# A Bayesian method for deriving population statistics from multiple imperfect data sources

John Bryant and Patrick Graham*
Statistics New Zealand, Christchurch, NZ
Corresponding author: John Bryant, email: john.bryant@stats.govt.nz

**Abstract**

Replacing a traditional census with an administrative census requires finding new ways of (i) generating population estimates, and (ii) assembling individual-level socioeconomic data. We describe Bayesian methods under development at Statistics New Zealand for dealing with both problems. Population estimation is carried out by setting up a large model containing a demographic account, models of the demographic processes, and models of the measurement processes. Coverage errors in individual-level administrative data are addressed using multiple imputation, based on output from the population estimation model.

**Keywords:** Bayesian, official statistics, demography, multiple imputation, missing data

## 1 Introduction

A traditional census provides an authoritative count of the population, disaggregated by basic demographic variables such as age, sex, and region. In countries that have a traditional census, population data from the census are a key input to population estimation. Population estimates have many uses, from allocating health funding to targeting housing investment. Any scheme to replace a traditional census with one based purely on administrative data needs to include a method for constructing accurate population estimates (Bycroft, 2013; Office for National Statistics, 2013).

A traditional census also yields an individual-level dataset covering almost every resident of the country. This dataset includes, in addition to basic demographic variables, a range of socio-economic information on matters such as education, occupation, and family status. The individual-level dataset, like census-based population counts, has a wide range of uses, from sociological analyses to market research. Any scheme to replace a traditional census would have to include a method for generating an equivalent individual-level dataset from administrative data (Bycroft, 2013; Office for National Statistics, 2013).

In this paper, we describe some new Bayesian population estimation methods that help with both problems: the construction of population estimates, and the construction of an individual-level socioeconomic dataset.

## 2 Population estimation

In the absence of a traditional census, population estimation has to rely on multiple noisy administrative datasets. Traditional demographic methods break down when con-

*The views expressed in this paper are those of the authors, and not necessarily those of Statistics New Zealand

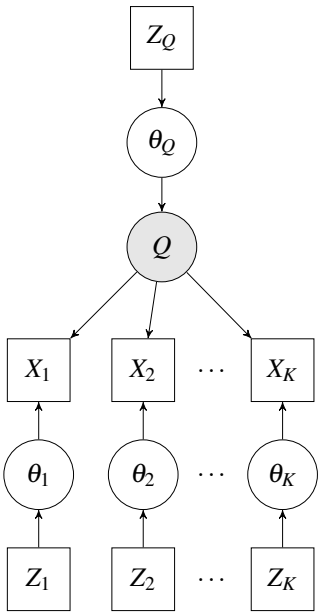


Figure 1: Bayesian population estimation. See the text for an explanation of the notation. Squares represent observed quantities, circles represent unobserved quantities, and arrows represent probabilistic relationships.

fronted with this type of data. In this section, we describe an alternative approach to population estimation that is being developed at Statistics New Zealand. A more detailed description is given in Bryant and Graham (2013). Figure 1 summarizes the framework.

At the core of the framework is a demographic account $Q$. The account consists of counts of people and events, linked by accounting identities. An account might, for instance, contain counts of births, deaths, international migrations, and population, all disaggregated by age, sex, region, and time, and all consistent with the identity that population at the end of the period equals population at the beginning plus births minus deaths plus net migration. We treat $Q$ as latent or unobserved.

Entries within $Q$ typically exhibit strong regularities. For instance, deaths follow a characteristic age profile. The model of the demographic account, $\theta_Q$, captures these regularities. Often there are auxilliary data $Z_Q$ that can assist with the estimation of parameters within $\theta_Q$. Data on regional income levels, for instance, can help explain variation in regional mortality rates.

Data sources $X_1, \ldots, X_K$ consist of counts of people or events, or of proxies for these counts. Examples include traditional sources such birth registrations, and more exotic sources such as tax data. A single data source $X_k$ may have been assembled by linking together several sets of administrative data. Models $\theta_1, \ldots, \theta_K$ capture the relationships between the demographic account and the data sources. Datasets $Z_1, \ldots, Z_K$ contain information that help explain variation in the relationship between account and data. For instance, data on housing type might explain variation in converage rates. If a coverage survey was taken for one of the $X_k$, data from this survey could be included in $Z_k$.

Inference is carried out via Markov chain Monte Carlo methods. A Gibbs sampler alternates between the full conditional distributions for $Q$, $\theta_Q$, and $\theta_X$. Sampling from the distribution for $Q$ is difficult because of the presence of the accounting identities; sampling from $\theta_Q$ and $\theta_X$ can be done using standard methods.

The approach offers some important advantages over traditional methods:

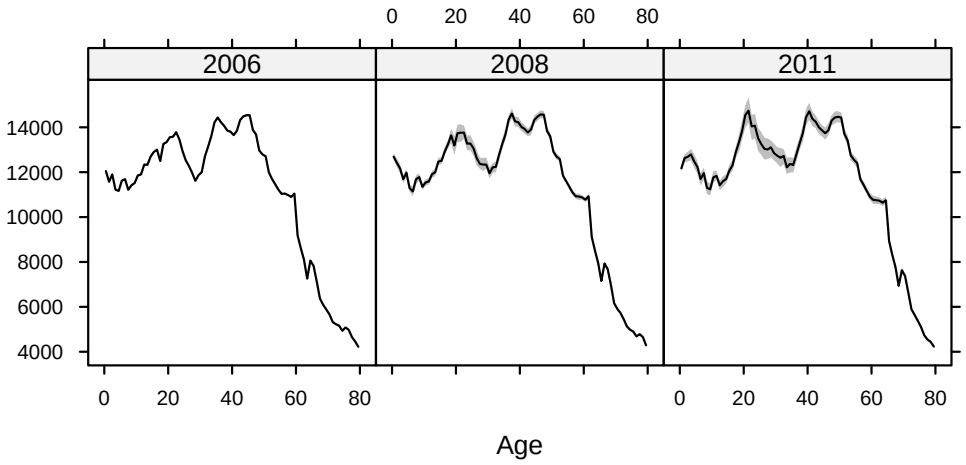


Figure 2: Population estimates for a single region of New Zealand. The black lines represent median estimated counts, and the grey bands represent 95% credible intervals. Uncertainty is low in 2006, a census year, higher in 2008, and higher again in 2011.

- Random variation in the demographic series and in the measurement of these series are properly accounted for.
- Detailed measures of uncertainty are produced.
- Extra datasets or extra dimensions can be easily be added.
- Missing or irregular data are accommodated naturally.

A prototype of the model has been built and tested, as described in Bryant and Graham (2013). Work is currently underway on software that is faster and more flexible than the prototype. Some illustrative results from the model are shown in Figure 2.

If a data source, say $X_1$, was very high quality, would the population estimation model be redundant? If $X_1$ included all the variables that the users of population estimates required, if it defined the population in the appropriate way, and if it was error-free, then population estimates could be read straight off it, and no population estimation model would be needed. If $X_1$ had minor errors, then the choice of whether to model or not would be less obvious. On the one hand, modelling requires extra work, and introduces errors of its own because of simplifying assumptions. On the other hand, any improvement in the accuracy of $X_1$ should lead to improvements in the accuracy of the model estimates. Moreover, the model can exploit additional information contained in the other $X_k$s and in the $Z_k$s, as well as taking account of demographic plausibility.

Outside of Scandanavia, most administrative datasets are likely to have large enough errors for modelling to be worthwhile. New Zealand, for instance, has unusually accurate data on births, deaths, and international migration, and has made rapid progress in linking together administrative datasets. However, most administrative datasets in New Zealand do not share a common personal identifier, and no administrative dataset covers all New Zealand residents. Moreover, data on residential addresses are generally poor (Gibb, 2013). The result is that all administrative datasets in New Zealand, linked or otherwise, miss eligible people, include ineligible people, and have missing data for important variables.

# 3 Producing an individual-level dataset

Under a traditional census, individual-level socioeconomic data are obtained by requiring everyone in the country to fill out a questionnaire. Under an administrative census, individual-level socioeconomic data are obtained by cleaning and linking administrative and survey datasets until enough of the desired variables are included, and coverage of the population is sufficiently high. As discussed above, almost all such datasets are subject to under-coverage, over-coverage, and missing values. If users were to analyse one of these datasets in its uncorrected form, they could easily be misled. An analysis of the assimilation of migrants, for instance, could easily go astray if the individual-level dataset incorrectly included migrants who had returned to their home country. To address these problems, statistical methods are needed. We are currently developing one such method, based on an extension of our population estimation model.

Let $X_1$ be the original individual-level dataset with the coverage errors and missing values. Let $X_1^{\text{true}}$ be the dataset we would like to have, containing no coverage errors or missing values. We get from $X_1$ to $X_1^{\text{true}}$ in two steps: we correct for coverage errors, and then we fill in the missing values. Let $X_1^{\text{cov}}$ be a dataset that has been corrected for coverage errors, but not for missing values. Using $X^{\text{obs}}$ to denote $X_1, \ldots, X_k$, we have

$$p(X_1^{\text{true}}|X^{\text{obs}},Z) = \int p(X_1^{\text{true}}|X_1^{\text{cov}},X^{\text{obs}},Z)p(X_1^{\text{cov}}|X^{\text{obs}},Z)\mathrm{d}X_1^{\text{cov}} \quad (1)$$

$$= \iint p(X_1^{\text{true}}|X_1^{\text{cov}},Q,X^{\text{obs}},Z)p(X_1^{\text{cov}}|Q,X^{\text{obs}},Z)p(Q|X^{\text{obs}},Z)\mathrm{d}X_1^{\text{cov}}\mathrm{d}Q. \quad (2)$$

Values for $X^{\text{true}}$ can be obtained by

1. drawing a value for $Q$, conditional on $X^{\text{obs}}$ and $Z$;
2. drawing a value for $X^{\text{cov}}$, conditional on $Q$, $X^{\text{obs}}$, and $Z$; and
3. drawing a value for $X^{\text{true}}$, conditional on $Q$, $X^{\text{obs}}$, $X^{\text{cov}}$, and $Z$.

Drawing values for $Q$ can be done using the methods sketched out in Section 2. Drawing values for $X^{\text{true}}$ can be done using standard methods for multiple imputation of missing data (Rubin, 1996). The imputation models would have available to them all the information in $Q$, $X^{\text{obs}}$, and $Z$. Drawing values for $X^{\text{cov}}$ is less standard, and the approach would need to be adapted to the data at hand.

Drawing values for $X^{\text{cov}}$ is easiest when an 'included-at-random' assumption is made: when it is assumed that, within each cell defined by $Q$, individuals are missing from $X_1$, and erroneously included in $X_1$, at random. This assumption can be made plausible by incorporating into $Q$ as many variables as possible that are associated with coverage.

Having made an 'included-at-random' assumption, a simple way of obtaining $X^{\text{cov}}$, for a given value of $Q$, would be to add or remove individuals until all cell counts in the corrected version of $X_1^{\text{obs}}$ matched the corresponding counts in $Q$. Individuals could be added by creating records with missing values on all variables except the variables used to define the cell. Missing values could be filled in when drawing from $p(X^{\text{true}}|X^{\text{cov}},Q,X^{\text{obs}},Z)$. Individuals could be deleted by removing randomly-selected records.

The method for obtaining $X^{\text{cov}}$ could be further refined to take advantage of extra information in $Z$. For instance, if $Z$ included data from a coverage survey, it might be used to give better predictions on the types of people likely to be missing. If $Z$ included information on patterns of misclassification, it might be possible align cell counts in

$X_1$ and $Q$ by randomly moving individuals between cells, as well as randomly adding and deleting. This would allow more of the original socioeconomic data in $X_1$ to be preserved.

The final product would be a set of $M$ versions of $X^{\text{true}}$ that could be treated like any other multiply-imputed dataset. Each individual $X^{\text{true}}$ could be analysed using ordinary complete-data methods, with the results for all $X^{\text{true}}$ being combined at the end (Rubin, 1996). With cheap computing power and the proliferation of software for multiply-imputed data, the practical obstacles to using such data are much smaller than they once were. The pay-off is sounder inferences than are possible with an uncorrected $X_1$ or a single corrected $X_1$.

# 4 Discussion

Fienberg (2011) and Little (2012) argue that Bayesian methods deserve a larger place in the production of official statistics. Both authors draw their examples mainly from surveys. Statistical methodologies for administrative data, Bayesian or otherwise, are still at an early development stage. However, we suspect that the arguments for bringing Bayesian methods into official statistics will turn out to be even stronger for administrative data than they are for survey data.

When making inferences from survey data, the biggest source of uncertainty is often sampling variability. When inferring from administrative data, the sources of uncertainty are more diverse, ranging from linkage errors, to misaligned target populations, to reporting lags (Zhang, 2011). Bayesians can represent all these sources of uncertainty in the same way, via probability distributions. Expert judgment about likely biases, for instance, can be captured in the form of prior distributions (Greenland, 2005).

It is rare for a single administrative source to contain all the information that a statistical office needs. Using administrative data typically means combining information from several datasets. A classic example of combining of information is record linkage. However, as illustrated in this paper, other types of linkage are also possible and useful. All of the $X$s and $Z$s in Figure 1, for instance, are linked in the sense that they influence each others' data models, via their influence on $Q$. This sort of general 'information linkage' poses no special difficulties within a Bayesian framework. Likelihoods are derived and multipled together, and then combined with prior distributions, much as in the case of a single data source. Bayesian methods are a natural way to deal with multiple imperfect administrative datasets.

# References


Bryant, J. R. and Graham, P. J. (2013). Bayesian demographic accounts: Subnational population estimation using multiple data sources. *Bayesian Analysis*.

Bycroft, C. (2013). Options for future New Zealand censuses: Census transformation programme. Technical report, Statistics New Zealand.

Fienberg, S. (2011). Bayesian models and methods in public policy and government settings. *Statistical Science*, 26(2):212–226.

Gibb, S. (2013). Evaluating administrative sources for population estimates. Technical report, Statistics New Zealand.

Greenland, S. (2005). Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2):267–306.

Little, R. J. (2012). Calibrated Bayes, an alternative inferential paradigm for official statistics. *Journal of Official Statistics*, 28(3):309–334.

Office for National Statistics (2013). Beyond 2011: Options explained 2. Technical report, Office for National Statistics.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.

Zhang, L.-C. (2011). A unit-error theory for register-based household statistics. *Journal of Official Statistics*, 27(3):415.