**The e-census challenge: how web affects quality. The Italian experience with a local focus on Tuscany**

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

**\*** ISTAT, Italian National Institute of Statistics

In recent years, also in the field of official statistic, the adoption of web survey techniques has greatly increased. It has been spurred by the evolution of technology, the changes in lifestyles, the wide-spread use of the Internet by people and enterprises, and the development of e-Government. Last but not least a further factor supporting this trend is due to the increasing pressure to find effective methods to reduce costs. The growing role played by the web data collecting techniques is stressed by the decisions of the Italian National Institute of Statistics (ISTAT) which adopted this approach in collecting data within the recent 2011-2012 Censuses. For the first time, the web mode (with the exception of certain telephone and face-to-face reminders) was the exclusive mode used to collect data within the Public Administration network. Instead, mixed mode prevails for population, enterprises and non-profit organizations (even though also in these cases web use was encouraged). This paper aims at illustrating the challenges of the web application in Populations and Housing Census (PHC) and Enterprises Census (EC) at regional level [NUTS-2] devoting particular care at investigating the differences in territorial web response rates. Indeed, Tuscany represents an interesting case of study due to its low web response rate associated with a high ICT penetration rate. Furthermore, the multi-level strategies adopted by Tuscany Istat Regional Office to promote the quality of statistical PHC process were briefly discussed: field work activities (training and supporting actions); elaboration of a set of paradata indicators (dashboard and road map); and ex post evaluation surveys addressed to census operators. Results can be useful for better planning the forthcoming (rolling) censuses and for highlighting the cooperation between NSI and local administrations as key factor to improve data quality and reduce costs.

1. **Introduction**

In recent years the adoption of web survey techniques increased greatly also in the field of official statistic. In Italy, 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 [3]. Specifically, web techniques affected the whole census data collection process of PHC and EC both in terms of questionnaires filling and in terms of monitoring of the process. Concerning the first point, returns of questionnaires followed, almost in all censuses, a mixed mode approach – web or paper– according to the choices of respondents. For this reason, the analysis of the geographical and individual characteristics of web respondents – performed here through a logit model – is a central step to perform due to the increasing role that web techniques are going to play in the near future also in official statistics. Concerning the second point, a web tool called SGR (Survey Management System) has been set up in order to monitor the data collection process and to adopt *in itinere* correction strategies oriented to increase data quality (DQ). Besides the *in itinere* actions, Istat Territorial Office for Tuscany and Umbria established also *ex post* surveys – namely IVALCENS and IVALCIS oriented to increase DQ. In the second part of the paper, these topics will be discussed.

1. **Main features of the Population and Housing and Enterprise Censuses**

All censuses set up by Istat during the period 2010-2013 - namely Agricultural Census (AC), Population and Housing (PHC), Enterprise Census (EC), Non-Profit Census (NPC) - are based on: a list of survey units that supported all the census process; the possibility for respondents to freely select the completion mode; the possibility for enumerators to collect data through face-to-face interviews (third channel of restitution) only during the follow up phase; and the utilisation of a survey management system (SGR), developed in house by Istat, to continuously monitor the process. Census operations could be grouped in four main temporal phases: network operators construction and training, mailing out of questionnaires, voluntary restitution phase and follow up phase. In more detail, field work activities have been partitioned in two main periods: voluntary restitution, and follow up phase. During the “voluntary restitution” phase respondents had the possibility to freely choice their preferred channel of restitution; no special incentives to respondents have been organized except general advertisement; enumerators delivery (i.e. face-to-face interview) was forbidden. Enumerators were employed only in back office activities such as monitoring of the process. During the follow-up phase, enumerators had the commitment to establish personalized strategies to increase coverage contacting the lacking survey units. If no reply signals were registered in SGR for a specific unit and in a sufficiently long period of time, they have to organize a “re-entry strategy”.

1. **Analysis of web response: data and methods**

PHC and EC represent the Censuses where firstly the mixed data collection mode (DCM) was adopted on 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), Italian National Institute of Statistics officially introduced the rolling census methodology [7]. For PHC rolling census will be completely paperless, beginning in 2016 and becoming fully operative in 2020. A first idea could be the investigation of the relationship between raw web response rates (WRR), the most suitable kind of response rate to monitor the quality of a DCM process [5], and the ICT penetration rate [4]. Figure 1, highlighting the weak inverse correlation level (r=-0,38 for Italy) between the WRR and the ICT penetration rates, underlines the necessity to include more covariates to explain the general behaviour of web respondents observed in Italy during the last censuses.



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

The driven idea is that the probability of web response depends on the specific characteristics of the complex social system where survey units, persons, household and enterprises, are settled. As literature shows [1] 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 neither for PHC nor for EC. For this reason, covariates have been limited to: some demographic and economic census data[[1]](#footnote-1); the IVALCENS and IVALCIS ex-post evaluation surveys data, the first one is referred to PHC and the second one to EC [6]; the ICT survey[[2]](#footnote-2) data (ISTAT, 2013). Municipality size, ageing index, quota of foreigner citizens for population and businessman for enterprises; and quota of ICT users are some examples. We investigated this hypothesis using a logistic regression model.

1. **Main results**

For 2011 Italian census the mean web response rate (WRR) by region (NUTS 2) is 33% in case of PHC and 74.7% in case of Enterprises. 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 shows 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-Romagna (78%). The lowest levels are those of Molise (66.2%) and Val d’Aosta (68.1%). More, WRR of PHC and of EC are substantially uncorrelated (R2=0.11). In fact, for PHC areas with the highest web response rates are located in the Southern part of Italy and in Islands; for EC, regions with the highest WRR are located in the North-Eastern part of the peninsula. At the same time, the lowest rates are located in the North for PHC and in the South for EC. To better catch WRR dissimilarities two distinct logistic regression models are used, with respectively six covariates (for the PHC) and three covariates (for the EC), following the standard formula

All explaining variables are bipartite using the median value. The baseline refers to values lower than the median.

|  |  |
| --- | --- |
| 1. **Population and Housing (PHC)**
 | 1. **Enterprises (EC)**
 |
|  |  |

**Figure 2 - Web response rates for Population and Housing and for Enterprises census in Italian regions. (Percentages, Quintiles)**

*4.1 Model and results for Population and Household*  *Census*

The selected covariates[[3]](#footnote-3) (at regional level) are the following:

1. Ageing index [Age[[4]](#footnote-4)]. The median value is 159.4, discriminating between “younger” and “older”regions.
2. Quota of foreigners [Foreigners]. The median value is 8%, discriminating between low and high presence of foreigners.
3. Average family size [Family]. The median values is 2.3, discriminating between regions with “large” and “small” family size.
4. Quota of municipalities (LAU 2) over 20,000 inhabitants [Large Munic]. The median quota is 4.3%, discriminating between “urban” and “rural” regions.
5. ICT users’ quota [ICT]. The median is 54.9%, discriminating between “cabled” and “not cabled” regions.
6. Ex-post evaluation level of web mode data collection [Evaluation]. This covariate comes from an ad hoc survey named IVALCENS (that states for ‘inquiry to evaluate census’) submitted to qualified census operators to catch their opinion about technical, organizational and methodological innovations of the census. Among the various items, they assess the web data collection mode in a scale between 0 (minimum) and 3 (maximum). The median value is 2.4; regions with a higher score have a good appreciation of the web channel.

Table 1 shows the results of the model in which all the coefficients are significant. The ex-post Evaluation variable and ICT variable have the main effects. A good appreciation of the census web tools has a positive effect on the odds ratio (47.02%) boosting web response rate compared to non web response rate. A larger quota of ICT users has a positive effect on the odds (12.33%). The others positive covariates are more linked to socio-demographic characteristics: family size (odds increases of 5.69% in families where average size is more than 2.3 components) and large municipalities (effect on the odds of 1.17% in municipalities with more than 20,000 inhabitants). Instead, two individuals’ characteristics have a negative impact on the odds-ratio: Age (-9.62%) and Foreigners (-11.61%). More in detail in elderly populations, WRR tends to reduce according to the “digital divide gap” which affects elderly people. The effect is similar in areas where the quota of foreigners is very high. As a consequence the quota of web users could increase via targeted actions on elderly populations and on foreigners.

**Table 1 – Results of the logit model for PHC. Italy**

*4.1.2 Model and results for Enterprises Census*

Covariates of the EC model are the followings:

1. Quota of foreigner entrepreneurs [Foreigners]. The median level is of 1.5%.
2. Quota of enterprises over 10 employed [Large Enter]. The median value is 4.4%.
3. ICT users’ quota [ICT], the same variable as for PHC.

Results of the model are shown in Table 2. Note that in the case of enterprises, the most important effect on WRR is determined by size: for larger enterprises odds ratio is 32.55%. This result is partially due to the organized census framework by which enterprises with 10 employees and more were invited to preferably answer via web. Even the use of ICT has a positive impact (+ 3%), but it is less intensive then in case of population. Probably in this case the individual effect of ICT is mitigated by being member of a more or less complex organization. The third covariate, citizenship of entrepreneur, has a negative impact on WRR: odds is reduced by 3.59%. Similarly than in case of population census, targeted actions on foreign entrepreneurs might help to increase the use of web.

**Table 2 – Results of the logit model for EC. Italy**

*4.1.3 General insights from the two models*

It is interesting to note that results for population and enterprises are quite similar for some covariates. This allows making some general considerations concerning the possible extension of WRR analysis. First, focusing on ICT knowledge: geographical areas where web use is more common seem to reach higher WRR. As a consequence increasing the use of technology will probably boost the web survey response rates. According the latest news released by Istat (Istat, 2013), the quota of web users markedly increased in the last year (more than 5%). In the next future this will probably imply a rise in WRR. Furthermore around 86% of families with at least a minor child (less than 18 years) use Internet at home. Vice versa only 13% of elderly (65 years and more) who lives alone has a connection to Internet. Policies aim to incentives the use of Internet especially for old people could be effective for use of web in official surveys and censuses.

A second insight from the results of the model is that larger dimensions (of enterprises or municipalities) seem to be another key factor to increase WRR. Probably, this type of correlation is affected by some typical organizational “biases” such as: the higher the number of units to collect is (or their complexity), the more significant are the actions realized by census operators to promote web completion. Indeed, being easier and faster for census operators to manage web questionnaires than paper ones, it is very convenient (or quite mandatory) for them to sustain the web strategy. Vice-versa, census operators working in areas with simpler or less numerous organizations spending less resources and time to manage questionnaires can do it in an easier way by “hand”. In planning new surveys it will be important to combine an appropriate web strategy both at micro (unit) and macro (municipality) level.

Finally the individual characteristic of citizenship (Italian vs. foreigner) has an impact on the use of web. Foreigners tend to have a flat approach to web. In a society with an increased presence of foreigners targeted actions on non Italian citizens could be the keystone to increase WRR.

5. **Fieldwork activities and paradata tools to increase DQ**

Activities adopted during PHC operations by Istat in Tuscany were essential to manage DQ. According to Wang [8] three approaches have been used: a) intuitive; b) theoretical; c) empirical. The *intuitive* approach is taken when the selection of data quality attributes is based on researcher experience or intuitive understanding about what attributes are important. The *theoretical* approach focuses on how data become deficient during the data collection process, through a set of a-priori indicators. The *empirical* approach captures the voice of users, weather data fit for use in their tasks.

All the three approaches where pursued by Istat in Tuscany for the Population census. As intuitive approach field activities were planned to train, support and monitor the whole network stressing the main dimensions of data quality (accuracy, timeliness, universality/stability, clarity, consistency): total of 69 training modules took place; more than 5 thousand operators have been trained; 11 thousand mail and telephone contacts were managed by the local Istat staff.

Concerning the theoretical approach indicators are calculated both as real-time indicators (dashboard) – useful to set up ongoing interventions on municipalities - and ex-post indicators (road-map), useful for future modifications of the implemented strategies. Dashboards indicators where mainly drawn from SGR data. A particularly interesting one was the response rate (QUES.LAC), calculated as the percentage ratio between number of questionnaires filled in and number of questionnaires in the municipality population registry lists (LAC). In all cases an operative a minimum threshold level has been identified. Interventions strategies continue till the threshold has been overcome. The judicious use of dashboard indicators acts both on timeliness and accuracy of the process. Timeliness is guaranteed via the immediate identification of the latecomer institutions and the subsequent actions to remove the reasons of delay. Accuracy depends on a properly evaluation of DQ to avoid systematic errors concerning both territorial and thematic issues.

As mentioned the IVALCENS ex-post survey and the corresponding ex-post survey for EC called IVALCIS have been established. Results about these last two enquiries will be presented in future works.

**6. Conclusions**

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. In fact web techniques, after the initial phase of software investment, are convenient in terms of costs, and are featured of better timeliness and lightening statistical burden. Also Census surveys are rapidly switching to this data collection mode, as reported in this paper for the 2010-2012 Italian wave. In the next future internet will be the only way to collect data for Italian “rolling” censuses. One of the major obstacles in setting up web surveys is represented by the lacking Internet penetration: not all enterprises and particularly not all households are internet connected. As a consequence the response rate often results unsatisfactory, lowering the quality of the process [2]. However, the recent evidences of the Italian population and economic censuses give a unique opportunity to take advantage from the very complex process, which allowed several concurrent data collection modes and territorial breakdowns, for investigating further factors that could affect the web response rate. The paper shows that the correlation between ICT penetration and web response rate is rather weak, and negative, suggesting to investigate other variables as underlying motivations of the effectiveness of web surveys. Furthermore the two censuses web response rates seem uncorrelated. In the case of Population and Housing Census regions with higher web response rates are concentrated in the South of Italy while regarding enterprises in the North of the country. This occurrence suggested to analyse them separately analyse. A logit model is thus applied selecting the following currently available 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. For the first group - population variables - WRR is positively affected by evaluation and ICT quotas and to a lesser extent by family size and by quota of bigger municipalities. Negative contributions are given by age and foreigner population quota. For enterprises, the most important variable affecting positively web response rate is the enterprise size, and to a lesser extent the ICT quota. Again the foreign entrepreneurship quota has a negative impact on the web response rate.

Results of this paper suggest that advisable policy actions aimed at increasing the web response rates have to focus on supporting the internet education of elderly people and foreigners, and on mainly concentrating on small dimension domains (municipalities /enterprises). Further developments are limited 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 features could play an important role to ensure an high participation rate.

**References**

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

[2] Bethlehem J. Biffignandi, S. (2011). *Handbook Web Surveys*: New York: Wiley & S Sons.

[3] 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 n.9/2007). Roma: Istat Servizio Editoria - Centro Stampa.

[4] Istat (2013). *Cittadini e nuove tecnologie.* Statistiche report (19 dicembre 2013), <http://www.istat.it/it/archivio/108009>.

[5] 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, Brussells, November.

[6] Stassi, G., Valentini, A. (2013). *L’Italia del censimento. Struttura demografica e processo di rilevazione.* Toscana, Roma: Istat.

[7] 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>

[8] Wang R.Y., Strong D.M. (1996), Beyond accuracy: What data quality means to data consumers, Journal of Management Information Systems
Vol.12, No.4 Spring 1996

1. First census data are available on: <http://dati.istat.it/> [↑](#footnote-ref-1)
2. Data are available on: <http://www.istat.it/it/archivio/48388> [↑](#footnote-ref-2)
3. Variable labels re displayed inside square brakets [↑](#footnote-ref-3)
4. The variable name used in the model is between square brackets. [↑](#footnote-ref-4)