CLASSIFYING HOUSEHOLDS BY SOCIO-ECONOMIC VULNERABILITY: AN APPLICATION TO AN ITALIAN MUNICIPALITY[1]

Daniela Bonardo, Sara Casacci, Adriano Pareto, Marco Dionisio Terribili

1. Introduction

The main objective of this work is to propose a method for classifying households in order to study their social and economic conditions at low territorial detail. The work has been conducted within the ARCHIMEDE (Integrated archives of economic and demographic microdata) project of the Italian National Institute of Statistics (Garofalo 2014). The collection of microdata we used is produced from the integration of information contained in administrative sources, properly treated, to study the socio-economic situation of households in Italy. The integration of several sources (Municipal Population Registers, Tax Returns Register, Central Register of Pensioners, Social Security Archives, Social Security Benefits Register, Student Registers) allows not only an informational enrichment through the creation of new variables, but also an improvement of data quality. In fact, administrative data are collected for administrative purposes and may be not of good quality when used for statistical purposes. In this project, the integration has the goal of compiling better information than is possible when using the separate sources. In practical, a set of decision rules was designed in order to (a) correct for under-coverage or over-coverage of some target populations (e.g., income earners), (b) harmonize data under a single common denominator (e.g., correct classification of income) and (c) correct for measurement errors, resolving inconsistencies in data (e.g., correction of incorrect amount of income). Nonetheless, an accurate assessment of data quality is still needed and future work should concern a measure of the impact of the errors affecting administrative sources on the results. Despite these limitations, the information produced within the ARCHIMEDE project

[1] The paper is the result of the common work of the authors: in particular, D. Bonardo has written Sections 5 and 6, S. Casacci has written Sections 1 and 2, A. Pareto has written Section 3, and M. D. Terribili has written Section 4.

allows to expand significantly the territorial detail (municipal level) to which data are disseminated.

In this paper a vulnerability composite indicator was calculated for each household resident in a municipality, and a Cluster Analysis was performed to detect homogeneous groups of households, in order to check the consistency of the results. Vulnerability can be defined as "exposure to contingencies and stress, and difficulty in coping with them" (Chambers 1989). Data are referred to the Italian municipality of Modena, in 2012. Modena is a municipality of the Emilia-Romagna region, in the Northern Italy, counting 84,632 resident households.

Theoretical framework and selection of indicators are discussed in Section 2; whereas technical steps for constructing the composite indicator (normalization, aggregation and validation) are reported in Section 3. Cluster Analysis is described in Section 4 and results are commented in Section 5. Finally, conclusions are drawn in Section 6.

2. Role of assets in reducing vulnerability: theoretical framework and selected indicators

The notion of "vulnerability" is a very broad one, encompassing a multiplicity of meanings and approaches among the disciplines. As remarked above, it can be defined as exposure to negative events, and difficulty in coping with them. In its broadest sense, vulnerability refers to the situation of individuals, households, or communities who are exposed to potential harm from one or more risks. It also refers to the capability to face negative shocks. Differences in approaches to vulnerability among the disciplines can be explained by their tendency to focus on different components of risk, household responses to risk and welfare outcomes.

One of the approaches to vulnerability is the asset-based approach, which is based on economics terminology, but it is multidisciplinary. The new literature on asset-based approach has its genesis in Amartya Sen's entitlement approach. This approach was assimilated into the sociological/anthropological literature by the late 1980s and entitlements were extended to include social capital and other forms of intangible assets. One of major conceptual focus of this literature is the ability of households to manage risk through enhanced responses to risk, whereas the treatment of risk is mostly implicit. In asset-based analyses, households with more income and other assets are considered to be more resilient to welfare losses caused by risky events. Vulnerability, therefore, is strictly linked to asset ownership: the more assets people have, the less vulnerable they are; the greater the erosion of assets, the higher the level of insecurity (Moser and Holland 1997). Also the definition of vulnerability adopted by OECD focuses on assets: "a person (or

household) is vulnerable to future loss of well-being below some socially accepted norms if he or she lacks (or is strongly disadvantaged in the distribution of) assets which are crucial for resilience to risks" (Morrone et al. 2011).

In this work, according to OECD asset-based approach, we focused on the resources that households can draw upon to reduce vulnerability and strengthen their resilience to a range of different risks. Vulnerability is defined as insufficient capital held by households, provided that the different forms of capital (material and financial capital, instruction, etc.) are taken into account. The selection of assets is based on the consideration that in a developed country most people will never experience the grave privations commonly faced by the world's poorest populations. Vulnerability is a multidimensional concept and it should be represented under different points of view. For this reason, we selected a set of indicators grouped in the following dimensions, considered on literature as assets preventing from vulnerability (Freyssinet 2009):

1. Income: affecting the possibility of households to purchase goods and services, it determines their resilience from adverse events (job losses, long-term illness, etc.). In this dimension we also included indicators referring to the share of income earners and to the income concentration within the household. They represent proxies of the household asset management, with multiple earners with high income levels as the optimum strategy.
2. Work: this dimension detects the quantitative aspect of labour market participation. It allows to highlight situations at higher risk of poverty and social exclusion, underlining the effects of a low-intensity occupation.
3. Education: educational attainment is a proxy for human capital. A high level of education is positively correlated with high standards of living, possibility to find work, to have healthier lifestyles and more opportunities to find jobs in a less risky (OECD 2010; Miyamoto and Chevalier 2010).
4. Structure of the household: some family structure are more likely to experience poverty than others (Cancian and Reed 2002). Actually, this dimension does not constitute in itself an asset preventing from vulnerability. However, it was included in the analysis to better classify Italian households.
5. Disadvantage: this dimension detects the existence of conditions damaging individual and household well-being. The presence of household members holding of a retirement benefit for occupational diseases, accidents at work, etc. have an impact on life conditions, social relationships, opportunities and prospects of individuals and of their families.

The indicators used for the composite indicator construction (with the respective 'polarity', i.e., the sign of the relation between the indicator and the protection from vulnerability) and Cluster Analysis are listed in Table 1.

The selection of the indicators represents a compromise between the availability of information in the data sources (bottom-up approach) and the literature review. Note that X_5, X_8, X_9, X_{10} and X_{11} were excluded from the composite indicator[2], whereas they were included in the cluster analysis.

Table 1 – *List of individual indicators.*

Dimension	Indicators	Labels	Composite indicator	Polarity
Income	Household equivalised gross income (€)	X_1	Yes	+
	Share of income earners	X_2	Yes	+
	Income concentration within the household (Gini Index)	X_3	Yes	-
Work	Household work intensity	X_4	Yes	+
	Share of household members receiving an unemployment benefit	X_5	No	
Education	Share of years in education of household members	X_6	Yes	+
	Share of household members aged 18-26 not in tertiary education	X_7	Yes	-
Work and education	Share of household members aged 15-29 not in education or employment	X_8	No	
Structure of household	Share of household members aged 0-14	X_9	No	
	Share of household members aged 65+	X_{10}	No	
	Share of household members with foreign citizenship	X_{11}	No	
Disadvantage	Share of household members holding of a retirement benefit for occupational diseases, accidents at work, etc.	X_{12}	Yes	-

3. Constructing the vulnerability composite indicator

As is known, constructing a composite indicator is a complex procedure that requires the following main steps (OECD 2008, Mazziotta and Pareto 2017):

1. Defining the phenomenon to be measured. This step requires the definition of the model of measurement, in order to specify the relationship between the phenomenon to be measured (concept) and its measures (individual indicators). If causality is from the concept to individual indicators we have a reflective model; if causality is from individual indicators to the concept we have a formative model (Diamantopoulos et al. 2008).

[2] They were excluded because X_5 has not a well-defined polarity, X_8 does not represent a single dimension (it concerns both work and education) and X_9-X_{11} are auxiliary variables about the household structure.

2. Selecting a group of individual indicators. The selection is generally based on theory, empirical analysis, pragmatism or intuitive appeal. Ideally, indicators should be selected according to their relevance, analytical soundness, timeliness, accessibility and so on.
3. Normalizing the individual indicators. This step aims to make the indicators comparable, as they often have different measurement units and/or different polarities. Normalized indicators are calculated by transforming individual indicators into pure, dimensionless, numbers, with positive polarity. There are various methods of normalization, such as re-scaling (Min-Max), standardization (z-scores) and 'distance' from a reference (index numbers).
4. Aggregating the normalized indicators. It is the combination of all the components to form one or more composite indices (mathematical functions). This step requires the definition of the importance of each individual indicator (weighting system) and the identification of the technique (compensatory or non-compensatory) for summarizing the individual indicator values into a single number. Different aggregation methods can be used, such as additive, multiplicative and non-linear methods. Multivariate techniques as Principal Component Analysis are also often used.
5. Validating the composite index. Validation step aims to assess the robustness of the composite index, in terms of capacity to produce correct and stable measures, and its discriminant capacity.

In this work, a formative measurement model was adopted, since indicators such as education, income, and work are items that cause or form the latent variable of social vulnerability. The individual indicators were normalized to ensure that they were all 'bounded' (i.e., ranging between fixed values) and with positive polarity. For each indicator, higher normalized values represent greater protection from vulnerability, i.e., lower levels of vulnerability.

An exploratory Principal Components Analysis was performed to study the overall structure of the dataset, as suggested in OECD (2008). Results show that the correlations among the indicators are generally very low (Table 2) and that the information given by the individual indicators is not redundant (the 1st principal component accounts for about 30% of the total variance). This supports the theoretical choice of a formative model rather than the reflective one.

In order to select the aggregation method, a comparison among six alternative methods - *compensatory* and *non-compensatory* - was performed (Istat 2015). The following methods[3] were tested:

1. *Additive methods*. Arithmetic mean of re-scaled values in the range [0,1] (AMR); arithmetic mean of *z*-scores (AMZ).

[3] For a review of the methods, see Mazziotta and Pareto 2017.

2. *Multiplicative methods*. Jevons Index (JI), i.e., geometric mean of *index numbers*; geometric mean of re-scaled values in the range [1, 199] (GMR).
3. *Unbalance-adjusted functions*. Mazziotta-Pareto Index (MPI); Adjusted MPI (AMPI).

Table 2 – *Correlation matrix of individual indicators.*

Individual indicator	X_1	X_2	X_3	X_4	X_6	X_7	X_{12}
X_1	1.000	0.453	0.045	0.271	-0.142	-0.108	0.034
X_2	0.453	1.000	-0.509	-0.042	0.282	-0.111	0.156
X_3	0.045	-0.509	1.000	0.191	-0.242	0.147	-0.105
X_4	0.271	-0.042	0.191	1.000	-0.443	0.065	-0.226
X_6	-0.142	0.282	-0.242	-0.443	1.000	-0.076	0.272
X_7	-0.108	-0.111	0.147	0.065	-0.076	1.000	-0.065
X_{12}	0.034	0.156	-0.105	-0.226	0.272	-0.065	1.000

Frequency distributions of composite indicators show a certain similarity, except for JI and GMR (Figure 1). AMR, AMZ, MPI and AMPI have negatively skewed distributions, whereas JI presents a strong positive skewness, due to the multiplicative aggregation of index numbers that penalizes low values of individual indicators. GMR has a very irregular distribution due to the use of the geometric mean with a Min-Max normalization. In addition to the similarities between the frequencies distributions, the rank correlations among the composite indicators are very high, except for JI (Table 3).

Since different weighting systems imply different results, no attempt is made to explicitly weigh the individual indicators. Implicitly then the dimensions are not equally weighted, but each of them is 'weighted' proportionally to the number of individual indicators that represent it. This introduces an element of subjectivity, but one that appears manageable because it relates to the relative importance of different aspects of vulnerability.

Results of the Influence Analysis[4] indicate that JI and GMR produce a lack of balance between individual indicators (i.e., the removal of an individual indicator produces a strong variation in the household ranking). AMR and AMZ, by contrast, turn out to be the most robust composite indicators.

On the basis of this information, the AMR was used for constructing the Vulnerability Composite Indicator (VCI) because it represents a good trade-off between robustness and interpretability (*z*-scores are 'unbounded', i.e., they do not range between fixed values, and then AMZ is more difficult to interpret). This means that, while there are differences between the properties of different kinds of assets, what is important is the compensability of different types of assets: "low

[4] Influence Analysis is a particular case of Uncertainty Analysis where individual indicators are iteratively removed from the composite indicator in order to assess its robustness (Mazziotta and Pareto 2017).

levels of one type of asset do not necessarily mean that an individual or household is inherently vulnerable; it is the composition of the overall 'asset portfolio' that matters" (Morrone et al. 2001). So, for example, it is reasonable to suppose that a low work intensity can be offset by a high value of household income.

Figure 1 – *Frequency distribution of composite indicators.*

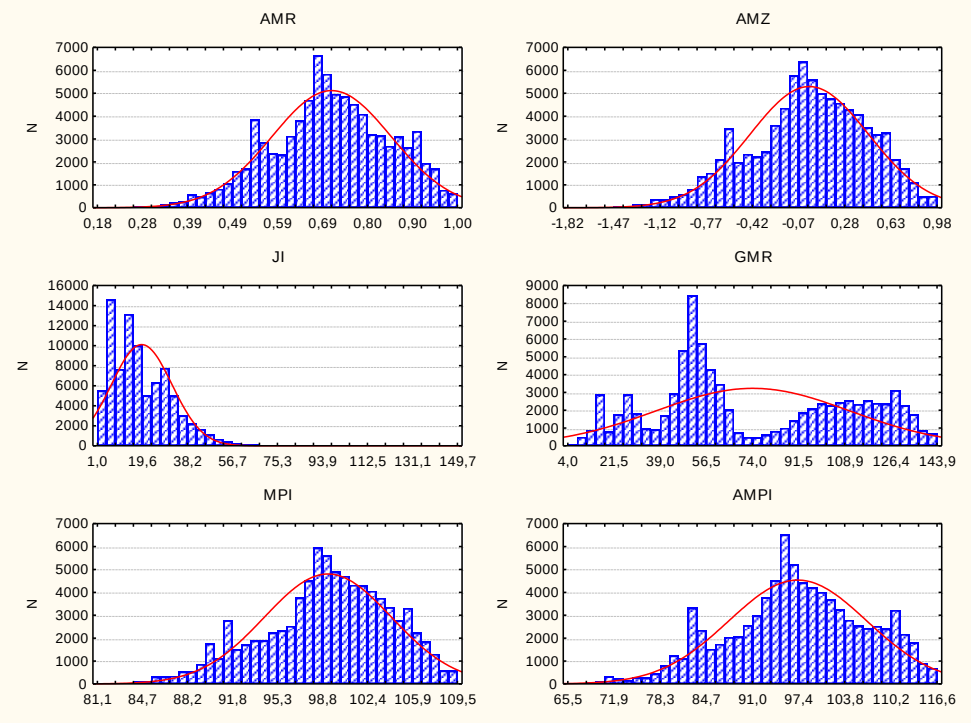


Table 3 – *Spearman correlation matrix of composite indicators.*

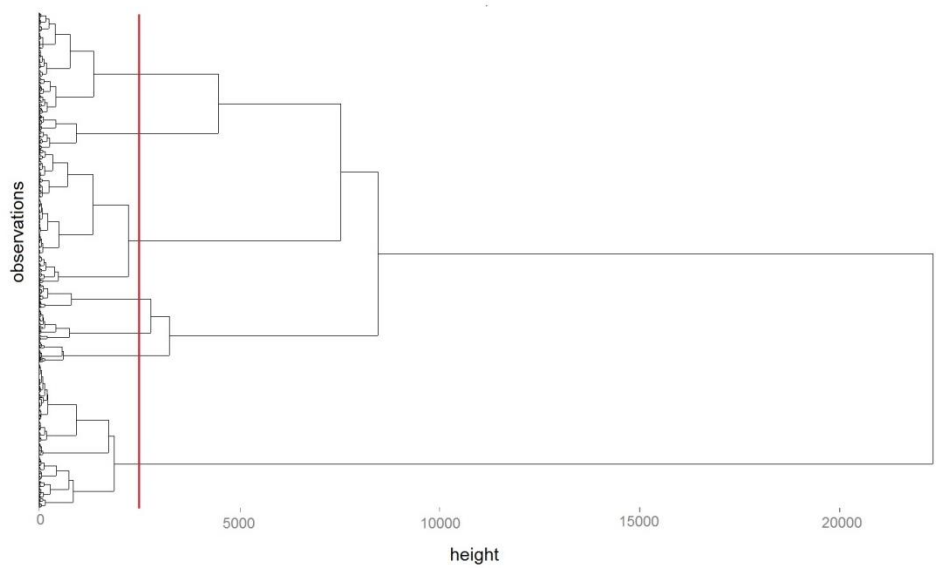
Composite indicator	AMR	AMZ	JI	GMR	MPI	AMPI
AMR	1.000	0.985	0.195	0.933	0.977	0.989
AMZ	0.985	1.000	0.160	0.906	0.996	0.982
JI	0.195	0.160	1.000	0.415	0.163	0.201
GMR	0.933	0.906	0.415	1.000	0.910	0.943
MPI	0.977	0.996	0.163	0.910	1.000	0.986
AMPI	0.989	0.982	0.201	0.943	0.986	1.000

4. Clustering households by vulnerability

Afterward the assignment of a vulnerability level to every household through the composite indicator, some multivariate techniques of clustering have been

applied to identify subpopulations or groups of units, as more homogeneous within, as more heterogeneous between each other. Our goal is to apply Cluster Analysis to point out distinct population segments sharing the same vulnerability level, based on the individual indicators used for constructing VCI and some other variables (Table 1). The clustering techniques are usually discerned in hierarchical and non-hierarchical. Several hierarchical clustering have been tested, by using two different distance functions (Euclidean and Manhattan distance) and three aggregation methods (Ward's method, complete linkage and average linkage). The choice of the number of clusters to be generated was made on the basis of our informational needs and of a graphical analysis of the resulting dendrograms (the output tree diagrams of the algorithm). Regarding the informational needs and the diagrams (in Figure 2 is shown an example), seven clusters have been pointed out, by pruning the dendrogram up to the red line. This solution represents a good compromise between a detailed clusters number and their within homogeneity.

Figure 2 – *Dendrogram of hierarchical clustering by Ward's method with Euclidean distance.*



Considering our informational needs, the established number of groups to generate and the high number of observations to cluster, a non-hierarchical method has been applied to point out, in a different way, seven clusters. So the clustering algorithm has been rerun by using the k-means method and exploiting the Euclidean distance function to aggregate the observations around the seven centroids, unobserved units representing the average position (in a m-dimensional

space, where m is the number of variables considered) of each group. These average positions are very important to define the distinct population segments, sharing the same condition about vulnerability. *K*-means method is quicker and more efficient than hierarchical methods, but it provides similar results. This is confirmed by the Chi-Square Test of Independence and also by the Cramer's V (Table 4). Null hypotheses of independence are all rejected (p-value<0.05) and Cramer's V values, ranging from 0 to 1, are in any case greater than 0.47. The aggregation method which gives the most similar partition is the Ward's method with Euclidean distance (Cramer's V=0.80), because both methods are based on the Sum of Squared Error (SSE). On the basis of these results, we classified households by *k*-means method, that is particularly able to find spherical clusters.

Table 4 – *Comparing 7-cluster partition of hierarchical methods with the one of k-means method*

Method	Distance	χ^2	p-value	Cramer's V
Ward	Manhattan	283,248.4	< 0.05	0.75
	Euclidean	324,305.0	< 0.05	0.80
Complete-linkage	Manhattan	187,649.3	< 0.05	0.61
	Euclidean	233,098.9	< 0.05	0.68
Average-linkage	Manhattan	153,493.2	< 0.05	0.55
	Euclidean	111,742.7	< 0.05	0.47

5. Results

The VCI for the households of Modena ranges from 0.18 (highest vulnerability level) to 1 (lowest vulnerability level). Overall, the degree of socio-economic vulnerability[5] is quite low: the distribution of the VCI is slightly negatively skewed with a mean of 0.71 and a standard deviation of 0.14. This result seems to be plausible, since Emilia-Romagna is a region with high levels of income and well-being. Cluster Analysis allows to detect specific groups of households in relation to their possession of asset, taking into account several socio-demographic characteristics (citizenship, structure of the households, age of members). One of the most interesting results of cluster analysis is that the elements of vulnerability often overlap. For each group of households, the median of VCI and other descriptive statistics were calculated, in order to assess the level of vulnerability (Table 5). The "Protected senior citizen" cluster is the largest, accounting for 26.8% of the total number of households. It is characterized by elderly people, perceiving a guaranteed income, with a medium-low level of vulnerability. Their

[5] Although a measurement of vulnerability should include the definition of a *cut-off* or *benchmark*, we did not choose a cut-off, since it should vary in different municipalities.

median degree of vulnerability is 0.67 since they lack in assets such as education and health. The "Well-to-do singles & couples" group (22%), is mostly composed by one-person households with both a high work intensity and a high income; this group has the lowest vulnerability risk (median=0.89). The "Leisure class" cluster (19.9%) is characterized by the presence of children, generally one income earner, high work intensity and low vulnerability risk (median=0.73). The "Scanty capital" cluster contains about nine thousand household (10.7% of the total) with low levels of education and lack of employment. This group presents a medium-high degree of vulnerability (median=0.66). Two clusters with different profiles are referred to foreigner households: the "At risk" cluster (7.1%), with high incidence of unemployed young people and low family income, and the one called "In gear" (6.9%), composed probably of long-term immigrants with high-intensity occupation. Whereas the "In gear" group is associated to a low exposure to vulnerability (median=0.80), the "At risk" one appears to be the most vulnerable (median=0.53). Lastly, the "Jobless" cluster (6.7%), which identifies households with children, is exposed to a greater risk of vulnerability (median=0.60) since adults are often unemployed.

Table 5 – *Absolute and percentage frequencies of clusters, and VCI statistics*.

Cluster	N	%	VCI			
			Min	Median	Max	Std dev
Protected senior citizen	22,662	26.8	0.34	0.67	0.86	0.08
Well-to-do singles & couples	18,611	22.0	0.61	0.89	1.00	0.06
Leisure class	16,812	19.9	0.41	0.73	0.90	0.08
Scanty capital	9,096	10.7	0.32	0.66	0.87	0.11
Foreigners at risk	5,998	7.1	0.18	0.53	0.77	0.09
Foreigners in gear	5,822	6.9	0.43	0.80	1.00	0.09
Jobless	5,631	6.7	0.21	0.60	0.84	0.12
Total	84,632	100.0	0.18	0.77	1.00	0.12

6. Conclusions: strengths and weaknesses

The main goal of this paper is to propose a combination of methods for classifying Italian households in relation to their socio-economic vulnerability, by using experimental microdata obtained from the treatment and integration of administrative sources. The core of the work is the construction of a composite indicator of vulnerability (VCI), by aggregating individual indicators concerning different dimensions (income, work, education and health) in order to assign a vulnerability level to every household. Measuring households' vulnerability by this approach has evident advantages, such as an one-dimensional representation of the phenomenon and an immediate interpretation and usability of data. However, the

reconstruction process of individual indicators in a composite indicator is complex in itself, since it needs a number of theoretically and methodologically oriented choices (e.g., variables used and indicators meaning, choice of the aggregation function) that have a significant impact on the final results. Furthermore, aggregation of the individual indicators implies a loss of information, as we are no more able to recognize the features of the vulnerable households. For this reason, a cluster analysis was conducted, trying to identify and characterize specific groups of households. The cluster analysis has highlighted distinct groups whose configurations in relation to socio-economic profiles and structure of households are fairly intuitive. Outcome of cluster analysis seems to be quite consistent with the results of the VCI, since the groups pointed out present different vulnerability levels and are able to discriminate among different types of households.

Turning to policy, the possibility of assessing the vulnerability degree of every Italian household (describing the size and characteristics of the vulnerable population) is a powerful tool for identifying the policy priorities required to reduce the incidence and intensity of vulnerability. Besides, the VCI is useful for analysis over time and among different groups or municipalities. Future work should concern the application of both methods on households of geographical areas with different characteristics, in order to verify the effectiveness of the choices made.

References

CANCIAN M., REED D. 2002. Changes in Family Structure: Implications for Poverty and Related Policy. In DANZIGER S.H., HAVEMAN R.H. (Eds) *Understanding Poverty*, Russell Sage Foundation Books at Harvard University Press, pp. 69-96.

CHAMBERS R. 1989. Editional introduction: vulnerability, coping and policy. In Chambers R. (Ed) *Vulnerability: How the Poor Cope*, I.D.S. Bulletin, pp. 1-7.

DIAMANTOPOULOS A., RIEFLER P., ROTH, K.P. 2008. Advancing formative measurement models, *Journal of Business Research*, 61, pp. 1203-1218.

FREYSSINET J. 2009. *How can social vulnerability be measured: a work in progress*. The 3rd OECD World Forum on «Statistics, Knowledge and Policy». Busan, October 27-30.

GAROFALO G. 2014. *Il Progetto ARCHIMEDE obiettivi e risultati sperimentali*. Istat Working Paper, 9. Available from: http://www.istat.it/it/files/2014/11/IWP-n.-9-2014.pdf. Accessed May 11, 2015.

ISTAT, 2015. Rapporto Bes 2015: il benessere equo e sostenibile in Italia. Available from: http://www.istat.it/it/files/2015/12/Rapporto_BES_2015.pdf

MAZZIOTTA M., PARETO A. 2017. Synthesis of Indicators: The Composite Indicators Approach. In MAGGINO F. (Ed) *Complexity in Society: From Indicators Construction to their Synthesis*, Social Indicators Research Series Vol. 70, Cham: Springer, pp. 159-191.

MIYAMOTO K., CHEVALIER A. 2010. *Education and health*, Chapter 4 of Improving Health and Social Cohesion through Education. OECD Publishing.

MORRONE A., SCRIVENS K., SMITH C., BALESTRA C. 2011. *Measuring vulnerability and resilience in OECD countries*. IARIW-OECD Conference on Economic Insecurity, Paris, November 22–23.

MOSER C., HOLLAND J. 1997. *Household Responses to Poverty and Vulnerability*. *Volume 4: Confronting Crisis in Cawama, Lusaka, Zambia*. Urban Management Programme, Report No. 24. Washington, D.C: The World Bank.

OECD 2008. *Handbook on Constructing Composite Indicators. Methodology and user guide*. Paris: OECD Publications.

OECD 2010. *The OECD Innovation Strategy. Getting a Head Start on Tomorrow*. Paris: OECD Publishing.

SUMMARY

Classifying households by socio-economic vulnerability: an application to an Italian municipality

The measurement of the socio-economic vulnerability of communities and households, especially at a low territorial detail, has important implications both in terms of the analysis of well-being and in terms of policy. This paper reports the results of a work conducted for classifying households of an Italian municipality in relation to their socio-economic vulnerability. Data are referred to the Italian municipality of Modena, in 2012. Since vulnerability is a multidimensional concept, a composite indicator approach was followed. A Cluster Analysis was also performed in order to identify and characterize specific groups of households. The empirical evidence shows that the degree of socio-economic vulnerability of households is quite low and that the elements of vulnerability often overlap. Translated into operational practice, the proposed framework facilitates policy makers to suitably target the local level interventions and to define the hierarchy of priorities to endorse the well-being of households and communities.

Daniela BONARDO, Istat, bonardo@istat.it
Sara CASACCI, Istat, casacci@istat.it
Adriano PARETO, Istat, pareto@istat.it
Marco Dionisio TERRIBILI, Istat, terribili@istat.it