A Bayesian framework for estimating fertility from multiple data sources

March 19, 2018

Extended abstract submitted to the "VIII Congreso Internacional de la Asociación Latinoamericana de Población" in Ciudad de Puebla, Mexico, from 10/23/2018 a 10/26/2018

Abstract

For decades, demographers have estimated fertility rates in Brazil using indirect demographic techniques. More recently, scholars have challenged the results of these techniques. This issue remains unsolved and there has been significant disagreement about what the levels of fertility for the past decades are, particularly for subnational levels. This paper proposes a Bayesian hierarchical method that combines different estimates of fertility and completeness of registered births, which are, in turn, confronted with other information, such as census coverage and infant mortality, to check for consistency. Application for Brazil in the years 1991, 2000 and 2010 shows that estimates resulting from this method have lower variability than the initial estimates. Furthermore, results seem to correct for some biases in the initial estimates. This is a promising approach that could be applied in a variety of contexts, and is particularly useful for regions with incomplete vital statistics systems. The proposed Bayesian method has the advantage of being flexible enough to allow fertility estimation based on independent information of either fertility rates, completeness of registered births, or the combination of both.

1 Introduction

For decades, demographers have estimated fertility rates in Brazil using indirect demographic techniques. Even though the limitations of these methods have been widely known, the lack of alternative data sources have led to a general consensus that these estimates reasonably describe the overall levels and trends in fertility (Berquó and Cavenaghi, 2014; Borges and Silva, 2015; Carvalho, 1982). More recently, with the continuous decline of fertility levels and change in the age schedule, in addition to a greater availability of alternative data sources due to the improvement of vital registration systems, scholars have challenged the results of these techniques (Carvalho, Gonçalves, and Silva, 2017; Castanheira and Kohler, 2015).

This issue remains unsolved and there has been significant disagreement about what the levels of fertility for the past decades are, particularly for subnational levels. Fertility estimates using different methods and data sources have led to different results. Indirect demographic methods have several limitation, but vital registration systems in Brazil are also limited. Even though vital registration systems have improved substantially oven in the last years, there are still a fair amount of births that are not registered, particularly in the less developed regions, which undermines their use without any adjustment.

This paper proposes a Bayesian hierarchical method that combines different estimates of fertility and completeness of registered births, which are, in turn, confronted with other information, such as census coverage and infant mortality, to check for consistency.

The results of this method provide a range of plausible values of the Total Fertility Rates (TFR). In addition to provide more precise estimates, this method also indicates which estimates are more implausible so that their results should be interpreted carefully.

Probabilistic approaches have been used to estimate fertility and mortality for contexts with incomplete vital registration systems. Alkema et al., (2012) and Liu and Raftery, (2017) developed methods to incorporate the uncertainty of past Total Fertility Rates (TFR) by using a method for estimating the bias and variance of different sources of data with varying data quality, mostly censuses and surveys. These approaches take into account sampling and non-sampling errors, which are evaluated through comparison with official estimates.

The appeal of using Bayesian analysis in these contexts lies in its potential to overcome the challenges of combining information from different sources and dealing with high stochastic variation, measurement errors and lack of identifiability in the models.

2 Methods and Data

2.1 Method

Demographic events such as births are subject to random variation and may be assumed to follow a Poisson distribution (Brillinger, 1986). Thus, the total number of births B_c women from cohort c (K_c) have is Poisson distributed as follows:

$$B_c \sim Poisson(K_c \cdot f_c) \tag{1}$$

where f_c is the age-specific fertility rate (ASFR) for cohort c.

To take underregistration of births into account, the number of registered births, B_c^{obs}, is modeled using a binomial distribution:

$$B_c^{obs} \sim Binomial(B_c, \beta_c) \tag{2}$$

where β_c is the probability of a birth being reported, which in turn may be modeled by the conjugate prior beta distribution. This mixture gives an algorithm for simulating from the beta-binomial: draw from the prior distribution $\beta_c \sim Beta(a_c^B, b_c^B)$ and then draw $B_c^{obs} \sim Binomial(B_c, \beta)$ (Gelman et al., 2013).

The models described below can be expressed hierarchically as:

$$B_c^{obs} \sim Poisson(K_c \cdot f_c \cdot \beta_c) \tag{3}$$

$$f_c \sim Gamma(a_c^f, a_c^f) \tag{4}$$

$$\beta_c \sim Beta(a_c^B, b_c^B) \tag{5}$$

Fertility rates f_c are often estimated by using information collected in surveys and censuses about fertility ((UN, 1983, chapter 2), (Moultrie, 2013)). Data about completeness of registered births are estimated directly, e.g. by surveys using capture and recapture methods.

A Bayesian approach is a natural choice to deal with this kind of problem, since identification problems arise from (3), since there is a range of values of K_c, f_c and β_c that maximizes the likelihood. In other words, the likelihood that derives from (3) only allows inference about the product $(K_c \cdot f_c \cdot \beta_c)$, and gives no possibility to estimate the parameters K_c, f_c and β_c individually, which are ultimately the measures of interest. The advantage of this Bayesian setup is that it is flexible enough to allow fertility estimation based on independent information of either fertility rates, completeness of registered births, or the combination of both.

It is often the case that the population at risk, K_c is taken as known. Census data are normally used as a proxy of the true population, which is a very strong assumption. The model proposed here relaxes this assumption by modeling fertility jointly with population counts, which is discussed below.

Demographic censuses are used for many purposes, serving as denominator of several rates, including fertility rates. Censuses, however, are not perfect and often presents coverage and quality problems. Census populations are usually smaller than the true population, meaning that census undercount exceeds overcount. The census counts K_c^{obs} could then be modeled as a binomial distribution:

$$K_c^{obs} \sim Binomial(K_c, \kappa_c) \tag{6}$$

where K_c is the true but unobserved population and κ_c is the census coverage for cohort c.

Even though the most common use of the binomial distribution is to estimate the probability of success κ_c given the number of successes K_c^{obs} in a series of experiments K_c, statisticians have also tried to make inference about the true but unobserved parameter number of trials (K_c). This issue is often called the "binomial n problem" and has been also addressed in the context of estimating total population through capture and recapture models in wildlife (Otis et al., 1978) and human populations (Wolter, 1986).

In human populations, K_c^{obs} normally comes from censuses and κ_c and K_c are parameters to be estimated. The Post-Enumeration Survey (PES) is the natural data source to model κ_c in equation 6. IBGE has carried out PES in Brazil since the 1970 Census. Information about K_c is much harder to obtain.

The binomial model is limited because it contains only one free parameter and the variance is determined by the mean. When estimating census counts, for example, both moments are calculated from the coverage estimation of the PES. More importantly, the model in equation 6 only accounts for random sampling errors. Sampling errors in the PES are relatively easy to control and quantify and depend mostly on the sample size. This tends not to be a problem for large population groups, as those used in this study. With increases in the sample sizes of the PES, sampling errors have been dominated by uncertainties due to systematic non-sampling errors. Non-sampling errors are harder to identify and measure and arise from many sources, such as correlation bias, processing and matching errors, among others (Wachter and Freedman, 2000) . Freedman and Wachter, (2003) suggest that large PES sample sizes not only increase the relative importance of non-sampling errors, but also make them more problematic, since bigger samples are harder to manage so that systematic errors are made more difficult to control and measure.

To take these issues into account, an over-dispersed version of the binomial distribution is required. In a Bayesian framework, the most used one is the beta-binomial distribution, where the probabilities of success, in this case κ_c, follow a beta distribution. The beta distribution is defined in the interval [0, 1] and parametrized by two shape parameters, a_c^K and b_c^K:

$$\kappa_c \sim Beta(a_c^K, b_c^K) \tag{7}$$

A conceptual difficulty with Bayesian analysis in the "binomial n problem" is to find sufficiently flexible and tractable family of prior distributions for the discrete parameter "n", in this case K_c (DasGupta and Rubin, 2005; Raftery, 1988). This has also some practical problems, since some statistical programs and modeling languages do not perform inference for discrete unknown parameters and these discrete parameter models need to be re-expressed as mixture models with continuous parameters (Team, 2017).

A natural alternative to overcome these difficulties is to model population counts with a Poisson distribution, which captures a feature often observed in count data, that is the observation error increases with the population size. More precisely, in the Poisson distribution, the variance is equal to the mean.

Thus, the true population counts K_c are assumed to follow a Poisson distribution with rate λ_c:

$$K_c \sim Poisson(\lambda_c) \tag{8}$$

As previously mentioned, prior information on λ_c is much harder to obtain, and a non-informative prior on this quantity is often required. This paper uses an improper uniform prior a for λ_c, which leads to a proper prior for K_c^{obs}.

The distributions described above (6, 7, 8) lead the following hierarchical structure:

$$K_c^{obs} \sim Poisson(\lambda_c \cdot \kappa_c) \tag{9}$$

$$\lambda_c \propto 1 \tag{10}$$

$$\kappa_c \sim Beta(a_c^K, b_c^K) \tag{11}$$

The posterior distribution of equation 9 results from the combination of the likelihood for K_c^{obs} and the priors for κ_c and λ_c. The resulting posterior distribution for K_c provides information about the plausibility of different values for the total population given the observed data and the prior knowledge about these parameters.

Figure 1 shows the summary of the model. The registered birth counts, B_c^{obs}, are modeled based on likelihood and prior information about coverage of census (κ_c), completeness of registered births (β_c) and fertility rates (f_c). If there is no prior information about either f_c or β_c, non-informative priors can be chosen. Notice that the population of children is modeled independently based on the observed population in the census (K_0^{obs}) and census coverage of this group (κ_c). It also uses information from fertility (f_c), population of women at reproductive ages (K_c) and survival (S_0).

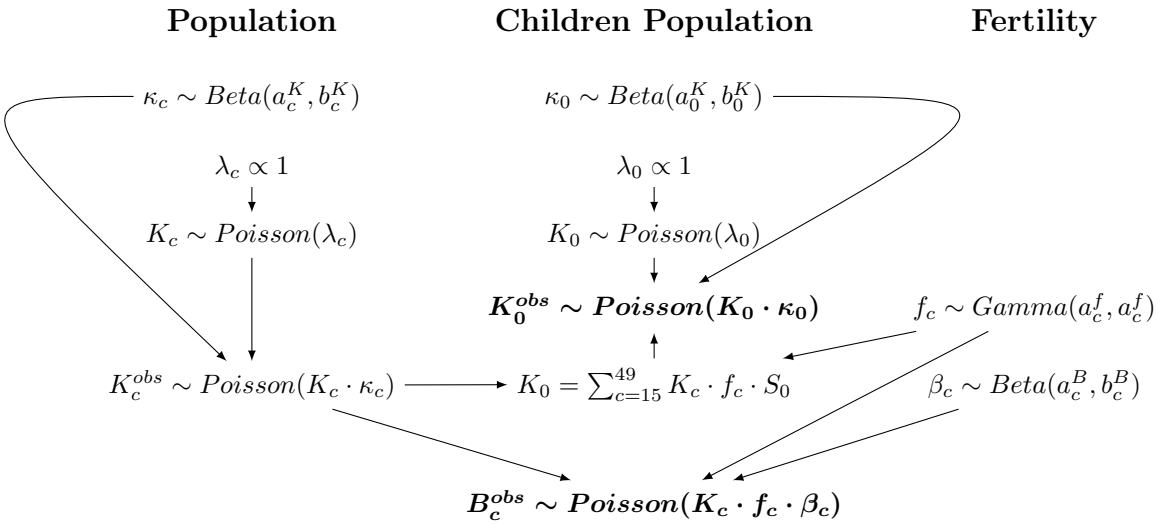

Figure 1: Diagram of the the relationship between likelihood, priors and fertility and population models

2.2 Data

The observed population (K_c^{obs}) of women at reproductive ages and children under age one comes from the 1991, 2000 and 2010 censuses carried out in Brazil. Prior information for the census coverage by age group (κ_c) estimates come from the PES of the 1991 and 2000 censuses. Since de 2010 PES has not been published, results for the 2000 PES have been used instead. The mean estimate of κ_c for 2000 is the same as that for 2010, but a higher variance was assigned to the 2010 estimate to take into account greater uncertainty.

Observed birth counts (B_c^{obs}) come from the Vital Statistics System produced by the Ministry of Health, the Information System on Live births (SINASC) and the prior on the completeness of registered births (β_c) is based on the estimates calculated by RIPSA, (2012).

Total Fertility Rates (TFR) come from a variety of data sources, which use different estimation methods [1].

2.3 Estimating prior distributions

This paper uses the methods of moments and percentiles to approximate distributions that represent the prior information about the parameters TFR and β_c, κ_c and κ_0. The proportion of the TFR for each age group is assumed to be the same for the different estimates. The birth completeness β_c is also assumed to be constant across age groups.

2.4 Simulating the posterior distribution

The model was fitted using the statistical softwares *R* and *Stan*. Samples from the posterior distributions of the parameters were draw via a Markov Chain Monte Carlo (MCMC) algorithm. More specifically, Stan uses the Hamiltonian Monte Carlo (HMC) algorithm to explore the target distribution. The HMC algorithm tends to explore the posterior distribution in a more efficient way. Efficiency in this context means that it requires fewer samples to describe the posterior distributions. HMC gains efficiency by reducing randomness when moving through the parameter space and exploiting knowledge of the target distribution. A practical advantage of the HMC algorithm is that, unlike Gibbs sampling and the Metroppolis algorithm, it makes easier to identify problems and divergences when sampling from the posterior (Carpenter et al., 2017; McElreath, 2016; Team, 2017).

3 Preliminary Results

Figure 2 shows prior and posterior distributions for the Total Fertility Rates (TFR) and other parameters used in the model (β_c, κ_c, κ_0) for the years 1991, 2000 and 2010. The point estimates of the TFR are also shown in the first panel.

The posterior distributions of the TFR are significantly different from the prior distributions. The posteriors also show much lower variability than the priors. These results indicate that including additional information changes the previous knowledge about fertility rates, in addition to increase precision of the estimates.

TFR estimates for 1991 have higher variance, since no information about completeness of registered births (β_c) was used. This is shown by the flat prior on this parameter for 1991. The only point estimate that seems highly implausible is 2.30, which is much lower than range of the posterior distribution.

For 2000, the posterior distribution is shifted to the right, compared to the prior. There is very low uncertainty in the fertility estimates for Brazil in 2000 and 95% of the posterior distribution is concentrated between 2.20 and 2.40.

For 2010, the posterior distribution of the TFR is highly concentrated below the prior mean. This indicates that the distribution of the TFR was shifted markedly to the left given the likelihood of population and births counts, in addition to their respective coverage priors. For 2010, 95% of the posterior distribution is concentrated between 1.67 and 1.86. This indicates that estimates of 1.6 and 1.9 seem highly implausible for Brazilian fertility in 2010. These two estimates result from the the question about children born in the last year, both unadjusted and adjusted by using the P2/F2 Brass ratio factor.

3.1 Discussion

This abstract presents a novel method to reconcile and estimate Total Fertility Rates based on multiple data sources. Results show that estimates have lower variability than the initial estimates. Furthermore, the method seem to be able to adjust for biases in the initial estimates. This is a promising approach that could be applied in a variety of contexts, but is particularly useful for regions with incomplete vital statistics systems. The proposed Bayesian method has the advantage of being flexible enough to allow fertility estimation based on independent information of either fertility rates, completeness of registered births, or the combination of both.

[1] The final version of this paper will detail the sources and methods used for each point estimate.

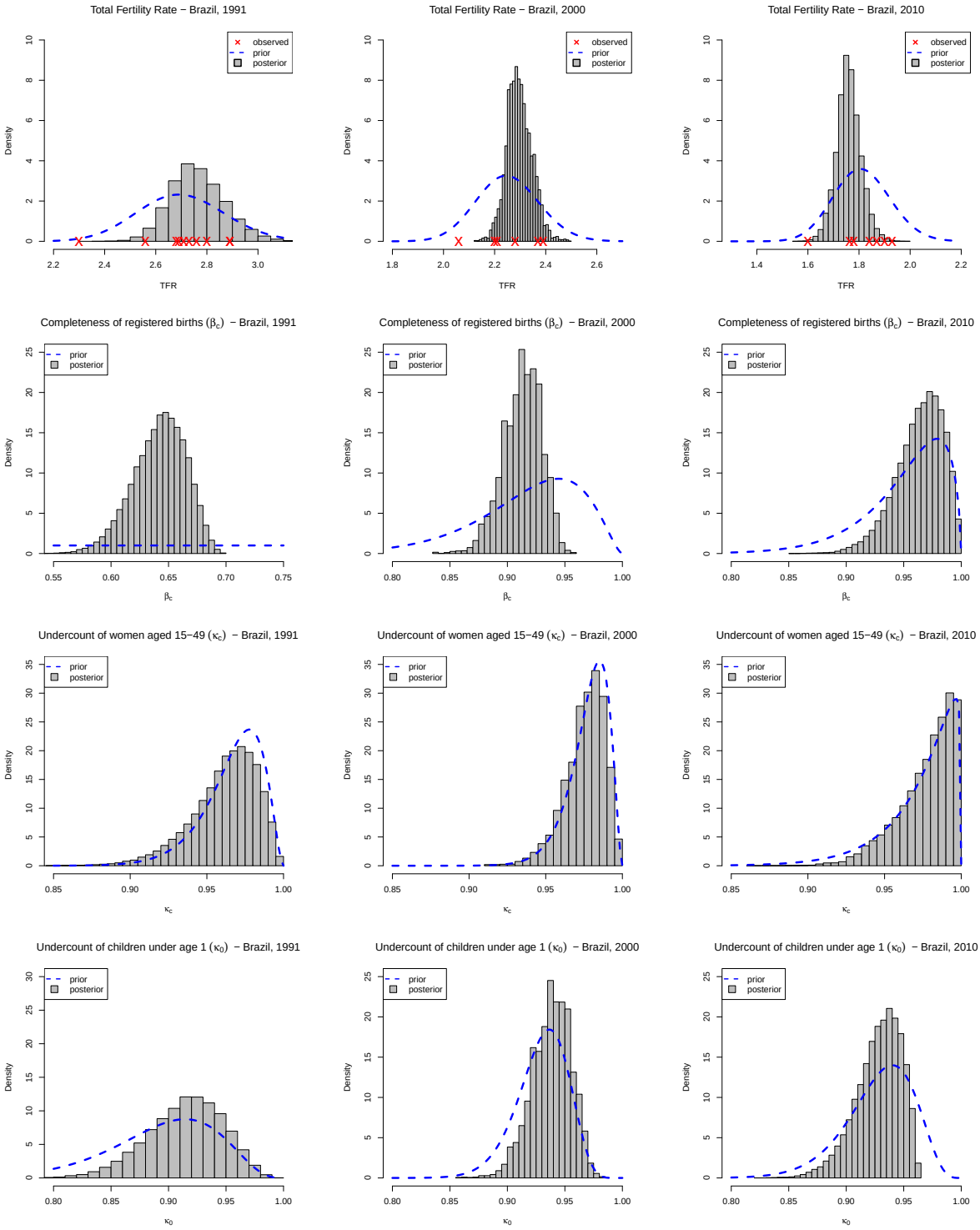

Figure 2: Prior and posterior distributions of the Total Fertility Rate (TFR), completeness of registered births (β_c) and adjustment factor of the census (κ_c, κ_0)

References

Alkema, Leontine et al. (2012). "Estimating Trends in the Total Fertility Rate with Uncertainty Using Imperfect Data: Examples from West Africa". In: *Demographic research* 26.15. ISSN: 1435-9871. DOI: 10.4054/DemRes.2012.26.15.

Berquó, Elza S. and Suzana M. Cavenaghi (2014). "Notas sobre os diferenciais educacionais e econômicos da fecundidade no Brasil". pt. In: *Revista Brasileira de Estudos de População* 31.2, pp. 471–482. ISSN: 1980-5519.

Borges, G. M. and L Silva (2015). "Fontes de Dados de Fecundidade No Brasil: Caracterısticas, Vantagens e Limitações". In: *Mudança Demográfica No Brasil No Inıcio Do Século XXI: Subsıdios Para as Projeções Da População*. Rio de Janeiro: IBGE, pp. 138–151.

Brillinger, David R. (1986). "A Biometrics Invited Paper with Discussion: The Natural Variability of Vital Rates and Associated Statistics". In: *Biometrics* 42.4, pp. 693–734. ISSN: 0006-341X. DOI: 10.2307/2530689.

Carpenter, Bob et al. (2017). "Stan: A Probabilistic Programming Language". en-US. In: *Journal of Statistical Software* 76.1. DOI: 10.18637/jss.v076.i01.

Carvalho, J. A. M. C (1982). "Aplicabilidade Da Técnica de Fecundidade de Brass Quando a Fecundidade Está Declinando Ou Quando a População Não é Fechada". In:

Carvalho, J. A. M. C, G. Gonçalves, and L Silva (2017). *Estimativas de Fecundidade No Brasil, Grandes Regiões e Unidades Da Federação, Em 2010, a Partir Da Aplicação Da Técnica P/F de Brass No Contexto de Rápida Queda Da Fecundidade Adolescente*. Textos Para Discussão Cedeplar-UFMG 564. Cedeplar, Universidade Federal de Minas Gerais.

Castanheira, Helena and Hans-Peter Kohler (2015). "It Is Lower Than You Think It Is: Recent Total Fertility Rates in Brazil and Possibly Other Latin American Countries". In: *PSC Working Paper Series, WPS 15-5*.

DasGupta, A. and Herman Rubin (2005). "Estimation of Binomial Parameters When Both n, p Are Unknown". In: *Journal of Statistical Planning and Inference*. Herman Chernoff: Eightieth Birthday Felicitation Volume 130.1, pp. 391–404. ISSN: 0378-3758. DOI: 10.1016/j.jspi.2004.02.019.

Freedman, David A and Kenneth W Wachter (2003). "On the Likelihood of Improving the Accuracy of the Census through Statistical Adjustment". In: *Lecture Notes-Monograph Series* 40, pp. 197–230. ISSN: 0749-2170.

Gelman, Andrew et al. (2013). *Bayesian Data Analysis, Third Edition*. en. CRC Press. ISBN: 978-1-4398-4095-5.

Liu, P and Adrian Raftery (2017). "Accounting for Measurement Error in Bayesian Probabilistic Projection of the Total Fertility Rate". In: Chicago: PAA.

McElreath, Richard (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. en. Boca Raton, FL: CRC Press.

Moultrie, Tom A. (2013). "Overview of Fertility Estimation Methods Based on the P/F Ratio". In: *Tools for Demographic Estimation*. Paris: International Union for the Scientific Study of Population.

Otis, David L. et al. (1978). "Statistical Inference from Capture Data on Closed Animal Populations". In: *Wildlife Monographs* 62, pp. 3–135. ISSN: 0084-0173.

Raftery, Adrian (1988). "Inference for the Binomial N Parameter: A Hierarchical Bayes Approach". en. In: *Biometrika* 75.2, pp. 223–228. ISSN: 0006-3444. DOI: 10.1093/biomet/75.2.223.

RIPSA (2012). *Indicadores e Dados Básicos - Brasil - 2012*.

Team, Stan Development (2017). *Stan Modeling Language: User's Guide and Reference Manual*.

UN (1983). *Manual X: Indirect Techniques for Demographic Estimation*. ST/ESA/SER.A/81. Population Studies 81. New York: United Nations.

Wachter, Kenneth W and David A Freedman (2000). "The Fifth Cell: Correlation Bias in U.S. Census Adjustment". en. In: *Evaluation Review* 24.2, pp. 191–211. ISSN: 0193-841X. DOI: 10.1177/0193841X0002400202.

Wolter, Kirk M. (1986). "Some Coverage Error Models for Census Data". In: *Journal of the American Statistical Association* 81.394, pp. 337–346. ISSN: 0162-1459. DOI: 10.1080/01621459.1986.10478277.