istat
working
papers

Italian e-census: a regional analysis of web response

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

Istat
Istituto Nazionale
di Statistica

istat
working
papers

Italian e-census: a regional analysis of web response

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

Istat

Istat Working Papers

Italian e-census: a regional analysis of web response

N. 10/2015

Italian e-census: a regional analysis of web response[1]

Linda Porciani, Luca Faustini, Alessandro Valentini, Bianca Maria Martelli[2]

Sommario

Negli ultimi anni nel campo della statistica ufficiale l'adozione di tecniche web per la raccolta dei dati si è sviluppata in maniera significativa in seguito alle evoluzioni tecnologiche, alla diffusione di Internet tra i cittadini, le istituzioni e le imprese e alla crescente necessità di ridurre i costi di rilevazione. Tenendo conto di questi aspetti, l'Istat ha adottato tecniche miste di rilevazione nella tornata censuaria del 2011. Il presente lavoro ha lo scopo di illustrare i tassi di risposta web e alcune determinanti socio-demografiche dei differenziali territoriali. Un caso di studio è rappresentato dalla Toscana, dove il tasso di risposta web è stato particolarmente basso nonostante l'elevata diffusione delle ICT tra la popolazione. I risultati dell'analisi indicano alcune possibili strategie efficaci per incrementare l'uso del web come canale di risposta nelle indagini statistiche. Gli stessi potranno essere utili per una migliore pianificazione del prossimo censimento permanente.

Parole chiave: censimenti, strategie multicanale di rilevazione, qualità del processo

Abstract

In the last years technological innovations have affected survey designs and data collection methods in Official Statistics. This was due to changes in lifestyles - the widespread use of the Internet by citizens and enterprises - and the development of e-government. A further factor supporting this trend was the increasing need to reduce Official Statistics costs. The Italian National Institute of Statistics (Istat) introduced the mixed-mode approach for 2011 Census. This paper aims at illustrating the challenges due to the application of the web approach in the General Population and Housing Census and the Census of Enterprises, investigating territorial differences of web response rates. A case study is the Italian region Toscana, due to its low web response rates associated with high ICT penetration rate. Results can be useful both for increasing the efficacy of the web response process and for better planning the next rolling paperless census.

Key words: Italian Censuses, Mixed-mode Data Collection, Process quality

---

Index

1. General framework

The Italian National Institute of Statistics (Istat) has 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, General 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 questionnaires followed a mixed mode approach – web or paper – according to 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. Within this framework, the main role of Istat Territorial Offices 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 next statistical future, a core question considered in this paper is the analysis of web respondents' features , both in terms of geographical and individual characteristics as items able to drive the web response rates, apart from the effective availability of the internet connection among firms and population.

The paper has been organized as follows: the first chapter describes data and methods, the second one focuses on main results, the third illustrates the case study of Toscana, which had the lowest web response rate during PHC, EC and NPC; and finally the last chapter debates some strategies for future census planning operations.

2. Analysis of web response: data and methods

PHC, EC and NPC are the first attempts to adopt a mixed data collection mode (DCM) on a large scale in Italy. Studying the impact of this innovation could be a key factor for better planning future censuses. Indeed recently, due to a specific law (D.L. no. 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.M., 2005), and ICT penetration rate.

WRR could be defined as the quota of web responses on the total eligible survey units; while ICT penetration - in the present paper - has been measured following three approaches:

the quota of household with Internet connection;

the quota of household with at least a mobile connected to Internet;

the quota of people aged 14 and over which use Internet to communicate with Public Administration and in particular people sending documents to Public Bureaus.

In the first two cases, analysis quite surprisingly shows a weak and inverse correlation between WRR of PHC and ICT measures: respectively r = – 0.30 (Figure 1) and r = – 0.39. Considering the third and more specific definition of ICT penetration, the same correlation is positive and modest: r = 0.32. For our purposes, the following analysis will be performed according to definition a).

Figure 1 – ICT penetration rate (a) (horizontal axis) and WRR of PHC (vertical axis) in the Italian regions (NUTS 2). Percentage values. Bubbles size is proportional to the households' number
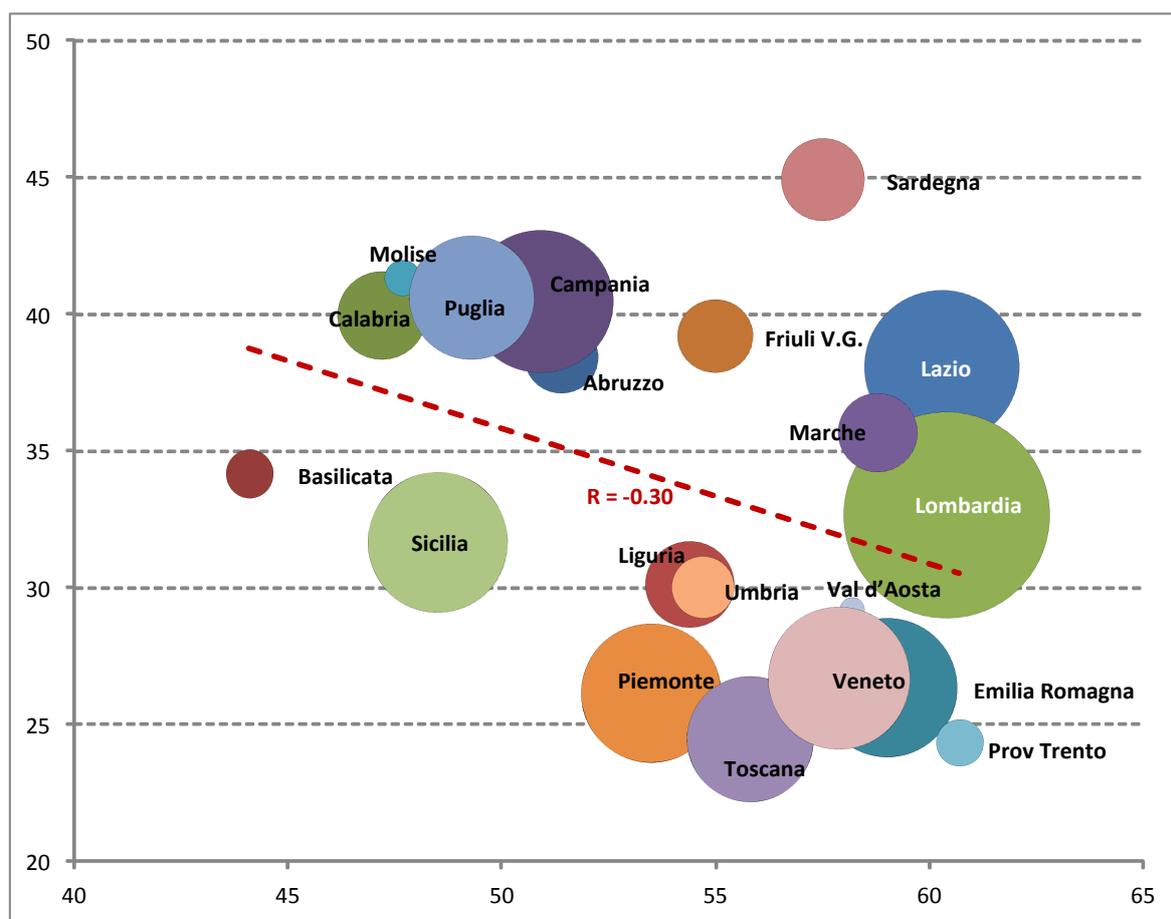
Source: our elaboration on Istat 2013, Statistiche Report n.19
(a) Quota of household th an Internet connection.

First results suggest a deeper analysis of this relation, mainly focused on the factors able to explain general behavior of Italian web respondents during the last censuses.

The driving idea to explain the determinants of WRR 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 items often included in this kind of analysis. Our data were gathered from various sources: census data, ex-post evaluation surveys (Stassi G. and Valentini A. 2013) and the ICT survey (Istat 2013). Unfortunately, linkage between individual web responses and individual charac-

teristics is not possible using these data. After a semantic analysis, variables were selected at an aggregate territorial level (NUTS2 and LAU2).

As well-known more developed societies display also a wider use of technology to communicate. This is the core of our research hypothesis, considering the web response process to official surveys as a kind of communication between citizens and public institutions.

To fulfil our task the following logit model, based on regional data[3], was applied both for population and enterprises:

$$\text{logit (WRR)} = \beta_o + \sum_{i=1}^{n} \beta_i x_i \qquad (1)$$

where WRR is the ratio between web and paper respondents and x_i are the covariates, specifically defined according to the main characteristics of the two censuses. Each covariate was dichotomized using the median value as cut-off criterion.

3. Main results

The mean WRR is 33% for PHC (namely HWR, Household Web Response rate) and 75% for EC (namely EWR, is Enterprises Web Response rate), even if a high variability across regions has been observed. HWR and EWR are not homogeneously distributed in the country (Figure 2).

Figure 2 – Household Web Response (HWR) and Enterprises Web Response (EWR) rates in Italian regions. Percentages: Quintiles distribution
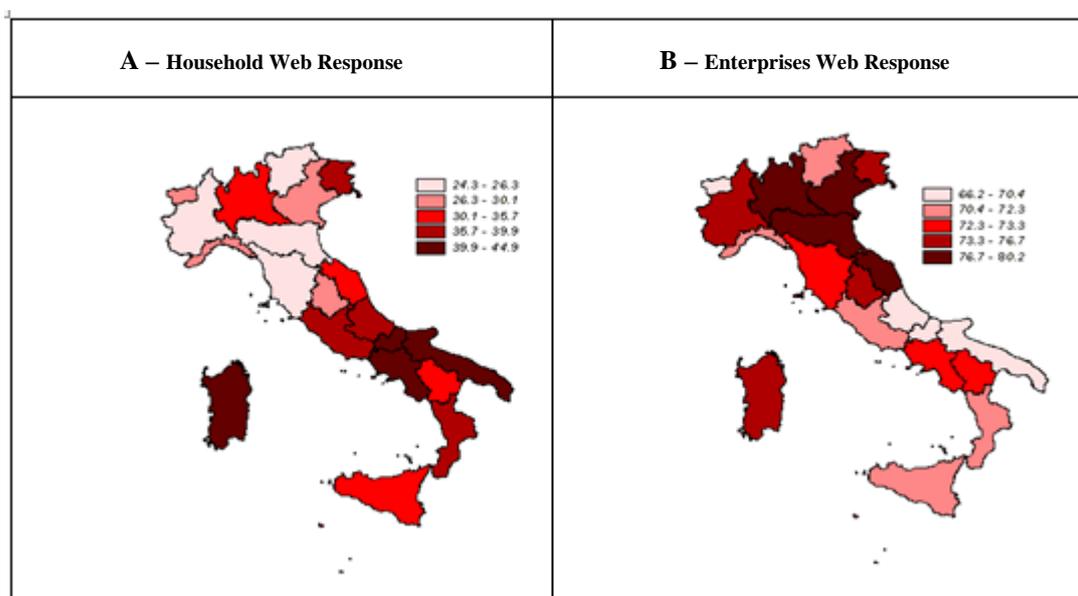
Regions with the highest HWR are Sardegna (44.9%), Molise (41.3%), Puglia (40.6%) and Campania (40.5%). Vice-versa Toscana and the Province of Trento show the lowest levels (24.5% and 24.3% respectively). For the Enterprises Census instead, Veneto and Emilia-Romagna are the regions with the highest EWR (80.2% and 78%); while Molise and Val d'Aosta display the lowest levels (66.2% and 68.1%). In general terms, the South of Italy and the Islands show the highest

_____

[3] Trentino Alto Adige has some specificities: for PHC it has been possible to include only the Province of Trento (Bolzano adopted a slightly different data collection process); for EC the two Provinces of Trento and Bolzano have been treated separately.

HWR ; and vice-versa, the North-Eastern part of the peninsula shows the highest EWR. . At the same time, the North for HWR and the South for EWR show the lowest rates. HWR and EWR are substantially independent (R^2=0.11).

To better study the variability of the web response rates distribution, two distinct logistic regression models are applied, with respectively seven covariates for HWR and five for EWR. Covariates for HWR analysis are the following[4]:

a) Quota of people with an academic degree on the population aged between 25 and 64 years [Degree]. The median value is 38.0%, which is the cut off between high and low educated regions.

b) Quota of foreigners [Foreigners]. The median value is 7.4%, which is the cut off between low and high presence of foreigners.

c) Quota of Head of Household aged 55 years and over on the total households [Age]. The median value is 54.1%, which is the cut off between young and old regions.

d) Average family size [Family]. The median value is 2.4, which is the cut off between large and small family size.

e) ICT users' quota [ICT] defined as the share of household with an Internet connection. The median is 54.9%, the cut off between cabled and not cabled regions.

f) Quota of municipalities (LAU 2) over 20,000 inhabitants [Large Munic], which is the cut off between high and low populated regions. The median quota is 4.3%.

g) Ex-post evaluation level of web mode data collection [Evaluation]. Evaluation surveys collect data about the assessment of census operators on 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.

Covariates a) and b) are proxy of individual characteristics; c), d) and e) are proxy of household characteristics; f) and g) are measures of the environmental effect.

Covariates of the EWR model are the followings:

a) Quota of foreigner entrepreneurs [Foreigners]. The median value is 1.7%.

b) Quota of entrepreneurs under 50 years [Age<50]. The median value is 60.4%.

c) Quota of enterprises (from 3 to 9 employees) which introduced innovations in the last three years [Innov]. The median value is 32%.

d) Quota of enterprises (from 3 to 9 employees) which are trading their products or services by electronic commerce [E-commerce]. The median is 24.6%.

e) Quota of enterprises with 10 or more employees [Large Enter]. The median value is 4.7%.

Covariates a) and b) are proxy of entrepreneurs characteristics; c) and d) are proxy of innovative enterprises; variable e) is a measure of enterprises complexity.

Table 1 and Table 2 show results for the PHC and EC model. It is interesting to note that results are quite similar for the effect of individual characteristics (age and citizenship), use of technology and environment variables (size of municipality or enterprises).

As a first insight, specific actions oriented to elderly people and foreigners are required to increase web rates. Indeed, the two sub-populations have a weak approach to web: this is confirmed both for population and enterprises. In other words, in a society with an increasing presence of foreigners and elderlies (65 years and more), targeted actions aimed at promoting the web response for these two groups of people could represent a central point to take in account. It is furthermore important to note that education has a relevant impact in HWR: the higher the quota of graduated people, the more relevant is the quota of internet respondents.

Secondly, areas where web use is more diffused have a higher level of HWR and EWR. As a consequence, increasing the use of technology will probably boost the web survey response rates. According to the latest data released by Istat, the quota of web users markedly increased during the year 2013 (more than 5%) and this should imply a rise in web responses.

_____

[4] In square brackets the name of variable used in the model.

Finally it is worth to take into account the positive correlation between the environmental variables (like municipalities and enterprises size) and web response rates. Probably, this type of correlation is affected by some typical organizational "biases" such as: the higher is the number of units to collect (or their complexity), the more substantial are the actions realized by census operators to promote web compilation, at least among a certain range of magnitude. 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 survey units can probably easily do it by hand.

A final consideration regards the positive effects connected to the evaluation of web use in census process by census operators. In PHC experience, the good appreciation of the whole census strategy, and in particular of the web tools, reported by census network operators determines the strongest effect on the HWR. In planning future surveys it should be very useful to continuously taking care of the cohesion of whole census network.

Table 1 – Results of the model for HWR. Italy

Code	Cut-off	Estimate (β)	P-value	Effect: Exp(β)
Intercept		-0.9482	<0.001	
[Degree]	38.0%	0.0336	<0.001	1.034
[Foreigners]	7.4%	-0.1261	<0.001	0.882
[Age]	54.1 %	-0.0539	<0.001	0.948
[Family]	2.4 components	0.1031	<0.001	1.109
[ICT]	54.9%	0.1106	<0.001	1.117
[Large Munic]	4.3%	0.0516	<0.001	1.053
[Evaluation]	2.4 points	0.4143	<0.001	1.513

Source: our elaboration from SGR data and from dati.istat.it

Table 2 – Results of the model for EC. Italy

Code	Cut-off	Estimate (β)	P-value	Effect: Exp(β)
Intercept		0.8540	<0.001	
[Foreigners]	1.7%	-0.1559	<.0001	0.856
[Age<50]	60.4%	0.1228	0.0003	1.131
[Innov]	32.0%	0.2027	<.0001	1.225
[E-Commerce]	24.6%	0.0596	<.0001	1.061
[Large enter]	4.7%	0.1556	<.0001	1.168

Source: our elaboration from SGR data and from dati.istat.it

4. The Toscana web response in case of Population and Housing Census

At local level, Toscana shows a lower HWR respect to the other Italian regions, despite the presence of two meaningful factors: (i) the relevant field work activities to support the web use performed by the staff of the Istat territorial office; (ii) the high diffusion (and use) of Internet services compared to the national level. In particular, Toscana households with an Internet connection (mobile, fixed phone, ADSL and so on) are 62.2% respect to the 60.7% at the national level; persons aged 6 or more able to use Internet services (communication services such as sending and receiving e-mails and phoning; creating and updating social networks profiles; discussing on line about political and social issues) are 58.1% in Toscana and 54.8% in Italy (Istat, 2013). In other words, inhabitants of Toscana are likely to use Internet - by tablet, smart phone and computer . for leisure necessities even if they don't use it to fill in the census forms.

The distribution of the response rates by channel of restitution is the same both at country and

regional level regarding municipal data collection centers (31.7%) and enumerators (12.4%). Instead, web and postal office channels show an opposite behavior at local and national levels: web channel collected the 33.4% of responses in Italy respect to the low 24.5% in Toscana, while Postal Offices gathered only the 22.6% in Italy respect to the 31.4% in Toscana. People living in Toscana prefer hand delivery rather than web one. This low web use suggests a deeper analysis of the reasons beyond this behavior through the application of the same logistic approach at a more detailed territorial level (LAU 2 for 287 municipalities). For this analysis, selected covariates are:

a) Quota of foreigners [Foreigners]. The median value is 7.4%;

b) Quota of persons aged 25-64 years with at least an high school diploma [Diploma]. The median value is 37.1%;

c) Quota of elderly couples with no child [Elderly_Alone]. The median value is 17.2%;

d) Demographic size of municipalities: less/over 20,000 inhabitants [Large Mun].

The first three covariate are dichotomized using the median value as cut-off criterion, in the last case cut-off has been defined at level over and under 20,000 inhabitants. The analysis of Toscana data confirms the expected behaviors: large municipality and presence of a relevant quota of people with a high school diploma have a positive impact on the web propensity while, foreigners and aged population (relevant quota of elderly couples with no child living with them) show an inverse relation with web use.

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

Code	Cut-off	Estimate (β)	P-value	Effect:Exp (β)
Intercept		-1.2917	<0.001	
[Foreigners]	7.4%	-0.00955	<0.001	0.990
[Diploma]	37.1%	0.068	0.0157	1.070
[Elderly_Alone]	17.2%	-0.0685	<0.001	0.934
[Large Mun]	20,000 inhab.	0.1733	<0.001	1.189

Source: our elaboration from SGR data and from dati.istat.it

More specifically, LAU 2 with a number of inhabitants higher than 20 thousands present an odds-ratio around 19% higher than for the less populated municipalities. The odds of areas in which the quota of people from 25 to 64 years with at least a high school diploma is over 37.1% is 7% higher than in the opposite case. Instead, foreigners quota has a weak inverse effect on the response variable: where the quota is over the median level (7.4%) the odds-ratio has a reduction of 1%. Finally, , in municipalities with a high presence of elderly couples, the effect on the odds-ratio is negative (-6.6%).

Despite the lack of data about ICT penetration at LAU-2 level, these 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 shows 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 linking individual channels of answer to other individual characteristics of respondents and extending the analysis to some other variables among them a more detailed 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 process, such as the place of compilation, the assistance received in compiling the form and other related 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.

5. 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, the Internet channel will be the only way to collect data for the Italian "rolling" censuses. One of the major constraints in setting up web surveys is represented by the low level of the 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 beyond the Internet availability and at the light of the unexpected 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. The applied regression model selected several socio-demographic and territorial covariates: i) for population: the municipality size, the age of the head of household, the foreigners' population quota, the quota of population with a high educational level, the households' size, the ICT users' quota and the web mode evaluation level; ii) for enterprises: the foreign entrepreneurs quota, the age of entrepreneurs, the firms size, the quota of enterprises which adopted electronic commerce, the quota of enterprises which adopted innovation in the last years and the ICT users quota. The model explained that for population census, WRR is positively affected by the evaluation and the ICT users' quota and, to a lesser extent, by the quota of graduated people, the family size and the quota of larger municipalities. A negative effect has been played by the age and the foreigner population quota. For enterprises, the most significant variable affecting positively web response rate is the ICT users quota and, to a lesser extent, the enterprise size (proxy of the organization system). Again, the foreign entrepreneurship quota has a negative impact on the web response rates.

Then, it could be useful to provide policies to support the dissemination of knowledge, and the use of computer tools for foreigners and elderly population, especially located in small domains (municipalities /enterprises).

More in detail, actions aimed at increasing web response rates should be planned taking into account various aspects of the survey design: communication strategies, data collection process and monitoring activities.

First the communication strategies should be realized through the mass media communication driving messages calibrated considering the characteristics of the specific population subgroup. For instance to reach elderly people, these messages could be expressed in simple words, using significant pictures recovered from daily life experiences. In case of foreigners people it is extremely important to have a deeper knowledge of their culture in order to coordinate messages and objectives of the survey. Further strategies at local level could be realized according to the citizenship of foreigners. For instance, in the case of Toscana, the municipality of Prato (Belluomini et al., 2012) adopted two main kind of strategies: for the well-established communities (Pakistan, China and Albany) meetings between census staff and the head of the main communities have been organized; for newly settled communities (Romanian and Ukraine) targeted advertising have been performed essentially by mail.

Furthermore, it is not worthy noting that a more comprehensive strategy oriented to promote statistical literacy in terms of awareness of the role that statistics plays in the society could be a core factor to boost response rates.

Referring to the data collection process, specific care has to be devoted to sampling procedures and data collection techniques. Indeed, one of the consequences of pushing web mode is the selection of respondents, and simultaneously the unintended consequence of lowering overall response

rates. To partially solve this bias, the sample design could be empowered using auxiliary information deriving from administrative archives (Istat, 2014) calibrating the sample according to some available dimensions, such as social, demographic, and territorial characteristics (age, educational level, place of living, household size, civil status).

Also data collection techniques should be tested in order to assess the more adequate mode for each target population. In this framework one of the goals of the current Pilot Survey for the forthcoming paperless rolling PHC is to measure the timing of the data collection, the best solicit strategies and the most suitable mix of data collection techniques.

The best combination of strategies has to be established balancing also its effectiveness (timeliness, accuracy, reliability) and its costs. A data collection strategy completely based on web probably is not able to totally realize this goal. Indeed, according to some recent experiences in the field of official statistics (Luiten A., 2014) it is not still possible to completely give up mixed mode approach in favor of the web one.

Finally, in order to reduce the gap between expected and realized results, the whole census strategy has to be maintained through a *just in time* monitoring activity performed by census network operators.

As a concluding remark, this work highlights that the census strategy is a complex system whose success depends on a strong cooperation among all involved actors: Istat staff at central and local level, field operators and stakeholders.

References

Bech, M., Bo Kristensen, M.B. 2009. *Differential response rates in postal and Web-based surveys among older respondents*, Survey Research Methods 3 (1): 1-6.

Belluomini, S., Carmagnini, S., Faustini, L., Porciani, L. 2012. *Data gathering for elusive population. The case of foreigners during the XV Italian Census. A focus on Prato*, Proceedings of the 46th Scientific Meeting of the Italian Statistical Society, Bologna.

Benassi, F., Bruno, M., Giacummo, M., Silipo, M., Vaste, G., Zindato, D. 2014. *Managing census complexity through highly integrated web system*. Rivista di statistica ufficiale 3: 43-59.

Istat. 2007. *La progettazione dei censimenti generali 2010-2011. Analisi comparative di esperienze censuarie estere e valutazione di applicabilità di metodi e tecniche ai censimenti italiani*. Documenti, 9. Roma.

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

Istat. 2011. *Istruzioni per la rilevazione del 15° Censimento generale della Popolazione e delle Abitazioni. Handbook of census*. Roma. Postel.

Istat. 2012. *Istruzione per la rilevazione – Censimento dell'Industria e dei Servizi 2011. Handbook of census*. Catanzaro. Rubettino Print.

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

Istat. 2014. *Linee strategiche del Censimento permanente della Popolazione e delle Abitazioni*. Metodi, letture statistiche. Roma. http://www.istat.it/it/archivio/137474.

Luiten, A. 2014. "Web response in mixed mode designs in official statistics". Presentation at 7th Internet survey methodology workshop. 1-3 December 2014. Bozen

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

Stassi, G., Valentini, A. 2013. *L'Italia del censimento. Struttura demografica e processo di rilevazione*. Istat. Roma.

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