Improving and Validating Survey Estimates of Religious Demography Using Bayesian Multilevel Models and Poststratification

Christopher Claassen* Richard Traunmüller†

December 8, 2017

Forthcoming in *Sociological Methods and Research*

*School of Social and Political Sciences, University of Glasgow
†Department of Social Sciences, Goethe University Frankfurt

Abstract

Religious group size, demographic composition, and the dynamics thereof are of interest in many areas of social science, including migration, social cohesion, parties and voting, and violent conflict. Existing estimates however are of varying and perhaps poor quality because many countries do not collect official data on religious identity. We propose a method for accurately measuring religious group demographics using existing survey data: Bayesian multilevel regression models with post-stratification, or MRP. We illustrate this method by estimating the demography of Muslims, Hindus, and Jews in Great Britain over a 20-year period, and validate it by comparing our estimates to UK census data on religious demography. Our estimates are very accurate, differing from true population proportions by as little as 0.29 (Muslim) to 0.04 (Jewish) percentage points. These findings have implications for the measurement of religious demography as well as small group attributes more generally.

Keywords: Religion, United Kingdom, demography, survey methods, multilevel regression with post-stratification (MRP), Bayesian methods.

1. Introduction

The size of religious groups, their demographic composition, and their dynamics over time are of interest in many areas of the social sciences, including social cohesion and change (DiPrete et al 2011; Putnam and Campbell 2010; Voas et al 2002; Voas and Chaves 2016), migration (Geddes 2003; Levitt 2003), parties and voting (Gelman et al 2008; Lipset and Rokkan 1967), and violent conflict (Fearon and Laitin 2003; Toft et al 2011). Religious demographic change is, if anything, even more of interest outside the academy, especially in Europe, where religious diversity is increasing rapidly and religion now plays a central role in vexing political debates around migration, citizenship, and identity (Adida et al 2016).

Yet, because of severe data limitations, we know surprisingly little about the demography of small religious minorities, both in Europe and beyond. For a variety of historical and political reasons, official census data on religious identity are gathered only in a few countries and are entirely lacking for major cases such as the United States, France, Germany, or the Netherlands.[1] While there have been a number of recent efforts to provide global measures of religious diversity (Brown and Patrick 2015; Johnson and Grim 2013; Maoz and Henderson 2013), these projects combine data of varying quality, often in an *ad hoc* fashion. They also only provide the marginal distributions of religious identity, rather than the full joint distributions. In other words, they provide no information regarding the socio-demographic composition of small religious groups, such as their age-structure, sex ratio, or socio-economic resources.

In this paper, we propose applying Bayesian multilevel modeling and post-stratification (MRP; Gelman and Little 1997; Park et al 2004) to existing survey data to measure the demography of small religious groups. Tighe et al (2010) pioneer the use of this method for estimating religious demography but focus only on the American Jewish population and only at a single point in time. Moreover, due to a lack of US census data on religious affiliation,

[1] Only 14 states among the EU27 countries include such a question in their most recent censuses (Johnson and Grim 2013).

Tighe and colleagues were not able to validate their estimates against a true benchmark. It thus remains unknown whether, and to what degree, MRP can help us measure religious demography.

The contribution of this paper is the provision of a dispositive test. To do so we estimate the size of three religious minorities – Muslims, Hindus, and Jews – in Great Britain[2] over a 20 year period. In addition to estimating the overall prevalence and dynamics of these religious identities, we also estimate their prevalence within age by gender by education subgroups. As such, we offer three contributions beyond the work of Tighe et al (2010). First, by focusing on three quite different groups, we provide a more general test of the method. In particular, while American (and European) Jewish populations are small, they are well-established groups that should pose no special sampling challenges. Second, we attempt to also estimate population dynamics, an especially challenging task for rapidly growing groups like British Muslims and Hindus. Finally, and most importantly, the UK government included questions on religious identity in the 2001 and 2011 censuses, allowing us to determine the accuracy of our estimates and the method we use.

Our results suggest that considerable optimism is warranted regarding the use of MRP applied to existing survey data to measure religious demography. Our estimates differ from the census estimates of the population prevalence of the three religious minorities by only .29 (Muslim), .14 (Hindu), and .04 (Jewish) percentage points on average.[3] Our estimates continue to be accurate, although somewhat less so, when we turn to the more arduous task of measuring the prevalence of the three religious identities within age by gender by education subgroups. Here, our Muslim estimates differ, on average, from the census values by .58 percentage points; Hindu estimates, by .22 percentage points; and Jewish estimates, by .11 percentage points.

The findings of our paper also have important implications that go beyond the study of

[2]We focus only on England, Wales, and Scotland for reasons we outline later.

[3]These are the estimates obtained from our preferred models, which are selected according to their accuracy and efficiency in estimating religious group prevalence within 32 demographic subgroups. Our most accurate estimates of overall size of religious groups are in fact even closer to the census estimates.

religious demography. Given the abundance of existing survey data, it is in principle possible to study small socio-demographic groups and their dynamics in considerable detail, even in the absence of census data. We envision that this approach will be immensely valuable for estimating the prevalence of various small-group attributes. This includes estimating the size of religious and other social groups, but also perhaps the prevalence of health and other behaviors.

The paper is structured as follows. We discuss existing approaches to measuring religious demography, the specific problems in using small-sample surveys to do so, and the potential that MRP holds in addressing these problems. We then describe our data and modeling approach before presenting and describing the results of our tests. We conclude with some advice to scholars interested in using this method to measure the size of religious or other demographic minorities.

2. Existing Measures of Religious Demography

Both researchers and policy-makers require accurate measures of religious demography and diversity. A number of data-collection projects have arisen to meet this demand, including the World Religion Database (Johnson and Grim 2013), the World Religion Dataset (Maoz and Henderson 2013), and the Religious Characteristics of States data set (Brown and Patrick 2015). Although the scope and comprehensiveness of these databases are admirable, and while they provide perhaps the only source of data for some regions and periods of time, there are nevertheless a number of limitations with their estimates.

First, they incorporate data of varying quality. In some countries census or large-sample survey data are used, while smaller surveys are relied upon for others. Where neither of these exist, or they are out of date, subjective expert opinions inform the estimates (Johnson and Grim 2013). Although these databases rightly respect the adage that some data is preferable to none at all, we have no way of ascertaining the degree of uncertainty attached

to any particular estimate because none are provided.[4] Without uncertainty estimates, analysts are led to treat census measures and expert opinions as equally valid.

Second, the methods used to adjust sample survey data, combine data, and obtain estimates when no data are available, are less than fully transparent. Adjusting, combining, interpolating, and extrapolating data requires modeling. Yet neither the assumptions underlying the model nor the exact methods for doing so are fully specified. In addition, the uncertainty induced by modeling is again ignored.

Finally, these project provide only the marginal, rather than also the joint demographic-religious distributions. Joint distributions are useful for at least three reasons. First, social scientific arguments on inter-group relations refer not only to uni-dimensional concepts based on mere group size (e.g. fractionalization) but on multi-dimensional concepts, such as religion by ethnicity or socio-economic status (e.g. cross-cutting cleavages; Selway 2011). Second, demographers require detailed information on the sex-specific age-structure of religious groups for information on religious group size to be useful in in population projections. Finally, religious-demographic joint distributions can be used to adjust or post-stratify survey opinions from unrepresentative samples of religious minorities.

3. Measuring Religious Demography Using Sample Surveys

We propose a method for measuring the size, demographic composition, and dynamics of religious minorities using secondary survey data. While official data on religious identity is rare, public opinion survey data is abundant. The vast and growing cache of social-scientific public opinion survey data provides a potentially rich source of data for measuring religious demography. Nationally-representative surveys that include questions on respondents' religious identity are now collected by a diverse range of agencies and organizations and for a diverse range of research projects. Thousands of such surveys have been fielded in dozens of

[4]Grim and Hsu (2011) develop a "data quality index" for religious demography data which rests on four criteria: geographic coverage, response rate, sampling, and questionnaire design but do not use this index to provide uncertainty estimates for any of these projects.

countries over the last two to three decades.

Two broad approaches to drawing population inferences from survey samples can be distinguished in the literature: a model-based and a design-based approach (e.g. Binder and Roberts 2009; Sterber 2009). The model-based approach involves building a model that is thought to capture the process by which the sample was collected and its attributes measured. The design-based approach, in contrast, restricts its attention to the sampling design and other methods used to gather the sample. Although the design-based approach has the appearance of rigor, it may prove to be inflexible and biased in practice.

In particular, there are two difficulties in using the design-based approach to measure religious demography using existing sample surveys. First, religious minorities are small groups – less than one percent of national populations in the case of European Jews. Standard public opinion survey samples are thus likely to include only tiny samples of religious minorities. Such small samples produce highly variable estimates.

A typical design-based response to this challenge is to combine survey data in some fashion. One method for doing so is pooling, which involves combining micro-level survey data, effectively creating a "mega-poll" with tens or hundreds of thousands of respondents, and then estimating quantities of interest (Korn and Graubard 1999; Roberts and Binder 2009; Thomas and Wannell 2009). An alternative method is to estimate the quantities of interest separately from each survey, before then combining the separate estimates.

Either method of combining surveys is a reasonable solution for estimating the demographic size and composition of static groups such as American or European Jews (e.g. Smith 2005; Tighe et al 2010). However, for rapidly growing groups such as European Muslims, neither pooling nor combining separate analyses are ideal, as both would produce biased estimates. In addition, neither method allows analysts to investigate the dynamics of the group's demography, which is frequently a topic of interest.

The second difficulty in using design-based approaches to extract estimates of religious diversity from sample surveys is that these samples are not, in fact, collected exactly as

mandated by the sampling design. Response rates are falling (Keeter et al 2006), even in well-funded, high quality national survey projects. For example, the UK sample of the European Social Survey saw response rates drop from 55.5% in 2002 to 43.6% in 2014, while the German sample faced a decline from 55.7% to 31.4%.[5]

To compound the problem, samples of religious minorities are particularly likely to be unrepresentative. Such individuals are more likely to have ethnic minority status and be of immigrant background, and less likely to speak the national language. As such, religious minorities constitute groups that are fundamentally more difficult to sample than majority populations (Font and Méndez 2013).

In sum, although decades of survey data exist on religious identity across virtually all European countries (and beyond), their samples of religious minorities are likely to be very small and additionally unrepresentative. Traditional design-based approaches to population inference are thus quite limited when it comes to measuring religious demography.

Instead, we propose modeling the survey data. In particular, we will use methods originally designed for measuring the attributes of small geographic units, known as small area estimation (SAE, e.g., Rao 2013). Although areal units – unlike demographic groups – have a spatial structure, there is considerable overlap in the problems posed by small area and small group estimation. Most notably, once one disaggregates survey data by group or areal unit, one typically finds small and unrepresentative samples. SAE methods address these issues by using regression models to "borrow strength" across time and areal units. They also often incorporate supplementary census or administrative data to improve estimates. We propose doing the same for small demographic groups.

Of particular interest is the method of Multilevel Regression with Post-stratification, or MRP (Gelman and Little 1997; Park et al 2004). MRP involves, firstly, multilevel modeling of some survey response using a set of geographic and demographic categories, such as areal unit, gender, and age. Secondly, the analyst then uses the model to predict the prevalence of

[5]See `http://www.europeansocialsurvey.org/data/deviations_7.html`.

that opinion within each cell of the joint demographic distribution (i.e. areal unit by gender by age); thirdly, weighs each of these predicted cells by its population size; and finally, aggregating the weighted estimates to the areal unit of interest.[6]

Although analysts could in principle use classical linear or generalized linear models for estimating the population attributes of small groups (or areas), the use of multilevel models offers four advantages. First, multilevel models allows estimates to be "partially pooled", or smoothed, across areal units and demographic groups to the extent that sample sizes in these subgroups are small (Gelman and Hill 2007). Partial pooling reduces the degree to which model predictions are affected by outlying observations and also provides for more efficient uncertainty estimates. It further allows for "deep" interactions among demographic-areal categories, which can lead to unreliable estimates in classical regression models (Ghitza and Gelman 2013). Second, because the multilevel framework includes higher-level models for the areal unit of interest, area-level covariates can be added. These allow for more accurate estimates by incorporating additional non-survey information (Warshaw and Rodden 2012). Third, when combining data from several survey projects, MRP can adjust for the project-specific differences in methodology, such as sampling frames and interview modes. Although classical regression methods can include project fixed effects, it is not straightforward to then produce opinion estimates within small areas because the effect of the omitted project category becomes embedded within the model intercept. Fourth, the post-stratification step uses census or other official data to adjust for possible non-representativeness among small area subsamples.[7]

[6]There is now an extensive literature on MRP. See, for example, Hanretty, Lauderdale, and Vivyan (2016); Lax and Phillips (2009); Park et al (2004); Selb and Munzert (2011); Warshaw and Rodden (2012); Leemann and Wasserfallen (2017)

[7]Although multilevel models are typically estimated using maximum likelihood or restricted maximum likelihood (e.g., Lax and Phillips 2009), Bayesian MCMC methods of estimation are believed to be preferable because they provide more accurate, and more readily accessible, measures of inferential uncertainty (Gill 2008; Park et al 2004; Tighe et al 2010).

4. The Present Study

Although MRP has been typically used to measure public opinion within areal units, Tighe et al (2010) demonstrate its utility for measuring religious demography. They apply MRP models to a pooled dataset of 50 surveys, finding that American Jews form 1.86% of the population of the United States. They go further than these marginal distributions by also estimating joint distributions: the prevalence of Jewish religious identity within demographic subgroups. Yet, although their estimates appear to be quite reasonable, Tighe and colleagues do not test their accuracy. Indeed, with such a wealth of survey data, it could well be the case that any method – even a simple mean of the raw data – would come close to the true population proportion.

We follow the example of Tighe et al (2010) in using MRP to measure the size and demographic composition of religious minorities. In contrast, however, we seek to test the accuracy of the method by comparing our estimates with official data on religious group size and demography extracted in the 2001 and 2011 UK censuses. This will not only allow us to determine the degree to which our estimates are accurate, but will also provide some guidance to other scholars interested in measuring religious demography using using existing survey data.

Our paper differs from Tighe et al (2010) in two additional ways. First, we aim to measure the size of several religious minorities, not just one: Muslims and Hindus, as well as Jews. Second, we also attempt to measure the dynamics of these three groups' demography over 20 years. While British (and American) Jewish populations are small,[8] they are long-established and demographically stable groups. Muslim and Hindu Britons, although slightly larger, are rapidly growing groups. Muslims, for example, increased from 2.8 to 4.8% of the UK population between 2001 and 2011, with Hindus showing an increase from 1.0 to 1.5%. In addition, survey data on Muslim Britons, in particular, is likely to

[8]According to the 2011 census, Jews formed 0.5% of the UK population.

suffer from poor sampling given that 92% of this group report ethnic minority status in 2011, 53% report being born outside the UK, and 48% are younger than 25 years of age. Thus, estimating the religious demography of these three groups over 20 years epitomizes the difficulties we have highlighted in measuring religious demography using survey data, and thus provide a stern, and hopefully generalizable test, regarding the utility of our method.

We estimate and test eight Bayesian multilevel models of Muslim, Jewish, and Hindu prevalence, thus providing eight sets of estimates for each group. Models all include a time component, but vary in the number and complexity of the demographic predictors, and whether the time effect is allowed to vary by demographic group. In addition to the simple additive model, we include a second model with all two-way demographic interactions, a third model where the effects of time vary by demographics, and a fourth model incorporating both two-way demographic interactions and varying time effects. Each of these four models will be tested both with and without ethnicity, producing eight models in total. Although the inclusion of ethnicity might be expected to increase the accuracy of estimates, it comes at a cost: the loss of survey data from projects where questions on ethnicity are not asked of respondents.

Because the method of MRP is both computationally challenging and requires a fair amount of data manipulation, we additionally model the data using a set of eight simple logit regressions. Each of these models corresponds as closely as possible to one of the eight multilevel models, with the exception that demographic factors are estimated using classical, non-multilevel methods, and the post-stratification weighting is done using a vector of survey weights – as one might find in an off-the-shelf survey dataset. These classical regressions with survey weights, which we refer to as CRSW estimates, offer a quick and convenient alternative for measuring demographics using survey data.

Finally, we also estimate religious group demographics by pooling the unweighted survey data before disaggregating by year and religious identity. We refer to these as "disaggregated" estimates.

5. Data and Methods

5.1. Data

We extract individual-level public opinion data from the UK components of three major cross-national survey projects: the International Social Survey Program (ISSP) and, in particular, its British component, the British Social Attitudes Survey (BSAS); the Eurobarometer (EB); and the European Social Survey (ESS).[9] We include all BSAS, ESS, and EB surveys that ask respondents their religious identity, and that were conducted between 1995 and 2014, both inclusive. Within these constraints, we obtain 20 survey samples from the BSAS, 15 from the EB and three from the ESS, with 91,862 respondents in total. These are summarized in Table 1. To validate our estimates, we use the publicly available five percent census samples for the 2001 and 2011 censuses.[10]

It is important for our purposes that these various sources of data have comparable religious identity questions and, indeed, comparable response sets. Yet as Table 2 indicates, although the three questions used in the different survey projects are virtually identical, the census question differs slightly. There is also some variation in response sets: while the census only includes one "Christian" response, all three survey projects allow respondents to choose among Christian denominations.

The UK Office of National Statistics (ONS) has tested whether these variations produce different results (Office for National Statistics 2009). They found significantly different percentages for the "Christian", "Sikh", and "No religion" responses when comparing the census and BSAS questions and response sets. Fortunately, these are not the groups we are fundamentally interested in. Nevertheless, to avoid bias, we do not attempt to estimate the size and demographics of all religious groups, but, rather, we treat Muslim, Jewish, and Hindu as separate categories and collapse all other responses, including refusals and "don't

[9]In order to not limit the generalizability of our results, we only use data from cross-national projects and do not use UK-specific data sources such as the UK Labour Force Survey.

[10]Census 2001: Small Area Microdata; Census 2011: Microdata Individual Safeguarded Sample.

knows," into an "Other" category.[11] We treat non-response in both census and survey data as a separate response category rather than removing these observations.

In addition, the public release of the Northern Ireland census data does not include detailed breakdowns of religious minorities. We therefore exclude Northern Ireland from all the census and survey data we collect. We also remove people younger than 16 from the census results because such respondents are excluded, by design, from the surveys we examine.[12] Our data and estimates are thus representative of the population aged 16 and older and living in mainland Britain (England, Wales, and Scotland).

Finally, we also extract data on several demographic factors from both the surveys and census datasets (more details in the next sub-section). These will be used as categorical predictors in our models and to post-stratify the resulting estimates. Non-response and "don't know" responses were coded as missing values, and all respondents with at least one missing value for these variables (5,162; 5.6%) were removed from the dataset, leaving a final sample size of 86,664 respondents.[13]

5.2. Model Specification

MRP entails predicting and post-stratifying the survey response of interest within demographic subgroups.[14] We selected these demographic variables using five criteria. First, we searched for variables which were available in both our survey and census data. We were interested in including both design-based variables, relating to the methods by which the survey samples were selected (such as household size and geographic region), and model-

[11] The census question is optional, so non-response does occur. This actually makes the census data more comparable to the survey data, where, of course, all questions are optional.

[12] While most of these surveys sample those 18 years of age or older, there are some 16 and 17 year olds in our survey dataset. Our choice of age categorization is also constrained by the fact that the publicly available census data is released with age already categorized (for example, 16-19).

[13] See the supplementary materials for further details on the type of coding and degree of missingness for each variable.

[14] As we have already noted, MRP has traditionally been used to estimate survey-based attributes (usually opinions) within small geographic areas. We follow Tighe et al (2010) in applying it to the estimation of demographic attributes within small demographic groups. As such, we do not focus on, or even include, small areal units in our models.

based variables, associated either with survey non-response or with respondents' selection of a religious minority identity (such as age and ethnicity). Next, we examined the degree to which the survey data were unbalanced compared to the census data on each of the demographic factors. Third, we examine the bivariate associations between demographic variables and each religious identity. Fourth, we also consider the extent to which survey data were missing for the variable in question. Finally, since our goal is to produce a model that might be used in settings where no official data on religion exist, we favored as few demographics, and thus as simple a model as possible. Using these criteria, we arrive at a basic set of four demographic variables that will be used in the regressions and the post-stratification step: gender (male and female); age (16-29, 30-49, 50-64, and 65 and older); education (degree or no degree); and household size (1, 2-4, and 5 or more adults).[15]

To address the changes in religious demography, we extend the MRP method by additionally modeling population growth (or decline). Given the deterministic nature of population dynamics, we find that a simple linear trend model effectively captures the dynamics of our religious group populations.

We are now in a position to describe our basic model more formally. Survey responses are coded as 1 if the respondent selects the particular religious identity in question (e.g. Muslim), 0 otherwise. The resulting variable is modeled as a Bernoulli,

$$y_i \sim \text{Bernoulli}(\pi_i)$$

with the probability of selecting a particular (e.g. Muslim) identity then being a function of a multilevel logit regression (where we use the notation of Gelman and Hill (2007) in referring to individual survey respondents i being nested within g gender groups or r age groups, for example):

$$\pi_i = \text{logit}^{-1}\left(\alpha + \beta \text{year}_{t[i]} + \mu^{gen}_{g[i]} + \mu^{age}_{r[i]} + \mu^{hhs}_{h[i]} + \mu^{edu}_{d[i]} + \mu^{proj}_{p[i]}\right).$$

[15]We provide further details on this process in the online supplementary materials.

Population dynamics are specified using a linear trend, captured with coefficient β.[16] The demographic grouping variables of gender, age, household size, education, and survey project are then modeled as random effects drawn from normal distributions with variances to be estimated from the data:

$$\mu^{gen}_{g[i]} \sim N(0, \sigma^2_{gen})$$
$$\mu^{age}_{r[i]} \sim N(0, \sigma^2_{age})$$
$$\mu^{hhs}_{h[i]} \sim N(0, \sigma^2_{hhs})$$
$$\mu^{edu}_{d[i]} \sim N(0, \sigma^2_{edu})$$
$$\mu^{proj}_{p[i]} \sim N(0, \sigma^2_{proj}).$$

Survey project is included as a grouping variable but we make no use of its parameter estimates when estimating predicted effects. Rather, we allow our model to partial out the particular effects of each survey project, leaving us with survey project-adjusted estimates.[17]

We can easily add further complexities to this basic multilevel model. We include three additional features: ethnicity as a grouping factor, two-way interactions among all demographic factors, and varying dynamic effects by (possibly interacting) demographics. Allowing for all combinations of these three features, we have eight models in total.

Ethnicity. We include an additional demographic category, ethnicity (white and non-white), in four of our eight models. Data on respondents' ethnicity are not available for a full 17% of the observations in our pooled dataset, mainly because this variable is not included on the Eurobarometer questionnaire. However, the strong relationship we observe between ethnicity and Hindu and Muslim identity[18] suggests that is worth considering whether including ethnicity increases predictive power enough to offset the negative effects of data

[16] The 20 year time period is standardized to range from 0 to 1.

[17] We exploit the fact that the set of project intercepts is modeled with mean of 0. By leaving out the project effects when using our model to predict, we in effect partial out the effects of survey project on religious minority identity.

[18] See the supplementary materials for evidence.

loss.[19] Model 2 extends Model 1 by including varying intercepts for ethnicity $\mu^{ethnic}_{n[i]}$.

Demographic interactions. We specify two models where we incorporate all two-way interactions between demographic categories, one with ethnicity included, and another, without. These models allow the relationships between demographic predictors and religious identity to depend on other demographic predictors. While such interactions are possible within a classical regression framework, they may result in unreliable parameter estimates because of increasingly sparse survey data within the joint demographic distributions. Multilevel models, in contrast, partially pool information across all categories of each predictor, allowing "deep interactions" (Ghitza and Gelman 2013). Thus model 3 extends model 1:

$$\pi_i = \text{logit}^{-1}\Big(\alpha + \beta \text{year}_{t[i]} + \mu^{gen}_{g[i]} + \mu^{age}_{r[i]} + \mu^{hhs}_{h[i]} + \mu^{edu}_{d[i]} + \mu^{gen.age}_{g\cdot r[i]} + \mu^{gen.hhs}_{g\cdot h[i]} + \mu^{gen.edu}_{g\cdot d[i]} + \mu^{age.hhs}_{r\cdot h[i]} + \mu^{age.edu}_{r\cdot d[i]} + \mu^{hhs.edu}_{h\cdot d[i]} + \mu^{proj}_{p[i]}\Big).$$

Time-varying demographics. We then allow the demographic group intercepts to vary across time. Again, we specify one such model with the basic four demographic categories and another with ethnicity added. These models allow differential rates of religious minority population growth within demographic subgroups. This helps to model situations such as higher rates of minority immigration among men or young people. Such a feature is again possible within a classical regression framework – in the form of time by demographic category interactions – but the risk is again that parameter estimates are too extreme. Model 5 extends model 1 through the addition of additional slope parameters γ for the time trend that vary by demographic group:

$$\pi_i = \text{logit}^{-1}\Big(\alpha + (\beta + \gamma^{gen}_{g[i]} + \gamma^{age}_{r[i]} + \gamma^{hhs}_{h[i]} + \gamma^{edu}_{d[i]})\text{year}_{t[i]} + \mu^{gen}_{g[i]} + \mu^{age}_{r[i]} + \mu^{hhs}_{h[i]} + \mu^{edu}_{d[i]} + \mu^{proj}_{p[i]}\Big)$$

[19] Although the model with ethnicity has to make do with less data, we consider these model comparisons to be "fair" in the sense that they represents the actual trade-off researchers face when using MRP to estimate religious demography.

The intercepts μ and trend slopes γ for each demographic category are modeled using bivariate normal distributions, with intercepts and slopes correlated ρ. For example, the gender level intercepts and slopes are modeled as follows:

$$\begin{pmatrix} \mu^{gen}_{g[i]} \\ \gamma^{gen}_{g[i]} \end{pmatrix} \sim N(0, \Sigma^{gen}), \quad \Sigma^{gen} = \begin{pmatrix} \sigma^2_{\mu^{gen}} & \\ \rho^{gen}\sigma_{\mu^{gen}}\sigma_{\gamma^{gen}} & \sigma^2_{\gamma^{gen}} \end{pmatrix}$$

Time-varying demographic interactions. Finally, our most complicated model specification includes all two-way demographic interactions *and* allows these interactions to vary by time. These models allow more finely-specified differential rates of religious minority population growth within demographic subgroups. Model 7 in effect combines model 5 and model 3:

$$\begin{aligned} \pi_i = \text{logit}^{-1}\Big(\alpha + (\beta + \gamma^{gen}_{g[i]} + \gamma^{age}_{r[i]} + \gamma^{hhs}_{h[i]} + \gamma^{edu}_{d[i]} + \gamma^{gen.age}_{g\cdot r[i]} + \gamma^{gen.hhs}_{g\cdot h[i]} + \gamma^{gen.edu}_{g\cdot d[i]} + \\ \gamma^{age.hhs}_{r\cdot h[i]} + \gamma^{age.edu}_{r\cdot d[i]} + \gamma^{hhs.edu}_{h\cdot d[i]})\text{year}_{t[i]} + \mu^{gen}_{g[i]} + \mu^{age}_{r[i]} + \mu^{hhs}_{h[i]} + \mu^{edu}_{d[i]} + \\ \mu^{gen.age}_{g\cdot r[i]} + \mu^{gen.hhs}_{g\cdot h[i]} + \mu^{gen.edu}_{g\cdot d[i]} + \mu^{age.hhs}_{r\cdot h[i]} + \mu^{age.edu}_{r\cdot d[i]} + \mu^{hhs.edu}_{h\cdot d[i]} + \mu^{proj}_{p[i]}\Big) \end{aligned}$$

The intercepts μ and trend slopes γ for each demographic category are again modeled using a bivariate normal distribution.

In addition to these eight multilevel models, we also specify and test a corresponding set of eight classical logit regressions (see Table 3). Rather than post-stratifying the estimates from the classical logit models, as we do with our multilevel models (see next section), we instead use survey weights to, in effect, find the weighted likelihoods. We calculate the vectors of weights in question ourselves, using our population data on age by gender by education by household size by (as appropriate) ethnic groups. These classical logit models with survey weights offer a computationally simple alternative to Bayesian MRP for analysts who might prefer the convenience of doing so.

5.3. Post-stratification

MRP requires estimates within demographic subgroups to then be post-stratified. To do so, we first obtain predictions, from each of our models of the proportion of adults in Great Britain holding each of the three religious identities within demographic subgroups j. For example, we denote model MRP1's estimates as $\hat{\pi}_j^{\text{MRP1}}$. For the models without ethnicity, there are 48 demographic subgroups (2 age by 2 gender by 4 education by 3 household size) for each of the 20 years, and thus 960 demographic-temporal subgroups in total. When ethnicity is included in the model there are 96 demographic subgroups for each of the 20 years, and thus 1,920 total subgroups. We next weigh each subgroup prediction by the proportion of the adult population in Great Britain each year that falls in that subgroup, N_j.[20] Finally, we obtain weighted estimates for a smaller set of $k \in 32$ (2 age by 2 gender by 4 education by 2 years) demographic subgroups by aggregating the set of $j \in 48$ subgroups across the three household size groups.[21]

Using our first model, MRP1, the post-stratified estimate of the proportion of the population ϕ_k^{MRP1} holding a particular religious identity within target demographic subgroup k is:

$$\phi_k^{\text{MRP1}} = \frac{\sum_{j \in k} N_j \hat{\pi}_j^{\text{MRP1}}}{\sum_{j \in k} N_j}$$

For comparison, we also produce "disaggregated" estimates by dividing the pooled but unweighted survey data into the 32 demographic subgroups to find the raw proportion of the population holding a particular religious identity within each these subgroups

$$\phi_k^{\text{Disag}} = \frac{\sum_{i \in k} y_i}{N_k}$$

[20] Our data for these population proportions come from the 2001 and 2011 censuses. We use linear interpolation to smooth the population estimates for each of the 48 subgroups between the census years of 2001 and 2011. Before and after these years, we use the unadjusted estimates from either the 2001 or 2011 census.

[21] When ethnicity is included we aggregate across household sizes and ethnicity to produce our 32 estimates.

5.4. Estimation

We fit our multilevel models using the `stan_glmer` function from the `RStanARM` library for R. This function calls the Stan modeling language (Carpenter et al 2017; Stan Development Team 2017), allowing the user to fit a fully Bayesian model using Hamiltonian Monte Carlo sampling and easily access the samples, all from within R. `RStanARM` also provides a convenient and user-friendly interface for specifying Bayesian hierarchical generalized linear models because it allows the user to specify models using the familiar modeling formulas utilized by the widely-used `lmer` and `glmer` functions from the `lme4` R library. `RStanARM` functions additionally allow the user to specify prior distributions, of which more later. Although it is not particularly onerous to code the models oneself in Stan, `stan_glmer` provides optimized code that runs more quickly and efficiently, not a small consideration given that our models took four to eight hours each to run.

We specify the following priors. The coefficient for the year linear effect is given a weakly informative Normal(0, 2.5) prior, which allows the group's rate of growth to be fairly strongly positive or negative, without being implausibly so. For example, given that we standardized year to range from 0 to 1, a coefficient of 1 would imply that the group had grown 170% over the period from 1995 to 2014.

The prior for the grand intercept should be chosen with care. In our specification, this intercept reflects the model estimates of the group size in the first year of analysis (1995) as all additional parameters are varying intercepts that are centered at zero. Even if the analyst does not know the exact size of the group being modeled (this is the purpose of the model after all), she would have some idea about the plausible range of sizes. For groups like British Jews, who are likely smaller than one percent of the population, a Normal(-5, 1) prior on the logit scale is suitable. This provides prior weight on small probabilities around 0.7% (the central 90% of the prior density lies between 0.13 and 2.37% on the probability scale). For slightly larger groups, such as British Hindus and Muslims, we use a Normal (-4, 1) prior on the intercept, where the central 90% of the prior density lies between 0.35 and

6.19% on the probability scale. For larger groups, analysts could use larger means on the prior normal.

Finally, the hierarchical variances (and covariances, for the models with varying slopes) are given the default prior in `RStanARM`. This decomposes the hierarchical covariance matrix for each set of hierarchical intercepts and slopes into a correlation matrix and a pair or variances. The variances are then decomposed into the product of a probability vector and a scale parameter. The correlation matrix is given an LKJ(1) prior (Lewandowski et al 2009; Stan Development Team 2017); the vector of probabilities is given a Dirichlet(1) prior; and the scale parameter a Gamma(1, 1) prior. For models without varying slopes, there are no hierarchical covariances, so the scale parameter is equivalent to the hierarchical variance parameter. Although Gelman (2006) recommends the use of half-Cauchy(0, 2.5) priors for such parameters, independent gamma distributions can be used to construct a multivariate Dirichlet distribution, and are thus preferable when intercepts and slopes vary by group. We retain the Gamma(1, 1) priors for comparability even for the models without hierarchical slope parameters.

The eight multilevel Bayesian models were fit using four parallel chains, run for 400 iterations each. The first 150 iterations in each chain used for adaptation and the remaining 1,000 samples saved and analyzed further. This number of iterations proved to be more than sufficient for convergence, with the $\hat{R}$ diagnostic reaching a value of between 0.95 and 1.05 for all parameters.

The eight classical logit regression models are estimated using the `svyglm` function from the `survey` library in `R`, which fits the classical logit model using inverse-probability weights. Two vectors of survey weights are used: the first uses the joint population age-gender-education-household size distributions; the second adds ethnicity. As such, the estimates obtained from the classical logit models are weighed *before* the subgroup predictions are made.

5.5. Empirical Strategy

The heart of our analysis is a comparison of the accuracy and efficiency of our 16 models by comparing their post-stratified estimates (e.g., ϕ_k^{MRP1}) to the "true" values obtained from the 2001 and 2011 censuses.[22] In particular, we compare the estimated and "true" proportions of British adults holding each of the three religious identities, in 2001 and 2011, within 32 gender by age by household size groups.[23] Such estimates of religious-demographic size are interesting in their own right, but with 32 points of comparison between estimates and "true" proportions, this empirical strategy also provide us with a fair, albeit rigorous test of estimation error.

Analysts face a number of options for measuring the performance of predictive models. A simple choice is to use an information criterion such as AIC, DIC, or WAIC. Such information criteria estimate the fit of a model using the log likelihood, employ a penalty for the number of parameters, and compare the fit to some baseline model. They attempt to approximate the out-of-sample predictive error of models using the in-sample model fit.

A better choice is to use internal validation, or cross-validation, which repeatedly splits the dataset into training and tests sets, fitting the model to the former and estimating the predictive error on the latter. Better still, however, is to use external validation to evaluate the predictive accuracy of competing models (Hastie, Tibshirani, and Friedman 2009). This is typically accomplished by splitting the dataset into a training and test set and "holding out" the latter: using it only once to estimate predictive error. Although this works well in principle, it is hard to know in practice whether analysts have indeed fit their models only to the training set and moreover whether they have in fact created only one test set. Best of all then, is when the test data are split from the training data by some actor or

[22] Although the five percent census samples we use do have sampling error, this error is very small, particularly for the tiny proportions we estimate. With approximately 3 million respondents and a proportion of .01, the standard error is .000057. Even when expressed as a standard error on the percentage scale (.0057%), this remains less than a rounding error in our results, and can be safely ignored.

[23] Thus, for each religious group, we produce 512 (32×16) estimates. We present additional analyses for all bivariate or trivariate distributions including household size and the full joint distributions in the online supplementary materials.

agency other than the authors themselves. This is the situation we are in: our test data are the census estimates of religious group size.

However, there are in fact two distinct goals when evaluating the predictive accuracy of a set of models: model selection and model assessment (Hastie, Tibshirani, and Friedman 2009). Model selection involves choosing the best-fitting model from those considered. This is our goal. Model assessment involves additionally estimating the predictive error that a single selected model will have using some external, as yet unseen, data. To accomplish both model selection and assessment, two separate test datasets are required, with the one dedicated to model selection usually labeled the validation set. Since we only have a validation dataset (the census data), and not a further test dataset, we cannot perform true model assessment. Our estimates of predictive error thus cannot be generalized beyond the case at hand. However, our use of an external validation dataset means that we avoid overfitting our models on the sample because our selection of models is based on the validation dataset, not the training dataset.

6. Results

6.1. Model Comparison and Selection

We use two metrics to test our models. First, we calculate the mean percent error (MPE). For each of the $k = 32$ estimates, we find the absolute value of the difference between the model-estimated values and the census value and divide this by the census value to convert to the percentage scale, before finding the mean. The MPE for our first set of estimates is:

$$\text{MPE}^{\text{MRP1}} = \frac{100}{32} \sum_{k=1}^{32} \frac{\left|\phi_k^{\text{Census}} - \phi_k^{\text{MRP1}}\right|}{\phi_k^{\text{Census}}}$$

The MPE thus rescales errors to a common metric. The simpler mean absolute error (MAE, i.e., $\frac{1}{32} \sum_{k=1} \left|\phi_k^{\text{MRP1}} - \phi_k^{\text{Census}}\right|$) does not, and would thus disadvantage estimates that were incorrect for larger proportions (such as young Muslims) and advantage estimates that

were accurate for smaller proportions (such as older Muslims).

We also calculate credible or confidence interval coverage (CIC) for our models. This test evaluates how accurately our 16 models capture the uncertainty in estimation. Put another way, CIC measures the accuracy of standard errors (Carsey and Harden 2014). We calculate the empirical coverage of our uncertainty intervals (whether confidence or credible intervals) by finding the percent of the 32 estimates from each model where the corresponding uncertainty interval includes the true, census value. We then compare this empirical coverage percentage to the nominal coverage implied by the confidence level, $1-\alpha$. The $1-\alpha\%$ CIC for model MRP1 is calculated as follows:

$$\text{CIC}^{\text{MRP1}} = \frac{100}{32}\sum_{k=1}^{32} \phi_k^{\text{Census}} \in \text{CI}_\alpha(\phi_k^{\text{MRP1}})$$

If the empirical coverage is substantially lower than the nominal $1-\alpha$ level then the estimator is overly precise, produces standard errors that are too small, and imparts an undue degree of confidence in results. If empirical coverage is substantially greater than the nominal level, the estimator is inefficient, produces standard errors that are too large, and imparts an undue degree of uncertainty in results. We use the $\alpha = 0.2$ level, to produce 80% credible / confidence intervals. This will allow us to ascertain both whether our standard errors are too small or too large.[24]

These two metrics are presented for the 16 models and three religious groups in Figure 1. A comparison of the results shows that the better performing models – those near the top of each plot – achieve roughly comparable accuracy across the three religious groups. They approach, and sometimes surpass, an error rate (MPE) of around 20% when compared with the census values. These deviations are modest for the estimation of such small subgroups.[25] The most accurate models for each group (in MPE terms) correspond to mean absolute errors

[24] In particular, the standard 95% confidence interval allows very little room to evaluate whether models produce inflated standard errors or overly conservative estimates of uncertainty, especially because we only have 32 observations (test values) per model.

[25] For example, only 0.26% of men aged 30-49 without a degree reported a Jewish identity in 2001.

of .58 percentage points for the estimates of Muslim identity, .22 for the Hindu estimates, and .10 for Jewish estimates.

Both the MRP and CRSW estimates substantially improve on the estimates obtained by disaggregating the raw, pooled survey data. As the figures clearly show, even the worst performing model offers a substantial improvement over the disaggregated results. The modeled results are up to 63% more accurate for Muslims, 77% for Hindus, and 84% for Jews. Modeling the raw data offers huge benefits for the accuracy of demographic estimates.

There is, however, considerable variation in the performance of particular models. The three most accurate estimates for each religious group are those derived from MRP, rather than CRSW models. Yet, the worst-performing estimates for Muslim and Hindu demographic estimation are also those based on MRP. MRP holds promise but perhaps peril as well. Model choice nevertheless clearly matters for improving survey estimates of religious demography. The best fitting MRP model is 16% more accurate than the best-fitting CRSW model with respect to the Muslim estimates, 18% more accurate for the Hindu estimates, and 22% more accurate for the Jewish estimates. Although the MRP models take considerably longer than the CRSW to run (hours versus seconds), the time spent does appear to pay off in a noticeable improvement in accuracy.

Which particular model is most accurate? It depends, first, on the religious group in question. For the newer religious minorities – Hindus and Muslims – the three most accurate estimates all include ethnicity. Despite the loss of data that follows from incorporating ethnicity in models of religious identity, doing so clearly helps improve the accuracy of estimates of the demography of newer religious minorities.

Two-way interactions also appear to increase modeling accuracy. Indeed, the most accurate model for Hindu and Muslim demographic estimation is MRP4, which includes ethnicity and two-way demographic interactions but no time-varying effects. Model MRP3, which has two-way interactions, no time-varying effects, and does not include ethnicity, is the second most accurate for Jewish demographic estimation.

Allowing the effects of demographics (either main effects or interactions) to vary by time has mixed effects. MRP models with varying effects are among the three most accurate for all religious minorities. These models are also theoretically appealing as they are the most flexible of those considered here. Such flexibility comes at a cost however, as models with varying trends also feature among the least accurate of the models we consider, especially when implemented within a classical regression framework. In sum, it appears that including varying slopes has volatile effects, and is thus best avoided.

Turning to the coverage of our uncertainty intervals, Figure 1 shows clearly that the MRP credible intervals generally approach the nominal 80% level. In particular, the two-way interaction models which we have already identified as being accurate (i.e. MRP4 for Muslims and Hindus; MRP3 for Jews) also have very good CI coverage properties, bolstering their claims to be the best model choice. In contrast, the CRSW uncertainty estimates are too precise and fall short, sometimes dramatically, of the 80% level. Thus, not only are the MRP point estimates more accurate than those obtained using classical regression models, their estimates of uncertainty are also substantially more reliable.

A more systematic analysis of the factors associated with lower MPE and better CI coverage is to regress the 48 (i.e., 16 models by three groups) results, for each of the two metrics, on indicators for whether interactions were included, whether time varies, whether MRP (or CRSW) was utilized, and which religious group was modeled. These results are included in Table 4, and support four conclusions regarding model choice. First, these regressions confirm that MRP produces lower error rates than CRSW and also less discrepancy between nominal and empirical CI coverage. Second, including ethnicity improves Hindu and Muslim estimates, both in terms of mean percent error and uncertainty interval coverage, but harms the Jewish estimates. Third, although including demographic interactions has no effect either way on percent error, it does reduce the discrepancy between empirical and nominal CIC. And finally, these results also confirm that including time-varying slopes reduces model accuracy, but does not harm uncertainty interval coverage.

In sum, we find that applying multilevel regression models, and post-stratification weighting, to survey data can provide accurate, perhaps even *very* accurate estimates of small religious-demographic subgroups. Our better estimates deviate from the corresponding census estimates by approximately 20% (or $0.1 - 0.6$ percentage points). In addition, the method of MRP, although more onerous than using classical regressions and survey weights, produces estimates that are $16 - 22\%$ more accurate. We also find that MRP models with two-way demographic interactions appear a good general choice for modeling religious demography, with among the best accuracy and uncertainty interval coverage. Varying the effects of time by demographics produces uneven results and is probably best avoided. Finally, for newer religious minorities, such as Hindus and Muslims – but not more settled minorities, such as Jews – including ethnicity can improve estimates despite some ensuing loss of data.

6.2. Descriptive Results

To complete the discussion of our model results, we examine more the results from our selected models more closely. We begin by examining the 32 predictions of religious group prevalence within demographic subgroups, before having a look at the estimates of overall religious group size over time.

Figures 2, 3, and 4 display the estimated and census prevalence of Muslim, Hindu, and Jewish identity within each of the 32 demographic subgroups. We have already confirmed that these estimates are reasonably accurate. The figures provide a visual confirmation of this accuracy, with our MRP point estimates closely corresponding to the census estimates in most instances.

The estimates do, however, show fairly substantial errors among a few demographic subgroups. Our estimates of Muslim prevalence, for example, are too high among young people (aged 16-29), particularly young men, in 2011. This suggests that the age structure of British Muslims is changing. Such a pattern of changing age structure might suggest

that a model with time varying effects would be more accurate. However, as we have shown in Figure 1, time-varying models are never more accurate than the two-way interaction model for Muslims, and often less accurate. Our estimates of the prevalence of Hindu identity among demographic subgroups (Figure 3) show a less systematic pattern of error, with slight underestimates of younger uneducated women in 2001 but slight overestimates of educated women in 2011. Finally, turning to the Jewish estimates (Figure 4), we see a declining population structure, in contrast to the growth shown by British Muslims and Hindus. However, our MRP model – which in this case does not include ethnicity – continues to perform well. There is overestimation of Jewish identity among older educated women in 2001, but otherwise estimates correspond closely to census proportions.

Although our empirical strategy called for measuring the prevalence of three religious identities within 32 demographic subgroups, analysts and practitioners might perhaps be more interested in our estimates of the overall size of these religious groups over time. Figure 5 thus plots these results. For each plot, we use the models that we have designated our most-preferred: the MRP estimates with two-way interactions, with ethnicity included in the cases of Muslims and Hindus.

Figure 5 contrasts the MRP estimates, in orange, with the disaggregated estimates, in grey. Both sets of estimates are shown with 80% uncertainty intervals. The actual census data for 2001 and 2011 are then presented using red dots. At the level of overall population size, this figure shows that the method of MRP applied to existing survey data produces very accurate estimates of Jewish population size. Our 2001 and 2011 estimates deviate from the census estimates by a tiny 0.04 percentage points, or 9.6% using the mean percent error metric. Jewish demographics are perhaps fairly easy to estimate as this group is stable and long-established in Britain. Yet our estimates of Hindu and Muslim population size remain very accurate, with mean absolute errors of 0.14 and 0.29 percentage points respectively, which translates into mean percent error of 11.8% and 7.9%.

These results show, unsurprisingly, that our method is even more accurate for mea-

suring the overall prevalence of religious minority identities than it is for measuring the prevalence of these identities within demographic subgroups. These estimates of overall group size differ from the census estimates by only around 10% on average across the three groups. Results as accurate as these confirm that MRP can indeed be used to reliably estimate the size of religious minorities.

7. Conclusion

This paper tests a method for improving estimates of the demographic size, composition, and dynamics of religious minority groups obtained from existing survey data: multilevel regression modeling with post stratification, or MRP. We compare the accuracy of eight MRP models to two simpler methods for measuring demographics with existing survey data: first, a corresponding set of eight classical logit regressions with simple survey weighting; second, pooled survey data disaggregated by religious group and year.

We find that MRP applied to existing survey data can indeed be used to accurately measure the size of small minority groups, and even the joint distributions of these minorities within other demographic subgroups (e.g. men aged 30-49 without degree in a particular year). In addition, we find a similar accuracy (in mean percent error terms) when measuring the prevalence of established minorities, British Jews in our case, and newer, rapidly growing minorities (British Muslims and Hindus) whose survey samples may be more suspect. We thus conclude that the size of small and increasing demographic groups can be reliably estimated using MRP.

Regarding model specification, we find that including two-way demographic interactions appears to generally aid in modeling religious demography. Varying the effects of time by demographic category produces uneven results and is probably best avoided. For newer religious minorities – but not more settled minorities – including ethnicity can improve estimates despite some ensuing loss of data.

Applied researchers will usually apply MRP and pick a model specification in the

absence of census data. We recommend that such researchers take additional contextual knowledge into account. In particular, if analysts believe that the populations of interest are reliably sampled by existing public opinion surveys (e.g. British Jews), then they might use methods of cross-validation (e.g., Broniecki, Leemann, and Wüest 2017; Warshaw and Rodden 2012). If however, the populations of interest are thought to be poorly represented in survey data (e.g., British Muslims), then the raw survey estimates are unreliable and cannot be used along with cross-validation to select models. In such cases, we recommend that researchers follow our general modeling suggestions.

Our choice of the UK was motivated by the availability of both a considerable quantity of existing survey data and two rounds of census estimates of religious minority prevalence. Similar rich troves of survey data are likely to exist for other highly developed countries and, indeed, for ethnic minorities as well as religious ones. However, where the quantity – but especially the quality – of existing survey data is diminished, analysts should expect accuracy to diminish as well.

In particular, where survey samples of minority groups are thought to be dramatically unrepresentative, analysts must rely more heavily on a model of group identity. In such situations, additional demographics and design-based variables might need to be added to our fairly stripped-down model. Fortunately the method of MRP, with its partial pooling and post-stratification, is designed for robustness. Indeed, MRP has been shown to produce accurate estimates even when used with non-random convenience samples to forecast elections (Wang et al 2015).

In future research, scholars could use MRP to measure the size of religious or other minorities in settings other than the UK, where official data are lacking. The method might even be implemented to measure other relatively rare attributes such as health behaviors, for which there are no census estimates. It would also be useful to investigate the efficacy of this method in contexts where survey data is sparser or of lower quality.

References

Adida, Claire L., David D. Laitin, and Marie-Anne Valfort. 2016. *Why Muslim Integration Fails in Christian-Heritage Societies.* Harvard University Press.

Binder, David A., and Georgia Roberts. 2009. "Design and Model-Based Inference for Model Parameters." In *Handbook of Statistics 29: Sample Surveys: Inference and Analysis*, Danny Pfeffermann and C.R. Rao, eds. London: Elsevier. pp. 33–54.

Broniecki, Philipp, Lucas Leemann, and Reto Wüest. 2017. "Improved Multilevel Regression with Post-Stratification Through Optimal Model Selection." Paper presented at the annual meeting of the European Political Science Association, Milan.

Brown, Davis and Patrick James. 2015. Religious Characteristics of States Data Set. Maryville University.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software*. 76(1): 1–32

Carsey, Thomas M., Jeffrey J. Harden. 2014. *Monte Carlo Simulation and Resampling Methods for Social Science.* Thousand Oaks, CA: Sage.

DiPrete, Thomas, Andrew Gelman, Tyler McCormick, Julien Teitler, and Tian Zheng. 2011. "Segregation in Social Networks Based on Acquaintance and Trust." *American Journal of Sociology* 116(4): 1234–1283.

Fearon, James, and David Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97(1): 228-90.

Font, Joan, and Mónice Méndez. (2013). "Introduction: The methodological challenges of surveying populations of immigrant origin." In *Surveying Ethnic Minorities and Immigrant Populations: Methodological Challenges and Research Strategies*, Joan Font and Mónice Méndez, eds. Amsterdam: Amsterdam University Press. pp. 11–30.

Geddes, Andrew. 2003. *The Politics of Migration and Immigration in Europe.* London: Sage.

Gelman, Andrew. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis* 1(3):515–33

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel / Hierarchical Models.* New York: Cambridge University Press.

Gelman, Andrew and Thomas C. Little. 1997. "Poststratification Into Many Categories using Hierarchical Logistic Regression." *Survey Methodology* 23:12735.

Gelman, Andrew, David Park, Boris Shor, and Jeronimo Cortina. 2008. *Red State, Blue State, Rich State, Poor State. Why Americans Vote the Way They Do.* Princeton: Princeton University Press.

Ghitza, Yair and Andrew Gelman. 2013. "Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups." *American Journal of Political Science* 57(3): 762-776.

Grim, Brian J and Becky Hsu. 2011. "Estimating the Global Muslim Population: Size and Distribution of the World's Muslim Population." *Interdisciplinary Journal of Research on Religion* 7: n/a.

Gill, Jeff. 2008. *Bayesian Methods. A Social and Behavioral Sciences Approach.* Chapman & Hall.

Hanretty, Chris, Benjamin E. Lauderdale, and Nick Vivyan. 2016. "Comparing Strategies for Estimating Constituency Opinion from National Survey Samples." *Political Science Research and Methods* doi:10.1017/psrm.2015.79

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition).* New York: Springer.

Johnson, Todd M. and Brian J. Grim. 2013. *The World's Religions in Figures. An Introduction to International Religious Demography.* Chichester: Wiley-Blackwell.

Keeter, Scott, Courtney Kennedy, Michael Dimock, Jonathan Best, and Peyton Craighill. 2006. "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey." *Public Opinion Quarterly* 70(5): 759–779.

Korn, Edward L., and Barry I. Graubard. 1999. *Analysis of Health Surveys.* Hoboken, NJ: Wiley

Lax, Jeffrey R. and Justin H. Phillips. 2009. "How Should We Estimate Public Opinion in the States? *American Journal of Political Science* 53:107-21.

Leemann, Lucas and Fabio Wasserfallen. 2017. "Extending the Use and Prediction Precision of Subnational Public Opinion Estimation." *American Journal of Political Science.* DOI: 10.1111/ajps.12319

Levitt, Peggy. 2003. " 'You Know, Abraham Was Really the First Immigrant:' Religion and Transnational Migration." *International Migration Review.* 37(3): 847–873.

Lewandowski Daniel, Dorota Kurowicka, and Harry Joe. 2009. "Generating Random Correlation Matrices Based on Vines and Extended Onion Method." *Journal of Multivariate Analysis* 100(9): 1989–2001.

Lipset, Seymour M. and Stein Rokkan. 1967. *Party Systems and Voter Alignments. Cross-National Perspectives.* New York: Collier-Macmillan.

Maoz, Zeev and Erol A. Henderson. 2013. "The World Religion Dataset, 1945-2010: Logic, Estimates, and Trends." *International Interactions* 39(3): 265-291.

Office for National Statistics. 2009. "Final Recommended Questions For the 2011 Census in England and Wales: Religion." London: Office for National Statistics. Retrieved 28 November, 2016 (http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/2011-census-questionnaire-content/final-recommended-questions-2011---religion.pdf.

Park, David K., Andrew Gelman and Joseph Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12: 375–385.

Putnam, Robert D. and David E. Campbell. 2010. *American Grace: How Religion Divides and Unites Us.* New York: Simon and Schuster.

Rao, J. N. K. 2003. *Small Area Estimation.* Hoboken, NJ: Wiley

Roberts, Georgia and David A. Binder. 2009. Analyses Based on Combining Similar Information From Multiple Surveys. *Survey Research Methods Section of the Joint Statistical Meetings (JSM)*, 2138–2147

Selb, Peter and Simon Munzert. 2011. "Estimating Constituency Preferences From Sparse Survey Data using Auxiliary Geographic Information. *Political Analysis* 19: 455–70.

Selway, Joel Sawat. (2011). "Cross-Cuttingness, Cleavage Structures and Civil War Onset." *British Journal of Political Science* 41(1): 111–138.

Smith, Thomas W. (2005). *Jewish Distinctiveness in America.* New York: The American Jewish Committee.

Stan Development Team (2017) *Stan Modeling Language: User's Guide and Reference Manual: Stan 2.15.0.* Stan Development Team

Sonya K. Sterba. 2009. "Alternative Model-Based and Design-Based Frameworks for Inference From Samples to Populations: From Polarization to Integration." *Multivariate Behavioral Research* 44(6): 711–740.

Tighe, Elizabeth, David Livert, Melissa Barnett, and Leonard Saxe. 2010. "Cross-Survey Analysis to Estimate Low-Incidence Religious Groups." *Sociological Methods & Research* 39(1): 56–82.

Toft, Monica, Daniel Philpott, and Timothy S. Shah. 2011. *God's Century: Resurgent Religion and Global Politics.* New York: W.W. Norton.

Thomas, Steve, and Brenda Wannell. 2009. "Combining Cycles of the Canadian Community Health Survey." *Health Reports* 20(1): 53–58.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. "Forecasting Elections With Non-Representative Polls." *International Journal of Forecasting* 31(3): 780–791.

Warshaw, Christopher and Jonathan Rodden. 2012. "How Should We Measure District-Level Public Opinion on Individual Issues?" *Journal of Politics* 74(1): 203–219.

Voas, David, Alasdair Crockett and David V.A. Olson. 2002. "Religious Pluralism and Participation: Why Previous Research Is Wrong." *American Sociological Review* 67(2): 212–230.

Voas, David and Mark Chaves. 2016. "Is the United States a Counterexample to the Secularization Thesis?" *American Journal of Sociology* 121(5): 1517–1556.

Table 1. Survey Data by Year and Project

Year	British Social Attitudes Survey	Euro-barometer	European Social Survey
1995	3,633	2,154	0
1996	3,662	0	0
1997	1,355	2,183	0
1998	3,146	1,066	0
1999	3,143	0	0
2000	3,426	0	0
2001	3,287	0	0
2002	3,435	0	0
2003	4,432	0	0
2004	3,199	0	0
2005	4,268	3,063	0
2006	4,290	3,021	0
2007	4,124	0	0
2008	4,486	1,005	2,352
2009	3,421	1,015	0
2010	3,297	1,009	2,422
2011	3,311	0	0
2012	3,248	1,001	2,286
2013	3,244	0	0
2014	2,878	0	0

Cell entries are the number of respondents that were asked a survey question regarding their religious identity by year and survey project. In some years more than one Eurobarometer survey asked respondents about religious identity. Total $N = 91{,}862$.

Table 2. Question Wording and Response Sets

Census[a]

Question: What is your religion?

Response set: 1) No religion; 2) Christian (Including Church of England, Catholic, Protestant and all other Christian denominations); 3) Buddhist; 4) Hindu; 5) Jewish; 6) Muslim; 7) Sikh; and 8) Any other religion or belief (WRITE IN).

European Social Survey

Question: Do you consider yourself as belonging to any particular religion or denomination? (IF YES) Which one?

Response set: 1) Yes, Roman Catholic; 2) Yes, Protestant; 3) Yes, Eastern Orthodox; 4) Yes, Other Christian denomination; 4) Yes, Jewish; 5) Yes, Islamic; 6) Yes, Eastern religions; 7) Yes, Other non-Christian religions; 8) No; 9) Don't Know

Eurobarometer

Question: Do you regard yourself as belonging to a religion? (IF YES) Which of them?

Response set: 1) Yes, Roman Catholic; 2) Yes, Protestant; 3) Yes, Orthodox; 4) Yes, Other Christian; 4) Yes, Jewish; 5) Yes, Muslim; 6) Yes, Buddhist; 7) Yes, Sikh; 8) Yes, Hindu; 9) Yes, Atheist; 10) Yes, Non-believer, agnostic; 11) Yes, Other (WRITE IN); 12) None

British Social Attitudes Study

Question: Do you regard yourself as belonging to any particular religion? (IF YES) Which?

Response set: 1) No religion; 2) Yes, Christian, no denomination; 3) Yes, Roman Catholic; 4) Yes, Church of England/Anglican; 5) Yes, Baptist; 6) Yes, Methodist; 7) Yes, Presbyterian/Church of Scotland; 8) Yes, Free Presbyterian; 9) Brethren; 10) United Reform Church (URC)/Congregational; 11) Other Protestant (WRITE IN); 12) Other Christian (WRITE IN); 13) Yes, Hindu; 14) Yes, Jewish; 15) Yes, Islam/Muslim; 16) Yes, Sikh; 17) Yes, Buddhist; 18) Yes, Other non-Christian (WRITE IN)

[a] England and Wales Census

Table 3. Models

Name	Modeling approach	Ethnicity	Demographic intercepts	Time trend
MRP1	MRP	Excluded	Main effects	Linear & fixed
MRP2	MRP	Included	Main effects	Linear & fixed
MRP3	MRP	Excluded	2-way interactions	Linear & fixed
MRP4	MRP	Included	2-way interactions	Linear & fixed
MRP5	MRP	Excluded	Main effects	Linear & varies by demogs
MRP6	MRP	Included	Main effects	Linear & varies by demogs
MRP7	MRP	Excluded	2-way interactions	Linear & varies by 2-way demogs
MRP8	MRP	Included	2-way interactions	Linear & varies by 2-way demogs
CRSW1	CRSW	Excluded	Main effects	Linear & fixed
CRSW2	CRSW	Included	Main effects	Linear & fixed
CRSW3	CRSW	Excluded	2-way interactions	Linear & fixed
CRSW4	CRSW	Included	2-way interactions	Linear & fixed
CRSW5	CRSW	Excluded	Main effects	Linear & varies by demogs
CRSW6	CRSW	Included	Main effects	Linear & varies by demogs
CRSW7	CRSW	Excluded	2-way interactions	Linear & varies by 2-way demogs
CRSW8	CRSW	Included	2-way interactions	Linear & varies by 2-way demogs
Disaggregated	Pool	NA	NA	NA

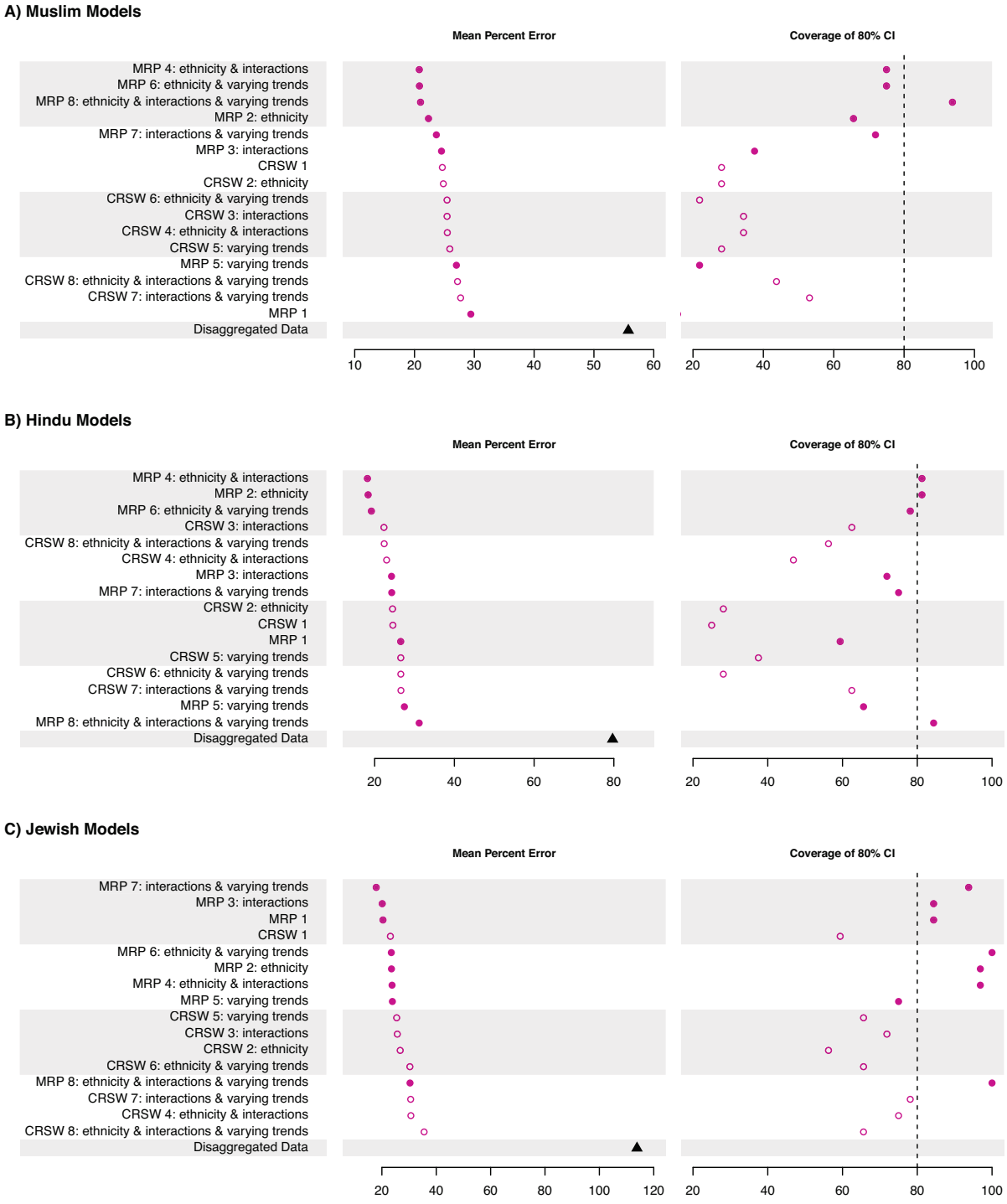
MRP: (Bayesian) Multilevel (logit) Regression with Post-stratification.
CRSW: Classical (logit) Regression with Survey Weighting.
Pool: Pooling of raw, unweighted survey data.

Table 4. OLS Regressions of Model Fit Metrics

	Mean Percent Error		Absolute Discrepancy in Uncertainty Interval Coverage	
	Model 1	Model 2	Model 3	Model 4
Intercept	24.28 (1.29)***	23.10 (1.14)***	36.56 (5.16)***	33.98 (5.17)***
Group: Jewish	1.59 (1.19)	5.16 (1.34)***	−9.14 (4.78)	−1.39 (6.12)
Group: Muslim	0.64 (1.19)	0.64 (1.02)	14.34 (4.78)**	14.34 (4.63)**
Method: MRP	−2.87 (0.97)**	−2.87 (0.83)**	−16.43 (3.90)***	−16.43 (3.78)***
Ethnicity: excluded	0.10 (0.97)	2.47 (1.02)*	2.99 (3.90)	8.16 (4.63)
Interactions: all 2-way	0.49 (0.97)	0.49 (0.83)	−12.32 (3.90)**	−12.32 (3.78)**
Time: varies	1.97 (0.97)*	1.97 (0.83)*	−3.62 (3.90)	−3.62 (3.78)
Jewish × Ethnicity excluded		−7.13 (1.76)***		−15.51 (8.02)
R^2	0.27	0.48	0.57	0.60
N	48	48	48	48

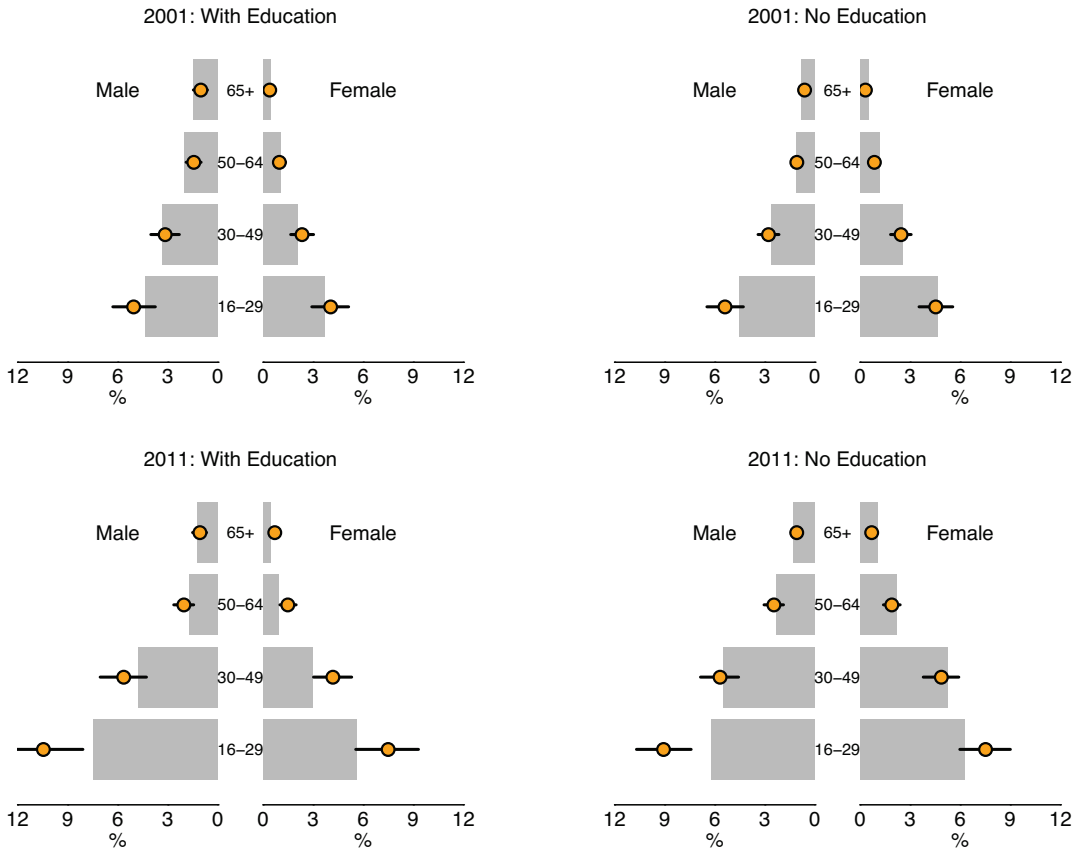
The columns are four OLS regressions of the two model fit metrics obtained for the 48 models (16 specifications × 3 groups). Cell entries are OLS coefficients with standard errors in parentheses. The absolute discrepancy in uncertainty interval coverage $\equiv |80 - \text{CIC}^{\text{emp}}|$, where CIC^{emp} is the empirical coverage of the census values by the 80% confidence or credible intervals of the 48 models. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Figure 1. Model Comparison and Selection



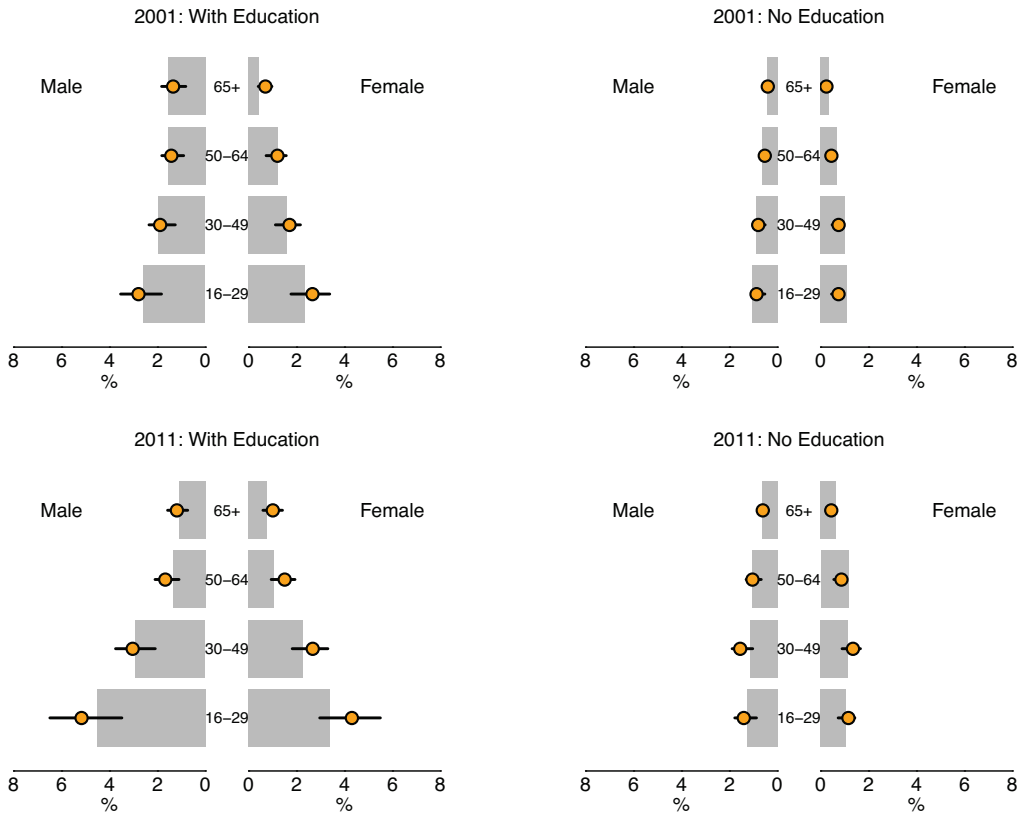
Mean percent error and confidence/credible interval coverage calculated using 32 model and census estimates of religious group prevalence within gender by age by education by year (2001 and 2011) subgroups. Filled circles indicate Multilevel (logit) Regression with Post-stratified (MRP) estimates; hollow circles, Classical (logit) Regressions with Survey Weighted (CRSW) estimates; filled triangle: disaggregated estimates. Models are listed on the y-axes and are ranked in descending order of accuracy by mean percent error.

Figure 2. MRP Estimates of Muslim Demographic Subgroup Size



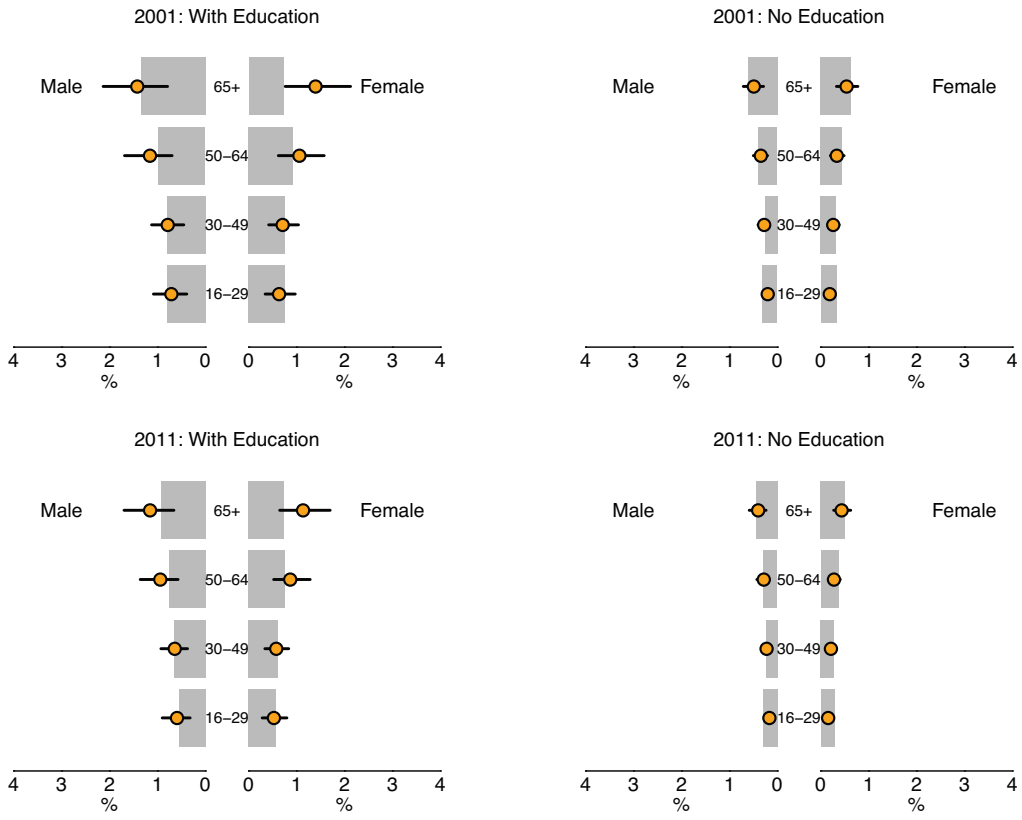
Plot indicates the estimated and "true" (census) prevalence of Muslims within census year (2001 and 2011) by education by age by gender category. MRP estimates are indicated using orange circles, with 80 percent credible intervals shown using black lines. Census estimates of Muslim prevalence within each subgroup are shown using grey bars. MRP4 estimates shown.

Figure 3. MRP Estimates of Hindu Demographic Subgroup Size



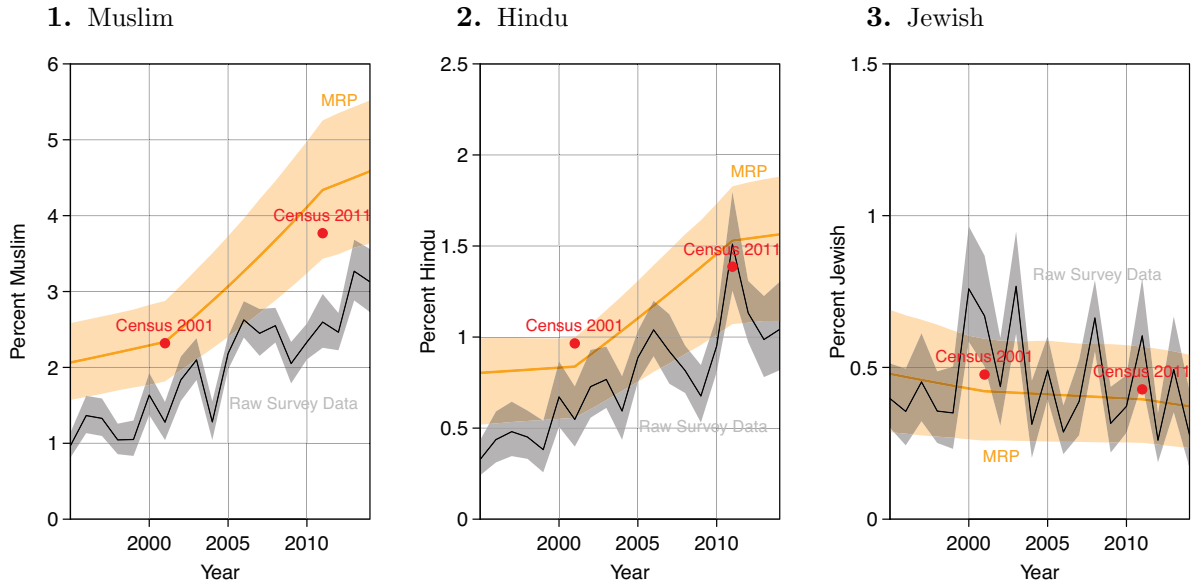
Plot indicates the estimated and "true" (census) prevalence of Hindus within census year (2001 and 2011) by education by age by gender category. MRP4 estimates shown.

Figure 4. MRP Estimates of Jewish Demographic Subgroup Size



Plot indicates the estimated and "true" (census) prevalence of Jews within census year by education by age by gender category. MRP3 estimates shown.

Figure 5. MRP Estimates of Religious Minority Identity in the UK, 1995–2014



Black lines and grey regions show the yearly religious group size estimates obtained by disaggregating the survey dataset by year, along with 80% confidence intervals. Orange lines and regions show the MRP estimates and attendant 80% credible intervals. Red circles indicate the 2001 and 2011 census estimates. Estimates displayed: Muslim and Hindu, MRP4; Jewish: MRP3. The steps in the Muslim and Hindu MRP estimates are due to the use of linearly interpolated census population estimates between the years of 2001 and 2011 and uninterpolated 2001 or 2011 estimates outside this window.