# Estimating adult mortality using sampled social network data: evidence from Brazil

Dennis M. Feehan* Matthew J. Salganik†

August 30, 2018

**Abstract**

Measuring adult mortality is fundamental to science and policy, yet hundreds of millions of people are affected by the *scandal of invisibility*: they live in places where death rates cannot be directly measured because most deaths are never formally recorded (Setel et al. 2007). Developing methods to estimate death rates has been challenging in part because it is rarely possible to collect data needed to estimate death rates in an environment where a gold standard is available for comparison. To help address this challenge, we conducted household surveys in 27 Brazilian cities ($n \approx 25,000$) to empirically validate several leading approaches to estimating adult death rates, including the new network survival method, the sibling survival method, and models based on those two methods. Our study will contribute to understanding both how data about deaths should be collected and also how models can be used to improve the accuracy and precision of estimated death rates.

## 1 Extended Abstract

### 1.1 Introduction

Measuring adult mortality is fundamental to science and policy, yet hundreds of millions of people are affected by the *scandal of invisibility*: they live in places where death rates cannot be directly measured because most deaths are never formally recorded (Setel et al. 2007). Developing methods to estimate death rates has been challenging in part because it is rarely possible to collect data needed to estimate death rates in an environment where gold standard death rates are available for comparison. To help address this challenge, we conducted household surveys in 27 Brazilian cities ($n \approx 25,000$) to empirically validate several leading approaches to estimating adult death rates, including the new network survival method, the sibling survival method, and models based

*UC Berkeley
†Princeton University

on those two methods. Our study will contribute to understanding both how data about deaths should be collected and also how models can be used to improve the accuracy and precision of estimated death rates.

### 1.2 Study design

Researchers who wish to estimate adult mortality are faced with different possible ways to collect information about deaths and different possible ways to estimate death rates from the collected information. There has been little empirical guidance available to help inform these decisions because it is typically difficult to collect the data needed to estimate death rates in a setting where gold standard rates are available for comparison. We designed our study to help overcome this challenge by collecting the data needed to estimate death rates using a variety of different methods in 27 different cities where gold standard estimates are available.

We used a multistage probability design to obtain a representative sample of adults over age 18 living in Brazil's 26 state capitals and in Brasilia (Figure 1; further details are in Appendix 3.1). The 27 cities in our sample have essentially complete death registration at adult ages, meaning that gold-standard adult death rates are available from the vital registration system[1].

In order to estimate death rates in each city, our survey collected two types of information[2]: reports about deaths in respondents' personal networks (Feehan et al. 2017); and respondents' sibling histories (Brass 1975; Graham et al. 1989; Rutenberg and Sullivan 1991).

Given the information we collected from survey respondents, two different statistical philosophies can be used to produce estimates: design-based and model-based. Design-based estimators rely upon the assumption that the observed data are a random sample from the population, where the sampling mechanism — i.e., the survey design — is known. In this design-based framework, uncertainty in a given estimate comes from sample to sample random variation due to the sampling mechanism (Sarndal et al. 2003). Design-based estimates are appealing because the estimators are simple and transparent; they rely upon a smaller number of assumptions than model-based estimates; and they are often amenable to sensitivity analyses that reveal the possible impact of errors in reporting or other factors affecting estimated death rates (Feehan and Salganik 2016a; Feehan et al. 2017).

However, design-based estimates can be inefficient – i.e., they can have large uncertainty intervals when sample sizes are not large. Design-based estimates can also perform poorly when the mech-

[1] A secondary source of adult mortality estimates in these 27 cities can be obtained from the data collected in Brazil's 2010 census.

[2] In this extended abstract, we present preliminary results from two types of estimates: design-based estimates from personal network reports and one set of model-based estimates from personal network reports. The paper will explore a wider range of models, and will also include estimates based on the sibling histories.

anism that produced the data differs from the assumed sampling design because of nonresponse, incomplete sampling frames, or other problems. Finally, design-based estimates may not take advantage of all of the information available to the researcher.

Research in statistics and demography has proposed a wide range of model-based approaches as an alternative to the design-based approach. A well-specified model has the potential to produce estimates that are more efficient and robust than design-based estimates. However, building a model requires adding complexity and assumptions, and a poorly specified model can lead to inaccurate estimates. Therefore, the extent to which a given model improves upon design-based estimates is an empirical question.

In our study, we use the results of our survey to produce design and model-based estimates, with the aim of comparing several different methods for estimating adult death rates in a setting where gold-standard estimates are available for validation. Critically, we will not compare any estimates to the gold-standard until after all of the estimates have been finalized and pre-registered. Thus, our project can be conceptually divided into three phases: (1) study design and data collection; (2) constructing estimated death rates; and (3) evaluating the accuracy of estimated death rates. By separating steps (2) and (3) – i.e., by not looking at the gold standard until all of the estimates have been computed – we mimic the situation that researchers estimating adult death rates usually face: the gold standard is not known.

### 1.3 Design-based estimates

One way to estimate adult death rates from a sample survey is the network survival method. Feehan et al. (2017) showed that the design-based network survival estimator for the death rate in group $\alpha$ in city $i$ is:

$$\widehat{M}_{\alpha}^{(i)} = \frac{\widehat{y}_{F,D_\alpha}^{(i)}}{\widehat{\bar{d}}_{F_\alpha,F}^{(i)}} \frac{1}{\widehat{N}_{F_\alpha}^{(i)}}, \tag{1}$$

where $\alpha$ is a demographic group (e.g. women aged 40-49 in 2012); $M_{\alpha}^{(i)}$ is the death rate in group $\alpha$ and city $i$ ($i \in 1, \ldots 27$); $\widehat{y}_{F,D_\alpha}$ is an estimate of the total number of reported deaths among respondents' personal network members; $\widehat{\bar{d}}_{F_\alpha,F}^{(i)}$ is the estimated average personal network size for people in group $\alpha$, and $\widehat{N}_{F_\alpha}^{(i)}$ is the estimated number of people in group $\alpha$. Details of the network survival method, including how the data are collected and the derivation of the estimator, can be found in Feehan et al. (2017).

We used the network survival method to estimate death rates in the 12 months before the survey for each of the 27 cities in our sample; the solid lines in Figure 2 show the results. Estimated death

rates for males are generally higher than for females, and that estimated death rates tend to increase with age. However, without comparing the estimates to the gold standard, it is difficult to assess how accurate they are. Moreover, some of the patterns in the design-based estimates look unusual; for example, in MG (Belo Horizonte, the capital of the state Minas Gerais), estimated death rates for males seem to decrease across the middle age range. While it is possible that these unusual patterns are the result of interesting trends in mortality, it is also possible that these design-based estimates are noisy and could be improved by modeling. Ultimately, this is an empirical question that can be resolved by comparing design- and model-based estimates to the gold standard.

### 1.4 Model-based estimates

Many different approaches have been proposed for modeling death rates at adult ages; most of these approaches are based on some combination of three big ideas: (1) information can be pooled across groups whose death rates are expected to be related to one another (e.g., Fay and Herriot 1979; Raftery et al. 2013; Schmertmann and Gonzaga 2018; Wakefield et al. 2018); (2) researchers may expect that true, underlying death rates will change smoothly from age to age, with no sharp jumps (De Beer 2012; e.g., Girosi and King 2008; Schmertmann and Gonzaga 2018); (3) researchers may expect that true, underlying death rates should follow the same general pattern that has been previously observed in other places or time periods (e.g., S. J. Clark et al. 2009; Coale and Demeny 1966).

The full paper will explore models that make use of all three ideas. In this extended abstract, we present the results from a preliminary model which makes use of the first idea by hierarchically pooling information for age-sex groups and cities. The preliminary model also accounts for the complex design we used to obtain our sample, using ideas from the literature on small area estimation (Fay and Herriot 1979; J. N. Rao and Molina 2015; Wakefield et al. 2018). Appendix 3.2 describes the preliminary model in detail.

Figure 2 shows estimated age-specific death rates using the model, comparing them to the design-based estimates. The hierarchical pooling in the model-based estimates leads them to be smoother and more similar across cities than the design based estimates.

Design- and model-based estimates can also be compared in terms of precision, or the amount of uncertainty associated with each estimate. Figure 3 compares the standard error of estimates produced for the design-based estimates (based on the bootstrap; x axis) and model-based estimates (from the estimated model posterior; y axis) for one of the modeled quantities: $\bar{y}_{F,D_\alpha}$ (results for $\bar{y}_{F_\alpha,\mathcal{A}}$ are qualitatively similar, but not shown here). Figure 3 reveals that model-based estimates have much smaller uncertainty intervals. However, the extent to which the model-based estimates are more accurate and have better-calibrated uncertainty estimates can only be revealed

by comparing the design and model-based estimates to the gold standard.

### 1.5 Next steps

The final paper will expand the discussion above in several ways. First, we plan investigate more model specifications, trying various combinations of the three big ideas described above. A focus of this model-building will be to assess the models themselves and also to assess techniques for checking and improving the models in the absence of gold-standard estimates.

We also plan to add the sibling history data to our analysis; we plan to produce both design and model-based estimates from the sibling history data.

Finally, after all estimates have been finalized, we plan to evaluate the quality of each approach to estimating death rates by comparing them to the gold standard estimates from the vital registration data. Our results will enrich our understanding of how adult death rates should be estimated in settings affected by the scandal of invisibility.


Figure 1: The cities in our sample: 26 state capitals and the Federal District (Brasilia).

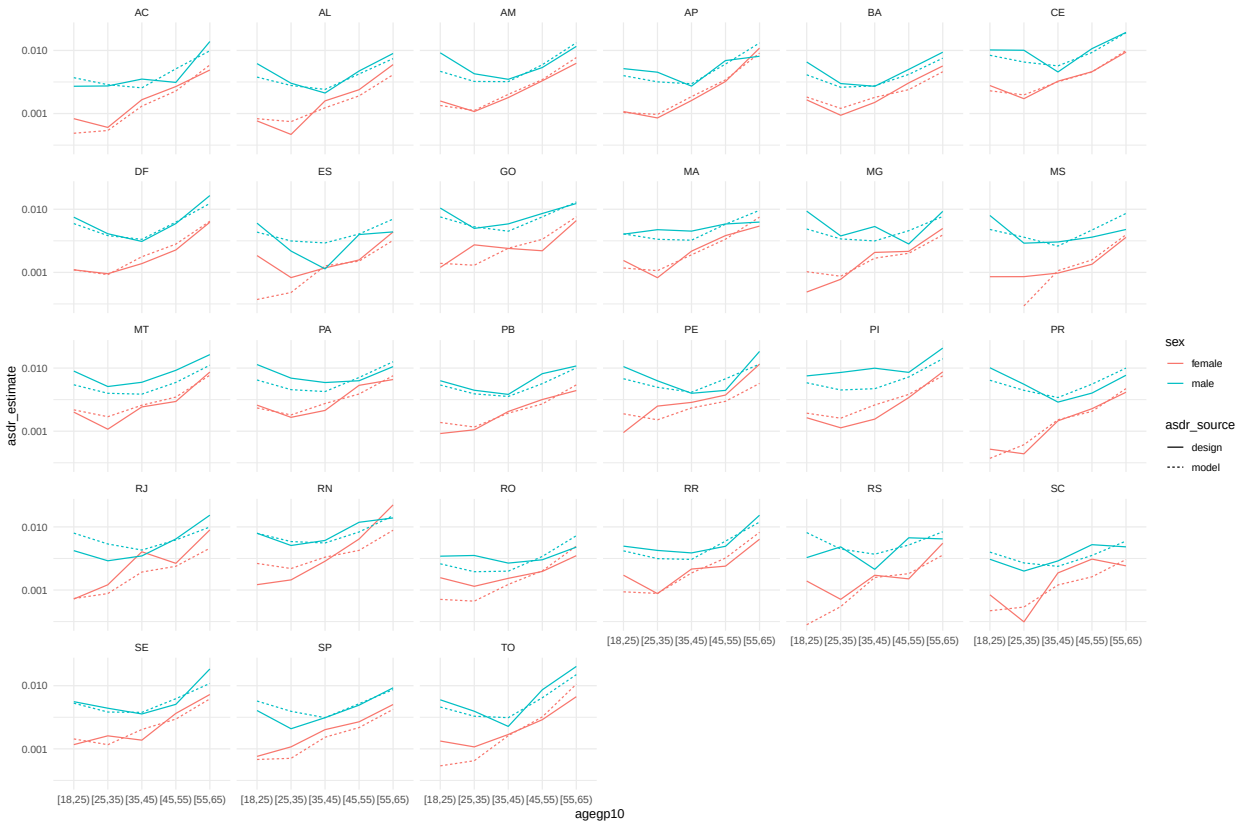


Figure 2: Preliminary design- and model-based estimates for age-specific death rates at adult ages using the network survival method in 27 cities across Brazil. (Note that the y-axis is on a log scale.) Uncertainty estimates are not shown in Figure 2 in order to reduce complexity; Figure 3 compares uncertainty estimates for the two estimators.

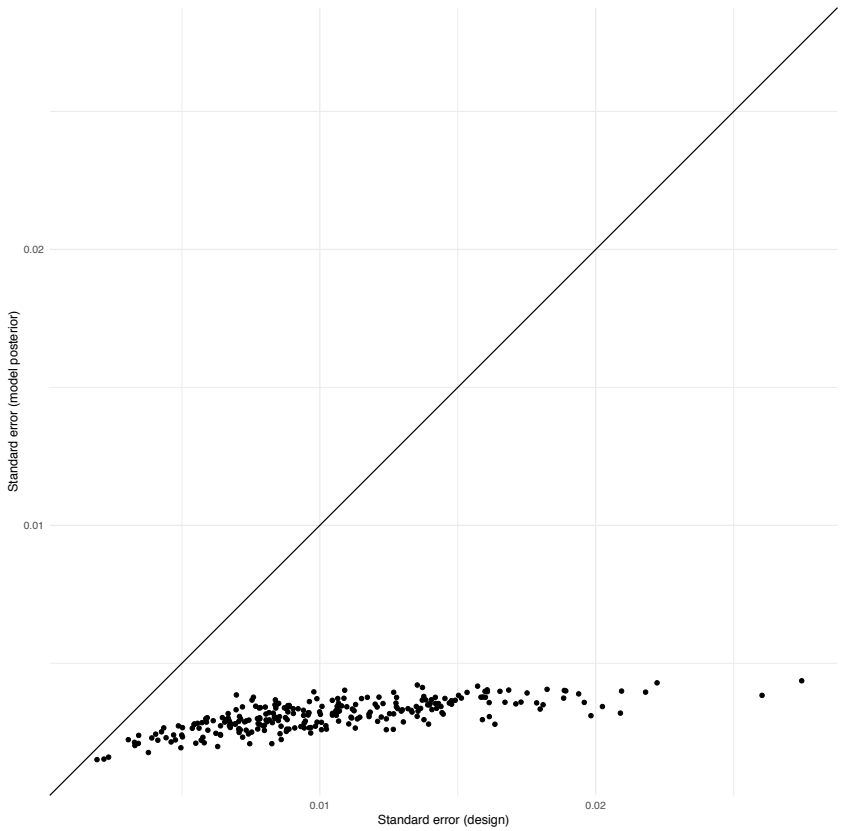


Figure 3: Comparison of estimated standard error for design-based estimates (x axis) and one set of model-based estimates (y axis) for death rates at adult ages using the network survival method in 27 cities across Brazil. Design-based estimates for sampling variance can be obtained from the rescaled boostrap, which accounts for the complex sample design (Feehan and Salganik 2016b; Feehan et al. 2017; J. N. K. Rao and Wu 1988; J. Rao et al. 1992). Model-based uncertainty estimates come from the model posterior.

## 2 References


Brass, W. (1975). Methods for estimating fertility and mortality from limited and defective data. *Methods for estimating fertility and mortality from limited and defective data.*

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

Clark, S. J., Jasseh, M., Punpuing, S., Zulu, E., Bawah, A., & Sankoh, O. (2009). Indepth model life tables 2.0. In *Annual Conference of the Population Association of America (PAA). Population Association of America (PAA).*

Coale, A. J., & Demeny, P. (1966). Regional model life tables and stable populations.

De Beer, J. (2012). Smoothing and projecting age-specific probabilities of death by TOPALS. *Demographic Research*, *27*, 543–592.

Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, *74*(366a), 269–277.

Feehan, D. M., & Salganik, M. J. (2016a). Generalizing the Network Scale-Up Method: A New Estimator for the Size of Hidden Populations. *Sociological Methodology*, *46*(1), 153–186.

Feehan, D. M., & Salganik, M. J. (2016b). *Surveybootstrap: Tools for the Bootstrap with Survey Data.*

Feehan, D. M., Mahy, M., & Salganik, M. J. (2017). The network survival method for estimating adult mortality: Evidence from a survey experiment in Rwanda. *Demography*, *54*(4), 1503–1528.

Girosi, F., & King, G. (2008). *Demographic forecasting.* Princeton University Press.

Graham, W., Brass, W., & Snow, R. W. (1989). Estimating maternal mortality: The sisterhood method. *Studies in Family Planning*, 125–135.

Raftery, A. E., Chunn, J. L., Gerland, P., & Ševčíková, H. (2013). Bayesian Probabilistic Projections of Life Expectancy for All Countries. *Demography*, *50*(3), 777–801.

Rao, J. N. K., & Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, *83*(401), 231–241.

Rao, J. N., & Molina, I. (2015). *Small area estimation.* John Wiley & Sons.

Rao, J., Wu, C., & Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, *18*(2), 209–217.

Rutenberg, N., & Sullivan, J. M. (1991). Direct and indirect estimates of maternal mortality from the sisterhood method. In *Proceedings of the Demographic and Health Surveys World Conference* (Vol. 3, pp. 1669–1696).

Salmon, C. T., & Nichols, J. S. (1983). The next-birthday method of respondent selection. *Public Opinion Quarterly*, *47*(2), 270–276.

Sarndal, C. E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer Verlag.

Schmertmann, C. P., & Gonzaga, M. R. (2018). Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. *Demography*, *55*(4), 1363–1388.

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

Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., & Clark, S. J. (2018). Estimating under-five mortality in space and time in a developing world context. *Statistical Methods in Medical Research*, 0962280218767988.

# 3 Appendix

## 3.1 Sample design

We used a multistage probability design to obtain a sample that is representative of adults over age 18 living in Brazil's 26 state capitals and in Brasilia (Figure 1). The first and second stages of our sample used the Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE), which is a list of the addresses of all Brazilian households that was produced by Brazil's Census agency (IBGE) in 2010. In the first stage, we selected census blocks with probability proportional to size. In the second stage, we used simple random sampling without replacement to select 6 households to be interviewed in each sampled census block. In the third stage, we selected a member of each household to be interviewed by using the next-birthday method: the interviewer listed the birthdays of all household members aged 18 or older, and selected as the respondent the household member whose next birthday was closest to the interview date (Salmon and Nichols 1983). If the selected resident was not at home or could not be interviewed during the study team's first visit, up to two additional visits were scheduled to try and reach the selected respondent. After three unsuccessful attempts, the household was replaced by another one previously selected from the same census block. The sampling design was the same within each of the 27 cities, except that the

target number of census blocks to be sampled in the first stage varied from city to city based on the city's total population.

Each respondent verbally consented to participate in the study and was then interviewed using a questionnaire that contained a basic socio-demographic module as well as the questions necessary to produce network scale-up estimates for the size of the three key populations. All data were collected on paper forms and then entered electronically, subjected to quality checks, and finally deposited into the analysis dataset. Data collection was carried out from April to November of 2012 and produced a final dataset with 24,977 interviews, distributed across the 27 cities.

### 3.2 Preliminary model

In order to model death rates based on network survival data, it is helpful to re-express the network survival estimator (Equation 1) using a different, equivalent expression:

$$\widehat{M}_{\alpha}^{(i)} = \frac{\widehat{y}_{F,D_\alpha}^{(i)}}{\widehat{y}_{F_\alpha,\mathcal{A}}^{(i)}} \frac{N_{\mathcal{A}}^{(i)}}{\widehat{N}_{F}^{(i)}} = \frac{\widehat{\bar{y}}_{F,D_\alpha}^{(i)}}{\widehat{\bar{y}}_{F_\alpha,\mathcal{A}}^{(i)}} \frac{N_{\mathcal{A}}^{(i)}}{\widehat{N}_{F_\alpha}^{(i)}}. \tag{2}$$

Equation 2 is useful because it expresses network survival estimates in terms of two quantities that are averages across survey respondents: $\widehat{\bar{y}}_{F,D_\alpha}^{(i)}$ and $\widehat{\bar{y}}_{F_\alpha,\mathcal{A}}^{(i)}$. We expect these averages to be more reasonable to model than totals, which will be influenced by city size.

We model each of these two average quantities separately; for concreteness, we discuss the model for $\bar{y}_{F,D_\alpha}$, which is the average number of reported connections to deaths in group $\alpha$; the model for $\bar{y}_{F_\alpha,\mathcal{A}}$ is analogous.

We assume that the observed, design-based estimate for the population average number of reported deaths in city $i$, $\widehat{\bar{y}}_{F,D_\alpha}^{(i)}$ is a function of the underlying population parameter $\mu_{D_\alpha}^{(i)}$ and sampling variation given by the design-based standard error $\widehat{\sigma}_{D}^{(i)}$:

$$\widehat{y}_{F,D_\alpha}^{(i)} \sim \text{Normal}(\mu_{D_\alpha}^{(i)}, \widehat{\sigma}_{D}^{(i)}).$$

We use the bootstrap-based estimate for the sampling variation $\widehat{\sigma}_{D}^{(i)}$ of the estimate $\widehat{\bar{y}}_{F,D_\alpha}^{(i)}$; thus, we treat the sampling variation $\widehat{\sigma}_{D}^{(i)}$ as a known parameter in the model.

The remainder of the model specifies a hierarchical relationship between the parameters $\mu_{D,\alpha}^{(i)}$ within a city across age-sex groups, and within an age-sex group across cities:

$$
\begin{aligned}
\mu_{D_\alpha}^{(i)} &= \alpha_0 + \beta_{city[i]}^{C} + \beta_{age-sex[\alpha]}^{AS} \\
\beta_j^C &\sim \text{Normal}(0, \sigma_C) && \text{for } j \in 1 \dots 27 \\
\beta_k^{AS} &\sim \text{Normal}(0, \sigma_{AS}) && \text{for } k \in 1 \dots K
\end{aligned}
$$

where $j$ indexes the 27 city-specific parameters $\beta_j^C$, and $k$ indexes the $K$ age-sex group parameters $\beta_k^{AS}$. Finally, we put diffuse priors on the overall mean $\alpha_0$, and on the two hierarchical variance parameters $\sigma_C$ and $\sigma_{AS}$.

This model can be seen as having two, non-nested random effects: one for the age-sex group, and one for the city. Roughly, estimates are regularized, or shrunk towards a common mean, across cities and across age-sex groups.

Given the model for $y_{F,D_\alpha}^{(i)}$ and an analogous model for $y_{F_\alpha,\mathcal{A}}$, we calculate death rate estimates using

$$
\widehat{M}_\alpha^{(i)} = \frac{\mu_{D_\alpha}^{(i)}}{\mu_{F_\alpha}^{(i)}} \frac{N^{(i)}}{\widehat{N}_{F_\alpha}^{(i)}},
$$

where $\mu_{F_\alpha}^{(i)}$ is the model-based parameter for $y_{F_\alpha,\mathcal{A}}^{(i)}$, analogous to $\mu_{D_\alpha}^{(i)}$ and $y_{F,D_\alpha}^{(i)}$; $N_{\mathcal{A}}^{(i)}$ is known from the study design (i.e., the size of the known populations; see Feehan et al. (2017)); and $\widehat{N}_{F_\alpha}^{(i)}$ is an unmodeled, design-based estimate for the size of the frame population in group $\alpha$ and in city $i$.

We fit the model using the Stan software package (Carpenter et al. 2017).