Discovering Similarities in the Time Use Patterns of the Spanish Autonomous Communities by Fuzzy Techniques

J. Palarea-Albaladejo[1], **J.A. Martín-Fernández**[2]
[1]Universidad Católica San Antonio, Murcia, Spain; *jpalarea@pdi.ucam.edu*
[2]Universitat de Girona, Girona, Spain

Abstract

In 2000 the European Statistical Office published the guidelines for developing the Harmonized European Time Use Surveys system. Under such a unified framework, the first Time Use Survey of national scope was conducted in Spain during 2002–03. The aim of these surveys is to understand human behavior and the lifestyle of people. Time allocation data are of compositional nature in origin, that is, they are subject to non-negativity and constant-sum constraints. Thus, standard multivariate techniques cannot be directly applied to analyze them. The goal of this work is to identify homogeneous Spanish Autonomous Communities with regard to the typical activity pattern of their respective populations. To this end, fuzzy clustering approach is followed. Rather than the hard partitioning of classical clustering, where objects are allocated to only a single group, fuzzy method identify overlapping groups of objects by allowing them to belong to more than one group. Concretely, the probabilistic fuzzy c-means algorithm is conveniently adapted to deal with the Spanish Time Use Survey microdata. As a result, a map distinguishing Autonomous Communities with similar activity pattern is drawn.

Key words: Time use data, Fuzzy clustering; FCM; simplex space; Aitchison distance.

1 Introduction

As a branch of microeconomics, time allocation theory deals with the topic of how individuals distribute their time among alternative activities. It has usually used the paradigm of the individual as a producing and consuming unit which seeks to find an optimal combination of market work, household work and leisure as to maximize its well-being—in terms of an utility function—subject to constraints on the total amount of time available, income and the price of goods. The theoretical modelling of time use patterns has been the subject of intense study for several decades. The seminal paper by Becker (1965) forms the basis for most current work in this area. The works by Juster and Stafford (1991), Klevmarken (1998) and Hamermesh and Pfann (2005), among others, provide overviews of the major developments within this field.

From an empirical point of view, the Multinational Comparative Time-Budget Research Project, promoted by Alexander Szalai in the middle of the 60's, means the origin of the modern time use studies. Thus, from the early 2000's the official statistical institutes of the majority of the OECD member countries—including the main European countries as well as the US and Canada—have collected large national time use data. This kind of data has become more and more relevant to understand human behavior and to monitor similarities, differences or changes in the way of life of populations or social groups. Also, to analyze the trade-off between work and leisure, the gender roles and obtaining valuable information in order to face problems related to urban environment, transportation, health, etc. How time is allocated to activities and how the allocation changes over time and across countries has implications for economic policy and welfare. Particularly, the differences in time allocation across countries or regions can not only help explain variations in economic growth, but can also clarify the influence of institutions and public policies on individual and household decisions.

If we look at the ways time allocation data are usually analyzed, we find that common descriptive, tabular and graphical methods are mostly applied. Converse (1972) introduces the use of multivariate techniques on such data, but the techniques both in this research and in the great majority of subsequent ones are used in its standard form, not taking into account the technical peculiarities of the data. On the other hand, discrete choice econometric models (see Greene, 2008, ch. 23) are the only generally used tool for modelling allocation of time among a set of alternatives depending on exogenous variables. A glance to any volume of the specialized journal *Electronic International Journal of Time Use Research*, to the collection in Pentland and others (2002) or to the recent review in Michelson (2006) may serve as evidence of such situation. In short, although some references are encountered, the mainstream literature on the subject has avoided the principles introduced in Aitchison (1986) to resolve the difficulties and incoherences of applying standard statistical tools for the analysis of data of compositional nature—as time allocation data are.

Time allocation data are *strictly* compositional in the sense that they are already originally non-negative data sum up to a constant—the total available time—which is exactly the same for all the individuals. That is, the vector of time shares—which reflects the distribution of time across D different tasks—does not require normalization to be defined on the simplex sample space. Formally, we will denote such a vector as $\mathbf{t} = [t_1, \ldots, t_D]$ defined on the $(D-1)$-dimensional simplex

$$\mathcal{S}^D = \{\mathbf{t} = [t_1, t_2, \ldots, t_D] : t_1 > 0, t_2 > 0, \ldots, t_D > 0; \sum_{i=1}^{D} t_i = T\},$$

where T represents the total available time, usually measured in hours—there are 24 hours in everyone's day. Nowadays, we know that the simplex $\mathcal{S}^D$ has an Euclidean vector space structure defined by two basic operations: *perturbation* and *powering* (see Billheimer and others, 2001; Pawlowsky-Glahn and Egozcue, 2001a, or Aitchison and others, 2002). Such a structure allows to define orthonormal bases from which any element of $\mathcal{S}^D$ can be obtained. The coordinates of a compositional vector $\mathbf{t}$ with respect to an orthonormal basis constitute a vector in $\mathbb{R}^{D-1}$. The key is that an isometry is established between $\mathbf{t}$ and its corresponding real coordinates. Hence, angles and distances in $\mathcal{S}^D$ can be associated with angles and distances in $\mathbb{R}^{D-1}$. This allows the analysis of data in the simplex to be based upon its associated real coordinates—using then standard multivariate data analysis techniques for real Euclidean spaces—and afterwards translate the results into the simplex. Such approach has become termed *modelling on coordinates* and it is currently an area of active research (see Egozcue and others, 2003; Pawlowsky-Glahn, 2003).

The objective of this paper is making a comparison of living patterns between the Spanish Autonomous Communities (from now on denoted by ACs) based upon determining a set of homogeneous clusters of them. For the clustering process we will follow the fuzzy approach which provides gradual memberships of objects to clusters, instead of the hard cluster assignments in ordinary clustering methods. This allows to express that objects—ACs for us—belong to more than one cluster at the same time, and can be useful for assessing how natural the cluster structure is. Also, fuzzy clustering offers a great flexibility in handling different types of distance measures. Concretely, we focus on the *fuzzy c-means* (FCM) algorithm (Bezdek, 1981), a flexible, efficient and broadly applicable algorithm which is here adapted to deal with time allocation data. As a result we provide a map wherein homogeneous ACs are distinguished.

The paper is organized as follows. Section 2 is devoted to a brief description of the FCM algorithm and of the way we will use it within a time allocation context. Section 3 presents the data set and deals with the problem of zero observations in it. Next, the clustering procedure is applied and a map of Spain distinguishing homogeneous ACs is drawn. Finally, Section 4 contains some concluding remarks.

2 FCM algorithm on coordinates

The fuzzy clustering problem can be expressed as classifying a given set of objects into c fuzzy subsets—from now on simply called clusters. In our context, let $\mathbf{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_n\}$ be a time use data set where $\mathbf{t}_i$ is the row vector describing the typical time allocation pattern of the population of an AC i, $i = 1, \ldots, n$. A c-partition of $\mathbf{T}$ is a set of values $\{u_{ik}\}$ that are arranged in a $n \times c$ matrix $\mathbf{U} = [u_{ik}]$, being c the number of clusters, with $1 < c < n$. We refer to $\mathbf{U}$ as a *possibilistic* fuzzy c-partition of $\mathbf{T}$ if

$$\sum_{i=1}^{n} u_{ik} > 0, \; k = 1, \ldots, c. \tag{1}$$

Then, u_{ik} is interpreted as a degree of typicality of AC i to cluster k. Alternatively, we refer to $\mathbf{U}$ as a *probabilistic* fuzzy c-partition of $\mathbf{T}$ if, in addition to (1),

$$\sum_{k=1}^{c} u_{ik} = 1, \; i = 1, \ldots, n, \tag{2}$$

holds. Condition (2) formally transform u_{ik} into a membership probability of AC i to cluster k relative to all other clusters.

Procedures derived from the FCM algorithm are commonly used to obtain a possibilistic or probabilistic partition in fuzzy clustering. The FCM recognizes a known or given number of c hyperspherical clouds of points in a data set. The clusters are assumed to be approximately the same size, and each one of them is represented by its centroid ν_k, $k = 1, \ldots, c$. The optimization problem that defines the c-means model is given by

$$\min_{\mathbf{U}, \nu, \mathbf{w}} \sum_{k=1}^{c} \sum_{i=1}^{n} u_{ik}^{m} d^2(\mathbf{t}_i, \nu_k) + \sum_{k=1}^{c} w_k \sum_{i=1}^{n} (1 - u_{ik})^m, \tag{3}$$

where the elements of $\mathbf{U}$ satisfy (1), or both (1) and (2), depending on the approach; $\nu = [\nu_1, \nu_2, \ldots, \nu_c]$ is the vector of cluster centroids; and $\mathbf{w} = [w_1, w_2, \ldots, w_c]$, where $w_k \in \mathbb{R}^+$, is the kth penalty term. This vector of penalties is included in the possibilistic case for avoiding the trivial solution $\mathbf{U} = 0$ (see Krishnapuram and Keller, 1993, for details). The parameter $m \geq 1$ is the degree of fuzzification. With higher values for m the boundaries between the clusters become softer, with lower values they get harder. It has been found from empirical results that values around 2 produce satisfactory results. Consequently $m = 2$ is usually chosen, and this will be used throughout this paper. Finally, $d(\mathbf{t}_i, \nu_k)$ denotes the measure of dissimilarity between an AC i—characterized by the vector $\mathbf{t}_i$—and the kth cluster centroid.

As for any other clustering method, fuzzy clustering is based on a distance measure. In classical applications, the ordinary distances are either Euclidean or Mahalanobis. These ones are well-behaved when dealing with real data, but this is not the case of time allocation data. In Aitchison (1992) the author provides a detailed account of criteria which must be satisfied by a meaningful measure of distance on the simplex space. We will not discuss here about the suitability of the Euclidean distance, or others, for measuring dissimilarity between compositions. We refer the reader to Bohling and others (1998), Martín-Fernández and others (1998), Barceló-Vidal and others (1999) Aitchison and others (2000), Rehder and Zier (2001) and Aitchison and others (2001) for an in-depth debate on the subject. Under our point of view, the main conclusion drawn from these papers is that assuming the fact that compositions contain only relative information—hence the adequacy of a log-ratio representation of them—and also considering the algebraic-geometrical structure of the simplex, the *Aitchison distance* (d_a) defined in Aitchison (1986) is a suitable measure of difference between two time use patterns $\mathbf{t}$ and $\mathbf{t}^*$:

$$d_a(\mathbf{t}, \mathbf{t}^*) = \sqrt{\frac{1}{D} \sum_{i<j} \left(\log \frac{t_i}{t_i^*} - \log \frac{t_j}{t_j^*} \right)^2}.$$

This distance satisfies the requirements of Aitchison (1992) and it can be derived from the own Euclidean structure of the simplex (Egozcue and others, 2003). The Aitchison distance is equivalent in the simplex to the standard Euclidean distance in real space, since geometrically speaking—although not visually—it draws spheres around any point of the simplex. In fact, following the modelling-on-coordinates approach, the next identity holds:

$$d_a(\mathbf{t}, \mathbf{t}^*) = d_e(\mathbf{y}, \mathbf{y}^*),$$

where $\mathbf{y}$ and $\mathbf{y}^*$ denote the vectors of coordinates of, respectively, $\mathbf{t}$ and $\mathbf{t}^*$ in relation to a given orthonormal basis and d_e refers to the ordinary Euclidean distance. In this way, our clustering problem in $\mathcal{S}^D$ can be handle in the real space of coordinates $\mathbb{R}^{D-1}$.

In this paper we focus on the probabilistic FCM which will provide a vector of cluster membership probabilities $\mathbf{u}_i$ for each AC i, and a classification of them into a selected number of clusters. The main difference with respect to the possibilistic approach is that probabilistic algorithms are forced to partition data, that is, the total membership probability, 1, is distributed across all the clusters. Thus, the membership probability of an AC to a cluster depends not only on the distance between them, but also on the distance between that AC and the other clusters.

From all the above, the minimization problem (3) is now written as

$$\min_{\mathbf{U},\mathbf{v}} \sum_{k=1}^{c} \sum_{i=1}^{n} u_{ik}^m d_e^2(\mathbf{y}_i, \mathbf{v}_k),$$

where $\mathbf{v}_k$ denotes the centroid ν_k in the space of coordinates. Notice that such an objective function cannot be minimized directly. Both the cluster centroids, $\mathbf{v}_k$, and the membership probabilities, u_{ik}, are then obtained by an iterative process. At the t-step they are calculated as follows:

$$\mathbf{v}_k^{(t)} = \frac{\sum_{i=1}^{n} u_{ik}^{m,(t-1)} \mathbf{y}_i}{\sum_{i=1}^{n} u_{ik}^{m,(t-1)}}, \quad k = 1, \ldots, c,$$

and

$$u_{ik}^{(t)} = \frac{1}{\sum_{j=1}^{n} \left(\frac{d_e(\mathbf{y}_i, \mathbf{v}_k^{(t)})}{d_e(\mathbf{y}_j, \mathbf{v}_k^{(t)})} \right)^{2/(m-1)}}, \quad i = 1, \ldots, n, \ k = 1, \ldots, c.$$

At the first iteration the membership probabilities are obtained from an initial set of centroids. These could be either selected at random or determined according to prior information about the groups structure. Note that the probabilistic FCM is quite insensitive to initial values and exhibits good convergence properties. Once the algorithm converges, the commonly-used criterion is to classify an object—an AC in our case—into the cluster with highest membership probability.

3 Application to regional time use pattern in Spain

3.1 The data and the problem of zero observations

In 2000 the European Statistical Office published the guidelines for developing the Harmonized European Time Use Surveys system. Under such a unified framework, the first Time Use Survey (TUS) of national scope was conducted in Spain during 2002–03. Time use data are collected by the TUS, along with many other background variables, through a daily activities diary. Originally, this diary concerns the activities collapsed in 10 primary activity groups exhaustive of all daily activities. Respondents record its activity category in 10 minutes time slots during a day—144 time slots in total. In this way the daily time allocated to each one of the activity groups could be obtained. Rather than deal with the large sample size of the complete TUS, attention is restricted to the subsample $\mathbf{T}$ of occupied working population (employees or self-employees) aged between 18 and 65—respectively the age of majority and the legal age of retirement in Spain—with secondary education or higher. So the size of the valid subsample is 6465 individuals. Note that this subgroup of population has a significant bearing on the determination of economic growth and distribution of national income inasmuch as they supply the most labour hours to the market. Thereby, the study of their time allocation decisions is of some importance.

A problem arises if individuals write down zero for any activity group. The presence of zeros prevent us for applying compositional data methods based upon ratios. Unfortunately in microeconomic data, such as collected in expenditure surveys, in panel studies or also in time use surveys, the occurrence of zero observations for shares is common. In fact, we have evidence that some time use investigators have moved away from the log-ratio approach because of the zeros problem. Under our point of view, investigating the reasons for the presence of zeros and applying a reasonable procedure to remove them, which allows for conducting a coherent analysis of the data, is worth more than using a misleading methodology.

In this paper we face the problem in two ways. On the one hand, the 10 primary activity groups are aggregated—amalgamated in compositional terminology—into 6 major categories:

1. *Personal care*: including feeding and sleeping time or some other basic personal care activities.
2. *Work*: including both remunerated and volunteer work (collaboration in organizations, informal casual work for family business, or similar). In this category we also include time to education as an investment in human capital to get better position, higher wages, etc.
3. *Household and family care*: including cooking, childcare, household management, and so on.
4. *Social life and entertainment*: this includes leisure time for sports, outdoors activities, hobbies, interests, etc.
5. *Mass media*: time for radio, television, newspaper, etc.

6. *Travel and unspecified time use.*

By the above amalgamation, the number of zeros is reduced whereas relevant time allocation information is retained. However, one must be cautious in using amalgamation since an excessive aggregation of shares might result in an unmeaningful data analysis. Table 1 contains the percentage of zero values for each of the new categories considered and also the overall percentage of zeros. We can see that this latter is moderate (11.1%), and also that groups 1 and 2 do not contain zero observations.

Table 1: Zeros in Spanish Time Use Survey microdata.

Activity group	% zeros
1. Personal care	0%
2. Work	0%
3. Household and family care	20.8%
4. Social life and entertainment	25.3%
5. Mass media	17.8%
6. Travel and unspecified time use	2.7%
Overall % zeros	11.1%

On the other hand, an imputation procedure is applied for the zeros remaining. We assume that these zeros are not truly zeros, but they result from the sampling process. For instance, it is the case of an individual who has allocate less than 10 minutes to a certain activity group. Several compositional zeros replacement procedures has been proposed in the literature (see Martín-Fernández and others, 2003; Palarea-Albaladejo and others, 2007; and Palarea-Albaladejo and Martín-Fernández, 2008). Here we apply the maximization-restoration method proposed in Palarea-Albaladejo and others (2007) by which zeros are replaced by estimated values taking into account the threshold value—10 minutes in our case, the same for all the activity groups—and the valuable information contained in the observed values. Only 16 iterations have been required by the algorithm to converge.

3.2 Finding homogeneous time use patterns

Firstly, recall that the analysis is focused on ACs, not on individuals, so a measure of their typical time allocation pattern is required. The mean vector is commonly used as such, but it is now well-known that the ordinary arithmetic mean is not well-behaved when it is applied to data of compositional nature. Instead, the *closed geometric mean* is proposed (Aitchison, 2001; Pawlowsky-Glahn and Egozcue, 2001b):

$$g(\mathbf{T}_h) = \mathcal{C}\left(\left[\prod_{i=1}^{n_h} t_{hi1}, \ldots, \prod_{i=1}^{n_h} t_{hiD}\right]^{1/n_h}\right), \quad h = 1, \ldots, 17,$$

where h indicates the subsample of $\mathbf{T}$ refers to every one of the 17 ACs[1] into which the Spanish territory is divided. The *closure* operator $\mathcal{C}$ divides each component of a vector by the sum of all the components, thus the vector is rescaled so that the sum of its components is T. Therefore, the final data set used for clustering is a data matrix of size 17×6 which comprises the representative time allocation pattern—closed geometric mean time allocation pattern—of the population of every AC among the $D = 6$ activity groups considered above. Note that for the whole country the mean time allocation pattern—expressed in proportion of total available time—is

$$[0.4747, 0.3556, 0.0389, 0.0302, 0.0403, 0.0603].$$

[1]Note that the autonomous cities Ceuta and Melilla have not been considered in this study.

Consequently, the great majority of daily time is dedicated to personal care and work, whereas a similar proportion of time is allocated to home care, leisure activities or media.

In order to apply the fuzzy clustering procedure described in the previous section, it is required to express the data in coordinates with respect to an orthogonal basis in $\mathcal{S}^6$. For our case, no special features of such a basis are demanded, so we will use that one provided by Egozcue and others (2003). Then the corresponding set of coordinates $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{17}\}$ in $\mathbb{R}^5$ is obtained by

$$y_{ij} = \frac{1}{\sqrt{j(j+1)}} \log \frac{\prod_{k=1}^{j} t_k}{t_{j+1}^{j}}, \quad j = 1, \ldots, 5, \;\; i = 1, \ldots, 17.$$

In clustering algorithms, it is usually assumed that the number c of clusters is known or given. The automatical determination of the optimal number of cluster is not a trivial task, and a large number of strategies has been proposed from different points of view. A simple procedure is to try with several values for c in order to discover a stable classification of objects. Then we have chosen a range of values for c from 3 to 6, and a set of stable homogeneous groups of ACs has been observed. In the following, we take $c = 4$ as representative of the underlying cluster structure. In this case the FCM algorithm converges in 14 iterations.

Table 2: Cluster membership probabilities for the Spanish Autonomous Communities.

	Cluster membership probability				
ACs	**1**	**2**	**3**	**4**	$d_a(\mathbf{t}_i, \nu_k)$
1. Andalucía	0.1280	0.5944	0.2104	0.0671	0.1332
2. Aragón	0.7885	0.0732	0.0945	0.0438	0.1151
3. Asturias	0.8165	0.0839	0.0699	0.0296	0.0914
4. Illes Balears	0.1664	0.4652	0.1914	0.1770	0.3414
5. Canarias	0.0874	0.1021	0.0471	0.7633	0.4864
6. Cantabria	0.2545	0.1970	0.4175	0.1310	0.7672
7. Castilla-León	0.2605	0.1749	0.5213	0.0432	0.4721
8. Castilla-La Mancha	0.1374	0.6632	0.1351	0.0642	0.1396
9. Cataluña	0.1091	0.1586	0.0707	0.6616	0.7048
10. Com. Valenciana	0.0988	0.7441	0.0915	0.0656	0.1294
11. Extremadura	0.2754	0.2182	0.4280	0.0785	0.8852
12. Galicia	0.0698	0.8348	0.0589	0.0365	0.0720
13. Madrid	0.0410	0.0490	0.0273	0.8827	0.6611
14. Murcia	0.8902	0.0434	0.0422	0.0241	0.0691
15. Navarra	0.0227	0.0244	0.9476	0.0053	0.6189
16. País Vasco	0.0419	0.0489	0.8977	0.0115	0.5631
17. La Rioja	0.5908	0.2209	0.1391	0.0492	0.1358

Table 2 shows the membership probabilities for every AC. There are ACs which clearly belongs to a certain cluster—as AC 15 to cluster 3—whereas this does not occur with ACs such as AC 6—due to its *low* highest membership probability. Another possible problem for clearly classifying an AC—which we have encountered for $c = 6$—is that of a highest and a second highest membership probabilities which were very close between them. Table 2 also includes in the last column the Aitchison distance between each AC and the centroid of its belonging cluster, and Figure 1 shows it graphically. It is observed that clusters 1 and 2 are the most homogeneous ones. For those ACs classified in the same cluster, one could expect an inverse relationship between both the magnitude of the highest membership probability and the distance from the corresponding centroid. In fact, this occurs for the members of clusters 1 and 2, but this is not true in all the cases. For example, Navarra is classified in cluster 3 with probability 0.9476—the higher one among those ACs in the same cluster—however the distance between Navarra and the centroid of cluster 3 is not the lowest one. This is due to the fact that membership probabilities are calculated taken into account the position of an AC with respect to all the clusters. Then a membership probability cannot be understood as a measure of typicality as might appear on first acquaintance.

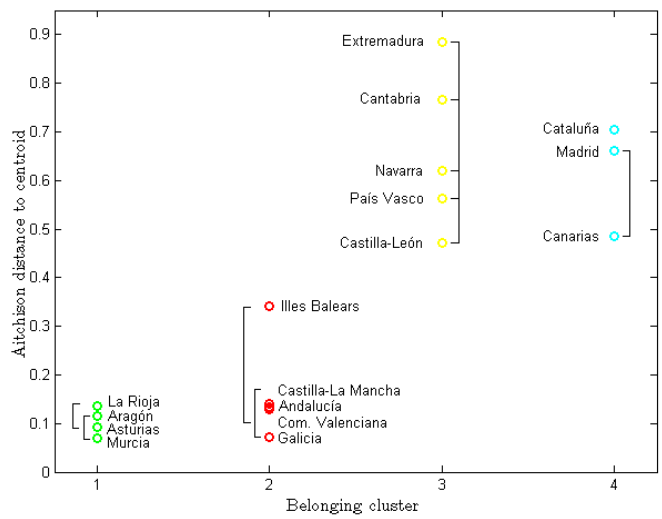

Figure 1: Aitchison distance between Autonomous Communities and cluster centroids.

Given the above membership probabilities, Figure 2 shows a map of Spain wherein ACs belonging to each cluster—ACs with similar time allocation pattern—are distinguished by different colors. There is not observed any definite geographical pattern—such as a north-south or coast-inland pattern—in the way the available time is distributed among activity groups. We finally highlight that our tests with several values of c have revealed some strong links between ACs in the sense that they are always clustered together. Such links have been represented in Figure 1 by vertical lines connecting those highly related ACs. Note that all the members of cluster 3 are highly related each other.

4 Concluding remarks

Time allocation data give a complete picture of a society by providing detailed information about how people use their time on different market and non-market activities. From a public policy point of view, the analysis of time use data could say much on the socioeconomic profile of a population, therefore a well-founded and well-conducted analysis is a must. Nevertheless, although compositional in nature, time allocation data are rarely analyzed using compositional data statistical methods.

In this work we illustrate how to adapt fuzzy clustering techniques to deal with compositional data by using microdata from the Spanish Time Use Survey. After treat the data for zero observations, a set of homogeneous groups of Spanish Autonomous Communities is obtained. The methodology is successfully applied and interesting information can be inferred from the results.

From an applied point of view, the immediate continuation of the work presented here consists in finding out the socioeconomic reasons underlying the daily time use patterns that have been discovered. Also, to analyze how that allocation might be affected by variables such as sex, age, day of the week or similar.

Figure 2: Homogeneous Spanish Autonomous Communities in terms of time allocation pattern.

Acknowledgements

This work has received financial support from the Dirección General de Investigación of the Spanish Ministry for Science and Technology through the project MTM2006-03040.

References

Aitchison, J. (1986). *The statistical analysis of compositional data.* London: Chapman & Hall. Reprinted in 2003 by Blackburn Press.

Aitchison, J. (1992). On criteria for measures of composition difference. *Mathematical Geology 24*, pp. 365–379.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A., Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. *Mathematical Geology 32*, pp. 271–275.

Aitchison, J. (2001). Simplicial inference In M.A.G. Viana, D.S.P. Richards (Eds), *Algebraic Methods in Statistics and Probability*, Contemporary Mathematics Series, v. 287, pp. 1–22. American Mathematical Society: Rhode Island.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A., Pawlowsky-Glahn, V. (2001). Reply to letter to the editor by S. Rehder and U. Zier on 'Logratio analysis and compositional distance' by J. Aitchison, C. Barceló-Vidal, J. A. Martín-Fernández and V. Pawlowsky-Glahn. *Mathematical Geology 33*, pp. 849–860.

Aitchison, J., Barceló-Vidal, C., Egozcue, J.J., Pawlowsky-Glahn, V. (2002). A concise guide for the algebraic-geometric structure of the simplex, the sample space for compositional data analysis. In *Proceedings of IAMG'02 — The Eighth Annual Conference of the International Association for Mathematical Geology*, pp. 387—392, Berlin.

Barceló-Vidal, C., Martín-Fernández, J.A., and Pawlowsky-Glahn, V. (1999). Comment on 'Singularity and nonnormality in the classification of compositional data' by G.C. Bohling et al. *Mathematical Geology 31*, pp. 581–586.

Becker, S.G. (1965). A theory of allocation of time. *Economic Journal 75*, pp. 493–517.

Bezdek, J. (1981). *Pattern recognition with fuzzy objective function algorithms.* New York: Plenum Press.

Billheimer, D., Guttorp, P., Fagan, W. (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association 96*, 1205—1214.

Bohling, G.C., Davis, J.C., Olea, R.A., Harff, J. (1998). Singularity and nonnormality in the classification of compositional data. *Mathematical Geology 30*, pp. 5–20.

Converse, P.E. (1972). Country differences time-use. In A. Szalai (Ed.), *The Use of Time: Daily Activities of Urban and Suburban Population in Twelve Countries*, pp. 145–177. The Hague: Mouton.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35*, pp. 279–300.

Greene, W. (2008). *Econometric analysis.* 6th Edition, Prentice Hall.

Hamermesh, D., Pfann, G.A. (Eds.) (2005). *The economics of time use.* Amsterdam: Elsevier.

Juster, F.T., Stafford, F.P. (1991). The allocation of time: empirical findings, theoretical models, and problems of measurement. *Journal of Economic Literature 29*, pp. 471–552.

Klevmarken, A. (1998). Microeconomic analysis of time-use data: did we reach the promised land? In J. Merz and M. Ehling (Eds.), *Time Use – Research, Data and Policy*, FFB-Schriftenreihe Band 10, Nomos Verlagsgesellschaft, Baden-Baden, pp. 423–456.

Krishnapuram, R., Keller, J. (1993). A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems 1*, pp. 98–110.

Martín-Fernández, J.A., Barceló-Vidal, C. and Pawlowsky-Glahn, V.(1998). Measures of difference for compositional data and hierarchical clustering methods. In A. Buccianti, G. Nard and R. Potenza (Eds.), *Proceedings of the Fourth Annual Conference of the International Association for Mathematical Geology*, pp. 526–531, Naples, Italy.

Martín-Fernández, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets. *Mathematical Geology 35*, pp. 253–278.

Michelson, W. (2006). *Time use: expanding explanation in the social sciences.* Paradigm Publishers.

Palarea-Albaladejo, J., Martín-Fernández, J.A., Gómez-García, J. (2007). A parametric approach for dealing with compositional rounded zeros. *Mathematical Geology 39*, pp. 625–645.

Palarea-Albaladejo, J., Martín-Fernández, J.A. (2008). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences*, (in press).

Pawlowsky-Glahn, V., Egozcue, J.J. (2001a). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment 15*, pp. 384—398.

Pawlowsky-Glahn, V., Egozcue, J.J. (2001b). BLU estimators and compositional data. *Mathematical Geology 34*, pp. 259–274.

Pawlowsky-Glahn, V. (2003). Statistical modelling on coordinates. In *Compositional Data Analysis Workshop – CoDaWork'03 Proceedings*, Girona, Spain.

Pentland, W.E., Harvey, A.S., Lawton, M.P., McColl, M.A. (Eds.) (2002). *Time use research in the social sciences.* Springer.

Rehder, S., U. Zier (2001). Comment on 'Logratio analysis and compositional distance' by J. Aitchison, C. Barceló-Vidal, J.A. Martín- Fernández and V. Pawlowsky-Glahn. *Mathematical Geology 33*, pp. 845–848.