Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

MF CALCULATOR: A Web-based Application for Analyzing Similarity

Célia M. D. Sales
Universidade Autónoma
de Lisboa (CIP / UAL),
CIS ISCTE-Instituto Universitário
de Lisboa (CIS / ISCTE-IUL)

Peter P. Wakker
Erasmus University

Paula C. G. Alves
CIS ISCTE-Instituto Universitário
de Lisboa (CIS / ISCTE-IUL)

Abstract

This paper presents the Metric-Frequency Calculator (MF Calculator), an online application to analyze similarity. The MF Calculator implements the MF similarity algorithm developed by Sales and Wakker (2009) for the quantitative assessment of similarity in ill-structured data sets. It is widely applicable as it can be used when there is little prior control of the variables to be observed, with regard to either their number or their content (qualitative or quantitative). A simulated example illustrates the implementation of the MF Calculator. An example with real data (n=150) of food consumer communication behavior in social media is also presented, in order to illustrate the potential of combining the MF Calculator with further statistical analysis. The MF Calculator is a user-friendly tool available free of charge. It can be downloaded from http://mfcalculator.celiasales.org/Calculator.aspx, and it can be used by non-experts from a wide range of social sciences.

Keywords: Metric-Frequency Calculator, metric similarity, feature similarity, software implementation, ill-structured data, metric distance, additive similarity trees.

1. Introduction

The measurement of similarity is important in many domains. For example, in perception, as studied in psychology, linguistics, and computer science, specific units are identified as a concept (table, horse) whenever they are sufficiently similar to a prototype (Paclik, Novovicova, and Duin 2006). In case-based decision theory, options are evaluated by the similarity-weighted average utility of their predecessors (Gilboa and Schmeidler 2001). Plagiarism accusations are based on similarity assessments (Lutz, Malpohl, and Philippsen 2003). Similarity is also central in categorizations (Sneath and Sokal 1973). Its use is so widespread that Ashby and Ennis (2007) wrote: "A review, or even a listing of all the uses of similarity is impossible."

Measures of similarity have been studied almost exclusively for well structured data sets, where the number of variables is specified a priori. However, in social sciences, data sets are often ill-structured. This is typically the case when people are given open-ended instructions and thus come up with items that are not possible to foresee in advance. We often compare subjective judgments, preferences, or emotional states where the type and number of alternatives is unknown beforehand, and can be diverse. In such situations, an approach that limits judgments to the same a priori set of items for every subject is not appropriate.

To deal with this issue, Sales and Wakker (2009) introduced a new theoretical measure for quantifying similarity, the Metric-Frequency measure (MF). They applied the MF to a psychotherapy study, comparing complaints of family members (Sales 2005; Sales and Wakker 2009, 2010, 2011). In this situation the issues raised by the members were unpredictable, which gave rise to the development of the measure. More recently, the MF has been integrated into the Individualized Patient Progress System (IPPS) (Sales and Alves 2012), a web-based monitoring system for psychological treatments that allows clinicians to compare patients based on their personal complaints. This system is currently being tested with family therapists, psychodramatists and drug addiction group therapists (Sales, Alves, Evans, Elliott, and Wakker 2011).

To facilitate the application of the MF by other researchers, a user friendly and effective tool is required. This paper presents such a tool, the MF Calculator, which is now freely available as a web application.

1.1. The MF

One approach for measuring similarity is a metric (or dimensional) approach, in which the variables can take a range of numerical values, and a metric distance measure is used to specify the similarity. Another approach is a qualitative (or featural) approach, used when the values are dichotomous (present or absent), and alternative measures have been developed (Tversky 1977). Carroll (1976) recommended that models of similarity should combine both metric and qualitative aspects, because the perception of similarity is a complex mental process that usually involves both aspects. Navarro and Lee (2003) first proposed such a model. They coded qualitative items as 1 (present) or 0 (absent), and could then apply metric distances, subsequently used in maximum likelihood fittings of data. The Metric Frequency measure (MF), defined below, combines metric and qualitative aspects in a different manner, designed to incorporate another generalization.

We begin with an example to illustrate the MF. Say two persons, A and B, are independently presented with a picture and are asked to list the emotions (items) associated with this picture and to rate the intensity of each emotion. As each person is free to raise any item according to his subjective evaluation, the items in the data set are unpredictable: there can be many emotions raised by one person and not by the other. The number of items raised by A and B can also differ. For instance, A may associate only 1 emotion with the picture and B may associate 7 emotions.

Suppose we want to compare A and B. A metric approach would compute the difference of scores in the items that both persons raise. However, there are additional aspects to be considered in the comparison, which are typically carried out by featural approaches.

First, when both people mention the same item, then this in itself is a signal of their similarity, while items raised by one person and not by the other in themselves reflect a difference between the two people (Tversky 1977). Second, the number of items raised also provides information about similarity. In the example above, person B raises 7 items, and person A raises only 1 item. This difference in itself entails a difference between the two people (Tversky 1977). For an efficient application of such frequency similarities in linguistic studies, see Maki, Krimsky, and Munoz (2006). Such information is not accounted for in traditional approaches.

The MF combines the metric component and the frequency component. This makes it suitable for situations where the number of aspects is unpredictable, and thus also contains information. For another efficient method to measure similarity, based on two-dimensional sorting on a computer screen, see Goldstone (1994).

The MF consists of the metric similarity, based on differences of the scores on joint items (the lower the differences, the more similarity there is), which is given by Eq. 1 below; and the frequency similarity (the more similar the number of items raised by both persons, the more similarity there is), given by Eq. 2; the MF formula is shown below in Eq. 3.

$$\frac{\sum(1-|\mathit{diff}|)}{j+f+m} \tag{1}$$

$$\frac{\sqrt{j/N}+1-\left|\sqrt{f/N}-\sqrt{m/N}\right|}{2} \tag{2}$$

$$\frac{1}{2}\frac{\sum(1-|\mathit{diff}|)}{j+f+m}+\frac{1}{4}+\frac{1}{4}\sqrt{j/N}-\frac{1}{4}\sqrt{f/N}-\sqrt{m/N} \tag{3}$$

Note:

Summation Σ in Eq. 1: over all items raised by either person A or person B[1].

$|diff|$: absolute value of the difference in 0-1 normalized scores that the two persons assign to the item under consideration, with $1-|diff|$ the resulting similarity.

j: the number of (*joint*) items raised by both person A and person B.

f: the number of items raised by person A and not by person B.

m: the number of items raised by person B and not by person A.

N: an upper bound for the number of items that can be raised by one person (explained later).

The score similarity (or metric similarity) proceeds as in most metric approaches to similarity measurement. Scores should be at an interval scale level, so that differences are meaningful. Some items may be raised by one person and not by the other. Accordingly when an item is not raised, it is entered as zero in the summation and this is taken as the minimum score. We

[1] Or, equivalently, over all conceivable items because those not raised will all contribute (0 - 0 =) zero to the summation.

thus assume no negative scores, and the items are uni-directional. For incorporating negative scores, see Sales and Wakker (2009).

The scaling of absence as 0, and its difference with the minimal score of items if present, should obviously agree with the other score levels and their differences. Those levels should therefore be chosen deliberately by the researchers when scoring the data (Sales and Wakker 2009). We next normalize to a 0-1 scale by dividing the scores of each item by the maximum of their range.

Then for each item we compute the difference between the scores that the two people give to this item, *diff*. The similarity is $(1 - |diff|)$ and the score similarity is the sum, $\Sigma(1 - |diff|)$ over the total number of items raised, $j + f + m$.

For the frequency similarity, we normalize the frequencies by dividing by N. We usually take N equal to the maximum number of items raised by any person in our sample. Then the frequencies are normalized to a 0-1 scale, and the frequency similarity is weighted likewise as the score similarity. Sometimes N can be chosen larger than the mentioned maximum. Then the frequency score is more compressed around 0.25, and is bounded away from the minimum 0 and the maximum 0.5. The overall effect is that the frequency similarity then has less influence on the MF. Such a procedure is appropriate if a researcher feels that the information contained in the observed frequencies is less reliable than that contained in the scores and, hence, should have less weight.

In principle, N could also be taken below the maximum frequency observed, in which case the frequency similarity would impact the MF more than the score similarity. The MF could then take negative values, but could be rescaled to a 0-1 scale. However it is not likely that the frequency similarity is more reliable than the score similarity. Hence we recommend taking N equal to the maximum frequency as an objective default, with N taken as larger if the researcher has reasons to give less weight to the frequency similarity. In general, the relative importance of the frequency information relative to the score information should be determined by researchers who know the context of the application. For this reason we leave liberty to researchers to choose the parameter N larger than the maximum frequency or, in exceptional cases, to choose it smaller.

By dividing all frequencies by N, we normalize results, ensure j/N, f/N, and m/N never exceed 1, and we properly weight the score similarity viz-a-viz the frequency similarity. In our example, B was the person raising most items, i.e., 7 emotional states. Therefore, N = 7. Instead of the number j/N, the frequency similarity model uses its square root, $\sqrt{j/N}$. This transformation is curved downwards (concave), meaning that similarity increases less for high values of j (and j/N) than for low values. For instance, if persons A and B list completely different emotional states ($j = 0$), and if then they both raise 1 same emotion, then this increase of 1 item ($j = 1$) has more impact than if there are already many items in common (concerning an increase from, say, $j = 18$ to $j = 19$). Such an evaluation is plausible. Following the same rationale, the square-root transformation is also applied to f and m.

Other concave transformations could obviously be considered. The square-root transformation, steep at the minimum 0, and with moderate derivatives in between that do not vanish at the maximum 1, fits this application well. The steepness at 0 captures the categorical difference between no and some overlap, and gives satisfactory results in applications.

The frequency similarity is the average of the similarity due to the number of items that both people raise, and the similarity due to the difference in the number of items raised by person A and person B. Finally, the MF results as the midpoint of the score similarity and the frequency similarity: $\frac{1}{2}\frac{\sum(1-|diff|)}{j+f+m}+\frac{1}{2}\frac{\sqrt{j/N}+1-\left|\sqrt{f/N}-\sqrt{m/N}\right|}{2}$, which can be re-written as Eq. 3.

2. Handling the MF Calculator

We now turn to the explanation of the MF Calculator. The MF is computed online at http://mfcalculator.celiasales.org/Calculator.aspx. In brief, the user prepares the input database in the CSV format, uploads it to the website, chooses between the available analyses (i.e. score similarity, frequency similarity, or overall similarity), and obtains the results on screen. Output files are provided in three formats (CSV, XLS and DOC), and can be used in subsequent analyses. To ensure confidentiality and data protection, all data and outputs generated by this website will not be stored automatically in any server. When the user closes the internet browser, everything is lost and cannot be retrieved.

2.1. Preparing input database

Our hypothetical example involves three individuals who were asked to identify their favourite colours and rate how much they liked them on a 7-point scale, ranging from 1 to 7 (Table 1). We want to compare their answers: how similar are they when it comes to their colour preferences?

The MF Calculator supports input databases in the CSV format with the following structure: a) Rows represent cases to be compared and b) columns represent items or variables. Cells display scores. Table 2 displays our data. Whenever an item was not raised by an individual, the score is 0. The score 0 should fit with the scale, which it does here. The MF calculator handles databases containing up to 500 cases. CSV files are generated by saving the database as a CSV file (*comma-separated values*).

Alternatively, we may opt for using a MFC macro which allows the selection of a few cases/variables for comparison only. This MFC macro is available on our website for MS Excel (versions 2007 and 2010) and SPSS and must be installed in the computer with administration privileges. After installation, a MFC icon will appear automatically in the program toolbar and the cases / variables for comparison are ready to be selected using the Ctrl key. The input database must be saved as a MFC file, by clicking on *Export to MFC* before uploading into the MF Calculator website. The MFC macro will remain permanently in the computer until uninstalled (Control Panel > Add / Remove Programs).

Person A		Person B		Person C	
Item	Score	Item	Score	Item	Score
Green	7	Brown	6	Green	7
Blue	6	Orange	6		
Red	4	Yellow	2		
Pink	3				

Table 1: Colour preferences of persons A, B and C, rated by intensity of preference

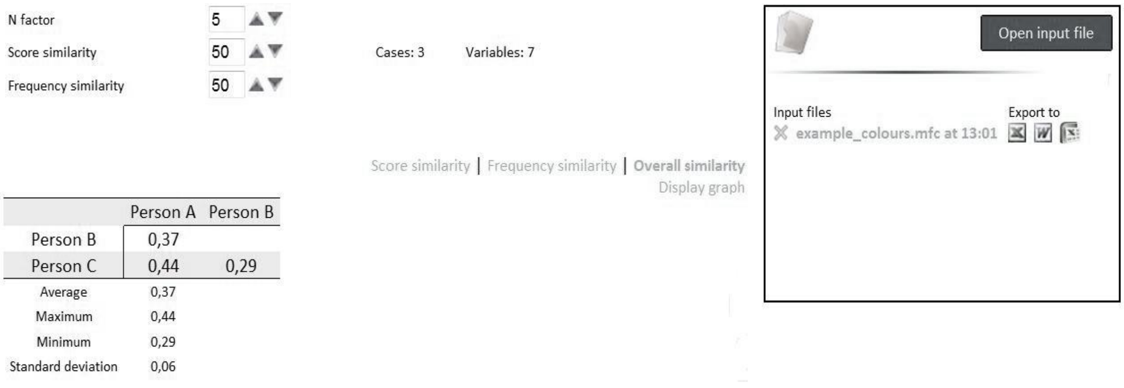
	Green	Blue	Red	Pink	Brown	Orange	Yellow
Person A	7	6	4	3	0	0	0
Person B	0	0	0	0	6	6	2
Person C	7	0	0	0	0	0	0

Table 2: Input database for MF Calculator

2.2. Calculating the MF Similarity

The MF similarity matrix is computed in the Calculator section of the website. To do so, it is necessary to upload the input database and to determine the N factor. In our example, person A raised the highest number of colours (4 colours), and N $\geq$ 4 is natural. For illustrative purposes let us assume that the mere presence or absence of colours is considered to be somewhat more coincidental and less informative than the scores given to the colors. We hence choose N = 5. With the database imported and the N factor chosen, the MF is ready to be calculated by clicking in the input database file.

The MF Calculator generates several outputs: 1) similarity matrixes (overall similarity, score similarity and frequency similarity) in DOC, XLS and CSV formats; 2) descriptive statistics (minimum similarity, maximum, mean and standard deviation); and 3) the graphical representation of results, using additive similarity trees (Sattath and Tversky 1977) (see Figures 1 and 2). Nodes in the tree represent the cases and the length of the path joining them represents their proximity. These similarity trees produced by the MF Calculator are recommended for visualization of samples of up to 50 cases. For larger samples, we recommend



N factor 5

Score similarity 50

Frequency similarity 50

Cases: 3 Variables: 7

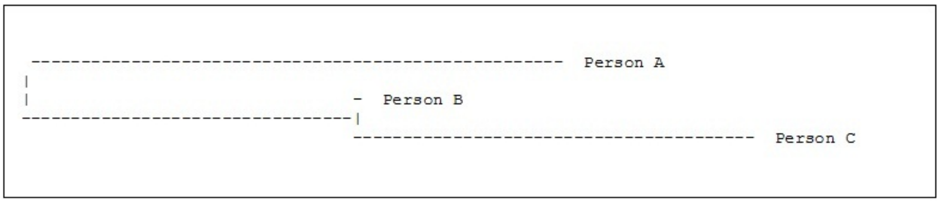
Input files

example_colours.mfc at 13:01

	Person A	Person B
Person B	0,37	
Person C	0,44	0,29
Average	0,37	
Maximum	0,44	
Minimum	0,29	
Standard deviation	0,06	

Open input file
Export to
Score similarity | Frequency similarity | Overall similarity
Display graph

Figure 1: MF similarity between colour preferences (rated by preference) as presented in the MF Calculator



```
 ------------------------------------------------------  Person A
|
|                                -  Person B
---------------------------------|
                                 ----------------------------------------  Person C
```

Figure 2: Similarity tree for persons A, B and C in terms of colour preferences (rated by preference)

downloading the MF output file and using statistical software for graphical representation (explained later).

The MF output files are available for download and must be saved to a local computer or to an external driver. Otherwise all data will be lost once the webpage is closed.

2.3. Using the MF calculator with nominal data

In a qualitative setting where only the presence or absence of items can be observed, the MF Calculator can also be used. Suppose that in the preceding example that individuals were only asked to identify their favourite colours, without rating how much they liked them (Table 3).

Person A	Person B	Person C
Item	Item	Item
Green	Brown	Green
Blue	Orange	
Red	Yellow	
Pink		

Table 3: Colour preferences of persons A, B and C, rated nominally (purely qualitative data)

The database then assigns scores 1 (present) and 0 (absent) to each item (Table 4). As in the previous example, the highest number of identified colours was 4. For consistency we choose N = 5 again, reflecting the same degree of reliability as before.

	Green	Blue	Red	Pink	Brown	Orange	Yellow
Person A	1	1	1	1	0	0	0
Person B	0	0	0	0	1	1	1
Person C	1	0	0	0	0	0	0

Table 4: Input database for MF calculator(1 = present; 0 = absent)

3. Contexts in which the MF Calculator can be applied

The MF Calculator covers a wide diversity of situations. It can be used with nominal, ordinal or interval data. Cases compared can either be people (as in our hypothetical example), or any other entity such as events, days (see our real life example in the next section), and so on.

The MF Calculator allows the comparison of user-generated items, such as contents in social networks (Kaplan and Haenlein 2010), patient-generated items in clinical settings (Sales and Alves 2012), content analysis judgments with no a priori categorization system, or virtually any kind of answers resulting from open-ended instructions. There is no limit to the content or number of resulting outputs.

The similarity matrix resulting from the MF Calculator can serve as a proximity matrix. It can be used this way in further statistical analyses including visualization and scaling (e.g., hierarchical cluster analysis, multidimensional scaling, factor analysis). The CSV format allows its direct use in most statistical software packages.

3.1. Illustration of the MF Calculator with a real example

We present a real life example illustrating the use of the MF Calculator to analyze social media data extracted from Twitter during the E.coli/EHEC food crisis in Europe in May-July 2011. During this period there was an E.coli contamination in Germany which the source was initially identified as Spanish cucumbers (CC), later changing to bean sprouts (BS). At the same time an E. coli outbreak was detected in France and eventually, many weeks after the initial outbreak, it was found to be caused by fenugreek seeds (FC) from Egypt.

In order to analyse consumer perceptions and communication of food risk/benefits during that period, the FoodRisC project (Ref. FP7-KBBE-2009-2-1-02) extracted from Twitter all individual tweets produced in Spanish and sent from a Spanish IP address, using keywords related to E.coli/EHEC as filters (examples: Escherichia coli; Zoonoses; Acute kidney failure; EHEC; VTEC).

A database was created describing the frequency of daily tweets in seven kinds of behaviour concerning consumer food communication (*eat CC*, *eat BP*, *buy CC*, *buy BP*; *other consumer behaviour CC*, *other consumer behaviour BP* and *other consumer behaviour Meat*). We used the MF Calculator to detect patterns of similarity within the 150 days under study (May 1-August 31, 2011).

In the input database, rows reflect the cases to be compared (days), and columns refer to the variables concerning consumer food communication. Each cell gives the number of tweets published daily regarding each consumer food communication behaviour. The database was built in Excel, saved in CSV format, and uploaded to the MF Calculator. N = 8 was used. The resulting MF similarity matrix was downloaded in CSV format and used in SPSS for a multidimensional scaling analysis (proxscal procedure). The original MF similarities were reliably reproduced in a two-dimensional perceptual map (S-Stress = 0.001; DAF = 0.998). The relative position of data points in the map reveals the existence of several days when tweets changed dramatically (see Figure 3).

When we analysed the days' sequences, at least two distinct patterns emerged: tweets began to change on the 45th day and became highly differentiated on the 46th day; they slowly returned to the baseline during the 47th-50th day period. A second highly distinct pattern of changes in tweets occurred in the 65th-69th day period. Its distribution in the perceptual map reveals that the communication pattern is clearly different from the previously identified pattern.

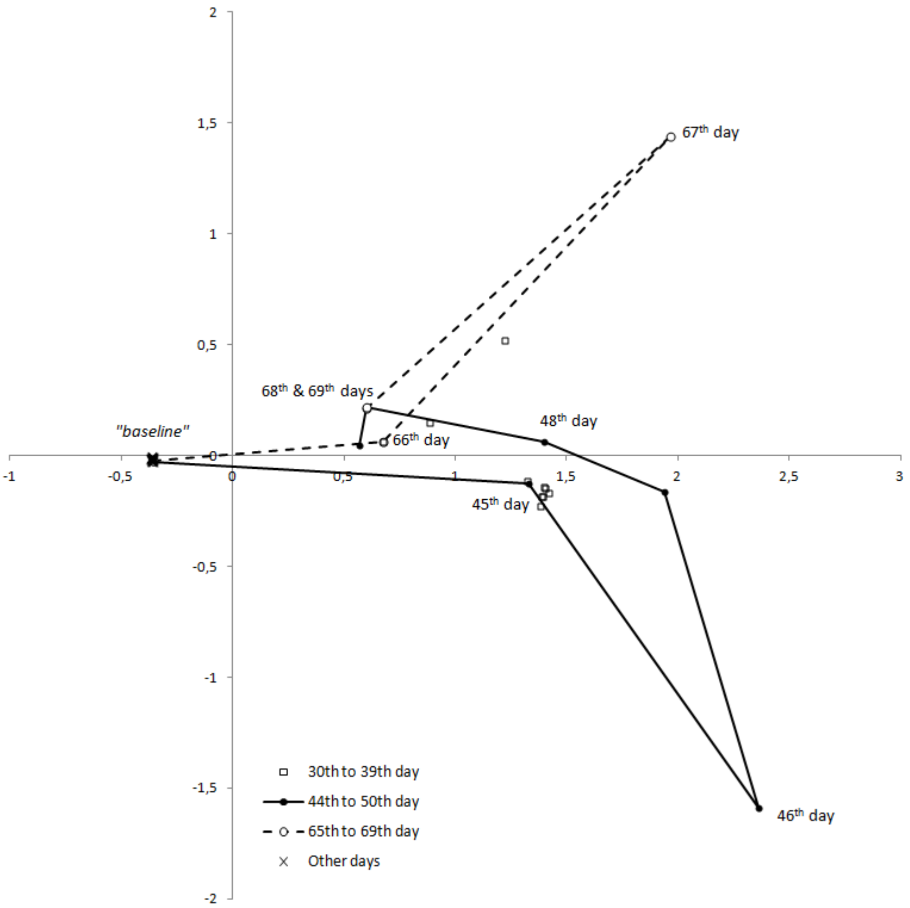


Figure 3: Two-dimensional representation of MF similarity matrix resulting from MDS analysis (FoodRisC project)

In conclusion, the combined use of MF similarity and multidimensional scaling revealed that the communications between consumers in the social media during the food crisis changed over time and in its nature. Further analysis can explore the relationship between these changes and, for instance, the pattern of published news related to the contamination.

4. Conclusion

The MF Calculator is the first implementable software for the computation of similarity in complex ill-structured settings where the number and content of variables are unpredictable. It is flexible and can be applied to both metric and qualitative variables. It, thus, can be widely applied in behavioral and social sciences. It greatly enhances our possibilities to measure and analyse similarity, which is a central concept in numerous fields.

5. Acknowledgments

The development of the software, as well as the preparation of this manuscript, was undertaken with the support of funds allocated by the Portuguese Foundation for Science and Technology (Grant nr. PTDC/PSI-PCL/098952/2008).

The social media data analysed in this article was collected as part of Workpackage 2 of an undergoing project funded by the European Commission under the 7th Framework Programme - FoodRisC: Perceptions and communication of food risk/benefits across Europe (Ref. FP7-KBBE-2009-2-1-02 http://www.foodrisc.org). We acknowledge the valuable input by the project partners involved in this task.

We also thank Luís Faísca (Departamento de Psicologia, Universidade do Algarve and CBME/IBB, LA), for his consultation on the combination of MF and other statistical analysis, as well as his comments on early versions of the paper.

Additionally, we thank Professor Bruno Maia (Universidade Autónoma de Lisboa) for his consultation in LaTeX software, and to Mr. Kenneth Gunn for his assistance in editing the English text.

References

Ashby FG, Ennis DM (2007). *Similarity measures.* Scholarpedia.

Carroll JD (1976). "Spatial, Nonspatial and Hybrid Models for Scaling." *Psychometrika*, **41**, 439–463.

Gilboa I, Schmeidler D (2001). *A theory of case-based decisions.* Cambridge University Press.

Goldstone R (1994). "An Efficient Method for Obtaining Similarity Data." *Behavior Research Methods*, **26**, 381–386.

Kaplan AM, Haenlein M (2010). "Users of the world, unite! The challenges and opportunities of Social Media." *Business Horizons*, **53**, 59–68.

Lutz P, Malpohl G, Philippsen M (2003). "Finding plagiarisms among a set of programs with JPlag." *Journal of Universal Computer Science*, **8**, 1016–1038.

Maki WS, Krimsky M, Munoz SOL (2006). "An Efficient Method for Estimating Semantic Similarity based on Feature Overlap: Reliability and Validity of Semantic Feature Ratings." *Behavior Research Methods*, **28**, 153–157.

Navarro DJ, Lee MD (2003). *Advances in Neural Information Processing Systems.* Massachusetts Institute of Technology.

Paclik P, Novovicova J, Duin RPW (2006). *Building road sign classifiers using trainable similarity measures.* IEEE ITS.

Sales CMD (2005). *Terapia Familiar en Contexto Psiquiátrico: Aportaciones para la Comprensión del Cambio Psicoterapéutico - Family Therapy in Psychiatric Context: Contributions for Psychotherapeutic Change Comprehension.* Ph.D. thesis, University of Seville.

Sales CMD, Alves P, Evans C, Elliott R, Wakker PP (2011). "IPPS: Finally, a Software to Make Therapists, Clients and Researchers Happy." In *4th Regional Mediterranean Congress* *of the International Association for Group Psychotherapy and Group Processes - Sociedade Portuguesa de Psicodrama.*

Sales CMD, Alves PCG (2012). "Individualized patient-progress systems: why we need to move towards a personalized evaluation of psychological treatments." *Canadian Psychology*, **55**, 115–121.

Sales CMD, Wakker PP (2009). "The Metric-frequency Measure of Similarity for Ill-structured Data Sets, with an Application to Family Therapy." *The British Journal of Mathematical and Statistical Psychology*, **62**, 663–682.

Sales CMD, Wakker PP (2010). "Combining Metric and Qualitative Approach in a Measure of Similarity for Ill-structured Data." In *Proceedings of the XVII Meeting of the Portuguese Association of Data Classification and Analysis, JOCLAD 2010.*

Sales CMD, Wakker PP (2011). "Similarity of Subjective Idiographic Data: the MF Calculator." In *42nd International Meeting of the Society for Psychotherapy Research.*

Sattath S, Tversky A (1977). "Additive Similarity Trees." *Psychometrika*, **42**, 319–345.

Sneath PH, Sokal RR (1973). *Numerical taxonomy: the principles and practice of numerical classification.* W. H. Freeman.

Tversky A (1977). "Features of Similarity." *Psychological Review*, **84**, 327–352.

Affiliation:

Célia Maria Dias Sales
CIS / ISCTE-IUL
Universidade Autónoma de Lisboa
Departamento de Psicologia e Sociologia
Rua de Santa Marta Nr. 47, 3, 1169-023, Lisboa, Portugal
E-mail: celiasales@soutodacasa.org

Peter P. Wakker
Econometric Institute
Erasmus University
P. O. Box 1738, Rotterdam, 3000 DR, The Netherlands
E-mail: wakker@ese.eur.nl

Paula Cristina Gomes Alves
Centro de Investigação e Intervenção Social (CIS)
ISCTE-Instituto Universitário de Lisboa
Avenida das Forças Armadas, Edifício ISCTE-IUL, 1649-026, Lisboa, Portugal
E-mail: paulagomesalves@hotmail.com

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd