Multivariate Hierarchical Modelling of Household Air Pollution

Oliver Stoner[1], Gavin Shaddick[1], Theo Economou[1], Sophie Gumy[2], Jessica Lewis[2], Itzel Lucio[2], Giulia Ruggeri[2], Heather Adair-Rohani[2]

[1] University of Exeter, United Kingdom
[2] World Health Organization, Switzerland

E-mail for correspondence: `ors203@exeter.ac.uk`

Abstract: Exposure to household air pollution has been attributed to an estimated 3.8 million deaths per year. A major contributor to this exposure is the reliance on various polluting fuels for cooking by almost half of all households in low and middle-income countries. We present a multivariate hierarchical model for surveys of the proportion of people relying on each fuel type, for the period 1990-2017, addressing several challenges with modelling the data including incomplete surveys and sampling bias.

Keywords: Multivariate proportion data; Bayesian methods; Sampling bias.

1 Introduction

Information on the proportion of people in each country relying primarily on each fuel for cooking is available in the form of nationally-representative household surveys. These surveys are collated in the World Health Organization (WHO) Household Energy Database. Statistical modelling of this data can be employed to estimate trends in fuel use from survey variability, to make predictions in countries and years with no surveys, and to forecast future fuel use.

Previous approaches to modelling this data, most notably Bonjour et al. (2013), have focussed on the overall proportion of people relying on any of wood, charcoal, coal, crop waste and dung, classified as solid fuels. This inhibits policy related to the use of specific fuels, such as the deployment of cleaner wood-burning stoves, and fails to take into account the different levels of harm caused by different fuels. For example, in Sub-Saharan Africa

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), Guimarães, Portugal, 7–12 July 2019. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

much of the population is switching from biomass fuels such as wood to charcoal, a change which has significant health implications but may not be detectable when only looking at overall solid fuel use. Instead, we present a multivariate hierarchical model for the use of eight fuel types (wood, crop waste, dung, charcoal, coal, kerosene, gaseous fuels and electricity).

2 Methodology

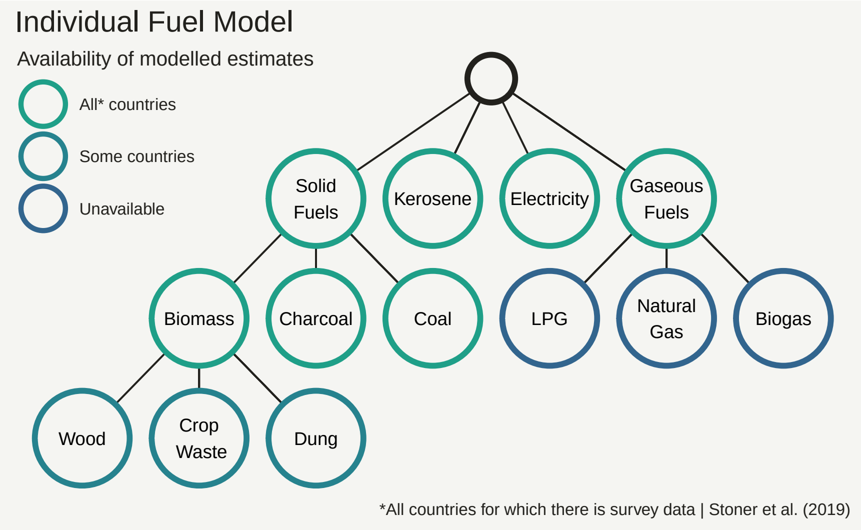

FIGURE 1. Chart illustrating the current fuel type hierarchy and output availability of the individual fuel model.

The model we employ is a version of the multivariate hierarchical model presented Stoner et al. (2019). Many surveys combine different fuel types into one option. For example, surveys often list ‘gas’ as an option, which may reflect the use of either LPG, natural gas or biogas. This aggregation means that the time series of survey values for certain individual fuels, such as LPG, is highly unstable in some countries, which makes it challenging to estimate trends. To address this, we have developed a tiered approach, illustrated in Figure 1, where some individual fuels are combined at the top level of the model, to be disaggregated at lower tiers. Specifically, we model the vector $\mathbf{y}$ of survey respondents relying primarily on any solid fuel (y_1), kerosene (y_2), any gaseous fuel (y_3) or electricity (y_4) as an imperfect sample of the population, arising from the Generalized-Dirichlet-Multinomial (GDM) family of distributions. Then, the vector of respondents using any biomass fuel, charcoal or coal is also modelled as GDM, out of all those using solid fuels, and so on. This ensures that any confusion or combination of fuels in the lower tiers, which is cancelled out when the affected fuels are aggregated, does not affect trend estimates for fuels in higher tiers.

From this point onwards, we focus only on the top tier GDM model, as the models for the lower tiers are identical. Many surveys do not provide values for all of the individual fuels, so the model must predict the missing fuels based on the values for the observed fuels. To do this, we implement the model using the implicit Beta-Binomial conditional distributions derived from the GDM:

$$\begin{aligned} y_1 &\sim \text{Beta-Binomial}(\nu_1, \phi_1, n) & (1) \\ y_i \mid y_1, \ldots, y_{i-1} &\sim \text{Beta-Binomial}(\nu_i, \phi_i, n - \sum_{j=1}^{i-1} y_j) & (2) \end{aligned}$$

Here n is the number of respondents in the survey, ν_i is the expected proportion of respondents using fuel i of those who are not using any of the fuels $\{1, ..., i-1\}$ and ϕ_i determines the variance around this expected proportion.

For a survey conducted in country c, area j (urban or rural) and year t, the relative mean for fuel i ($\nu_{i,j,c,t}$) is modelled by:

$$\log\left(\frac{\nu_{i,j,c,t}}{1-\nu_{i,j,c,t}}\right) = f_{i,j,c}(t) \qquad (3)$$

where each $f_{i,j,c}(t)$ is a smooth function of time with a thin-plate spline basis (Wood, 2016). This affords a high degree flexibility in capturing non-linear and non-monotonic trends in fuel use, as will be highlighted in Section 3. To avoid over-fitting, each spline is penalized for smoothness by a parameter $\lambda_{i,j,c,t}$. The degree of smoothing is be controlled by means of an informative prior distribution for the smoothing parameters.

Whilst most surveys are divided into urban and rural populations, some surveys only report an overall value for the whole population of a country. For these surveys to inform the urban and rural trends, we incorporate a layer in the model to constrain the marginal mean proportions as follows:

$$\begin{aligned} \boldsymbol{\mu}_{i,j,c,t}^{overall} &= \pi_{c,t}\boldsymbol{\mu}_{i,j,c,t}^{urban} + (1-\pi_{c,t})\boldsymbol{\mu}_{i,j,c,t}^{rural} & (4) \\ \log\left(\frac{\pi_{c,t}}{1-\pi_{c,t}}\right) &= \log\left(\frac{P_{c,t}}{1-P_{c,t}}\right) + g_c(t) & (5) \end{aligned}$$

The weights $\pi_{c,t} \in (0,1)$ represent the expected proportion of survey respondents living in an urban area. United Nations estimates of the proportion of people living in an urban area $P_{c,t}$ are used as offsets in a model for $\pi_{c,t}$. Systematic deviations from these estimates are modelled using penalized thin-plate splines $g_c(t)$, to allow for potential under- or over-sampling of urban populations in the survey data.

Finally, whilst the WHO Household Energy Database is constantly improving, there are some recorded survey values which truly defy the trend for a given country, to the extend that modelled estimates are substantially skewed. To address this, we incorporate a layer in the model which mixes each Beta-Binomial with a discrete uniform distribution. The extent of mixing is determined by a different parameter ρ for each survey, where large values of ρ correspond to a greater contribution from the uniform to the likelihood. This effectively allows the model to decide if a survey is overwhelmingly different to other data in the same country and area, subject to a very strong prior distribution for each ρ, assuming that each survey is very unlikely to be an outlier. This greatly increases the model's robustness to outliers.

The model is implemented using `nimble`, a package for flexible implementations of Markov Chain Monte Carlo (MCMC). Further details and model checking can be found in Stoner et al. (2019).

3 Results

The model is applied to over 1100 surveys from the WHO Household Energy Database, covering more than 150 countries globally, and is used to predict fuel use trends for each country. For example, Figure 2 shows the median predicted proportion using each fuel in urban and rural areas of Bangladesh, with 95% survey prediction intervals. The model is capable of capturing well differing levels of survey variability between fuels, areas and countries. Here this is evident in the disparate widths of the prediction intervals between, for example, the use of gaseous fuels in urban and rural areas.

Similarly, Figure 3 shows the predicted fuel trends in Indonesia. Here, the advantage of using splines is clear from the way the model captures nonlinear trends in the use of gaseous fuels with ease.

The model also provides information on systematic trends in the sampling of urban and rural respondents. Figure 4 shows the model's prediction of mean trends in the proportion of urban respondents for India (left) and Malawi (right). The model estimates that there is systematic oversampling of urban respondents in India (compared to U.N. estimates of the true urban proportion), while the model estimates very little systematic deviation in Malawi.

4 Summary

We have discussed the need for a multivariate predictive model for the reliance on individual fuels, to allow for better informed policy relating to the use of specific fuels. We have presented a highly flexible hierarchical model, applied to surveys in the period 1990-2017 contained within the WHO Household Energy Database, where trends in the use of individual fuels are modelled jointly whilst taking into account sampling bias, incomplete surveys and probable outliers. As well as predicting trends in fuel use

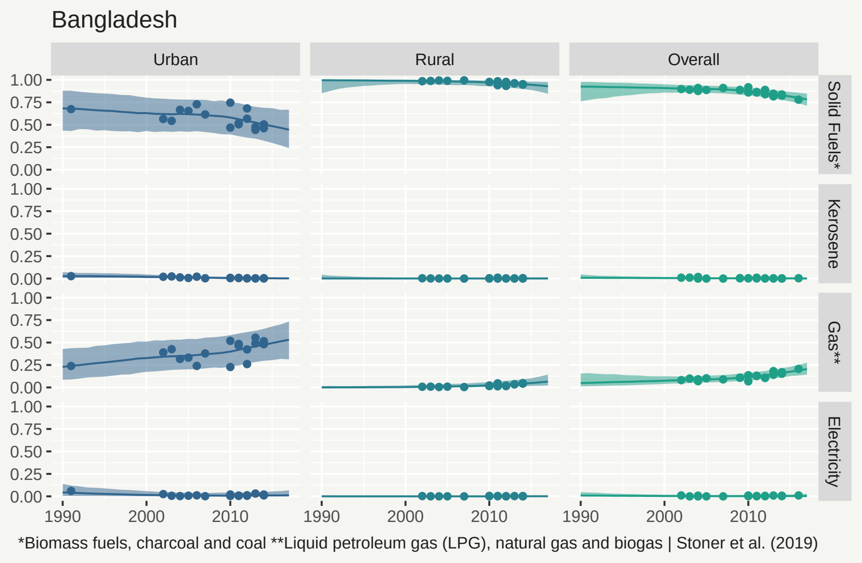

FIGURE 2. Median predicted fuel usage with associated 95% posterior predictive intervals for Bangladesh, 1990-2017.

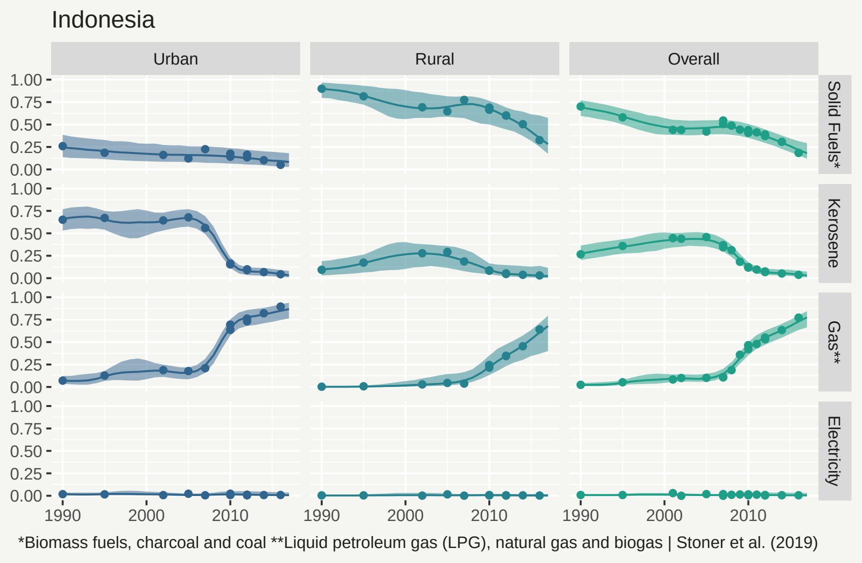

FIGURE 3. Median predicted fuel usage with associated 95% posterior predictive intervals for Indonesia, 1990-2017.

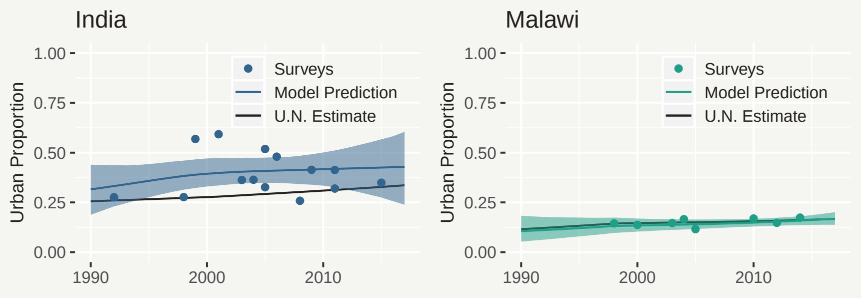

FIGURE 4. Plot of the urban proportions of fuel survey respondents in India (left) and Malawi (right), shown as points, compared to the U.N. estimates of the proportion of the respective populations and the model's predictions, with 95% credible intervals.

for each country, the model has been adopted as a key tool for monitoring progress towards UN Sustainable Development Goal 7.1, to 'ensure universal access to affordable, reliable and modern energy services' by 2030. Furthermore, predictions from the model will form the basis of future work in quantifying the burden of disease caused by exposure to each fuel type.

Acknowledgments: This work was supported by a NERC GW4+ Doctoral Training Partnership studentship from the Natural Environment Research Council [NE/L002434/1] and the World Health Organization contract APW 201790695.

References

Bonjour, S., Adair-Rohani, H., Wolf, J., Bruce, N., Mehta, S., Prüss-Ustün, Lahiff, M. Rehfuess, E., Mishra V. and Smith, K. (2013). Solid Fuel Use for Household Cooking: Country and Regional Estimates for 1980-2010 *Environmental Health Perspectives*, **121(7)**, 784 – 790.

Stoner, O., Shaddick, G., Economou, T., Gumy, S., Lewis, J., Lucio, I., Ruggeri, G., and Adair-Rohani, H. (Submitted 2019). Estimating Household Air Pollution: A Multivariate Hierarchical Model for the Use of Polluting Fuels for Cooking. https://arxiv.org/abs/1901.02791

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Lang, D. and Bodik, R. (2017) Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE, Journal of Computational and Graphical Statistics, **26(2)**, 403 – 413.

Wood, S. (2016). Just Another Gibbs Additive Modeler: Interfacing JAGS and mgcv *Journal of Statistical Software, Articles*, **75(7)**, 1 – 15.