Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

Towards the 2021 UK Census imputation strategy: Response mode as a matching variable in a donor-based approach?

Steven Rogers, Lawrence Dyer, Brian Foley[1]

Abstract

Since July 2014, the Office for National Statistics has committed to a predominantly online 2021 UK Census. Item-level imputation will play an important role in adjusting the 2021 Census database. Research indicates that the internet may yield cleaner data than paper based capture and attract people with particular characteristics. Here, we provide preliminary results from research directed at understanding how we might manage these features in a 2021 UK Census imputation strategy. Our findings suggest that if using a donor-based imputation method, it may need to consider including response mode as a matching variable in the underlying imputation model.

Key words: census, imputation, donor, internet capture, CANCEIS

1. Introduction

Adjusting Census data for item-level inconsistencies and non-response (from here on referred to as imputation), is an important part of Census data processing. The 2011 UK Census imputation strategy was a donor-based approach implemented in CANCEIS (Aldrich and Rogers, 2012; Wardman and Rogers, 2012; Wardman, Aldrich, and Rogers, 2014). The Census data were processed in 101 Delivery Groups (DGs) containing on average 241,000 households with 530,000 people, implemented through a series of modules, namely Demographics, Culture, Health, and Labour Market.

In CANCEIS, the variables within a module, along with other auxiliary information, served as matching variables representing the underlying imputation model. A Nearest Neighbour algorithm constructed a pool of potential donors statistically similar to the record being imputed based on available data in the matching variable set. From that pool, the donor that supplied an imputed value was chosen randomly from a set of near minimum change imputation actions (NMCIAs) (Bankier, Lachance, and Poirier 1999; CANCEIS, 2009). This strategy provides point and variance estimates of the distribution of the missing data conditioned by the underlying imputation model (from here on referred to as imputation estimates).

Since completion of the 2011 Census, following a detailed research programme, the National Statistician recommended a predominantly online UK Census in 2021, supplemented by the use of administrative data. This was approved by the UK Government in July 2014. The ONS is now fully committed to the development of a different kind of Census in 2021. Exploring the impact that administration data and internet capture may have on the accuracy of imputation estimates are two important facets of that commitment.

Previous research suggests that internet capture may have two promising features. It appears to yield data that has less non-response and more consistency than that captured through a paper questionnaire (e.g., Côté and Laroche, 2009). It also appears to attract respondents with characteristics often associated with those that are hard to count, such as working age males; higher levels of education; a professional occupation; a country of birth outside of the UK; a second home abroad (Ghee, 2014). In addition to having less data to impute overall, an increase in observed hard to count characteristics can only serve to reduce uncertainty in Census population estimates usually associated with imputed data.

Despite these promising aspects, notable distributional differences in the characteristics of internet and paper based responders also suggest that when imputing mixed mode data using donor-based methodology it may be necessary to include mode as a matching variable in the underlying imputation model. Unspecified, there is risk of introducing a sampling bias into imputation estimates through randomly drawing a donor from a donor pool that is potentially

[1] Steven Rogers, Office for National Statistics, Segensworth Road, Fareham, England, PO15 5RR, steven.rogers@ons.gsi.gov.uk; Lawrence Dyer, lawrence.dyer@ons.gsi.gov.uk; Brian Foley, brian.foley@ons.gsi.gov.uk

heterogeneous. As sampling bias tends to be more likely with subsample size inequality, the higher rate of clean and consistent records in internet responses, coupled with a higher rate of records that need imputing in paper based responses, may serve to increase that risk.

The overarching aim of the current research is to establish how to manage mixed mode data and any associated risks to the accuracy of imputation estimates in the 2021 UK Census imputation strategy. Here we present early findings of some preliminary work based on 2011 UK Census data directed at evaluating the differences between imputation estimates from a donor based imputation strategy that includes mode as a discrete matching variable to one that does not. The analyses focuses on DG513 which consists of four local authorities; Westminster, Kensington & Chelsea, Camden, and City of London. Data capture in this inner-city region is notoriously difficult, placing it firmly in the highest category of the ONS Hard to Count Index.

The analyses explore the data for households containing 1 to 6 people. It also focuses on variables from the 2011 UK Census imputed in the Demographics module: Age; Sex; Marital status; Activity last week; Second address indicator (UK, Other, or None); and Country of birth indicator (UK & Rep. Ireland, Other, or None). When exploring the differences between the characteristics of internet and paper respondents a few additional variables implicated in other research such as Highest level of qualification, Ethnicity, and Industry, are also included.

2. Results & Discussion

2.1 Comparison of internet and paper responses

Table 2.1-1 provides highlights of the distributional differences in the characteristics of individuals in DG513 responding by the internet compared to those responding by paper questionnaire. All categories within variable are shown where there was an absolute difference of at least 3% between response modes. Positive differences in the table indicate higher proportions in the internet data for that category.

Table 2.1-1
Highlights of distributional differences in the characteristics of internet and paper respondents

		Paper		Internet		Difference
		(n)	(%)	(n)	(%)	(%)
Age Group	25 to 29	37.3k	9.87	18.2k	13.90	+4.03
	30 to 34	35.6k	9.42	18.5k	14.09	+4.67
	Over 79	14.7k	3.89	1.1k	0.80	-3.09
Sex	Male	178.3k	46.99	67.3k	51.11	+4.12
	Female	201.2k	53.01	64.4k	48.89	-4.12
Marital status	Single	207.8k	55.62	81.1k	61.58	+5.96
Activity last week	Retired	40.3k	11.13	5.4k	4.08	-7.05
	Working	179.6k	49.65	40.4k	53.62	+3.97
	Student	22.5k	6.23	12.5k	9.55	+3.32
Second address	Other (not in UK)	28.5k	7.80	17.0k	12.94	+5.14
Country of birth	UK & Rep.Ireland	209.7k	56.13	57.5k	43.77	-12.36
	Other	161.1k	43.13	104k	54.34	+11.21
Industry	Finance	27.5k	8.13	15.7k	12.24	+4.11
Ethnicity	English	158.0k	44.86	40.0k	32.23	-12.63
	Other white[1]	78.8k	22.37	37.8k	30.50	+8.13
Highest qualification	None	40.1k	11.02	8.0k	6.06	-4.96
	Level 4+[2]	148k	40.66	61.7k	46.94	+6.28

1 Not English, Irish, Gypsy or Irish Traveller
2 Degree (BA, BSc), Higher Degree (MA, PhD, PGCE), NVQ Level 4-5, HNC, HND, RSA Higher Diploma, BTEC Higher level, Foundation degree (NI), Professional Qualifications (Teaching, Nursing, Accountancy)

The differences observed between internet and paper responses for DG513 are consistent with that of other research. The internet data contained higher proportions of people who were between 25 and 34 years old; male; single; working or student; and with a high level of education. There were also higher proportions of people who reported

being white but not English or Irish; born outside of the UK; and who had a second address outside of the UK. These differences represent the conditions that may lead to a risk of bias in imputation estimates.

Table 2.1-2 shows a detailed breakdown of how the households in DG513 were distributed relative to the aim of imputing for missing and inconsistent data.

Table 2.1-2
Households in DG513: Counts and key proportional comparisons

Committed to imputation (n)				**Row Comparisons (%)**	
	paper	internet	totals	paper	internet
Clean donors	135,717	59,192	194,909	69.63	30.37
Missing & inconsistent	52,807	1,922	54,729	96.49	3.51
totals	188,524	61,114	249,638	75.52	24.48

Proportion clean (%)		***Other (n)***	
paper	internet	*Clean non-donor*	*paper: 39*
71.99	96.86		*internet: 16*
		Mixed mode	*1,195*

It is worth noting first that data for 1,195 households contained responses provided through both modes. For the purpose of the current study, these records were left out of any further analyses. Of the records being committed to imputation 24.48% were internet responses. Consistent with findings from other research, 96.86% of all internet responses were clean and consistent compared to only 71.99% of all paper responses. Notably, 96.49% of all records needing imputation were paper responses whereas 30.37% of all potential donors were internet responses. Proportional inequalities such as this represent the conditions that may lead to an increased risk of bias in imputation estimates.

2.2 Differences in imputation estimates

A comparison of imputation estimates from a donor based imputation strategy that includes mode as a discrete matching variable compared to one that does not was obtained through a simple experiment. The Demographics module for DG513 was imputed in CANCEIS under two conditions. In the controlled condition, donors could only be selected who responded through the same mode as the record being imputed. In the free condition, any donor could be selected, regardless of mode.

Differences in imputation estimates can be measured by comparing the distributions of the imputed data obtained from each of the two imputation strategies. However, to ensure that any potential differences were not simply attributable to imputation variance the data were imputed 10 times for each condition and comparisons were based on jackknife point and variance estimates of the imputed distributions over those 10 runs (see appendix 1).

In general, there were systematic differences in the imputation estimates from each of the strategies that could not be explained away by imputation variance. Table 2.2-1 shows the sum of the absolute differences in distributional estimates for each of the variables in the Demographic module. It also shows the range of 99% confidence intervals associated with those estimates.

The largest overall difference in imputation estimates was for Age Group, followed by Sex, Country of birth, Activity last week, and Second address. The smallest overall difference was for Marital status.

Table 2.2-2 provides detailed highlights of the distributional differences between the data imputed through the controlled and free conditions. All discrete categories within variable are shown where there was an absolute difference of at least 0.5%. Positive differences in the table indicate that where donors were selected freely, regardless of mode, there were higher proportional estimates in the imputed data for that particular category. Negative differences indicate lower proportional estimates.

Table 2.2-1
Absolute differences in imputed distributions and range of 99% confidence intervals

	|Difference| (%)	99% confidence intervals (to 2 decimal places)					
		Controlled condition			Free condition		
		Min	Mean	Max	Min	Mean	Max
Age Group[1]	5.22	0.01	0.02	0.03	0.01	0.02	0.03
Sex	3.31	0.04	0.04	0.04	0.08	0.08	0.08
Country of birth	2.54	0.01	0.02	0.03	0.01	0.02	0.03
Activity last week	1.48	<0.01	0.01	0.02	<0.01	0.01	0.02
Second address	1.04	0.01	0.01	0.01	0.01	0.01	0.02
Marital status	0.88	<0.01	0.01	0.02	<0.01	0.01	0.02

1 Age was imputed for single year and clustered: <1; followed by 5 year equal intervals; ending in >79

Table 2.2-2
Highlights of distributional differences in imputation estimates

		Controlled condition (%)	Free condition (%)	Difference (%)
Age Group	15 to 19	9.890	9.046	-0.843
	20 to 34	29.292	30.802	+1.510
	35 to 44	13.440	14.032	+0.593
Sex	Female	53.252	51.597	-1.655
	Male	46.748	48.403	+1.655
Country of birth	UK & Rep.Ireland	51.062	49.956	-1.106
	NCR[1]	19.916	21.182	+1.270
Activity last week	Working	36.205	36.903	+0.698
Second address	Outside UK	4.704	5.223	+0.519

1 No Code Required: Student with term time address outside of UK

Compared to the imputation strategy that included mode in the imputation model, not controlling for mode led to higher imputation estimates for 20 to 44 year olds; males; and people who were working. The increase in estimates for the 20 to 44 age groups seemed largely to be at the expense of lower estimates for 15 to 19 year olds. Not controlling for mode also led to higher estimates for students living outside of the UK during term time and people with a second address outside of the UK. The increase in estimates for students living outside of the UK seemed mainly to be at the expense of lower estimates for people born in the UK or Republic of Ireland.

Tellingly, the results show that the characteristics of people represented with higher proportions in the internet data (Table 2.1-1) were similar, if not the same, as those characteristics estimated more frequently when imputing without mode serving as a constraint on donor selection (Table 2.2-2). This seems to confirm that by not conditioning the imputation on mode the risk of introducing a sampling bias into imputation estimates through randomly drawing a donor from a heterogeneous donor pool is likely to be realised. Overall, the results of the current study suggest that when imputing mixed mode data using donor-based methodology, particularly where respondents are free to choose their mode of response, it may indeed be important to consider including mode as a matching variable in the underlying imputation model.

References

Aldrich, S., Wardman, L., and Rogers, S. (2012), The practical implementation of the 2011 UK Census imputation methodology. Available online: [06/01/2015] http://www.unece.org/stats/documents/2012.09.sde.html#/

Bankier, M., Lachance, M., & Poirer, P. (1999), A generic implementation of the new imputation methodology. Accessed online: [07/12/2011] http://www.ssc.ca/survey/documents/SSC2000_M_Bankier.pdf

CANCEIS (2009), Users Guide V4.5. CANCEIS Development Team. Social Survey Methods Division, Statistics Canada

Côté, A-M., and Laroche, D. (2009), The internet: A new collection method for the Census. Available online: [06/01/2015] http://www.statcan.gc.ca/pub/11-522-x/2008000/article/10986-eng.pdf

Ghee, K. (2014), Internet versus paper mode effects in the 2011 Census of England and Wales: Analysis of Census Quality Survey agreement rates. Available online: [06/01/2015] http://www.unece.org/stats/documents/2014.09.census1.html#/http://www.unece.org/stats/documents/2014.09.census1.html#/

Wardman, L., Aldrich, S., and Rogers, S. (2012), Item imputation of Census data in an automated production environment; advantages, disadvantages and diagnostics. Available online: [06/01/2015] http://www.unece.org/stats/documents/2012.09.sde.html#/

Wardman, L., Aldrich, S., and Rogers, S. (2014), 2011 Census item edit and imputation process. Available online: [06/01/2015] http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/quality-measures/response-and-imputation-rates/item-edit-and-imputation-process.pdf

Appendix 1

$\bar{\theta}_{Jack} = \frac{1}{n} \sum_{i=1}^{n} (\bar{\theta}_i)$ $\qquad$ $Var(\theta) = \frac{n-1}{n} \sum_{i=1}^{n} \left(\bar{\theta}_i - \bar{\theta}_{Jack}\right)^2$