RESPONDENTS AND NON RESPONDENTS TO POPULATION AND HOUSING CENSUS: SOME STRATEGIES FOR DATA COLLECTION DESIGN IN THE ERA OF LOW RESPONSE RATE AND HIGH RESPONSE BURDEN. AN APPLICATION OF A DECISION TREE MODEL.[1]

Manuela Bussola, Novella Cecconi, Elena Donati, Linda Porciani

Abstract. The Italian National Institute of Statistics (Istat) has adopted the methodology of permanent Population and Housing Census in 2018, which is planned as an annual cycle. This study aims to identify the profile of respondents and non-respondents to the Census, in order to put respondents at the centre of the design and management of data collection and to develop specific strategies focused on subsets of the sample population. The methodology used was the 'Tree decision model', which was able to present the relative weight of the included variables on the independent variable through a good and direct visual layout. The first results confirm the relevance of socio-demographic variables, such as the age structure, education level, and occupational status of the sampled household, and present some new analytical insights capable of describing first the profile of non-respondents and then the profile of respondents by channel.

1. Introduction

The information collected by the Census does not simply provide a count of the people living in a given area but allows a picture to be drawn of how different groups, with specific characteristics, are located within the national territory. Moreover, census data form the statistical pillar of other official surveys and the backbone of the knowledge about different aspects of a population: households, transport systems, schools, hospitals, neighbourhoods and cities.

The essential need for this information must be combined with the challenges that are driving changes of the Census strategies (from a traditional to a permanent strategy): the management of a reduced financial budget, the respect of the timeliness of the data information; the decrease in the response rate. The answers are: a

[1] The paper was carried out by the joint work of the authors. More in detail, the single paragraphs are attributed as follows: paragraphs 1 and 3.1 to Manuela Bussola; paragraphs 2, 3, 3.2 to Novella Cecconi; paragraphs 2.1 and 4 to Elena Donati; paragraph 3.3 to Linda Porciani.

sampled, annual and integrated (both as data sources, survey and register, and as techniques, CAWI, CAPI and CATI) Population Census.

One of the most focused aspects of the Census process is the web response rate, because of its potentialities for responding to the real challenges: the 2022 edition of the Population Census shows a CAWI response rate of 44.3 percent (48.7 percent of responding households), 3 to 5 points lower than previous editions, with a huge difference between northern and southern regions. According to the literature, this gap could be due to several factors: territorial digital divide, different proximity of institutions to citizens in the territory, different organisational environment of municipalities regarding census operations, delays in the delivery of information letter, different living contexts and characteristics of households.

A major challenge is therefore to understand how to increase the CAWI response, by overcoming on the one hand the declining level of trust and confidence of people to respond to surveys, and on the other hand the lower participation of people living in hard-to-reach areas, those with low literacy levels, or those who by definition escape the official count, making the enumeration of the whole population a long, slow, and costly process (UNECE, 2021).

The aim of the present work is to outline the profile of responding and non-responding households in the 2022 Census of Population and Housing, in order to develop specific strategies aimed at increasing the household participation rate and, in particular the CAWI response rate.

To this end, the analysis of the characteristics of the survey sample is based on three aggregates by the most relevant variable: sampled households by response rate; respondents by technique; non-respondents by those contacted by the survey network.

This study presents the methodology used, and the main results of the three aggregates. At the end, it is possible to define some specific profiles of the census respondents, which could be the core for improving future editions of the census.

2. The model: the decision tree model

The population of sampled households in the Census is very heterogeneous in terms of socio-demographic characteristics, household conditions and place of residence.

In order to classify this population, a decision tree model is used because it allows to have homogeneous subgroups of predictors for the three aggregates: sampled households, respondents and non-respondents.

The classification algorithm is CHAID (CHI-squared Automatic Interaction Detection), which detects the interaction between variables in a dataset and it is best suited to the objectives of the analysis and the nature of the variables selected. CHAID identifies discrete groups and then, by examining the responses to the explanatory variables, attempts to predict the effect on the initial variable (KASS, 1980).

This multiple segmentation technique is based on the $\chi 2$ test to test the hypothesis of statistical independence between the dependent variable and the explanatory variables. For each modality of selected variables and for each combination of modalities, the model generates a contingency table (starting from the dependent variable) by calculating the $\chi 2$ and the corresponding p-value.

The Xi attribute with the smallest p-value (pmin) is compared to the threshold value α (this could be the maximum tree size, the maximum number of levels, or the minimum number of elements in a node):

If pmin$< \alpha$ the modality Xi is considered as an attribute of the partition
If pmin $> \alpha$ a leaf is identified.

If an attribute has more than 2 values, the model allows to group them in order to have homogeneous values with respect to the dependent variable.

2.1. The variables selected

The total number of sample households in 2022 is 998,540, of which 58,952 are non-target (moved, deceased, homeless), of the remaining 939,588 households, 874,976 were contacted and 64,612 were not reached by the survey network.

The analysis was carried out on "valid" sampled households (939,588 units) by response rate, on respondents (855,595 units) by channel; and on non-respondents (84,293 units) by whether or not they were contacted by the survey network.

Specifically, a dependent variable has been defined for each of the three aggregates: the response rate for sampled households; the percentage of response by channel for respondent households; the contact rate for non-respondent households[2].

The first step in the analysis is to define the theoretical dimensions behind the Census response rate. The propensity to respond to the Census is at the core of NSIs' the data collection strategies of the (ONS, 2011). In Italy in particular, it has been studied since the traditional population census in 2011 (Bernardini *et al*., 2014).

According to the literature (ISTAT, 2022a) and the working experience of the research group, they could be classified into: 1.the living context of the household, 2. the organizational characteristics of the Municipal Census Office (MCO), 3. the socio-demographic characteristics of the sample households.

The first dimension includes geographical variables in the form of administrative boundaries and the Inner Classification[3], which takes into account both the demographic dimension and the proximity of citizens to essential services (health, education and mobility) (ISTAT, 2022b).

The organisational dimension of the MCO is represented by the presence or absence of a statistical office, the number of sampled households per operator, the size of the MCO in relation to the sample size, the typology of participation in the census (every year/only one year in the census cycle), the completeness of the training activities of the MCO operators.

The third dimension, the socio-demographic characteristics of the households, is described as age structure, citizenship of household members, educational level and occupational status.

Other contextual information is also taken into account: trust in institutions, use of the Internet, sending of information letters, postal monitoring, and declaration of receipt of Istat letters by households.

[2] In 2022 there were 855,595 responding units and 84,293 non-responding units: 416,476 households responded via web, 81,932 households responded via telephone interview, and 356,887 were interviewed face-to-face by a municipal surveyor/operator. Among the nonresponding households, 19,681 were contacted at least once by a municipal surveyor/operator and 64,612 were never contacted.

[3] Inner Areas classifies municipalities according to their distance from three essential services: health, education and mobility. The map of Inner Areas is a tool that looks at the entire Italian territory in its articulation at the municipal level and identifies municipalities with a combined supply of three types of services - health, education and mobility - named Multi Municipality Service Center/Single Municipality Service Center. It also presents all other municipalities according to their distance from these Centeres (in terms of actual average road travel time), and classifies them into four bands of increasing relative distance – Belt Area, Intermediate Area, Remote Area, Ultraremote Area - and, therefore, with potentially greater inconvenience in accessing services. The municipalities classified as Intermediate, Remote and Ultraremote represent all of the Inner Areas of our country (www.istat.it/it/archivio/273176).

Based on the selected variables (Table 1), composite variables were developed to optimize the analysis without losing core information. Some other variables were excluded because of their excessive complexity or low variability.

The strategy of reclassification of variables was relevant for the dimension related to the household. In particular, it is necessary to construct variables for multi-component households that combine information on the age, citizenship, occupational status, and educational level of the members. The variable "Households by generation" takes into account both the number of members and their age (before/after the baby boom, i.e. is 1964) (ISTAT, 2017). The educational level has been identified as the highest level of education of the household members; and the same process has been adopted for the occupational status, defined as the most qualified occupation in the family.

Table 1 – *The selected variables.*

Household	**Territory**	**Fieldwork organisation**
Types of households by citizenship	**Territorial aggregation**	**Statistical Office in Municipalilty**
- Italian	- North	- Yes
- Non Italian	- Centre	- No
- Italian + non Italian	- South	
Households by generation	**«Inner Areas»**	**Field Workload**
- 1 member until baby boom	- Single-Municipality Service Centre	*[households for interviewer]*
- 1 member after baby boom	- Multi-Municipality Service Centre	- Less than 50
- More than 1 member until baby boom	- Belt Area	- 50 - 100
- More than 1 member after baby boom	- Intermediate Area	- 100 - 150
- More than 1 member mix	- Remote Area	- 150 - 200
Occupational status	- Ultraremote Area	- 200 - 250
- Low		- More than 250
- Medium		
- High		
Educational level		
- Low		
- Medium		
- High		

3. The main results

The main results can be represented through three tree decision models focused on:

1. the profile of households by response rate;
2. the profile of households by channel of questionnaire return
3. the profile of households not responding by contact with survey network.

3.1. The profile of sampled households by response rate

Figure 1 shows the results of the *decision tree* where response or non-response to the Census is considered as an independent variable[4].

The most significant variable for response rate is citizenship. Italian households are more likely to respond to the census. On the other hand, a quarter of all foreigners or mixed households did not respond.

The tree decision model makesit possible to identify specific significant variables that affect the three groups: Italians, foreigners and mixed households. Italian households are more influenced by the organisational context of the MCOs: they have a higher propensity to respond if they live in a more structured MCOs; while foreign households are influenced by the age composition of the household, and finally the response rate of mixed households is more related to the territory, according to the classification in "Inner Areas."

For Italian respondents, it is common that a higher workload for the municipal operators corresponds to a higher level of non-response: a high percentage of non-responding households have not been contacted by the local survey network, and have only received postal reminders and via the “IO App”[5] carried out by Istat. This is the "hard core" of people who are not used to responding online and who need to be "reminded" to respond.

On the other hand, if the workload of the MCOs is not particularly heavy, the area is likely to have an impact on the Census participation; while if the workload is high, the presence of a statistical office, with staff dedciated to statistical activities and probably better organised, facilitates contact and participation by Italian households, more than in a municipality without a statistical office.

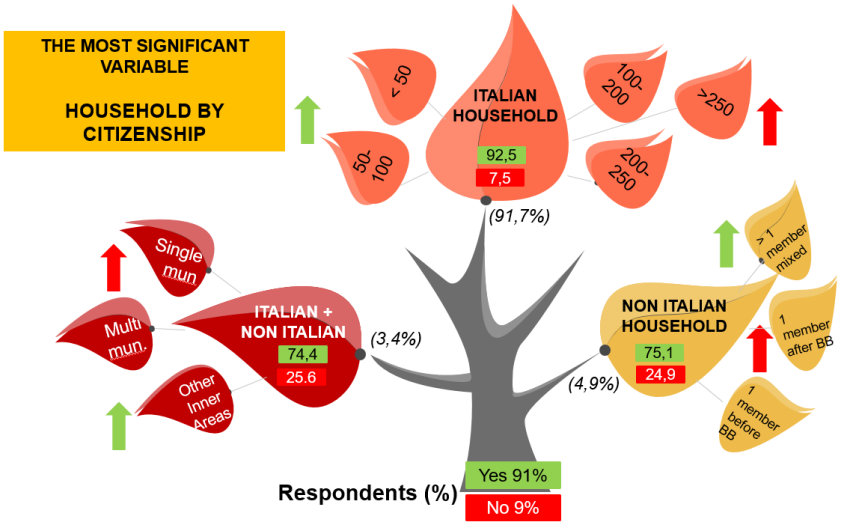
Foreign households, on the other hand, are the least likely to respond if the household consist of one person, regardless of the age: almost half of the one-person member households did not respond to the questionnaire. In particular, foreigners living in Single Municipality Service Centre (SMSC), Multi Municipality Service

[4] Respondents are 91 percent and non-respondents are 9 percent, the latter adjusted for non-targets.
[5] Io App is the APP for the communication among citizens and public administration.

Centre (MMSC) (both characterised by little or no distance from the three basic services) or Ultraremote Area (UA) (municipalities more distant from basic services) are more reluctant to partecipate in the Census.

On the other hand, living in municipalities with essential services related to health, education, and mobility seems to be a barrier to responding for mixed-citizenship households: the response rates are 65.5 percent of households in SMSC, 72.3 percent in MMSC and almost 80 percent in more remote areas.

Figure 1 – *Decision tree sampled households by response rate.*



3.2. The profile of respondents by channel of questionnaire return

The second *decision tree* was applied to the respondents according to the technique used to answer the questionnaire: CAWI, CATI, CAPI.

Educational level is the most important variable in the choice of response mode to the Census. As the level of education increases, so does the likelihood of responding online: almost 65 percent of households with at least one member with high level of education opted to complete the questionnaire via web, while 34 percent of households with a low level of education opted for a face-to-face interview at home or in Municipal Survey Centres (MSCs). The telephone is the less used channel and, moreover, it is almost exclusively an “outbound” response, i.e., due to

interviewer's reminder rather than the household's own request. The use of CATI increases as the level of education decreases.

For the most and least educated household the influence of the living places on the choice of the channel used to complete the census form is common. Living in municipalities provided with basic services (or not far from them), increases the propensity to respond via web. In more peripheral areas, the propensity decreases by at least ten percentage points.

Even if the effect of living places is the same, the response rate differs according to the level of education: the most highly educated households living in SMSC and MMSC show an increase in the CAWI response, especially if the highest occupation in the household is medium-high (about 70 percent of the households living the "central" municipalities).

On the other hand, more educated households living in a peripheral territorial context are more influenced by the number of members and generation: multi-component households composed of people born after the baby boom prefer web-based responses.

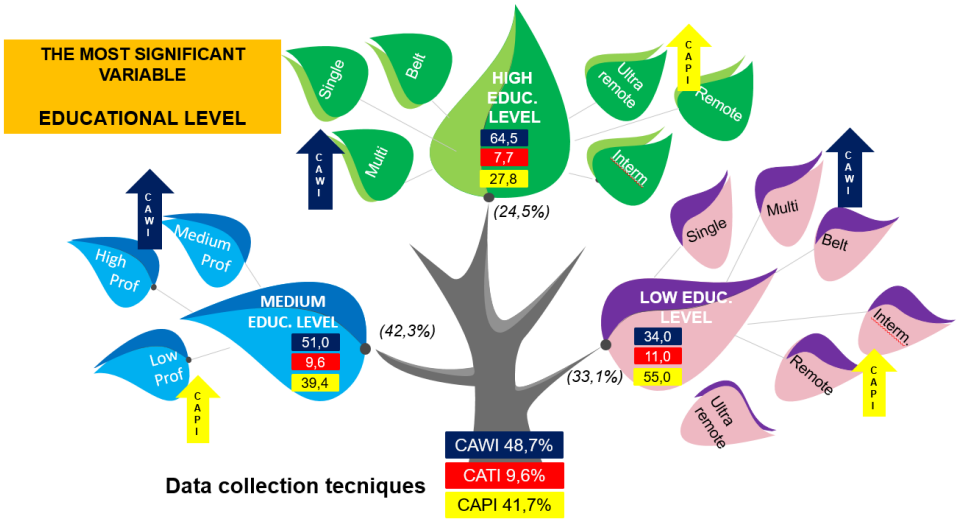
Finally, education level and living context are the most important variables in the choice of the web channel, regardless of the age of the respondent.

For households living in SMSCs, regardless of the number of household components, the propensity to use the face-to-face interview is higher for younger households, while the propensity to use the web is higher for older households. It is possible to observe the same factor influencing the web response of low educated households in the other territorial contexts, except for those living in MMSC, which are more affected by citizenship.

Occupational status is a significant factor in web use for households with a high school degree as the highest educational qualification.

A higher percentage of CAWI respondents is observed when the highest occupation is medium (64.5 percent) or a high (56.1 percent). The second influencing factor in this "leaf" is the living place: households living SMSC show a high propensity to use the web channel.

Figure 2 – *Decision tree responding households by channel of questionnaire return.*



3.3. The profile of non-respondents by contact with the survey network

The third *decision tree* was applied to non-responding households with the aim of describing their characteristics and to optimising the design process, especially at the stage of planning targeted field recovery activities.

The majority of non-responding households were not contacted by the local survey staff, especially if the households lived in SMSC and MMSC, where 80.0 percent of non-responding households had no contact with the survey staff: 83.6 percent and 79.8 percent respectively.

The probability of being contacted increases for non-respondent households living in the Belt or Intermediate Areas: 15 percent and 11 percent higher respectively compared to Single/Multi Centre.

Finally, households living in more remote municipalities (representing 6.8 percent of non-respondents) are also more likely to be contacted, at least once, by the municipal survey network: 71.8 percent of peripheral households escaped any kind of contact with census staff.

Following the tree model, the second most significant variable is the workload of the municipal census network, which is positively correlated whit the absence of contact with the sampled household.

In the case of the SSMCS, this relationship is evident: the higher the number of households assigned per operator, the higher the percentage of non-responding never

contacted households. When the local census network is overloaded, more than 250 households per operator (0.7 percent of the initial population) were no reached by the network.

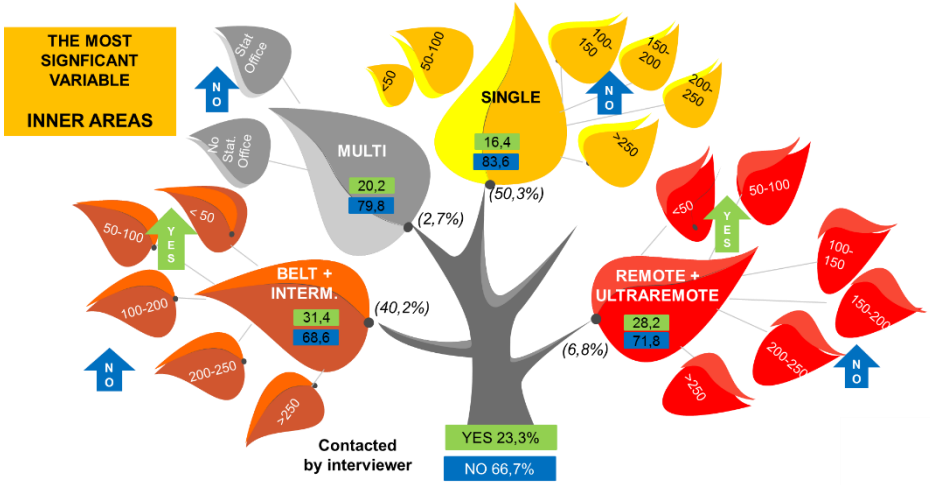
Households living in the Belt or Intermediate Areas (40.2 percent) show a lower non-contact rate when the workload is lower; in particular, the probability of being contacted is higher when an operator manages than 100 households, and when the workload is minimal (less than 50 households per operator). In the latter case, 41.7 of non-respondents are contacted at least once.

Looking at the more distant leaves of the tree, another important factor is the generational composition of the households.

In general, households with a one-member born after the baby boom are the most difficult subgroup to reach, regardless of the workload per operator. In fact, in the municipalities with the highest proportion of non-respondents, the generational composition becomes the second significant factor explaining the probability of not being contacted: 85.3 percent and 88.0 percent respectively in the municipalities with low and medium workload municipalities.

The same influence is present in the suburban and ultra-peripheral municipalities, where, although the presence of non-respondents is lower than in the Single and Multi-Service Centre, households with one or more members born after the baby boom are more likely to escape contact than other households (75.1 percent against 64.6 percent in municipalities with 50 to 100 households per operator).

Figure 3 – *Decision tree households not responding by contact with survey network.*



4. Conclusions

The Population Census is the largest statistical operation carried out by the NSIs and it presents complexity in each phase of the statistical process design and management; it should face the societal changes in terms of work organisation and household behaviour. These are some of the reasons of the need to have study and reaserch about Popolutation Census data and process to improve the quality and to update the process steps. The present study provides some insights based on the census results.

From the analysis of the *decision trees*, it is clear that the behaviour of non-respondents and respondents and the survey techniques show a homogeneity that crosses geographical boundaries. This means that strategies for implementing Computer-Assisted Web Interviewing (CAWI) and optimising non-response recovery can be applied across the country, following paths that do not necessarily correspond to administrative boundaries.

The results show us that the core of a renewal of the design process should be the characteristics closest to people's lives and to the organization of fieldwork. In particular:

- a high workload has a negative impact on both response rates and the probability of households being contacted;
- households with at least one foreign member have a low propensity to complete the census questionnaire, especially if they are single-member households or households living in Single Municipality Service Centre or Ultraremote Area;
- living in a Single Municipality Centre has a negative effect on both response rates and the probability of households being contacted, while at the same time, it has a positive effect on the use of the CAWI technique.

These results suggest that the planning of the future censuses edition could be enriched by the inclusion of some new elements in order to increase the response rate, especially the web response, and to reduce the number of sampled units not contacted:

- reducing the workload of the local survey network;
- an improvement in communication campaigns targeted at households with low levels of education and suburban contexts;
- the introduction of a smartphone questionnaire: "responsive" and accessible via QR Code;
- increasing of the contact information of the sampled units: i.e. mobile phone numbers available to the local census staff;
- setting up of an on line booking system for the sampled units to make an appointment to complete the questionnaire.

Further and detailed analysis of the different profiles of sampled households should be developed to be able to better manage this complex operation.

References

BERNARDINI A., FASULO A., TERRIBILI M.D. 2014. A model based categorisation of the Italian municipalities based on non-response propensity in the 2011 census, *Rivista Italiana di Economia Demografia e Statistica*, Vol. LXVIII, No. 3/4 (Luglio-Dicembre 2014), pp. 77-84.

ISTAT, 2017. *Rapporto annuale 2017*. Roma: Istat.

ISTAT, 2022a. *BES | 2021. Il benessere equo e sostenibile in Italia*. Roma: Istat.

ISTAT, 2022b. *La geografia delle aree interne nel 2020: vasti territori tra potenzialità e debolezze*. Statistiche Focus. Roma: Istat.

KASS G.V. 1980. An explanatory tecnique for investigating large quantities of categorical data, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 29, No. 2, pp. 119-127.

UNECE, 2021. *Keeping count. Conducting censuses during the covid-19 pandemic*. Geneve: Unece.

ONS, 2011. *Predicting patterns of household non response in the 2011 Census.* London: ONS.

Manuela BUSSOLA, Italian National Institute of Statistics, bussola@istat.it
Novella CECCONI, Italian National Institute of Statistics, ncecconi@istat.it
Elena DONATI, Italian National Institute of Statistics, eldonati@istat.it
Linda PORCIANI, Italian National Institute of Statistics, porciani@istat.it