Journal of Statistics & Applied Science, Volume -2 Issue -2
ISSN 2718-0999

Article

Received: date: 14.09.2021
Accepted: date: 24.10.2021
Published: date: 28.11.2021

Visualizations and Compositional Data Analysis for the European Time-Use

Cenk İÇÖZ[1]

[1]Eskişehir Teknik Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Eskişehir Turkey, cicoz@eskisehir.edu.tr
Orcid: 0000-0002-0219-1088[1]
* Correspondence Author, e-mail: cicoz@eskisehir.edu.tr

Abstract: Time-use data is a specific data that is of interest to many researchers such as sociologists, psychologists, statisticians etc. It is multivariate by nature and can be seen as a subtype of multivariate data called compositional data. Thus, time-use data can also be analyzed and visualized by compositional data analysis techniques. This is rarely mentioned in literature; usually analyses and visualizations of time-use data are shown in multivariate real space. However, this may lead to false interpretation of the data and even some multivariate statistical analysis cannot be directly applied before a transformation is made. The main contribution of this study is to show how an analysis might be used for time-use data and what benefits might be gained from this new setting of the data set as compositional parts. Indeed, the results show that there are some differences between traditional techniques and compositional data analysis.

Keywords: Time-use data, Compositional data, Log-ratio analysis, Clustering Analysis, Ternary Diagrams

1. Introduction

Recent advances in computing and visualization software help data scientists working in several fields examine the data easily. Statisticians name the visual interpretation of the data as exploratory data analysis (EDA), which is pioneered by John W. Tukey [1] in his seminal book. EDA is the primary analysis for many statisticians to visually examine data first-hand before testing it. Using powerful statistical software, implementation of interactive visualizations, and statistical analyses are not bothersome anymore. Interactive visualizations are a handy tool and can give the researcher or data scientist especially a first impression. Interactive graphics on a computer screen allow a person to select individual observation values and the relationships between different variables. The adaptation of modern visualizations and analyses to time-use data may enlighten the patterns behind the time-use data and enable the scientist to understand what might be happening behind the curtain.

Time-use surveys have been conducted by the national statistics offices of many European countries. These surveys, which are designed to get information about several activities of people's daily life, are a great data source for investigating the differences of time-use according to gender and also the differences among countries. Time-use data that consist of several variables of activities of time-use for a sample of individuals chosen from different countries are naturally a form of multivariate data. Additionally, time-use data set are defined so that each row of the data adds up to 100% for the activities investigated. These type of data sets can be analyzed in a different space by using compositional data analysis techniques. In the literature, this property of time-use data sets is rarely mentioned. Compositional data has a correlation structure and other drawbacks for classic multivariate statistical analysis. Therefore, before applying multivariate analysis to compositional data some sort of transformations might be needed. Furthermore, visualizations like ternary diagrams, which are often used in the literature for other data types, may be used to subject time-use data to visual inference.

https://doi.org/10.52693/jsas.994860

The literature will be given below with the property of similar data use and the similar statistical analysis use with the aim of visualizing the data in lower dimension and clustering the objects (countries). Gálvez-Muñoz et. al. [2] examined Harmonised Time-Use Surveys (HETUS) from the perspective of gender inequality. They formed a difference data of gender by dividing women's time-use by men's time-use. The attractive property of this data is that it can be a subject for lower dimensionality visualization of gender inequality. By only looking at the ratio data one can make inferences about gender inequality. Higher ratios refer to higher differences for the concerned attribute. The target people of this study are in the age of 20-74 and under employment. Hierarchical clustering was performed using Euclidean distances as a proximity matrix and formed a grouping of countries by using this difference data. Furthermore, group mean differences were tested with the employment of analysis of variance. The study aimed to find out whether unpaid work is the main reason of gender inequality in all countries or not. Finally, they realized that Eastern European countries notably differ in clustering. Furthermore, in this study the differences among countries were not taken into account within each gender.

Moreno-Colom [3] examines gender segregation in domestic work. In this study, gender's role in contrasting the influence of welfare regimes and employment status on daily life organization is compared. First, they made country-wise comparisons of time-use patterns of European countries. Afterwards, they presented a special case about the effect of employment status on distribution of housework in Spain. One interesting result of the study is that daily task completion restricts equal share of housework between genders.

Coffe [4] examines the time-use of the Members of Parliament (MPs) of New Zealand. They claim the study of her is an innovative one, which is the first study ever to investigate the MPs' time-use. They asks to MPs both for completion of a time diary and a questionnaire. Using correspondence analysis as a main tool for cross tabular data, they find interesting results like differences according to gender and seniority between related activities of MP's and make inferences of MPs' time-use. For instance, Female MPs tend to spend more time on communication, meetings, travelling, and research and reading. On the other hand, male MPs decide to participate in social activities as well as attending House sessions.

Robinson and Gershuny [5] applied a multidimensional scaling (MDS) to diary collections of Oxford MTUS (Multinational Time-Use Study) data archive and compared these mappings with the 1965 Multinational Time-Budget Study. MDS is a technique that spatially maps the data using a proximity matrix.

Wight et al. [6] examines American teenagers' time-use by using ATUS data with a focus of variables that may affect children's well-being. They use OLS (Ordinary Least Squares) regression or logistic regression to find out the relationship between parental and household characteristics and teenagers' use of time.

None of the aforementioned studies [2-6] include any information about the space that the time-use data belongs to and all the analyses are performed without taking the compositional nature of the data into consideration. There are some recent studies that employs compositional data analysis for time-use data. Dumuid et al.[7] use compositional data analysis from Wave 6 of the Longitudinal Study of Australian Children diary data to explore the relationships between daily time-use. They also mentioned the drawbacks of using compositional data in time-use analysis. Gupta et. al. [8] compared time spent sedentary and in physical activity between age groups and sexes. They also showed and emphasized the difference of compositional data analysis results from the results that obtained through standard analytical procedures.

The aim of this study is to create interactive and classical visualizations of time-use data in a restricted space and search for patterns in data by using several statistical analyses techniques such as clustering analyses and log-ratio analysis (LRA). In this study, the compositional data analysis setting of the time-use data is considered. The statistical techniques used in this study get the benefits of using compositional data analysis approach as compared to the conventional data analysis. The compositional version of the related methods like LRA and distance measure is employed in clustering analysis. In doing so, this new setting might reveal completely different results and inferences compared to other studies using the traditional techniques for the same data. For clustering analysis, dissimilar formation of groups is obtained through two distance measures in different measurement spaces. LRA is an analysis based on log-ratios so it will give us a chance to make pairwise comparison of the activities of

countries' time-use neither principal component analysis (PCA) nor correspondence analysis (CA) will give. It is the only method among which has the sub-compositional coherence property. The article will be organized as follows. The theory of the statistical analyses is explained briefly in the next section of the study. Then, in the data section, categories of activities and the information about structure and gathering of the data are given. In the results section, the results of the analyses and interesting visualizations are shown, and some interpretations are given. Finally, a discussion is given about the results and findings, why to use compositional data analysis with time-use data, and other key points of the article.

2. Materials and Methods

A composition is often defined as a vector of D positive components $\mathbf{x} = [x_1, x_2, \dots, x_D]$whose total is equal to a constant K. Compositional data are strictly non-negative data that have a constant sum and its components consist of only relative information. Generally, the sum of the data is not of interest to researchers. Examples can be given as percentages of workers in different sectors, portions of the chemical elements in a mineral, concentration of different cell types in a patient's blood, portions of species in an ecosystem or in a trap, concentration of nutrients in a beverage, portions of working time spent on different tasks, portions of types of failures, percentages of votes for political parties, sources of pollution in air or in a water source, expenditures of households to different spending item categories, etc [9,10,11].

There are some disadvantages of using conventional multivariate analysis techniques for compositional data. Therefore, some transformations have to be applied before implementing multivariate analysis. These disadvantages can be found in van den Boogaart & Tolosana-Delgado [11]. For instance, variance-covariance matrices are singular so that some of the statistical analysis like Hotelling's T-square test and linear discriminant analysis cannot be applied directly; components do not fit into normal distribution which is a key assumption for many statistical analyses; and lastly, the data is correlated because of the restricted total sum. One can easily write a component in terms of other *(D-1)* components.

A subset of data that doesn't contain at least one component of the original compositional data is called a subcomposition. Closure operation is expressing each component of a compositional data in proportions, by just dividing them with the constant sum, or in percentages, by dividing them with the constant sum multiplied 100. In this way, the total is also transformed into 1 for proportions and 100 for percentages. A subcomposition has subcompositional coherence property after closure operation is performed if the calculated values (statistics or ratios) do not change when a subcomposition is used instead of full compositional data. Hence, the planned analysis results will not change for the subcomposition. Greenacre & Primicerio [12] proved numerically that log-ratios have the property of subcompositional coherence.

Clustering analysis is the grouping of similar objects according to proximity matrices into clusters; the objects in the same cluster are homogenous, and objects in different clusters are heterogeneous. There are several clustering methods. The most commonly used ones in statistical applications are k- means algorithm and hierarchical clustering.

In hierarchical clustering, as the name suggests, clusters are formed by a hierarchy at each level of the hierarchy by merging or dividing clusters or objects at the next lower level. Hierarchical clustering can be divided into two basic paradigms: agglomerative (bottom-up) and divisive (top-down). The agglomerative one begins at the bottom with single observation as a cluster and at each level recursively merges a selected pair of clusters or objects into a single cluster. The pair chosen for merging includes the two groups with the smallest intergroup dissimilarity. At the lowest level, each cluster contains a single observation. At the highest level, the final clusters are composed of all elements. Divisive strategy can be considered as the opposite of the agglomerative strategy [13].

The problem in the hierarchical clustering is the merging after the first iteration: merging a cluster and a single object. The distance matrix needs to be upgraded according to several rules. In the single linkage method, the minimum distance between a cluster and a single object is considered, while an average distance is taken into account in the average linkage method. Last of all, the maximum distance between a single object and a cluster is taken into account in the complete linkage method.

The average linkage method will be employed in this study for clustering analysis by using an adaptation of a Euclidean distance matrix for compositional data. Therefore, the results for the

clustering analysis will be completely different in real multivariate space. The clustering will be conducted in Aitchison's sample space called simplex [12]. Simplex is given with the constant sum, K, and the non-negativity constraint as in equation (1):

$$\mathbb{S}^D := \left\{ \boldsymbol{x} = [x_1, x_2, \dots, x_D] : \; x_i > 0 \; ; \sum_{i=1}^{D} x_i = K \right\} \quad [1]$$

To perform the specific operations in the simplex, the data must be converted via a transformation called centered log-ratio (clr) transformation. This transformation can be thought of as reweighting the data according to its geometric mean *g(x)*, which is considered a general tendency measure used for ratio-scale data. The clr transformation is defined as in equation (2):

$$clr(\boldsymbol{x}) = \; log\left(\frac{x_i}{g(\boldsymbol{x})}\right)_{i=1,2,\dots,D} \quad [2]$$

The distance matrix will be organized between composition "ith" and composition "jth" through Aitchison distance given in equation (3). Furthermore, it can be easily interpreted as Euclidean distances of clr transformed data.

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left[\sum_{k=1}^{D} \left(log\left(\frac{x_{ik}}{g(\boldsymbol{x}_i)}\right) - log\left(\frac{x_{jk}}{g(\boldsymbol{x}_j)}\right) \right)^2 \right]^{\frac{1}{2}} \quad [3]$$

$g(\mathbf{x}_i)$ in equation (3) is the geometric mean of the component ith. Martin-Fernandez et. al [15]compared the measures of difference for compositional data in use for hierarchical clustering and mentioned that the unit constraint has to be taken into account while performing hierarchical clustering on compositional data. The Aitchison distance has also got the subcompositional coherence property and other properties of compositional data according to the study.
Martin-Fernandez et al. [15] gave an important example of the difference between Aitchison distance and the Euclidean distance. The paper shows that a three-part (D=3), compositional data consist of only 4 compositions. These are given in equation (4):

$$\boldsymbol{x_1} = [0.1, 0.2, 0.7] \, , \boldsymbol{x_2} = [0.2, 0.1, 0.7], \boldsymbol{x_3} = [0.3, 0.4, 0.3], \qquad \boldsymbol{x_4} = [0.4, 0.3, 0.3] \quad [4]$$

Euclidean distances between the first two and last two compositions are equal according to the compositional data set given. On the other hand, their compositional distances are not equal, although they have the same amount of differences, ±0.1, in the first two components. This is the result of the compositional nature of the data that one part is dependent on the other parts. In the calculation of the distance between composition 1 and composition 2, the residual part is 0.7 while it is 0.3 between compositions 3 and 4. It can be interpreted that the first distance is produced over a residual of 0.7 that could produce a greater distance, while the second distance is produced over 0.3 residual. Therefore, the distance between the first two compositions is greater than the last two. Thus, the two dendrograms that are obtained through Euclidean and Aitchison distance will be presented to examine the differences between them. Different clustering occurs for the same data because of the reasons mentioned above.
Ternary diagrams are the fundamental graphs to visualize a 3-part compositional data. They are the analogy of the scatterplots for the three-dimensional display. All points will be located in a planar triangle in a ternary diagram with the edges at (1,0,0), (0,1,0), and (0,0,1). The interpretation of ternary diagrams is not complicated, and ternary diagrams can be used as a tool of visual analysis for compositional data. Through an edge of the triangle, the value in that edge can increase up to 1 for the associated component, whereas the other two components' values are approximately zero (because of the constant sum property). Furthermore, through the middle of the triangle, each components' value approximates to each other like (0.33, 0.33, 0.33) for a composition.

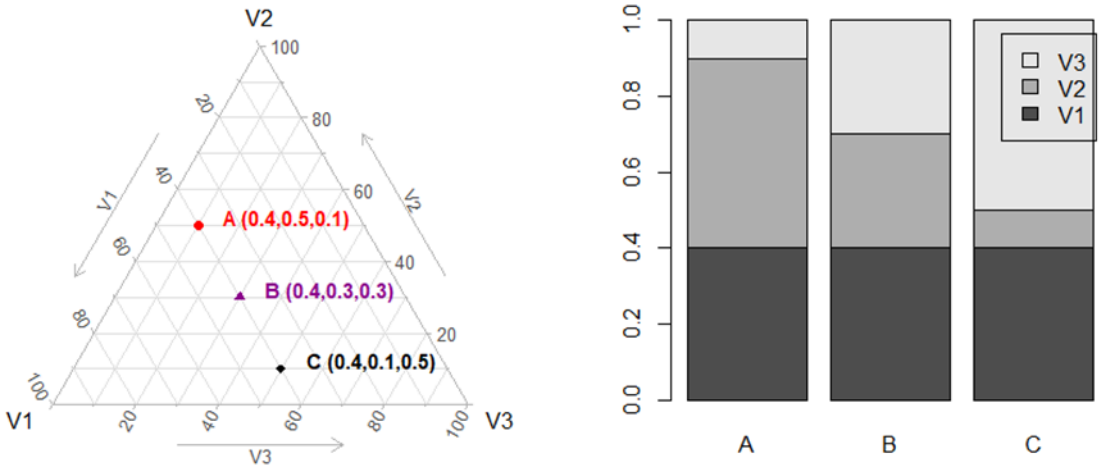

Figure 1. Ternary diagram example and bar graph

In Figure 1, some compositions are given for the interpretation of a ternary diagram in the left and it's counterpart as a stacked barplot in the right. A component's value remains the same along a line parallel to the side of a triangle, which is also located opposite to an edge. To illustrate this property, three compositions (A, B, and C) are given in Figure 1. All of them have equal component values in V1 component so they are in the same line that is parallel to the side of triangle located opposite of V1 edge. The arrows outside the triangle show the direction of the increase through the edge of the related component's value. If the composition gets closer to an edge, it gets a higher value on that component up to 1. Let's say that V1 is gainful work activity and V3 is domestic work activity. Compositions A, B, and C are in a line parallel to V2-V3 side and opposite to the V1 edge. Therefore, the values on gainful work component for these three composition stay the same with an amount of 0.4, and only the other two components' values change. Furthermore, when going down from composition A through composition C along the same line, the location of a component is getting closer to the V3 edge. Therefore, from composition A to composition C, domestic work's value is increasing from 0.1 to 0.5. Figure 1 is plotted by using the "ggtern" package in R.

In the bar graph, it may be inferred that the exact values from the scale of the axis on the other hand for a three part composition it could be known for sure in a ternary diagram with given scales. As the size of the variable number increase for a compositional data, the 3-part combinations also will increase and it will hard to interpret the data with ternary diagrams either.

In Figure 2, the stacked barplots according to genders with values given as labels in each individual bar representing each variable. By this way, it is easy to see the whole variable values together in a visualization without inferring the variable values from the scales of the graph.

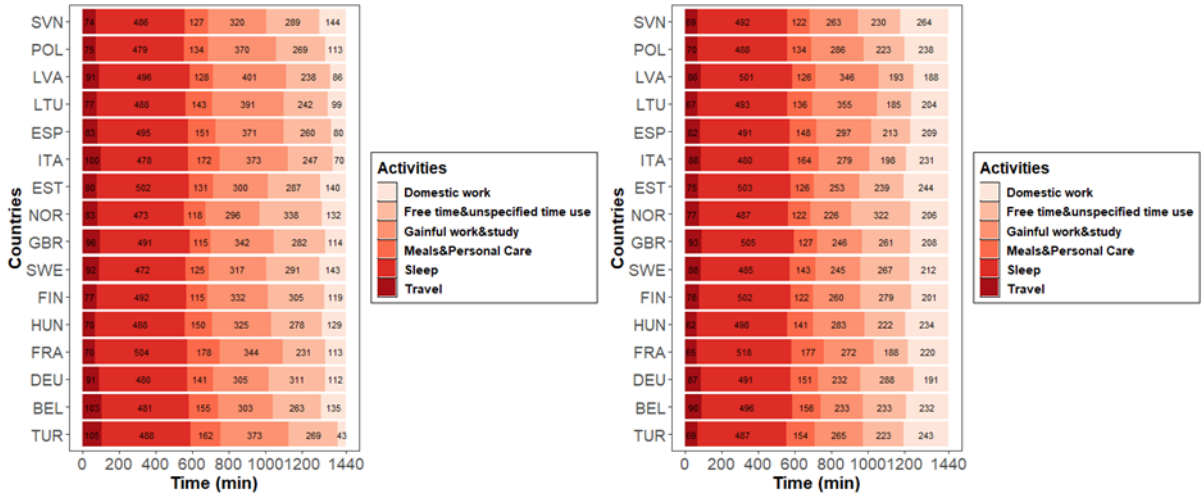

Figure 2. Stacked barplots according to genders with labels

The same kind of relationship between scatterplots and scatterplot matrices is also valid for ternary diagrams and ternary diagram matrices. For a compositional data that is composed of more than three components, ternary diagram matrices are employed in literature. The "ggtern" package provides a margin for the third component on the top edge, which can be both a fixed component itself or the geometric mean of the rest of components. Likewise scatterplot matrix ,the part in the right of the diagonal are the same except for the replacement of the two components located in the bottom side of the triangle.

Log-ratio analysis (LRA) is a method which is used to visualize compositional data in lower dimensions. When all the variables of the data are in interval scales, PCA is the preferred analysis. However, for the ratio scale data, multiplicative differences can be the subject of the analysis. Log-ratio transformation can be used to convert multiplicative differences to the additive differences and enable to application of PCA to ratio-scale data.

LRA is similar to PCA and CA that is used to visualize pairwise log-ratios in a biplot along with the sample points. Usually, standardization should be applied before PCA for data which is measured in different units, and the variables need to be reweighted through their standard deviations. When all variables are measured in the same scale as in our case, log-ratios are the perfect standardization. They have a special feature of comparability between both within variables and both within objects like in our case. Thus, no further standardization is necessary when log-ratio transformation is conducted once. The main difference between LRA and CA/PCA is that it has subcompositional coherence, because log-ratios remain the same for a subcomposition. Instead of biplot axes in PCA and CA link vectors are used for interpretation of biplot in LRA [10]. These link vectors demonstrate the outperforming objects in each side. When a composition is located close to a component, it means that it has got a higher ratio when compared to other objects in that link. For instance, if a composition has a higher ratio of component A over B, it is located closer to the component A side of the link vector as compared to other compositions, whereas it is far away from side B of the link vector and vice versa.

The algorithm of the weighted version of LRA is given in Greenacre [14] below:

1) Calculate the row and column margins, **r** and **c**, respectively, when $n = \sum_i \sum_j n_{ij}$ is the grand total of the compositional data matrix **N**: $\mathbf{r} = (1/n)\mathbf{N1}$, $\mathbf{c} = (1/n)\mathbf{N}^{\mathrm{T}}\mathbf{1}$
2) Logarithmic transformation of the elements of the matrix **N**: $\mathbf{L} = log(\mathbf{N})$
3) Weighted double-centering of L: $\mathbf{Y} = (\mathbf{I} - \mathbf{1r}^{\mathrm{T}})\mathbf{L}(\mathbf{I} - \mathbf{1c}^{\mathrm{T}})^{\mathrm{T}}$
4) Weighted singular value decomposition (SVD) of **Y**: $\mathbf{S} = \mathbf{D}_{\mathrm{r}}^{1/2}\mathbf{Y}\mathbf{D}_{\mathrm{c}}^{1/2} = \mathbf{U}\mathbf{D}_{\varphi}\mathbf{V}^{\mathrm{T}}$
5) Calculation of the coordinates

The weighted version of LRA has several advantages like giving a different weight for each row and column (proportional to row and column margins). The positions of the objects in the biplot are determined by the log-ratio distances given in equation (3).

2.1. Data

The sources of the data are given in Aliaga [17]. Data for Turkey is taken from the Turkish Statistical Institute (Turkish Statistical Institute 2006). The data which consist of 16 countries and 6 variables of aggregated activity categories from time-use diaries. Components represent average time spent on each activity for the countries given. Time-use data is a multivariate data, which includes 6 different activities concerning the time-use as variables (components in compositional data case). Each variable has strictly positive values, and each of them carry relative information. The total time is 1440 minutes in a day, which is the constant sum of the each composition. Therefore, the data fits into the classic definition of compositional data.

Time Use Surveys provide statistics about differences between women and men in gainful and domestic work, their participation in educational and cultural activities, and other aspects of life. A representative sample of individuals completes a diary during one weekday, and one weekend day is distributed over the whole year [17]. Daily time is divided into 6 categories, which includes same kind of tasks for ease of interpretation. These are gainful work & study, domestic work, travel, sleep, meals & personal care, and free time and unspecified time-use. Explanations of these categories are given in[15], as written below:

- Gainful work and study includes time spent on primary and secondary jobs and related activities, breaks, and travel during working hours, and during job seeking. The time spent on study at school and during free time is combined with gainful work.
- Domestic work includes housework, child and adult care, gardening and pet care, construction and repairs, shopping and services, and household management.
- Travel includes commuting and trips connected with all kinds of activities, except travel during working hours.
- Sleep includes sleep during night or daytime, waiting for sleep, naps, as well as passive lying in bed because of sickness.
- Meals, personal care includes meals, snacks and drinks, dressing, personal hygiene, making up, shaving, sexual activities, and personal healthcare.
- Free time and unspecified time use includes all other kinds of activities, e.g. volunteer work and meetings, helping other households, socializing and entertainment, sports and outdoor activities, hobbies and games, reading, watching TV, resting, or doing nothing.

There are some methodological notes concerning how the surveys were conducted. The age range of the respondents differs between countries. Long time periods spent on resting is counted as sleep in France, whereas in other countries, rest is included as free time. Also, the national data is rounded, which may result in some discrepancies (see methodological notes [17] for more information).

Although the data is outdated, the results and interpretations regarding to it may differ slightly when compositional analysis methods are employed. Furthermore, the collection of this sort of data is not an easy task as completion of a diary is needed for citizens chosen as a sample for each country. Indeed, great collaboration is necessary between national statistics offices of countries to fulfill the instructions which are constituted by EUROSTAT. There are still nuances in the application of the instructions which may affect the studies conducted with the data and must be taken into consideration.

3. Results

Descriptive graphics like boxplots, bar plots, and thematic maps can be examined by using the link [18]. All the interactive plots were drawn using Tableuau software. The dashboard in the link below can be examined online, or the tableau workbook can be downloaded to examine each sheet separately. To draw a different thematic map according to a time-use activity, a user must select “variable section” to desired time-use activity. One can freely download Tableuau software’s free trial version or student version by proving a valid student identity.

Compositional calculations and ternary diagram matrixes are made by using the package “compositions” in R. Ternary diagram matrix is given in Figures 3. In the first row and first column of the ternary diagram matrix, free time, domestic work, and gainful work are the selected activities. The genders are shown with different colors to be separated from each other in each ternary diagram, respectively blue and red for women and men. Also, It is seen that there are obvious differences between

men and women in spent time on domestic work and gainful work. In some of the ternary diagrams, the points in the diagram for the men and women coincides which means that there no obvious difference in this 3-part variable selection. However, for other ternary diagrams including 3 variable parts like domestic work, gainful work, free time and meals&personalcare, domestic work and free time the separation is more noticeable between genders because women spent more time on meals&personalcare and domestic work when compare to men had more working hours when compare to women. A clustering dendrogram shows how the units are separated from each other according to a similarity measure. When the height of the dendrogram is high, the dissimilarity of units increases in the dendrogram. In Figures 4 and 5, cluster dendrograms of the hierarchical clustering analysis, which are obtained by using two different distance measures, are given for both genders' time-use. Turkey is seen as a single cluster according to dendrogram of Aitchison distance and is very far away from other countries in Figure 4(a). On the other hand, it is located with Italy and Spain in the same branch at a lower height of the dendrogram in Figure 4(b).

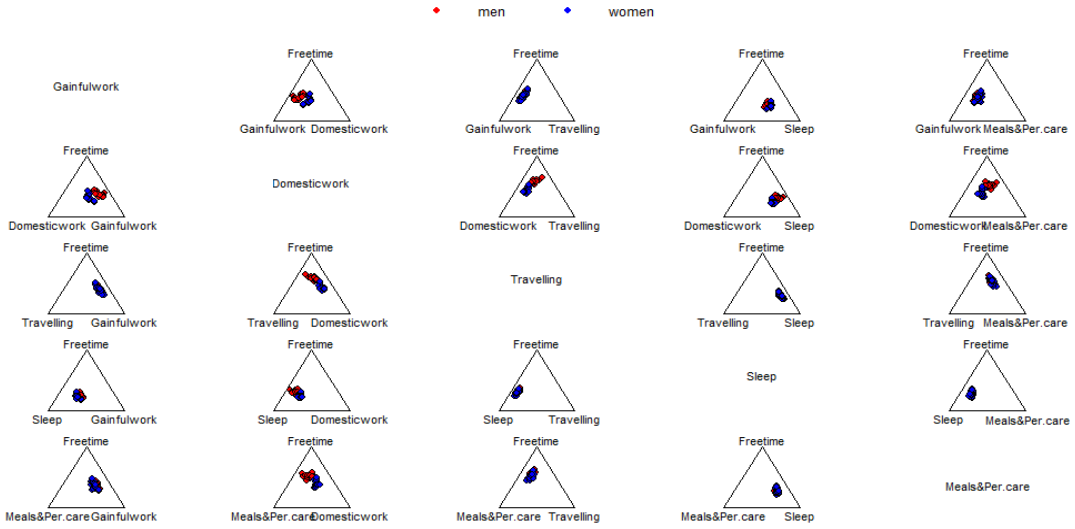

Figure 3: Ternary diagram Matrix of men and women time-use

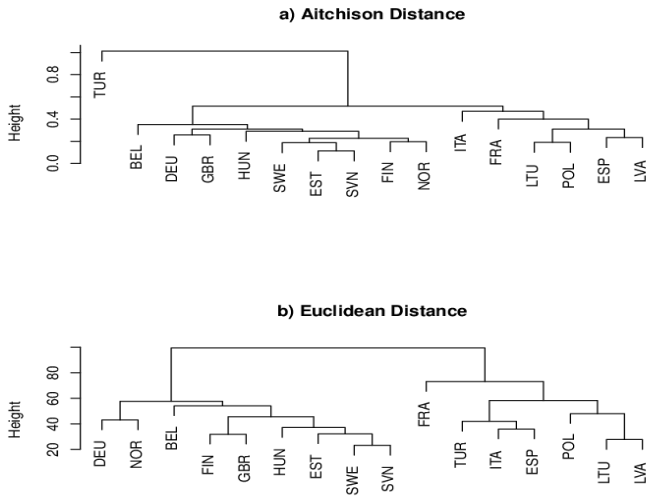

Figure 4. Cluster dendrogram of men's data according to two distance measures.

The women's dendrogram obtained through Aitchison distance in Figure 5(a) is more homogenous compared to mens' when the dendrogram is divided from lower height. When the dendrograms of each gender in Figures 4 and 5 are cut through at a lower height with greater similarity, different clusters show up according to the Aitchison and Euclidean distances.

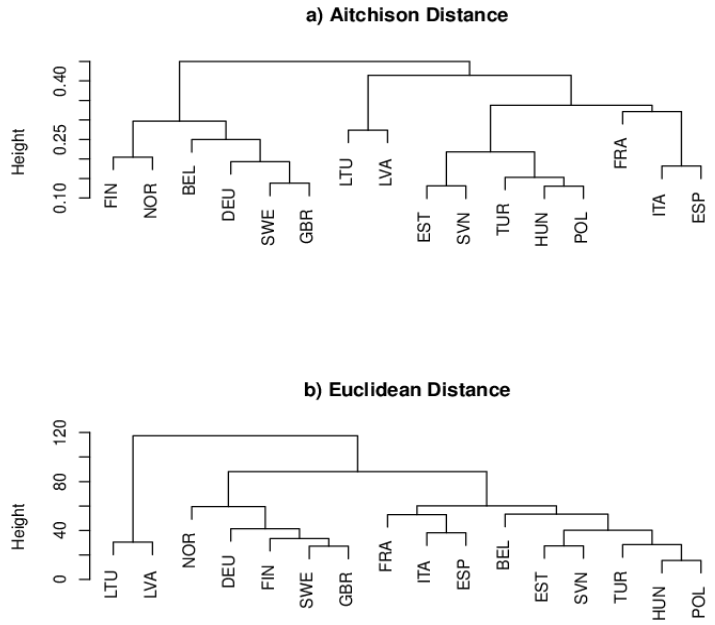

Figure 5. Cluster dendrogram of women's data according to two distance measures.

In Tables 1 and 2, clusters which are the results of the hierarchical cluster analysis using Aitchison distance are given for men and women, respectively. Countries are divided into three clusters by hierarchical clustering analysis according to their time-use for men while for women they are divided into four clusters. Eastern European countries seem to fall into same clusters for both genders. Furthermore, Scandinavian countries are in the same clusters so one can interpret that being spatially close can be a factor influencing the cluster formation. Therefore, a spatial relation can also be considered in clustering time-use. Turkey is listed as a single cluster in men's time-use because its domestic time use is so low as compared to others, and it might be an outlier compare to an average composition.

Dendrograms have different structures according to both distances with major and minor key points. For instance, the time-use of Turkish men completely differs according to two distance measures that would most probably result in a different cluster after the analysis. Therefore, the clusters formation by using different distance measures also differs from one another.

Table 1. Clusters of men according to time-use

Clusters	Countries
1	Turkey
2	Belgium, Germany, Hungary, Finland, Sweden, Great Britain, Norway, Estonia, Slovenia
3	France, Italy, Spain, Lithuania, Latvia, Poland

Table 2. Clusters of women according to time-use

Clusters	Countries
1	Turkey, Hungary, Estonia, Poland, Slovenia
2	Belgium, Germany, Finland, Sweden, Great Britain, Norway
3	France, Italy, Spain,
4	Lithuania, Latvia

In Figures 6 and 7, boxplots of clusters for genders are given. The difference in domestic work and gainful work can also be recognized after the cluster analysis. They can be the underlying cause for the clustering with free time activities. In both genders, the 2nd clusters contain the countries which have got more free time than others. The leading countries in these clusters are the Scandinavian countries of Finland, Norway, and Sweden. The most homogenous clusters are seen in terms of sleep activity,

according to the boxplots. In Norway, diary construction is also considered as socializing, a subcategory of free time, and may be the reason of its higher value in that category.

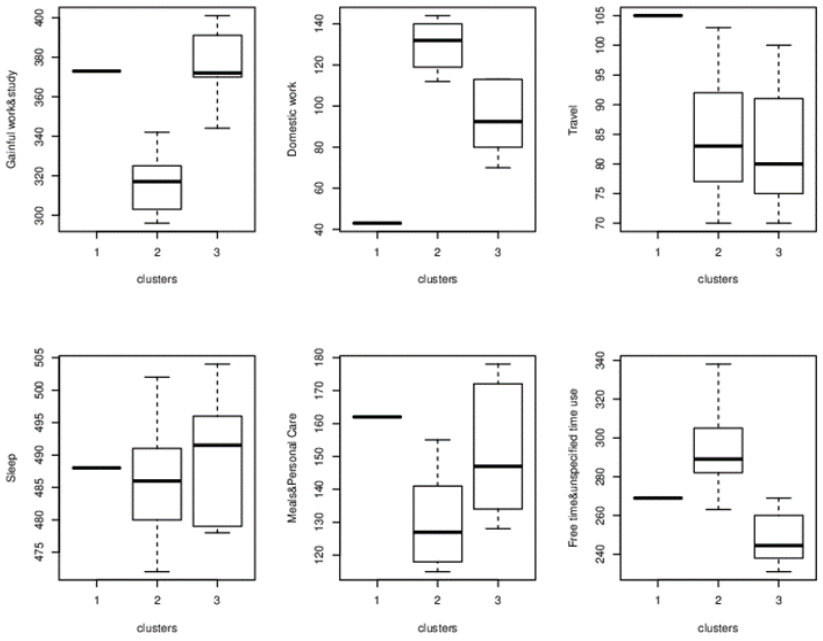

Figure 6. Boxplot of men's time-use data according to clusters

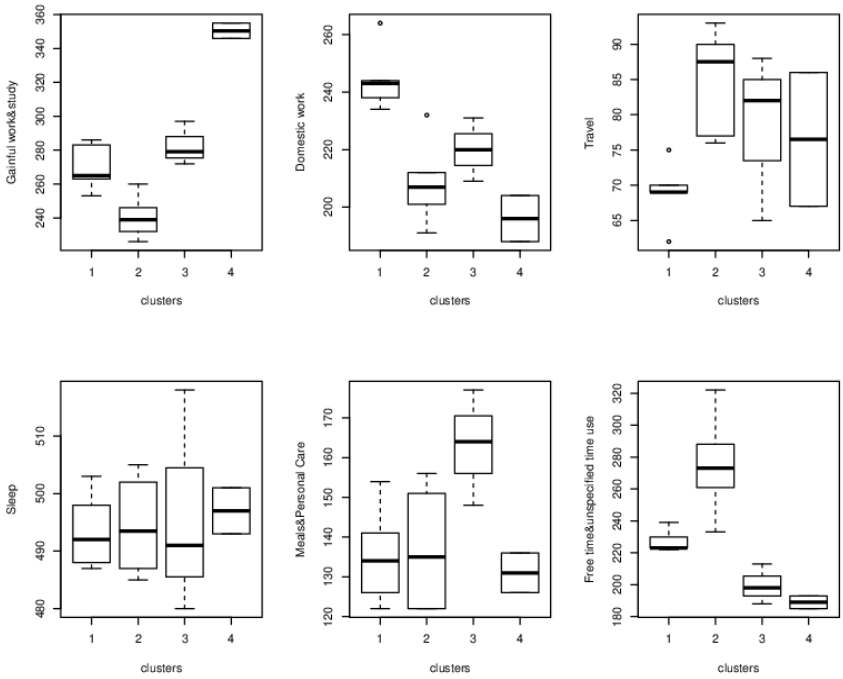

Figure 7. Boxplot of women's time-use data according to clusters

To interpret the log-ratio biplots, the link vectors that connect all the combinations of pairwise components must firstly be drawn. In our example, only three of the pairwise components that are thought of as the most important are taken in consideration. These pairwise components are also selected for an easier interpretation when there are 15 pairwise components in total. A link vector takes the place of the biplot axes in CA but with a slight difference in meaning. A link vector is a two-sided arrow whose component value dominates the other or vice versa through the concerned variable. Therefore, the ratios are reversed according to direction of the arrow.

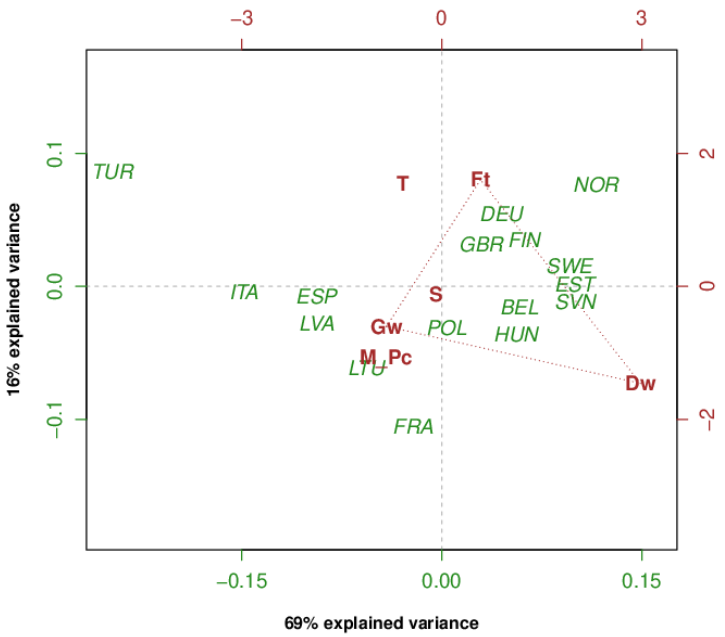

Figure 8. Log-ratio biplot of men's time-use data

In Figure 8, the scattering of the units in log-ratio biplot is obtained. Turkey is located in the upper left portion of the graph, which is obtained as a single cluster in Table 1. Thanks to its lower domestic work value and high gainful work value, the log-ratio of gainful work to domestic work is higher than any other country. It can be inferred that when moving from domestic work to gainful work component, different clusters are revealed. One cluster is located close to free time because of their high values in this component. Examples are Norway, Germany and Finland.

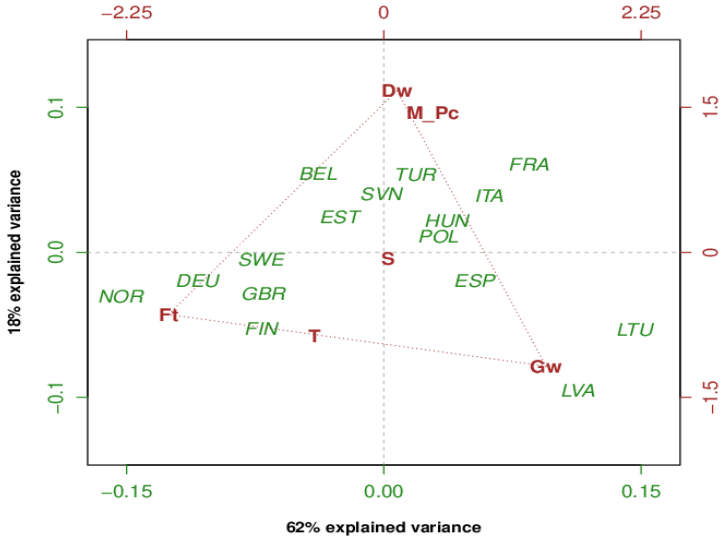

Figure 9. Log-ratio biplot of women's time-use data

In Figure 9, gainful work to domestic work link and gainful work to free time link are the link vectors that affects the cluster formation for women's time use. There are not so diverging countries from components of time as compared to those of men. There are many countries located near the domestic work component. Important difference can be noticed that Latvian and Lithuanian women has got a higher working hours than any other country because they are located in the edge of gainful work variable in the gainful work- domestic work link vector. Furthermore, Scandinavian countries and Germany has got much higher free time when compared to other countries as they are located closer to the edge free time variable on the link vector freetime and domestic work. Turkey, Belgium Slovenia

and other have have high domestic value to freetime and they are located on the other edge of the same link vector.

4. Conclusions

In this study, a bunch of visualizations and related statistical analyses are given to examine the time-use data of European countries. Data includes average time-use of the residents of EU countries and Turkey, which is obtained through surveys (time-use diaries). Even though the surveys are conducted according to some rules, there are still methodologic differences between countries because of the segmentation of time-use Also, the sample size and sample diversity (spatial sampling of the respondents) for a country are the key issues for national representability of the data. Another issue is the homogeneity of the data. The age gap of residents who participate in the survey is so large that it may play a role in a completely different time-use data set.

A pioneering part of this study is that it may be a study in which interactive visualizations is mentioned and the importance of the need to treat the time-use data as a compositional data. It should treat the data according to its own space rather than treating every multivariate data in the multivariate real space.

Furthermore, visualizations resulting from compositional thinking called "ternary diagrams" and log-ratio biplots for time-use compositional data are given. Ternary diagrams can be easily interpreted for three-part, compositional data. However, when the component number increases, the interpretation would be difficult for the researcher. Moreover, LRA interpretation is different from and more complicated that of CA and PCA, which results in a difficult understanding of the findings; however, it is the true visualization tool for the compositional data. Also, it reflects the distance measure given in equation 3, in the determination of the country locations in the biplots. Parallel results are found with a study in literature, mostly giving importance to the gender inequality of time-use.

The differences of using two distinct distance measures are demonstrated with the dendrograms derived from the cluster analysis, and the major differences of the clusters are shown. The proof of getting different results from implementing the statistical analysis in different measurement space is also given in our study. It is important to use an adaptation of conventional multivariate statistical analysis of compositional data, because it fits with the compositional setting. In this way, one may treat a country like a composition as a whole, rather than an object that has independent attributes. Also, a lower dimensionality visualization method which is suitable for compositional data is given for the visual inference of the data.

Clusters obtained in our study have some similarities with those in the study by Gálvez-Muñoz et. al. [2]. However, the cluster analysis is applied to each gender in this study, whereas Gálvez-Muñoz et al. [2] applied cluster analysis on the difference data to investigate gender equality. Furthermore, Gálvez-Muñoz et al. relates countries in the same cluster have the same level of GDP. This inference can also be valid for our study. In this study, I tried to give all the visualizations and analyses separately for each gender so the reader can also see the differences of countries within gender.

Usage of compositional data analysis reveals that Turkey's time-use among men slightly differs with the conventional multivariate counterparts. It can be classified as an outlier after compositional data analysis, but it appears as an ordinary observation in clustering in multivariate real space.

The quantitative analysis part of the study is far more developed than the institutional setting builds upon only gender equality in the given literature. When the social researchers and data scientists get together as a team to investigate time-use, much more trustworthy and consistent results can be obtained. It is hoped that all the visualizations and analyses can be a reference guides to researchers working with time-use data.

Author Contributions:

Funding: No financial resources were provided for this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] J.W. Tukey,. Exploratory data analysis. Reading, Mass, Addison-Wesley Pub. Co, 1977

[2] L. Gálvez-Muñoz, P. Rodríguez-Modroño, and M. Domínguez-Serrano, "Work and Time Use By Gender: A New Clustering of European Welfare Systems". Feminist Economics, 17(4), pp.125–157,2011

[3] S. Moreno-Colom, "The gendered division of housework time: Analysis of time use by type and daily frequency of household tasks". Time & Society, pp.1–25,2015, Available at: http://tas.sagepub.com/cgi/doi/10.1177/0961463X15577269.

[4] H. Coffe, "Time use among New Zealand Members of Parliament." Time & Society, 2015 Available at: http://tas.sagepub.com/cgi/doi/10.1177/0961463X15579578

[5] J.P. Robinson and J. Gershuny, "Visualizing multinational daily life via multidimensional scaling (MDS)", Electronic International Journal of Time Use Research, 2013, 10, issue 1, p. 76-90, Available at : http://eijtur.org/pdf/volumes/eIJTUR-10-1-5_Robinson_Gershuny.pdf#page=76

[6] V. R. Wight, J. Price, S. M. Bianchi, & B. R. Hunt "The time use of teenagers". Social Science Research, 2009. 38(4), pp.792-809.

[7] D. Dumuid, Ž. Pedišić, J. Palarea-Albaladejo, J.A. Martín-Fernández, Hron, K. and T. Olds, "Compositional data analysis in time-use epidemiology: what, why, how". International journal of environmental research and public health, 2020. 17(7), p.2220.

[8] N. Gupta, S.E. Mathiassen, G. Mateu-Figueras, M. Heiden,, D.M. Hallman, M.B. Jørgensen, and A. Holtermann, "A comparison of standard and compositional data analysis in studies addressing group differences in sedentary behavior and physical activity." International Journal of Behavioral Nutrition and Physical Activity, 2018. 15(1), pp.1-12.

[9] J. Aitchison,, A Concise Guide to Compositional Data Analysis, 2005. Available at :http://ima.udg.edu/activitats/codawork05/A_concise_guide_to_compositional_data_analysis.pdf

[10] Bacon-Shone J., A short history of compositional data analysis: In: Pawlowsky-Glahn V and Buccianti A (eds) *Compositional Data Analysis Theory and applications.* 2011, New Delhi: John Wiley & Sons, Ltd

[11] K.G. van den Boogaart, R. Tolosana and M. Bren, Compositions: Compositional Data Analysis. R package version 1.40-1. 2014, Available at: https://CRAN.R-project.org/package=compositions

[12] M. Greenacre, and R., Primicerio, Multivariate Analysis of Ecological Data. 2013. Bilbao: Fundación BBVA

[13] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning 2nd ed. ,2008 ,Springer New York Inc. New York NY USA

[14] J. Aitchison, "The Statistical Analysis of Compositional Data". Journal of the Royal Statistical Society, 1982, 44(2), pp.139–177.

[15] J.A. Martın-Fernandez, C. Barcelo-Vidal, and V. Pawlowsky-Glahn,Measures of difference for compositional data and hierarchical clustering methods In: Buccianti A, Nardi G and Potenza R (eds.) Proceedings of IAMG'98, The Fourth Annual Conference of the International Association for Mathematical Geology: De Frede, 1998 Naples, p. 526–531

[16] M Greenacre, *Biplots in Practice,* 2010, Bilbao: Fundación BBVA

[17] C.Aliaga, 2006, How is the time of women and men distributed in Europe?, EUROSTAT

[18] https://public.tableau.com/app/profile/cenk.i.z/viz/EuropeanTimeUseData/Story1 (accessed Oct. 19, 2021)