Commentary on "Targeting with machine learning: An application to a tax rebate program in Italy"*

B. Kaan Basdil† Nikolaos Papoulias‡

Toulouse School of Economics *Toulouse School of Economics*

December 1, 2022

Abstract

This paper intends to alleviate concerns of misreporting in Andini *et al.* (2018) by doing a principal component analysis on the variables of the decision tree in the same paper. The results of the principal component analysis indicate that it is possible to replicate the indicators in Andini *et al.* while alleviating concerns of misreporting at the same time.

*Group project for the Machine Learning for Economists course offered in Toulouse School of Economics as part of the Econometrics and Empirical Economrtrics program in the fall term of 2022/2023. We would like to thank Alessio D'Ignazio for his clarifications during various stages of our investigation. Relevant R code and data are included in the authors' GitHub pages: https://github.com/bkbasdil

†Available through baris.basdil@ut-capitole.fr.

‡Available through nikolaos.papoulias@ut-capitole.fr.

Contents

1 Introduction

This paper intends to offer a brief commentary on Andini *et al.* (2018). Andini *et al.* intended to build a decision tree for targeting consumption-constrained individuals in Italy, as part of a tax rebate program. The authors trained the decision tree for a subsample of the 2010 and 2012 samples of the Survey of Household Income and Wealth. The point of departure of this paper from that of Andini *et al.* is that the authors had used a reported variable "consumption-constrained" as a dependent variable in their growing of the decision tree. The paper at hand argues that the utilization of a reported variable in a period of prolonged economic distress may cause certain households to overstate their economic distress in order to prompt the government to engage in redistributive policy.

To this end, this paper develops an index of "economic well-being" using principal component analysis and compares the indications from this index with those of the consumption-constraint indicator developed by Andini *et al.*. The index uses the same variables as in the decision tree and is expected to offer a more objective measure of economic well-being (and by extension economic distress), alleviating any concerns of misreporting or dishonesty in the answering of the survey.

2 Methodology and Data

This paper uses principal component analysis to build an alternative measure of economic well-being (and by extension economic distress) to replace the self-reported consumption-constraint indicator. Principal component analysis is a commonly-used dimension reduction tool that projects the dataset onto a new orthogonal basis to reduce the amount of unnecessary information, or variance. Using the coefficients of the first principal component (which explains by itself the biggest portion of variance among all components), it is possible to reduce the dimension of a given dataset to one. Building on this unidimensional measure, it is possible to create an index with any range, the most common of which is the scale of

one hundred. In line with Andini *et al.* did, this paper will train the principal component analysis by selecting two-thirds of the entire sample at random, and testing the analysis on the remaining third. Then, the coefficients will be transferred to the 2014 sample.

The variables used in the principal component analysis are the same variables used in the decision tree grown by Andini *et al.*: Minimum income within household, maximum income within household, annual financial income, annual financial assets, disposable income. These quantities are all measured in Euros. The 2010 and 2012 samples include 4451 households with nonzero labor income, and the 2014 sample includes 2054 such households.

3 Results

3.1 2010 and 2012 Samples

The results of the principal component analysis of variables used in Andini *et al.* are omitted from this table in an effort to save space[1]. For simplicity, the resultant index is constructed using the first component only[2]. This component explains 44 percent of all variation, and therefore is useful in the construction of such an index. The corresponding factor loadings are given in Table 1.

To create an indicator of economic distress atop the economic well-being index, it is possible to set a cutoff point, under which households will be deemed "in economic distress". The specification of this cutoff can be done in different ways. One of these ways is to plot the distribution of the index and check whether there is a systematic break-off point in this distribution. Figure 1 plots the distribution of the index. It is seen that most index observations are situated below the index value of 25. A finer specification is given in Figure 3, where the index is plotted for values lower than ten. In that figure, it is evident that there is a breaking point around an index value of three: The left-side of the cutoff and

[1] All such omissions in the paper are included in the code. The reader is kindly asked to refer to the companion R code and data for any clarifications and further explanations.

[2] Sendhil *et al.* (2017) explains this process very simply.

the right-side of the cutoff seem to come from different distributions. As such, the indicator (henceforth PCA-constraint) is created with this cutoff: Individuals with a PCA-constraint indicator above three are deemed not to be consumption-constrained, and therefore not eligible for the tax rebate.

After the building of the index, it is reasonable to check whether the two indicators are in agreement. In other words, it is in order to check whether the old indicator and the new indicator determine consumption-constraint in a similar manner. The cross-tabulation of the two indicators indicates that the two indexes indicate the same way for 3800 households (3048 constrained and 651 not constrained) and the otherwise for 651 households (586 constrained only by Andini and 65 only by PCA). A match rate of around 85 percent is a good indication that the new index is of value: The index is available to replicate the results of the previous indication without any need for self-reporting. Furthermore, comparing economic outcomes (minimum income, maximum income, disposable income, financial assets, financial income) between the two indicators yields very similar average outcomes, in support of the previous positive finding.

3.2 2014 Sample

The component parameters set and the cutoff determined, it is in order to construct the index and indicate households eligible for the tax rebate (or equivalently consumption-constrained) in the 2014 survey. The cross-tabulation of the two indicators indicates that the two indexes indicate the same way for 1817 household (1388 constrained and 429 not constrained) and the otherwise for 237 households (64 constrained only by Andini and 173 only by PCA).

In support of the cross-tabulation featured above, the new index is able to simulate and improve upon the aforementioned economics outcomes. Table 2 shows average economic outcomes for households deemed to be constrained by Andini *et al.* or by the new index individually. It is seen that average minimum income, average maximum income, and dis-

posable income are all lower for the new index, whereas for financial assets and financial income, the opposite case holds. A similar trend is observed for Table 3, which shows average economic outcomes for households deemed not to be constrained by Andini *et al.* or by the new index individually. In this case, it is seen that all average economic outcomes are higher for the new index than for Andini *et al.*. Overall, it seems that the index is as successful in predicting non-constrained individuals as Andini *et al.* whereas in the case of constrained individuals, the new index performs better in three of the five variables. These differences can be accounted for by sampling variation, therefore no conclusive judgements are drawn from this comparison.

4 Conclusion

Motivated by the probability of a household misreporting their consumption-constraint status during a prolonged period of economic distress, this paper intends to offer an objective measure of consumption-constraint based on the decision tree grown by Andini *et al.*. Utilizing the same variables used in the decision tree, this paper uses principal component analysis to develop an index of economic well-being, and distinguishes between constrained and non-constrained households by a cutoff value within the index. The index categorizes individuals the same way as Andini *et al.* around 85 percent of the time for the 2014 wave of the survey. In the case of economic variables, the index is able to replicate average economic outcomes for individuals deemed constrained and non-constrained by Andini *et al.*. As such, it is possible to alleviate any concerns of misreporting while retaining a high level of precision.

References

[1] Andini, Monica Ciani, Emanuele De Blasio, Guido & D'Ignazio, Alessio Salvestrini, Viola. (2018). Targeting with machine learning: An application to a tax rebate program

in Italy. Journal of Economic Behavior Organization. 156. 10.1016/j.jebo.2018.09.010.

[2] Sendhil R, Anuj Kumar, Satyavir Singh, Ajay Verma, Karnam Venkatesh and Vikas Gupta (2017). Data Analysis Tools and Approaches (DATA) in Agricultural Sciences. ICAR-Indian Institute of Wheat and Barley Research. pp 1-126.

5 Appendix: Tables

Table 1: Principal Component Analysis for the Training Sample of the 2010 and 2012 Surveys.

	PC1	PC2	PC3	PC4	PC5
MinIncome	-0.5629	0.3942	-0.0808	0.1396	0.7083
MaxIncome	-0.1475	0.1223	0.9788	-0.0137	-0.0709
Financial Assets	-0.4861	-0.4763	-0.0225	-0.7321	0.0205
FinancialIncome	-0.3553	-0.6579	0.0348	0.6617	-0.0426
Disposable Income	-0.5467	0.4122	-0.1835	0.0809	-0.7007

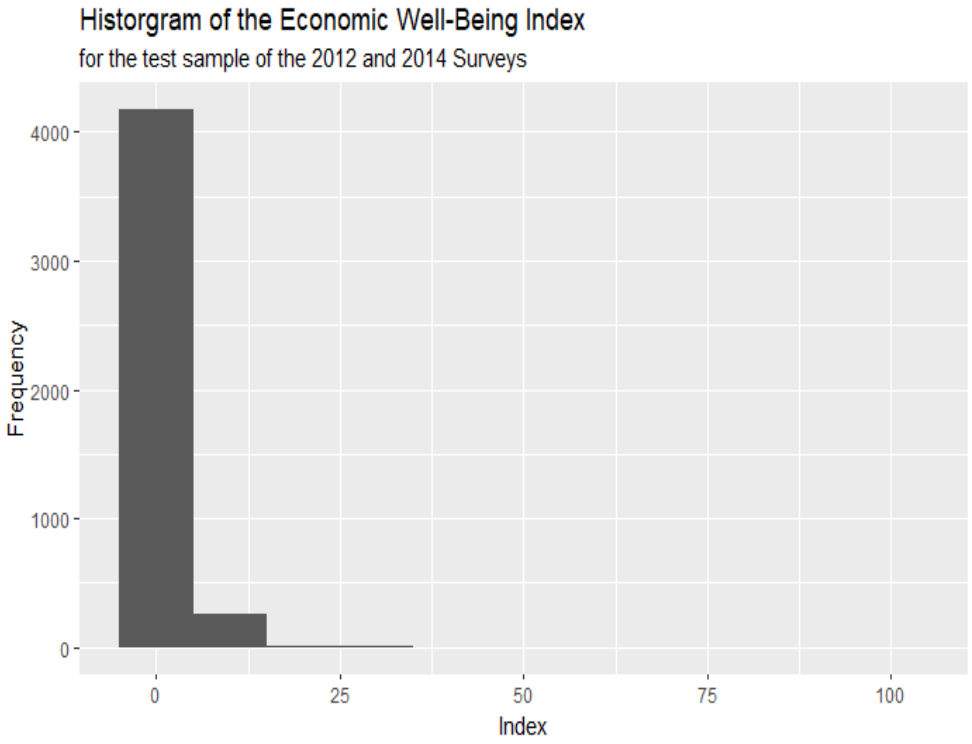

Figure 1: Histogram of the Economic Well-Being Index

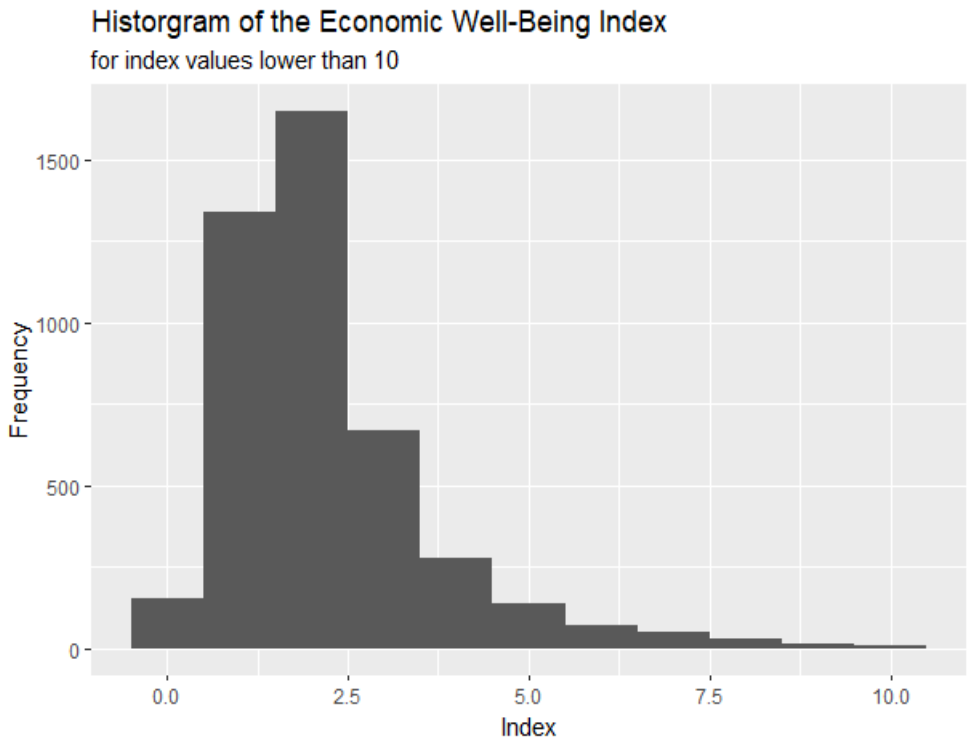

Figure 2: Histogram of the Economic Well-Being Index For Index Lower Than 10

Table 2: Average Economic Outcomes For Constrained Households

	Min. Income	Max. Income	Fin. Assets	Fin. Income	Disp. Income
Andini *et al.*	16018	17064	3497	-402.95	20744
New Index	15671	16848	5325	-354.52	19858

Table 3: Average Economic Outcomes For Non-Constrained Households

	Min. Income	Max. Income	Fin. Assets	Fin. Income	Disp. Income
Andini *et al.*	26693	20419	68106	511.1	38040
New Index	30153	21843	76603	559.81	44670