ISSN: 2148-9173 **Vol: 9 Issue:3 September 2022**

International Journal of Environment and Geoinformatics (IJEGEO) is an international, multidisciplinary, peer reviewed, open access journal.

Comparison of Model-Based Clustering to Other Clustering Methods: An Example on Meteorological Data

Selim DÖNMEZ

Chief in Editor

Prof. Dr. Cem Gazioğlu

Co-Editors Prof. Dr. Dursun Zafer Şeker, Prof. Dr. Şinasi Kaya,

Prof. Dr. Ayşegül Tanık and Assist. Prof. Dr. Volkan Demir

Editorial Committee (September 2022)

Assoc. Prof. Dr. Abdullah Aksu (TR), Assoc. Prof. Dr. Uğur Algancı (TR), Prof. Dr. Levent Bat (TR), Prof. Dr. Paul Bates (UK), İrşad Bayırhan (TR), Prof. Dr. Bülent Bayram (TR), Prof. Dr. Luis M. Botana (ES), Prof. Dr. Nuray Çağlar (TR), Prof. Dr. Sukanta Dash (IN), Dr. Soofia T. Elias (UK), Prof. Dr. A. Evren Erginal (TR), Assoc. Prof. Dr. Cüneyt Erenoğlu (TR), Dr. Dieter Fritsch (DE), Prof. Dr. Manik Kalubarme (IN), Dr. Hakan Kaya (TR), Assist. Prof. Dr. Serkan Kükrer (TR), Assoc. Prof. Dr. Maged Marghany (MY), Prof. Dr. Micheal Meadows (ZA), Prof. Dr. Masafumi Nakagawa (JP), Prof. Dr. Burcu Özsoy, Prof. Dr. Hasan Özdemir (TR), Prof. Dr. Chyssy Potsiou (GR), Prof. Dr. Erol Sarı (TR), Prof. Dr. Maria Paradiso (IT), Prof. Dr. Petros Patias (GR), Prof. Dr. Barış Salihoğlu (TR), Assist. Prof. Dr. Başak Savun-Hekimoğlu (TR), Prof. Dr. Elif Sertel, (TR), Prof. Dr. Füsun Balık Şanlı (TR), Dr. Duygu Ülker (TR), Prof. Dr. Seyfettin Taş (TR), Assoc. Prof. Dr. Ömer Suat Taşkın (TR), Assist. Prof. Dr. Tuba Ünsal (TR), Assist. Prof. Dr. Sibel Zeki (TR)

Abstracting and Indexing: TR DIZIN, DOAJ, Index Copernicus, OAJI, Scientific Indexing Services, International Scientific Indexing, Journal Factor, Google Scholar, Ulrich's Periodicals Directory, WorldCat, DRJI, ResearchBib, SOBIAD

International Journal of Environment and Geoinformatics 9(3):187-191 (2022)

Research Article

Comparison of Model-Based Clustering to Other Clustering Methods: An Example on Meteorological Data

Selim Dönmez

Eskişehir Osmangazi University, Faculty of Arts and Sciences, Department of Statistics, Statistical Information Systems

E-mail: sdonmez@ogu.edu.tr

Received 24.03.2022
Accepted 19.06.2022

How to cite: Dönmez, S. (2022). Comparison of Model-Based Clustering to Other Clustering Methods: An Example on Meteorological Data, *International Journal of Environment and Geoinformatics (IJEGEO)*, 9(3): 187-191. doi. 10.30897/ijegeo.1092672

Abstract

Using a methodology comprised of model-based clustering and panel data analysis, we have tried to draw conclusions from a real-life meteorological data based on philosophy of science. In this study, we used model-based clustering on a real-life meteorological data consisting of yearly observations of annual mean temperatures gathered from the 58 stations in regions of Turkey and compared the results to the results of other agglomerative clustering methods that were derived upon the results of an earlier study on the aforementioned real-life meteorological data. We then configured the clusters as separate dummy variables the effects of which was put to the test by longitudinal data analysis. The agglomerative clustering results were more successful according to the between R^2 obtained from longitudinal data analysis compared to the earlier study but the best results were obtained from the model-based clustering of the aforementioned real-life meteorological data. The comparative clustering analysis demonstrated that the climate change occurs differently across regions of Turkey.

Keywords: Longitudinal Data, Model-Based Clustering, Climate Change, Philosophy of Science

Introduction

Climate Change is one of the hot topics in the scientific community and it has been the topic of approximately 52000 articles by the end of 2021. Even though there are a lot of articles based on climate change, the lack of longitudinal studies providing empirical results appears to be a clear problem in scientific literature. When we searched for articles in web of science database including topics such as "Climate Change", "longitudinal" and discarded any article based on the inclusion of keywords: "Questionaire" and "Survey", we ended up with 271 results. The number of articles we ended up when searched in web of science database without excluding the keywords: "Questionaire" and "Survey" were 1214. When we used the keyword "panel" instead of "longitudinal", the following number of results was 1187 and 1293, respectively. These numbers correspond to %58 of articles being based on the climate change, surveys and questionnaires and the actual number of longitudinal studies providing information on public opinion of climate change may be much higher. This can be explained by the state of the global politics and the media coverage of the environmental disasters but longitudinal studies focusing on the climate change's effects on human life seem to have lowered the interest on providing information on understanding the natural phenomenon. Using the state of the art longitudinal data clustering methods, it is possible to describe and detect the variability of the temperatural changes in regions. The purpose of this study is to provide information of climate change effects on different regions, as in the local examples leading to the civil war in Syria (Ülker et al., 2018).

There are a few exemplary studies that have used longitudinal data on climate change. For example, Bhaumik and Sengupta (2020) studied the performance of the Nadaraya-Watson estimator of the mean function from a pair of paleo climactic functional data. Their study established that registering one data set with respect to the other helps the Nadaraya-Watson estimator to be consistent under a few additional conditions and improve estimation despite error in registration. Another such study is an article for fire science. Dexen et al. (2019), jointly modelled duration and area burned, in terms of days and hectares respectively, from ground attack to final control of a fire as a bivariate survival outcome using both copula model and a joint modelling framework that connects the two outcomes with a common random effect. The study was performed in order to compare the two methodologies with respect to their utilities and predictive power. Our study provides an insight to longitudinal data clustering and analysis as well as a critical approach to cluster analysis. In our study, we described our methodology in the section named Methodology and Models. We demonstrated our results of methodology in the section of Results and lastly, we provided insights in the section of Conclusion and Some Remarks.

Methodology and Models

The methodology we used is comprised of panel data analysis and panel data clustering. The clustering method

we used is based on McNicholas (2017). McNicholas (2017) developed a clustering method that can be used on time series. The method was based on Cholesky decomposition. Using Cholesky decompositon method on the matrix notation of a autoregressive time series formula, one can obtain the following formula: Where φ_{ts} is the value of T in uth row and sth column, ε_u is normally distributed with 0 mean and 1 standard deviation and d_t is the diagonal value of D in uth row. Assuming different conditions on T_g and D_g, one can derive various iteration algorithms for the clustering method which can be tabulated in Table 1.

$$\hat{X}_u = \mu_u + \sum_{s=1}^{t-1}(-\varphi_{us})(X_s - \mu_s) + \sqrt{d_u}\varepsilon_u$$

Table 1: The attributes of T_g and D_g that correspond to the clustering types in McNicholas (2017)

Model	T_g	D_g
EEA	equal	equal & anisotropic
VVA	variable	variable & anisotropic
VEA	variable	equal & anisotropic
EVA	equal	variable & anisotropic
VVİ	variable	variable & isotropic
VEİ	variable	equal & isotropic
EVİ	equal	variable & isotropic
EEİ	equal	equal & isotropic

The clustering method of McNicholas (2017) uses an iterative algorithm based on mixture models. The mixture models can be based on Student's t distribution and Gaussian distrbution and the algorithm has 16 settings which we determine including the mixture model. The more detailed explanation of these technicalities can be found by researching the longclust package.

Table 2: The original clusters that are gathered from Celebioğlu (2018)

cluster id	Station Names included in cluster
1	Adıyaman, Afyonkarahisar, Bartın, Bayburt, Bolu, Gaziantep, Iğdır, Isparta, Kırklareli, Malatya, Muş, Ordu, Sivas
2	Aksaray, Antakya, Ardahan, Bilecik, Bingöl, Erzincan, Kayseri, Kilis, Rize, Siirt, Uşak, Yozgat
3	Adana, Amasya, Burdur, Çankırı, Düzce, Elazığ, Gümüşhane, Karaman, Kütahya, Mardin, Yalova
4	Ankara, Artvin, Bursa, Edirne, Kastamonu, Muğla, Samsun, Sinop
5	Aydın, Denizli, Giresun, Kocaeli
6	Niğde, Şanlıurfa, Van
7	Kırşehir, Nevşehir
8	Tokat
9	Tunceli
10	Kahramanmaraş
11	Mersin
12	İzmir

Table 3: The 3-cluster result from the agglomerative clustering based on original clusters

Cluster id	Station Names Included in Cluster
1	Adana, Adıyaman, Afyonkarahisar, Aksaray, Amasya, Antakya, Ardahan, Bartın, Bayburt, Bilecik, Bingöl, Bolu, Burdur, Çankırı, Düzce, Elazığ, Erzincan, Gaziantep, Gümüşhane, Iğdır Isparta, Karaman, Kayseri, Kilis, Kırklareli, Kırşehir, Kütahya, Malatya, Mardin, Mersin, Muş, Nevşehir, Niğde, Ordu, Rize, Siirt, Şanlıurfa, Sivas, Tokat, Tunceli, Uşak, Van, Yalova, Yozgat
2	Ankara, Artvin, Aydın, Bursa, Denizli, Edirne, Giresun, İzmir, Kahramanmaraş, Kastamonu, Kocaeli, Muğla, Samsun, Sinop

Results

Our data consists of the annual mean temperatures gathered from 58 meteorological stations between 1965-2012. The data was discretized as 0 representing the nonincreasingness of the continuous data and 1 representing the opposite of it. After this preprocessing, the data was transformed into an aggregate of Markov chains and was subjected to cluster analysis based on the frequency of states of discretized data. Celebioğlu (2018) had obtained the following results in Table 2.

The clusters were later merged using the agglomerative clustering methods based on Euclidean distance function and the steady-state distribution was created based on the values of discretized data. The results can be seen in Table 3 and Table 4:

Table 4: The 2-cluster result from the agglomerative clustering based on original clusters

Cluster id	Station Names Included in Cluster
1	Adana, Adıyaman, Afyonkarahisar, Aksaray, Amasya, Antakya, Ardahan, Bartın, Bayburt, Bilecik, Bingöl, Bolu, Burdur, Çankırı, Düzce, Elazığ, Erzincan, Gaziantep, Gümüşhane, Iğdır Isparta, Karaman, Kayseri, Kilis, Kırklareli, Kırşehir, Kütahya, Malatya, Mardin, Mersin, Muş, Nevşehir, Niğde, Ordu, Rize, Şanlıurfa, Siirt, Sivas, Tokat, Tunceli, Uşak, Van, Yalova, Yozgat
2	Ankara, Artvin, Bursa, Edirne, Kahramanmaraş, Kastamonu, Muğla, Samsun, Sinop
3	Aydin, Denizli, Giresun, İzmir, Kocaeli

Table 5: The 2-cluster result from the model-based clustering of the data

Cluster id	Station Names Included in Cluster
1	Adana, Adıyaman, Afyonkarahisar, Aksaray, Amasya, Antakya, Ardahan, Bartın, Bayburt, Bilecik, Bingöl, Bolu, Burdur, Çankırı, Düzce, Elazığ, Erzincan, Gaziantep, Gümüşhane, Iğdır, Isparta, Karaman, Kayseri, Kırklareli, Kırşehir, Kilis, Kütahya, Malatya, Mardin, Mersin, Muş, Nevşehir, Niğde, Ordu, Rize, Siirt, Sivas, Şanlıurfa, Tokat, Tunceli, Uşak, Van, Yalova, Yozgat
2	Ankara, Artvin, Aydın, Bursa, Denizli, Edirne, Giresun, İzmir, Kahramanmaraş, Kastamonu, Kocaeli, Muğla, Samsun, Sinop

The membership probabilities demonstrate the likeliness of a station falling in one cluster with reference to another and the clusters not being predetermined provide a comparative approach to other clustering methods. The membership probabilities are calculated blindfoldedly using a iterative algorithm based on a underlying mixture model. The blindfoldedness of the algorithm provides an objective way to describe the variabilities across the time series. When we used the clusters mentioned in Çelebioğlu (2018), we came across the following results in Table 6.

The model-based clustering procedure was carried out using the longclust package in R studio. 16 types of algorithms were evaluated based on Bayesian Information Criteria (BIC) and came across the following results in the Table 5:

Where Y_lag represents the lagged values of Y and s1_12, s2_12, s3_12, s4_12, s5_12, s6_12, s7_12, s8_12, s9_12, s10_12, s11_12 variables represent the clusters from the first to the eleventh in 12-cluster situation, respectively. The 11th and 12th cluster is demonstrated to be insignificant as their p-value is above %5. After application of the agglomerative clustering, the original number of clusters were first reduced to 3 and later 2. The results improved significantly in terms of between R^2 which can be seen in Table 7 and Table 8:

Table 6: Panel data analysis based on results from Çelebioğlu (2018)

within R^2	0,0254	
between R^2	0,9523	
overall R^2	0,6847	
	Coefficients	p-value
Y_lag	0,6394	0,0000
s1_12	-2,1635	0,0000
s2_12	-2,0690	0,0000
s3_12	-1,7266	0,0000
s4_12	-1,7608	0,0000
s5_12	-0,8000	0,0050
s6_12	-1,8772	0,0000
s7_12	-2,5926	0,0000
s8_12	-2,0271	0,0000
s9_12	-1,9345	0,0000
s10_12	-0,4598	0,2020
s11_12	0,4091	0,2560
constant term	6,7115	0,0000

Table 7: Panel data analysis based on results after agglomerative clustering which aimed for reducing the number of clusters to 3

within	0,0254	
between	0,9870	
overall	0,6647	
	Coefficients	p-value
s1_3#Y_lag		
0	0,6891	0,0000
1	0,6309	0,0000
s2_3#Y_lag		
0	0,0477	0,0000
constant term	4,3143	0,0000

Table 8: Panel data analysis based on results after agglomerative clustering which aimed for reducing the number of clusters to 2

within	0,0253	
between	0,9927	
overall	0,6618	
	Coefficients	p-value
s1_2#Y_lag		
0	0,7160	0,0000
1	0,6851	0,0000
constant term	4,2252	0,0000

Where Y_lag represents the first order lagged values of Y; the s1_3, s2_3 represents first cluster, second clusters in 3-cluster situation; s1_2 represents first cluster in 2-cluster situation and s1_2#Y_lag represents the interaction of s1_2 and Y_lag. In Table 7, the value 1 for s2_3#c.Y_lag has been omitted when we analyzed the data. The comparison of the coefficients for s1_3#Y_lag demonstrates that autoregressive effect is much smaller for the stations falling in the first cluster than for those that don't. This would mean that the data from the stations that don't fall in the cluster 1 should be used in an independent study. The results from Table 8 can also interpret in that way. The following results in Table 9 have been obtained using the clusters from model-based clustering. The d1 variable represents the first cluster in 2-cluster situation

Table 9: Panel data analysis based on results of model-based clustering

within	0.0254	
between	0.9997	
overall	0.6585	
	Coefficients	p-value
d1#Y_lag		
0	0,6981	0,0000
1	0,7038	0,0000
constant term	4,1250	0,0000

Where Y_lag represents the first order lagged values of Y; d1 represents the stations falling in the cluster 1 based on model-based clustering and The d1#Y_lag variable is interaction of d1 and Y_lag. The d1#Y_lag variable has coefficient values 0,6981 and 0,7038 when d1 takes values 0 and 1, respectively. This indicates that the stations in the first cluster have a more significant autoregressive effect than those in the second cluster.

When we compared the last 2 results of clustering analysis in Table 8 and Table 9, the results from the model-based clustering procedure performed better in terms of between R^2 and the stations included in clusters differed as displayed in Table 10:

Table 10: The comparison of cluster structures in 2 cluster situations based on both agglomerative clustering and model-based clustering

Cluster Id	Station names included in cluster based on agglomerative clustering	Station Names included in cluster based on model-based clustering
1	Adana, Adıyaman, Afyonkarahisar, Aksaray, Amasya, Antakya, Ardahan, Bartın, Bayburt, Bilecik, Bingöl, Bolu, Burdur, Çankırı, Düzce, Elazığ, Erzincan, Gaziantep, Gümüşhane, Iğdır, Isparta, Karaman, Kayseri, Kırklareli, Kırşehir, Kilis, Kütahya, Malatya, Mardin, Mersin, Muş, Nevşehir, Niğde, Ordu, Rize, Siirt, Sivas, Şanlıurfa, Tokat, Tunceli, Uşak, Van, Yalova, Yozgat	Adana, Afyonkarahisar, Amasya, Ankara, Artvin, Aydın, Bartın, Bayburt, Bilecik, Bolu, Burdur, Çankırı, Denizli, Erzincan, Iğdır, Isparta, İzmir, Kahramanmaraş, Kayseri, Kırşehir, Kocaeli, Malatya, Mardin, Muğla, Niğde, Ordu, Rize, Siirt, Şanlıurfa, Van, Yozgat
2	Ankara, Artvin, Aydın, Bursa, Denizli, Edirne, Giresun, İzmir, Kahramanmaraş, Kastamonu, Kocaeli, Muğla, Samsun, Sinop	Adıyaman, Aksaray, Antakya, Ardahan, Aydın, Bingöl, Bursa, Düzce, Edirne, Elazığ, Gaziantep, Giresun, Gümüşhane, Karaman, Kastamonu, Kırklareli, Kilis, Kütahya, Mersin, Muş, Nevşehir, Samsun, Sinop, Sivas, Tokat, Tunceli, Uşak, Yalova

In total, 41 stations have changed clusters and changed the effectiveness of clustering analysis in terms of between R^2 and coefficients. In Table 9, the coefficient value of d1 was 0,7038 which represented the first cluster's effectiveness and was higher than the base value of 0,6981. However in Table 8, the coefficient value was 0,6851 which was lower than the base value. These findings demonstrate that cluster analysis can change the results without giving any evidence to what results data may provide. Therefore, only the most significant of criteria and a model to serve as a base for comparison should be used in order to provide results. The following model in Table 11 has served as a base to compare the last 2 models:

Table 11: The base model with lagged values of Y being the explanatory variable

within	0,0254	
between	1,0000	
overall	0,6584	
	Coefficients	p-value
Y_lag	0,7015	0,0000
constant term	4,1206	0,0000

Where Y_lag represents the lagged values of Y. The coefficient for Y_lag variable is lower than 0,7038 but also higher than 0,6851. This creates a philosophical problem with the clustering method because in Table 10, we demonstrate that no matter how many stations clusters have in common adding a station can have an effect in terms of between R^2 and in general, a problem must be formulated according to falsifiability criterion. Table 11 demonstrates that the coefficient of Y_lag is 0,7015 which provides a central measure for the coefficients of clusters to be around but the difference of cluster results provide bias which comes from cluster analysis and cluster analysis also becomes ineffective as it produces heterogeneous clusters which is the opposite of the aim of the cluster analysis. The ineffectiveness of cluster analysis can be understood from the difference between numbers of stations in clusters. Nevertheless, the model-based clustering has provided a better solution to understand the variability across stations.

Conclusion and Some Remarks

In this study, we demonstrated that cluster analysis can have some usage in terms of between R^2 and on future studies in the field of meteorology. The purpose that model-based clustering should serve is providing potential places to collect data and formulate a theory based on that. In order to make advancements on meteorological data, the results in Table 10 can be used. The common stations, e.g. Adana, from both of clustering methods can be examined to formulate a theory and the uncommon ones can be used for cross-validation.

Acknowledgements

We thank Prof. Dr. Salih Çelebioğlu for his contributions. Some of the work of the article has been provided by him.

References

Bhaumik, D., Sengupta, D. (2020), Estimating Historic Movement of a Climatological Variable From a Pair of Misaligned Functional Data Sets, *Environmental and Ecological Statistics*, 27, 729-751.

Çelebioğlu, S. (2018), On Some Climatic Scenarios For Turkey From The Perspective of Changes in the Annual Mean Temperatures via Aggregation by Steady-State Distribution, *International Journal of Environment and Geoinformatics*,5(2), 197-217.

Dexen, D.Z.Xi, Dean, C.B., Taylor, S.W.(2019), Modeling The Duration and Size of Extended Attack WildFires As Dependent Outcomes, *Environmetrics*, 31(5).

McNicholas, P.D. (2017), *Mixture Model-Based Classification*, CRC press, New York.

Ülker, D., Ergüven, O., Gazioğlu, C. (2018). Socio-economic impacts in a Changing Climate: Case Study Syria, *International Journal of Environment and Geoinformatics*, 5(1), 84-93.