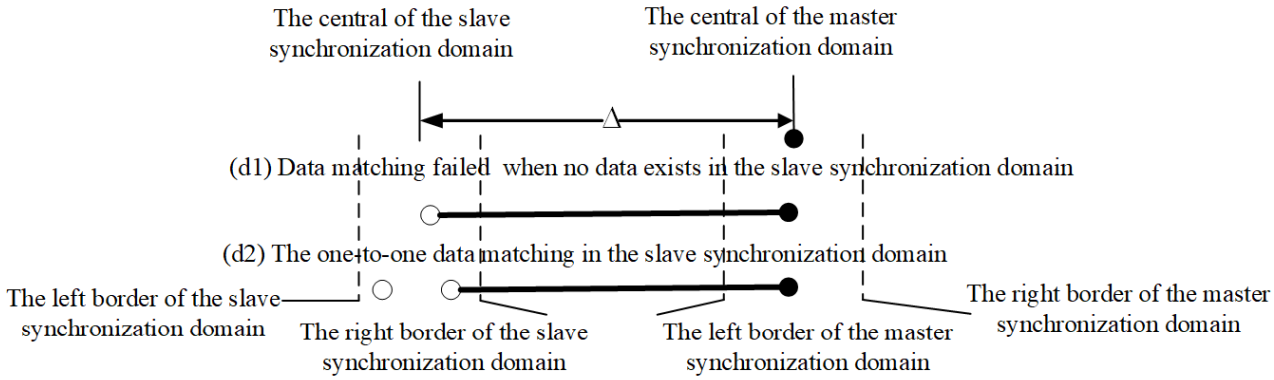
(c3) Data matching when item with less data is the master item



(c4) Data matching based on the mean value

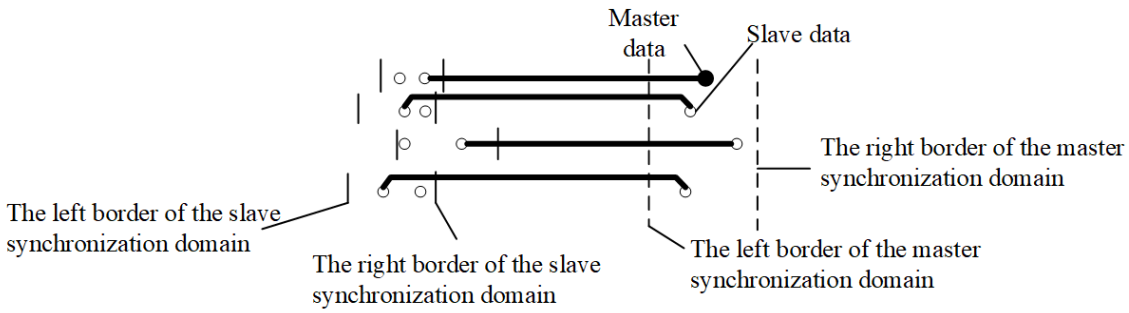
(c5) Data matching when item with more data is the master item

(c) Five example for data matching based on correlation coefficient in the synchronization domain



(d3) Data matching based on closest measurement time in the slave synchronization domain

(d) Three examples for data matching in the single item rate calculation



(e) Data matching in the multi-item rate calculation

Figure 1. *Data matching diagram.*

3.3. Correlation Coefficient

The method to calculate the correlation coefficient has been described above. Figure 1(c) shows various data matching methods for the measured data in the synchronization domain. In Figure 1(c), we consider that the data retest of item 1 and item 2 is at equally interval, whereas we retest twice for item 1, and six times for item 2 in the synchronization domain. Therefore, the data is still considered as unequally interval data when calculating the correlation coefficient. In theory, there is no master and slave when calculating the correlation coefficient between the two groups of data, but we define item 2 as the master to enable more input data to be applied directly, and Figure 1(c5) shows the optimal selection. We consider the item with a large amount of data as the master item, which can discard less measurement value and obtain more trial counts.

3.4. Single Item Rate (or Difference) Calculation

The single item calculation rate is a basic algorithm in the data processing of cross-fault deformation measurement. The calculation of rate and the cumulative intensity are all single item calculation. For the rate calculation, the synchronization domain where the minuend is located in the master synchronization domain, and the synchronization domain where the subtrahend is located is the slave synchronization domain. The widths of the master and slave synchronization domain are the same. The difference Δ between the center value of the master synchronization domain and the slave synchronization domain is set by the user. For example, when calculating the annual rate, Δ is set as 12 months. Based on the definition, we set the measurement time as the central value of the master synchronization domain when calculating single item rates. Under this condition, the data matching is failed when there is no data in the slave synchronization domain. Instead, when input data exists in the slave synchronization domain, we calculate the difference between each measurement time in the slave synchronization domain and the central value of the master synchronization domain, and match the data pairs of which the difference is the closest to Δ. Figure 1(d1) shows the case that no data exists. Figure 1(d2) shows the one-to-one data matching. Figure 1(d3) shows the one-to-many data matching. When calculating the single item rate, selecting a wider synchronization domain will have more opportunities for adaptation and fewer data to be discarded.

3.5. Multi-item Rate (or Difference) Calculation

Calculation tasks such as fault activity, strain, synthesis, etc. are multi-item calculations. When calculating the multi-item rate, the first process is to determine the master synchronization domain. The master synchronization domain should generally meet the following conditions: 1) The item with the most data can be selected as the main item, and all measurement values are traversed in order. The time of the main measurement value is the output time of the result. Other items in the master synchronization domain are the slave items. 2) The time difference between all measured values in the master synchronization domain should be less than or equal to the width of the synchronization domain. 3) When a slave item has two or more optional measurement values in the master synchronization domain, we select the measurement value of which the measurement time is closest to that of the master measurement value. 4) When a slave item has no data to be matched, we should decide to make trade-offs, continue or exit based on the needs of the algorithm. Figure 1(e) shows how to select the master synchronization domain in the multi-item rate calculation. It is calculated by each measurement value in the master synchronization domain, and the method is the same as that in the single item rate. The slave synchronization domains of each item do not overlap in time.

4. Conclusion

Generally, it is more reasonable to regard cross-fault deformation measurement data as unequally interval data, and the unequally interval data should be directly used for calculation during data processing. Considering some algorithm requirements, data characteristics, and practical experience, this paper tries to solve some problems encountered in data matching by convention. In this paper, we consider 1) the measurement time of cross-fault measurement should be represented by ‘year-month-day’, and the parameters such as window length and step length should be represented by month. The data selected according to different time calculation methods will also be different, causing different calculation results. Unless otherwise stated, the same software convention uses the same algorithm. 2) The synchronization domain is the basic restriction to achieve data matching. We find that it is not necessary to require all the time properties in the synchronization domain to be equal, and we just treat them as synchronous. The data outside the synchronization domain and the data in the synchronization domain are asynchronized. A synchronization domain is a time zone determined by algorithms, data characteristics, and practical experience. 3) For the master-slave relationship, the minuend is the master and the subtrahend is the slave for rate calculation. The master-slave can be used for subtraction and for the determination of the synchronization domain in the multi-item calculation. In order to meet the requirement of directly using unequally interval data and matching as many data pairs as possible, we can select an item with high reliability and a large amount of data as the master item in the calculation of the correlation coefficient without a master-slave relationship. 4) The difference calculation of time series is comparable during the processing of equally interval data, but not theoretically comparable for unequally interval data. We should replace the difference with the rate as much as possible. If the replacement cannot be achieved, we need to evaluate the temporal dispersion of unequal interval data to determine whether preprocessing is required.

The design of this paper has been realized by the cross-fault deformation measurement data processing software. We regard the item with a large amount of data as the master item for using unequally interval data directly. However, to study the normal variation of time series, we should take the item that is observed once in each retest period as the master item. The widening of the synchronization domain can get more matching opportunities, but it will increase the time dispersion and decrease the comparability of unequal interval data. Different time calculation algorithms are suited for different scenarios, so the data selected are also different. The uniform algorithm is more scientific but is not easier to understand. While the 'year-month-day' algorithm is easier to understand, it cannot guarantee the closest matching time. Therefore, we need to make trade off in practice.

Acknowledgements

This work is found by the Spark Program of Earthquake Technology of CEA, No. XH14057; Fund of China Earthquake Administration, No.JC1704021026; Fund of Beijing Earthquake Agency. No.BJMS-2022008; Natural Science Foundation of Beijing Municipality, No. 8041001.

References

[1] The Journal of Antibiotics. doi: 10.1038/s41429-021-00430-5.

[2] State Seismological Bureau. Cross-Fault Measurement Standard. Beijing: Seismological Press, 1991.

[3] China Earthquake Administration. Earthquake Industry Standard of the People's Republic of China: The Method of Earthquake-Related Crust Monitoring-Fault-Crossing Displacement Measurement. Beijing: Seismological Press, 2012.

[4] Lu M., Liu T., & Huang B., et al (2011). Discussion of Environment and Monitoring Technology for Cross-Fault Mobile Deformation Monitoring. Journal of Geodesy and Geodynamics, 31 (5), 141-145. doi: 10.14075/j.jgg.2011.05.031.

[5] Software Technology Group, State Seismological Bureau. The Software System for Earthquake Prediction in China. Beijing: Seismological Press, 1994.

[6] Jiang J., Li S., & Zhang Y., et al (2000). Earthquake Precursor Information Processing and Software System. Seismological Press.

[7] Lu Y., Li S., & Deng Z., et al (2002). GIS-Based Seismic Analysis and Forecasting System. Chengdu Cartographic Publishing House.

[8] Peng Y., Wu A., & Li S., et al (2012). Dynamic Display of Observation Curve on China Earthquake Precursor Network. Journal of Geodesy and Geodynamics, 32 Supp., 49-52. doi: 10.14075/j.jgg.2012.S1.015.

[9] Qu J., Zhang S., (2014). Study on Cross-Fault Site Information System Based on GIS. Recent Developments in World Seismology, 424 (4), 27-34.

[10] Lu M., Xiong D., & Yu H., et al (2015). The Building of Temporary Cross-Fault Deformation Basis and Monitoring Data and It's Environmental Information Database in The Capital Region. Recent Developments in World Seismology, 442 (10), 32-38.

[11] Liu W., Lu M., & Luo S., et al (2020). Design and Implementation of Data Management and Pre-processing System for Cross-Fault Flow Deformation. Technology for Earthquake Disaster Prevention, 15 (3), 635-642. Doi: 10.11899/zzfy20200318.

[12] Xu S. (1994). C Common Algorithms. Tsinghua University Press.

[13] Lu M., Wu P., & Li Q., et al (2019). Software Construction of Data Processing for Cross Fault Flow Deformation. Recent Developments in World Seismology, 486 (6), 14-19.