Evaluation of Clusters for Climate Data

DZENANA DONKO, NEJRA HADZIMEJLIC, NIJAZ HADZIMEJLIC
Faculty of Electrical Engineering
University of Sarajevo
Zmaja od Bosne bb, Kampus Univerziteta, 71000 Sarajevo
BOSNIA AND HERZEGOVINA
ddonko@etf.unsa.ba, nejra.hadzimejlic@gmail.com, nhadzimejlic@etf.unsa.ba

Abstract: - Climate data analysis is a progressive research area that focuses on analysis of change of climate conditions, investigation of climate phenomena and evaluation of interconnections of climate conditions. Data mining techniques introduce the effective and efficient way to analyze large amount of data in climatology. In this paper is presented the algorithm for climate data analysis using the clustering data mining techniques. The developed solution represents evaluation of climate data from the different points of view in order to provide a complete view of the data. Climate research experts can use these results to draw their own conclusions and perform detailed climate change analysis. Climate data is represented graphically as the map of measured climate parameters, the map of climate clusters identified in specified moment of time and the map of evolution steps identified between the consecutive time slices.

Key-Words: - Data mining, clustering, climate, hierarchical clustering, meteorology, evolving clusters

1 Introduction

Climate data analysis, performed in order to understand climate change process and effect of different environmental factors in that change, has been focus of interest of researches for many years [1-3], Industrialization, increase of population, destruction of environment and global change of many factors that influence the state of atmosphere changed climate conditions in the long term. Climate change disturbs natural balance, leading to endangerment of many species and their habitats.

Atmospheric conditions can be represented with numerous parameters (temperature measurements, humidity, precipitation, sunshine duration, wind direction) and are influenced by numerous factors such as geographical location, forest coverage or urbanization of the area, water surfaces presence. All these factors should be considered in order to perform complete analysis of the state of the atmosphere and in order to properly detect interconnection of these factors. Climate data analysis is performed on large amount of data and data mining techniques can be used to extract different information, such as the region detection, identification of the relevant atmosphere factors and their role in atmospheric conditions, data for atmosphere change analysis.

Data mining process includes the two main steps: data pre-processing / preparation of data for analysis, and the process of analysis itself performed by one of the data mining methods, depending on the goal of analysis. Clustering, as one of data mining methods, can identify groups of similar objects in the data set, where similarity is determined based on the distance function [4]. The goal is to group objects in the data set in a manner that most similar objects are in the same group and the difference between objects in the different groups is maximal. Clustering process can be performed using different approaches, depending on how similar objects are identified. Some of the clustering methods are partitioning methods, hierarchical methods, grid-based methods, density-based methods.

This paper presents solution for climate data analysis using clustering methods in order to identify atmospheric conditions in one time slice and change of those conditions between two consecutive time slices (atmosphere state evolution). Initial problem was analyzed and separated into three simpler sub problems that needed to be solved: presentation of climate data measurements for specified time slice, presentation of atmospheric conditions by identifying regions with similar state of the atmosphere and presentation of change of those conditions between two consecutive time slices. By putting together these three algorithms a complete solution that addresses initial problem is provided.

The main reason for selection of this topic is the importance of understanding climate change in

order to understand correlation of different factors in the nature. That kind of knowledge contributes in solving problems regarding sustainable development of mankind, and it can be said that understanding atmospheric conditions and processes occurring in the atmosphere is a small but important segment in solving problems that involve all nature.

Currently, this kind of research, data mining on climate data, is in its initial phase since there were not enough real measurements data on which analysis could be performed. Climate prediction models based on the complex physics calculations were and continue to be in use, but there are too many correlated parameters involved to model them correctly in formulae. Advantage of the data mining over prediction models is that it uses real data. In this way model is proven to be exact, and makes conclusions based on that. Its downside, which is becoming less present, is obtaining large amounts of real climate parameters measurements.

In the next chapter is defined problem that will be solved, which is the climate data analysis in one time slice and change of atmospheric conditions between consecutive time slices. The main ideas on how to perform detection of evolving clusters and work of researchers regarding clustering methods in analysis of changing clusters are presented. Third chapter gives the idea of the algorithm developed in this paper and provides a brief analysis of applicability of the different clustering methods in solving the initial problem. Implemented algorithm for climate data analysis is given in chapter four. It consists of the three algorithms that are explained in details and results of each algorithm are given. Last chapter summarizes the work presented in this paper and propose some ideas for further development in this area.

2 Atmospheric conditions evolution

Development of the methods for climate data analysis is a progressive research area regarding the fact that the understanding of atmosphere conditions and their changes is one of the key components in sustainable development of mankind [5]. This paper is based on the need of identifying states through which atmosphere passes in consecutive time slices in order to further analyse rules of those changes and influence of different parameters on those changes. It is extension version of some previous work and analysis of available data set [6].

The goal of this paper is to develop a solution to analyse atmospheric conditions in each specified time slice and to analyse evolution of those conditions between consecutive time slices through the visual representation of state of the atmosphere in specified time slice T_X and representation of evolution of it from slice T_X to T_{X+1}. Identification of evolution of atmospheric conditions implies determining presence of change of the state of the atmosphere in two consecutive time slices: state is unchanged in certain areas and in some areas climate characteristics have spread or disappeared or have changed in other ways. Clustering methods are applicable for this problem because in each time slice, areas with similar enough state of the atmosphere can be observed as weather clusters and in that case evolution of state of the atmosphere is seen as the evolution of weather clusters. Two basic requirements can be identified based on what has been said:

- Identification of clusters of areas with similar state of the atmosphere in each time slice
- Identification of similar weather clusters in consecutive time slices in order to determine cluster evolution

Applicability of different clustering algorithms in identifying state of the atmosphere in specified time slice will be analysed later. In this chapter is given review of some papers that work on cluster evolution since it is a problem that not many researches discuss and is presented in different research areas in some form.

One of the papers in this area of research evaluate problem of identifying groups of migrating animals [7]. Animals usually move in the groups, but certain animals leave the group and/or others join it. A group of animals can be observed as a cluster that always has the same identity, but moves through space in time and portion of its objects is removed or added to it during that time. In order to solve this problem a moving cluster is introduced and methods for detection of such clusters are represented. A moving cluster is defined as a sequence of spatial clusters that appear in consecutive time slices of momentarily layout of objects in space and each two consecutive spatial clusters have a certain amount of objects in common. Improvement of this paper is the fact that spatial movement of cluster is considered and cluster in two consecutive moments can have partially different data set of objects in it.

Addition to this change mining is introduced in order to emphasize the need that, beside the consideration regarding the adaptation of current data set to existing one (e.g. data set of the same cluster is different because new items have been added to it over time), rules of changes of data set between time slices also have to be analysed [8]. In other words, change mining focuses on determining

the rules of changes of data model. Change mining is defined as an example of temporal data analysis in order to discover, model and interpret changes in the model that best describes data set. Cluster evolution can be analysed by tracking changes in data and determining the rules of those changes.

The above mentioned papers suggest methods for dealing with certain problems regarding changing clusters and provide theoretical knowledge for solving problems such as climate change tracking. In those methods criterion of cluster similarity, where one cluster is from time slice T_X, other one from T_{X+n}, is based on analysis of objects that are in clusters and two clusters were determined as similar if they had enough objects in common. Such criterion is also called cluster data set similarity in which clusters are formed based on similarity of objects regarding all dimensions. Downside of that approach is that, in case of high-dimensional data, clusters could be hidden in subspaces that consider only some dimensions. Also, tracing similar clusters through time and space and defining it as a cluster evolution does not represent well enough the problems in the climate data analysis, since the focus there is tracing similar atmospheric conditions.

For example, correlation of smoke concentration and air temperature can be considered as probability of the forest fire. High temperatures and increased smoke concentration phenomena occurs both on northern and southern hemisphere in summer time, which means that the difference between occurrences of these phenomena is half year and location is different. Algorithms for analysis of such climate phenomena should identify this similarity, but because of mentioned disadvantages this pattern would not be recognised.

Cluster in climate data analysis is more of generalised climate phenomena, rather than just a set of similar objects and alteration of understanding what cluster is crucial for implementation [9]. Their solution overcomes problems that have been present so far by cluster evolution as the tracing is based on similarity of values of objects. Clusters are traced in subspaces and similarity of objects is determined based on some, rather than all dimensions. Similarity of subspace clusters is not new but the innovation is its application in clustering algorithms for evolution tracking. With such understanding of cluster evolution it is possible to solve problem mentioned in previous section.

3 Selection of the clustering method for climate data analysis

In previous section we have defined the problem to be solved: representation of state of the atmosphere for specified time slice and identifying evolution of it from one moment to the other, and all this in order to provide high-quality data for climate analysis.

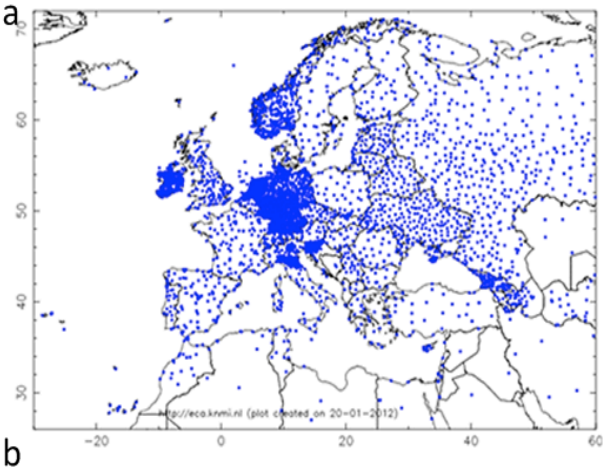
Initial problem can be split into two sub problems: first one is identifying points in data set that have similar atmosphere conditions in specified time slice T_X and second one is tracking changes of those conditions from T_X to T_{X+1}. Obviously, algorithm for solving the initial problem can also consist of two parts: first algorithm which will determine climate clusters in one specified time slice (and in this climate cluster can be observed as a cluster of objects from data set that are geographically near enough and have similar measurements of atmospheric conditions parameters: temperature, sea level pressure etc.), and second algorithm to solve the defined sub problem of tracking climate clusters evolution over the time.

It is first necessary to identify points in the data set that have similar values of the attributes (and attributes are in this case different atmosphere parameters) which implies usage of the clustering algorithm. That algorithm has to take in account geographical proximity and measured atmospheric parameters similarity. Two points which are on opposite ends of Europe cannot be in the same cluster even if they have exactly the same atmospheric conditions and two points of measurement, even if they are geographically very near, can be in different clusters if they don't have similar enough atmospheric conditions (e.g. due to the fact that they are separated by a mountain). If it is snowing and is cold in one and warmer and without precipitation in other area those points cannot be in the same cluster, regardless of their geographical proximity. Once atmospheric conditions are evaluated for several consecutive time slices, it is necessary to identify changes of those conditions that have occurred. Solution of that sub problem has to take in account the time and the spatial scale of data to be analysed in order to provide correct results.

Defined sub problems determine time and spatial scale of data: for these analyses we need measurements of atmospheric parameters for each day from enough locations considering the size of the area. Data set available from European climate assessment and data project was analysed [10]. Data set contains meteorological stations specifications

and measurements from each station. Measurements of 12 atmospheric parameters (figure 1.b) from 4824 stations (figure 1.a) are available for a period of over 20 years.

a



b

Maximal temperature TX	Minimal temperature TN	Mean temperature TG
Humidity HU	Percipitation RR	Sea level pressure PP
Sunshine duration SS	Cloud cover CC	Snow depth SD
Wind gust FX	Mean wind speed FG	Wind direction DD

Fig.1 Map of meteorological stations [10] (a) and measured atmospheric parameters (b)

First step in providing the solution is data pre-processing in order to prepare them for further analysis [11]. Pre-processing involved several steps:

- Unifying data in one data source: data has been imported from textual files to database, each row representing measurements for one day from one station
- Deleting rows that don't have a single valid measurement (and by valid we mean that measurement exists - is not NULL, and it is not a code for missing value, which is -9999): lots of rows like this are present in last few years due to the fact that quality assurance of data is not effective
- Deleting all measurements for stations that don't have enough valid measurements
- Deleting measurements of dimensions that don't have enough consecutive measurements: it is important that analysed measurements are available in consecutive time slices for all stations on a given period of time in order to identify climate clusters evolution on a daily basis

Filtered data set after pre-processing consist of the measurements of parameters TN, TX and RR with geolocation of stations, forming five dimensional space defined by numerical attributes. A subset of data set consisted of measurements during one year (1990.) from 689 stations was used in further analysis because in that period complete consecutive daily measurements were available.
Based on what has been mentioned so far about data set and with knowledge of problem we can say that:

- Numerical type of data is analyzed
- Similarity of points can be determined easily since attributes are numerical values
- Density of points in space is uneven
- Number of clusters to be formed is not known in advance
- Data set to be analyzed is not very large since in one algorithm execution measurements of only one or two time slices are processed
- Clusters to be formed are irregularly shaped
- Each point can be assigned to only one cluster in clustering process

Clustering methods, depeding of approach in cluster formation, are [4]:

- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Models-based methods

If we analyze applicability of certain method in solving defined problem we can conclude that:

- All mentioned methods can process numerical data
- None of the methods should have problem regarding dimensionality and data set size
- Partitioning methods and certain algorithms cannot detect irregularly shaped clusters and thus they are not convenient in solving this particular problem (hybrid methods that combine different algorithms and can detect irregular clusters could be used)
- Clustering process should be intuitivelly manageable in a way that change of certain parameter of algorithm and its consequences are predictable
- Algorithms that require knowledge of number of clusters to be formed are not applicable
- Density-based methods are not applicable because density of points is uneven in space

- Methods that eliminate outliers are not applicable and those points should be identified as a new cluster or should belong to a cluster already formed with similar atmospheric conditions, even if points in that cluster are a bit further geographically (if there are no other clusters between the point and the cluster)
- Methods that represent their results as a percentage of how much each point belongs to each cluster are not applicable (eg. fuzzy partitioning methods)

Regarding the way clusters are formed and considering everything mentioned so far, agglomerative hierarchical algorithm was used.

4 Development of the solution for climate data analysis

Based on the previous chapter we can point out that the goal is to create a solution that will provide information for climate data analysis for each time slice specified and information regarding evolution of state of the atmosphere from one to another time slice. Data set for one time slice consists of one-day measurements data of all parameters from all stations. Solution is developed in Matlab [12].
The initial problem solution is implemented through three algorithms:

- Algorithm for graphical representation of measured data
- Algorithm for clustering of one day measurements data (climate clusters detection for each time slice)
- Algorithm for determining climate clusters evolution

First algorithm gives us graphical representation of measured parameters TN (minimal temperature), TX (maximal temperature) and RR (precipitation) displayed on a map, where colour intensity denotes intensity of measured value (figure 2 is the output of this algorithm: measurements of TN (a), TX (b) and RR (c) for 01.04.1990.).

Graphical representation of the measured values is necessary in order to validate the results of clustering algorithm, since there is no data set that has verified clustering results: the easiest way to check clustering results of algorithm for climate cluster identification in one day data is by comparison of images representing measured data and image that represents identified clusters.

Second and third algorithm mentioned address two requests mentioned in chapter two and are main focus of this paper. From problem definition we can conclude that algorithm for cluster evolution identification uses clustering data for two consecutive dates, and that clustering data is the result of algorithm for climate cluster identification in one-day data.

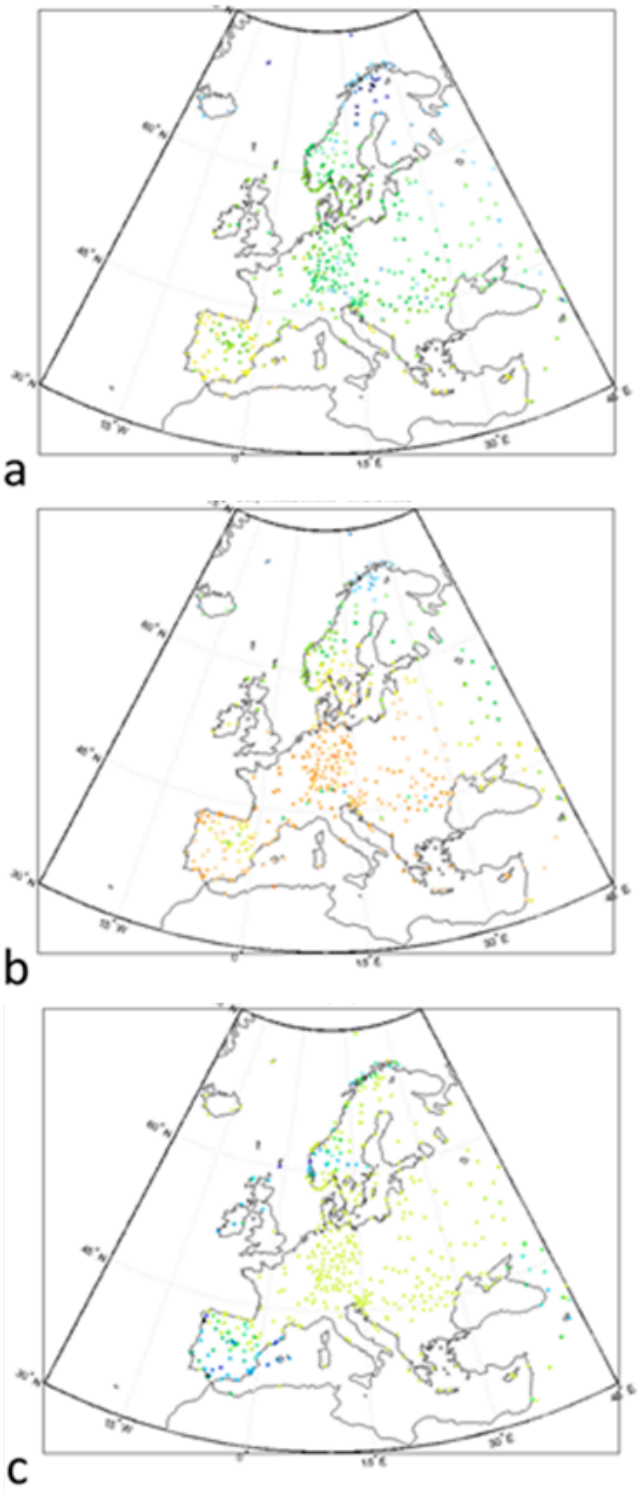


Fig. 2 Graphical representations of measurements for 01.04.1990. : TN (a), TX (b) and RR (c)

4.1 Algorithm for climate cluster identification in one-day data

Goal of algorithm for climate cluster identification in one day data is to, based on measurements of atmospheric parameters from all stations in one day, determines climate clusters for that day.

Algorithm executes in three steps:

- Data preparation
- Clustering process
- Representation of clustering results

First step is to prepare the data for clustering process. For simpler data manipulation and analysis one day measurements data is saved in .mat file and the first thing to do is to read the data. After that additional steps are required in order to prepare the data. These steps have been determined iteratively and lead to better clustering results. Steps to prepare the data for clustering algorithm are:

- Alteration of RR parameter measurements in a way that all measurement values above 420 are cut down to 420: reason for this is the fact that not many measurements had values above 420 and range of values measured in 1990. was zero to 2668 - very high values influenced clustering results too much, so they have been altered.
- Normalization of all dimensions (TN, TX, RR, LAT and LONG): ranges of all attributes have been changed iteratively in order to determine the combination of ranges that gives us the best clustering results.

Values have been normalized so that each dimensions contribution in calculation of distance function is proportional to significance of it in state of the atmosphere. Also, data normalization guarantees that measured data range is always the same, regardless of the year that measurements are from. For example we could analyse measurements from summer and winter time: normalization guarantees that values that will be analysed and processed in algorithm are similar for both periods, which simplifies selection of clustering algorithm parameters and enables us to select unique values of them, no matter the time of the year.

Second step in this algorithm is clustering process. We selected agglomerative hierarchical clustering algorithm. In the beginning of the clustering process each object is a separate cluster and cluster are unified iteratively into larger clusters based on similarity. Process ends once all objects are in the same cluster [13]. Resulting clusterization depends of parameters defined and is actually clusterization formed at certain iteration of clustering algorithm. In order to execute agglomerative hierarchical clustering algorithm in Matlab there are two key steps: determining objects linkage (creating cluster tree) and creating clusterization based on that linkage [12].

Objects linkage calculation involves two steps: calculating objects distance by applying selected metric [14] and determining objects linkage by application of selected method for linkage evaluation. Metric and method for linkage evaluation are agglomerative hierarchical clustering algorithm variables. Linkage of objects can be represented by hierarchical cluster tree and depends on selected method and metric (figure 3 represents cluster tree for data represented on figure 2 using different method and metric).

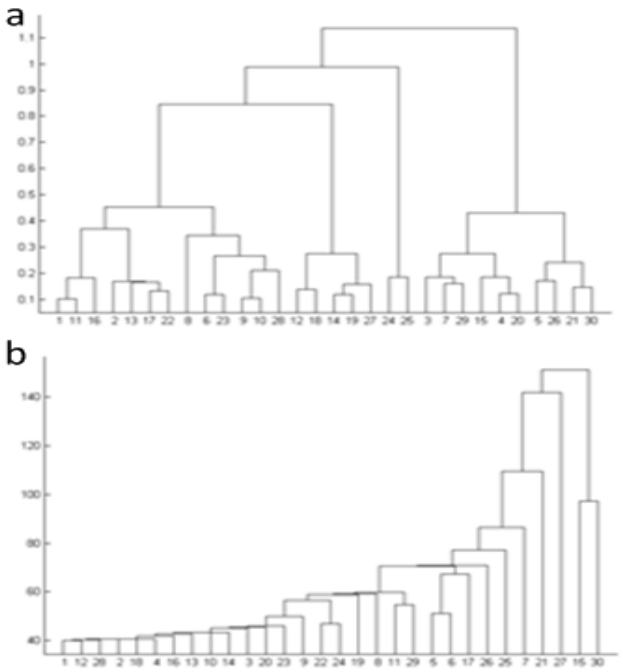


Fig.3 Formed hierarchical cluster tree if correlation metric and average - linkage method are used (a) and in case Euclidean metric and single - linkage method are used (b)

Connection height in the tree is a measure of similarity of objects that it connects - that height is also called cophenetic distance. How well hierarchical tree represents actual data set can be evaluated by comparison of those distances with original distances calculated using distance function in Matlab. If clusterization represented by that tree is valid values of these distances is approximately the same. In order to evaluate the similarity of distances a function that calculates correlation of these distances is called. It returns cophenet

coefficient which denotes similarity of distances calculated: if it is near 1, data is represented more accurately.

By altering metric and method for cluster similarity evaluation data set can be represented better or worse - cophenet coefficient gives is a measurement of how well is the data represented. By testing different metrics and methods on a range of data set we came to conclusion that the best data representation is provided when Chebyshev metric and average-linkage method is used. Chebyshev metric calculates distance of objects as the greatest of their differences along any coordinate dimension. Average-linkage method of cluster similarity evaluation calculates the difference between clusters as average of differences of each two objects where first object is from first and second object is from second cluster and in each iteration clusters with minimal distance between them are united.

After hierarchical tree has been formed we need to determine the clusterization. Clusterization can be determined in two ways: by specifying maximal number of clusters formed or by defining cutoff parameter which implies similarity between objects in the same cluster. First way of clusterization formation results by creation of specified number of clusters or less, depending on data set. This way is not good in initial phases of algorithm development since we have no knowledge of how many clusters should be formed in a certain day and wrong choice of this parameter can lead to creation of clusters of low quality (objects in it are not similar enough). During the algorithm creation we concluded that no more than 30 clusters are formed in each time slice, regardless of the time of the year or weather conditions and every clusterization with over 30 clusters is too grainy for real applications, so this mechanism of clusterization formation is used as secondary mechanism (it is applied if more than 30 clusters are formed when cutoff parameter is applied). Clusterization formation based on specification of cutoff parameter is primarily used in algorithm created for this paper. It is a more suitable solution since it guarantees us that objects in the same cluster are similar enough or similar as we want them to be.

In order to understand how cutoff parameter is applied in the clusterization formation, first we have to understand inconsistency. Inconsistency coefficient is a value related to each link in hierarchical tree and it gives us comparison of height of that link with average height of links below it in the tree (if it is approximately the same then difference between clusters that are connected by that link is small enough). Since agglomerative hierarchical clustering algorithm terminates once all clusters are unified in one, it is our task to determine the iteration of algorithm in which the clusterization was the best (objects in all clusters were as similar as we want them to be). We could easily spot that in a graphical representation of the tree, but since we need a numerical value in algorithm, inconsistency coefficient is used: links between clusters that aren't similar have high inconsistency coefficient and vice versa.

Cutoff parameter defines inconsistency and can be applied in two manners when determining clusterization. First is that cutoff defines inconsistency in leaf nodes for which all objects are grouped into on cluster, and second one is that cutoff defines height of the line of horizontal cut in the tree and all leaves at height which is equal or less than cutoff are grouped into a cluster. Through numerous iterations we concluded that the most appropriate value for cutoff parameter is 27. As we pointed out before, data normalization was necessary in order to be able to set fixed values for algorithm parameters, cutoff parameter for instance.
Representation of clustering process results is the most important thing for the user and based on an appropriate results representation we can make conclusions and further hypothesize about analysed phenomena. Clustering results are given in numerical and graphical form. Graphical representation is displayed as a map of locations of measurements where colour of the point denotes cluster to which it belongs. Figure 4 represents clusterization of one-day measurements data for 01.04.1990.

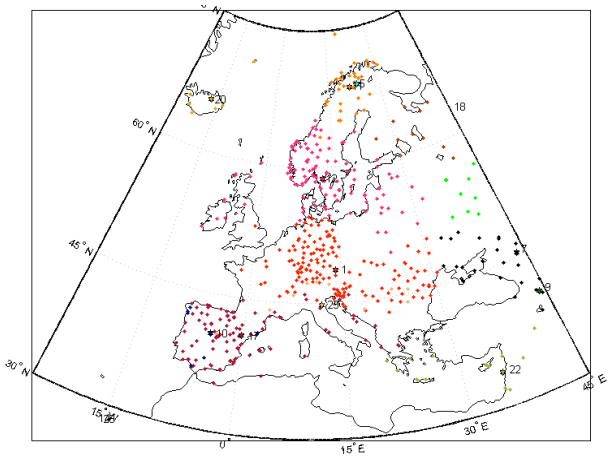


Fig. 4 Identified climate clusters for 01.04.1990. measurements

Numerical representation of clustering results is given in a form of several values. First value is already mentioned cophenetic coefficient. Second

one is vector of clusterization in which for each object in data set ID of cluster to whom it belongs is noted. Useful value is number of clusters formed and number of objects in each cluster in order to better evaluate clusterization quality. Last numerical value that describes clusterization is representatives vector. Cluster representative is calculated as mean value of objects in it for each dimension. Representative data is needed for algorithm for climate cluster evolution identification since analysis of such data set is much faster than analysing the clusters interaction as an interaction of objects that belong to them.

4.2 Algorithm for climate cluster evolution identification

After desired level of clusterization quality has been assured in climate clusters identification process, it is necessary to create an algorithm for climate cluster evolution identification. This algorithm analyses representatives data from two time slices obtained by the previous algorithm in order to determine which one of them are similar enough to be considered evolutionally connected (in other words, cluster C_{1m} from time slice T_1 and cluster C_{2n} from slice T_2 are evolutionally connected if cluster C_{2n} is similar enough to the cluster C_{1m}).

Algorithm for climate cluster evolution identification consists of three steps: data preparation, evolution identification and results representation.

During the data preparation process algorithm for climate cluster identification in one day data is executed on data from two consecutive time slices in order to provide representatives data for those two time slices. Each representative is given in a form of its geolocation and averages of measurements of TN, TX and RR parameters of objects which belong to that cluster. For algorithm to be applicable with same parameters, regardless of time of the year for which the data is analysed, representatives data has been normalized in the same manner as measurements data in the algorithm before.

Next step is to determine cluster evolution. The goal is to identify representatives that are similar enough based on their attributes values, so agglomerative hierarchical clustering algorithm was used once again to group similar objects together. Formed Clusters do not have many objects in them and if two objects are in the same cluster that means that objects (in this case clusters formed by the first algorithm) from second time slice have evolved from objects from the first time slice.

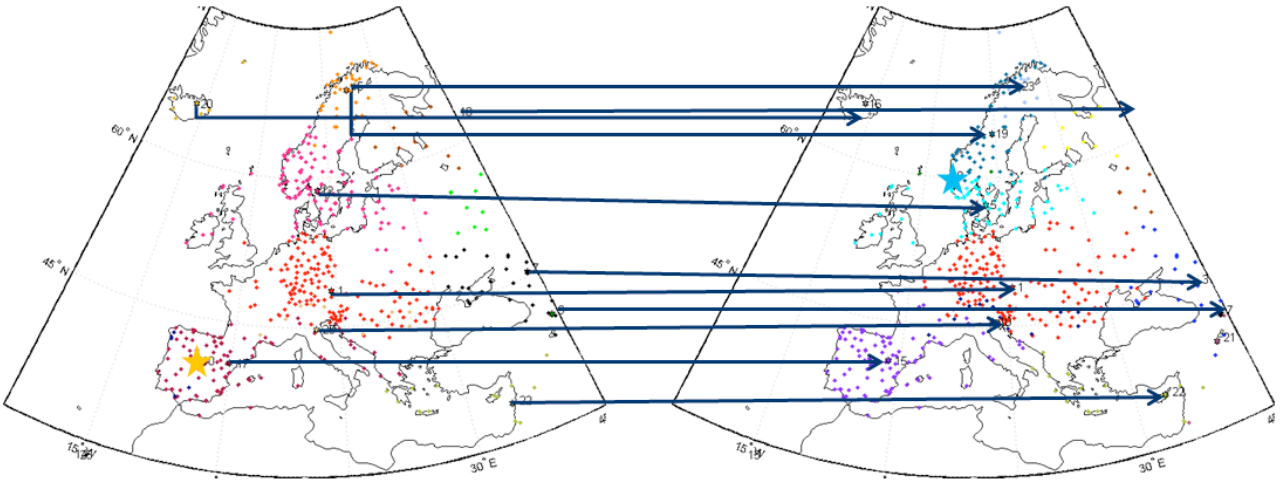


Fig. 5 Expected cluster evolution from time slice 01.04.1990. (left) to slice 02.04.1990. (right)

Creation of this algorithm first required that we manually determine expected evolutionary steps on a range of data sets (figure 5) and then algorithm parameters were altered in order to get the result that is the most similar to the expected result.

Since the data is normalized, clustering algorithm parameters have fixed values. The best combination of parameters was determined by analysing algorithm results on different data sets. Since Chebyshev metric was used in the previous

algorithm, the same metric was used in this one. Method for cluster formation was selected by analysing the number of correctly identified evolutionary steps that we determined manually before (represented in figure 5). Also, by analysing hierarchical tree formed for different data sets we concluded that the best cutoff value is 30 and it is applied as a horizontal cutoff in the tree.

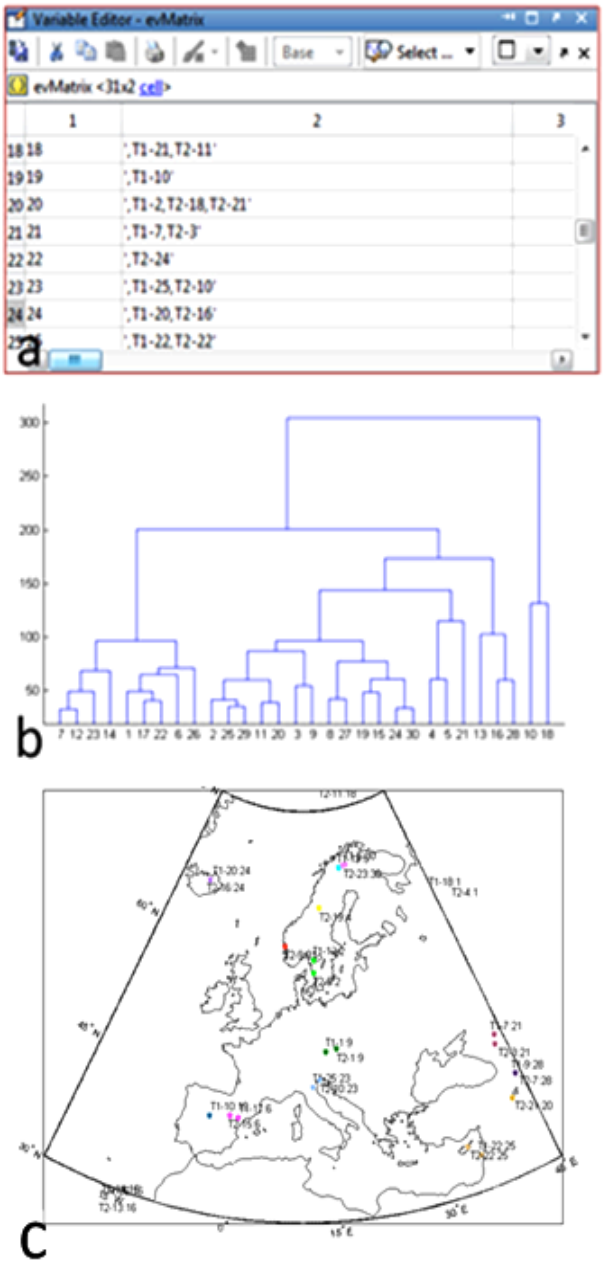


Fig. 6 Identified cluster evolution from slice 01.04. to 02.04. 1990.: cluster groups matrix (a), hierarchical tree (b) and a map of evolutionary connected clusters (c)

Agglomerative hierarchical clustering algorithm adapted in this way and executed on representatives data for two consecutive time slices gives us evolutionary connections between climate clusters.

Third step of this algorithm is representation of data to the user. Identified evolutionary steps are represented in numerical values and graphically.

Since we are mining data that has not been processed in this way by any other researchers, there was not a reliable result or resource to which we could compare our results. Providing both graphical and numerical representation give us means to verify the result by their human inspection (in which case graphical representation is quicker and easier to inspect), as well as make them easier to process by computer (for which numerical representation is more adequate).

In results T1-X means that the cluster was formed in time slice T1 and in that time slice it had ID X. Numerical representation is given as a matrix where representatives of evolutionary connected clusters from two time slices are grouped together (figure 6.a) and graphical representation of algorithm results is hierarchical tree (figure 6.b) and map on which evolutionary connected clusters are pointed out (figure 6.c).

In figure 6.c marking T1-X:Y means that cluster X from time slice T1 is in evolution group Y and if cluster Z from time slice T2 in evolutionary group Y marked as T2-Z:Y exists that means that these two clusters are evolutionary connected (cluster Z evolved from cluster X). Different clusters are displayed in different colour.

5 Evaluation of developed solution and future work

Implemented algorithms are one way to solve defined problems. It is possible to use other clustering methods, set the algorithm parameters differently or even use other data mining techniques to get the same information about the state of the atmosphere on a daily basis and about change of that state from one day to another. Further work would be comparison of quality of the results obtained using different methods and techniques for solving the same problem.

Created algorithm for climate cluster evolution identification is used in order to discover evolutionary connections in measurements from consecutive time slices. If this same algorithm with parameters defined for daily data analysis was used to discover evolutionary connections in data which does not satisfy our condition of measurements

being from consecutive time slices, it would give us the results which are incorrect in the area of research, even though they are correct from algorithmic point of view. E.g. if we provide as an input to the algorithm measurements from April and September which are similar enough, algorithm would provide us the information that those two conditions are connected on a daily basis - that would not be correct from climatological point of view. However, modification of parameters of agglomerative hierarchical clustering algorithm could be performed in order to perform same type of analysis on data of different temporal and spatial scale.

Implemented algorithm identifies climate cluster evolution from one to another time slice, but a modification is possible in order to discover cluster evolution from time slice T_X to slice T_{X+n}, where $n > 1$: all that would be necessary is to execute created algorithm for pairs of time slices and chain the results.

If we would provide information about type of change that cluster passed through results of algorithm would be of better quality. If we compare values of the attributes of cluster T1-X representative from time slice T1 which evolved to T2-X in time slice T2 we could identify that the state of the atmosphere has changed in a way that, for example, temperature has increased, decreased or has remained unaltered. And by comparison of number of objects in cluster T1-X and in cluster T2-X we could conclude that climate cluster X has expanded or disappeared or remained almost the same.

Another solution for detection of emerging and disappearing climate conditions is by comparison of objects that are extremes from geographical point of view (objects that have minimal and maximal latitude and longitude) for clusters T1-X and T2-X: position of those objects could tell us if cluster X had spatial movement of any kind from time slice T1 to slice T2.

Identified evolutionary steps could be represented as a graph where nodes are cluster representatives in each time slice and two nodes are connected if they are one another successor / predecessor.

Another improvement would be if algorithms could process data sets that have missing values or if clustering results are refined by exploring subspaces for objects that do have valid measurements of additional attributes.

Primary purpose of the algorithm created in this paper is analysis of historical data of atmospheric conditions in order to identify climate clusters and analyse change that they go through over the time. The algorithm could be upgraded in a way that, based on historical data analysis and recognition of patterns of climate changes in historical data, atmospheric state prediction is performed. Created algorithm can provide input data about cluster evolution for another algorithm that would identify climate state change patterns and by comparison of those patterns with present state of the climate one could predict state of the atmosphere in the future.

Ideas mentioned here are just a few of many for further work in the area of climate data analysis using clustering data mining techniques. Of course, there are many more options for exploration and improvement of data mining methods and techniques and application of them in climate or any other data analysis.

6 Conclusion

This paper shows one of many possible ways of applying clustering data mining techniques in order to solve the problem of identification of state of the atmosphere and determine change of that state over the time. The paper gives the steps in the development of algorithm for climate cluster identification and climate cluster evolution identification, but the same process can be apply for any other similar domain of the problem:

- precisely define the problem to be solved
- explore existing solutions and algorithms
- gather necessary resources (data)
- analyse which existing algorithms are applicable in solving defined problem
- adapt existing or create the new algorithm for solving the problem

Data mining unifies techniques from different areas in order to extract information from large amount of data. Data mining process can provide different information depending on the used method. Clustering data mining methods are applied in order to discover the groups of similar objects in the data without prior knowledge of the groups and their characteristics. Clustering process can be implemented in different ways depending of the goal of the analysis and data to be analysed: difference between clustering methods is the approach to cluster formation.

This paper presents a new solution for climate data analysis using clustering data mining techniques. The way in which data have been processed is unique and the other resources to which provided results could be compared were not available at the time when this paper is published.

The presentation relays more on human verification of the graphs given above. Created algorithm detects climate clusters in the specified time slices and identifies evolution of those clusters over the time. We defined the problem to be solved, provide review of existing solutions and algorithms and analysed whether they could be used in the algorithm created in this paper. Other models such as climate prediction model that apply complex physics calculations use too many correlated parameters and are difficult to derive correct model. The paper gives full procedure of creating an algorithm for solving the defined problem. Creation of the algorithm is given step by step and selections of different mechanisms and/or parameters in that process have been explained.

Also, we presented several ideas for future work in this area, e.g. improving developed solution by implementing mechanisms to improve quality of the results. Combining created algorithm with another could solve more complex problems than the one defined here. For example, implemented algorithm could be combined with another one to identify patterns in climate change in order to predict climate changes. Implemented solution can be used to create data sets for other algorithms (e.g. creating testing data sets for neural network).

References:

[1] David P. Schneider, Clara Deser, John Fasullo and Kevin E. Trenberth, *Climate Data Guide Spurs Discovery and Understanding*, EOS, Transactions American Geophysical Union, 2013, pp. 121-122

[2] B. Hennemuth, S. Bender, K. Bulow, N. Dreier et al., *Statistical methods for the analysis of simulated and observed climate data, applied in projects and institutions dealing with climate change impact and adaptation*, CSC Report 13, Climate Service Center Germany, 2013

[3] P. Saikia, H. Tahbilder et al., *Prediction of rainfall using datamining technique over Assam*, Indian Journal of Computer Science and Engineering, 2014

[4] J. Han and M. Kamber, *Data Mining Concepts and Techniques,* 2nd edition, Morgan Kaufman, 2006

[5] *UN Department of Economic and Social Affairs, Division for Sustainable Development - Area of Climate Change and Sustainable Development*, http://www.un.org/esa/dsd/dsd_aofw_cc/cc_index.shtml

[6] Nejra Hadzimejlic, Dzenana Donko, Nijaz Hadzimejlic, Climate Data Analysis Using Clustering Data Mining Techniques, *Proceedings of the 3rd International conference on Applied Informatics and Computing Theory (AICT '12)*, WSEAS Press, 2012, pp. 96-100

[7] P. Kalnis, N. Mamoulis and S. Bakiras, On Discovering Moving Clusters in Spatio-temporal Data, *SSTD proceeding*, 2005, pp. 364-381

[8] M. Bottcher, F. Hoppner and M. Spiliopoulou, On Exploiting the Power of Time in Data Mining, *ACM SIGKDD Explorations Newsletter 10 (2),* 2008, pp. 3-11

[9] S. Günnemann, H. Kremer, C. Laufkötter and T. Seidl, Tracing Evolving Subspace Clusters in Temporal Climate Data, *Data Mining and Knowledge Discovery Journal (DMKD*), Vol. 24, Nr. 2 2012, pp.387-410)

[10] Klein Tank, A.M.G. and Coauthors, *Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment,* 2002.

[11] Project team ECA&D, Royal Netherlands Meteorological Institute KNMI, *Algorithm Theoretical Basis Document v10.4*, European Climate Assessment & Dataset project, 2011

[12] Matlab R2012a Documentation. Link: http://www.mathworks.com/help/techdoc/

[13] I.H. Witten, E. Frank and M.A. Hall, *Data Mining - Practical Machine Learning Tools and Techniques*, 3rd edition, Morgan Kaufmann, 2011

[14] C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008