

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Bonn, Germany, 25-27 September 2006)

Topic (i): Editing nearer the source

DIAGNOSTICS FOR THE EVALUATION OF IMPUTED DATA

Supporting Paper

Submitted by the Office for National Statistics, UK¹

I. INTRODUCTION

1. The effective evaluation of the performance of an imputation process is important for any NSI or other data provider. The choice of both the imputation method, and the associated parameter settings, directly affect the quality of the output data. When evaluating an imputation process, the prime consideration is to ensure that the statistical properties of the data have been maintained. Ideally, measures of performance should provide information about the quality of the imputation process itself together with a broad indication of how it might be improved.

2. ONS have endorsed the use of CANCEIS (Bankier, 2000) as the Corporate tool for the editing and imputation of surveys and other data sources where the collected data is mainly nominal. The system has been successfully implemented on a number of household surveys and other statistical sources where the data are collected during the registration of Life Events. Based on the findings of an extensive three year research programme by the Methodology Directorate (MD), CANCEIS will also be implemented in the 2011 UK Census of Population and Housing (Wagstaff et al, 2004, Wagstaff and Rogers, 2005).

3. Prior to implementing CANCEIS in a live environment MD undertake an evaluation study to validate the system outputs. To facilitate the process, a test deck is constructed which represents a set of definitive criterion variables which are considered as a 'gold standard'. The accuracy of the output data is then assessed by repeatedly imputing the test deck in a fully controlled simulation environment. To date, the number of independent repetitions has been limited by the need to review a number of measurements which are taken to ensure that the imputation process has maintained the statistical properties of the data. The main aim of the current research is to develop a set of diagnostics which would facilitate, say, 1000 independent repetitions of the imputation process and clearly identify the quality of the outputs. This would also allow identification of the imputations which are of lower quality for closer scrutiny.

4. Chambers (2001) describes five desirable properties for an imputation procedure. For each of these, he specifies a series of general performance measures which are dependent on the scale of measurement of the variable being imputed. Since NSI's typically produce tabular outputs constructed from large datasets, emphasis is necessarily placed on the distributional accuracy of the imputed data with predictive accuracy being of secondary importance. Chambers comments that, of the five desirable properties, predictive accuracy is the hardest to achieve. Much of the data collected

¹ Prepared by Heather Wagstaff & Steven Rogers (Heather.Wagstaff@ons.gov.uk, Steven.Rogers@ons.gov.uk)

in household surveys and from administrative sources is nominal. For such data, Chambers proposes that the extent to which an imputation procedure preserves the marginal distributions can be assessed by calculating the value of a Wald-type statistic that compares the imputed and ‘true’ distributions of the variable across all levels of the categories simultaneously. This is essentially the Stuart-Maxwell significance test which follows a Chi-squared distribution (Stuart, 1955; Maxwell, 1970). To measure predictive accuracy, Chambers proposes a measure based on the proportion of true values recovered, which he denotes by ‘D’, together with an estimator for the variance. However, the variance is only valid when there is no evidence to reject the hypothesis that distributional accuracy has been maintained.

5. Despite the clarity of Chambers’ guidelines, in recent research, we observed conditions under which the Stuart-Maxwell test became unstable. Significant results were observed, implying that distributional accuracy was not maintained, when an exceedingly high proportion of the “true” values were recovered. The Stuart-Maxwell test, based on the Chi-squared distribution, like any test of significance has limited usefulness. Observing a small p-value provides evidence of an association, conversely, observing a large p-value indicates that little or no association exists. We are cautious of depending on the results of a Chi-squared test alone but rather seek to scrutinise the nature of the association. To investigate the observed phenomena in relation to the Stuart-Maxwell test further, we calculated the non-parametric Kappa coefficient (Cohen, 1960) which accounts for the agreement which might have arisen by chance (Fleiss, 1975). The value of Kappa ranges from -1 to 1 dependent on the strength of agreement, with zero implying that the agreement could have arisen by chance. The large sample variance for Kappa was also calculated (Everitt, 1968, Fleiss, 1969). Concerns about the utility of Kappa are well documented, and arise primarily because its value is dependent on the marginal distributions (Agresti, 1996). Hence, the proportion of true values recovered for each variable were also scrutinised in order to provide supporting evidence of the level of predictive accuracy actually achieved. Perfect predictive accuracy, by definition, implies perfect distributional accuracy. However, as predictive accuracy tends towards perfection the Stuart-Maxwell test becomes unstable and produces misleading results. These are the issues which we seek to address in this paper.

6. In the current paper we provide a brief overview of our recent research context and describe the conditions under which instability in the Stuart-Maxwell test was observed and also the identification of two paradoxes for the Kappa coefficient. An initial solution to the instability is presented in the form of a broad heuristic which forms a first step towards accounting for the relationship between the two measures of accuracy and provides a broad indication of the overall quality of the imputed data. We present an example of the utility of the heuristic before providing some concluding remarks about future development work.

II. RESEARCH CONTEXT

Simulation Environment

7. In order to fully evaluate the output from CANCEIS a simulation environment was developed based on a set of clean records, with known statistical properties. To achieve this, survey data in respect of a single UK administration area was identified as reference data, analysed and its properties documented. From the 2001 ONS Area Classifications, further administration areas were identified which were classified as having properties similar to the reference data. A synthetic dataset was then constructed, from the data for these further areas, by selecting a series of fully consistent records, as measured by a pre-specified set of edit rules, which contained no missing values. The resultant data forms the ‘truth deck’ which mirrored the structure of the original reference data and contained hierarchical records in respect of about 170,000 households and 400,000 individuals. Finally the data was perturbed by approximating the patterns of missing values observed in the reference data. When using the synthetic data to evaluate the efficacy of the imputation process we must always be mindful that an element of dependence exists between the “true” and imputed data. The “true” status of each record is known and hence, when considering the transition matrix of “true” versus imputed values, one marginal is always fixed whilst the may other vary.

8. When preparing CANCEIS for implementation much of the preparatory work is concerned with tuning the system parameters. The process begins by defining a benchmark from which to evaluate subsequent tuning efforts. In this setting, benchmarking involves applying CANCEIS to all variables in the dataset simultaneously, with the system parameters constrained to their default settings. This initial step facilitates the measurement of improvements which may be achieved by: identifying relationships amongst the variables; constructing meaningful subsets of the data; and adjusting the tuning parameters. In order to evaluate the quality of the outputs, the perturbed synthetic data is repeatedly imputed in a fully controlled simulation environment. Successful tuning is evidenced by the improved recovery of the statistical properties.

Baseline Analysis

9. Since CANCEIS implements a stochastic process we expect a measure of variation in the outputs. In order to control for the expected variation, we repeatedly impute the data and evaluate the outcomes. The synthetic data contains 25 mainly nominal variables which were imputed simultaneously 30 times with the system parameter settings fixed at their default values. The analyses focus on measuring both the distributional and predictive accuracy of the outputs. Distributional accuracy was measured by the Stuart-Maxwell test, at the 0.05 significance level, for each of the 25 variables within each of 30 imputed datasets. Predictive accuracy was assessed by the Kappa coefficient, together with its large sample standard error, and the proportion of agreement (p) between the “true” and imputed values was also calculated. Table 1 depicts the results.

Variable	Stuart-Maxwell % non sig	Kappa Coefficient		True Values p	Variable	Stuart-Maxwell % non sig	Kappa Coefficient		True Values p
		mean	(s.e.)				mean	(s.e.)	
Var_1	100.00	0.50	(0.017)	0.75	Var_19	0.00	0.85	(0.003)	0.92
Var_2	100.00	0.71	(0.008)	0.77	Var_20	0.00	0.88	(0.003)	0.93
Var_3	0.00	0.73	(0.008)	0.88	Var_21	0.00	0.84	(0.006)	0.91
Var_4	0.00	0.87	(0.003)	0.92	Var_22	0.00	0.74	(0.011)	0.86
Var_5	0.00	0.80	(0.005)	0.90	Var_23	10.00	0.66	(0.021)	0.79
Var_6	26.67	0.25	(0.011)	0.88	Var_24	100.00	0.56	(0.031)	0.73
Var_7	0.00	0.29	(0.012)	0.89	Var_25	100.00	0.26	(0.053)	0.41
Var_8	10.00	0.25	(0.008)	0.70					
Var_9	0.00	0.23	(0.006)	0.87	Var_1*2	96.67	0.73	(0.006)	0.76
Var_10	3.33	0.44	(0.009)	0.86	Var_1*3	0.00	0.78	(0.005)	0.83
Var_11	0.00	0.32	(0.008)	0.80	Var_1*4	0.00	0.87	(0.002)	0.89
Var_12	0.00	0.63	(0.004)	0.86	Var_1*5	0.00	0.84	(0.003)	0.87
Var_13	0.00	0.84	(0.003)	0.93	Var_2*3	0.00	0.80	(0.005)	0.83
Var_14	0.00	0.32	(0.002)	0.47	Var_2*4	0.00	0.87	(0.002)	0.89
Var_15	0.00	0.92	(0.002)	0.96	Var_2*5	0.00	0.84	(0.003)	0.97
Var_16	13.33	0.60	(0.003)	0.74	Var_3*4	0.00	0.89	(0.002)	0.91
Var_17	100.00	0.52	(0.005)	0.67	Var_3*5	0.00	0.81	(0.004)	0.89
Var_18	0.00	0.85	(0.003)	0.91	Var_4*5	0.00	0.89	(0.002)	0.91

Table 1 - Baseline Evaluation: measuring distributional and predictive accuracy following 30 repeated applications of CANCEIS to 25 variables simultaneously.

10. Initial analysis indicated that the imputation system achieved marginal homogeneity for only 5 of the variables in all 30 imputed datasets. For example, the marginal distribution of Var_2 was maintained in all 30 imputed datasets, as measured by the Stuart-Maxwell test and achieved good mean predictive accuracy ($\kappa=0.71$, $se(\kappa)=0.008$, and $p=0.77$). The joint distribution of Var_1*2 was maintained for 29 of the 30 imputations and also achieved good mean predictive accuracy ($\kappa=0.73$, $se(\kappa)=0.006$, and $p=0.76$). In respect of marginal homogeneity, the results for the other variables were poor. However, there were instances where exceedingly high proportions of the “true” values were recovered and yet the Stuart-Maxwell test produced a significant result. For example, the results for Var_15 indicated a 95.93% recovery of the “true” values ($\kappa=0.92$, $se(\kappa)=0.002$) and yet the results of the Stuart-Maxwell test indicated that the marginal distributions were not maintained for any of the 30 datasets.

Limitations of the Stuart-Maxwell Test and Kappa Coefficient

11. In order to investigate the limitations of the Stuart-Maxwell test, a categorisation of the Kappa coefficient was introduced. Landis and Koch (1977) give some arbitrary but useful benchmarks for evaluation which are depicted in Table 2.

Kappa Coefficient	Strength of Agreement
0.00 - 0.20	Poor
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1.00	Almost Perfect

Table 2: Kappa Coefficients: benchmarks for strength of agreement.

12. The imputed data were reanalysed with reference to the strength of agreement for Kappa and the proportion of “true” values recovered. The analysis provided evidence that, in the presence of high predictive accuracy, the Stuart-Maxwell test becomes unstable. This result is hardly surprising since the significance test makes no allowance for the proportion of cases with an exact recovery of the “true” values. If we consider the transition matrix, formed from the cross classification of “true” versus imputed values, only the discordant pairs, the values which sit in the upper and lower triangles, directly contribute to the significance test. The patterns of symmetry, or conversely asymmetry, have significant influence on the test result. For example, the imputation system apparently failed to maintain marginal homogeneity for 3 key demographic variables (Var_3, Var_4, Var_5) in any of the 30 imputed datasets. Table 2 deems their respective mean strength of agreement to be substantial or almost perfect (Var_3: $\kappa=0.73$, $se(\kappa)=0.008$, $p=0.88$; Var_4: $\kappa=0.87$, $se(\kappa)=0.003$, $p=0.92$, Var5: $\kappa=0.80$, $se(\kappa)=0.005$, $p=0.90$). On closer scrutiny of the outputs we observed that the failure of the significance test was attributable to a small proportion of off-diagonal asymmetry. This was also true for a number of other variables. Table 3 depicts example output in respect of Var_4.

True Values	Imputed Values										
	cat_1	cat_2	cat_3	cat_4	cat_5	cat_6	cat_7	cat_8	cat_9	Total	
cat_1	n	3795	0	0	0	0	0	0	0	0	3795
	%	48.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	48.17
cat_2	n	35	20	0	12	7	9	19	9	0	111
	%	0.44	0.25	0.00	0.15	0.09	0.11	0.24	0.11	0.00	1.41
cat_3	n	0	0	0	1	0	1	0	0	0	2
	%	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.03
cat_4	n	24	7	1	753	1	49	55	21	0	911
	%	0.30	0.09	0.01	9.56	0.01	0.62	0.70	0.27	0.00	11.56
cat_5	n	24	10	0	0	178	3	0	2	0	217
	%	0.30	0.13	0.00	0.00	2.26	0.04	0.00	0.03	0.00	2.75
cat_6	n	52	12	0	56	2	143	21	14	0	300
	%	0.66	0.15	0.00	0.71	0.03	1.82	0.27	0.18	0.00	3.81
cat_7	n	13	19	0	65	0	23	49	35	0	204
	%	0.17	0.24	0.00	0.83	0.00	0.29	0.62	0.44	0.00	2.59
cat_8	n	21	11	0	25	4	15	20	9	0	105
	%	0.27	0.14	0.00	0.32	0.05	0.19	0.25	0.11	0.00	1.33
cat_9	n	0	0	0	0	0	0	0	0	2233	2233
	%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	28.34	28.34
Total	n	3964	79	1	912	192	243	164	90	2233	7878
	%	50.32	1.00	0.01	11.58	2.44	3.08	2.08	1.14	28.34	100.00

Table 3: Example of Transition Matrix: Var_4 to demonstrate the instability of the Stuart-Maxwell test in the presence of a high predictive accuracy.

13. In Table 3 the rows represent the “true” values and the columns the imputed values. Of the 7878 imputed records some 91.14% (7180) of the “true” values were recovered exactly. The significant Stuart-Maxwell results were primarily attributable to just 2.15% (169) of the records which were imputed into category 1 but with no reciprocal in the upper triangle. The result is largely due to the presence of 2 dominant categories in the variable together with just 2 missing values for category 3 of which none were recovered exactly in this realisation of imputed data. In the synthetic dataset, the

overall distributions of the donor and recipient records are similar but there are also instances where missing values were introduced in rare categories leaving few or no exact matches amongst the donor records. However, this situation reflects what we might expect in real life and the data in Table 3 is just one such realisation. When there are insufficient records which match the recipient exactly, then simplistically, the imputation system selects a donor based on proportional representation of the donor records. It is interesting to note that, when collapsing the transition matrix during analysis, we observed that the value of κ increased from that calculated for the full table. This is simply a function of the categorisation of the variable and not problematic in our current research.

14. The baseline analysis indicated the need to define an acceptable threshold for predictive accuracy, above which the results of the Stuart-Maxwell test is invalid, but below which they might be assumed stable. Initial analysis focussed on the proportion of non-significant Stuart-Maxwell tests observed for each variable which was plotted and overlaid, firstly with the associated normalised Kappa coefficients, then secondly with the proportion of “true” values recovered, both ordered by predictive accuracy. Figure 1 depicts the results.

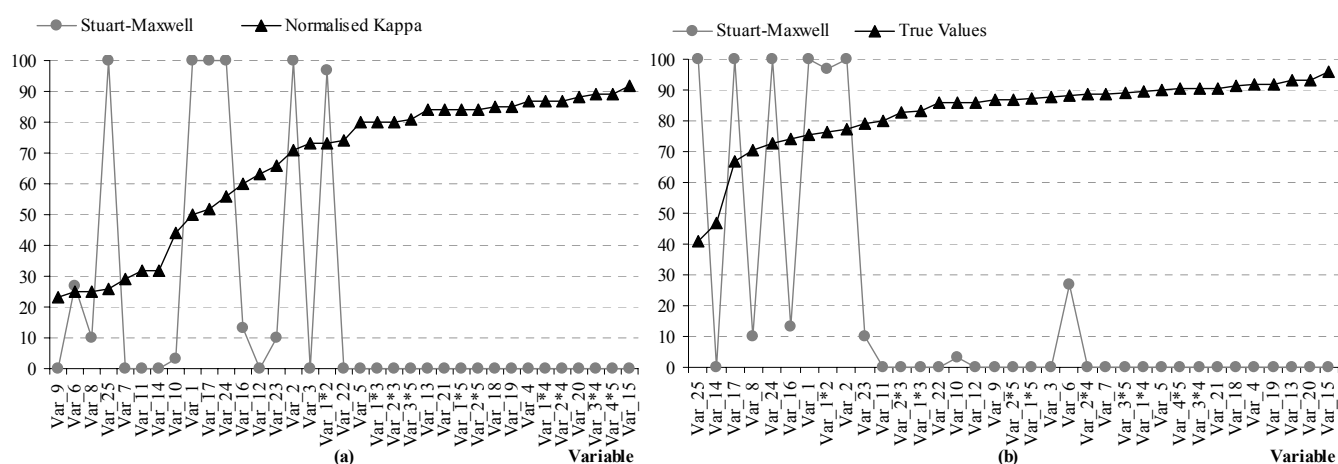


Figure 1. Baseline Imputation: graph of proportions of non-significant Stuart-Maxwell results by (a) normalised mean Kappa coefficients; (b) mean proportion of true values recovered; ordered by increasing predictive accuracy.

15. The graphs in Figure 1 both support the theory that the results of the Stuart-Maxwell test are unreliable in the presence of high predictive accuracy. It is interesting to note the difference in the ordering of the variables between the Kappa coefficient and the simple proportion of “true” values recovered. Figure 1(a) implies that when the Kappa coefficient is around 0.80, indicating high predictive accuracy, the Stuart-Maxwell test consistently returns a significant result. Figure 1(b), which shows the simple proportions recovered, provides supporting evidence of the instability of the Stuart-Maxwell test. The result is hardly surprising given that the Stuart-Maxwell test is dependent on the values in the discordant cells. However, the two figures clearly demonstrate the correspondence between the values of the Kappa coefficient and the proportion of “true” values recovered. They also clearly demonstrate a paradox associated with the Kappa coefficient: high levels of agreement with low Kappa values. However, in the current research, this was not problematic since if the Kappa coefficient is considered alone it might lead to false negatives and never to false positives. However, accounting for the false negatives was managed by directly observing the proportion of “true” values recovered. A further issue was identified with the Kappa: as the imputed distributions changed it was difficult to predict the associated changes. But then this is the case for any statistic with dependence on the marginals. In the simulation environment the “true” marginal is fixed for each variable.

III. PROPOSAL FOR HEURISTIC

16. In the simulation environment, when preparing an imputation system for implementation, we seek to independently repeat the imputation process, say, 1000 times and evaluate the outputs. The purpose of the current research is to find diagnostics, or summary measures, which indicate the quality of the imputed data. Here, the Stuart-Maxwell test is considered in combination with a categorisation of the Kappa coefficients proposed by Landis and Koch (1977), but with supporting evidence of predictive accuracy from the proportion of “true” values recovered. Based on our research findings, we propose that the following heuristic provides a good starting point to begin resolving the issues identified in the previous section.

Case	Condition	Imputation Status
	Stuart-Maxwell test = not significant	
1	$\kappa = 0.80 - 1.00$: almost perfect predictive accuracy	Accept
2	$\kappa = 0.60 - 0.79$: substantial predictive accuracy	Accept
3	$\kappa \leq 0.59$: moderate or less predictive accuracy	Scrutinise
	Stuart Maxwell test = significant	
4	$\kappa = 0.80 - 1.00$: almost perfect predictive accuracy	Scrutinise
5	$\kappa \leq 0.80$: substantial or less predictive accuracy	Reject

Table 4 – Proposed Heuristic: evaluating the quality of imputed data

17. The optimal values for the proportion of true values recovered has yet to be defined but is calculated for all cases. Cases 1 and 2 are straight forward: where a non-significant Stuart-Maxwell test is observed together with almost perfect (case 1) or substantial (case 2) predicted accuracy we would accept the imputed data. However, data falling into cases 3 and 4 should always be scrutinised.

Case 3: when there is a non-significant test combined with poor to moderate predictive accuracy we should scrutinise the data before simply accepting the imputations. During research, this situation occurred under conditions where matching variables were insufficient and not accurately predicting the correct outcome i.e. unable to discriminate between categories of the donor records. In other cases records contained rare combinations of characteristics and insufficient donors with matching characteristics were present in the data. The issue here is, in the presence of marginal homogeneity, is any level of predictive accuracy acceptable no matter how low? For this reason we closely scrutinise the outputs.

Case 4: when there is a high recovery of the true values, leaving a very low proportion of discordant records, the Stuart-Maxwell test can give a spurious significant result and must be interpreted in conjunction with the Kappa coefficient and/or proportions of “true” values recovered. Both of the conditions identified in Case 3 above can be present here. Where the level of missingness in the data is low the marginal distributions of the complete dataset are not really affected and the higher order distributions are more likely to be preserved.

The final case relates to a significant test result combined with a lower level of predictive accuracy where the imputed data should be rejected as being unsatisfactory. The crucial point is to understand the detailed structure of the data and to accurately interpret the results. It is straightforward to investigate such cases within a simulation environment but not practicable within the confines of a live production process.

IV. EXAMPLE APPLICATION OF THE HEURISTIC

18. By applying logistic regression to the synthetic data, 5 key variables were identified which formed a self-predicting set i.e. each acted as a predictor for the others although the strength of the

association was sometimes weak. The variables comprised 4 nominal and one scalar variable which was categorised but the ordering ignored. The 5 variables were repeatedly imputed simultaneously 30 times with the system parameters fixed to the default settings. The results are compared with those from the baseline analysis. Table 5 depicts the results.

Imputation Strategy	Variable	Categories of Heuristic						
		Stuart-Maxwell: not significant				Stuart-Maxwell: significant		
		1	2	3		4		5
Accept	Accept	Accept	Reject	Accept	Reject	Reject		
1. Baseline (25 vars simultaneously)	Var_1			30				
	Var_2		30					
	Var_3							30
	Var_4					30		
	Var_5					24		6
2. 5 key variables only	Var_1	2		28				
	Var_2		30					
	Var_3		30					
	Var_4	17				13		
	Var_5	28						2

Table 5: Proposed Heuristic: categorisation of results from imputation process against differing imputation strategies.

19. The results were categorised by the levels of the proposed heuristic and against the differing imputation strategies: 1. impute 25 variables simultaneously; 2. impute 5 key variables only. Table 5 clearly shows the improvement gained from imputing the 5 key variables alone whilst also demonstrating the utility of the heuristic. We scrutinised the 125 transition matrices for the variables that fell into the heuristic categories 3 or 4 and concluded that they all related to acceptable imputations. An example of case 4, from the baseline imputation, was depicted in Table 3 above and Table 5 depicts an example of case 3.

True Values		Imputed Values		
		cat_1	cat_2	Total
cat_1	n	544	135	679
	%	44.99	11.17	56.16
cat_2	n	120	410	530
	%	9.93	33.91	43.84
Total	n	664	545	1209
	%	54.92	45.08	100

Table 6: Example of Heuristic Case 3: significant Stuart-Maxwell (McNemar=0.76) in combination with a moderate predictive accuracy ($\kappa = 0.573$, $se(\kappa) = 0.0006$, with 78.9% true values recovered).

20. Since the variable is comprised of 2 levels the Stuart-Maxwell (McNemar) statistic simply tests the difference of the proportion of records in category 1 before and after imputation. The proportion of concordant pairs accounts for 78.9% whilst the sum of the discordant pairs (11.17% and 9.93% respectively) is somewhat high. However, the relatively even spread of the discordant pairs about the diagonal ensures marginal homogeneity. There was insufficient predictive power amongst the other matching variables to adequately discriminate between the categories of this variable. Hence, we conclude that the data represents an acceptable imputation.

21. Figure 1(a) above showed that when Kappa is around 0.80, indicating high predictive accuracy, the Stuart-Maxwell test consistently returned a significant result. This result was investigated further when the 5 key variables were imputed alone. The proportion of non-significant Stuart-Maxwell tests observed for each of the 5 key variables, and their joint distributions, was plotted and overlaid with the associated normalised Kappa coefficients, firstly for the baseline analysis, and secondly when imputed alone, both ordered by predictive accuracy. Figure 2 depicts the results.

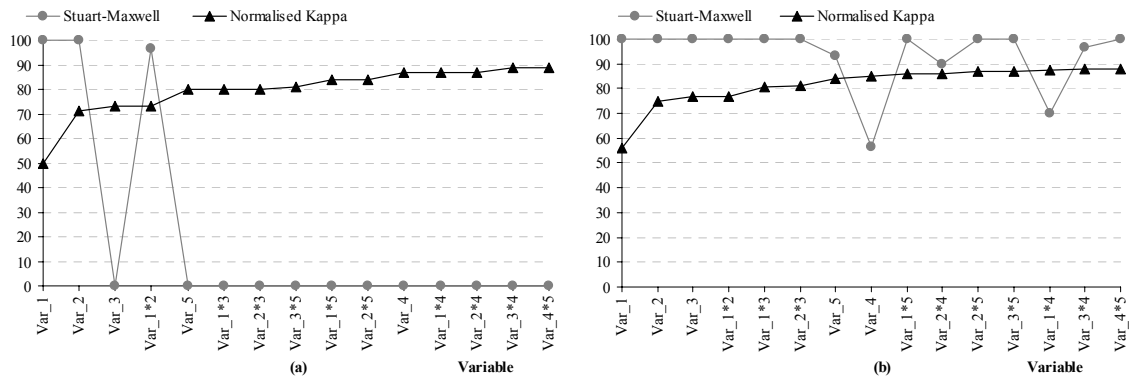


Figure 2. 5 Key Variables: Graph of proportions of non-significant Stuart-Maxwell results and normalised mean Kappa coefficients, ordered by increasing predictive accuracy: (a) as part of the baseline analysis; and (b) when imputed alone as a self predicting set.

22. Figure 2 indicates that a substantial improvement was achieved by changing the imputation strategy. Note that the ordering of predictive accuracy changed between Figures 2 (a) and (b). For the 5 univariate and 10 joint distributions, Figure 2(b) shows that 10 passed the Stuart-Maxwell test in all 30 imputed datasets. Further, Var_4, and 3 of its joint distributions, achieved better Stuart-Maxwell results in Strategy 2 and yet the associated average Kappa values fell slightly. Var_1*4 passed 70% of the Stuart-Maxwell tests, against failing all in Strategy 1, yet there was no change in the Kappa coefficient ($\kappa=0.87$, $se(\kappa)=0.002$). To investigate the results further, the proportion of non-significant Stuart-Maxwell tests for the 5 key variables, and their joint distributions, was plotted and overlaid with the associated proportion of “true” values recovered, firstly for the baseline analysis, and secondly when imputed alone, both ordered by predictive accuracy. Figure 3 depicts the results.

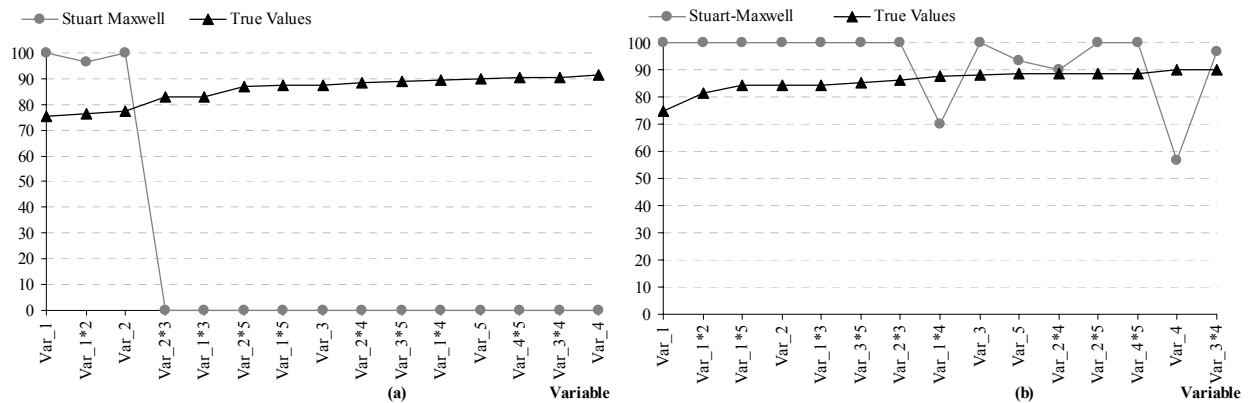


Figure 3. 5 Key Variables: Graph of proportions of non-significant Stuart-Maxwell results and proportion of true values recovered, ordered by increasing predictive accuracy: (a) as part of baseline analysis; and (b) when imputed as a self predicting set.

23. Figure 3 confirms the findings that, overall a substantial improvement was achieved by changing the imputation strategy. However, the earlier logistic regression indicated that Var_4 and Var_5 had weaker association than the remaining three variables. This is clearly shown in Figure 3 from the changes in the proportion of “true” vales recovered between the two graphs. From Figure 3(a), from the baseline imputation Var_4, Var_5, together with their joints with the other key variables, failed all the Stuart-Maxwell tests. However, from Figure 3(b) those same variables, including their joints, all passed the Stuart-Maxwell test more than 50% of the time with some passing 100%. However, the associated proportion of “true” values recovered was lower. For example, Var3*5 failed all Stuart-Maxwell tests in the baseline, recovering 89.17% of the true values, yet passed all 30 Stuart-Maxwell tests when the 5 key variables were imputed alone, recovering only 85.46% of the “true” values. As the Stuart-Maxwell test moved from ‘all fail’ to ‘all pass’, the proportion of “true values” fell by an average of 3.71%. However, Var_1 passed all Stuart-Maxwell

tests in both imputation strategies and yet the proportion of true values fell from 75.37% to 74.53%. Amongst the self-predicting set there was little predictive power to discriminate between the levels of this variable. Table 7 depicts the detailed results for both imputation strategies together with the differences.

Variable	Baseline - All 25 Variables			5 Key Variables Only			Difference		
	Stuart-Maxwell (a)	Kappa Coeff't (b)	% True Values (c)	Stuart-Maxwell (d)	Kappa Coeff't (e)	% True Values (f)	Stuart-Maxwell (d) - (a)	Kappa Coeff't (e) - (b)	% True Values (f) - (c)
Var_1	100.00	0.50	75.37	100.00	0.56	74.53	0.00	0.06	-0.84
Var_2	100.00	0.71	77.45	100.00	0.75	84.39	0.00	0.03	6.94
Var_3	0.00	0.73	87.60	100.00	0.77	88.23	100.00	0.04	0.63
Var_4	0.00	0.87	91.66	56.67	0.85	89.96	56.67	-0.02	-1.70
Var_5	0.00	0.80	89.84	93.33	0.84	88.34	93.33	0.04	-1.50
Var_1*2	96.67	0.73	76.36	100.00	0.77	81.45	3.33	0.04	5.09
Var_1*3	0.00	0.78	82.99	100.00	0.81	84.47	100.00	0.03	1.48
Var_1*4	0.00	0.87	89.41	70.00	0.87	87.84	70.00	0.00	-1.57
Var_1*5	0.00	0.84	87.45	100.00	0.86	84.36	100.00	0.02	-3.09
Var_2*3	0.00	0.79	82.75	100.00	0.81	86.40	100.00	0.02	3.65
Var_2*4	0.00	0.87	88.59	90.00	0.86	88.46	90.00	-0.01	-0.13
Var_2*5	0.00	0.84	87.01	100.00	0.87	88.55	100.00	0.03	1.54
Var_3*4	0.00	0.89	90.61	96.67	0.88	90.05	96.67	-0.01	-0.56
Var_3*5	0.00	0.81	89.17	100.00	0.87	85.46	100.00	0.06	-3.71
Var_4*5	0.00	0.89	90.52	100.00	0.88	88.77	100.00	-0.01	-1.75

Table 7 - Evaluation of Imputation Strategies: measuring distributional and predictive accuracy following 30 repeated applications of CANCEIS to: (1) 25 variables simultaneously; and (2) 5 key variables alone.

24. Table 7 depicts the patterns of Kappa coefficients and the proportions of “true” values recovered between the two imputation strategies. In some cases the results are counterintuitive: in order to maximise the number of Stuart-Maxwell passes the level of predictive accuracy decreases. The evidence from Table 7 clearly shows that imputing a subset of 5 key variables alone achieves an improvement in the recovery of the statistical properties of the data from that achieved in the baseline.

V. CONCLUDING REMARKS

25. Chambers (2001) clearly states that the evaluation of an imputation method should be based on measures of distributional and predictive accuracy. For large datasets, where the outputs are typically based on aggregates, emphasis is placed on distributional accuracy whilst predictive accuracy is often viewed with lesser importance. However, our current research has indicated that such thinking does not always lead to clear conclusions about the quality of the imputed data or the efficacy of the tuning parameters.

26. A key aim of the current research is to develop diagnostics to facilitate large numbers of repetitions of the imputation process whilst also providing information about the quality of the output data. Whilst the proposed heuristic is not fully developed it has proved to be effective and facilitated the evaluation process. The thresholds at which the Stuart-Maxwell statistic becomes unstable is dependent on the marginal values and the distribution of responses within the cells. However, the threshold is relatively straightforward to identify for 2 level nominal variables, especially given that the row marginals are fixed in the simulation environment since the “true” values are known. Even so, it is not straightforward for larger tables. The Kappa coefficient is based on the observed and expected proportions of records on the leading diagonal of a contingency table. Hence the value of κ is clearly dependent on the proportion of records in each category. For any given value of the proportion of records that are recovered exactly, and sit on the leading diagonal, we can expect any number of differing values of κ since, although the marginal of “true” values is fixed, the imputed marginals can vary. With large contingency tables, more than 2 categories, it is even harder to judge comparably. Part of our analysis involved collapsing large contingency tables to try to understand the data, it was

observed that the value of κ increased from that calculated for the full table. The proportions of “true” values recovered were always calculated as a safety net for the known paradoxes of Kappa coefficient.

27. In conclusion, in ongoing research we are seeking to make the heuristic more robust and form the basis of a set of diagnostics. We continue to apply the Stuart-Maxwell, in combination with the Kappa coefficient, and supported by the proportion of “true” values recovered. Despite the short comings of the Kappa coefficient it is undoubtedly the right type of approach. Kappa may be interpreted as the proportional agreement, corrected for the expected values, which appears to be the best approach to this type of problem. However, what is clear is that, as always, it is important to view the raw data.

References

Agresti, A. (1996), “An Introduction to Categorical Data Analysis”, Wiley Series in Probability and Statistics.

Bankier, M. (2000) “2001 Canadian Census minimum Change Donor Imputation Methodology”. Working Paper No. 17, UN/ECE Work Session on Statistical Data Editing, Cardiff.

Chambers, R. (2001) “Evaluation Criteria for Statistical Editing and Imputation”. National Statistics Methodology Series No. 28.

Cohen, J. (1960) “A measure of agreement for nominal scales”. Educational and Psychological Measurement, 20: 37-46.

Everitt, B.P. (1968) “Movements of the statistics kappa and weighted kappa”. British Journal of Mathematical and Statistical Psychology, 21, 97-103.

Fleiss et al. (1969) “Large sample standard errors of kappa and weighted kappa”. Psychological Bulletin, 72, 323-7.

Fleiss, J.L. (1981) “The measurement of interrater agreement.” In: Fleiss J L., editor. “Statistical methods for rates and proportions”. New York, N.Y: Wiley. pp. 212–236.

Landis, J.R. and Koch, G.C. (1977) “The measurement of observer agreement for categorical data”. Biometrics, 33, 159-174.

Maxwell, A.E. (1970) “Comparing the classification of subjects by two independent judges”. British Journal of Psychiatry, 116, 651-655.

McNemar, Q. (1947) “Note on the sampling error of the difference between correlated proportions or percentages”. Psychometrika, 12, 153-157.

Stuart, A. (1955) “A test for homogeneity of the marginal distributions in a two-way classification”. Biometrika, 42, 412-416.

Wagstaff, H.F., Spicer P.C., Skentlebery, R., (2004) “Report on the initial evaluation of CANCEIS on 2001 Census data ”. ONS Internal Report.

Wagstaff, H.F. and Rogers, S., (2006) “Application of CANCEIS to 2001 Census data: Technical Report”. ONS Internal Report.