A Mapping of Investor’s Risk Profile

Tommaso Lando, Lucio Bertoli-Barsotti

VŠB-Technical University of Ostrava
Department of Finance, Faculty of Economics
Sokolská 33
Ostrava, 70121
Czech Republic
e-mail: tommaso.lando@unicatt.it

University of Bergamo
Department of Management, Economics and Quantitative Methods
via dei Caniana 2
Bergamo, 24127
Italy
e-mail: lucio.bertoli-barsotti@unibg.it

Abstract
In this paper we analyze a questionnaire, aimed to evaluate the investors’ risk profile. The questionnaire focuses on the following fundamental aspects: investor’s knowledge and financial experience; financial objectives; personal predisposition to risk/earn; temporal horizon. The analysis consists in a visualization of the dataset, provided by the new VOS method, and it is based on similarity measures between the individuals’ response patterns. The result is a pure descriptive statistical analysis which does not involve any model assumption.

Keywords: Map, Similarity
JEL codes: C19 C45 C58

1. Introduction and data

In this paper we propose to analyze data from a questionnaire submitted by the Italian bank “UBI><Banca” to its customers. The questionnaire is aimed to investigate about several aspects of the customers’ background and investment preferences, in order to propose suitable financial products for each category of investors. Especially, the questionnaire involved is conform to a particular legislation from the European Parliament called MiFID (Markets in Financial Institute Directive). This law, acknowledged in Italy in 2007, modifies the rules of financial intermediation in order to eliminate the asymmetric information and reach well-functioning markets. So, the main aim of this directive is to protect investors, especially the ones that have a poor knowledge about finance, from taking wrong decisions in terms of financial investments.

The questionnaire, already studied by Caviezel et al. (2012), consists in 37 items, submitted to 19685 individuals: some of these items are focused on the “background” of the investors (Job, Title of study…), others are more specifically aimed to investigate about their risk profile. We propose to analyze a part of this dataset, focusing in particular on the similarities between response patterns of different individuals: for this purpose we use a similarity measure studied by Lando and Betoli-Barsotti (2013). After determining the appropriate similarity measure for this situation, we make use of recent mapping method VOS (Van Eck, Waltman, 2007) just to provide a graphical representation of our results.

We chose only a subset of the original 37 items in the questionnaire, for several reasons. First of all, we believe it is important to focus on the items which are really and directly linked to risk profile: for this reason we considered only the items aimed to evaluate the following fundamental aspects: investor’s knowledge and financial experience, financial objectives, personal predisposition to risk/earn and temporal horizon. Anyway, the more items we consider, the more complicated is to interpret the results, in particular if we want to represent several aspects in a singular bi-dimensional representation. So, to avoid some interference in the results, items which are not strictly related with the aim of our research have been dropped out. This is the final set of items that we consider:

Knowledge and financial experience:

1. Title of study (Polytomous-5 response categories)
2. Experience in financial instruments (Multiple choice-12 response categories)
3. Financial knowledge (Multiple choice-8 response categories)

Financial objectives:

4. Purpose: % of investment for liquidity (Polytomous-5 response categories)
5. Purpose: % of investment for capital growth in medium-long period, with slight fluctuations (Polytomous-5 response categories)
6. Purpose: % of investment for capital growth in medium-long period, with strong fluctuations (Polytomous-5 response categories)
7. Temporal horizon: % of assets for very-short period investments (Polytomous-5 response categories)
8. Temporal horizon: % of assets for short period investments (Polytomous-5 response categories)
9. Temporal horizon: % of assets for medium period investments (Polytomous-5 response categories)
10. Temporal horizon: % of assets for long period investments (Polytomous-5 response categories)
11. Temporal horizon: % of assets for very-long period investments (Polytomous-5 response categories)

As a large part of the selected items (9 over 11) are polytomous, in the next section we propose a similarity measure which can deal with both polytomous and multiple choice data.

2. Similarity formula

Let n be the number of individuals and k the number items. Suppose that items have a polytomous structure (with mutually exclusive response categories). We define the *response pattern* of individual v as:

$$y_v=(y_{v1}, y_{v2},\dots,y_{vk})' \qquad (1)$$

where $y_{vi}\in\{0,1,\dots,m_i\}$, for $v=1\dots n$ and $i=1\dots k$. The integer number m_i is simply the number of response-categories of item i, notice that it does not have to be equal for all the items.

We believe that similarity should be an increasing function of co-occurrence data (Van Eck, Waltman, 2009a). Co-occurrence, in this situation, is simply the number of times when a couple of individuals give the same response to the same item. We can compute co-occurrence C_{vw} between individual v and w in the following way:

$$C_{vw}=\sum_{i=1}^{k} c(y_{vi},y_{wi}), \qquad (2)$$

where:

$$c(y_{vi},y_{wi})=\begin{cases}1 & if \quad y_{vi}=y_{wi}\\ 0 & if \quad y_{vi}\neq y_{wi}\end{cases}, \text{ for } i=1\dots k. \qquad (3)$$

Anyway our idea is to recode the responses dividing item i in m_i sub-items and so obtaining, for any response y_{vi}, the following vector:

$$\bar{x}_{vi}=(x_{vi}(1), x_{vi}(2),\dots, x_{vi}(m_i))', \qquad (4)$$

where:

$$x_{vi}(j) = \begin{cases} 1 & if \quad j = y_{vi} \\ 0 & if \quad j \neq y_{vi} \end{cases}, \text{ for } j=1\dots m_i. \tag{5}$$

So, we can define:

$$C^*{}_{vw} = \sum_{i=1}^{k}\sum_{j=1}^{m_i} c\big(x_{vi}(j), x_{wi}(j)\big) = \sum_{l=1}^{K} c(x^*{}_{vl}, x^*{}_{wl}), \tag{6}$$

where:

$$x^*{}_v = (x^*{}_{v1}, \dots, x^*{}_{vK})' = (\bar{x}'_{v1}, \dots, \bar{x}'_{vk})'; \;\; K = \sum_{i=1}^{k} m_i. \tag{7}$$

So, vector $x^*{}_v$ is the v-th row of a new data matrix, that we call X^*.
It is easy to verify that

- $c(y_{vi}, y_{wi}) = 1$ if and only if $\sum_{j=1}^{m_i} c\big(x_{vi}(j), x_{wi}(j)\big) = m_i$
- $c(y_{vi}, y_{wi}) = 0$ if and only if $\sum_{j=1}^{m_i} c\big(x_{vi}(j), x_{wi}(j)\big) = m_i - 2$

So, the following relation holds:

$$C_{vw} = \frac{C^*{}_{vw} - minC^*{}_{vw}}{2} \tag{8}$$

where $minC^*{}_{vw} = \sum_{i=1}^{k}(m_i - 2)$.
Essentially, the two approaches are equivalent. So, from the original polytomous response patterns y_v and y_v it is possible to get a couple of "dichotomized" response patterns x^*_v and x^*_v. Then, we can simply deal with these vectors using any similarity measure for dichotomous response patterns. In particular, here we use a formula that has been studied by Lando and Bertoli-Barsotti (2013) that is based also on the weights given by the *item marginal sums*. Item marginal sum s_l is simply the sum of the l-th column of data matrix X^*, so, in our situation, it corresponds not only to a particular item, but also to a particular response category. For this reason, it should be more correct to call it *item-category marginal sum*. Anyway, it is preferable to use the following procedure only when the number of response categories is fixed for all the items, otherwise we could have the one item "weigh" more then another, only because its response categories are more. Since the details can be found in the cited paper, below we only provide a brief overview of the procedure for computing this "weighted" similarity measure. From pattern x^*_v we create a vector that we call *conformity distribution* $f^*_v = (f^*_{v1},..,f^*_{vl},\dots,f^*_{vK})$, where, for l=1…K, we have:

$$f^*_{vl} = [\, x^*_{vi}s_l + (1-x^*_{vl})(n-s_l)]/n. \tag{9}$$

According to the total scores s_i we can also define the "most conform" pattern x , which has conformity distribution f defined by:

$$f_l = \max\{\, s_l \,,\, n\text{-}s_l \,\}/n \tag{10}$$

for l=1…K (notice that the "most conform" pattern x could not belong to the dataset, but this is not necessary for our purpose). Then, taking inspiration from statistical *divergence measures* (Ali, Silvey, 1965), we measure the "relative divergence" between f and any distribution f^*_v by:

$$\Phi(f_v^*, f) = \sum_{l=1}^{K} \phi\left(\frac{f_{vl}^*}{f_l}\right) f_l \tag{11}$$

where ϕ is a decreasing and convex function (we take ϕ = ln). For example $\Phi(f_v^*, f) \leq \Phi(f_w^*, f)$ means that f^*_v is "closer" to f respect to f^*_w. This suggests to use, as a "distance" between two conformity distributions f^*_v and f^*_w, the function:

$$d(f_v^*, f_w^*) = |\Phi(f_v^*, f) - \Phi(f_w^*, f)|. \tag{12}$$

Finally, the similarity measure $S(y_v, y_w)$ between patters y_v, y_w (or $x*_v$,$x*_w$), is given by:

$$S(y_v, y_w) = \frac{C^*{}_{vw}}{1+d(f_v^*, f_w^*)^\alpha} \tag{13}$$

for $0 < \alpha$, which is an arbitrary weight (we take α=0.5).

Notice that, if some of the items have a multiple-choice structure (as, in our case, for item 2 and 3) we can also deal with them in the same way, "dichotomizing" and computing the similarity index explained above. Further, the same formula can be employed to measure the similarity between the items.

Now that we have defined a measure that seems to suit to our data, the next step is to explain how to plot the obtained similarities in an appropriate way: for this purpose we make use of a recent method, called VOS, for visualizing similarities between objects. The aim of VOS is to provide a low dimensional (two, in our case) visualization in which the distance between a pair of objects is coherent with their similarity. Objects with high similarity will be located close to each other, objects with low similarity will be located far away from each other. Let there be n objects, and a symmetric n×n similarity matrix $\boldsymbol{S}=(s_{pq})$. The coordinates of object i (used for the visualization in a bi-dimensional space) are contained in an n×2 matrix $\boldsymbol{Y}$. The vector $\boldsymbol{y}_i=(y_{i1}, y_{i2})$ denotes the ith row of $\boldsymbol{Y}$ and indicates the coordinates of object p. The idea of VOS is to find these coordinates by minimizing a weighted sum of the squared Euclidean distances between all pairs of objects, imposing the constraint that the sum of all distances must equal some positive constant. Note that the weight in this minimization is given by the similarities, and the constraint is obviously to avoid solutions where all the objects are located to the same coordinates.

So, the mathematical problem is:

$$\min_y \sum_{p<q} s_{pq} \left\| y_p - y_q \right\|^2 \tag{14}$$

where || . || denotes the Euclidean norm, subject to

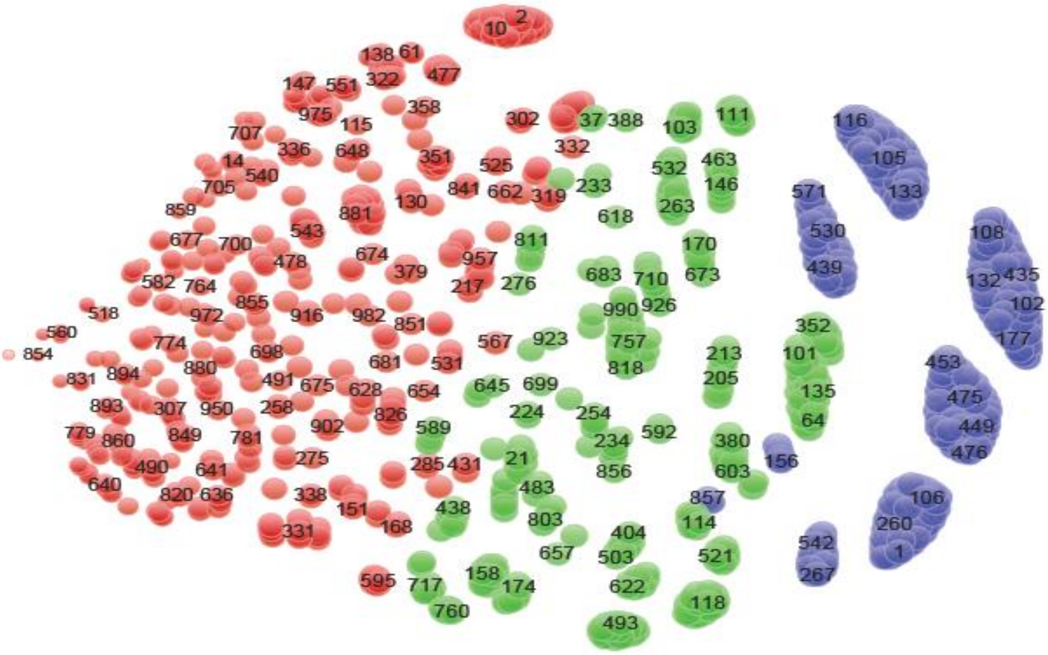
$$\sum_{p<q} \left\| y_p - y_q \right\| = 1. \tag{15}$$

Since the similarity matrix have a central role in this minimization, the final results depend mainly on the chosen similarity measures. Essentially, given a similarity matrix, we can map the whole dataset in a way which is consistent with the similarity measures. In our particular case, it will be possible to visualize similarities and dissimilarities between the bank costumers basing on their background and their investment preferences. In the next section we show the results of this analysis, provided by the computer program VOSviewer (Van Eck, Waltman, 2009b).

3. Analysis

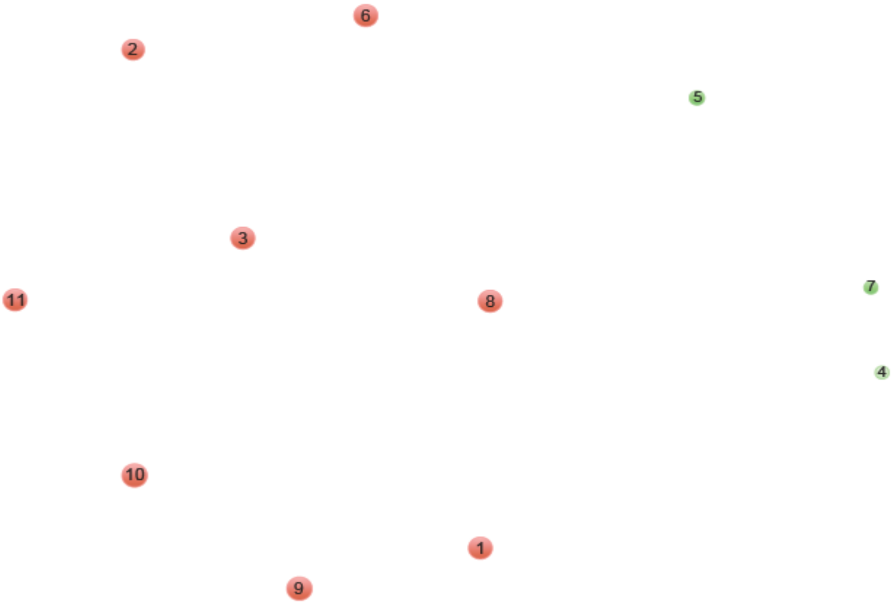
We extracted a random sample of n=1000 individuals from the original dataset and we tried to classify them basing on the 11 items described in section 1. The first step in our analysis consists in an overall visualization of the dataset, basing on all the 11 items. As, considering the whole set of items, some items have more response categories respect to others, the procedure explained in the previous section could give more emphasis to the items with more response categories: for example item 2 (about financial experience) will weigh more than items from 4 to 11 (which are about financial objectives). The idea the we propose to solve this problem is to simply dichotomize the responses for each item, according to the response given respect to the median response category. Essentially, if the response exceed the median value we take 1, otherwise 0. This is surely a "rough" solution that solves the problem at the expense of a significant loss of information. Anyway, the graphical results are represented below in Figure 1.

Figure 1: Similarity computed on items 1-11, “dichotomized” data



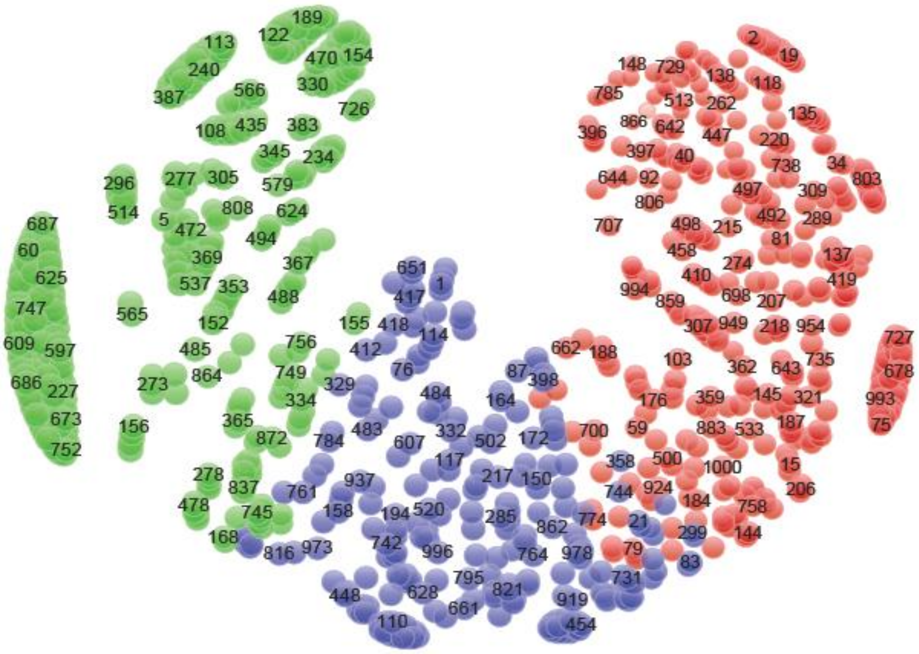
The result showed in Figure 1 is not easy to interpret. First of all, we can see that there is a remarkable clusterization in the plot: this is due to the loss of information caused by dichotomization, which produces a large number of very similar, or identical, patterns. Further, it is difficult to detect the dynamics underneath the data: the main reason for this difficulty can be that we are trying to visualize in a bi-dimensional space an event which involves a higher number of variables and dimensions. In this circumstance, some items could interfere with the others, resulting in a confused and not clear classification. To really understand the relations between items, we applied the similarity analysis also to the items (simply computing similarities using the transposed data matrix X^T), and obtained the following result.

Figure 2: Similarities between items



From Figure 2 it is possible to analyze items about financial objectives: in particular we notice a clusterization where items 4, 5, 7, (about liquidity, very-short period and slight fluctuation investments) that might represent the preferred choices for risk-adverse investors, are slightly separated from items 6, 8 (short-medium period investments) and from items 9, 10, 11 (medium-long period investments). Between these three clusters we can notice items 1, 2 and 3, which are on a totally different dimension, (financial knowledge-experience). This result can be interpreted in the following way: item 3 (financial knowledge) can be similar to item 11 (very-long period investments) as, generally, individuals behave similarly in relation to these two different issues. In particular, this means that if an individual has a poor financial knowledge, he will tend to assign a low percentage of his assets to very long period investments. On the other hand, if his knowledge is good, he will be inclined to invest more. In a similar way, we can interpret the similarity between item 1 (title of study) and item 9 (medium period investment). Anyway, in spite of this logical interpretation, we can't deny that items 1-3 create some are interference with the others. This suggested us to repeat the analysis between the individuals focusing only on their financial objectives, so we will consider items from 4 to 11. Further, since these items have all the same number of response categories (5) we can apply the procedure explained in section 2: this will surely produce a more detailed and deep analysis, since there is no loss of information in this case. We show the results below.

Figure 3: Financial objectives: *S* computed on items 4-11, polytomous data



Although, in this case, we don't consider the investor's knowledge or experience, which can be a drawback, we can be quite sure that the similarity formula works appropriately and accurately and also we can find an interesting way to interpret the data. In particular we found that, in Figure 3, the key for the interpretation is the diversification of investments. The red cluster on the right represent those individuals who prefer to concentrate their investment preferences on a singular and particular direction: as comprehensible these individuals, especially the ones located on the outer edge of the plot, usually chose to assign large part of their assets for liquidity or short-period investments. We can also identify a large set of individuals on the right part of the red cluster: these individuals simply have response patterns that are not acceptable, in particular they declare to assign 0% of their assets for all the possible investment choices, which is mathematically not possible. This underlines that a quite remarkable part of the bank's customers do not understand the questions or simply refuse to answer. After detecting them, we could eventually drop out these individuals form the dataset. In the green cluster, individual generally diversify their investments, in particular the big cluster on the left side of the plot represent those individuals who declare to assign less than 30% of their assets for all the possible investment choices (this is mathematically acceptable). Then, the blue cluster in the middle of the plot represent those individuals whose intentions are simply half-way between the green and the red ones, which means opting for the diversification of investments but also focusing on a particular temporal horizon (generally medium period). Finally, one last remarkable observation is that individuals which are disposed to risk are located in the left part (or the inside part) of the red cluster. In particular, these individuals concentrate their investments but, unlike the other red ones (which focus on liquidity or short-period investments and are located in the right part of the red cluster) they choose to assign large part of their assets for generally long period investments.

4. Conclusion

The proposed method seems to provide interesting results in mapping risk profile investor, especially if the analysis focuses only on the "financial objectives" set of items. It can be used to classify individuals by clustering them in different subsets, detecting, for example, those who give unacceptable responses, those who are disposed to risk, and those who are completely risk-adverse. Another advantage of this procedure is that the same analysis can also be done with respect to the items. This can be helpful, especially for someone who is not expert in finance, to get the correct interpretation and collocation of the items. Therefore, the procedure can be useful to interpret the data from the analyzed questionnaire and represents an alternative approach respect to "traditional" descriptive statistic analysis. Anyway, a drawback of this method is that the interpretation gets difficult when a large number of items, corresponding to different "dimensions" are considered. As a future development of this study, it would be interesting to obtain a new similarity formula for polytomous and (maybe) multidimensional response data, specifically aimed to measure similarities in this particular dataset.

Acknowledgements

This paper has been elaborated in the framework of the project Opportunity for young researchers, reg. no. CZ.1.07/2.3.00/30.0016, supported by Operational Programme Education for Competitiveness and co-financed by the European Social Fund and the state budget of the Czech Republic.

The authors would like to thank the UBI>< Banca group, that made available a sample of the questionnaire responses obtained by their customers, only for research purposes.

References

ALI S. M., SILVEY S. D., (1965). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, Vol. 28, No. 1, pp. 131-142.

CAVIEZEL V., ORTOBELLI S., BERTOLI-BARSOTTI L. (2012). Risk profile versus portfolio selection: a case study, *Proceedings of the XLVI Scientific Meeting*, La Sapienza University of Rome, June 20-22, 2012, CLEUP Editrice, Padova, pp. 1-4.

LANDO T., BERTOLI-BARSOTTI L. (2013). Similarity measures for response patterns on dichotomously scored items, *Proceeding of the 31st Conference on Mathematical Methods in Economics,* Jihlava, Czech Republic, pp. 529-534.

VAN ECK, N.J., WALTMAN, L. (2007). VOS: a new method for visualizing similarities between objects. In Lenz H.-J., Decker R. (eds.), *Advances in Data Analysis: Proceedings of the 30th Annual Conference of the German Classification Society*, pp. 299-306, Springer.

VAN ECK, N.J., WALTMAN, L.. (2009a). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology* **60**, pp. 1635-1651.

VAN ECK, N.J., WALTMAN, L. (2009b). VOSviewer: A computer program for bibliometric mapping. In: Larsen, B., Leta J. (eds.), *Proceedings of the 12th International Conference on Scientometrics and Informetrics*, pp. 886-897.