GIORNATE DELLA RICERCA IN ISTAT | 10-11 NOVEMBRE 2014

SESSIONE II
PREVENZIONE, VALUTAZIONE E TRATTAMENTO DEGLI ERRORI NON CAMPIONARI

E-census in Italia: un'analisi regionale dei tassi di risposta web

Linda Porciani, Luca Faustini, Alessandro Valentini e Bianca Maria Martelli

Istat

E-census in Italia: un'analisi regionale dei tassi di risposta web

Linda Porciani, Luca Faustini, Alessandro Valentini, Bianca Maria Martelli

Istat

porciani@istat.it; faustini@istat.it; alvalent@istat.it; bmartelli@istat.it

Sommario

Negli ultimi anni anche nel campo della statistica ufficiale, l'adozione di tecniche web per la raccolta dei dati si è sviluppata in maniera significativa grazie alle evoluzioni tecnologiche, ai cambiamenti negli stili di vita, alla diffusione dell'uso di Internet da parte dei cittadini, delle istituzioni e delle imprese. Un altro fattore determinante nella diffusione del web nelle surveys riguarda la crescente necessità di ridurre i costi di rilevazione. tenendo conto di questi aspetti l'Istat ha adottato tecniche miste di rilevazione nella raccolta dati nella tornata censuaria del 2011. Questo lavoro ha lo scopo di illustrare i tassi di risposta web e alcune determinanti da rintracciare nel tessuto socio-demografico, ponendo particolare attenzione ai differenziali territoriali. A questo proposito un caso di studio è rappresentato dalla Toscana, dove si è registrato il più basso tasso di risposta web nonostante l'elevata diffusione delle ICT tra i cittadini, le imprese e la pubblica amministrazione. I risultati dell'analisi sottolineano l'importanza della cooperazione tra l'Istat e le amministrazioni locali come fattore chiave per incrementare la qualità dei dati e ridurre i costi. Inoltre gli stessi potranno essere utili per una migliore pianificazione del prossimo censimento permanente.

Parole chiave: Censimento della popolazione e delle abitazioni, strategie multicanale di rilevazione, valutazione

Abstract

In the last years several technological innovations affected survey designs and data collection methods in Official Statistics. It has been spurred by changes in lifestyles, a wide spread use on Internet by people and enterprises, and the development of e-government. A further factor supporting this trend is the increasing pressure to find effective methods to reduce costs. In this framework, the Italian National Institute of Statistics (ISTAT), in collecting data for the whole 2011 Census wave, strengthened its mixed mode approach introducing also the web techniques. This paper aims at illustrating the challenges of the web application in Population and Housing Census and Enterprises Census, devoting particular care at investigating the differences in territorial web response rates. An interesting case study is Tuscany, due to its low web response rates associated with a high ICT penetration rate. Results can be useful for better planning the forthcoming (rolling) census and for highlighting the cooperation between NSI and local administrations as key factor to improve data quality and reduce costs.

Key words: Population and Housing Census**,** Web Mode Data Collection Mode, Process Evaluation

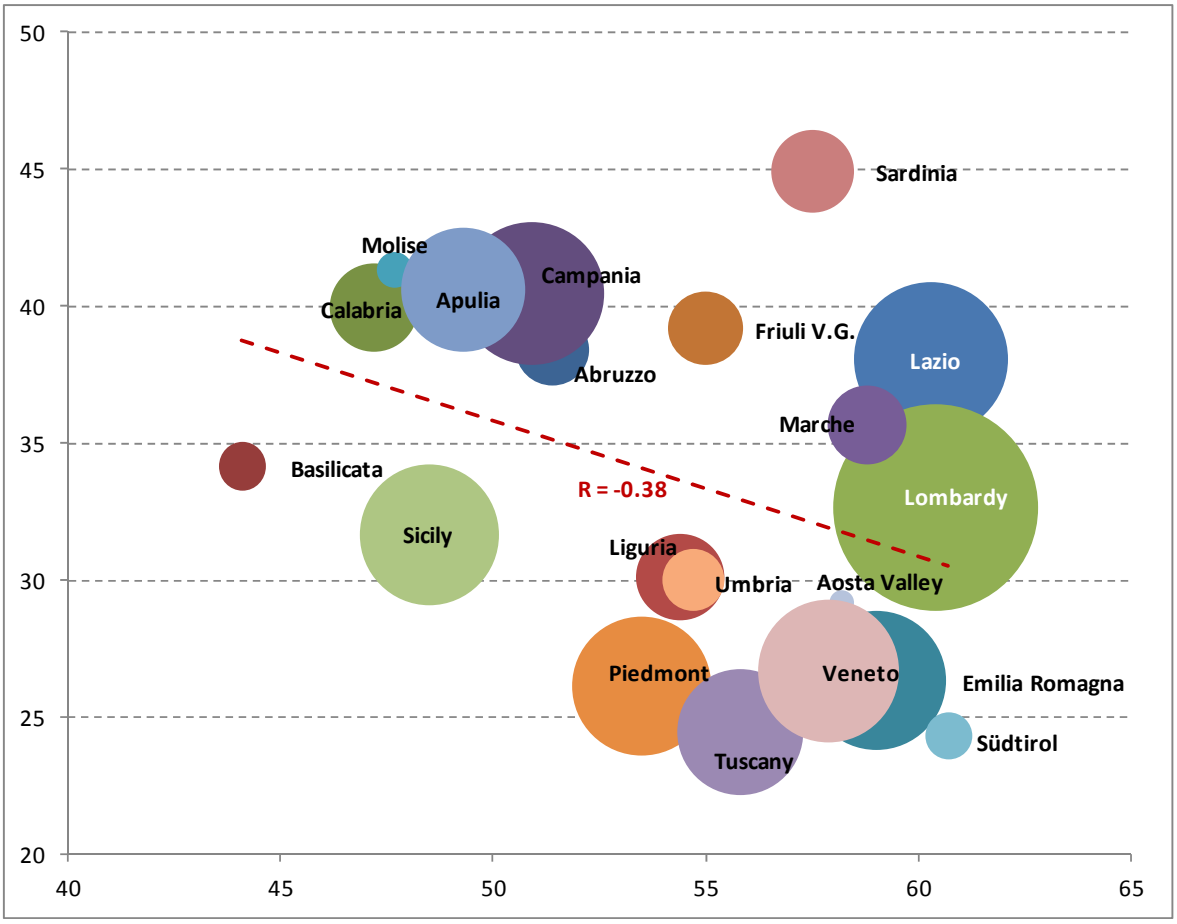
General framework

The Italian National Institute of Statistics (ISTAT) adopted web technologies with the general purpose to improve data quality, reducing costs and maximizing data timeliness and accuracy (ISTAT, 2007). These choices marked the transition from the traditional door-to-door census to the e-census. Specifically, web techniques affected the data collection processes of the censuses conducted since 2010, namely Agricultural Census, Population and Housing Census (PHC), Enterprises Census (EC) and Non-Profit Institutions Census (NPC), and Public Institutions Census. (ISTAT 2010, 2011, 2012). In all cases, except for the Public Institution Census, returns of the questionnaires followed a mixed mode approach – web or paper– according to the choices of respondents. For the Public Institution Census web was the only reply mode allowed. Furthermore, a web tool called SGR (Survey Management System) has been organized in order to monitor each step of the data collection process. In this framework, the main role of ISTAT Territorial Offices – regional branches of ISTAT – was to guarantee a constant monitoring of all census operations. A team of people specifically devoted to census activities supervised the data collection network and guaranteed the training of operators. Furthermore, after the end of the PHC, EC and NPC data collection process, census operators replied to an on line questionnaire (named IvalCens for PHC and IvalCis for EC and NPC) focused on the evaluation of technical, organizational, and methodological innovations, including the adoption of web techniques. Given the increasing importance of the web in the near statistical future, a core question considered in this paper is represented by the analysis of features of the web respondents, both in terms of geographical and individual characteristics as items able to drive the web response rates, apart from the effective availability of internet connection among firms and population.
The paper has been organized as follows: the first chapter has been devoted to the description of data and methods, the second one describes the main results of the applied models and illustrates the case-study of a specific Italian region, Tuscany, which had the lowest web response rate during PHC,EC and NPC; and finally the last chapter debates some thoughts for future census planning operations.

1. Analysis of web response: data and methods

PHC, EC and NPC represent the Censuses where firstly the mixed data collection mode (DCM) was adopted on a large scale in Italy. Studying the impact of this innovation could be a key factor for better planning future censuses. Indeed recently, thanks to a specific law (D.L. 83/2012), Istat officially introduced the rolling census methodology (U.S. Bureau of Census, 2001). PHC rolling census will be completely paperless; it will start in 2016 and will become fully operative in 2020.
A first research idea was the investigation of the relationship between raw web response rate (WRR), the most suitable kind of response rate to monitor the quality of a DCM process (Martelli B., 2005), and the ICT penetration rate (Istat, 2013). WRR for Population Census and Internet penetration rate shows a weak and inverse correlation level (r=-0,38 for Italy) suggesting to focus the analysis on other factors in order to better explain the general behavior of web respondents observed in Italy during the last censuses (Figure 1).

Figure 1 – ICT penetration rate (horizontal axis) and WRR of PHC (vertical axis) in the Italian regions (NUTS 2). Percentage values. Bubbles size is proportional to the households' number



The driving idea is that web response could be linked to the specific characteristics of the complex social system where survey units, persons, household and enterprises are settled. As literature shows (Bech M. et al., 2009), socio-demographics features influence the web propensity to surveys reply: gender, age, income, education level, civil status and health status are some items often included in this kind of analysis. Unfortunately, at present a release of the complete micro census dataset is not yet available; for this reason, covariates have been limited to some demographic and economic census data[1], the IVALCENS and IVALCIS ex-post evaluation surveys (Stassi G. et al., 2013), and the ICT survey data[2] (ISTAT, 2013).

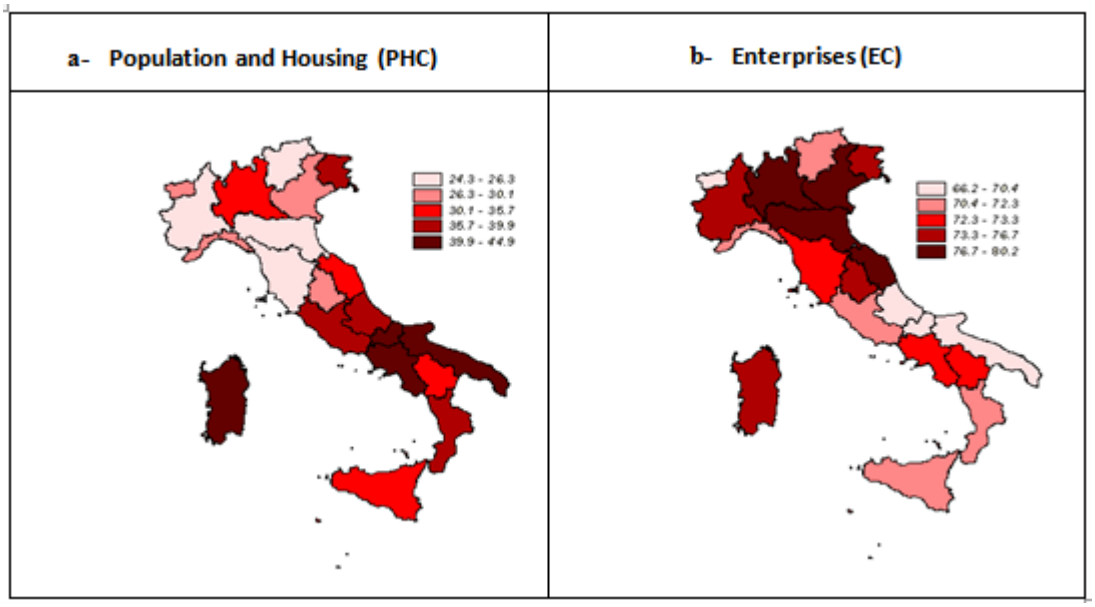
2. Main results

The web response rate mean by region is 33% in case of PHC and 74,7% in case of EC, even if a high variability across regions has been observed. As shown in Figure 2, WRR are not homogeneously distributed in the whole country. Following the regional level of analysis, areas with the highest WRR in the PHC are Sardinia (44.9%), Molise (41.3%), Apulia (40.6%) and Campania (40.5%). Vice-versa Trentino Alto Adige/Südtirol and Tuscany show the lowest recorded levels (24.3% and 24.5% respectively). In case of Enterprises Census, the region with the highest WRR is Veneto (80.2%), followed by Emilia-

[1] Census data are available on: http://dati.istat.it/
[2] Data are available on: http://www.istat.it/it/archivio/48388

Romagna (78%). The lowest levels are those of Molise (66.2%) and Val d'Aosta (68.1%). Moreover, WRR of PHC and of EC are substantially uncorrelated (R^2=0.11). In fact, for PHC areas with the highest web response rates are located in the Southern part of Italy and in the Islands; for EC, regions with the highest WRR are located in the North-Eastern part of the peninsula. At the same time, for EC the lowest rates are located in the North for PHC and in the South.

Figure 2 - Web response rates for Population and Housing and for Enterprises census in Italian regions. Percentages: Quintiles distribution



WRR has been analysed at regional level through a logistic regression model applied to the odds of WRR with six covariates for the PHC and three covariates for the EC. The model is:

$$\log\left(\frac{\text{WRR}}{1-\text{WRR}}\right) = \beta_o + \sum_{i=1}^{n} \beta_i x_i$$

All explaining variables are dichotomized using the median value and the baseline group (reference) is that with values lower than the median.
The covariates for PHC analysis are the following[3]:
- Ageing index[4] [Age]. The median value is 159.4: cut off between "young" and "old" regions.
- Quota of foreigners [Foreigners]. The median value is 8%, cut off between low and high presence of foreigners.
- Average family size [Family]. The median value is 2.3, cut off between large and small family size.
- Quota of municipalities (LAU 2) over 20,000 inhabitants [Large Munic]. The median quota is 4.3%, cut off between urban and rural regions.

[3] In square brackets the name of variable used in the model
[4] Ageing index is the ratio between the population over 64 years and the population less than 15 years.

- ICT users' quota [ICT]. The median is 54.9%, cut off between cabled and not cabled regions.
- Ex-post evaluation level of web mode data collection [Evaluation]. The evaluation surveys collect data about the assessment of census operators also about the web data collection mode in a scale between 0 (minimum appreciation) and 3 (maximum appreciation). The median value is 2.4; regions with a higher score have a good appreciation of the web channel.

The covariates of the EC model are the followings:
- Quota of foreigner entrepreneurs [Foreigners]. The median value is 1.5%.
- Quota of enterprises with 10 or more employees [Large Enter]. The median value is 4.4%.
- ICT users' quota [ICT], the same variable used in PHC model.

Table 1 and Table 2 show the results for the PHC and EC model. It is interesting to note that results are quite similar for ICT use and size of municipalities (or enterprises). Areas where web use is more spread have higher level of WRR. As a consequence, increasing the use of technology will probably boost the web survey response rates. According to the latest data released by Istat (Istat, 2013), the quota of web users markedly increased in the last years (more than 5%), this should imply a rise in WRR. Furthermore around 86% of families with a child less than 18 years use Internet at home. Vice versa only 13% of alone elderly (65 years and more) has a connection to Internet. Policies aimed to promote the use of Internet especially for old people could be effective for the use of web in official surveys and censuses.

A second insight regards the size of enterprises or municipalities, which has a positive correlation with WRR. Probably, this type of correlation is affected by some typical organizational "biases" such as: the higher the number of units to collect (or their complexity), the more significant the actions realized by census operators to promote web compilation. Indeed, management of web questionnaires respect to paper ones is easier and faster for census operators, so it is very convenient for them to support the web strategy. Instead census operators which work in areas with simpler or less numerous organizations have to manage simpler questionnaire and they can easy do it by hand. In planning new surveys it should be also important to take into consideration the workload for survey units and for the different actors involved in the data collection process, such as municipalities or chambers of commerce, in other words at micro and at macro level.

Finally individual characteristics such as citizenship or age affect the use of web. Foreigners and old people tend to have in general a weak approach to web: this is confirmed by applied models. So, in a society with an increased presence of foreigners and elderly targeted actions focused on those sub set of population could be the keystone to increase WRR.

Table 1 – Results of the model for PHC. Italy

Code	Name Variable	Parameter	Estimate (β)	P-value	Effect:Exp (β)
	Intercept		-0.8517	<.0001	
[Age]	**Ageing Index**	Age>159.4	-0.1011	<.0001	0.9038
[Foreigners]	**Foreigner quota**	Foreigners>8.0	-0.1234	<.0001	0.8839
[Family]	**Average Family Size**	Family>2.3	0.0553	<.0001	1.0569
[Large Mun]	**Municipalities quota**	Large Munic>4.3	0.0116	<.0001	1.0117
[ICT]	**ICT users quota**	ICT>54.9	0.1163	<.0001	1.1233
[Evaluation]	**Ex-post evaluation level of web DCM**	Evaluation 2.4	0.3854	<.0001	1.4702

Source: our elaboration from SGR data

Table 2 – Results of the model for EC. Italy

Code	Name Variable	Parameter	Estimate (β)	P-value	Effect:Exp (β)
	Intercept		-0.9268	<.0001	
[Foreigners]	**Foreigner quota of ent.**	Foreigners>1.5	-0.0366	<.0001	0.9641
[Large Enter]	**Quota over 10 employees**	Big Enter>4.4	0.2818	<.0001	1.3255
[ICT]	**ICT users quota**	ICT>54.9	0.0296	<.0001	1.0300

Source: our elaboration from SGR data

2.1 The Tuscany web response in case of Population and Housing Census

At the local level, Tuscany shows a low web response rate to PHC census in comparison to other Italian regions, despite the presence of two meaningful factors: (i) the field work aimed at supporting the web use performed by the staff of the territorial office of the Istat; and (ii) a high diffusion and use of Internet services, higher than at national level. In particular, households with an Internet connection by various channels (mobile, fixed phone, ADSL and so on) are 62.2% in Tuscany and 60.7% in the whole country, persons aged 3 or more able to use Internet services (communication services such as sending and receiving e-mails and phoning; creating and updating social networks and blog; discussing on line about political and social themes and so on) are 56.9% in the region and 54.3% in Italy (ISTAT, 2013). In other words, inhabitants of Tuscany are likely to use Internet by tablet, smart phone and computer, but they don't use it to fill in the census forms.

The distribution of response rate by channel is the same at country and regional level for municipal data collection centers (31.7%) and enumerators (12.4%), while the use of web (33.4% in Italy versus 24.5% in Tuscany) and – conversely – the use of Postal Offices (22.6% versus 31.4%) are instead steadily differentiated. In Tuscany people prefer hand delivery rather than web compilation: the self-confidence to see the hands in which the form goes rather than the inconsistency of the web. This low web use suggests a deeper analysis of the reasons beyond this behavior through the application of the illustrated regression model at a detailed geographical level. The analysis has been conducted at municipality (LAU 2) level, for the 287 municipalities of Tuscany. Covariates are 4: [Age], expressed in terms of ageing index; [Foreigners], that is the quota of foreigners; [Family] as average family size; [Large Mun] which is the size in terms of inhabitants of municipalities. Again, each covariate was dichotomized using the median value as cut-off criterion. The model converged and the test of null hypothesis ($\beta=0$) was satisfied ($p<0.0001$) for the intercept and for all variables, except for quota of foreigners (Table 3).

The main effects in Tuscany are dual: large municipality and presence of foreigners have a positive impact on the web propensity; while living in a numerous family and aged context show an inverse relation with web use.

More specifically, LAU 2 with a number of inhabitants higher than the median (5,6 thousands) present an odds-ratio of the quota of Internet answers around 8% higher than that of the less populated municipalities. Instead, the quota of foreigners has an effect on the response variable, but of limited extension and not statistically significant: in LAU 2 where the quota of foreigners is over the median level (7.5%) the odds-ratio has an increase of only 1%. On the other side, in municipalities where household dimensions were higher than the median level (2.3 member per household), the effect on the odds-ratio is negative (-17%). An ageing index higher than the median (200.3%) shrinks the odds-ratio of around 8%.

Table 3 - Results of the model for PHC. Tuscany municipalities

Code	Name Variable	Parameter	Estimate (β)	P-value	Effect:Exp (β)
	Intercept		-10.856	<.0001	
[Age]	**Ageing Index**	Age>200.3	-0.079	<.0001	0.9240
[Foreigners]	**Foreigner quota**	Foreigners>7.5	0.0089	<.0001	1.0089
[Family]	**Average Family Size**	Family>2.3	-0.187	<.0001	0.8294
[Large Mun]	**Municipalities quota**	Large Munic>5.6	0.0783	<.0001	1.0814

Source: our elaboration from SGR data

Despite the lack of data about ICT penetration at LAU-2 level, this preliminary results show that official statistic has to take into consideration also social and demographic features in planning web mode or mixed mode methods to collect data. In particular, the local analysis show that for increasing the quota of web respondents, Istat should devote primary attention to specific subsets of population less familiar with new technologies: the traditional families, where the number of components is 3 or more and the elderly living in small LAU 2.
The model could be enriched as soon as further individual results of census will be available, extending the analysis to some others variables, among them educational level, professional skill, household typology (family with or without children or other relatives) and others. Furthermore, it could be good for deeper analysis having details about the web response processes, such as place of compilation, assistance received in compiling the form and other factors. These paradata could represent key factors to enrich the understanding of web DCM and to better plan the strategies of forthcoming Italian rolling census.

4 Concluding remarks

The sharp and to some extent unpredictable increase of Internet users in the very recent years led both private and public institutions to extensively adopt the web survey techniques for collecting statistical information. Web techniques, after a first phase of software investment, are convenient in terms of costs, timeliness and reduction of statistical burden. Also Census surveys are rapidly switching to this data collection mode, as reported in this paper since the 2010-2012 Italian experience. In the next future, Internet will be the only way to collect data for Italian "rolling" censuses and one of the major obstacles in setting up web surveys is represented by the low level of Internet penetration. Indeed, not all enterprises and households are connected. As a consequence, the response rate often results unsatisfactory, lowering the quality of the process. However, the last evidences from the Italian population and economic censuses give a unique opportunity to investigate further factors that could affect the web response rate, besides the Internet availability, beginning from the weak and negative correlation between ICT penetration and web response rate. In addition, the two censuses web response rates are uncorrelated. In the case of Population and Housing Census, regions with higher web response rates are in the South of Italy while regarding enterprises in the North of the country. This evidence suggested two ad hoc analyses. A regression model is thus applied selecting several socio-demographic and territorial covariates: i) for population: municipality size, age of population, foreigners' population quota, households' size, ICT users' quota and web mode evaluation level; ii) for enterprises: foreign entrepreneurs quota, firms size and ICT users quota. The model explained that for population census, WRR is positively affected by evaluation and ICT users' quota and, to a lesser extent, by family size and by quota of bigger municipalities. A negative effect has been played by age and foreigner population quota. For enterprises, the most significant variable affecting positively web response rate is the enterprise size (proxy

of the organization system), and to a lesser extent the ICT users quota. Again, the foreign entrepreneurship quota has a negative impact on the web response rates.
Then, in order to increase the web response rate, it could be useful to provide policies to support the dissemination of knowledge, and use of computer tools for foreign and elderly population, especially located in small domains (municipalities /enterprises). Further developments of analysis will be conditioned by the availability of additional information; as soon as more data will become available it will be possible focusing on further features besides the ICT literacy. This latter in fact is a very important motivation in web participation but also a numerous set of social, demographic, and territorial characteristics could play an important role to ensure an high participation rate.

References

Bech M., Bo Kristensen M.B. *Differential response rates in postal and Web-based surveys among older respondents*, Survey Research Methods: 2009, Vol.3, No. 1, pp.1-6.

Istat. La progettazione dei censimenti generali 2010-2011. Analisi comparative di esperienze censuarie estere e valutazione di applicabilità di metodi e tecniche ai censimenti italiani. Roma: 2007. (Documenti n.9/2007).

Istat. Istruzioni per la rilevazione del 6° censimento generale dell'Agricoltura. Handbook of census. Catanzaro: 2010, Rubettino Print.

Istat. Istruzioni per la rilevazione del 15° censimento generale della popolazione e delle abitazioni. Handbook of census. Roma: 2011. Postel.

Istat. Istruzione per la rilevazione – censimento dell'industria e dei servizi 2011. Handbook of census. Catanzaro: 2012. Rubettino Print.

Istat. Cittadini e nuove tecnologie. Roma: 2013 (Statistiche report,19 dicembre), http://www.istat.it/it/archivio/108009.

Martelli, B.M. Relationship between response rates and data collection methods. Paper presented at OECD Workshop / European Commission Working Group on Business Tendency and Consumer Opinion Surveys, Brussells: 2005, November.

Stassi, Giuseppe, Valentini, Alessandro. L'Italia del censimento. Struttura demografica e processo di rilevazione. Roma: Istat,2013.

U.S. Bureau of Census. Introduction to Census 2000 Data Products. MSO/01-ICDP. Washington D.C.:2001Available on http://www.census.gov/prod/2001pubs/mso-01icdp.pdf