How plausible are small-area estimates of fertility in Sub-Saharan Africa?

Dennis Feehan
dfeehan[at]princeton.edu

DRAFT: March 1, 2011*

Abstract

In many parts of the world, reliable direct estimates of demographic quantities like births and deaths are unavailable for small geographical areas, yet this information is essential to good policy and research. Model-based strategies for combining different data sources may help in this situation, provided the assumptions they rely upon hold. I investigate the plausibility of some of these methods using the example of district-level fertility in Uganda. I fit a Bayesian hierarchical model to birth history data from a national survey, and then estimate fertility rates at the district level. I evaluate the model's performance using district-level fertility rates computed from two census microsamples. The results from this simple specification suggest that the model has the potential to be useful, but that it overstates the precision of its estimates.

*I would like to thank Kosuke Imai, who first introduced me to these methods through their applications in political science, and whose class served as a starting point for this project. I also benefitted from many helpful discussions with Matthew Salganik. Please do not cite this version of the paper; it is very much a work in progress. I would be grateful to receive any comments or suggestions at dfeehan[at]princeton.edu.

Draft: results for presentation at PAA, March 31 2011.
Please do not cite

1 Introduction

Many countries suffer from a lack of adequate data on the distribution and welfare of their populations. Although wealthy nations have vital registration systems that record the occurence and location of virtually all of their citizens' births and deaths, in the rest of the world policymakers and researchers must often rely on other strategies to produce the estimates of these quantities (Setel et al., 2007). These alternatives typically involve some combination of household surveys, decennial censuses, and administrative records. For the measurement of population processes like births and deaths, household surveys are very often the most timely and complete resources avaialble. Although some surveys - notably the Demographic and Health Surveys - are generally regarded as very high quality, they are often designed to provide estimates only at the national, or regional level.

In such cases, certain model-based estimators may help (Pfeffermann, 2002; Rao, 2005). If the assumptions they rely upon are met, these strategies allow the analyst to combine detailed household survey data with less complete, but more extensive, data from censuses or administrative records to produce estimates of demographic quantities for small areas[1]. Methods similar in spirit to the one explored below have been applied, with varying degrees of success, in several fields, including the estimation of poverty (Elbers, Lanjouw and Lanjouw, 2003), political opinions (Park, Gelman and Bafumi, 2004), and health insurance coverage (DeNavas-Walt, Proctor and Smith, 2009). However, critics have cautioned that these methods make assumptions that are often unreasonable in practice and that they can lead to a misleading sense of precision (Tarozzi and Deaton, 2009).

So how do we know if the assumptions these model-based strategies depend upon are reasonable? The answer will, of course, depend on the data sources available and on the quantity of interest. In any application of these methods, an important first step is to determine whether or not they produce accurate

[1] In this report, we will only investigate geographically small areas; however, there is no reason why the methods described here could not be used for other purposes, for example making inferences about a small ethnic group.

and reliable estimates in an environment where we know the correct value of the quantity being estimated. This paper investigates the plausibility of methods for estimating small-area fertility from a combination of survey and census data, using the example of Uganda. I fit a simple Bayesian hierarchical model to the 2000-2001 Uganda Demographic and Health Survey (UBOS and Macro, 2001). I then use the model to predict fertility rates by age at the district level for five-year time periods from 1970 to 2000. I evaluate the model's performance using district-level fertility rates computed from the 1991 and 2002 microsamples of Uganda's population censuses.

2 Data

The 2000-2001 Uganda Demographic and Health Survey consists of a sample of 7,246 women aged 15-49 (UBOS and Macro, 2001). The sample was designed to provide estimates for the national level, at the level of each of Uganda's four regions, and for urban and rural areas. Full birth histories were obtained as part of the interviews, permitting the computation of fertility rates using standard techinques (Rutstein and Rojas, 2003).

In order to evaluate the accuracy of the district-level fertility rates produced by the model, I employ the public-use microsamples of the 1991 and 2002 Uganda Population and Housing Censuses (Minnesota Population Center, 2010; Uganda Bureau of Statistics, 1991, 2002). The 2002 microsample is a simple random sample of 10 percent of the records collected in the 2002 Census, which gives us very precise estimates of age-specific fertility in each of Uganda's 54 districts. The 1991 microsample is a stratified sample of 10 percent of the records collected in the 1991 census. For the 1991 census, I used the sampling weights produced by the national statistical office to compute the district level age-specific fertility rates; again, the quantity of records available make the estimates very precise. Below, I use these Census-derived fertility rates to evaluate the performance of the model's predictions[2].

[2] Uganda has created twenty-four additional districts since 2002, making a total of 80 today; this analysis concerns itself with the districts as of 2002, since this permits us to evaluate the model estimates using the census microsample.

3 Methods

The is a large literature on using surveys and other data sources to produce estimates for small areas; thorough reviews of the statistical literature are provided by Ghosh and Rao (1994), Pfeffermann (2002), and Rao (2005). In this paper, I present results from one of the most simple models that could be applied for our purposes. Other strategies are available, but this approach is comparatively straightforward and is a natural starting point in considering how likely these methods are to produce estimates that would be useful to researchers and policymakers.

I model the number of births observed in the survey from a given district, five-year age group, and five-year time period as Poisson distributed with separate terms for age, time period, and district; that is, I assume is that

$$\begin{aligned} b_i &\sim \text{Poisson}(\theta_i n_i) \qquad (1) \\ \log(\theta_i) &= \mu + \beta^{\text{age}}_{\text{age}[i]} + \beta^{\text{year}}_{\text{year}[i]} + \beta^{\text{dist}}_{\text{dist}[i]}, \end{aligned}$$

where b_i is the number of births in age-year-district i, n_i is the amount of exposure, θ_i is the fertility rate, age$[i]$ is the age 5-year category that i is in, year$[i]$ is the 5-year time period, and dist$[i]$ is the district. Fitting this model to the survey data available using techniques like least squares or maximum likelihood would be challenging because of the limited information available for any one age-year-time cell from the survey responses. In order to obtain estimates of the coefficients in the model, I therefore adopt a simple hierarchical Bayesian approach. In this case, in order to estimate the age, year, and district terms in a hierarchical fashion, I propose the following model:

$$
\begin{aligned}
b_i &\sim \text{Poisson}(\theta_i n_i) \\
\log(\theta_i) &= \mu + \beta^{\text{age}}_{\text{age}[i]} + \beta^{\text{year}}_{\text{year}[i]} + \beta^{\text{dist}}_{\text{dist}[i]} \\
\mu &\sim N(0, 10,000^2) \\
\beta^{\text{year}}_j &\sim N(0, \sigma^2_{\text{year}}),\ j \in \{1, \ldots, n_{\text{year}}\} \\
\beta^{\text{age}}_k &\sim N(0, \sigma^2_{\text{age}}),\ k \in \{1, \ldots, n_{\text{age}}\} \\
\beta^{\text{dist}}_l &\sim N(0, \sigma^2_{\text{dist}}),\ l \in \{1, \ldots, n_{\text{dist}}\},
\end{aligned}
\tag{2}
$$

where b_i is the observed number of births for district-age-year i, n_i is the amount of exposure for i, θ_i is the fertility rate for i, and the β are the parameters for age, year, and district. I assume diffuse, Uniform$(0, 100)$ priors on σ_{year}, σ_{age}, and σ_{dist}.

A key advantage of this approach is that the estimation of the distribution of district terms permits us to make inferences, though possibly imprecise ones, for districts that have a very small sample in the survey dataset, or even for those that do not show up at all (Gelman and Hill, 2007; Lynch, 2007; Park, Gelman and Bafumi, 2004). Hierarchical models are well-suited to this situation, since they permit partial pooling, meaning that estimates for cells that have few or no survey responses will be influenced mainly by overall means from all of the cells, while the estimates for cells that have lots of responses will mainly be informed by the data (Gelman and Hill, 2007; Lynch, 2007).

I obtain samples from the posterior distribution of the parameters through a Markov-chain Monte-Carlo algorithm implemented using JAGS (Plummer, 2003); I ran three chains for 25,000 iterations each as a burn-in period, and then obtained a further 2,500 samples from each chain to use for inference. Diagnostics of the samples I obtained suggested that the sampler had converged, so that the draws it was producing were coming from the posterior we are interested in. For example, traceplots indicated that the chains had thoroughly mixed, and the Gelman-Rubin $\hat{R}$ for all of the model parameters was less than or equal to 1.01 (Gelman, 2004; Lynch, 2007).

There are at least two quantities of interest that the model produces for each district. The first is the age-specific fertility rates, and the second is the number of births. Predictions for the age-specific fertility rates for each district can be made directly from the model, using the posterior distributions of the β^{age}, β^{year}, and β^{dist}. In the case where I wish to produce estimates for a district that was not in the survey, I repeatedly draw $\beta^{\text{dist}}_{\text{new}}$ from the posterior distribution of the β^{dist} and use those draws in my predictions.

In order to obtain estimates for the number of births, rather than the rates, I post-stratify the model's predicted rates for each age-year-district cell on the total number of women in that cell. This involves the following steps: for each district I obtain the total number of women in each age category for the year we are interested in, P_{day}. If the district we wish to predict for is found in the survey dataset, I use the values of θ_i sampled from the posterior, where i is the survey cell matching P_{day},[3] to obtain predictions for the number of births, $\hat{b}^m_{day} \sim \text{Poisson}(P^m_{day}\hat{\theta}^m_i)$, for $m = 1, \ldots, 2,500$. The procedure for districts that are out of sample is the same, except in that case I draw the district effect from its posterior distribution, $\beta^{\text{dist},m} \sim N(0, \sigma^{2,m}_{\text{dist}})$ again for $m = 1, \ldots, 2,500$. I add the sampled district effects to the posterior draws of the age and year parameters, and proceed to post-stratify on P_{day}. The method is very similar to Park, Gelman and Bafumi (2004), who apply it to estimate state-level public opinion from national polls in the US.

Comparators

In evaluating the performance of the model, it is useful to consider alternative estimators that might be used to obtain district-level fertility rates. Below, I consider two of these. The first is the survey estimator, which is the estimates produced for a given district using only the data from the survey produced by respondents in that district. Since not all of the districts in Uganda had any surveys, this estimator does not exist in all districts; for districts with large samples, though, it may be reasonable. The second comparator is what I call the national rate estimator. This is simply obtained by using the national-level

[3] That is, θ_i is the fertility rate derived from the tally of exposure and births for year y from the women in the survey who live in district d, and are in age group a.

fertility rates, derived from the household survey, as estimators for the fertility schedule in each district.

4 Results

As an example of some of the output, Figure 1 shows marginal and bivariate summaries of the posterior samples for the age parameters, which are the β^{age} in Equation 2. Below the diagonal, we see estimated bivariate relationships between the samples of the parameters for each pair of age groups; the diagonal has a histogram of the marginal distribution of each age parameter, and above the diagonal, we see the posterior correlations. Although these were assumed to be independent in the prior, we can see that the posterior indicates that many of these pairs are correlated in demographically sensible ways: for example, Figure 1 shows that high estimates in the age group 15-19 are often estimated together with high estimates in the age group 20-24. On the other hand, the negative relationships between posterior estimates of the parameters from age groups 15-19 and 45-49, suggests that when fertility at young ages is relatively high, fertility at older ages tends to be relatively low.

The district-specific terms, along with their 95% credible intervals are shown in Figure 2. Substantively, these give us a rough idea of the relative level of fertility estimated for each district. The district of Kampala, Uganda's largest city, is estimated to have the lowest fertility. Note that the size of the uncertainty interval varies from district to district, since varying amounts of data are available from the survey for each one.

Similar plots, containing posterior means and 95% credible intervals for the age-group and time-period terms, are shown in Figure 5. On the left we see the posterior estimates of the age-group terms, which follow a characteristic age-fertility pattern. Estimates for the younger ages are more precise because there is more data available for them, while estimates for the age group 45-49 are not very precise. On the right, we see estimates of the time-period terms. Again, estimates for the most recent time period, 2000-2001 (the year of the survey) are relatively imprecise, since there is less data available for them. Nonetheless, these estimates suggest that fertility overall increased markedly over the period from 1985-1994, and then decreased dramatically between 1995-2001. Although part of this may be due to recall problems and other sources of non-sampling error in the survey responses, declines in the fertility of several countries in the region over the 1990s have been noted in the literature (Ezeh, Mberu and Emina, 2009).

Figure 4 shows the age-specific fertility rates from the model, the survey, the national estimator, and the two census microsamples, for the district of Kiboga. The grey bands show the 95% credible interval for the posterior estimates produced by the model. The green and light red lines show the observed age-specific rates from the 1991 and 2002 censuses, respectively. Evidently, in this case the model has done a good job of recovering the values observed in the census; both are contained within the model's predicted band. On the other hand, direct estimates from the survey, that is, estimates computed using survey responses from the district of Kiboga alone, are quite noisy, as we see in the pink lines. The national rate estimator, which is national-level age-specific fertility rates used as estimators of the rates for the district of Kiboga, perform worse than the model does as well.

On the other hand, Figure 5 shows that the model fares somewhat less well in the case of Kampala. Here, it uniformly oveestimates the age-specific rates for the time period 2000-01; moreover, the 95% intervals for the model's predictions in Kampala are quite narrow, giving a misleading impression of high precision. Nonetheless, the model's estimates do appear to be better than the ones for the national rate estimator.

As Figure 5 suggests, an evaluation of the performance of the uncertainty intervals is a critical part of understanding how well a model like this one would work in practice. I computed the coverage of the 95% credible intervals produced for each age-district fertility rate prediction from the model for the time period of 2000-01 and 1990-94, taking the true values to be the ones from the census microsamples. That is, for each prediction in an age-district cell for the time periods that have census data, I examined whether or not the 95% credible interval contained the census rate. The fraction of times that it did is what I call the coverage. Figure 6 shows the results. It indicates that the model's 95% intervals never included the census value more than about 70%

of the time, and sometimes did so much less. The precision of the model's estimates is thus overstated.

5 Discussion

The results from this application of a simple hierarchical model to produce district-level estimates of fertility rates from a household survey in Uganda are mixed. We see that in some cases the model can perform quite well, producing plausible patterns that do a good job of recovering fertility rates measured independently in two national censuses. In other cases, however, the model's predictions are less accurate and, perhaps more concerning, the precision of its estimates is overstated.

Efforts to refine these models should focus on the second of these problems, with particular attention paid, first to the issue of interactions: the model proposed in Equations 1 and 2 could be made more flexible by allowing interactions between district and year, permitting more detailed district time trends to emerge. Another important avenue of exploration is the inclusion of overdispersion terms, which would explicitly model extra-Poisson variation.

References

DeNavas-Walt, C., BD Proctor and JC Smith. 2009. "US Census Bureau current population reports, P60-236: income, poverty and health insurance coverage in the United States, 2008." *Washington, DC: Government Printing Office* .

Elbers, C., J.O. Lanjouw and P. Lanjouw. 2003. "Micro-level estimation of poverty and inequality." *Econometrica* pp. 355–364.

Ezeh, A.C., B.U. Mberu and J.O. Emina. 2009. "Stall in fertility decline in Eastern African countries: regional analysis of patterns, determinants and implications." *Philosophical Transactions of the Royal Society B: Biological Sciences* 364(1532):2991.

Gelman, A. 2004. *Bayesian data analysis.* CRC press.

Gelman, A. and J. Hill. 2007. *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press Cambridge.

Ghosh, M. and JNK Rao. 1994. "Small area estimation: an appraisal." *Statistical Science* 9(1):55–76.

Lynch, S.M.I would like to thank. 2007. "Introduction to applied Bayesian statistics and estimation for social scientists.".

Minnesota Population Center. 2010. *Integrated Public Use Microdata Series, International: Version 6.0 [Machine-readable database].*

Park, D.K., A. Gelman and J. Bafumi. 2004. "Bayesian multilevel estimation with poststratification: state-level estimates from national polls." *Political Analysis* 12(4):375.

Pfeffermann, D. 2002. "Small Area Estimation: New developments and directions." *International Statistical Review/Revue Internationale de Statistique* pp. 125–143.

Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, March.* Citeseer pp. 20–22.

Rao, JNK. 2005. *Small Area Estimation.* Wiley-Interscience.

Rutstein, S.O. and G. Rojas. 2003. "Guide to DHS statistics." *Calverton, MD: ORC Macro* .

Setel, P.W., S.B. Macfarlane, S. Szreter, L. Mikkelsen, P. Jha, S. Stout and C. AbouZahr. 2007. "A scandal of invisibility: making everyone count by counting everyone." *The Lancet* 370(9598):1569–1577.

Tarozzi, A. and A. Deaton. 2009. "Using census and survey data to estimate poverty and inequality for small areas." *The Review of Economics and Statistics* 91(4):773–792.

UBOS and ORC Macro. 2001. *Uganda Demographic and Health Survey 2000-2001.* Uganda Bureau of Statistics (UBOS) and ORC Macro.

Uganda Bureau of Statistics. 1991. *1991 Uganda Population and Housing Census.* Uganda Bureau of Statistics.

Uganda Bureau of Statistics. 2002. *2002 Uganda Population and Housing Census.* Uganda Bureau of Statistics.

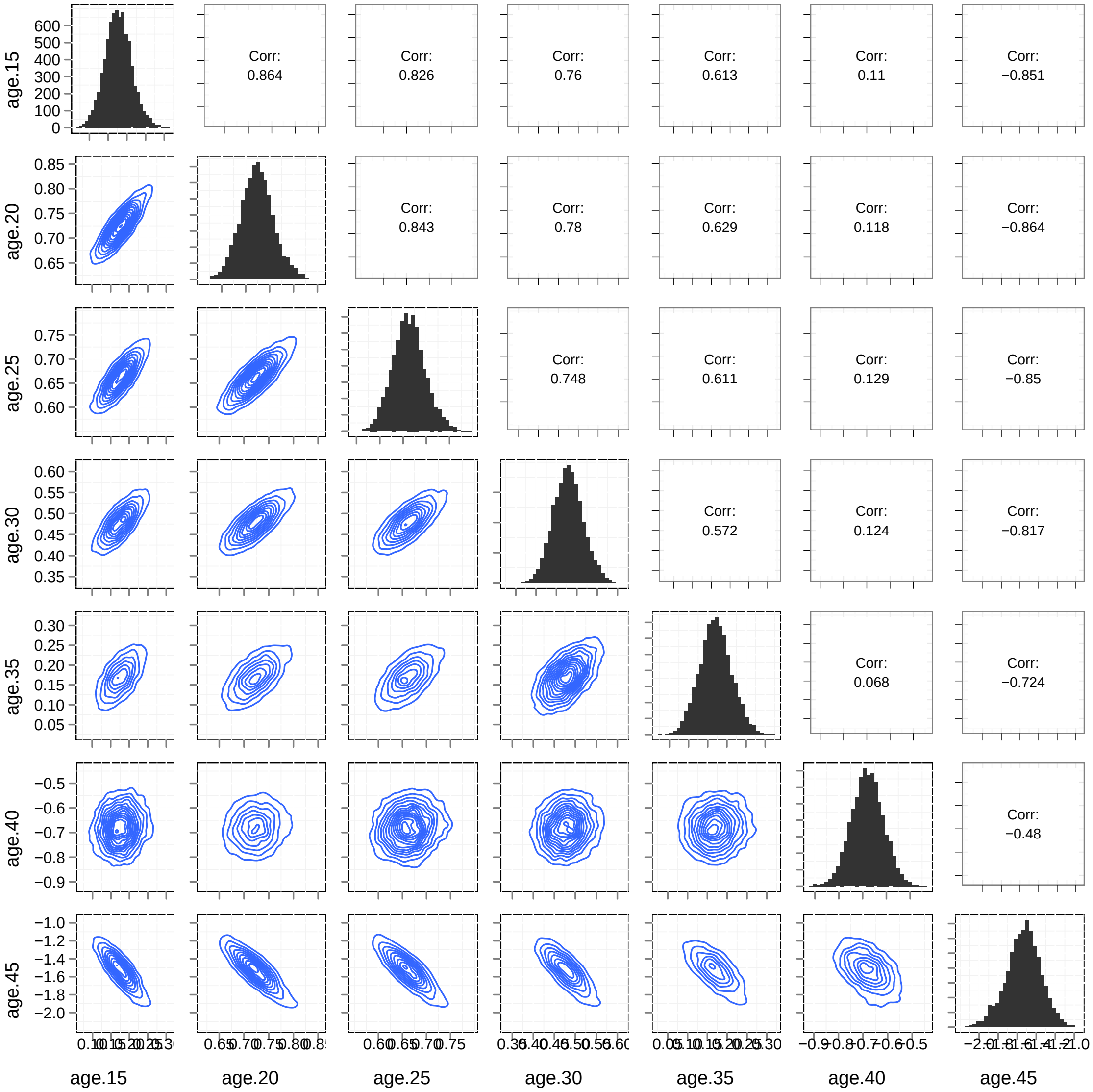

Figure 1: Bivariate and marginal relationships between the posterior draws of the age terms, β^{age}. Below the diagnal, in blue, are contour maps of the bivariate relationship between each pair of age terms, and above the diagonal are correlations summarizing them. For some pairs, like younger age terms, there is a strong positive relationship; for example, the correlation of 0.864 between posterior draws of the 15-19 and 20-24 age group terms indicates that when one of those groups was estimated to be higher than average, the other often was as well. On the other hand, there is a less strong or negative relationship between some of the younger age group terms and the oldest ones; for example, the estimated posterior correlation between the draws from age group 15-19 and 45-49 was -0.851, indicating that when the lowest age group was estimated to be relatively high-fertility, the oldest was estimated to be relatively low-fertility, and vice-versa.

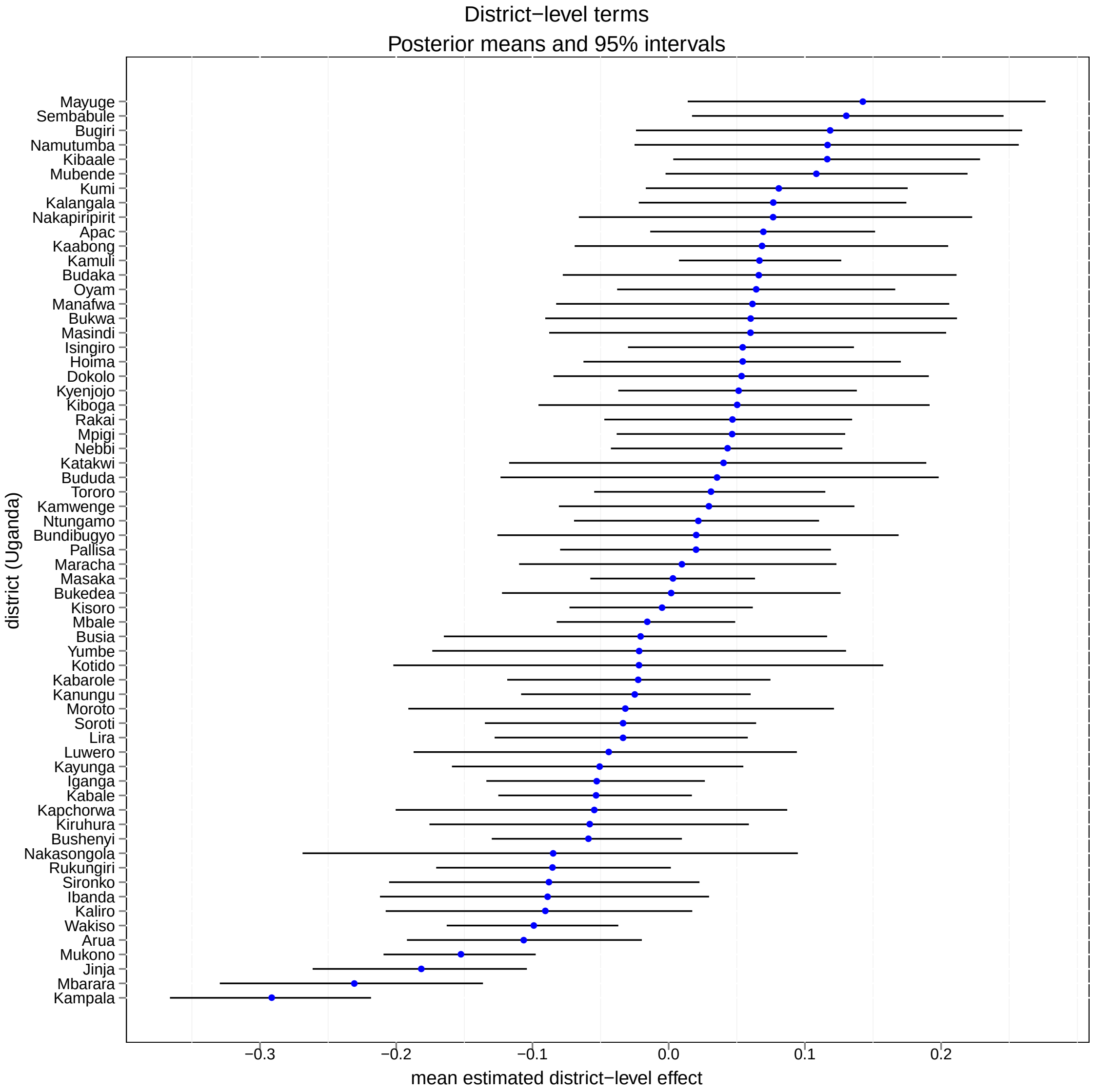

Figure 2: Means and 95% credible intervals for the posterior estimates of the district terms, β^{dist}. We see that the overall level of fertility is estimated to vary considerably across different districts, with Kampala, Uganda's largest city, having the lowest estimate and Mayuge estimated to have the highest. Note also that some of the terms are estimated more precisely than others because different amounts of data were available in the survey for different districts.

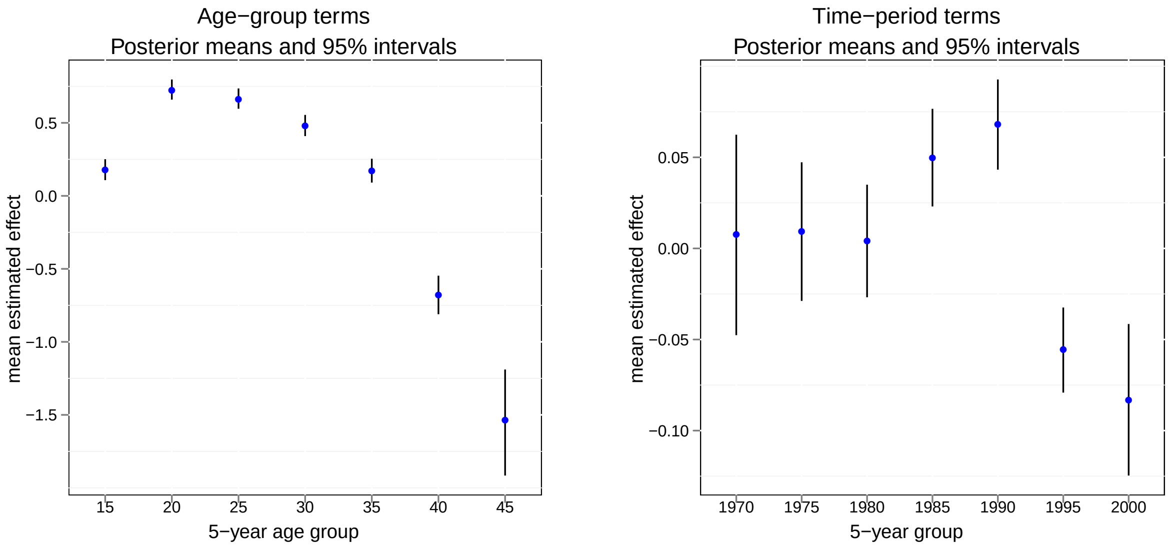

Figure 3: Means and 95% credible intervals for the posterior estimates of the age-group terms, on the left, and the time-period terms, on the right. These show the estimated average pattern of relative fertility by age and time period. From the age-group terms, on the left, we see a plausible fertility pattern, with much more precise estimates at the younger ages, where there are more births in the data. On the right, we see that the model estimates a marked increase in fertility during 1985-1994, but then a sharp decline from 1994-2001. Parts of this pattern may be due to recall bias in the survey responses, though a decline in fertility over this time period in some parts of the region has been noted in the literature. Again, the oldest and most recent time periods are least precisely estimated, since the survey has the least amount of data about them.

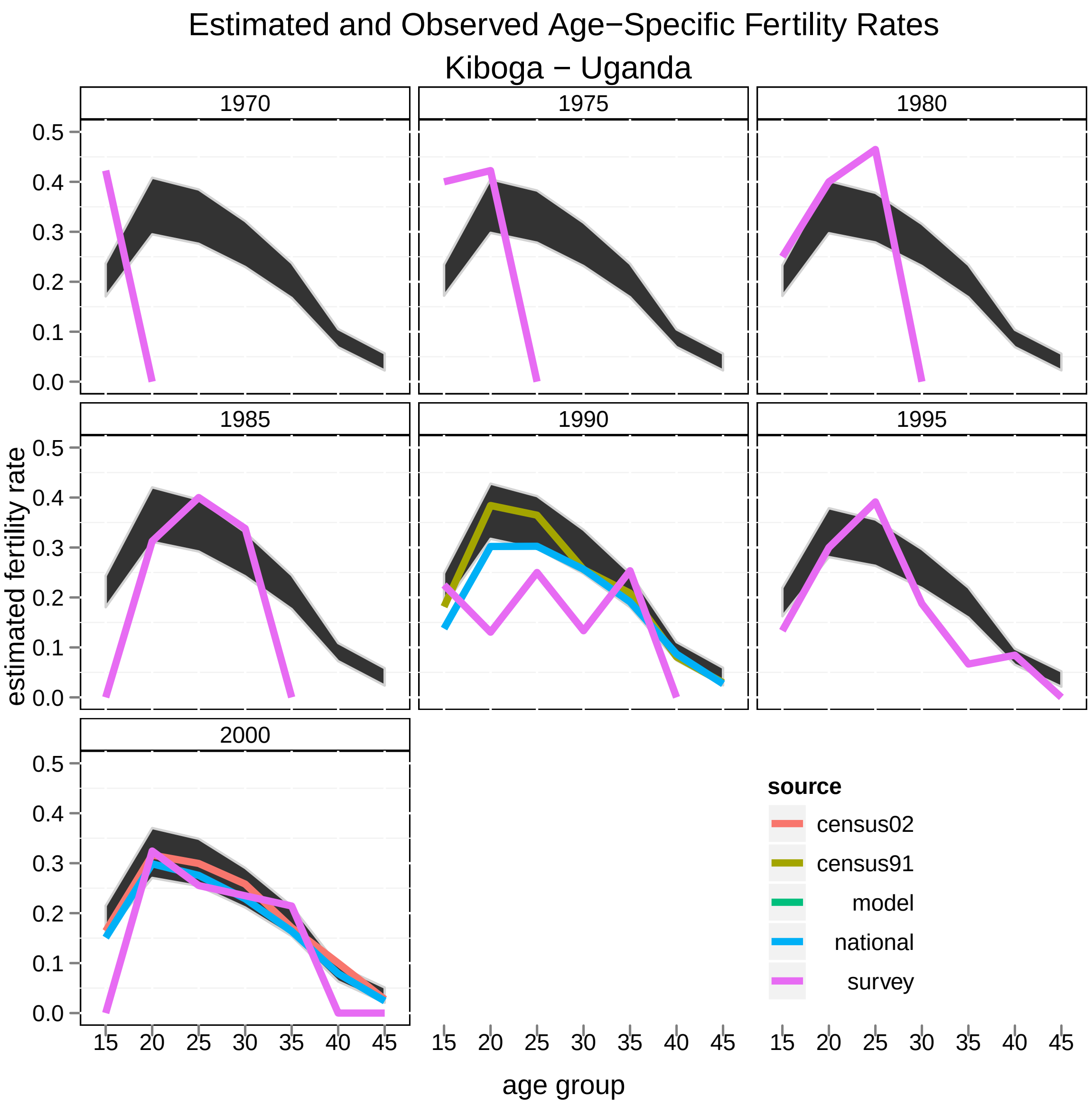

Figure 4: Age-specific fertility rates for the district of Kiboga, Uganda from the model, the survey, national rates, and two census microsamples. The 95% credible interval of the estimates produced by the model are shown in the grey bands. Direct estimates computed from the survey (ie, just using survey data from this district) are shown in pink. National rates, which might be used in practice if district rates are not available, are shown in blue. Rates from the 1991 census are shown in green, and the rates from the 2002 census are shown in light red. In this district, the simple model I tested appears to do quite a good job: although the direct estimates, in pink, are generally very noisy and would be unlikely to be usable in practice, the model's estimates recover the rates observed in both census quite well, and are an improvement over the national rates.

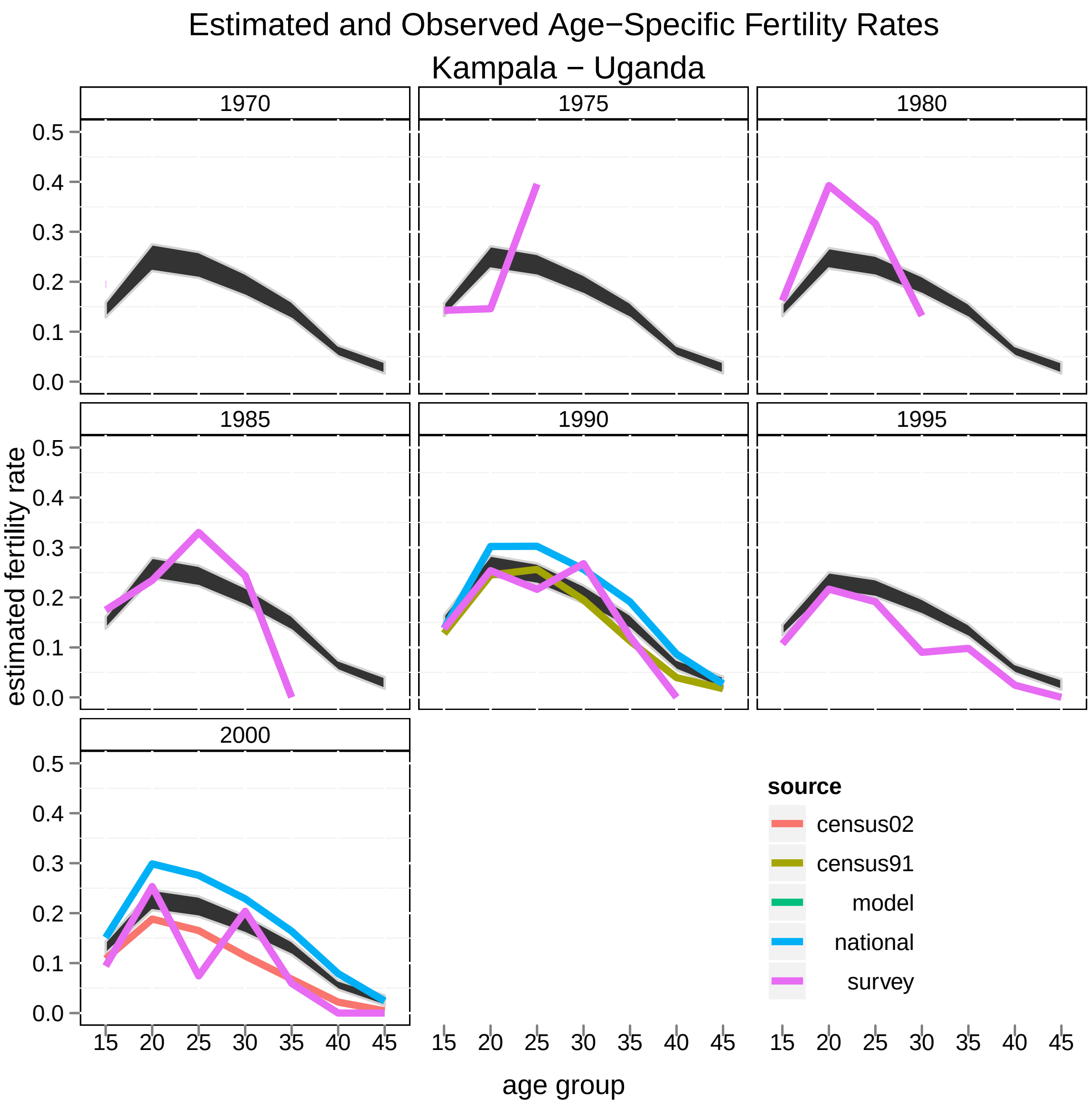

Figure 5: Age-specific fertility rates for the district of Kampala, Uganda's biggest city, from the model, the survey, national rates, and two census microsamples. The 95% credible interval of the estimates produced by the model are shown in the grey bands. Direct estimates computed from the survey (ie, just using survey data from this district) are shown in pink. National rates, which might be used in practice if district rates are not available, are shown in blue. Rates from the 1991 census are shown in green, and the rates from the 2002 census are shown in light red. In this case, the model does not do so well. The grey bands are quite narrow, indicating that it produces precise estimates; however, the estimates do not recover the rates from the 2000 census very well. However, the model does appear to be an improvement over using national rates as estimators for the rates in Kampala, since the national rates are farther from the census estimates than the model's predictions in both time periods when census data are available.

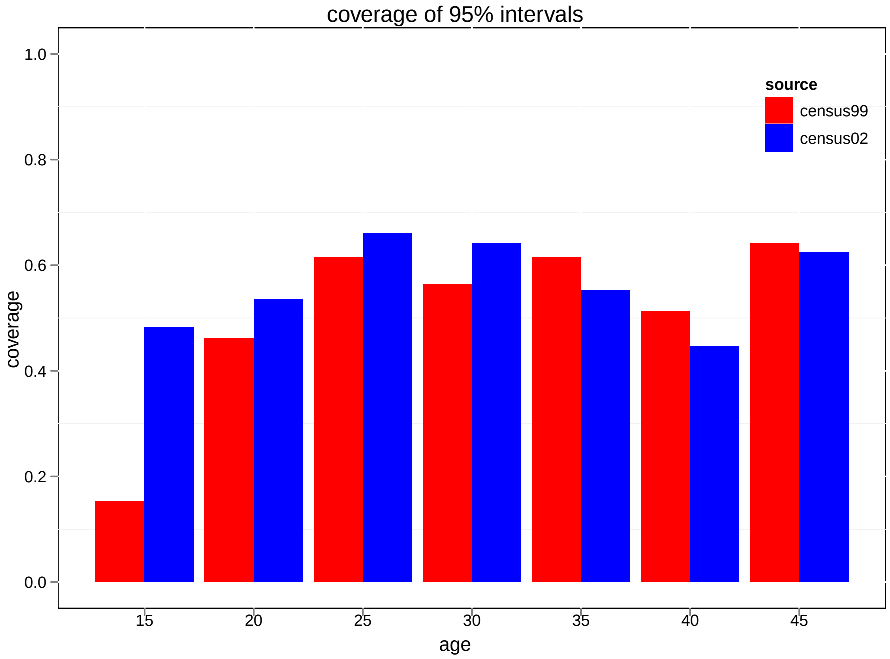

Figure 6: Observed coverage of the 95% credible intervals from the model's predictions, where the true values are taken to be the rates from the census in each district for 1991 and 2002. The performance of the model's uncertainty estimates is not very good: in none of the age groups or time periods does the model's 95% credible interval include more than 70% of the true values. This is an indication that a more elaborate model may be warranted; however, other factors may also contribute, including imperfections in the census-based fertility rates.