Imputation Variance Estimation for Statistics New Zealand's Accommodation Occupancy Survey

Raazesh Sainudiin* and Richard Penny†

Department of Mathematics and Statistics*,
University of Canterbury,
Private Bag 4800, Christchurch, New Zealand
r.sainudiin@math.canterbury.ac.nz and

Statistics New Zealand†,
Private Bag 4741
Christchurch, New Zealand
Richard.Penny@stats.govt.nz

June 15, 2009

Abstract

We formulate the problem of imputation variance estimation for the Accommodation Occupancy Survey (AOS) run on behalf of the Ministry of Tourism by Statistics New Zealand and develop a methodology and the accompanying code to address this problem. We use nonparametric blocked bootstrap techniques to provide consistent estimates of the imputation variance under the assumption of homogeneity within the predefined imputation cells.

— This work was sponsored by The New Zealand Ministry of Tourism —

Contents

List of Figures

List of Tables

1 Imputation Variance of AOS Statistics

1.1 Introduction and Background

The Accommodation Occupancy Survey (AOS) is run on behalf of the Ministry of Tourism by Statistics New Zealand. It is administered monthly, and is not a sample survey, but rather a full census of accommodation providers in New Zealand (for more details on the survey see `http://www.stats.govt.nz/products-and-services/info-releases/accom-survey.htm`). The AOS data is the set of responses to the AOS questionnaire from each accommodation provider for each month. The Ministry of Tourism is interested in various statistics or transformations of the AOS data. These AOS statistics, being key indicators of the country's tourism sector, are the fundamental objects of interest in this study. By running a census rather than a sample survey there is no sample error, but there continues to be non-sample error in the AOS statistics, which implies that there is still some uncertainty in any outputs from this data collection. It is well known [6, Part IV] that non-sample error contains many components and is not easy to estimate. In this report we investigate the non-sample error that arises from non-response to AOS, and investigate a procedure currently used by Statistics New Zealand to account for the non-response in its AOS statistics.

In the AOS, as for almost all surveys, there is a certain level of non-response to the survey which will contribute to overall non-sampling error. In the AOS, this non-response can either arise from an accommodation provider not returning the questionnaire (unit non-response), or returning the questionnaire but not providing answers to all the questions (item non-response). In either case, some decision is required on what to do regarding the missing data as it will influence the AOS statistics produced. One possible approach is to only use the responses, termed a "complete case analysis". This has the assumption that the nonrespondents are overall similar to the respondents. This assumption is similar to Missing Completely at Random [3].

An alternative approach is to estimate any missing response for each survey unit. This process is termed imputation. Imputation may use any data given by the respondent, data from other respondents, previous responses, or other data sources to estimate the missing responses. In AOS, a combination of current responses from all accommodation providers and the previous responses from the accommodation provider being imputed are used. The resulting mix of responses and imputations are used to produce a point estimate of a complete data file or a set of AOS statistics of interest. All AOS statistics of interest in this study are merely functions of the AOS data. Recall that a point estimate is our single best guess for the object of interest. This object is the missing data or non-responses and/or the AOS statistic of interest that further depend on all of the data, i.e., both responses and non-responses. An imputation method should also provide a confidence set for the missing data or a confidence interval for an AOS statistic. Recall that a confidence set or interval will contain the quantity of interest with a high probability. Thus, the confidence set is a formal way to incorporate the uncertainty inherent in the imputation process and necessary for realistic interpretations of various AOS statistics that are obtained from further transformations of the imputed data. This is a particularly relevant issue for censuses, such as AOS, where there is no sample error to report. To generate a confidence interval arising from the uncertainty associated with the imputation one needs to understand the imputation model used.

For the AOS, available data is used to provide an estimate of the non-responses generally by modelling the heterogeneity in survey response within the whole population by using multiple homogenous sub-populations. The respondents and non-respondents are first grouped into relatively homogeneous sub-populations or imputation cells. The respondents from a given imputation cell are assumed to provide their responses according to the same underlying distribution for the purposes

of model building, and furthermore that this model is assumed to apply to the non-responses. This is how homogeneity is typically exploited in imputation. For the AOS, we have monthly data with not only correlation between the responses in any given month but also responses between months. Currently in AOS the imputation cells are not changed from month to month. Thus, the assumption is that the homogeneity within an imputation cell and heterogeneity between imputation cells is consistent over time.

The AOS imputation procedure generates point estimates of the expected response(s) of any particular accommodation provider with missing data. It uses the current month and sometimes the previous month's data. As such, it is reliant on the assumption that any month's data is homogeneous within the sub-population, an assumption that is known to be untrue for some months. Statistics New Zealand has an approach to address this problem, but it is done separately from the imputation model. Also the imputation procedure currently used by Statistics New Zealand does not give a confidence set for the imputed data and therefore neither for any of the statistics produced from the AOS data. In this project we aim to provide methodologies for consistent point estimates as well as confidence sets for various AOS statistics that potentially fully and efficiently utilises all information over space and time from the AOS.

There has been considerable work done on estimating the variance due to various imputation models. One approach is to apply resampling techniques to imputation either with jackknife [2, 1] or bootstrap [7]. Another approach is to use multiple imputation where the imputation is repeated several times, i.e. bootstrapped either parametrically or nonparametrically, for each non-response. This results in several possibly distinct realisations of the imputed data [4, 5]. Then the statistics of interest are computed for each imputed data. This yields the empirical distribution of the statistics under the imputation model and gives the asymptotically consistent point estimate as well as the confidence set that correctly reflects the imputation variance. We use such a non-parametric bootstrap strategy for imputation variance estimation in this study.

2 The Data Collected in AOS

2.1 General Introduction

When a new accommodation provider (AP) is identified (termed a "birth") Statistics New Zealand starts collecting information on this accomodation provider. This collected information includes some data which does not change over time.

1. its geographical location
2. type of accommodation, such as,
   (a) motel
   (b) hotel
   (c) backpackers
   (d) camping ground
   (e) hosted accommodation — though this type of accommodation provider was no longer surveyed after August 2009

The geographical location and type of accommodation are used for editing or pre-processing as well as for subsequent imputation and statistical analysis purposes since the spatial proximity of accommodation providers reflect the geographical location of tourist attractions or centres of business or government, and the behaviour of the accommodation types vary as they cater for different types of guests.

The accommodation provider is then surveyed every month (see § 2.2) until the accommodation provider ceases to exist (termed a "death"). The accommodation provider is defined as the business entity that provides the accommodation, not the physical entity. Thus, if a business is sold this will result in a death and a birth even though the physical units are the same.

The quality and internal consistency of the data supplied by accommodation providers is another contributor to non-sample error. Editing or pre-processing of the data can identify and fix obvious inconsistencies and errors, but it is possible that minor mistakes by the accommodation providers will not be identified in the pre-processing step as it can be difficult to distinguish incorrect data from anomalous data. For example, is the response zero through a respondent mistake, or are there unusual and one-off circumstances for that month that mean there were no guests that month? For the purposes of this work we assume that all responses in the pre-processed data are correct.

2.2 Monthly Survey

Every month Statistics New Zealand asks each accommodation provider to provide data for the preceding month on 5 variables.

1. Number of Stay Units (SU)
2. Total Stay Nights (SN)
3. Total Guest Nights (GN)
4. Origin of Guests is a tri-partition of GN
   (a) Domestic Guest Nights ($\mathsf{GN_D}$)
   (b) International Guest nights ($\mathsf{GN_I}$)
   (c) Unknown Guest Nights ($\mathsf{GN_U}$)
5. Total Guest Arrivals (GA)

A copy of the current questionnaire is available on the Statistics New Zealand website [8]. To understand the imputation methodology it is necessary to understand what each variable measures and its possible relationship to the other variables.

2.2.1 Number of Stay Units

A stay unit (SU) is the physical entity that the accommodation provider provides to the guest. That is, it can be a bed, a room, a collection of rooms, a cottage, an area of land, or some other entity that can be occupied by a guest or guests overnight. The number of SU for any particular accommodation provider generally does not change from month to month as any change would arise from the accommodation provider physically expanding or contracting. As such it is used as a basis for much of the imputation of other variables, and thus is imputed first. If the respondent does not

supply a response to SU for any given month it can generally be safely assumed to be the same as the last response.

For most accommodation providers any particular SU could have from zero to several people staying in that SU on any given night of a month. At one extreme, the range of possible guests per SU is largest for camping grounds, whereas at the other extreme, it is generally one for most backpackers, as most of the SU are defined as a bed, and one guest can occupies that SU on any given night. Therefore, it is necessary to use to include the type of accommodation providers when determining the sub-populations for imputation, i.e., the imputation cells.

2.2.2 Total Stay Nights

Each night, a certain number of the SU at the accommodation provider will be occupied. For the purposes of providing stay night (SN) it does not matter how many guests are in the occupied SU, but only that the SU is occupied and not empty. The number of SU occupied any night is the SN for that date. The monthly SN value is the sum of the individual SN for each day in the month.

Thus, SN for a given month can be between 0 (no guests that month) to $\mathsf{SU} \times d$, where d is the numbers of days (or nights) in the month of interest (i.e. all SU occupied every night in that month). Associated with this value is the statistic termed Occupancy Rate (OR) which is:

$$\mathsf{OR} := \frac{\mathsf{SN}}{\mathsf{SU} \times d}$$

Thus, the range of OR is the unit interval $[0, 1]$. The OR can be regarded as the average fraction of units occupied over the month and is a provider-specific measure of accommodation occupancy due to the normalisation by the provider's SU. However it is expected that accommodation occupancy rates are very likely to be similar for accommodation types in the same area. The similarity of this derived statistic enables the responses from many accommodation providers to be used to estimate missing responses.

2.2.3 Total Guest Nights

Guest Nights (GN) is the sum of the number of nights each guest stays at the accommodation provider over a given month. For example, if 3 guests are in a SU for 4 nights then $\mathsf{GN} = 3 \times 4 = 12$.

The minimum value for GN is 0 (no guests that month). For many accommodation providers the number of guests in any SU can vary from night to night. Therefore only broad relationships between the variables can be deduced, though these will differ across different accommodation types. For example for backpackers the SU is generally a single bed where only one guest can occupy the SU. Thus for most backpacker APs SN will be close to GN.

2.2.4 Origin of Guests

The accommodation provider is asked to disaggregate GN into three classes

1. Domestic Guest Nights ($\mathsf{GN_D}$) - The number of GN where the guests are normally resident in New Zealand
2. International Guest Nights ($\mathsf{GN_I}$) - The number of GN where the guests are normally resident overseas

3. Unknown Guest Nights ($\mathsf{GN_U}$) - The number of GN where the accommodation provider does not know the residency of the guests

These variables are merely disaggregating GN into 3 mutually exclusive classes $\mathsf{GN} = \mathsf{GN_D} + \mathsf{GN_I} + \mathsf{GN_U}$. In the final data used for outputs the $\mathsf{GN_U}$ are allocated to $\mathsf{GN_I}$ and $\mathsf{GN_D}$, but for the purposes of this project we have used the values for the three variables as provided by the respondents.

2.2.5 Total Guest Arrivals

When a guest, or guests, first occupies a SU, the number of guests is classed as a Guest Arrival (GA). That is, if 2 people check into a SU then $\mathsf{GA} = 2$, irrespective of how long they stay, whereas each extra night they stay will increment GN. If guests occupy a SU for a period of time, then leave for at least one night before reoccupying the SU this will be regarded as two different occurrences of GA. The values of GA are assigned to the month when they first occupied that SU. The minimum value for GA is 0 (no guests that month). GN will be greater than or equal to GA. If $\mathsf{GN} = \mathsf{GA}$ this is equivalent to saying that no-one stays more than one night.

2.2.6 Synthetic Data

Statistics New Zealand has provided us with two sets of synthetic data that resemble the responses and non-responses for December 2008 and January 2009 (i.e. prior to changes in the survey in September 2009). The November 2008 data has also been synthesised but all non-responses have been replaced with values. This allows us to model the effect of imputation over time beginning from an agreed start point that is free of missing data, as we intend our approach to use the data stuctures of the responses from the beginning of the AOS. Data from the first month of the AOS will have to be completed by imputation but, as eventually more than 100 months of responses will be available, when looking at current data the influence of the initial imputations on the current imputation will be negligible. Thus we focus here on imputing the missing data for the next two months: December 2008 and January 2009. These two files are named `SYN0812.txt` and `SYN0901.txt`, respectively. We encode the missing data with -1 in order to use fixed dimensional arrays for efficient matrix processing in `Matlab` or `NumPy` or a similar numerical computing environment.

The format of the data is given in Table 1. The data is an $n \times 12$ matrix and we represent this in a computer by an $n \times 12$ array data-structure. The number of accommodation providers for a given month are denoted by n and this corresponds to the number of rows. The first and second columns are the year and month of the survey respectively. The third column give the unique identity number or ID that is assigned to each accommodation provider at birth and allows us to connect the responses from any particular accommodation provider over time.

For a concrete appreciation of the encoding of our data, consider the following three possible combinations of missing data for the sub-array made of the last four columns, $(\mathsf{GN}, \mathsf{GN_D}, \mathsf{GN_I}, \mathsf{GN_U})$, where $*$ stands for a response:

1. $(-1, -1, -1, -1)$ – no response to all four questions
2. $(*, -1, -1, -1)$ – response to GN, but not to the decomposition
3. $(*, *, *, *)$ – response to all 4 questions

Year	Month	ID	C_G	C_I	SU	SN	GA	GN	GN_D	GN_I	GN_U
2008	11	1	16	11	182	1790	722	1790	1790	0	0
2008	11	2	34	2	86	71	111	142	69	73	0
2008	11	3	33	15	6	23	36	57	20	37	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2008	11	4016	31	11	45	540	285	540	138	402	0

Table 1: Format of the 3961×12 array of the synthetic AOS survey data for November 2008. Here $n = 3961$. The missing values for other subsequent months are encoded as -1.

Thus, there should be no cases of types 1. or 2. in `SYN0811.txt` as we assume the absence of any non-response in that dataset. In the rare event that some of the data is still inconsistent in the file `SYN0811.txt` that was provided by Statistics New Zealand, we simply ignore them in this study.

3 Current Imputation Methodology

3.1 Introduction

The basic idea behind any imputation methodology is to assume distributional homogeneity within sub-populations and impute a missing value for a nonrespondent from the data provided by the respondents in the same sub-population as the nonrespondent. This basic idea is justified by the assumption that the nonrespondents would provide similar responses as the respondents of the same sub-population. By definition, it is difficult to confirm this assumption as information on the nonrespondents is required to test this hypothesis. For the purposes of this project we assume that the responses within each sub-population are randomly distributed according to the distribution for this sub-population, where each sub-population is allowed to have a distinct distribution. This is equivalent to assuming that non-response is Missing At Random (MAR) [3].

In this section we briefly describe the current imputation methodology of Statistics New Zealand, namely, the point estimates of the missing values. We provide extensions of the estimates of the missing values to confidence sets and confidence intervals in order to obtain the variance in the imputed estimates and some AOS statistics in § 4.

3.2 Imputation Cells and Homogeneous Sub-Populations

The type of guest is expected to vary across accommodation providers (e.g. people travelling on business in hotels, families in campgrounds), and thus the accommodation occupancy patterns across accommodation providers will vary. Thus to impute for a non-response from a hotel, one should use the responses from other hotels rather than use the responses from motels, backpackers or camping grounds. Thus, a greater homogeneity is expected among accommodation providers of the same accommodation type and this homogeneity should be exploited during imputation. It also seems likely that the general accommodation occupancy patterns of accommodation providers who are geographically close to one another are more similar than a random accommodation provider in New Zealand. For example, Queenstown is expected to have a different accommodation pattern than Auckland.

This implies that we want the imputation to be based on as homogenous a population of respondents and nonrespondents as possible. The total population of accommodation providers in New Zealand cannot be regarded as homogenous so we divide the population into a set of homogeneous sub-populations, which we term imputation cells. The imputation cells are based on a combination of accommodation type and some spatial classification defined using a discrimination analysis technique. It is possible that a particular accommodation provider in an imputation cell has anomalous responses which are not typical of other accommodation providers in that imputation cell (e.g. it is closed for that month). Statistics New Zealand has methods to identify these anomalous respondents and minimise their effect on the imputed values, but as we do not have this information we have used all the responses from each imputation cell.

For our analysis, we use the imputation cells created by Statistics New Zealand. These cells are based on a combination of the 24 Regional Tourism Organisations (RTO) and the 5 accommodation types (there are modifications after August 2009). This implies a possible 120 imputation cells, but some of these 120 possible imputation cells will contain very small numbers of accommodation providers, so Statistics New Zealand has combined some of these possible imputation cells to ensure a reasonable number of expected responses. As some of the neighbouring RTO will have similar accommodation patterns, some merging of the possible cells has been done to create the imputation cells actually used by Statistics New Zealand.

This minimum imputation cell size constraint is mainly to ameliorate the effect of an anomalous response on imputation outputs by having a large enough sample size of responses. In other words, if the number of valid responses for a given month is too small within a sub-population corresponding to a given imputation cell then imputation cells with relatively homogeneous sub-populations are combined to overcome the lack of information. However, this is at the expense of decreasing the homogeneity of the newly combined sub-population used for imputation. With our approach this would become less of an issue over time as we would draw strength from the past responses, as opposed to relying solely on the current responses.

Currently, AOS has 45 imputation cells for imputation of GA, GN and SN. The same imputation cells are used for these 3 variables as it can be shown that these variables are highly correlated. That is, the sub-populations are homogeneous for all three variables. There are 52 imputation cells for $\mathsf{GN_D}$, $\mathsf{GN_I}$ and $\mathsf{GN_U}$. Having created the distribution of imputation-based estimates that are required to create the confidence sets of the missing values or non-responses as well as the confidence intervals of various AOS statistics, it is possible that they could be used to identify those responses that have the greatest effect on the outcomes of the imputation. Further research is required to investigate the feasibility of this approach.

3.3 Estimators of Missing Data in AOS

There are several imputation methods used by Statistics New Zealand. The one used to impute for any particular accommodation provider for any variable depends on at least the following two questions.

- Has the unit responded previously?
- Are we imputing an integral value or a percentage or a probability mass function?

Next we visit some specific point estimates that are currently used by Statistics New Zealand and give our extension for confidence sets for the missing values and AOS statistics.

3.3.1 Mean Ratio of an Imputation Cell

Each month we can effectively assume that we know SU for every accommodation provider, even if they have not responded in the current month, as SU is collected when an accommodation provider is first surveyed. SU does not change in the data unless a different value is given by the accommodation provider in their monthly questionnaire. Based on available data, a change in the value for SU is very rare. Thus, if there is no response for SU, Statistics New Zealand can impute the SU value of the AP from its last response since it is known to change very rarely.

Clearly there is a relationship between many of the other variables and SU, and therefore it is imputed first. For any particular accommodation provider Statistics New Zealand can calculate the monthly average of a variable per SU (e.g. SN per SU) for that accommodation provider. If Statistics New Zealand calculates this average over all the respondents in an imputation cell it applies this average to the SU for any nonrespondent to impute their value for the missing variable for that month. For example, if those in an imputation cell that responded have a total SN of 3000, and their total SU is 300, then that imputation cell's average is $\mathsf{SN}/\mathsf{SU} = 10$. Therefore for a nonrespondent with an SU of 12, their imputed SN value will be $10 \times 12 = 120$. This method is used to impute values for SN, GN and GA where the accommodation provider has no previous value (i.e. a birth) or for those accommodation providers where the previous month's value was imputed (i.e. did not respond in previous month).

A similar method is used for $\mathsf{GN_D}$, $\mathsf{GN_I}$ and $\mathsf{GN_U}$, but as $\mathsf{GN_D}$, $\mathsf{GN_I}$ and $\mathsf{GN_U}$ is effectively the partition of GN into three categories, and thus there is an additivity constraint, these are imputed as a whole distribution, rather than variable by variable. We use the respondents to calculate the 'average distribution' of GN within an imputation cell. For example, in Table 2 for $\mathsf{GN_D}$ we calculate $(10 + 30 + 3)/(17 + 90 + 18) = 43/125 = 0.344$. This proportion is applied to the total GN of the nonrespondent. Thus, the point estimate of $\mathsf{GN_D}$ for the missing response in Table 2 would be $0.344 \times 40 = 13.76$. Note that non-integer $\mathsf{GN_D}$, $\mathsf{GN_I}$ and $\mathsf{GN_U}$ are acceptable for imputed values, under the current imputation scheme, as the aggregated outputs from AOS are rounded. In other words, the overall proportion of $\mathsf{GN_D}$, $\mathsf{GN_I}$ and $\mathsf{GN_U}$ over the respondents in an imputation cell are applied to the value of GN, either from a response, or a previously imputed value for GN. Since imputation for guest nights happens first, it does not matter that many nonrespondents to origin of guests are also nonrespondents to guest nights.

AP	$\mathsf{GN_D}$	$\mathsf{GN_I}$	$\mathsf{GN_U}$	GN
1	10	2	5	17
2	30	50	10	90
3	3	10	5	18
4	-1	-1	-1	40

Table 2: Imputing a point estimate of $(\mathsf{GN_D}, \mathsf{GN_I}, \mathsf{GN_U})$ given GN for the nonrespondent 4 based on the respondents 1, 2 and 3 from the same imputation cell.

3.3.2 Weighted Historical

If an accommodation provider has given a response for SN, GN and GA for the previous month a different method of imputation is used. A forward movement factor (FMF) is calculated from the

respondents in the imputation cell that have responded for both months.

$$\mathsf{FMF} = \frac{\sum_{i=1}^{r} x_{it}}{\sum_{i=1}^{r} x_{i(t-1)}} \tag{1}$$

The previous unimputed (i.e. actual) value for the nonrespondent is multiplied by the FMF. For example, if the FMF is 1.1, a 10% rise for the month, and the previous month's value for the AP was 50, then the imputed value will be 55.

3.3.3 Beyond Point Estimation

The imputation methodologies currently used by Statistics New Zealand only produce point estimates for the non-responses without any formal accounting of the inherent uncertainty in the imputation procedure. Thus, we need some way of introducing the inherent uncertainty, caused by the response/nonresponse mechanism, into the estimation process. We use non-parametric bootstraps to directly obtain samples from the distribution of responses for each imputation cell. In our new approach, we propagate all uncertainties formally. For example, when GN is imputed prior to imputing the origin of guests, i.e., $\mathsf{GN_D}$, $\mathsf{GN_I}$ and $\mathsf{GN_U}$, we need to formally account for the uncertainty in the imputation of GN in our subsequent conditional imputation of $\mathsf{GN_D}$, $\mathsf{GN_I}$ and $\mathsf{GN_U}$. This is explained in the next section.

4 Confidence Sets for AOS Statistics of Missing Data

4.1 Introduction

The previous section briefly described the current imputation methodologies of Statistics New Zealand. All these methodologies only provided a point estimate of the missing data and did not prescribe a confidence measure of the the point estimates. In this section, we describe and provide proof-of-concept implementations of methods that compute confidence sets and confidence intervals that contain the missing data with a high probability. This gives the desired variance in the imputed estimates and some AOS statistics that depend on them.

Our basic methodology here involves the use of the non-parametric bootstrap algorithm within each imputation cell, based on the assumption that APs within an imputation cell respond homogeneously according to the same distribution. We only bootstrap from the respondents to impute the missing values of non-respondents as described in § 6.1. Specifically, the blocked-bootstrap method for the problem can be summarized as follows: For each sub-population or imputation cell $i \in \{1, 2, \ldots, k\}$, we assume:

$$X_1^{(i)}, X_2^{(i)}, \ldots, X_{n_i}^{(i)}, X_{n_i+1}^{(i)}, X_{n_i+2}^{(i)}, \ldots X_{n_i+m_i}^{(i)} \overset{i.i.d}{\sim} F^{(i)}$$

where the first n_i are not missing while the remaining m_i are missing. We simply use $\widehat{F_{n_i}^{(i)}}$, the empirical distribution function (EDF) of the i-th imputation cell or sub-population from the non-missing data to impute missing data:

$$X_{n_i+1}^{(i)}, X_{n_i+2}^{(i)}, \ldots X_{n_i+m_i}^{(i)} \overset{i.i.d}{\sim} \widehat{F_{n_i}^{(i)}}$$

This is our sub-population-specific non-parametric blocked bootstrap methodology that is statistically consistent as $n_i \to \infty$.

We need to first look at the imputation of SU because, as noted above, the other variables are imputed on the basis of their relationship to SU, and thus SU must be imputed first.

4.2 Estimating SU

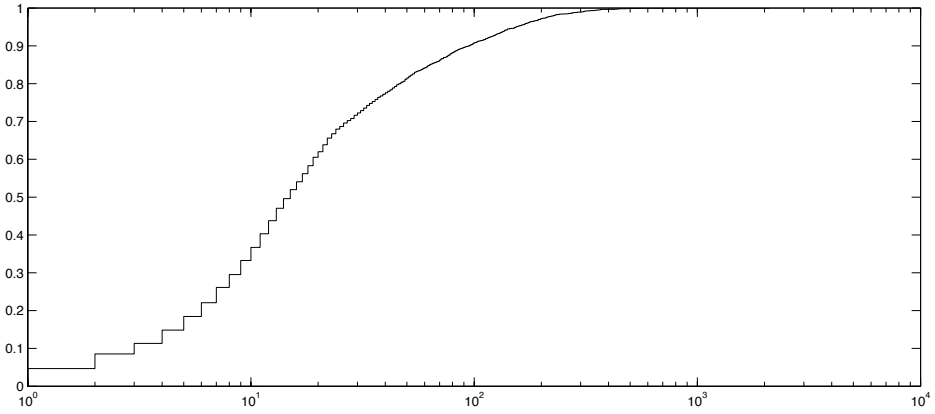

Figure 1: The empirical distribution function of SU imputed from historical data.

Figure 1 shows the empirical distribution of the SUs for the month of December 2008 based on the values in November for each AP. As can be seen there is a great range of values for SU. We have used a log scale for the x-axis to emphasise the differences within orders of magnitude. Such a broad distribution generally leads to large uncertainty in imputed values unless other information is used to divide the population into dissimilar homogeneous sub-populations. For example, there are a large number of SUs with a value of 1 ($\sim 5\%$). Most of these SUs will be hosted accommodation since most hosted accommodation providers are small. Further disaggregation of the population could be attempted, but there will still be a broad range of responses for some of the sub-populations.

However as noted previously, the number of stay units for any particular AP is expected to change rarely, thus the imputation of SU has been designed as a simple look-up problem. That is, when an AP does not provide a response for SU in a given month, we simply look at the past records of this AP for the most recent response for SU. Such a response will exist since Statistics New Zealand collects SU at least once in the history of the AP. In other words, none of the responses from other AP are used for imputation. As a result, though the imputed response to SU is used for imputation of the other variables and its contribution to uncertainty of the estimates for the other variables may not be zero for all AP, overall it is highly likely to have a negligible contribution to overall imputation variance.

4.3 Estimates of SN, GA and GN

The problem of imputing SN, GA and GN for a given AP with SU many units is less trivial due to the fact that the variables in the response vector (SU, SN, GA, GN) are inter-dependent. In other words, not only must the imputed values for each variable be consistent with the distribution of the responses that have been given by AP in the imputation cell, but also all the imputed values for the variables for a given AP provider must be internally consistent. This point is important because typically an AP either answers all questions or does not respond to several questions. We model this 4-variable response vector as a realisation of some underlying sub-population specific distribution over the four dependent variables.

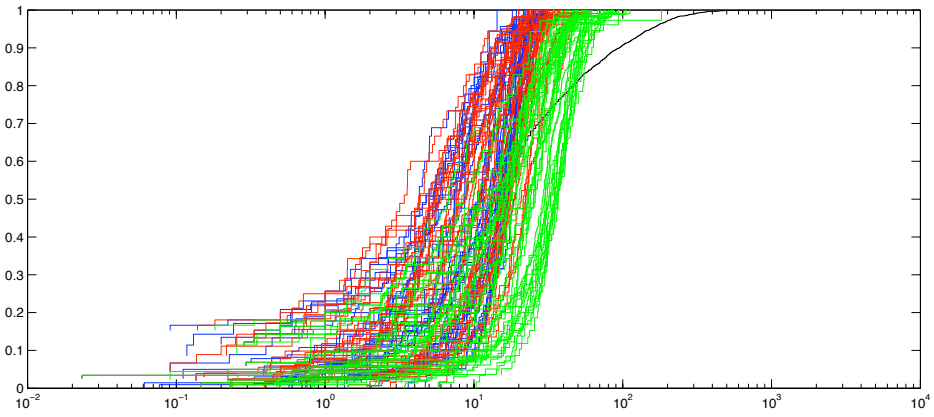

Figure 2: The empirical distribution function of SN/SU (blue), GA/SU (red) and GN/SU (green) from the responses in each of the 45 imputation cells or sub-populations for December 2009 AOS survey data.

SU, SN, GA and GN are highly correlated to SU so it is their ratio to SU that is of interest for imputation as much of the variation between total SU, SN, and GN between sub-populations will merely arise from the differences in the number of SU in that sub-population. Using the ratios can be considered equivalent to standardisation of the sub-populations, and thus the imputation cells.

As our approach is to use the empirical distribution of these ratios for any particular imputation cell for further imputation, it is important to examine the empirical distribution functions (EDFs) of these ratios for all respondents in each imputation cell. Figure 2 shows the empirical distribution functions for all the ratios as well as SU. We have once again used a log scale for the x-axis to emphasise the differences within orders of magnitude. Such a scale allows for a better visualisation of the data. As expected, since the number of guests in a SU has a wide range; SN/SU EDFs are generally to the left, GN/SU EDFs are to the right and GA/SU EDFs are in between. SU EDF has by far the widest range. Figure 2 is the summary of the full data that is used for subsequent imputation variance estimation of AOS statistics of interest to Statistics New Zealand.

While the EDF for each of the variables SN, GA and GN are similar in shape it can be readily seen they differ in location, thus showing that the imputation cells are heterogeneous in terms of

their ratios to SU. Statistics New Zealand is primarily interested in the totals of the variables for various populations of interest (e.g. national, regional) which are related to the means by the total number of SU in the population of interest. We impute values by drawing from the distribution and thus "fill in" for the nonresponse, but are not interested accuracy of any single imputation. Rather it is the statistics from the data completed by imputation that is of interest. By performing a number of non-parametric bootstraps we end up with a number of realisations of what the AOS data could be like if all APs responded.

In Figure 3 we have plotted the sub-population-specific non-parametric bootstrap of the empirical distributions of SN (blue), GA (red) and GN (green) for December 2008. This is obtained by imputing responses for each item non-response using our blocked bootstrap methodology. As can be seen, these bootstrapped EDFs are almost vertical. This indicates that the imputation is performing well in terms of minimising imputation variance. As we are plotting total SN, GA and GN the spread of location on the x-axis is mainly a result of the differences in SU between imputation cells.

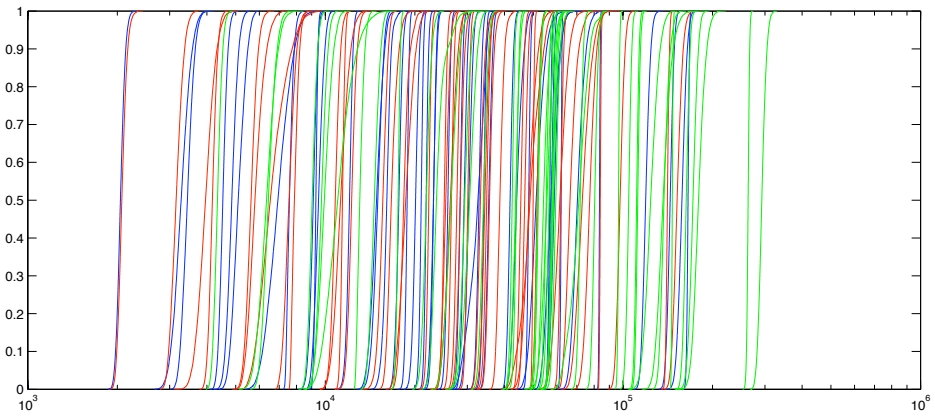

Figure 3: The sub-population-specific non-parametric bootstrap of the empirical distributions of SN (blue), GA (red) and GN (green) for December 2008.

In Figure 4, we show the empirical distributions of the total SN (blue), GA (red) and GN (green) along with those of the sub-populations' that was summed to obtain the total for December 2008. It is of interest that most of the EDF are parallel, which suggest the imputation variance within most imputation cells are approximately the same. However there are a few which are much less vertical and it is clear that the imputation variance in these imputation cells are considerably higher than most. Whether this arises from a higher non-response, an anomalous value that is increasing the initial variance of the respondents or sub-population-specific inadequacies of the imputation technique would require further investigation.

Figure 5 shows the EDF of national total SN, GA and GN along with all the EDFs used in their imputation. As noted earlier, it is not the accuracy of an particular imputed value for an AP that is of interest, but rather the realised total resulting from the responses and imputations in an imputation cell. As such it can be seen the EDF of the realised national totals from the bootstraps

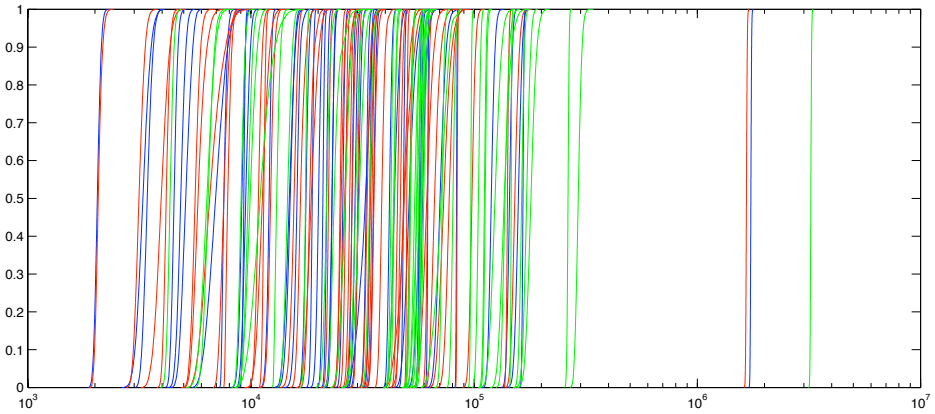

Figure 4: The non-parametric bootstrap of the empirical distributions of the total SN (blue), GA (red) and GN (green) along with those of the sub-populations' that was summed to obtain the total for December 2008.

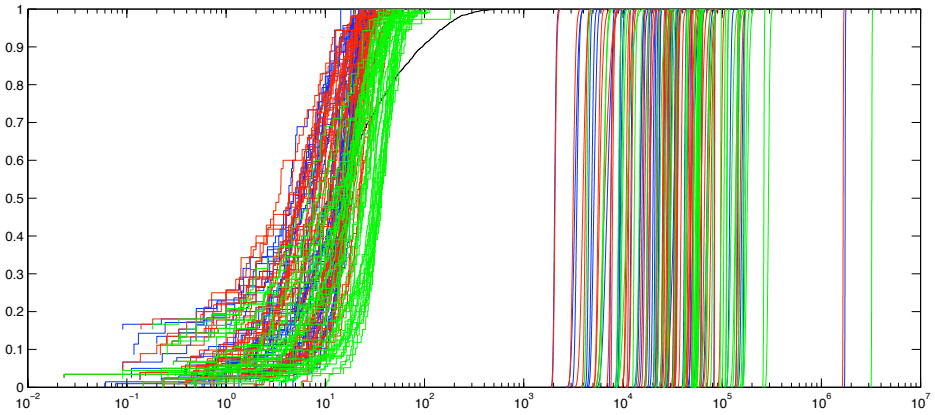

Figure 5: The entire bootstrap process to get the empirical distributions of total SN, GA and GN for December 2008.

is considerably smoother than the EDF from which they have been drawn.

4.4 Bootstrap-based Variance Estimates of the total SN, GA and GN for December 2008

Using the bootstrap method outlined above, we can obtain the three basic AOS statistics of interest with confidence statements related to the imputation methodology. We can thus say that the true value of the total GN for December 2008 (after accounting for missing data) lies in the interval $[3,184,482\ ,\ 3,264,655]$ and our single best guess — point estimate form the 0.5-th quantile — for this month's GN is $3,223,676$. This compares well with the *mean ratio* point estimate produced by the methodology of § 3.3.1. While there is no sample error in the AOS, as it is a census of AP, the uncertainty resulting from imputation for nonresponse appears non-negligible. Other statistics at a national level (e.g. SN and GA) can also be easily produced by our method and are summarised in Table 3.

AOS Stats.	Mon'YY	0.025-th quantile	0.5-th quantile	0.975-th quantile	Mean Ratio
SN	Dec'08	1,717,345	1,736,786	1,757,482	1,736,955
	Jan'09	2,187,283	2,213,689	2,241,581	2,214,119
GA	Dec'08	1,637,324	1,659,149	1,682,624	1,659,325
	Jan'09	2,023,080	2,053,760	2,088,124	2,054,361
GN	Dec'08	3,184,482	3,223,676	3,264,655	3,224,020
	Jan'09	4,366,033	4,430,805	4,508,966	4,432,837

Table 3: Point estimate based on the median or 0.5-th quantile as well as 95% confidence intervals based on the 0.025-th quantile and the 0.975-th quantile of the non-parametrically bootstrapped distribution of the estimator of the total SN, GA and GN for December 2008 and January 2009.

5 Discussion

We have extended the current Statistics New Zealand imputation methodology for the Accommodation Occupancy Survey (AOS) to not only provide point estimates, but also to provide confidence intervals that account for the uncertainty in the imputation process. From the assumption that the non-respondents are similar to the respondents within a sub-population, we can use the sub-population-specific responses to infer the response distribution. The uncertainty due to imputation appears similar in magnitude to the sample error for many of Statistics New Zealand's other surveys. This provides quantitative information to users of the AOS outputs on the quality of the information, as Statistics New Zealand in the *Technical Notes* that accompany an AOS release does comment that there is uncertainty arising from non-response while not currently quantifying this. While we have only calculated the imputation variance for the total it is straight-forward to extend our approach to sub-populations. By doing this, Statistics New Zealand would find which sub-populations have the highest imputation variance. As imputation variance depends on the distribution of the responses used for imputation as well as the nonresponse rate, Statistics New Zealand could better target improvements in its response rates to those areas of interest with the

highest imputation variance, rather than by simply focussing on those with the lowest response rates.

We also see that our approach is flexible enough to extend the current imputation methods to better utilise *all* the information collected in the AOS since its inception. Using more past information could protect the imputations from an existing problem induced by heterogeneity arising within an imputation cell for a particular month (e.g. an accommodation provider being closed for the month and thus its responses being zero). In fact, using our approach, it would be easier to identify such changes in the characteristics within any imputation cell. However, instead it could allow the imputation cells to be dynamically developed from the currently most homogenous subpopulations defined by similarity or dissimilarity measures over any pair of APs in New Zealand. It seems to us that the spatial nature of the AOS in particular has not be fully utilised. Given that the location of each AP is known to some degree of accuracy — at least to a city block in urban areas — it would appear that this knowledge should be used more effectively for localised, targeted and geographically refined tourism statistics. Such, geographically fine resolutions of GN for instance can directly shed more light on effective management decisions in the tourism sector. With our methodology it is also possible to update imputation cells over time, though more work is required to see if there is enough change over time to merit updating imputation cells more frequently than currently done by Statistics New Zealand.

To formally approach such time-dependent and spatially-dependent statistics one has to use non-parametric spatio-temporal blocked bootstrap techniques in conjunction with interactive visualisation of non-parametric moving density estimates of the basic AOS statistics. Human visualisation of appropriate AOS statistics that is depicted spatially over the two islands on the basis of the geographic location of each provider as well as temporally through the months may shed light that is not captured by simple summary statistics and numerical tables. This approach will use more information in the surveys and therefore lead to significantly better managerial and administrative decisions. By the use of appropriate non-parametric techniques to impute missing data one can make confidence-qualified estimation and prediction of spatio-temporal flows of accommodation occupancy measures and statistics. Such a detailed data-centered nonparametric approach is beyond the scope of this study but is a feasible topic for future work on the AOS.

6 Appendix

6.1 Non-parametric Bootstrap of the responses for Confidence Sets

Let $T_n := T_n((X_1, X_2, \ldots, X_n))$ be a statistic, i.e. any function of the data $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} F^*$. Suppose we want to know its variance $\mathbf{V}_{F^*}(T_n)$, which clearly depends on the fixed and possibly unknown DF F^*.

If our statistic T_n is one with an analytically unknown variance, then we can use the bootstrap to estimate it. The bootstrap idea has the following two basic steps:

Step 1: Estimate $\mathbf{V}_{F^*}(T_n)$ with $\mathbf{V}_{\widehat{F}_n}(T_n)$.

Step 2: Approximate $\mathbf{V}_{\widehat{F}_n}(T_n)$ using simulated data from the "Bootstrap World."

For example, if $T_n = \overline{X}_n$, in Step 1, $\mathbf{V}_{\widehat{F}_n}(T_n) = s_n^2/n$, where $s_n^2 = n^{-1}\sum_{i=1}^n (x_i - \overline{x}_n)$ is the sample variance and $\overline{x}_n$ is the sample mean. In this case, Step 1 is enough. However, when the statistic T_n

is more complicated (e.g. $T_n = \widetilde{X}_n = F^{[-1]}(0.5)$), the sample median, then we may not be able to find a simple expression for $\mathbf{V}_{\widehat{F}_n}(T_n)$ and may need Step 2 of the bootstrap.

$$\begin{array}{rrcl} \text{Real World Data come from} & F^* & \implies X_1, X_2, \ldots, X_n & \implies T_n((X_1, X_2, \ldots, X_n)) = t_n \\ \text{Bootstrap World Data come from} & \widehat{F}_n & \implies X_1^\bullet, X_2^\bullet, \ldots, X_n^\bullet & \implies T_n((X_1^\bullet, X_2^\bullet, \ldots, X_n^\bullet)) = t_n^\bullet \end{array}$$

Observe that drawing an observation from the ECDF $\widehat{F}_n$ is equivalent to drawing one point at random from the original data (think of the indices $[n] := \{1, 2, \ldots, n\}$ of the original data $X_1, X_2, \ldots, X_n$ being drawn according to the equi-probable de Moivre$(1/n, 1/n, \ldots, 1/n)$ RV on $[n]$). Thus, to simulate $X_1^\bullet, X_2^\bullet, \ldots, X_n^\bullet$ from $\widehat{F}_n$, it is enough to drawn n observations with replacement from $X_1, X_2, \ldots, X_n$.

In summary, the algorithm for Bootstrap Variance Estimation is:

Step 1: Draw $X_1^\bullet, X_2^\bullet, \ldots, X_n^\bullet \sim \widehat{F}_n$

Step 2: Compute $t_n^\bullet = T_n((X_1^\bullet, X_2^\bullet, \ldots, X_n^\bullet))$

Step 3: Repeat Step 1 and Step 2 B times, for some large B, say $B > 1000$, to get $t_{n,1}^\bullet, t_{n,2}^\bullet, \ldots, t_{n,B}^\bullet$

Step 4: Several ways of estimating the bootstrap confidence intervals are possible:

(a) The $1 - \alpha$ percentile-based bootstrap confidence interval is:

$$C_n = [\widehat{G^\bullet}_n^{-1}(\alpha/2), \widehat{G^\bullet}_n^{-1}(1 - \alpha/2)],$$

where $\widehat{G^\bullet}_n$ is the empirical DF of the bootstrapped $t_{n,1}^\bullet, t_{n,2}^\bullet, \ldots, t_{n,B}^\bullet$ and $\widehat{G^\bullet}_n^{-1}(q)$ is the q^{th} sample quantile of $t_{n,1}^\bullet, t_{n,2}^\bullet, \ldots, t_{n,B}^\bullet$.

6.2 Mean Ratio based Point Estimates for SN, GA, and GN

```
%% This Matlab script obtains Mean-Ratio Point Estimates For missing data
% point estimator of sub-population specific (C_I-specific) monthly totals of
% SU, SN, GA and GN
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% raw data from 2 months -- 0811 is "complete" survey 0812 has missing
% values as -1 and Columns are:
% Year        Month ID C_G C_I SU        SN GA GN GN_D GN_I GN_U
A=dlmread('SYN0811.M.txt'); % has the "complete" data
%B=dlmread('SYN0812.M.txt'); % the first month with missing data
B=dlmread('SYN0901.M.txt'); % the next month with missing data

%%some preprocessing of errors in synthetic data
%  fixing su=0 to su=1 in col 6 of 'SYN0811.M.txt'
ZeroSURows=find(A(:,6) == 0);
A(ZeroSURows,6)=ones(length(ZeroSURows),1);
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Impute missing su first by look up of previous month
% get the row numbers of missing data in cols 6, i.e., missing SU
Missing6Rows = find(B(:,6) <= 0);
Missing6IDs = B(Missing6Rows,3);
%B(Missing6Rows,6:9)
if(length(Missing6Rows)==length(Missing6IDs))
    LenMissing6=length(Missing6IDs);
else error('lenght(Missing6Rows)~=length(Missing6IDs)');
end
```

```
for imp6 = 1:LenMissing6
    %Missing6IDs(imp6)
    RowA = find(A(:,3) == Missing6IDs(imp6));
    assert(A(RowA,6) >= 1);
    B(Missing6Rows(imp6),6) = A(RowA,6);
end
%B(Missing6Rows,6:9);
% check that all su values are > 0 in imputed B
assert(length(find(B(:,6) < 1))==0);
% plotting EDF of SUs
% semilogx(0,0); hold on;[x1 y1]=ECDF(B(:,6), 0, 0,1);stairs(x1,y1,'color','k');
% end of imputing the missing su values from previous complete data
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Mean-Ratio Point Estimate of missing data in the next three columns: 7,8,9
%  corresponding to SN,GA,GN one at a time, i.e., marginally

% get range of imputation cell numbers
disp('min and max imputation cells in col 4 (assume contiguous numbers)');
%CellIDCol=4; % For Geographic Imputation Cell Numbers in column 4
CellIDCol=5; % For Imputation Cell Numbers in column 5
FirstImpCell = min(B(:,CellIDCol));
LastImpCell = max(B(:,CellIDCol));

%% make arrays for Monthly Total of MissingCol = 7, 8 or 9
Colors=['b','r','g'];% blue,red,green for col 7,8,9
for MissingCol=7:9
    TotalKnownImpCells = zeros(1,LastImpCell-FirstImpCell+1);
    TotalMeanRatioImpCells = zeros(1,LastImpCell-FirstImpCell+1);
    % loop over imputation cells contiguously from first to last start at 1
    % to get the SU-averaged mean-ratio measure
    for ImpCell = FirstImpCell:LastImpCell
        %ImpCell; %used as array index: FirstImpCell=1,2,...,LastImpCell !!!
        %
        %filled imp cell specific indices
        ImpCellIndicesF = find(B(:,CellIDCol)==ImpCell & B(:,MissingCol)>=0);
        assert(min(B(ImpCellIndicesF,6))>0); % check that each filled SU>0
        %sum of filled imp cell specific indices over MissingCol
        TotalKnownImpCells(1,ImpCell)=sum(B(ImpCellIndicesF,MissingCol));
        EmpiricalAvgBySU = B(ImpCellIndicesF,MissingCol)./ B(ImpCellIndicesF,6);
        % %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
        %missing imp cell specific indices
        ImpCellIndicesM = find(B(:,CellIDCol)==ImpCell & B(:,MissingCol)<0);
        SUForMissing=B(ImpCellIndicesM,6);
        NumMissing=length(ImpCellIndicesM);
        assert(NumMissing==length(SUForMissing));
        MeanRatioVector = ones(1,NumMissing) * mean(EmpiricalAvgBySU);
        NumFilled=length(ImpCellIndicesF);
        assert(NumFilled==length(EmpiricalAvgBySU));
        % mean ratio imputation step
        TotalMeanRatioImpCells(1,ImpCell)=...
            (MeanRatioVector * SUForMissing)+TotalKnownImpCells(1,ImpCell);
    end

    %TotalMeanRatioImpCells

    %begin stem plotting
    subplot(1,2,1)
    stem(TotalMeanRatioImpCells,'fill','--',...
        'MarkerFaceColor',Colors(MissingCol-6));
    hold on;
    format('long');
    TotalMeanRatioAllCells = sum(TotalMeanRatioImpCells)
    subplot(1,2,2)
    stem(TotalMeanRatioAllCells,'fill','--',...
        'MarkerFaceColor',Colors(MissingCol-6));
    hold on;
    % end stem plots
end
```

6.3 Bootstrapping Variance Estimates for SN, GA, and GN

```
%% This Matlab script nonparametrically obtains samples from imputation-based
% estimator of sub-population specific (C_I-specific) monthly totals of
% SU, SN, GA and GN
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% raw data from 2 months -- 0811 is "complete" survey 0812 has missing
% values as -1 and Columns are:
% Year        Month ID C_G C_I SU        SN GA GN GN_D GN_I GN_U
A=dlmread('SYN0811.M.txt'); % has the "complete" data
%B=dlmread('SYN0812.M.txt'); % the first month with missing data
B=dlmread('SYN0901.M.txt'); % the next month with missing data

%%some preprocessing of errors in synthetic data
%  fixing su=0 to su=1 in col 6 of 'SYN0811.M.txt'
ZeroSURows=find(A(:,6) == 0);
A(ZeroSURows,6)=ones(length(ZeroSURows),1);
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Impute missing su first by look up of previous month
% get the row numbers of missing data in cols 6, i.e., missing SU
Missing6Rows = find(B(:,6) <= 0);
Missing6IDs = B(Missing6Rows,3);
%B(Missing6Rows,6:9)
if(length(Missing6Rows)==length(Missing6IDs))
    LenMissing6=length(Missing6IDs);
else error('lenght(Missing6Rows)~=length(Missing6IDs)');
end
for imp6 = 1:LenMissing6
    %Missing6IDs(imp6)
    RowA = find(A(:,3) == Missing6IDs(imp6));
    assert(A(RowA,6) >= 1);
    B(Missing6Rows(imp6),6) = A(RowA,6);
end
%B(Missing6Rows,6:9);
assert(length(find(B(:,6) < 1))==0); % check that all su values are > 0 in imputed B
% plotting EDF of SUs
 semilogx(0,0); hold on;[x1 y1]=ECDF(B(:,6), 0, 0,1);stairs(x1,y1,'color','k');
% end of imputing the missing su values from previous complete data
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Imputation of missing data in the next three columns: 7,8,9
%  corresponding to SN,GA,GN one at a time, i.e., marginally
%  get the row numbers of missing data in cols i=7,8,9
MissingCol=9;
MissingRows = find(B(:,MissingCol) < 0);
FilledRows  = find(B(:,MissingCol) >= 0);
MissingIDs = B(MissingRows,3);
%B(MissingRows,7:9)

% for checking
%disp('all rows for missing data in cols '); disp(MissingCol);
%B(MissingRows,MissingCol)
%disp('all rows for filled data in cols '); disp(MissingCol);
%B(FilledRows,MissingCol)

% get range of imputation cell numbers
disp('min and max imputation cells in col 4 (assume contiguous numbers)');
%CellIDCol=4; % For Geographic Imputation Cell Numbers in column 4
CellIDCol=5; % For Imputation Cell Numbers in column 5
FirstImpCell = min(B(:,CellIDCol));
LastImpCell = max(B(:,CellIDCol));

%% make arrays for Monthly Total of MissingCol = 7, 8 or 9
Colors=['b','r','g'];% blue,red,green for col 7,8,9
for MissingCol=7:9
    TotalKnownImpCells = zeros(1,LastImpCell-FirstImpCell+1);
    Bootstraps=10000;
    TotalBootImpCells = zeros(Bootstraps,LastImpCell-FirstImpCell+1);
    %semilogx(1000,0); hold on;%plotting
```

```
    % loop over imputation cells contiguously from first to last start at 1
    % to get the empirical distributions of SU-averaged measure
    for ImpCell = FirstImpCell:LastImpCell
        %ImpCell; %used as array index: FirstImpCell=1,2,...,LastImpCell !!!
        %
        %filled imp cell specific indices
        ImpCellIndicesF = find(B(:,CellIDCol)==ImpCell & B(:,MissingCol)>=0);
        assert(min(B(ImpCellIndicesF,6))>0); % check that each filled SU>0
        %sum of filled imp cell specific indices over MissingCol
        TotalKnownImpCells(1,ImpCell)=sum(B(ImpCellIndicesF,MissingCol));
        EmpiricalAvgBySU = B(ImpCellIndicesF,MissingCol)./ B(ImpCellIndicesF,6);
        % plotting %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
         [x1 y1]=ECDF(EmpiricalAvgBySU, 0, 0,1);
         assert(min(y1>=0)); assert(min(x1>=0))
         stairs(x1,y1,'color',Colors(MissingCol-6)); hold on;
        % %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
        %missing imp cell specific indices
        ImpCellIndicesM = find(B(:,CellIDCol)==ImpCell & B(:,MissingCol)<0);
        SUForMissing=B(ImpCellIndicesM,6);
        NumMissing=length(ImpCellIndicesM);
        assert(NumMissing==length(SUForMissing));
        NumFilled=length(ImpCellIndicesF);
        assert(NumFilled==length(EmpiricalAvgBySU));
        % bootstrap step

        TotalBootImpCells(:,ImpCell)=...
            (EmpiricalAvgBySU(ceil(rand(Bootstraps,NumMissing)*NumFilled))...
            * SUForMissing)+TotalKnownImpCells(1,ImpCell);
    end

    %semilogx(2.0e+06,0);
    hold on;%plotting
    for ImpCell = FirstImpCell:LastImpCell
        %plotting
        BootDistImpCell=TotalBootImpCells(:,ImpCell);
        [x1 y1]=ECDF(BootDistImpCell, 0, 0,1);
        stairs(x1,y1,'color',Colors(MissingCol-6));
        hold on;
    end
    TotalBootAllCells = sum(TotalBootImpCells');
    [x1 y1]=ECDF(TotalBootAllCells, 0, 0,1);
    stairs(x1,y1,'color',Colors(MissingCol-6));
    SortedTotalBootAllCells = sort(TotalBootAllCells);
    format('long');
    ConfInt95BootPercentile = ...
        [qthSampleQuantile(0.025,SortedTotalBootAllCells),...
        qthSampleQuantile(0.5,SortedTotalBootAllCells),...
        qthSampleQuantile(0.975,SortedTotalBootAllCells)]
end
```

6.4 Auxillary code and functions for § 6.3

6.4.1 Empirical Distribution Function

```
function [x1 y1] = ECDF(x, PlotFlag, LoxD, HixD)
% return the x1 and y1 values of empirical CDF
% based on samples in array x of RV X
% plot empirical CDF if PlotFlag is >= 1
%
% Call Syntax: [x1 y1] = ECDF(x, PlotFlag, LoxD,HixD);
% Input      : x = samples from a RV X (a vector),
%              PlotFlag is a number controlling plot (Y/N, marker-size)
%              LoxD is a number by which the x-axis plot range is extended to the left
%              HixD is a number by which the x-aixs plot range is extended to the right
% Output     : [x1 y1] & empirical CDF Plot IF PlotFlag >= 1
%
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

R=length(x);       % assume x is a vector and R = Number of samples in x
x1=zeros(1,R+2);
y1=zeros(1,R+2);        % initialize y to null vectors
for i=1:1:R        % loop to append to x and y axis values of plot
y1(i+1)=i/R;                    % append equi-increments of 1/R to y
end                % end of for loop
x1(2:R+1)=sort(x);        % sorting the sample values
x1(1)=x1(2)-LoxD; x1(R+2)=x1(R+1)+HixD;   % padding x for emp CDF to start at min(x) and end at max(x)
y1(1)=0; y1(R+2)=1;         % padding y so emp CDF start at y=0 and end at y=1

% to make a ECDF plot for large number of points set the PlotFlag<1 and use
% MATLAB's plot function on the x and y values returned by ECDF -- stairs(x,y)
if PlotFlag >= 1       % Plot customized empirical CDF if PlotFlag >= 1
    %newplot;
    MSz=10/PlotFlag;   % set Markersize MSz for dots and circles in ECDF plot
                       % When PlotFlag is large MSz is small and the
                       % Markers effectively disappear in the ecdf plot
    R=length(x1);       % update R = Number of samples in x
        hold on            % hold plot for superimposing plots

    for i=1:1:R-1
        if(i>1 && i ~= R-1)
            plot([x1(i),x1(i+1)],[y1(i),y1(i)],'k o -','MarkerSize',MSz)
        end
        if (i< R-1)
          plot(x1(i+1),y1(i+1),'k .','MarkerSize',2.5*MSz)
        end
        plot([x1(i),x1(i+1)],[y1(i),y1(i)],'k -')
        plot([x1(i+1),x1(i+1)],  [y1(i),y1(i+1)],'k -')

    end

    hold off;
end
```

6.4.2 q-th Sample Quantile

```
function qthSQ = qthSampleQuantile(q, SortedXs)
%
% return the q-th Sample Quantile from Sorted array of Xs
%
% Call Syntax: qthSQ = qthSampleQuantile(q, SortedXs);
%
% Input      : q = quantile of interest, NOTE: 0 <= q <= 1
%              SortedXs = sorted real data points in ascending order
% Output     : q-th Sample Quantile, ie, inverse ECDF evaluated at q

% store the length of the the sorted data array SortedXs in n
N = length(SortedXs);
Nminus1TimesQ = (N-1)*q; % store (N-1)*q in a variable
Index = floor(Nminus1TimesQ); % store its floor in a C-style Index variable
Delta = Nminus1TimesQ - Index;
if Index == N-1
    qthSQ = SortedXs(Index+1);
else
    qthSQ = (1.0-Delta)*SortedXs(Index+1) + Delta*SortedXs(Index+2);
end
```

6.5 Posterior Means for the frequencies of GN_D, GN_I and GN_U

```
%% raw data from 2 months -- 0811 is complete survey 0812 has missing
%% values as -1
```

```
%% Columns are:
%% Year        Month ID C_G C_I SU        SN GA GN GN_D GN_I GN_U
A=dlmread('SYN0811.M.txt');
B=dlmread('SYN0812.M.txt');

%% take just first 10 rows for now
As = A(1:10,:)
Bs = B(1:10,:)

%% let's get the missing data imputed in the last three columns: 10, 11, 12
Bs(:,10:12)
[Bs(:,9),sum(Bs(:,10:12),2)]%col 9 against sum of cols 10,11,12
% get the row IDs of missing data in cols 10,11,12
Missing101112 = find([Bs(:,9),sum(Bs(:,10:12),2)] * [1; -1] ~= 0)
disp('all cols for missing data in cols 10,11,12');
Bs(Missing101112,:)
disp('previous month data for these IDs (row number and Id are assumed to be the same)')
As(Missing101112,:)
disp('get the row IDs of filled data in cols 10,11,12');
Filled101112 = find([Bs(:,9),sum(Bs(:,10:12),2)] * [1; -1] == 0)
disp('all cols for filled data in cols 10,11,12');
Bs(Filled101112,:)
disp('get min and max imputation cells in col 4 (assume cells are numbered contiguously from min to max)');
CellIDCol=4; % For Geographic Imputation Cell Numbers in column 4
%CellIDCol=5; % For Imputation Cell Numbers in column 5
FirstImpCell = min(B(:,CellIDCol))
LastImpCell = max(B(:,CellIDCol))

%% make arrays for posterior distribution over (GN_D, GN_I, GN_U)
Posterior1=zeros(LastImpCell-FirstImpCell+1,3);
Posterior2=zeros(LastImpCell-FirstImpCell+1,3);
%PrePosterior2=zeros(LastImpCell-FirstImpCell+1,3);

% loop over the imputation cells contiguously from first to last
for ImpCell = FirstImpCell:LastImpCell
    ImpCell
    ImpCellIndices = find( ...
        B(:,CellIDCol)==ImpCell & ...
        [B(:,9),sum(B(:,10:12),2)] * [1; -1] == 0 & ...
        B(:,9) >= 0 & ...
        B(:,10) >= 0 & ...
        B(:,11) >= 0 & ...
        B(:,12) >= 0);
    Post1=sum(B(ImpCellIndices,10:12)) + [1,1,1];
    Posterior1(ImpCell,:)=Post1/sum(Post1);

    ImpCellIndicesA = find( ...
        A(:,CellIDCol)==ImpCell & ...
        [A(:,9), sum(A(:,10:12),2)] * [1; -1] == 0 & ...
        A(:,9) >= 0 & ...
        A(:,10) >= 0 & ...
        A(:,11) >= 0 & ...
        A(:,12) >= 0);
    A(ImpCellIndicesA,10:12)
    pause
    % the posterior of previous month
    Post2=sum(A(ImpCellIndicesA,10:12)) + [1,1,1];
    %assuming current and previos months form a set of homogeneous seb-pops
    %Post2=sum(B(ImpCellIndices,10:12)) + sum(A(ImpCellIndicesA,10:12));

    %assuming the sub-population's posterior of previous month is prior for this month
    %PrePost2=sum(A(ImpCellIndicesA,10:12)) + [1,1,1]
    %PrePosterior2=PrePost2/sum(PrePost2)
    %Post2=sum(B(ImpCellIndices,10:12)) + PrePosterior2

    Posterior2(ImpCell,:)=Post2/sum(Post2);

end
 Posterior1
```

```
Posterior2

%% now we do (Poisson, if col 9 is missing) multinomial sampling to fill
% (col 9) and cols 10,11,12
```

References

[1] J. Chen and J. Shao. Jackknife variance estimation for nearest neighbour imputation. *Journal of the American Statistical Association*, 96:260–269, 2001.

[2] J. Rao and J. Shao. Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79:811–822, 1992.

[3] D. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[4] D. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley, New York, 1987.

[5] D. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91:473–520, 1996.

[6] C.-E. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer-Vaerlag, 1992.

[7] J. Shao and R. Sitter. Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91:1278–1288, 1996.

[8] Statistics New Zealand. AO/BI/01 - Accommodation Survey 2007. Distributed by Statistics New Zealand at `http://www2.stats.govt.nz/`, 2007.