How to cite: Magyari-Sáska, Z., Dombay, Ş. (2022) Experimental Method to Assess the Looseness or Compactness in Climate Changing for Several Major Cities of Hungary. *2022 "Air and Water – Components of the Environment" Conference Proceedings*, Cluj-Napoca, Romania, p. 116-127, DOI: 10.24193/AWC2022_12.

EXPERIMENTAL METHOD TO ASSESS THE LOOSENESS OR COMPACTNESS IN CLIMATE CHANGING FOR SEVERAL MAJOR CITIES OF HUNGARY

Zsolt MAGYARI-SÁSKA[1]*, Ștefan DOMBAY[1]

DOI: 10.24193/AWC2022_12

ABSTRACT. This research wants to investigate the possibility to highlight and track the looseness or compactness of climate change using a network model based on long-term weather changes for five major cities in Hungary, for which an important daily scale dataset is available starting from 1901 to 2020. Climate study based on network models is a novel approach, it does not have a well-developed research, methodological and literature background. Even if several network models can be developed, in that used in the present study the edges of a network connect the same periods of different years based on the greatest aggregate weather similarity. One of the results of this research was actually the development of the model itself using functions developed in the R CRAN system. In this case of large-scale data processing, it was very important to use programming methods that could do all this in a time-efficient manner. Using three different study intervals there were results that converged, but there were also results for which in the larger study period disappeared or equalized the results obtained for the shorter periods. The most pronounced looseness occurs in in the late autumn and winter months and in early spring periods. In addition to all these main trends, high looseness values can be observed for some settlements in the summer as well as in the late spring or early autumn periods, while summer typically appears as a more compact period.

Keywords: climate, network model, data aggregation, R CRAN, Hungary

1. INTRODUCTION

The scientific literature regarding climate change is very extensive as its consequences affects more and more people. However, these felt effects are locally differentiated which suggests that it should be analyzed and assessed according to local specificities. (Meresa et al., 2017).

For climate study there are many models and methods but one of the most important possibilities, that of network models (Barabási 2002, 2013) can hardly be found in the literature for climate studies. Network models are used in many different domains (Li and Maini, 2005; Light et al, 2005; Borgs et al., 2007; Hopkins, 2007;

[1] Babeş-Bolyai University Cluj, 535500 Gheorgheni, Str. Grădina Csiky, nr. 53; zsolt.magyari@ubbcluj.ro, stefan.dombay@ubbcluj.ro

Emmert-Streib et al., 2018; Arquilla and Ronfeldt, 2001) and might have be successful in weather analysis. Since the weather that generates climate implies a succession of dates, and these periods, large or small, form a temporal chain, it seemed logical to ask how far the spatial similarity expressed in Moran's first law (Moran, 1950) could be applied to time (events closer in time are more similar than events further away in time). It was soon realized that this was not always true; synthesized data for the weather of a later day, week or month might be more similar to a past period of similar duration than to the weather of a day, week or month immediately before it.

The existence of the connection between periods that may be distant in time is a real situation, the expression, representation and study of which may prove useful as it indicates the dynamics of weather change.

The aim of this research was to develop a new network model based on which the dynamic of climate change can be characterized. The terms of looseness or compactness wants to express whether the climate evolution is much similar to a flowing river different time periods are related to each other based on their weather similarity, or the climate is compact having multiple separate time periods which aren't connected to each other. The new network model was applied for five large cities in Hungary. At the city level, there are only a few studies either for Hungary or for other countries (Probáld, 2014; Stone, 2012) that include climate studies. However, climate change at the city level is important because cities are both generators and victims (Bulkeley, 2012; Hunt and Watkins, 2011). With the availability of daily meteorological datasets for five major cities in Hungary (Budapest, Debrecen, Pécs, Szeged and Szombathely) for the period 1901-2020 in 2021, a dataset was available that seemed appropriate, both in terms of time period and detail, to carry out the climate studies that were set out.

2. DATA AND METHODS

As mentioned in the introduction the analyzed daily dataset was freely available on the site of Hungary's National Meteorological Service for all the five stations covering the period between January 1, 1901 and December 31, 2020.

2.1. Homogeneity of data

In all data processing it is important to ensure homogenization of the data (reference). As measurements at different municipalities have changed a lot over the past 120 years, both in terms of location, measurement instruments and even methodology the homogeneity of data has to be ensured.

Starting from metadata a temporal representation of the changes was constructed, changes that may have contributed to the inhomogeneities for each of the five municipalities (Fig. 1). Changes in location, in measuring instruments, changes in measuring methodology or data reconstruction based on data from other stations could all have contributed to the non-homogeneous data series. In figure 1 different colors indicate a new measurement site or a change of measurement instrument. The

time bands in black represent lack of measurements, the values in these periods are the result of posterior data reconstruction.

Knowing that the data sets cover a period spanning two world wars, it was necessary to map these metadata over time. It is noticeable that while the First World War only caused outages in the data collection at Pécs, the Second World War interrupted the measurement processes for shorter or longer periods for the other three stations, with the exception of Szombathely and Budapest.

The homogeneity of the data was assured by Hungary's National Meteorological Service for the daily air temperature and precipitation amount, data which were used in the current research.

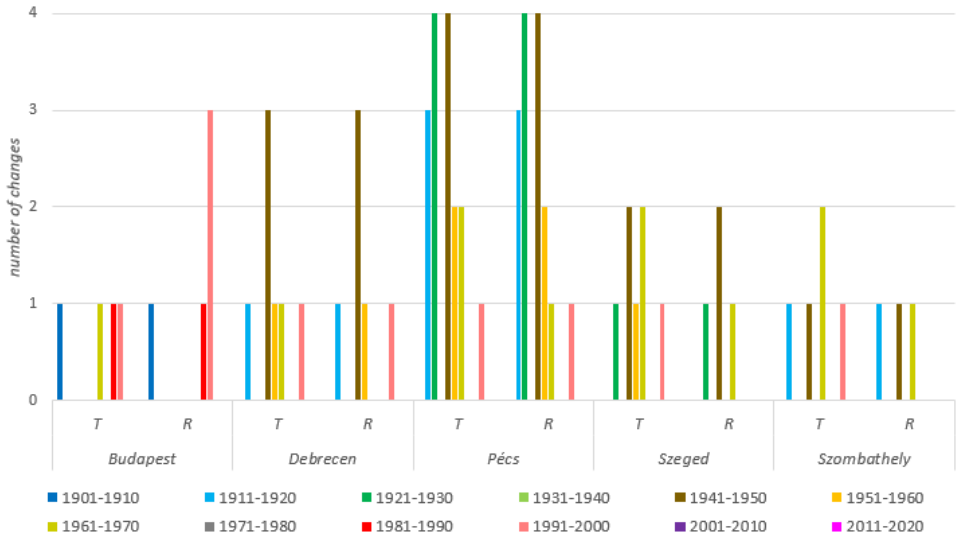


Fig. 1. ***Number of changes (location/ instrument) in data measuring for every decade***
T – (daily mean, maximum and minimum temperatures), R – daily preceipitation amount

2.2. Data aggregation

Starting from homogenized data derived and aggregated datasets were created to characterize and at the same time amplify the elements of weather that allows to track changes. Using and R CRAN script a total number of 32 values were derived and aggregated for different time periods to characterize the climate.

Three aggregation interval was defined: a seasonal breakdown, a monthly breakdown and the half-month breakdown. The idea of seasonal aggregation was given by the fact that it has often been heard in recent times that winter or summer used to be different in the past then nowadays. The seasons appear with their meteorological time limits in the research. Aggregation at months level it may be useful because it defines a time interval that describes perhaps in the best way our time scale for characterizing the weather within a year. Findings about the weather for a given month are often used. Using the half-month time scale may reveal some

changes which occurs on a shorter time scale, changes that might be lost over a larger time interval.

2.3. The network model

The construction rules of the network data model are the key element in the evaluation of looseness or compactness of weather. Each aggregation period will represent a node in the network, while these nodes will be connected or not based on a weather similarity index discussed below.

2.3.1. Similarity index

To have a way for deciding whether two nodes should be connected or no a similarity index had to be defined expressing the higher or lower similarity for the different aggregation periods.

Since the values of each derived weather characteristic is presented on a different scale and orders of magnitude, it was necessary to normalize these values (formula 1). As their distribution does not necessarily follow a normal distribution, standardization didn't represent a possibility.

$$n_i = \frac{\min(v) - v_i}{\max(v) - \min(v)} \quad [1]$$

where,

n_i – normalized value

v_i – original value

min(v), max(v) – minimum and maximum of original data series

A similarity index can be define concentrating into a single value all the characterizing components, which would have a significant drawback the way deciding the importance, the weight of each component, or using other possibilities. There are several methods to determine the similarity between two value vectors. One of the best known is correlation, but also the Jaccard index, multidimensional Euclidean, Minkowski or Hamming distance, or cosine similarity (Tan et al., 2005).

In this research the latter was used since this value is often used in data mining (Han et al., 2012) and shows how the values of the two vectors have the same orientation, representing the same direction in evolution. The cosine similarity index is the cosine of the differences between the vectors defined in a multidimensional (number of dimensions equals the number of features in the vectors) space (formula 2).

$$sim(v, w) = \frac{v \cdot w}{\|v\| \times \|w\|} = \frac{\sum_{i=1}^{n} v_i \times w_i}{\sqrt{\sum_{i=1}^{n} v_i^2} \times \sqrt{\sum_{i=1}^{n} w_i^2}} \quad [2]$$

where,

v and w – vectors holding the different derived characteristics of an aggregation period

v_i *and* w_i – individual values of the vectors

The R CRAN function calculating the similarity values between different nodes, based on the data table passed as a parameter, which contains the meteorological characteristics of the different aggregation periods in rows, returns the similarity matrix whose (i,j) element indicates how similar aggregation period i is to

aggregation period j is presented in figure 2. Higher value indicates a higher similarity.

Parameters of the function:

- *df* - the data table

- *idcols* - the number of initial columns of the data table containing the identifiers of the aggregation period (e.g. year, month, etc.) which are not used in the determination of the similarity index. Default value for the dataset used in this research is 4.

- *norm* - provides the possibility to normalize data. It is **on** by default, but since the similarity matrix can be generated with other methods than cosine similarity it can be switched **off**.

- *SIM* - function type parameter, name of the function that is used to determine the pairwise similarity index. In this research this is the cosine similarity but other functions can be used

- *simetric* - can be used to control whether the similarity matrix has to be filled below the main diagonal or not. The similarity matrix is symmetric by nature but when choosing a period with the highest similarity to another aggregation period, and if we wish to take into account chronology of the aggregation periods it's preferable not to fill in the elements below the main diagonal, i.e. not to symmetrize the matrix.

```
create_similarity_matrix<-
function(df,idcols=4,norm=TRUE,SIM=cosine_similarity,simetric=TRUE)
{
  norm_df<-df
  factno<-length(df[1,])
  if (norm) {
   for(i in (idcols+1):factno)
     norm_df[,i]<-normalize(df[,i])
  }
  norm_df[is.na(norm_df)] <- 0
  recno<-length(df[,1])
  sim_mat<-matrix(rep(0,recno^2),nrow=recno,ncol=recno,byrow=TRUE)
  den <- c()
  for(i in 1:recno)
    den <- c(den,sqrt(sum(norm_df[i,(idcols+1):factno]^2)))
  for(i in 1:(recno-1)) {
    for(j in (i+1):recno) {
      sim_mat[i,j] <- SIM(norm_df[i,(idcols+1):factno],norm_df[j,
                      (idcols+1):factno],den[i],den[j])
      if (simetric)
        sim_mat[j,i] <- sim_mat[i,j]
    }
  }
  return (sim_mat)
}
```

Fig. 2. ***R CRAN function to create the highest similarity matrix***

2.3.2. Network creation

In the network model each aggregation period is a node that can be connected to any other whether it appears before or after it in time. The only criterion is the highest

similarity, each node will be connected to the one node it is most similar to. There may be situations where, for example considering nodes A and B, that A is most similar to B and B is most similar to A and not to any other node; in this case they will form a closed system. In fact, this is also the role of this model, to reveal the compactness or looseness of the weather.

The R CRAN function creating this network model is presented in figure 3, having the following parameters:

- *path* - path to the input data file folder
- *filename* - name of the input data file containing the aggregated dataset
- *base* - the base year after which the data series start; its role is only for possible data filtering, which can be controlled by the following two parameters
- *yearS* - first year of the study interval (relative offset from the base year)
- *yearF* - last year of the study interval (relative offset from the base year)

The result is a file in gml format, which keeps the name of the input file, but adds the ***_M2_y1-y2*** part, where ***y1*** and ***y2*** are the sequence numbers of the first and last year of the analyzed period.

```
create_similarity_network<-function(path,filename,base=1900,yearS=1,yearF=120)
{
  dataSet<-read.csv(paste(path,"//",filename,".csv",sep=""))
  dataSet<-dataSet[dataSet$ye>=base+yearS & dataSet$ye<=base+yearF,]
  simMatrix <- create_similarity_matrix(dataSet,simetric=FALSE)
  len <- length(dataSet[,1])
  g <- make_empty_graph()
  g <- add.vertices(g,len,
          attr=list("year"=dataSet$ye,"month"=dataSet$mo,"part"=dataSet$part))
  for(i in 1:(len-1))
    for(j in (i+1):len)
      simMatrix[j,i] <- simMatrix[i,j]
  for(i in 1:len) {
    y <- which(simMatrix[,i]==max(simMatrix[,i]))
    g <- add.edges(g,c(i,y),
               attr=list("weight"=max(simMatrix[,i]),"from"=i,"to"=y))
  }
  write_graph(g,paste(paste(path,"\\",filename,"_M2_",
              yearS,"-",yearF,".gml",sep="")),format="gml")
  }
```

Fig. 3. ***R CRAN function creating the network model***

The network model was created for the five municipalities using the presented R functions (Ihaka and Gentlman, 1996). In t addition to the basic R system, the igraph package (Csárdi and Nepus, 2006) was used to perform all the network-related operations, such as network creation, adding edges, determining shortest paths, graph weighting, etc. Later, for network visualization the Cytoscape software (Shannon et al., 2003), was used which in addition to visualization also provides some basic network analysis facilities.

3. RESULTS AND DISCUSSION

Based on the R CRAN script explained in the previous chapter the network models were created for all five cities at seasonal, monthly and semi-month time resolution.

As it's observable from figure 4 showing the seasonal scale representation for Szombathely in the network appear has many components. The coloring is according to seasons, blue is for winter, green for spring, red for summer and yellow for autumn. The size of the node is related to time distance from the beginning of study, recent years having higher diameter.

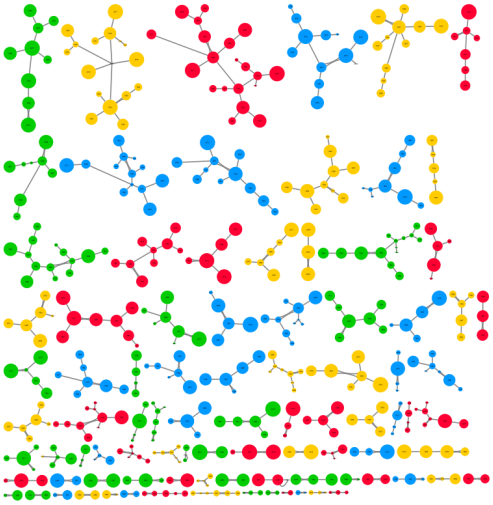


Fig. 4. ***Season level network model for Szombathely***

The more components the network consists of, the more fragmented or loose the weather is in the given settlement.

At season level network only in a few cases appear components in which nodes belonging to different seasons are mixed. Using a smaller time scale the mixing phenomenon is amplified and the components can no longer be separated to belong to a single month or half month. Because of this, separate networks were created for each studied time level. In these cases, the following indicators were used to characterize the compactness or looseness of the weather:

- the number of components in the network
- the standard deviation of network components' size
- the variability index of the year values for each component, defined as the ratio between the standard deviation of the year values and their average value.

Because higher values for all three indicators show a higher looseness for a given period, combining them with multiplication the operation amplifies the result and the higher the final value is the more loose the period is.

At monthly scale the vertical axis of figure 5 shows the product of the above-mentioned characteristic. The loosest weather appears for winter at Budapest, autumn at Debrecen, Pécs and Szeged, while winter seems to be the most loose in Szombathely. On the other hand, summer is the most compact season at Budapest, spring at Debrecen, Pécs and Szeged, and autumn at Szombathely. In some cases, the differences between the values that decide first and last place are very close. Winter stands out at Budapest, but the other three seasons have almost the same values. At Debrecen, the indicator values for spring and summer are almost the same, as in the case of Pécs. At Szeged, with the exception of spring, all three other seasons show almost the same high values. In the case of Szombathely, summer and winter show a significantly larger looseness than spring or autumn.

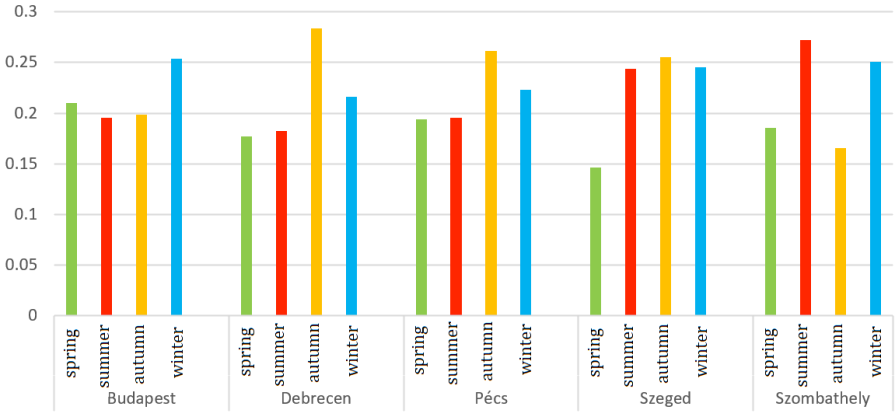


Fig. 5. ***Instability (loose weather) indicator at season level***

At monthly level (fig. 6) the highest values representing the looseness are for January and May at Debrecen, February and December at Szeged, July at Pécs and February at Szombathely. The lowest values representing the compactness are appears for February and March at Budapest and for May at Pécs. It is also noticeable that there are months that manifest themselves in a similar way in different locations based on this complex indicator. Such a month is October and in a lesser way April, August and September, while for other months different things are experienced at different locations. Significant differences can be seen for January, February, but May and July manifest themselves differently in some locations.

In case on the half-month level analysis the three defined indicators were plotted separately (fig. 7) in order to examine the locations and months where the two half-months differ significantly in terms of that indicator. This also makes it possible to better follow the dynamics of successive periods. In case of Budapest between the instability of the second half of January and the last half of February includes a compact, stable manifestation for the second half of January and the first half of February. Similarly, we can see that the half-months of March at Pécs or Szeged, shows up differently. The looseness of the first half-month is much more pronounced than in the case of the second half-month.

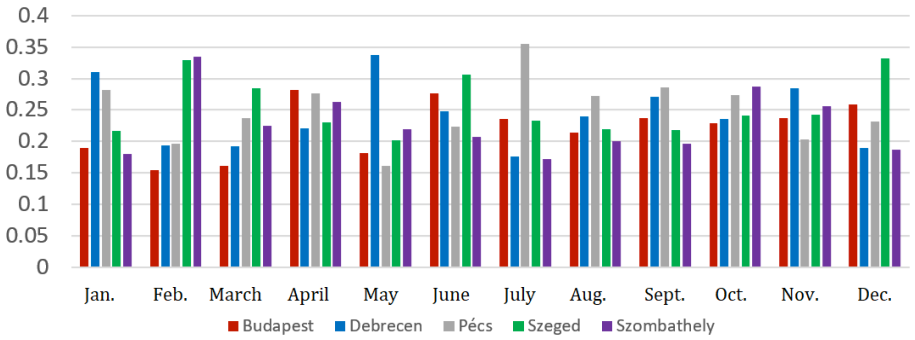
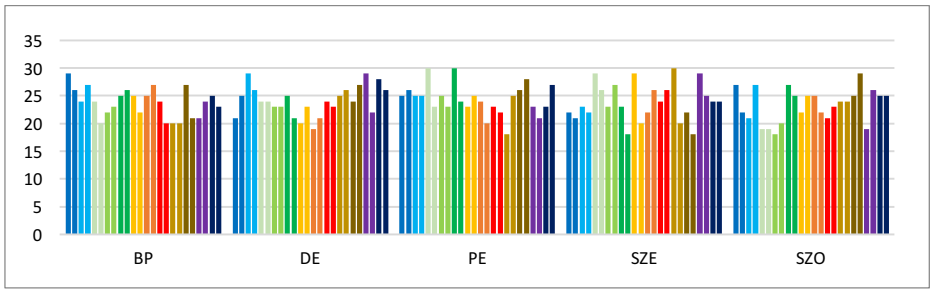
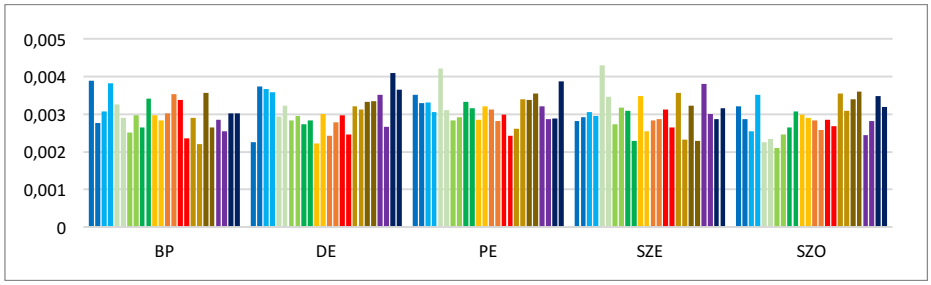


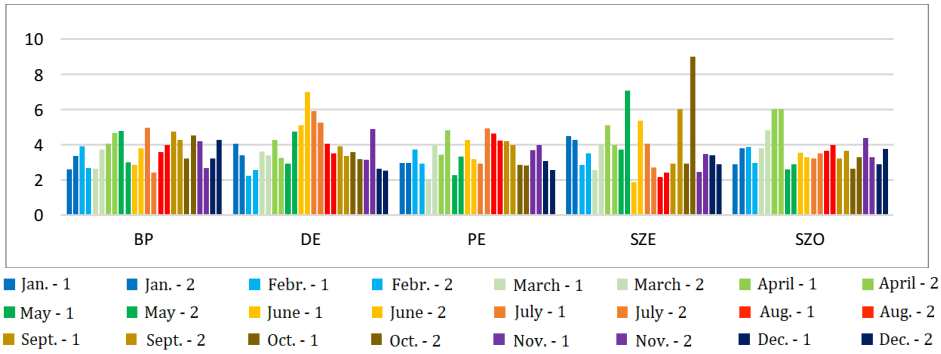
Fig. 6. ***Instability (loose weather) indicator at monthly level***



a) number of components



b) variability index



c) standard deviation of network components' size

Fig. 7. ***Individual indicators of monthly level network***

It can also be observed that for this analysis level the number of components of the networks and the average number of nodes participating in the components correlate much better with each other than for longer aggregation periods. We can hardly find a situation where these two indicators determined from completely different data sources would not show the same relationship, or if it occurs (ex. Pécs in October or Szombathely in June) the differences are not significant.

It can be observed that in case of Budapest, Debrecen and Szombathely the winter half-months are most compact at the beginning of the year, while in the case of Pécs and Szeged the early spring months, the first half of March shows outstanding looseness. In these settlements March is the month in which the first and second half-months differ the most. If we look at these differences within a month, October shows the biggest differences in case of Budapest, November in case of Debrecen and February in case of Szombathely.

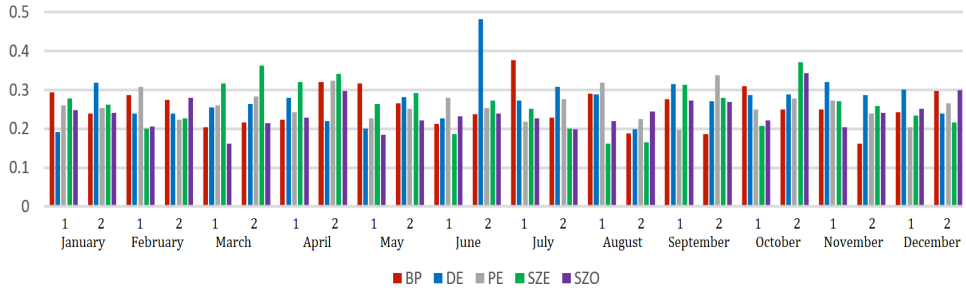


Fig. 8. ***Loose weather indicator at half-month level***

At the end of July and the beginning of August in case of Debrecen, considering the standard deviation of the components compared to the number of components (which is not very high), we can see that some large components appear, which means the compactness of the given period. Similar situations occur in the second half of May at Szeged and in April at Szombathely. The real remarkable value, however, appears at Szeged in the second half of October. The high standard deviation and the low component number indicates that there are one or just a few huge components and several small components with about the same number of nodes. In any case, a strong compactness characterizes this period.

Based on the aggregated indicator (fig. 8.) which includes all three indicators presented above, four outstanding values can be observed. These are the case of Debrecen in the second half of June, then the first half of July at Budapest and in case of Szeged, two outstanding values, which refer to the second half of March and October. These are the half-months that appear to be most loose. If we look at the settlements, the second half of April and the beginning of May are loose at Budapest, outside the first half of July already mentioned. In case of Debrecen the second half of January and July and the first half of September and November show looseness. Pécs shows the highest value in the second half of September, but the first half of

February and August and the second half of April are also worth mentioning. At Szeged both halves of March and April and the second half of October are the loosest periods, while in case of Szombathely the only outstanding value occurs in the second half of October.

4. CONCLUSION

Climate study based on network models is a novel approach, it does not have a well-developed research, methodological and literature background. Based on own research experience so far, different network data models provide different opportunities for data mining, which suggests that it is worth doing more research on different network data models. In this model the edges of a network connect the same periods of different years based on the greatest aggregate weather similarity.

The preparatory operations required to carry out the research required the handling of a considerable amount of data, which would not have been possible with existing software, as there is no software to do all this on a click. Therefore, one of the results of this research was actually the development of the model itself using functions developed in the R CRAN system. In this case of large-scale data processing, it was very important to use programming methods that could do all this in a time-efficient manner. In this direction a number of optimizations were made through the research, as a total of more than 100 network data models had to be created and analyzed for the five settlements at the level of seasons, months and half-months.

Using three different study intervals there were results that converged, but there were also results for which in the larger study period disappeared or equalized the results obtained for the shorter periods. What can be concluded as a general appreciation for the five studied settlements is that changes considering the compactness or looseness of weather can be detected based on network data model. The most pronounced looseness occurs in in the late autumn and winter months and in early spring periods. In addition to all these main trends, high looseness values can be observed for some settlements in the summer as well as in the late spring or early autumn periods, while summer typically appears as a more compact period.

The present research was based on four weather characteristics covering which have covered the entire study period. By expanding them (eg humidity, number of hours of sunshine), the similarity indicator will be able to show a more precisely the individual characteristics of shorter aggregation periods (months, half-months) even more accurately making the changes more easier to recognize. In this case, even the aggregation periods could be reduced and new indicators could appear in the analyzes process, for example, the internal homogeneity of each month could be examined, which in this case could only be done at season level.

Acknowledgment. The presented research was supported by the DOMUS scholarship program of the Hungarian Academy of Sciences

REFERENCES

1. Arquilla and D. Ronfeldt. *Networks and Netwars: The Future of Terror, Crime, and Militancy*. RAND: Santa Monica, CA, 2001.
2. Barabási, A.L. (2013), *Behálózva*, Helikon, ISBN 9789632272580
3. Barabási, A.L. (2002). *Linked: The New Science of Networks*. Perseus Books Group. ISBN 9780738206677.
4. Borgs, C., Chayes, J., Daskalikis, C., Roch, S. (2007), First to market is not everything: an analysis of preferential attachment with fitness, *STOC'07 Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, 135-144, DOI: 10.1145/1250790.1250812
5. Bulkeley, H. (2012). *Cities and Climate Change*. Routledge. DOI: 10.4324/9780203077207
6. Csardi G, Nepusz T (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems, 1695. https://igraph.org
7. Emmert-Streib F., Tripathi S., Yli-Harja O., Dehmer M. (2018), Understanding the World Economy in Terms of Networks: A Survey of Data-Based Network Science Approaches on Economic Networks, *Frontiers in Applied Mathematics and Statistics*, 4, DOI: 10.3389/fams.2018.00037
8. Hopkins, L. (2007), Network Pharmacology. *Nature Biotechnology*, 25: 1110-1111, 2007.
9. Han. J., Kamber, M., Pei, J. (2012), *Data Mining: Concepts and Techniques*, Elsevier, DOI: 10.1016/C2009-0-61819-5
10. Hunt, A., Watkiss, P. Climate change impacts and adaptation in cities: a review of the literature. *Climatic Change* 104, 13–49 (2011). DOI: 10.1007/s10584-010-9975-6
11. Ihaka, R., Gentlman, R. (1996), R: A Language for Data Analysis and Graphics, *Journal of Computational and Graphical Statistics*. 5 (3): 299–314. DOI:10.2307/1390807.
12. Li, C., Maini, P.K. (2005), An evolving network model with community structure, *Journal of Physics: a mathematical and general*, 38 (45), 9741-9749 DOI: 10.1088/0305-4470/38/45/002
13. Light, S., Kraulis, P., Elofsson, A. (2005), Preferential attachment in the evolution of metabolic networks, *BMS Genomics*, 6,159, DOI: 10.1186/1471-2164-6-159;
14. Meresa, H.K., Romanowicz, R.J. & Napiorkowski (2017), Understanding changes and trends in projected hydroclimatic indices in selected Norwegian and Polish catchments, *J.J. Acta Geophys.* (2017) 65: 829. https://doi.org/10.1007/s11600-017-0062-5
15. Probáld F. (2014), The urban climate of Budapest: past, present and future, *Hungarian Geographical Bulletin* 63 (1) (2014) 69–79. DOI: 10.15201/hungeobull.63.1.6
16. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–2504.
17. Stone, B. (2012) The City and the Coming Climate. *Climate Change in the Cities we Live*. Cambridge, Cambridge Univ. Press, 198 p.
18. Tan PN, Steinbach M, Kumar V (2005). *Introduction to Data Mining*. ISBN 0-321-32136-7