УДК 550.83:624.13 DOI: 10.33744/0365-8171-2021-109-157-165

CONSTRAINTS OF OPTIMIZATION OF STATISTICAL ANALYSIS OF DATA OF ENGINEERIG MONITORING OF TRANSPORT NETWORKS

ОБМЕЖЕННЯ ОПТИМІЗАЦІЇ СТАТИСТИЧНОГО АНАЛІЗУ ДАНИХ ТЕХНІЧНОГО МОНІТОРИНГУ ТРАНСПОРТНИХ МЕРЕЖ

***Kuzminets Nikolai P.**, Doctor of Engineering Sciences, associate professor, National Transport University, chairman at Dept. of computer, engineering drawing and design, e-mail: kuzminecmp@ukr.net, phone +380442849713, Ukraine, 01103, Kyiv, M. Boychuk street 42, office 705*

http://orcid.org/0000-0002-9636-919X

***Dubovenko Yurii I.**, Doctor of Phylosophy in Physics and Maths, associate professor, National Transport University, senior teacher at Dept. of computer, engineering drawing and design, e-mail: nemishayeve@ukr.net, phone +380665979384, Ukraine, 01103, Kyiv, M. Boychuk street 42, office 709*

http://orcid.org/0000-0002-8128-5989

Summary. Time series for the technical monitoring of the transportation networks include interference and omissions. Their analysis requires the special statistical analysis. Known statistical packages do not contain a full cycle for processing of large time series. The linear timeline for the processing of periodic data is not available in digital statistics.

The sliding window approach is suitable for processing of interrupted time series. Its disadvantage is the restriction on the length of the row and the sensitivity to data gaps. The graph of the time series processing needs the internal optimization. The necessary steps for optimizing of the time series processing graph are determined. They are as follows: store data in an internal database, build the data samples on a single time scale, sampling based on the meta-description of the series, averaging in a sliding window, calendar bindings and omission masks, generalization of graphs, storage of graphics in vector format and so on.

The conditions for the study of series are revealed such as the database, calendar structure of data, processing of the gaps, a package of numerical methods of analysis, processing in a sliding window.

Keywords: monitoring, time series, optimization, averaging, sliding window.

Introduction

The data time series obtained during the continuous technical monitoring of engineering and transport networks, along with the useful signal, also contain the highly variable interference, data gaps, and non-stationary effects. These properties of the series require the use of the special methods of the statistical analysis for the initial data as well as their transformations. To do this, one can use the statistical analysis packages (SPSS, WinStat, etc.), but they do not contain a complete cycle of operations for the processing of time series. In particular, for processing of the large-dimension data, a database management system (DBMS) for time series, a computing unit, and a data visualization unit are required. An example of the implementation of such a block is presented in the Simulink package for Matlab [1]. It is necessary to provide a linear timeline in the data processing [2]. It is most accessible implemented in the modern visual editors (such as MS Visio, 3D Studio Max, etc.) [3]. This greatly increases the ergonomics of data analysis and the joint processing of the series with unequal start dates or inconsistent frequency of observations. Such series often accumulate when monitoring the acoustic, seismic, electromagnetic, mechanical and other types of vibrations.

To analyze the structure of the series, it is necessary to identify the dependencies and relationships between the individual signals, which are burdened by noise. Many methods of digital processing of imperfect data are known [4]. The moving time window technique is often used [5] to study the development of the process in time and to identify the variation of useful signal anomalies as a reaction to external influences. The drawback of the moving window technique is the limitation on the row length and high sensitivity to data gaps and omissions.

Materials and Methods

The graph for processing of the series of data monitoring in the general case should contain the following blocks: the data collection and storage; preliminary data processing; processing of the data gaps; the data synchronization; the averaging in a moving window; visualization; report creation. In addition, each of these steps requires the internal process optimization. This article will be devoted to the analysis and generalization of the main approaches in this direction.

Optimization of storage of source data. The data must be stored in an internal database. Each row is an infinite number of cells with a clear time reference and a fixed interval between the observations (from 10^{-6} sec to 10^{+6} sec). When storing the data in a compact format (32 bits per value), it is possible to process rows containing ~ 10^9 points. When importing, the data must be added to the existing series (while monitoring the compliance with the chronology) in order to process the ongoing series of monitoring technical data. The advantage will be the ability to read data from digital catalogs of earthquakes [6].

The formation of the primary database of monitoring allows us to work not with the file names, but with the data samples. When setting up the data processing parameters and visualizing their analysis results, a clear time scale is required, rather than the point numbers, which greatly increases the quality and the speed of the processing. For the joint processing of the series with mismatching intervals and / or the periodicity of data, it is necessary to recalculate the data in a single time scale, leading them to the same calculation level.

In the general case, the primary flowchart of the processing of time series of the technical observations can have the following form: creation (import) of a data sample – preliminary filtering (data editing) – import into a calculation unit – mathematical transformations – analysis of new data – the data visualization – the report generation.

Optimization of source data preparation. While performing the various operations on the source data, we change them, creating the new series. To protect the primary data from accidental change, we need to distinguish between the accesses levels to them. Primary digital information is a series of instrumental observations imported to database. The *service utilities* can have access to it (import, export, sorting, editing, etc.). In addition, the numerical data analysis algorithms [7] should work not with primary data, but with series copies such as the data sampling.

Not all information from the database is recorded into the sample, but the data needed for a specific analysis (in the set of series and in the sampling time interval). Such a selection is convenient for the calculations: it is easier to specify the data that should be used by the necessary procedure and to set the real processing parameters. When creating a sample, data series must be recalculated at a given polling frequency. Thus, it is possible to build a joint analysis of data with the different frequency of observations [8].

The data series must have a *meta description* (a special attribute like exif / iptc in image files). New series created during the calculations should retain a meta-description of the original series, supplementing it with the new information about the processing performed. It is important to record the user's operations within the data processing file (such as the calculation method, the model parameters, the calculation results, the statistical characteristics of the original series and those created during the calculation). This saves the processing information in a convenient form for subsequent analysis in other programs for engineering and / or mathematical modeling.

It is necessary to enter the processed (changed) rows from the workspace into the database separately, adding a new series to the main database, and not changing the primary data.

The well-known statistical analysis packages (SPSS, Statistica, Stadia, Statgraphics, Matlab) store the workspace in RAM, which imposes the appropriate restrictions on the size of the data sample [9]. Moreover, if one save the operational information from the workspace onto a separate (external) hard disk of a computer, one can load any amount of the data into the sample (up to filling the entire disk). The created sampling is not lost during the abnormal end of session, but automatically opens the next time the program starts.

Results and Discussion

Optimization of the data loss processing. The real series of the technical measurements in the monitoring of engineering and transport networks without stopping their current operation contain the data gaps. The gaps are artificially filled during the preliminary data preparation. Data gaps are filled in according to an *a priori model* of the measured value prediction. There are a number of valid models for interpolating the data gaps. The choice of one of the models at the stage of data preprocessing is not always optimal

A different approach to the processing of the observation gaps allows for their presence at any stage of the analysis. The average value can be estimated from the measured data, when alternating them with gaps or not (Fig. 1). One can not fill in the data gaps, but one need to calculate the data averaging in a moving window, while entering the necessary corrections (such as the recalculation of the weights and norms, taking into account other effects, the correction for obtaining of the unbiased estimates).

The FFT method [10] does not allow the gaps; it is necessary to fill in the gaps artificially before calling the method. Alternatively, one can to provide that the method fills in data gaps by default before calculating it. This requires more sophisticated time series calculation algorithms. However, you can use the data analysis methods that are not available in packages of the statistical processing of observations. Moreover, the analysis results are more statistically substantiated than when they are filling in the blanks manually when preparing the data.

The rejection of the incorrect measurements can be performed during the preliminary preparation of the data series (as far as the misfit looks like a sharp outlier) and at any stage of the analysis. Adaptive procedures are required for rejecting the defective observations at all the stages of the data processing.

Data sync optimization. While working with the monitoring data series, one need to calculate the arithmetic expressions (to subtract one signal from another, to determine the average value of the variables) element by element, taking into account the synchronization of measurements. It is necessary to add with each other the data that are measured at one moment in time. If the first parameter was measured daily from 1995 until 2015, while the second is obtained from 2003 until 2019, then the sum of the signals will be calculated for the time interval varying from 2003 until 2015. We take the information about the start and end dates of each row involved in the calculations from the meta description of the series. It is important for the user to type in the formula window the desired expression with the names of the series and the rest should be done automatically.

The variables in the formula are the identifiers of the data series, and the result of the calculations is a new time series, each value of which is calculated from the values of the original series by the formula given. In the computations, a “reasonable” processing of data gaps and special situations (dividing by 0, etc.) is needed. To correctly account for the experimental and the computational errors, it is necessary to prohibit the dividing by zero and also by any small numbers less than a given threshold. This threshold is also needed for comparing values, search-and-replace, etc.

The data synchronization is also needed in other calculations such as the estimating the correlation coefficient in a moving window. This allows for joint signal processing not to look for the start-end dates of the series, but to focus on the data analysis. The calculation results get the calendared binding (a method setting). In calculating of the moving average, the result can be compared with the middle of the window and its right border [11].

In addition to the arithmetic operations, the operations with calendar time series are needed: we need to "transfer" the series to another time zone (given the switch to summer / winter season) or to deploy it "backwards" or to "stretch-squeeze" the time axis, if the clock by which the observations were synchronized was behind or in a hurry.

One need to combine the rows without breaking the calendar data binding. While calculating the value of the combined series, we look in the given order for the values of the original series for the desired time instant; thus, we put in the result series the first value, which is not a gap. Therefore, you can build complex algorithms for filling gaps. It is necessary to form the several rows of aggregates, using the certain methods depending on the different conditions, and then merge all the rows into one.

The gaps mask of the certain series is used in the numerical modeling on the model of a real series survey or generated by applying to it the *spreading* or the *drying* procedures. In the first case, the mask includes all the values adjacent to the gaps in time (*n*-th order neighbors for *n*-folded spreading). In the second case, all the values that have at least one missing value among the neighbors are deleted from the gaps mask. Masks allow us to build the flexible algorithms for filling the gaps, while combining the different methods, depending on how far in time the gap is from the nearest measured value.

In the numerical modeling, a quasiperiodic and pseudorandom implementations of a series are needed, the length and moments of gaps of which correspond to a real series of measurements. They are used as a "control group" in assessing the significance of the various effects. By mixing the values of the series, it is possible to destroy all the correlation relationships without changing its distribution function. Replacing the values of a number by their ranks, one can maintain the ordering of values, transforming any distribution function into a uniform one. The numerical modeling makes it possible to evaluate the *stability* of the results, their dependence on the characteristics of the signal, and reasonably interpret the observations date series.

These data series transformations can be performed in any statistical analysis package (SPSS, Statistica, Stadia, Statgraphics, Matlab) [8]. However, you need to perform all the calculations taking into account the synchronization of observations, and the flexible processing of the gaps and the special situations.

Optimization of the processing in a moving window. One of the monitoring problems is the tracking of the changes in a controlled system. The structure of the series changes, the response to variations of the external factor (atmospheric pressure, tide, etc.) does it too another way [12]. In order to detect the changes, it is necessary to evaluate the properties of the signal not in the whole row, but within the time window. Then one can move the window to the right and repeat the calculations. Therefore, they apply the adaptive noise filtering, adjusting the filter parameters to the current properties of the series

The drawback of moving window methods is that they are difficult to combine with each other. These methods require that the entire the window is placed "inside" the row. The full "run" of the window is always less than the length of the original row. Therefore, if the row is limited in length and the window is wide, then after applying several methods, nothing will remain of the signal to process.

The moving window methods should work without reducing the length of the filtered signal. At the beginning of time, the center of the moving window is aligned with the first point of the signal; the values in the left half of the window are considered gaps. If the allowed number of passes in a moving window is more than 50%, the calculations are performed for all points in the series, from its first point. In addition, the length of the processed signal does not decrease. By varying the number of gaps, one can adjust the size of the "exit" of the window outside the row. If gaps are prohibited (0%), the calculations are performed only within a series. The gaps inside the row are similarly processed (Fig. 2).

This algorithm is better than the pre-filling the gaps procedure, or rejecting the data array in the presence of an insignificant amount of “spoiled” observations.

If the moving window is shifted by 1 point, the time step of sampling of the processed signal is equal to the step of sampling the original row (not the size of the window). This allows us to accurately track the changes in environmental monitoring parameters. By combining moving window methods, one can implement almost any data analysis algorithms.

Data visualization optimization. With the enormous power of computing and signal processing, a visual analysis of the graphs, and estimation of the signal characteristics goes by the wayside. The basis of statistical analysis packages are the data processing methods, and visualization is an addition to them. In the graphs, attention is paid to the design, not the content, which is convenient for the reports, but not in the analysis procedures.

Work with a signal should be started by studying its graph, and by choosing the optimal processing methods, as well as their parameters and the settings. The processing results must be displayed in the form of graphs in order to evaluate the effectiveness of the method and make the necessary corrections.

In the study the most interesting often are the unexpected effects of deviation of signal characteristics from the model ones. If these deviations are associated with interference, their detection is also important in order to eliminate the source of interference and to improve the quality of observations. Formal algorithms come from a predefined model of the perturbation and the signal, and they are ineffective in detecting such deviations. Using the qualified data visualization, deviations can be detected by the researcher [13].

During the visualization procedure, the data series (of a million-dot length) must be displayed quickly, on the fly, detailing any section of the series in a real calendar scale. Labels and fields must be minimized to use all the screen area to display the data.

In working with the large data sets, generalization of graphs is needed. This speeds up image retrieval, reduces the file size of the graph when stored in the vector format. Vector drawings are convenient for large rows processing and visualization, providing the image quality and ease of its design.

In addition to the time series processing units studied above, the scattering diagram (correlation field), spectra and periodograms (for the cyclic processes) are also used in the processing of technical monitoring data, as well as the fractal properties of the series are analyzed. However, these characterristics are beyond the scope of our study.

Conclusions and Recommendations

The problems of the data processing for the technical monitoring of engineering and transport networks require different processing methods. The statistical analysis packages do not have the necessary tools; it is difficult to maintain the databases, to process the gaps, and to perform the data synchronization. Matlab uses the ready-made built-in functions (such as the spectrum calculation, the plotting, etc.), or the algorithms for solving the non-standard problems. However, in Matlab there are no tools for organizing the database, most of the available data series functions do not work with the gaps, and one need to program the synchronization of data series by itself.

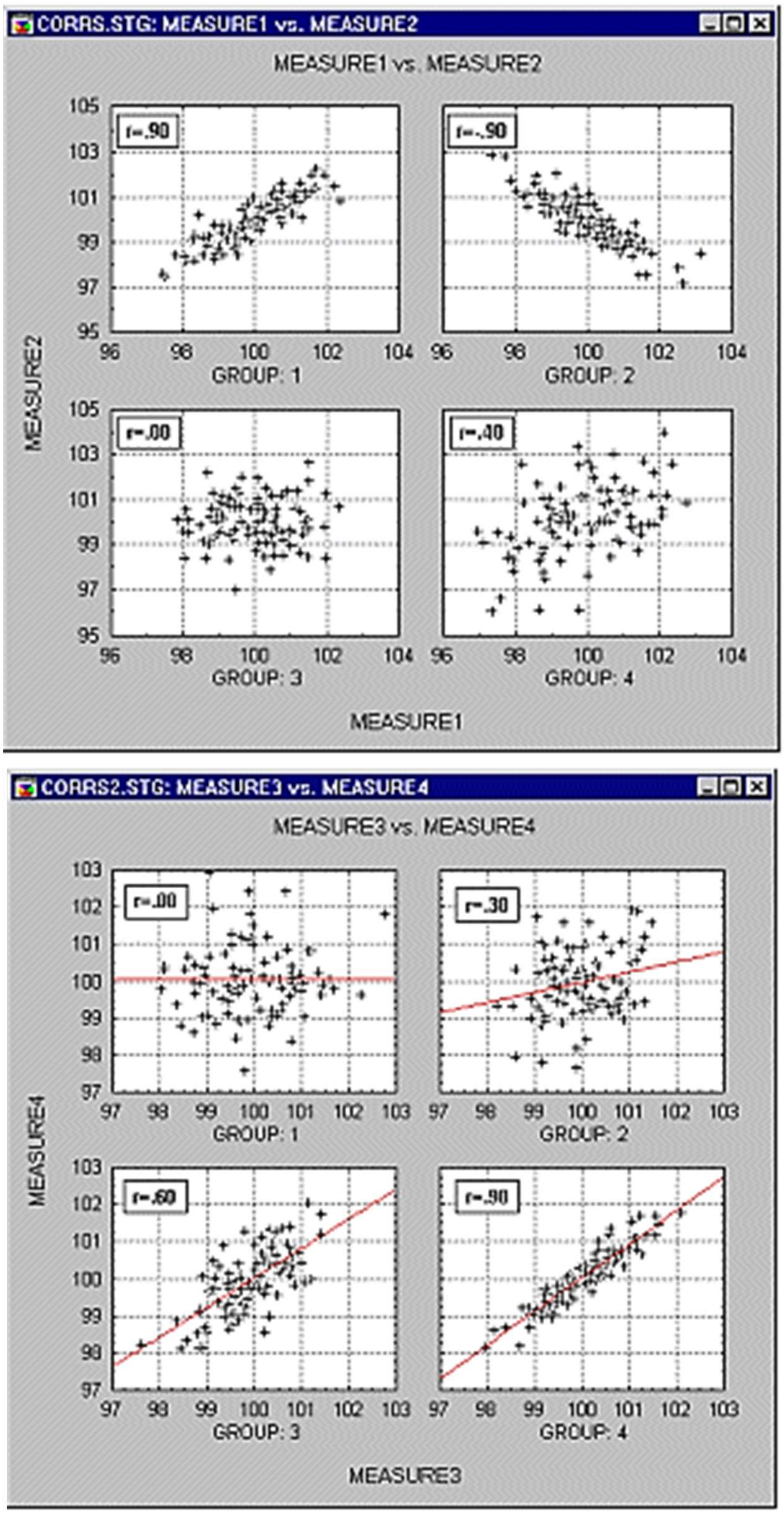

Figure 1 – Points with data gaps and moving averaging in a window with a given averaging radius
Рисунок 1 – Точки з пропусками даних та осереднення у ковзному вікні
з заданим радіусом осереднення

The alternative programs for the analyzing of the data series are poorly documented and require the constant author support. They cannot be adapted to solve the new problems while studying the data of the technical monitoring of the engineering and transport networks. None of the row tools has the necessary functionality. To study the time series, one need a database of time series, a calendar data structure, the processing of the gaps without restrictions, a powerful package of the numerical analysis methods, the processing in a moving window of time series of technical measurements.

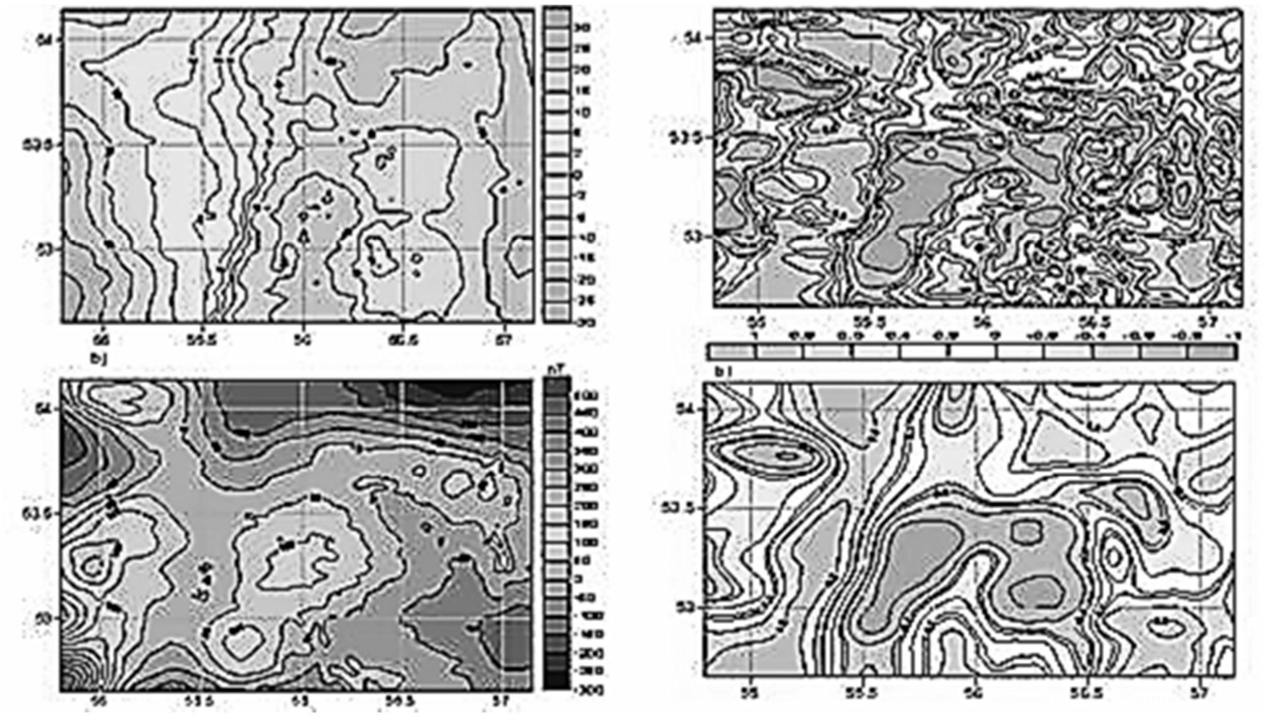

Figure 2 – Estimation of the correlation coefficient between geophysical fields in a moving windows: on the left is the gravity (top) and magnetic (bottom) fields; on the right is the correlation coefficient calculated in a window of 11x11 (top) and 33x33 (bottom)

Рисунок 2 – Оцінювання коефіцієнта кореляції між геофізичними полями у ковзному вікні: зліва – поле сили тяжіння (вгорі) і магнітне (внизу) поле; справа – коефіцієнт кореляції, обчислений у вікні 11x11 (вгорі) і 33x33 (внизу)

References

1. Yakimov I.M., Kirpichnikov A.P., Mokshin V.V., Mukhutdinov T.A., 2015. Simulation training in the Simulink package of the Matlab system. Bulletin of Kazan technol. univ. Vol. 18. 5: 184–188. (in Russian).

2. Trofimov A.V., 2015.The time index is a criterion for assessing the impact of traffic load on the quality of its functioning. Bulletin of IrSTU. 10 (105): 181–185. (in Russian).

3. Belenkiy A.V., 2008. Timelines that allow you to see the time. ComputerPress. 12: 5–12. https://compress.ru/article.aspx?id=19860 (in Russian).

4. Glinchenko A.S., 2005. Digital signal processing [Text]: textbook / A.S. Glinchenko. 2nd ed., Revised. and add. Krasnoyarsk: PH KSTU, 234 p. (in Russian).

5. Lychkina, NN, 2005. Simulation of economic processes: manual / NN. Lychina. Moscow. 220 p.

6. Chuba M.V., Keleman I.N., Garanja I.A., Stasyuk A.F., Verbitsky Yu.T., Nishchimenko I.M., Plishko S.M., Verbitskaya O.Ya., Davydyak O.D., Oleinik G.I., Simonova N.A., Burlutskaya A.M., Evdokimova O.V., 2011. Catalog and detailed data on earthquakes in the Carpathian region for 2011. Seismologist. Bull. of Ukraine for 2011. Sevastopol, SPC "Ecosee-Hydrophysics", pp. 115–182. (in Russian).

7. Bakhvalov N.S., Zhidkov N.P., Kobelkov G.M., 2008. Numerical methods. - 5th ed. Moscow: Binom. (in Russian).

8. Anderson T., 1976. Statistical time series analysis. 755 p. Moscow: Mir, (in Russian).

9. Velichko V.V., 2016. Comparative analysis of statistical software packages. Innovative science. 5: 32–35. ISBN 2410-6070. (in Russian).

10.Cooley James W. and Tukey John W., 1965. An algorithm for the machine calculation of complex Fourier series. Mathematics of Computation, p. 297–301.

11.Shishov V.V., Ivanovsky A.B., 2006. Comparative analysis of moving synchronization coefficients in the analysis of time series. Bulletin of the Siberian state. Aerospace University by Academician M.F. Reshetnev. Pp. 29–33. (in Russian).

12.Dubovenko Yu.I., Chorna O.A. On the ambiguity of 4D gravity monitoring of geological media. Геофізичний журнал. 2010. 3**2**, № 4. С. 41–46.

13.Dubovenko Yu.I., 2015. Notices on the strategy for creating of digital gravimetry databases in Ukraine. Geoinformatika. 4(56): 65–74. (in Russian).

ОБМЕЖЕННЯ ОПТИМІЗАЦІЇ СТАТИСТИЧНОГО АНАЛІЗУ ДАНИХ ТЕХНІЧНОГО МОНІТОРИНГУ ТРАНСПОРТНИХ МЕРЕЖ

Кузьмінець Микола Петрович, доктор технічних наук, доцент, Національний транспортний університет, завідувач кафедри комп’ютерної, інженерної графіки та дизайну, e-mail: kuzminecmp@ukr.net, тел. +380442849713, Україна, 01103, м. Київ, вул. М. Бойчука 42, к. 705. http://orcid.org/0000-0002-9636-919X

Дубовенко Юрій Іванович, кандидат фізико-математичних наук, доцент, Національний транспортний університет, викладач кафедри комп’ютерної, інженерної графіки та дизайну, e-mail: nemishayeve@ukr.net, тел.: +380665979384, Україна, 01103, м. Київ, вул. М. Бойчука 42, к. 709. http://orcid.org/0000-0002-8128-5989

Анотація. У процесі неперервного технічного моніторингу транспортних мереж та об’єктів дорожньо-мостового господарства накопичуються великі обсяги даних неперервних вимірювань різноманітних технічних показників. Ці бази даних містять окремі часові ряди записів технічного моніторингу транспортних мереж, що включають, крім корисних сигналів, також випадкові завади (шуми) та пропуски даних різної природи. Аналіз таких даних, що ускладнені нелінійними похибками, вимагає спеціального статистичного інструментарію. Широко вживані статистичні пакети для обробки випадкових даних, здебільшого не містять повного циклу засобів та алгоритмів, необхідних для обробки великих часових рядів та інших спостережень великої розмірності. Зокрема, лінійна шкала часу для обробки періодичних даних в поширених програмах цифрової статистики відсутня. А тим часом, накопичено позитивний досвід обробки даних, що ускладнені похибками, в зовсім іншій галузі знань – геофізичному та геоінженерному моніторингу. І звідти можна запозичити деякі прийоми та засоби.

Зокрема, шляхом аналізу технічних спостережень із подібними початковими умовами та часовими обмеженнями, виявлено, що методика ковзного вікна підходить з достатньою для практики точністю для обробки перерваних часових рядів. Проте її головними недоліками є істотне обмеження щодо довжини рядка запису та підвищена чутливість до прогалин даних. Нехтування цими обмеженнями методу призводить до істотних спотворень моделі корисного сигналу під час потокової фільтрації вхідної інформації. Тому графік обробки часових рядів потребує певної внутрішньої оптимізації, щоб адаптуватись до накопичення похибок під час послідовних ітерацій з первісної обробки даних.

Визначено компактний алгоритм, який містить необхідні кроки для оптимізації графіка обробки часових рядів. Ці кроки полягають у наступному: необхідне зберігання даних у внутрішній базі даних системи спостережень; побудова зразків даних для порівняння має виконуватись лише у єдиному масштабі часу; цільову вибірку даних для конкретної задачі доцільно здійснювати на основі метаопису усієї серії спостережень; осереднення найбільш оптимально здійснювати у ковзному вікні; крім того, доцільно задіяти прив’язку масиву даних та окремих серій до календарних дат та передбачити маски для генералізації пропусків (як при пошуку у файлових базах даних). Також не зайвим буде можливість побудови зведених графіків спостережень та зберігання графіки у поширеному векторному форматі.

Розкриваються умови вивчення часових рядів технічного моніторингу, а саме такі, як база даних, календарна структура даних, обробка прогалин, оптимальний пакет чисельних методів аналізу даних, послідовна їх обробка у ковзному вікні тощо. Всі пропозиції проілюстровано конкретними прикладами технічного та геофізичного моніторингу, що виконувались авторами у попередні роки.

Ключові слова: моніторинг, часовий ряд, оптимізація, усереднення, ковзне вікно.

References

1. Якимов И.М., Кирпичников А.П., Мокшин В.В., Мухутдинов Т.А. Обучение имитационному моделированию в пакете Simulink системы Matlab. Вестник Казанского технолог. ун-та. 2015. Т. 18, № 5. С. 184–188.

2. Трофимов А.В. Временной индекс – критерий оценки влияния загрузки улично-дорожной сети на качество ее функционирования. Вестник ИрГТУ. 2015. № 10 (105). С. 181–185.

3. Беленький А.В. Timelines, которые позволяют увидеть время. КомпьютерПресс. 2008. № 12. С. 5-12. https://compress.ru/article.aspx?id=19860

4. Глинченко А.С. Цифровая обработка сигналов [Текст]: учеб. пособие / А.С. Глинченко. 2-е изд., перераб. и доп. Красноярск: ИПЦ КГТУ, 2005. 234 с.

5. Лычкина, Н.Н. Имитационное моделирование экономических процессов: учеб, пособие / Н.Н. Лычкина. Москва, 2005. 220 с.

6. Чуба М.В., Келеман И.Н., Гаранджа И.А., Стасюк А.Ф., Вербицкий Ю.Т., Нищименко И.М., Плишко С.М., Вербицкая О.Я., Давыдяк О.Д., Олейник Г.И., Симонова Н.А., Бурлуцкая А.М., Евдокимова О.В. Каталог и подробные данные о землетрясениях Карпатского региона за 2011 год. Сейсмолог. Бюлл. Украины за 2011 год. Севастополь, НПЦ «Экоси-Гидрофизика», 2011. С. 115–182.

7. Бахвалов Н.С., Жидков Н.П., Кобельков Г.М. Численные методы. – 5 изд. Москва: Бином, 2008.

8. Андерсон Т. Статистический анализ временных рядов. Москва: Мир, 1976. 755 с.

9. . Величко В.В. Сравнительный анализ статистических пакетов программ. Инновационная наука. 2016. № 5. С. 32–35. ISBN 2410-6070.

10.Cooley James W. and Tukey John W. An Algorithm for the Machine Calculation of Complex Fourier Series. Mathematics of Computation, 1965 p. 297–301.

11.Шишов В.В., Ивановский А.Б. Сравнительные анализ скользящих коэффициентов синхронности при анализе временных рядов. Вестник Сибирского гос. аэрокосмич. ун-та им. академика М.Ф. Решетнева. 2006. С. 29–33.

12.Dubovenko Yu.I., Chorna O.A. On the ambiguity of 4D gravity monitoring of geological media. Геофізичний журнал. 2010. 3**2**, № 4. С. 41–46.

13.Дубовенко Ю.И. Замечания о стратегии создания цифровых баз данных гравиметрии в Украине. Геоинформатика. 2015. № 4(56). С. 65–74.