Development and study of a parallel algorithm of iteratively forming latent functionally-determined structures for classification and analysis of meteorological data

V A Sorokin¹, Yu V Volkov², A I Sherstneva³, and I A Botygin²

¹ Undergraduate, National Research Tomsk Polytechnic University, Tomsk, Russia
² Senior Researcher, Institute of Monitoring of Climatic and Ecological Systems SB RAS, Tomsk, Russia
³ Assistant professor, National Research Tomsk Polytechnic University, Tomsk, Russia

E-mail: vas1@tpu.ru, yvvolkov@tpu.ru, sherstneva@tpu.ru, bia@tpu.ru

Abstract. This paper overviews a method of generating climate regions based on an analytic signal theory. When applied to atmospheric surface layer temperature data sets, the method allows forming climatic structures with the corresponding changes in the temperature to make conclusions on the uniformity of climate in an area and to trace the climate changes in time by analyzing the type group shifts. The algorithm is based on the fact that the frequency spectrum of the thermal oscillation process is narrow-banded and has only one mode for most weather stations. This allows using the analytic signal theory, causality conditions and introducing an oscillation phase. The annual component of the phase, being a linear function, was removed by the least squares method. The remaining phase fluctuations allow consistent studying of their coordinated behavior and timing, using the Pearson correlation coefficient for dependence evaluation. This study includes program experiments to evaluate the calculation efficiency in the phase grouping task. The paper also overviews some single-threaded and multi-threaded computing models. It is shown that the phase grouping algorithm for meteorological data can be parallelized and that a multi-threaded implementation leads to a 25-30% increase in the performance.

1. Introduction
The fundamental scientific researches include tasks of forecasting weather, predicting climate and global changes in the atmosphere. Their effective solution requires high-performance computing systems that allow parallelization. Due to global changes in climate, it is very relevant to study the multidimensional geophysical data and detect important dependences (usually, in time) of different meteorological parameter fields or other parameters for regions of different scales. For example, studying various time trends in a given area of distribution, since trend mathematical description allows to make prognoses on studied parameters. Clustering technology is one of the basic approaches for solving such tasks, especially for processing experimental data of geophysical nature. For this problem the parallel computing is necessary due to big volume of processed data. For a single location, the number of geophysical data bins may reach a few thousands, while properly analyzing climatic changes might require processing a huge number of locations. Besides having to store a significant amount of data, it is also necessary to have a non-trivial approach to mathematical
processing. To summarize, implementing parallel computing is one of the most important tasks for meteorological data classification and analysis [1-4].

As mentioned above, the cluster analysis is a possible approach to mathematical processing of geophysical data (meteorological data in particular). In general, the method consists of the following steps: selecting objects for cluster analysis, forming the criteria and metrics for object evaluation, determining objects’ identity according to metrics and criteria, using identical objects to form clusters (groups). The described steps are executed iteratively. The optimal result might require altering the key parameters at every clustering step.

Depending on the way of data processing, the cluster analysis methods are divided into hierarchical and non-hierarchical. The hierarchical methods consequently merge smaller clusters into the bigger ones (agglomerative) or split big clusters into the smaller ones (divisive) [5].

Depending on the way of data analysis, there are non-overlapping (precise) and overlapping (fuzzy) clustering algorithms, as well as single-step or multi-step methods depending on the number of iterations [6].

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) clustering algorithm is based on double-step hierarchical clustering and can effectively be used for very large volume of data. This algorithm, however, cannot work with anything except numerical data.

K-means method decomposes a set of objects into a preset number of clusters. The algorithm is easy to use while sensitive to emissions (which can distort mean values) and requires knowing the exact number of clusters [7].

CLOPE (Clustering with slope) algorithm is intended to work with transaction data, well-scalable and can boast a high performance and clustering quality. The method automatically selects a required number of clusters and is widely used in data mining [8].

PAM (Partitioning around medoids) algorithm is a modification of K-means method. It is very effective for smaller amounts of data [9] and widely used in such fields, as text analysis, bioinformatics and intelligent transport systems [10].

Konohen self-organizing maps is one of the neural algorithms without teacher. It allows to obtain a set of objects with reduced dimensionality compared to the original [11].

Concluding this brief overview, we have to point that there is no universal clustering algorithm. It is proven by the existence of a whole number of heuristic criteria as well as algorithms that do not have a clear criterion, yet are capable of doing a reasonable clustering. Moreover, the clustering result significantly depends on metrics that is usually selected by expert’s subjective judgement [12]. As such, finding the optimal approach for a given task is only possible by analyzing the nature of data and task specifics.

2. Grouping algorithm

This work studies thermal fields with the goal of finding latent functionally determined structures, which would allow a new way to analyze and classify sets of measurement data. Generally, natural climate changes show a cyclic behavior, so it is possible to do a research using signal spectral analysis methods. The authors of this work suggest an original phase grouping algorithm based on one of the possible ways of representing oscillations (analytical signal of Nobel laureate Zoltán Pál Dienes Gabor). Phase is the main parameter of oscillation process. While amplitude describes variable parameter range, phase contains an information on periodicity of variable change over time. Using cluster analysis methods and algorithms, researching process phase allows to obtain new estimates that describe the behavior of meteorological parameters over a given period of time.

A study of monthly average temperature data over the 56 years at 818 weather stations shows that data changes form an oscillating process with quasi-period of one year. The temperature values deviating from an annual cycle may possibly describe seasonal climate changes. However, finding consistent dependencies proves a complicated task [13]. The researches show that frequency spectrum of temperature oscillating process is narrow-banded and has only one mode for most weather stations, which allows to use causality conditions and introduce oscillation phase [14]. The yearly phase
component, being a linear function, was removed from the phase for 56 years period of time. Afterwards, the remaining phase fluctuations allow to consistently research their coordinated behavior and timing using the Pearson correlation coefficient for evaluation.

We’ve developed an algorithm to compute phase values for each set of temperature values. Its general scheme is shown in figure 1 below.

Step 1. Fast Fourier transform (moving from time domain to frequency domain). After input data is uploaded, every set of temperature values is converted into a sum of sinusoidal functions with Fourier discrete transform (1).

\[ X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}kn}, k = 0 \ldots N - 1 \]  

Fourier discrete transform is based on fast Fourier transform algorithm.

Step 2. Filtering. Median filtering is used to reduce data noise. One-dimensional median filter, based on heuristic processing method, uses sliding-window method to average segments of temperature data.

Step 3. Inverse Fourier transform. This algorithm is used to return from phase domain into time domain.

\[ x_n = \frac{1}{N} \sum_{n=0}^{N-1} X_k e^{\frac{2\pi i}{N}kn}, n = 0 \ldots N - 1 \]  

Step 4. Computing phase. Using the analytic signal theory, the phase value is found as an argument to the ratio of the real part to imaginary part. From the computing point of view, the analytic signal corresponding to the given real function is found through Fourier transform and following inverse Fourier transform by positive phases. The phase value obtained at this step is one of the main values of arc tangent function within yearly period of time (figure 2).
Step 5. Computing continuous phase function over the entire observation interval.

Step 6. Computing the linear component. It is subtracted from the continuous phase function, which allows to form phase fluctuations of the given temperature data set (figure 3).

An algorithm to properly group weather stations with similar climate was developed. It compares phases by calculating pairwise correlation coefficient for each one, uses them to obtain standard phases and generates the final station grouping accordingly.

A general scheme of standard phase calculation is shown in figure 4 below.
Stage 1. The phase values computed for each group are sent to the correlation table calculation node input, which outputs pairwise correlation coefficients.

Stage 2. The resulting phase correlation table is sent to the group forming node. The groups are created from the temperature data sets which, alongside group-forming data set, have correlation coefficients above given level.

Stage 3. The standard phase computing node uses the created groups to calculate standard phase values for the current iteration.

The described procedures are done iteratively, with previous level standard phase values as input data and next level standard phase values as output data. The algorithm is executed for each set of temperatures and, consequently, for each weather station. Additionally, the groups are changed according to the given threshold criterion – primarily, through inclusion of new elements which did not belong to the group at the previous stage. The threshold determines group size and coherence between current phase and previous ones during the iteration process.

Every calculation in algorithm is divided into two parts: temperature data phases are computed during the preparation part, grouping is performed during the main part.

3. Parallel computing
Let us look into the possibility of parallel computing in the phase clustering algorithm suggested above. During the preparation phase the calculations are linear and each step requires results from the previous step. Therefore, parallel computing is not possible here. However, since it is necessary to process a lot of temperature data sets (above three hundred) it might be expedient to use parallel computing to prepare each set.

All of the parallel computing steps are based on MapReduce model shown in figure 5 below.
Figure 5. General scheme of parallel computing.

The task distribution block decomposes input data by available computing power to calculate phases of temperature data sets (Map procedure). Afterwards, the results obtained by each node are collected in data gathering block to prepare for further analysis (Reduce procedure).

The parallel computing is also not possible for all of the phase grouping algorithm, because this process is iterative as well (each iteration requires result from the previous step). However, it is possible to implement parallel computing for certain algorithm steps. In the phase grouping block diagram (figure 2), the parallel computing is possible for «Calculation of correlation table» and “Calculation of standard phases” steps. The general scheme of parallel computing (figure 5) is being used in these cases.

The process of correlation table calculation can be distributed between available calculating powers. After the data is gathered into a single table (based on final table analysis), it is possible to implement parallel computing for standard phase calculation process.

- Multi-threaded computing in standard mode;
- Multi-agent computing based on distributed computational system.

The efficiency analysis has shown that phase grouping algorithm for mete

Over the course of this work, we’ve conducted program experiments aimed to evaluate computation efficiency in phase clustering task. The following computing models were reviewed:

- Single-threaded computing in standard mode.

Phase algorithm groups using meteorological data lends itself to parallelization, and a multi-threaded implementation provides increased productivity by 25-30%.

4. Results
The steps described above were used as a phase clustering task solution for 818 weather station over the 56 year observation period. A text file with every row containing monthly average temperatures for weather stations was used as an input data.
To form the groups of stations with similar climate, we’ve developed an algorithm based on selecting standard phases, related to main groups with the help of pairwise correlation table.

The first stage of algorithm inputs computed phase values into correlation table calculation node, resulting in pairwise correlation coefficient for each weather station. It uses Pearson correlation coefficient:

\[ r_{xy} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}} \] (1)

After obtaining the correlation table, algorithm forms weather station groups with coefficients of correlation to other weather stations above given \( r \) value. The phase value for every group is average for every phase in the group. The default phase function for grouped stations are replaced with current ones. Then the process is repeated and groups change according to \( r \) criterion by including new elements that group didn’t contain in the previous iteration. Eventually, current phases will have their standard properties defined and stop changing due to process convergence.

Afterwards, the duplicates are removed from the group list and it allows to calculate the correlation table of default phases together with computed standard phases. The result of the algorithm is shown in figures 6 and 7.

**Figure 6.** First result of phase grouping algorithm.

**Figure 7.** Second result of phase grouping algorithm.

5. Conclusions

The monthly average changes over the studied period form an oscillation process with quasi-period of one year. The temperature change deviation from yearly cycle is a definitive feature of weather. However, determining consistent dependencies is complicated. The analysis of climate field structure (zoning, classification etc.) is aimed to find regularities of climate type formation in global climate system. On the other hand, defining the territorial boundaries of climate types varying by properties
allows to decompose the massive amounts of stored climate data into a smaller number of information structures.

This work suggest an original method for determining climate fields which allows to solve both tasks for any spatial and temporal scales. The method is based on the fact that thermal oscillation process phase spectrum is narrow-banded and has only one mode for most weather stations. This allows to use the causality conditions and introduce the oscillation phase by using analytic signal theory. The yearly phase component is a linear function which was removed from phase by least square method. The remaining phase fluctuations allow to consistently research their coordinated behavior and timing, using the Pearson correlation coefficient for dependence evaluation.

In the case of surface air temperature, the developed algorithm allows to find climate structures with coordinated changes in thermal field, to make conclusions on climate uniformity for given region and to research climate changes over time by analyzing standard group deviations.

The results are obtained within the framework of the project "Development and research of intelligent information-analytical system for analysis and prediction of climatic processes on the basis of high-performance clusters" of RAS 1.33P fundamental research program "Fundamental problems of mathematical modeling. Fundamental problems of factorization methods in different fields. Algorithms and software for ultra-high performance computing systems".

References
[1] Sherstnev V S, Sherstneva A I, Botygin I A and Kustov D A 2016 High Technology: Research and Applications (Key Engineering Materials vol 685) (Switzerland: Trans Tech Publications) pp 867–871
[2] World Community Grid http://www.worldcommunitygrid.org/ Accessed: January 2016
[3] Botygin I A, Popov V N, Tartakovsky V A and Sherstnev V S 2015 Proc. 21st Int. Symp. Atmospheric and Ocean Optics: Atmospheric Physics vol 9680 (Tomsk: SPIE) pp 96800J1–96800J4
[4] Popov V N, Botygin I A and Koshelev N V 2016 High Technology: Research and Applications (Key Engineering Materials) vol 685 (Switzerland: Trans Tech Publications) pp 925–929
[5] Hierarchical agglomerative methods http://stu.sernam.ru/book_fan1.php?id=81 Accessed: January 2016
[6] Neysky I M Classification and comparison of clustering methods http://it-claim.ru/Persons/Neyskiy/Article2_Neisky.pdf Accessed: January 2016
[7] K-Means Clustering means http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html Accessed: January 2016
[8] Clustering of categorical data: scalable algorithm CLOPE https://basegroup.ru/community/articles/clope Accessed: January 2016
[9] Introduction to Modern Data Mining http://statistica.ru/local-portals/data-mining/ Accessed: January 2016
[10] Rechkalov T V 2015 Supercomputer days in Russia 2015 (Moscow: Publishing house of the Moscow State University) pp 632–640
[11] Kohonen self-organizing maps: mathematical apparatus https://basegroup.ru/community/articles/som Accessed: January 2016
[12] Clustering http://www.machinelinearning.ru/wiki/index.php?title=%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F Accessed: January 2016
[13] Volkov Yu V and Tartakovsky V A 2013 Energy: Efficiency, reliability, safety (Tomsk: SCAN) pp 177–178
[14] Vakman D E and Weinstein L A 1977 Physics-Uspekhi 123 657–682