WP.44
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (vii): Editing and imputation of census data

Editing and Imputation of the 2011 Abu Dhabi Census

Prepared by Glenn Hui and Hanan Ibrahim AlDarmaki
Statistics Centre - Abu Dhabi

Abstract: For the 2011 Abu Dhabi Census of population, Statistics Centre - Abu Dhabi (SCAD, the emirate's official statistics office) used a variety of imputation methods–primarily donor imputation–to ensure a complete and consistent data set. This paper begins with a brief overview of how SCAD conducted the 2011 Census, some of the challenges unique to the region's demography, and how editing and imputation were developed to account for these factors. The discussion moves on to differences between SCAD's methodology and that of other modern statistics offices, and follows with a thorough analysis of results and performance. SCAD's imputation methodology is a significant change from traditional row-by-row manual imputation, a contrast this paper examines.

I. Introduction

1. In October 2011, Statistics Centre - Abu Dhabi (SCAD) conducted its first census of population.[1] SCAD, the official statistics office of the emirate of Abu Dhabi, was created in 2008 to modernise the collection, processing, and dissemination of statistics in the emirate; as a result, the census was developed and run almost entirely from scratch, with minimal carryover from previous censuses.

2. As part of the modernisation effort, editing and imputation–the processes of identifying and correcting erroneous responses, respectively–were implemented primarily via the Canadian Census Edit and Imputation System (Canceis). Canceis was developed by Statistics Canada to impute its Censuses, and has since been adopted by several other countries, such as the UK, Italy, and Brazil.

3. This paper intends to demonstrate how modern, statistically sound methodology can be adapted for non-Western cultures without much difficulty. We hope that this will be of benefit to developing statistical offices looking to modernise their editing and imputation, and of interest to others as a case study.

A. Census Overview

4. The bulk of the 2011 Abu Dhabi Census was collected via computer-assisted personal interview. Close to 3,000 enumerators went out door-to-door over a four week period, equipped with Apple iPad tablet computers. Other collection methods, primarily paper form with manual data entry, were used for certain subsets of population, such as worker camps and hotels.

[1] The first official census of Abu Dhabi was conducted in 1975, as part of the census of the UAE. Federal censuses were conducted in 1980, 1985, 1995, 2001 and 2005.

5. As of writing, coverage studies are still on-going and public release of Census results has yet to occur, so total counts are not provided. Where appropriate we mention percentages of the enumerated population.

II. Editing and Imputation Methodology

6. Editing and imputation were performed via three different methodologies, which we refer to as donor imputation, deterministic imputation, and manual imputation. Each methodology was used for specific reasons, and is described below.

A. Donor Imputation and Canceis

7. Terminology in the literature varies somewhat, so in our multi-lingual environment we simply refer to Canceis' methodology as donor imputation. Statistics Canada describes it as "minimum change nearest neighbour imputation".[2] The key characteristics are:

- Donor imputation in general substitutes a value in the failed record with a value from another record. (Sometimes also called hot-deck imputation, a reference to donors being selected from the same dataset being imputed, as opposed to a different dataset.)
- Minimum change is the philosophy that imputation should retain as much of the collected data as possible when correcting a failed record.
- Nearest neighbour indicates that a donor record is selected based on similarity to the failed record.

8. The editing stage of Canceis identifies households needing imputation (failed records). To do so, users define validation sets to indicate which values are valid for a particular variable. Invalid responses–generally missing values and out-of-range responses, such as an age above 115 or a value not in the code set–are immediately flagged for imputation.

9. To flag inconsistent data, consistency rules are created via Decision Logic Tables (DLTs), essentially a custom-designed programming syntax. One advantage Canceis has over other systems is that DLTs can easily identify inconsistencies between persons of a household, not just within an individual record. For example, a between-person edit might check that mother-child pairings have an appropriate age gap–say, between 15 and 50 years–while a within-person edit might check that a college graduate is old enough to have graduated.

10. Households that pass all edits are called donors, and donor households are thus selected from this pool of passed households.

11. In the imputation stage, Canceis selects donors that are closest to the failed household using custom distance measures and weights defined for each variable. There are two goals in imputing the failed household: ensuring the household passes all edit rules after imputation, while retaining the smallest distance between the original and imputed data. Canceis offers both deterministic and donor imputation, but we used only the donor (hot deck) portion.

12. Over 100 DLTs were developed by SCAD staff, using both methodological expertise and local knowledge to produce rules that would be familiar to other Canceis users but also foreign in many ways. Three modules were used, as described below:

- Demography: Gender, Age, Relationship to Head of Household, Marital Status, Nationality
- Education / Employment: Enrolment Status, Current Stage, Current Grade (if Stage in grades 1-12), Attainment; Activity Status, Type, Occupation, Industry, Sector

[2] Statistics Canada, CANCEIS User's Guide.

- Migration: Absent Flag, Current Region if Absent, Absence Duration, 2010 Region of Residence

13. The Demography module was applied to all data streams, although Relation was out of scope for non-family housing units. A revised version of the Education / Employment module was applied to worker camps. Most non-family data streams were not run through the Migration module, as most of its questions were out of scope.

B. Deterministic Imputation

14. Deterministic imputation simply applies logical rules to modify data. This method was implemented primarily before donor imputation to eliminate out of scope responses. Because deterministic imputation carries risks of oversimplifying data or wiping out rare but valid cases, it was used very sparingly and only after verifying its validity.

15. Example edits include erasing marital status or employment data for children below 15 years old. Base assumptions–that some fields, particularly age, were more reliable than others–were verified both by analysing samples of micro-data as well as aggregate data. Additional rules were applied for nationality: single missing nationalities in households were corrected deterministically based on the nationality of other members in the household.

16. Deterministic imputation for out-of-scope responses affected around 8,000 records in family households prior to donor imputation. Applying deterministic imputation before donor imputation improved the performance of Canceis by increasing the number of possible donors, as well as ensuring that some difficult-to-impute variables were set correctly.

C. Manual Imputation

17. Households deemed particularly difficult to impute automatically–mostly very large households–were collected and checked manually for errors. SCAD staff looked over this data using names and other fields to correct Gender and Relationship errors. Over 3,000 large households (around 40,000 person records) were checked manually and corrections applied where obvious errors were found. Actual changes were made sparingly and only where there was strong evidence of an error before applying corrections.

18. Manual imputation was also applied in a few complicated cases of conflicting data that did not have obvious automated solutions, but the total number of these households reviewed was only a few hundred.

III. Past Data Editing in the UAE

19. The 2005 UAE Federal Census, conducted in Abu Dhabi by the Department of Economic Development, used a combination of automated editing and deterministic and manual imputation. The first phase applied validation edits via SQL queries, and made a small number of deterministic imputations. The majority of failed records were handed to a team of about 15 data entry staff who manually corrected them. In the next phase, outlier detection rules were applied via SQL and all outliers were manually reviewed. The manual editing process took over four months on top of development time. For comparison, all 2011 imputation processing took two methodologists less than five months, with a little aid from two local analysts.

20. Many of the rules used in 2005 were preserved and implemented as Canceis DLTs in 2011. It is worth noting that, even using similar sets of rules, donor imputation can have very different results.

21. For example, one of the 2005 deterministic rules involving relationship and marital status was stated in this manner: “if relationship is Spouse, then marital status must be Married”. This rule always changed marital status and not relationship, which may have introduced a bias in the 2005 dataset. In

contrast, in Canceis a similar edit rule is defined for error detection, and the Relationship variable is given greater weight–but the imputation procedure could still change Relationship depending on the entire household. This resulted in more accurate micro-data, and likely better-preserved aggregate distributions.

IV. Cultural differences and imputation challenges

22. Several cultural differences between the UAE and Canada added a level of complexity to the implementation of Canceis. These challenges were overcome by thorough analysis of the data and the aid of local demography experts.

23. The main challenge was a high proportion of very large households: even counting only private families, over 10% of the population lives in households of over 12 people. Due to the high error rates proportional to family size, households with 10 or more people were edited using a separate set of DLTs with less strict rules. Technical challenges were faced while attempting donor imputation on households of 17 or more persons, so they were all imputed in one set as individual records. (For comparison, in the 2006 Canadian Census, households of 9 or more persons were imputed as individuals.)

24. Large households are inherently more error-prone, due to having more opportunities for error as well as human aspects such as impatience with interviews lasting over an hour. They also have complicated family structures that are difficult to capture via the 11 possible Relationship responses.

25. Further complicating family structure are multiple-wife households, which are fairly rare but common enough to affect editing. The most significant challenge was modifying all rules involving spouses or possible mother/child relationships. Additionally, because multiple-wife households are generally valid, some common errors were hard to identify–particularly multiple couples living together and listing each wife as Spouse, or the Household Head's children-in-law being listed as Spouses. For households with four or more Spouses, we simply applied manual imputation.

26. In addition, Abu Dhabi has a large expatriate community that outnumbers Emiratis. Even excluding worker camp residents, about a third of expatriates live in shared living arrangements–that is, multiple families (sometimes over a dozen) living in one housing unit. The relationship variable was out of scope for these households. Highly varying nationalities also required complicated distance matrices.

27. Other challenges included unrelated servants living with families and systematic enumerator errors.

V. Results and Performance

28. This section focuses on the results and performance of Canceis, as the performance of the other two imputation methodologies is less measurable. By "performance" we refer to both the accuracy of editing (detecting errors where they exist and not falsely flagging good data) as well as the accuracy of imputation (imputing to the "true" value).

29. A total of 13.1% of person records were imputed by Canceis, the majority of which failed due to missing values. By module, 5.4% of records were imputed in Demography, 7.0% in Education / Employment, and 3.1% in Migration. (A record can be imputed in multiple modules, of course.) These numbers are roughly in line with other agencies' experiences.

30. Several statistical measures were used to evaluate imputation performance, most of them adopted from Chambers' 2003 paper on the evaluation of editing and imputation, part of the Euredit project on data editing[3]. All measures are defined in detail in Appendix A. Where appropriate, the results in this section are compared to the Euredit results which used Canceis on a subset of the 1991 UK Census.

A. Test Datasets

[3] Chambers, R., "Evaluation Criteria for edit and imputation methods".

31. Evaluation datasets were generated from a subset of 4-person households that passed all edits, around 30,000 households. This subset, denoted Y*, is defined to be complete and correct. Errors were then introduced to Y* by randomly replacing values with new modified data: either missing values, interchange errors, or both. (Interchange errors are introduced by replacing the original value with a random value in the valid range.) Three evaluation dataset versions were generated; the table below shows the notation used for each dataset and whether it contains missing values or errors:

Missing? \ Errors?	No	Yes
No	Y*	Y1
Yes	Y2	Y3

Table 1: Evaluation Datasets

32. The error rates were controlled for each variable by randomly selecting a percentage of records and replacing one of the variables with a missing value or error[4]. The variables we used for evaluation are the four demographic variables in Table 2.

Variable	Valid Range
Age	0-115
Gender	1-2
Relationship to household head	1-11
Marital status	1-6

Table 2: Demographic Variables

33. The observed error rates for each evaluation dataset are shown in Table 3 below. The missing rate for marital status is lower than the rest since children under 15 were out of scope, and the rate is calculated over all records.

	Missing values				Interchange errors			
dataset	Age	Gender	Relation	Mstatus	Age	Gender	relation	Mstatus
Y1	0%	0%	0%	0%	4%	4%	4%	4%
Y2	4%	4%	4%	2.5%	0%	0%	0%	0%
Y2b	2%	2%	2%	1.25%	0%	0%	0%	0%
Y3	2%	2%	2%	1.25%	2%	2%	2%	2%

Table 3: Person-Level Error Rates

34. Using this setup, the household-level error rates as reported by Canceis can reach 50%, which is unrealistically high, but experiments showed that performance is only slightly dependent on error rates.

B. Editing Performance

35. Error detection performance is measured here mainly on Y1 which contains only errors, with around 16% error rate at the person level. Missing values are not counted as errors since they are all correctly identified by Canceis.

36. There are two possible types of classification errors: Type 1 errors are false positives (correct values classified as errors), and Type 2 errors are false negatives (incorrect values accepted as valid).

37. Alpha α measures the proportion of Type 2 errors, Beta β measures the proportion of Type 1 errors, and Delta δ measures the proportion of Type 1 and Type 2 errors in the whole dataset. Table 4 shows the performance of the editing procedure for all variables. Table 4b shows the corresponding results from the Euredit project for comparison.

[4] Wagstaff, H., "Descriptions of datasets and their perturbations".

	Age	Gender	Relation	Mstatus
Alpha	0.643	0.666079	0.364711	0.259398
Beta	0.003383	0.000517	0.017313	0.007762
Delta	0.028482	0.027262	0.031197	0.017771

Table 4: Editing Performance by Variable- Y1

	Age	Gender	Relation	Mstatus
Alpha	0.593281	0.078518	0.435005	0.243563
Beta	0.004183	0.000275	0.000821	0.000255
Delta	0.045277	0.005354	0.028296	0.011676

Table 4b: Euredit Results[5]

38. It is immediately apparent that, while most of our results are comparable to Euredit's, Alpha for Gender is an order of magnitude higher. We quickly realized that this is because there is only one edit for Gender in the Demography module, for married couples. The vast majority of Genders are ignored.

Dataset	Alpha	Beta	Delta
Y1	0.522778	0.01901	0.078299
Y3	0.501027	0.011713	0.039376

Table 5: Overall Editing performance

39. The editing efficiency measures in Table 5 above are calculated across all demographic variables. The error rates are lower in Y3, and the performance is slightly better than the performance on Y1.

C. Donor Imputation Performance

40. For the purpose of imputation evaluation, the experiments are conducted on Y2 (containing missing values only) and Y3 (containing both missing values and errors). The measures described below are used to evaluate the imputation procedure, so they are calculated over the set of n imputed values only. Non-imputed values, including the undetected errors, are not included in these calculations.

41. For categorical variables, two measures are used to measure the predictive accuracy of the imputation procedure. Distance measure D is simply the proportion of imputed values that differ from their original values, while D_{gen} is the distance between imputed and real values.

42. Table 6 shows the results on evaluation datasets. The results are averaged over 5 experiments for each dataset.

	D			D_{gen}		
Experiment	Gender	Relation	Mstatus	Gender	Relation	Mstatus
Y2	0.249118	0.040729	0.075526	0.249118	0.02811	0.052591
Y2b	0.246315	0.041776	0.073222	0.246315	0.028982	0.051059
Y3	0.192884	0.262819	0.146578	0.192884	0.183615	0.104665

Table 6: D Statistic for Categorical Variables

43. Y2 and Y2b results demonstrate that the performance is not highly affected by the error rate. Results for Gender are somewhat worse than other variables, again explained by the fact that there are few edits checking against Gender. We did note that the distribution of Gender was preserved despite some individual records not always being imputed correctly.

[5] Charlton, "Evaluating New Methods for Data Editing and Imputation - Results from the Euredit Project (UK)"

44. In Y3, D is significantly reduced for Gender due to the high accuracy of error correction when an edit fails, since there are only two possible values in the valid range. Therefore, all detected errors in Gender have an imputed value equal to the original value. On the other hand, all true values classified as errors will be different from the original value, so due to the high Beta values for Relation and Mstatus, D is increased in the presence of errors. Undetected errors are not included in these measures, so high alpha values do not affect the results.

45. For the continuous variable Age, the following measures are used:

- **Predictive Accuracy:** Using a simple linear regression model, R^2 is calculated to assess the correlation between real and imputed values. Values for R^2 range from 0 to 1, with a good fit being close to 1. D_{L2} is a distance measure which calculates the average distance between real and imputed values, so a smaller distance indicates better preservation of true values.

- **Estimation Accuracy:** These measures assess how well the imputation method reproduces the lower-order moments for the distribution of true values. m_1 measures the preservation of the mean, and m_2 measures the preservation of the variance of the empirical distribution.

- **Distributional Accuracy:** The Kolmogorov-Smirnov statistic, KS, confirms the overall preservation of the empirical distribution of true values.

Details for calculating these measures are described in Appendix A.

Dataset	R^2	D_{L2}	Relative D_{L2}	m_1	m_2	KS
Y2	0.84958	6.891341	0.906017	0.315374	22.68677	0.012343
Y2b	0.86314	6.693901	0.94789	0.295742	27.99995	0.016372
Y3	0.81282	7.856174	1.344191	0.53128	34.41243	0.020611

Table 7: Evaluation Measures for Age

46. R^2 is larger than 0.8 in all datasets, indicating a reliable correlation between imputed and real Ages. Low values for the distance measures D_{L2} indicate the preservation of true values after imputation. The preservation of the mean and variance of the distribution of the imputed values is indicated by the low values for the first and second moments m_1 and m_2. The Kolmogorov-Smirnov statistic for all datasets is less than 0.1, suggesting that imputation did not distort the distribution of Age. In spite of the high error rate and presence of errors in Y3, Canceis performed reasonably well in all measures.

47. On the whole, our editing and imputation achieved respectable results for all the demographic variables analysed above. Brief experiments in 6-person households suggested that performance did not noticeably decrease with the larger household size.

VI. Conclusions

48. Overall, editing and imputation in the 2011 Abu Dhabi Census was a success. Performance measures, as well as both micro- and macro-analysis, indicate that imputation was statistically sound and performed to a high standard.

49. We feel that our imputation serves as a good example of adapting best statistical practices in a culture unlike that which they were designed for. Traditional manual imputation will always have its place; certainly, some degree of micro-data analysis is always necessary as part of evaluation. Nonetheless, the methodology SCAD implemented for 2011 has the desirable properties of improving the accuracy of the data, producing measurable changes, and being more efficient and easily adapted for future use.

Appendix A: Evaluation Measures

To measure editing accuracy we used Alpha, Beta, and Delta. Let n be the total number of person records, n_e the number of errors, n_{ec} the number of errors identified as correct, n_c the number of correct records, n_{ce} the number of correct records identified as errors. Then

$$\alpha = n_{ec}/n_e \qquad \beta = n_{ce}/n_c \qquad \delta = (n_{ec} + n_{ce})/n$$

Because imputation performance was considered more crucial, a number of different measures were used in its evaluation. All measures in this section are defined over the set of n imputed values, and not the whole dataset. For categorical variables, we start with indicator variable $I(\hat{Y}_i \neq Y_i^*)$ that is equal to 1 if the imputed value $\hat{Y}_i$ is different from the real value Y_i^* and 0 otherwise. The predictive accuracy of the imputation process D is simply the proportion of incorrect imputations:

$$D = n^{-1}\sum_{i=1}^{n} I(\hat{Y}_i \neq Y_i^*)$$

For the ordinal categorical variables Relation and Mstatus, the distance between the imputed and real value is taken into account using the following distance function

$$d(\hat{Y}_i, Y_i^*) = \frac{1}{2}\left[\frac{\left|\hat{Y}_i - Y_i^*\right|}{\max(Y) - \min(Y)} + I(\hat{Y}_i \neq Y_i^*)\right]$$

which is dependent on the relative difference between the imputed and real values. This measure is based on the assumption that the categories are assigned numeric values corresponding to the distance between categories. Thus, a generalized formula D_{gen} is defined as

$$D_{gen} = n^{-1}\sum_{i=1}^{n} d(\hat{Y}_i, Y_i^*)$$

For Age, different measures were required. To measure the preservation of true values, a distance function is used as described in the formula:

$$d_{L2}(\hat{\mathbf{Y}}, \mathbf{Y}^*) = \sqrt{\sum_{i=1}^{n}(\hat{Y}_i - Y_i^*)^2 / n}$$

For a measure that takes into account the relative distance between true and imputed values, the differences $\hat{Y}_i - Y_i^*$ above is replaced by, $\left(\hat{Y}_i - Y_i^*\right)/Y_i^*$. This will be denoted as **Relative D_{L2}**
To assess the preservation of aggregates, the following formula is used

$$m_k = \left|\sum_{i=1}^{n}(Y_i^{*k} - \hat{Y}_i^k)/n\right| = \left|m(Y^{*k}) - m(\hat{Y}^k)\right|$$

m_1 measures the preservation of the mean, and **m_2** measures the preservation of the variance.
The Kolmogorov-Smirnov distance (KS) measures the maximum distance between the empirical distribution functions of the real and imputed values. The distribution functions are measured by:

$$F_{Yn}(t) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq t)$$

The distribution function above is evaluated for each possible value t_j. The KS distance is then

$$KS = \max_j\left(\left|F_{Y^*n}(t_j) - F_{\hat{Y}n}(t_j)\right|\right)$$

References

CANCEIS User's Guide, version 4.5, Statistics Canada. Ottawa, Canada, 2007.

Chambers, R., "Evaluation Criteria for edit and imputation methods", in *Methods and Experimental Results from the Euredit Project*, Euredit Deliverable D6.1, Volume 2, Chapter 8.

Wagstaff, H., "Descriptions of datasets and their perturbations" , in *Methods and Experimental Results from the Euredit Project*, Euredit Deliverable D6.1, Volume 2, Appendix B.

Charlton, J. C., "Evaluating New Methods for Data Editing and Imputation - Results from the Euredit Project (UK)", *UNECE Statistical Data Editing Work session*. Madrid, Spain, 2003.

Bankier, M., Poirier, P. and Lachance, M., "Efficient Methodology within the Canadian Census Edit and Imputation System (Canceis)", *Proceedings of the Annual Meeting of the American Statistical Association*, August 5-9, 2001.