A CLUSTERING APPROACH FOR DETERMINING STRATIFICATION VARIABLES IN SBS SURVEYS

Ilaria Bombelli, Giorgia Sacco, Alessio Guandalini

Abstract. Many Structural Business Statistics (SBSs) surveys, according to European Regulations, must move from considering the Legal Unit (LU) as unit of interest towards considering the Enterprise (ENT) as such. This transition is not trivial, as many NSIs still need to provide estimates at a LU level, for comparability through the time.

Consequently, to modify and enhance the standard sample design based on LU to address this shift, it could be required to investigate an alternative stratification of the sample. To address this task, we propose to use a clustering algorithm, i.e., the K-prototype, to obtain groups of ENT and assess the variables' importance in the clustering result.

The algorithm is applied to several input datasets, obtained by sub-setting the ASIA ENT 2021 register, which includes all enterprises carrying on economic activities. The input datasets include ENT working on different sections of the statistical classification of economic activities in the European Community (NACE) and ENT included in the target population of the Community Innovation Survey (CIS) carried out by ISTAT. The clustering is applied separately to each aforementioned dataset. From the clustering result, we assess the variables' importance and identify the variables that mostly influence the obtained partition.

The most influential variables are used to build the new stratification of the ENT, hence they contribute to a new definition of the strata. The proposed stratification is used to allocate a sample of the same dimension as the one extracted with the current stratification. From the sample, we estimate some of the survey's target variables and their coefficient of variation (CV). The CVs are compared with the ones resulting from the current stratification. The comparison reveals that the efficiency of the estimates is preserved. In addition, the new stratification allows for reducing the number of strata and therefore also the processing time is limited.

1. Motivation of the study

Structural Business Statistics (SBS) surveys are carried out periodically by the Italian National Institute of Statistics (ISTAT), as well as the other National Statistical Institutes (NSI) in Europe. SBS surveys are used to collect information about, among the others, the organisation's structure, the economic activity and performance in terms of turnover, production and innovation of businesses over

time[1]. In addition, as required by the EUROSTAT regulations, definitions and harmonisation policies, the statistical units in SBS surveys carried out by European NSI must have the same definition to make comparable estimates.

Therefore, nowadays, all European National Statistical Institutes (NSI) are facing a new challenge: the transition from considering the legal unit (LU) as the statistical unit of analysis towards considering the Enterprise (ENT) as such. The LUs include *legal persons whose existence is recognized by law independently of the individuals or institutions which may own them or are members of them* and *natural persons who are engaged in an economic activity in their own right.* Moreover, *the legal unit always forms, either by itself or sometimes in combination with other legal units, the legal basis for the statistical unit known as the enterprise*[2]. The ENT is defined as *the smallest combination of legal units that is an organizational unit producing goods or services, which benefits from a certain degree of autonomy in decision-making, especially for the allocation of its current resources. An enterprise carries out one or more activities at one or more locations. An enterprise may be a sole legal unit*[3].

It is important to mention that each ENT can be made up of one or more LU, which in turn can belong to one or more ENT simultaneously according to a given *share*. Specifically, in the ASIA ENT register related to 2021, which includes all enterprises carrying on economic activities in 2021, almost all ENT are made up of a single LU: among the 4462146 ENT, 4409019 (98.8%) are made up of only one LU. For a complete description of the register, the reader can refer to Section 2 of the current paper.

However, even though NSI must provide estimates at the ENT-level, as required by EUROSTAT, they also want to provide estimates at LU-level, for comparability through the time. This makes the aforementioned transition not trivial. For this reason, a solution which has been adopted is to use a cluster sampling strategy: in this way, the ENT are considered as the statistical units during the sampling process; then, once a sample of ENT has been selected, all the LUs belonging to the sampled ENT will be provided with the questionnaire. This sampling strategy will allow to collect data at a LU-level, but the estimates are possible at both LU and ENT levels, by properly adjusting the sampling weights in the calibration step. This strategy has

[1] Source: https://ec.europa.eu/eurostat/web/structural-business-statistics#:~:text=Structural%20business%20statistics%20(SBS)%20describe,European%20business%20statistics%20.

[2] EUROSTAT glossary accessible at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Legal_unit.

[3] Regulation (CEE) n. 696/93 accessible at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31993R0696.

been adopted for selecting the sample of the Community Innovation Survey (CIS) of 2022, as it will be discussed in Section 2.1.

What is proposed in this paper are some suggestions for leading the transition process. More in detail, we want to highlight some insights to adapt the sampling design to this change and to stratify ENT efficiently. In order to study a new stratification, we apply cluster analysis on ENT; then, from the clustering result, we assess the variables' importance, and we identify the variables that mostly influence the obtained partition. In this way, the most influential variables are considered as the new stratification variables.

The data and the methodological tools descriptions are provided in Section 2, 3, respectively. Results are presented and discussed in Section 4. Finally, Section 5 sums up the main outcome of the research.

2. Data

The Business Register ASIA includes all enterprises carrying on economic activities contributing to gross domestic product at market prices, in the fields of industry, commerce and services[4]. The register is therefore used to collect information about the ENT.

Table 1 - *Variables' description.*

Variable	Description	Type
number_ug_ent	number of UG belonging to the ENT	numeric
employees	number of employees working for the ENT	numeric
turnover	turnover	numeric
s_shares	sum of the shares of the LU belonging to the ENT and sharing with it the same NACE code	numeric
s_shares_diff	sum of the shares of the LU belonging to the ENT and with a NACE code different from the ENT's code	numeric
ls_ent	legal status	categorical
region_ent	region	categorical
ateco5_ent	Categories NACE code (5 digits)	categorical
ateco4_ent	Classes NACE code (4 digits)	categorical
ateco3_ent	Groups NACE code (3 digits)	categorical
ateco2_ent	Divisions NACE code (2 digits)	categorical
section_ent	Sections NACE code (1 digit)	categorical
cladd3_ent (cis)	class of employees: 10-49, 50-249, 250+	categorical

[4] For a complete description of the register the reader may refer to Siqual website accessible at: https://siqual.istat.it/SIQual/visualizza.do?id=7777976.

The most updated version of the register available, i.e. ASIA ENT 2021, includes the ENT active in 2021. The number of units (ENT) is 4462146, while the number of variables is 17.

We considered only some of the available variables to carry out our analysis: *number_ug_ent, employees, turnover*[5]*, ls_ent, region_ent, ateco5_ent*. In addition, by furtherly manipulating some of the variables we determined additional variables such as *s_shares, s_shares_diff, ateco4_ent, ateco3_ent, ateco2_ent, section_ent* and *cladd3_ent*. The variables *ateco5_ent, ateco4_ent, ateco3_ent, ateco2_ent, section* are related to the ENT's economic activity according to the *statistical classification of economic activities in the European Community* (NACE classification). The variables' descriptions and information are summarised and reported in Table 1.
Instead of considering the whole ASIA register, we implemented our study on several subsets of this register. The filtering variables were those related to economic activity.

- ENT working in the C section of the NACE classification ('manufacturing activities');
- ENT working in the M section of the NACE classification ('Professional, Scientific and Technical Activities');
- ENT working in the R section of the NACE classification ('arts, sports, entertainment and recreation activities);
- ENT working in the division codes (NACE 2-digits) of the NACE classification usually survey in Community Innovation Survey (CIS).

The following Section provides a full description of the aims and the current sampling design for the CIS survey.

2.1. CIS Survey

The Statistical Survey of Business Innovation (CIS) aims to collect information on the strategies, behaviours and innovative activities carried out by enterprises. Since 2023, the sample survey is based on a stratified cluster sampling design with equal inclusion probabilities for all population units. The enterprises were stratified, and from each stratum they were selected by simple random sampling. All legal units belonging to enterprises were sampled. In this way, data are collected at LU-level and therefore direct estimates at LU-level are easily obtained; at the same time, given

[5] To handle missing data in this variable, we used the Multivariate Imputation by Chained Equation (mice R package, van Buuren and Groothuis-Oudshoorn, 2011), even though we are aware that there exist many other imputation techniques which are used in Official Statistics.

the LU-ENT structure, it is also possible to obtain estimates at ENT-level by properly adjusting the weights in the calibration process.

The goal of the survey is to collect information on the number of employees, turnover and total innovation spending of enterprises. In particular, the target population consists of all enterprises with more than 10 employees and operating in the following economic activity by NACE Rev. 2 at one-digit (section) level: B, C, D, E, F, G, H, J, K, L, M (excluding division 75). The survey is sample-based for enterprises with 10 to 249 employees and census-based for those with 250 or more employees. The stratification variables are: a) the economic activity by NACE Rev. 2. Stratification by NACE was done at a two-digit (division) level, except for section F; b) the enterprise size according to the number of persons employed (the size-classes used were the following ones: between 10 and 49; between 50 and 249; 250 and more); c) the regional variable. The breakdown of national territory into regions was performed based on the NUTS level 2.

The estimates are provided, according to regulations, for the following 4 domains: a) Estimation domain 1: economic activity by NACE Rev. 2 at two-digit (division) level, except for section F; b) Estimation domain 2: economic activity by NACE Rev. 2 at two-digit (division) level, except for section F cross-tabulated with 3 classes of employees (3 classes: 10-49, 50-249, 250 and above); c) Estimation domain 3: the regional variable cross-tabulated with Macrosector (Eu-core sectors - Other sectors) and 2 classes of employees (2 classes: small and medium-sized enterprises - large enterprises); d) Estimation domain 4: only for Bolzano province, Trento province, and Friuli Venezia Giulia region: economic activity by NACE Rev. 2 at one-digit (section) level cross-tabulated with Eu-core/Others.

A multi-variable and multi-domain sample allocation is used. In particular, the adopted procedure is an application of the Bethel algorithm (Bethel, 1989) . It is an optimum allocation since it aims at minimising survey costs under the constraint that sampling errors (in terms of coefficient of variation, CV) of estimates of each variable of interest and for each domain don’t exceed the given upper bounds assigned to each of them.

3. Methodological Tools

Cluster analysis is used to obtain clusters of ENT and to identify the most influential variables in the partition which can be used as stratification variables: indeed, influential variables in clustering are the variables that discriminate and partition the ENT homogeneously, so it is reasonable to use them to build new strata of ENT. In this way, ENT will be stratified efficiently.

Since our dataset includes both numerical and categorical variables, we applied the K-prototype clustering algorithm (Huang, 1998), allowing mixed-data type as input. The algorithm is implemented in the clustMixType R package (Szepannek, 2018) and it is considered as an extension of the K-Means (McQueen, 1967) and K-Modes (Huang, 1997) algorithm.

To apply cluster analysis to the four subsets of interest, we decided to use a bootstrap-like procedure for handling huge data dimensions. In particular, from each dataset, 500 random samples of 1000 units without replacement have been selected. On each sample, we performed clustering. Practically, on each sample, by letting the number of clusters vary in [2,20], we ran the K-prototype and we selected the optimal number of cluster k* using the Silhouette index (Rousseeuw, 1987). The resulting partition with k* clusters has been evaluated and analysed. To study which variables were the most influencing in the resulting partition, the feature importance of each variable has been computed by using the R function FeatureImpCluster (Pfaffel, 2021), which measures the importance of a variable in terms of misclassification rate relative to the baseline cluster assignment due to a random permutation of feature values. Finally, variables are ranked according to the feature importance score.

This bootstrap-like procedure helped us to deal with high-dimensional datasets, as in our application. However, to validate the results, we implemented the clustering algorithm on the whole CIS survey target population and we assessed the variables' importance. The most influential variables can indeed be considered for a new stratification.

4. Results

For each dataset, Table 2 reports the variables that were shown to have the greatest influence in most of the iterations of the study. Specifically, the percentage distribution of the most influencing variable obtained in each of the 500 iterations has been computed. Then, the two variables which have the two highest percentage frequencies are reported.

Table 2 – *Two most influencing variables in the simulation study. Datasets C, M, R refer to the subsets of ENT of ASIA2021 working in section C, M, R of NACE classification, respectively.*

Dataset	Two most influencing variables	% of occurrences
C	*ls_ent, s_shares*	82.2, 6.0
M	*region_ent, ateco5_ent*	74.6, 21.8
R	*ateco5_ent, region_ent*	60.0, 39.4
univ_CIS	*section_ent, region_ent*	92.0, 8.0

Table 2 indicates that the region and the section, which pertain to economic macro-activity, are the two most influential variables in the dataset associated with the CIS survey. We found that location information is crucial for the two economic sectors that sections M and R identify. The location of the enterprises may have an impact on their economic activities. Additionally, the variable *ateco5_ent*, which refers to a very detailed economic activity description (being the finest economic activity classification), turns out to be a very discriminant variable given that those two sectors are very heterogeneous, including a large number of quite different economic sub-activities. Therefore, in this sector, the factors that have the biggest effects on differentiating amongst ENTs are the locations of the enterprises and the specific economic sectors in which each of them operates.

As mentioned in Section 3, we also used the K-prototype for the whole CIS dataset to properly validate the findings of the previously described study. The results indicate that the ENT in the CIS dataset is divided into 18 clusters, with *region_ent* and *section_ent* having the most influence in the obtained partition. These variables are consistent and support the output of the simulation.

Furthermore, given the peculiar composition of ENT, we divided the dataset into two subgroups to examine the results in more detail: the first subset (*ENT-Mono*) only takes into account ENTs composed of a single LU, while the second subset (*ENT-Pluri*) only takes into account ENTs composed of more than one LUs. Indeed, in ASIA ENT 2021, among the 4.462.146 ENT, only 1.2% is *ENT-Pluri*, and the remaining 98.8% is *ENT-Mono*, while in the CIS dataset, among the 156619 ENT, 15.67% is *ENT-Pluri*, and the remaining 84.43% is *ENT-Mono.*

We decided to implement two separate analyses for *ENT-Mono* and *ENT-Pluri*, as we thought that the structure and characteristics of those two types of ENT were extremely different. The K-prototype algorithm was applied to both subgroups.

Section_ent and *region_ent* are the two most important variables for *ENT-Mono*, as the simulation study also highlights. Rather, the variables with the highest relevance scores for the *ENT-Pluri*, a very small subset of the entire dataset, are *section_ent, ateco2_ent, cladd3_ent*, and *region_ent*. What is therefore underlined is

the potential for a new stratification that is more wide and comprehensive when the ENT is an *ENT-mono* and more precise and specific when the ENT is an *ENT-pluri*.

Therefore, in order to wrap up the experiment, we choose to test the newly proposed stratification, where the strata variable for each ENT i is specified as follows:

$$stratum_i = \begin{cases} region_i \times section_i & if\ the\ ENT\ i\ is\ a\ ENT-mono \\ region_i \times ateco2_i \times cladd3_i & if\ the\ ENT\ i\ is\ a\ ENT-pluri \end{cases} \quad (1)$$

It has to be noticed that we choose to include *ateco2*, which is a subset of the *section* and more accurate in characterising economic activity, in place of the *section* variable for the *ENT-pluri*.

Table 3 – *Median values of estimates' CVs resulting from the samples selected either using the standard allocation or the new one***.**

DOM	Variable	CV_alloc_stand	CV_alloc_new
DOM 1	*employees*	0.017	0.064
DOM 2	*employees*	0.016	0.013
DOM 3	*employees*	0.015	0.009
DOM 4	*employees*	0.013	0.035
DOM 1	*turnover*	0.037	0.110
DOM 2	*turnover*	0.045	0.066
DOM 3	*turnover*	0.028	0.022
DOM 4	*turnover*	0.024	0.057

The proposed stratification has been employed to allocate a sample of the same dimension (39000 LU) as the one extracted using the existing stratification, and the optimal allocation in a multi-variable and multi-domain framework has been used[6] (Fasulo et al., 2021; Barcaroli et al., 2023). Compared to the current stratification, there are about 800 fewer new strata (1769 instead of 2541).

Given the new allocation, we derived the estimates of employees and turnover variables and their related CV. A common and extensively used metric of an estimate's mistake in various NSIs is the CV, which is defined as the ratio of the standard deviation to the mean. Publicising estimates with a CV less than 0.33 and classifying as totally dependable those with a CV less than 0.15[7] is standard procedure.

[6] The optimal allocation in a multi-domain, multivariate framework implemented in the R package R2BEAT has been used (see https://barcaroli.github.io/R2BEAT/index).

[7] See Statistics Canada. Sampling error, https://www150.statcan.gc.ca/n1/pub/62f0026m/2010003/section1-eng.htm

More in detail, we implemented a Monte Carlo simulation with 100 iterations: in each iteration, two samples have been selected: the former is extracted using the current allocation, while the latter is selected using the proposed allocation. Given the two samples, the estimates of the target variables and their related CV are computed. The estimates are obtained at the four different domains discussed and presented in Section 2.1.

Table 3 reports the median level of the CVs of the estimates for each combination of target variable and domain of interest.

The estimates derived from the samples extracted following the proposed allocation have median values of the CVs that are marginally higher than those associated with the current allocation in some combination but turn out to be below the acceptance threshold.

To conclude the proposed assessment, two are the most important advantages of the methodological proposal. On the one hand, the new stratification allows to obtain a lower number of strata; on the other hand, the efficiency of the estimates of the target variables for each domain of interest is still preserved.

The reduction in the number of strata turns out to be an important gain in stratified sampling strategy: indeed, one of the main concerns and issues which arise when adopting stratified sampling is related to the fact that after the survey, several strata turned out to be empty, i.e. no-one of the sampled unit responded to the survey; the *empty-strata issue* is a big concern as units belonging to the strata are no-more represented by any respondent. Instead, by considering a lower number of strata, the strata dimension at the population-level (number of population units, *N_h*) is higher, as well as therefore also the strata dimension at the sample-level (number of sample-allocated units, *n_h*); as a consequence, the risk of incurring in empty-strata issue is considerably reduced.

Furthermore, because the number of strata is reduced, the sample is less widespread and can be allocated more efficiently.

Finally, considering the editing and imputation phase of the whole statistical survey process, when using hot-deck imputation, if imputation classes exactly match the strata, then the number of potential donors inside each imputation class is high, given that the strata have a higher size and hence a higher number of respondents is obtained in each stratum.

It is worth mentioning that a reduced number of strata allows also to reduce the processing time of the whole allocation, sampling and estimation processes.

5. Final consideration

This work focuses on addressing the use of a new stratification to select a sample of ENT in SBS surveys. More in detail, a clustering approach for determining stratification variables has been proposed: K-prototype clustering algorithm has been applied to several input datasets, obtained through a subset of the ASIA ENT 2021 register, which is a mixed-data type dataframe. The input datasets include ENTs working in different sections of the statistical classification of economic activities in the European Community (NACE) and ENTs included in the target population of the CIS survey conducted by ISTAT. Finally, the most influencing variables in determining the clustering partition have been used as stratification variables.

The analysis of the CIS target population has been deeply discussed; the new proposed stratification has been used to allocate a sample of ENT of the same dimension as the one selected with the current stratification. Through a Monte Carlo simulation, the variables of interest related to the CIS survey (turnover and employees) have been estimated as well as their coefficients of variation. By comparing the results obtained by sampling using the proposed stratification and the ones obtained by sampling using the current stratification two important aspects are revealed: on the one hand, the efficiency of the estimates is preserved; on the other hand, the number of strata has been reduced, and several related advantages have been discussed in the previous Section. Among the others, it's worth recalling the reduction of the risk of incurring in empty-strata issue and the reduction in processing time.

This work aimed to provide some hints to help NSI go through the transition process from considering the Legal Unit as the statistical unit in SBS surveys towards considering the ENT as such. The results are promising and suggest that it may be necessary to think about a new sampling design. However, further investigations are needed.

References

BARCAROLI G., FASULO A., GUANDALINI A., TERRIBILI M.D. 2023. Two Stage Sampling Design and Sample Selection with the R Package R2BEAT. *The R Journal*, Vol. 15, No. 3, pp. 191-213.

BETHEL J. 1989. Sample allocation in multivariate surveys. *Survey Methodology*, Vol. 15, pp. 47 -57

FASULO A., BARCAROLI G., FALORSI S., GUANDALINI A., PAGLIUCA D., TERRIBILI M.D. 2021. R2BEAT: Multistage Sampling Allocation and Sample

Selection *R package version* 1.0.4. url:https://CRAN.R-project.org/package=R2BEAT

HUANG J. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, Vol. 2, No. 3, pp. 283-304.

HUANG Z. 1997. A fast clustering algorithm to cluster very large categorical data sets in data mining. *DMKD*, Vol. 3, No. 8, pp. 34-39.

MCQUEEN J. B. 1967. Some methods of classification and analysis of multivariate observations. In: *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob*, Vol 2, No. 3, pp.281-297.

PFAFFEL O. 2021. FeatureImpCluster: Feature Importance for Partitional Clustering. *R package version 0.1.5*. url: https://CRAN.Rproject.org/package=FeatureImpCluster.

ROUSSEEUW P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, Vol. 20, pp. 53-65.

SZEPANNEK G. 2018. ClustMixType: User-Friendly Clustering of Mixed-Type Data in R. *R Journal*, Vol. 10, No. 2, p. 200.

VAN BUUREN S., ROOTHUIS-OUDSHOORM K 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, Vol. 45, No. 3, pp. 1-67.

Ilaria BOMBELLI, ISTAT, ilaria.bombelli@istat.it
Giorgia SACCO, ISTAT, giorgia.sacco@istat.it
Alessio GUANDALINI, ISTAT, alessio.guandalini@istat.it