Integrating representative and non-representative survey data for efficient inference

by

Nathaniel Wu Dyrkton

B.Sc., Simon Fraser University 2022

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Statistics)

The University of British Columbia

(Vancouver)

April 2024

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Integrating representative and non-representative survey data for efficient inference

submitted by **Nathaniel Dyrkton** in partial fulfillment of the requirements for the degree of **Master of Science** in **Statistics**.

Examining Committee:

Paul Gustafson, Professor, Statistics, UBC
Supervisor

Harlan Campbell, Adjunct Professor, Statistics, UBC
Supervisory Committee Member

Abstract

Non-representative surveys are commonly used and widely available but suffer from selection bias that generally cannot be entirely eliminated using weighting techniques. Instead, we propose a Bayesian method to synthesize longitudinal representative unbiased surveys with non-representative biased surveys by estimating the degree of selection bias over time. We show using a simulation study that synthesizing biased and unbiased surveys together out-performs using the unbiased surveys alone, even if the selection bias may evolve in a complex manner over time. Using COVID-19 vaccination data, we are able to synthesize two large sample biased surveys with an unbiased survey to reduce uncertainty in now-casting and inference estimates while simultaneously retaining the empirical credible interval coverage. Ultimately, we are able to conceptually obtain the properties of a large sample unbiased survey if the assumed unbiased survey, used to anchor the estimates, is unbiased for all time-points.

Lay Summary

In the age of the internet, surveys are becoming easily accessible and fast. The vast majority of said surveys suffer from being non-representative, meaning the views expressed in the survey generally do not reflect those of the greater population even after adjustment. Does this mean these surveys are unhelpful? We argue that the answer is no if we are able to combine these surveys with a survey that is believed to be representative of the population. We propose a method that synthesizes non-representative surveys and representative surveys that make measurements at several time-points. We show that combining the surveys generally leads to better performance than solely using the representative survey. We apply the proposed method to COVID-19 vaccination data to show that we may use the large-sample non-representative to improve estimates when a representative survey is available.

Preface

This thesis includes work done by Nathaniel Dyrkton in the Department of Statistics at the University of British Columbia. The research was supervised by Dr. Paul Gustafson and Dr. Harlan Campbell. The methodology was proposed Dr. Harlan Campbell and Dr. Paul Gustafson. The author worked closely with Dr. Paul Gustafson and Dr. Harlan Campbell to finalize the model, develop the simulation study, and provide the inspiration to apply the method to the vaccination data. The author was responsible for coding all the simulations, coding proof of concepts, and applying the method to real-world data. The simulations were run on the Cedar cluster from the Digital Alliance of Canada. The code to reproduce the results found in this thesis can be found at: `https://github.com/NDyrkton/SurveyTogether`.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to acknowledge my amazing supervisors Dr. Paul Gustafson and Dr. Harlan Campbell for guiding me through this journey in graduate school. They supported and encouraged me through all the events in the past year and provided me with an amazing research and learning opportunity that I am very grateful for. I would also like to thank the entire department for giving me an amazing experience. I learned an incredible amount in a breadth of different areas that really helped develop my thinking in statistics. Thank you to all the friends I made in graduate school, relating to each-other in tough times really helped me through the periods of difficulty. A special thanks to Clayton for being a great friend and for teaching me that I'm no good at table tennis. Most of all, thank you to my father, who has supported me my entire life. He has always been there for me no matter the situation, and has given me confidence and encouraged me to believe that I can achieve.

Dedication

To my Family

Chapter 1

Introduction

Representative survey methods rely on the conceptually attractive principle of probability-based sampling and have long been considered the gold standard (Lohr 2019). However, new developments have brought an evolving comprehension of the steepening cost-bias trade-off between non-representative and representative surveys. This is due to the fact that online opt-in surveys are becoming widely available and cheap to collect while representative probability surveys are becoming increasingly more expensive and difficult to properly conduct. At the same time, we are better understanding the effects and limitations of post-sampling adjustments.

It has long been recognized that obtaining unbiased representative estimates through post-sampling adjustment hinges on a crucial condition: systematic differences between the study sample and the target population must be known and measured, either directly or through proxies (Lohr 2019). However, the practical implications of this condition remain somewhat unclear. Two recent papers, Wang et al. (2015) and Bradley et al. (2021), shine a light on this ambiguity with their completely contradictory conclusions based on their experiences with recent longitudinal surveys.

Wang et al. (2015) conclude that non-representative surveys can be used to generate accurate election forecasts by demonstrating that post-stratification and regression techniques can be used to correct for biases in a distinctly non-representative sample: voluntary participants solicited via the Xbox gaming platform. The results of Wang et al. (2015) suggest that post-sampling adjustment can correct for even extreme biases given that enough information about demographics, census, and historical data is available. On the other hand, Bradley et al. (2021) conclude that non-representative surveys (even when sample sizes are very very large) can be misleading after reviewing two large online surveys about COVID-19 vaccine uptake: (1) the Delphi-Facebook survey which recruited active Facebook users, and (2) the Census Household Pulse which, despite randomly sampling households (for which contact information was available), arguably failed to obtain a representative sample due to a very low response rate (5-7%). These two surveys drastically overestimated the number of vaccinated Americans, even after careful post-sampling adjustment. The difference in Wang et al. (2015) and Bradley et al. (2021)'s conclusions is striking since both consider online surveys that use similar adjustment methods. Is there a fundamental difference in how these surveys were collected and/or adjusted that explains this stark contrast in representativeness?

An intuitive explanation for the contradictory conclusions may be that Wang et al. (2015) were able to adjust for a large number of key demographic variables while the two "biased" surveys that Bradley et al. (2021) consider lacked certain key variables. Indeed, neither the Delphi-Facebook nor Census Household Pulse surveys adjusted for the political partisanship of respondents, nor did they adjust for urbanicity (and the Delphi-Facebook survey did not explicitly adjust for education). Even if these "biased" polls adjusted for urbanicity, education, and political partisanship, it is not certain that these estimates would

become fully representative. For example, there may be one or more variables that dictate the propensity to be vaccinated, the affinity to join the Facebook platform, and/or the willingness to respond to the survey (while on the Facebook platform). Selecting and accessing a sufficient set of variables on which to adjust is not trivial. Indeed, in certain cases, perhaps counter-intuitively, adjusting for an additional variable seemingly important variable can increase rather than decrease bias.

To demonstrate this, let us review an example about polling for an American election by Mercer et al. (2017). Consider taking a sample that primarily consists of older individuals and those who live in urban areas. Older individuals are more likely to vote Republican, but at the same time, individuals who live in urban centres are more likely to vote Democrat. In this scenario, the biases may "cancel-out" to some extent, leaving a relatively unbiased representative estimate of the proportion of individuals who would vote Democrat. Consequently, only adjusting for either urbanicity or age may lead to an increase in the bias of the survey estimate. This example is perhaps an edge case, but it still emphasizes the idea that domain specific knowledge is required to identify (and measure) the set of variables on which to adjust.

If we draw our attention to the similarities between causal inference in observational studies and non-representative sampling, as suggested by Mercer et al. (2017), we can better understand the main assumptions required to obtain unbiased estimates. The key assumption for post-survey adjustments is that of "conditional ignorability" (Mercer et al. 2017, Schuessler & Selb 2023). Much like the assumption of "no unmeasured confounders" that is required to obtain unbiased causal estimates in an observational study, the conditional ignorability assumption cannot be guaranteed. In other words, there is no sure way of choosing a sufficient set of adjustment covariates (based on available data alone), and ultimately there is no test or way to

validate when the assumption of conditional ignorability is met.

Does this mean non-representative surveys can never be entirely reliable and are of little value? Not necessarily. when combined with representative surveys, non-representative surveys may still be valuable if they can be appropriately leveraged.

1.1 Previous work

There is a lack of literature about combining non-representative and representative surveys. Elliot (2009) propose a method that adjust the estimates via covariates based on estimating the probability of inclusion by using Bayes rule and utilizing the probability of inclusion for the representative probability survey. Wiśniowski et al. (2020) consider a Bayesian regression approach by combining small unbiased surveys with larger biased surveys, but find that bias can still be introduced by the inclusion of non-probability surveys. Another strategy is a method called blended calibration (DiSogra et al. 2011). This strategy weights the probability survey and combines it with unweighted non-probability surveys which are bench-marked against the probability survey. A common theme between these studies is the use of the unbiased survey to "benchmark", or what we will refer to as "anchor" the non-probability survey estimates. One problem we see with all these previous methods is that they all weight, to some degree, either the survey itself or the individuals within each survey, and thus relying on covariates. However, we wish to propose a model that does not rely on any type of weighting procedures, which would in turn remove the need for any adjustment covariates. Our proposed research also differs from the previous methods in that the surveys we analyze are longitudinal, where the degree of selection bias may evolve in some manner over time.

In this thesis, we propose a method that, instead of using

weighting (of either individuals within surveys or of the entire surveys themselves), takes advantage of how selection bias evolves over time within longitudinal surveys. Specifically, we develop a simple and fast Bayesian evidence synthesis method for combining non-representative longitudinal surveys with representative ones. This method neither assumes previously applied weighting, nor requires any informative covariates. One potential complication is that biases can change over time and we consider three possible ways to incorporate the evolution of bias starting from simplest to most complex.

In Chapter 2, we outline our Bayesian synthesis model, state all assumptions, and provide an illustrative example where the inclusion of non-representative surveys reduces the uncertainty of estimates. In Chapter 3, we conduct a simulation study to show that using conservative assumptions, over-parameterizing the model is not overly detrimental. In Chapter 4, inspired by the dataset presented by Bradley et al. (2021), we show that the proposed method performs well in estimating the vaccination rate by combining two "biased" surveys with a presumably "unbiased" survey. The results show a noticeable reduction in uncertainty (compared to only using the unbiased survey). We conclude in Chapter 5 by discussing potential uses and limitations for using large amounts of unrepresentative survey data to improve the precision of survey estimates.

Chapter 2

A Bayesian Synthesis Model

In this chapter we define notation, describe the data, and propose a Bayesian model that synthesizes non-representative and representative surveys. Suppose we have data from K surveys, each of which ask individuals a binary "Yes or No" question, and N is the population size (assumed to be constant across time-points and surveys). Let P_t be the total number of positive individuals (i.e., those who would, if asked, answer "Yes" to the survey question) at time-point t, where $t = 1$ is the first surveyed time-point. For group k in $1, \ldots, K$, suppose:

- n_{kt} is sample size of the k-th survey at time-point t;
- Y_{kt} is the number of positive ("Yes") individuals amongst the n_{kt} individuals surveyed in the k-th survey at time-point t. We assume that individuals are surveyed such that they participate in only a single survey at only a single time-point.

The goal is to estimate $\text{logit}^{-1}(\theta_t)$, the proportion of the (super-)population which is positive at time t, which is the expected value of P_t/N. Specifically, one might be interested in

one of three goals: general inference, "now-casting", and forecasting. General inference refers to collecting data up to time T and making inferences on all time-points up to and including T (i.e., inference on $\theta_0, \theta_1, ..., \theta_T$). Now-casting refers to making inference on a specific time point using only the data collected prior to, and at, that time-point. Lastly, forecasting refers to making inferences about potential future parameters: $\theta_{T+1}, \theta_{T+2}, ...$. In this thesis, we focus on general inference and now-casting.

If the number of time-points is sufficiently large, a random walk model can be used to model how θ_t changes over time (Heidemanns et al. 2020). For instance, suppose P_t changes over time according to:

$$P_t \sim \text{Binomial}(\text{logit}^{-1}(\theta_t), N), \tag{2.1}$$

where

$$\theta_t | \theta_{t-1} \sim \text{Normal}(\theta_{t-1}, \sigma^2), \tag{2.2}$$

and σ^2 represents the variance of the jump in the proportion of "Yes" from the previous time-point (on the logit scale).

In non-representative surveys, estimating θ_t may be challenging due to selection bias. That is, individuals who would likely answer "Yes" might be more inclined to participate in the survey than those who would likely answer "No" (or vice-versa).

Let the degree of selection bias correspond to the ϕ_{kt} non-centrality parameter, where Y_{kt} follows the non-central hyper-geometric distribution with

$$(Y_{kt} | P_t) \sim \text{NCHyperGeo}(P_t, N - P_t, n_{kt}, \phi_{kt}). \tag{2.3}$$

The non-central hyper-geometric distribution is a generalization of the hyper-geometric distribution and describes the probability of drawing Y_{kt} from P_t positive individuals (from a fixed

population of size N) using a sample size of n_{kt}. The generalization occurs where positive individuals are more likely to be be sampled (or vice-versa). This is represented by the ϕ_{kt} parameter, which can be interpreted as an odds ratio. That is, $\phi_{kt} > 1$ means that positive individuals are more likely to be selected, $\phi_{kt} < 1$ implying positive individuals are less likely to be selected, and $\phi_{kt} = 1$ reduces to the typical hyper-geometric distribution.

There are two models that are both referred to as the "non-central hyper-geometric distribution": Fisher's, and Wallenius'. The difference between the two distributions lies in how the ϕ_{kt} parameter changes when an individual is sampled. Wallnius' model assumes that individuals are sampled one at a time. Meaning if a positive individual is sampled, the odds ratio ϕ_{kt} must be adjusted accordingly to account for the fact that there is one less positive individual in the population. Having this property poses a challenge and requires some careful mathematics. Wallenius' non-central hyper-geometric has the following PMF derived by Wallenius (1963):

$$\begin{aligned} P(Y_{kt}|P_t) = &\binom{P_t}{Y_{kt}}\binom{N-P_t}{n_{kt}-Y_{kt}}\times \\ &\times (P_t - Y_{kt} + \phi_{kt}(N - P_t - n_{kt} + Y_{kt}))\times \\ &\times \int_0^1 (1-t)^{Y_{kt}}(1-t^{\phi_{kt}})^{(n_{kt}-Y_{kt})}\times \\ &\times t^{P_t - Y_{kt} - 1 + \phi_{kt}(N-P_t-n_{kt}+Y_{kt})}dt. \end{aligned} \tag{2.4}$$

Obviously (2.4) presents computational difficulties and many Markov Chain Monte Carlo (MCMC) samplers do not support it.

Instead, we opt to employ Fisher's non-central hyper-geometric distribution, which assumes the odds ratio (ϕ_{kt}) is fixed regardless of the number of positive individuals sampled. This can interpreted as the sampled individuals are all sampled/drawn at

once. This distribution has a more manageable PMF, but it still presents computational challenges:

$$P(Y_{kt}|P_t) = \frac{\binom{P_t}{Y_{kt}}\binom{N-P_t}{n_{kt}-Y_{kt}}\phi_{kt}^{Y_{kt}}}{\sum_{y=max(0,n_{kt}-(N-P_t))}^{min(n_{kt},P_t)}\binom{P_t}{y}\binom{N-P_t}{n_{kt}-y}\phi_{kt}^{y}}. \tag{2.5}$$

For a more detailed explanation of both the non-central hyper-geometric distributions and their differences see Fog (2023, 2008). Also see Ballerini & Liseo (2022) for details on the application of Fisher's non-central hyper-geometric distribution to population size estimates in surveys, and the Bayesian estimation of its parameters.

In many cases, N will be large (in relation to n_{kt}) so both distributions will be indistinguishable. Moreover, in this large population case, we may completely disregard sampling with replacement. Thus, we may use an approximation to Fisher's non-central hyper-geometric distribution proposed by Harkness (1965):

$$(Y_{kt}|\theta_t) \sim \text{Binomial}\left(n_{kt}, \frac{p\phi_{kt}}{1-p+p\phi_{kt}}\right), \tag{2.6}$$

where $p = \text{logit}^{-1}(\theta_t)$. This allows us to reduce potential computational difficulties associated with the probability mass function of the non-central hyper-geometric distribution(s).

Defining prior distributions is often controversial, as their choice can substantially influence the posterior when few data are available; see Gelman et al. (2006). We proceed by adopting a truncated normal prior for σ^2:

$$\sigma^2 \sim \text{Normal}(0, \eta_0^2)\text{T}(0, \infty), \tag{2.7}$$

where a larger η_0^2 expresses a greater prior uncertainty about σ^2. A suggested value is $\eta_0^2 = 1$, which corresponds to a prior belief that large changes in the positive rate, while possible, are rather

unlikely. Specifically, if the positive rate at time $t-1$ is 0.5, at the next time point $\text{logit}^{-1}(\theta_t)$ is between 0.26 and 0.76, 90% of the time. For most purposes this is quite wide, unless the real-world time between T and $T-1$ is in the vicinity of years. We may also adopt a normal prior on θ_0:

$$\theta_0 \sim \text{Normal}(\nu_0, \Gamma_0^2), \tag{2.8}$$

where the value for ν_0 directly encodes the belief of where the initial positive rate may lie and Γ_0^2 corresponds to the degree of certainty of this belief. For example, setting $\nu_0 = 0$ assumes the median of $\text{logit}^{-1}(\theta_0)$ is 0.5, which is likely appropriate for a tight presidential poll. Note that if the value for Γ_0^2 is too large, then the majority of the density of $\text{logit}^{-1}(\theta_0)$ will lie near the extremes (0 and 1). Thus a general prior we propose for most cases is $\nu_0 = 0$, and $\Gamma_0^2 = 2$, where roughly 90% of the mass of $\text{logit}^{-1}(\theta_0)$ lies between 0.14 and 0.86, and the density of $\text{logit}^{-1}(\theta_0)$ is approximately uniformly distributed, but its density decreases slightly near 0 and 1.

The only remaining component is a prior for ϕ_{kt}. The degree of selection bias might vary considerably across surveys and may also vary, but perhaps to a lesser degree, across time. Given this parameter is latent and not expected to drastically change over time, we refrain from explicitly reviewing the interpretation of its prior. Instead, we propose simple wide priors that can be used for all but the most extreme purposes (such as the real-world time between time-points is years or decades). We consider three options for movement of ϕ_{kt} over time. Firstly, we consider that ϕ_{kt} is constant in time for each survey. Thus, for all t

$$\begin{aligned} \phi_k &= \exp(\gamma_k), \\ \gamma_k &\sim \text{Normal}(0, 1). \end{aligned} \tag{2.9}$$

Next we may consider that ϕ_{kt} increases or decreases slightly

(and consistently) over time, thus ϕ_{kt} follows a linear model:

$$\begin{aligned} \phi_{kt} &= \exp(\gamma_{k0} + \gamma_{k1} t), \\ \gamma_{k0} &\sim \text{Normal}(0, 1), \\ \gamma_{k1} &\sim \text{Normal}(0, 0.25). \end{aligned} \tag{2.10}$$

Lastly, if ϕ_{kt} changes slightly (and more haphazardly) up or down relative to the previous survey collection:

$$\begin{aligned} \phi_{kt} &= \exp(\gamma_{kt}), \\ \gamma_{kt} | \gamma_{k(t-1)} &\sim \text{Normal}(\gamma_{k(t-1)}, \pi^2), \\ \gamma_{k0} &\sim \text{Normal}(0, 1), \\ \pi^2 &\sim \text{Normal}(0, 1)\text{T}(0, \infty), \end{aligned} \tag{2.11}$$

for the k-th survey at time-point t, for k in 1,...,K. The prior specification therefore assumes that the degree of selection bias may be different for different studies and change over time but likely change slowly over time for a given study.

In order to reasonably estimate ϕ_{kt} and the other model parameters, we need at least one unbiased survey to "anchor" the estimates. In this case we would have a subset of surveys for which ϕ_{kt} is known (usually equal to 1 for all t). Without loss of generality, suppose this subset is the first $k^{'}$ surveys, such that for $k = 1, \ldots, k^{'}$, we have $\phi_{kt} = 1$, for all t. In a situation where all surveys are probability-based, $k^{'} = K$.

2.1 Bayesian Estimation

We may use Bayes rule to derive the posterior up to a proportionality constant. Supposing we fit the model using the binomial approximation (2.6), we review the three models for ϕ_{kt}, all as described in this chapter. We start off with the constant

model for ϕ_{kt} (2.9):

$$P(\theta_0, ..., \theta_T, \sigma^2, \gamma_1, .., \gamma_k | Y_{kt}) \propto \Big[\prod_{t=1}^{T}\Big[\prod_{k=1}^{K} P(Y_{kt}|\theta_t, \gamma_k)P(\gamma_k)\Big] \times$$
$$\times P(\theta_t|\theta_{t-1}, \sigma^2)\Big] P(\theta_0)P(\sigma^2).$$

For the linear model (2.10):

$$P(\theta_0, ..., \theta_T, \sigma^2, \gamma_{10}, ..., \gamma_{K0}, \gamma_{11}, ..., \gamma_{1T} | Y_{kt}) \propto$$
$$\Big[\prod_{t=1}^{T}\Big[\prod_{k=1}^{K} P(Y_{kt}|\theta_t, \gamma_{k0}, \gamma_{k1})P(\gamma_{0k})P(\gamma_{k1})\Big] \times$$
$$\times P(\theta_t|\theta_{t-1}, \sigma^2)\Big] P(\theta_0)P(\sigma^2).$$

Lastly, for the random walk model (2.11):

$$P(\theta_0, ..., \theta_T, \sigma^2, \gamma_{10}, ..., \gamma_{K0}, \gamma_{k1}, ..., \gamma_{K1}, ..., \gamma_{KT}, \pi^2 | Y_{kt}) \propto$$
$$\Big[\prod_{t=1}^{T}\Big[\prod_{k=1}^{K} P(Y_{kt}|\theta_t, \gamma_{kt})P(\gamma_{kt}|\gamma_{k(t-1)}, \pi^2)\Big] P(\gamma_{0k})P(\pi^2) \times$$
$$\times P(\theta_t|\theta_{t-1}, \sigma^2)\Big] P(\theta_0)P(\sigma^2).$$

To take draws from the posteriors, we employ JAGS (Just Another Gibbs Sampler) (Plummer et al. 2003). This sampler uses Gibbs sampling to draw from the full conditional distribution for each parameter. If the full conditional distribution is only known up to a proportionality constant, (and this is most likely the case), then JAGS picks the best MCMC technique. For our proposed models, JAGS uses a variant of slice sampling (Neal 2003) for almost all parameters.

2.1.1 A note on MCMC computation

As we have mentioned before, the non-central hyper-geometric distributions cause considerable computational challenges. JAGS

does support Fisher's non-central hyper-geometric distribution, but it can be ill behaved for our proposed model unless the data is "nice," so to speak. There are several cases where JAGS will not run if we use the hyper-geometric distribution (2.5) as the likelihood. The exact reason is not fully understood, but we will attempt to provide a good guess from reviewing the JAGS source code.

The density calculation of (2.5) is not possible for most values given the many factorials; however, Liao & Rosen (2001) present a recursive algorithm to approximate the density, which is implemented in JAGS. This algorithm (for the un-normalized portion) involves recursive multiplication between n and N, and the number of multiplications appears to scale with N. Thus, when attempting to draw a value from the full conditional distribution using slice sampling, the log density needs to be evaluated. The log density for the non-central hyper-geometric distribution does not have a mathematical form, so the log density is evaluated as log(approximate un-normalized density) rather than a stable mathematical form. In our experience, any large value for N and n will result in an error in log density calculation (even if ϕ_{kt} is fixed). Ballerini & Liseo (2022) have also noted that if N or n are large, the computation becomes quite laborious, and instead propose using an MCMC algorithm that avoids evaluating the likelihood entirely called ABC (Approximate Bayesian Computation) for sampling P_t. This could potentially be used, but would likely require hand-coding the sampling algorithm.

Moreover, if ϕ_{kt} is unidentifiable, for example, if $Y_{1t} \leq n_{1t}$ for the anchor survey and $Y_{kt} = n_{kt}$ for the biased surveys, then JAGS, for the most part, won't run. The initial values to start the chain, usually picked from the prior (Plummer 2017), this matters for ϕ_{kt}. Suppose our initial value is $\gamma_{kt} = 0$, then the conditional distribution $P(P_t, |\cdot)$ can initialize a P_t outside the bounds of its distribution (i.e $P_t > N$). This can be alleviated,

by narrowing the priors on the true parameters, but we risk introducing significant bias.

On the other hand, the binomial approximation (2.6) has a closed form expression for the log likelihood and is thus stable for most values across the parameter space. It lacks the discrete parameter P_t that may have initialization issues, and is easy to implement in JAGS. Thus, we recommend using the binomial approximation (2.6) in almost all cases.

2.2 An Illustrative Example

To illustrate the proposed method, consider the goal of performing inference for $T = 10$ time-points and $K = 3$ surveys, with a population of $N = 10,000$. Survey 1 is unbiased, whereas survey 2 and 3 are compromised by selection bias. Specifically, for these two surveys we generate data under the assumption of a linear ϕ_{kt} (2.10). All the parameters and data are generated jointly using (2.1) (2.2), (2.5), (2.7), and (2.8), and the priors used in analysing the dataset are as such. The data are illustrated in Table 2.1. Analysis of each survey in isolation assumes $\phi_{kt} = 1$

	time-point									
	1	2	3	4	5	6	7	8	9	10
1	$\frac{9}{100}$	$\frac{18}{100}$	$\frac{4}{100}$	$\frac{14}{100}$	$\frac{20}{100}$	$\frac{3}{100}$	$\frac{8}{100}$	$\frac{3}{100}$	$\frac{6}{100}$	$\frac{12}{100}$
2	$\frac{66}{1000}$	$\frac{48}{1000}$	$\frac{7}{1000}$	$\frac{19}{1000}$	$\frac{30}{1000}$	$\frac{2}{1000}$	$\frac{10}{1000}$	$\frac{2}{1000}$	$\frac{2}{1000}$	$\frac{6}{1000}$
3	$\frac{207}{1000}$	$\frac{293}{1000}$	$\frac{102}{1000}$	$\frac{208}{1000}$	$\frac{345}{1000}$	$\frac{117}{1000}$	$\frac{185}{1000}$	$\frac{145}{1000}$	$\frac{174}{1000}$	$\frac{441}{1000}$

Table 2.1: Data generated for the illustrative example, each cell corresponds to $\frac{Y}{n}$ of the respective survey and time-point.

for all t, but retain all the same priors. In Figure 2.1, we see that using the biased surveys ($k = 2$ & 3, assuming they are unbiased) in isolation, completely miss the true positive rate.

On the other hand, the unbiased survey ($k = 1$) tracks the true positive rate well. The proposed method (magenta) also tracks the true positive rate, but the 95% equal-tailed credible intervals more tightly encapsulate the true positive rate.

In this thesis all point-estimates are calculated as the posterior median, and the interval estimates are equal-tailed credible intervals . The choice of the median as a point estimate is to avoid Jensen's Inequality: $g(E[X]) \leq E[g(X)]$, where g is some monotone convex function. The median (and any quantile) is invariant to monotone convex transformations, which are used often in this thesis.

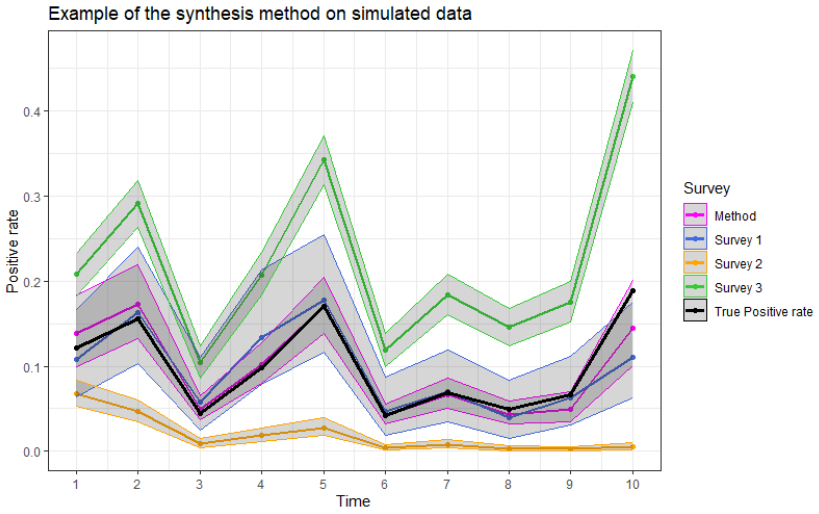

Figure 2.1: Example of the method on simulated data. The ϕ_{kt} assumption is linear (2.10) where $k = 1$ is unbiased, and $k = 2$ & 3 correspond to biased surveys. Estimates are posterior medians and equal tailed 95% credible intervals.

The reduction in credible interval width is quite noticeable. To quantify this, we take the mean ratio of the unbiased survey's credible interval width (in isolation) to that of the pro-

posed method. Across all time-points this mean ratio is 2.18, meaning the method cuts the credible interval width in half. This reduction is present to a lesser degree in the now-casting case. The credible interval for the last time-point ($t = 10$) using the unbiased survey is only 1.12 times larger than that of the synthesis method. Moreover, the method's now-casted credible interval does contain the true positive rate whereas the unbiased survey's interval does not.

Chapter 3

Simulation Study

To assess the extent to which the synthesis method reduces the now-casted MSE (Mean Squared Error) and the effects of over-parameterizing the generating mechanism of ϕ_{kt}, a 3 x 3 simulation design is performed. We consider 3 generating mechanisms for ϕ_{kt}, where ϕ_{kt} either is constant in time (2.9), linear in time (2.10), or follows a random walk (2.11), all as described in the Chapter 2.

3.1 Parameter and Data generation

To get a good representation of how the methods perform under many parameter conditions we generate the parameters from the prior distributions. That is, each simulated dataset arises from different underlying parameter values. One issue from this technique is that priors presented in Chapter 2 are wider than what might be realistic. For example, if one or more of the parameters are extreme in value, then Y_{kt} may be equal to 0 or n_{kt} for many values of t. To avoid this, we narrow the priors to

more plausible values:

$$\begin{aligned} \theta_0 &\sim \text{Normal}(0, 1), \\ \sigma^2 &\sim \text{Normal}(0, 0.1)\text{T}(0, \infty), \\ \gamma_1 &\sim \text{Normal}(0, 0.01), \\ \pi^2 &\sim \text{Normal}(0, 0.01). \end{aligned}$$

This translates to $\text{logit}^{-1}(\theta_0)$ having roughly 90% of its prior density between 0.22 and 0.78. Meaning most of the time we believe the initial positive rate is near 50%. We also set narrow our belief about σ^2 which means $\text{logit}^{-1}(\theta_t) - \text{logit}^{-1}(\theta_{t-1})$ is ± 0.15 or less, 90% of the time, assuming $\text{logit}^{-1}(\theta_{t-1}) = 0.5$. For the linear model, $\phi_{kt} - \phi_{k(t-1)}$ in between -0.33 and 0.33 90% of the time when $\phi_{k(t-1)} = 1$. Lastly, for the random walk model $\phi_{kt} - \phi_{k(t-1)}$ is between -0.37 and 0.40, 90% of the time, where $\phi_{k(t-1)} = 1$.

We simulate the data with $K = 3$ surveys and with a population of $N = 10,000,000$. Let $k = 1$ be the unbiased survey, and $k = 2$ and 3 be the non-probability surveys. Given that non-probability surveys are often larger we set $n_{1t} = 100$, and $n_{2t} = n_{3t} = 1000$ for all t. We repeat this experiment for 3 sets of time-points: ($T = 5$, 10, and 15).

3.2 Simulation Metric

The goal is to compare the MSE, i.e., the averaged squared difference between $\text{logit}^{-1}(\theta_T)$ and the posterior median across the 2000 parameters generated. We also wish to be cognisant about the possibility of Monte Carlo error in the simulations. Ideally, a larger number of repetitions would give more confidence in numerically approximating the MSE.

3.2.1 MCMC and simulation uncertainty

Let us discuss all potential sources of Monte Carlo error, and how we attempt to minimize it. For each of the 2000 repetitions, 12 models are fit in JAGS 4.3.x (Plummer et al. 2003) with 10 parallel independent chains using the dclone package(Sólymos 2010). For the $T = 5$ simulations, there are $20,000$ burn-in with $50,000$ draws for the synthesis model, and $15,000$ burn-in with $50,000$ draws for the unbiased-only model. For the $T = 10$ and 15 simulations, the synthesis models have $25,000$ burn in and $70,000$ draws while the unbiased-only model have $20,000$ burn in and $50,000$ draws. For all simulations, a thinning interval of 5 is used to limit auto-correlation. To assess the convergence of the chains, we may employ the potential scale reduction factor, denoted $\hat{R}$, which measure how the much the variance between the chains differs from the within chain variance; see Gelman et al. (2013). Asymptotically, $\hat{R}$ converges to 1 and suggests that the Markov Chain has converged.

In preliminary testing, $\hat{R}$ and the corresponding upper confidence interval both are 1 for the positive rates, providing strong evidence that the chains have converged. Moreover, with a minimum of $50,000$ draws, the Monte Carlo estimated posterior mean has an error of $O(1/\sqrt{\text{draws}})$ (Gelman et al. 2013). We can assume that the posterior median is close the posterior mean in the case of large n, so this should be sufficiently small.

We are also numerically approximating the MSE by using 2000 repetitions. To visualize our uncertainty we add confidence intervals for the Monte Carlo estimates. Let $\nu_i = \text{logit}^{-1}(\theta_T)_i$ be the true positive rate for repetition i (now-cast), and let $\hat{\nu}_i$ be the now-casted positive rate estimate for repetition i. Then our MSE estimate is:

$$\widehat{MSE} = \sum_{i=1}^{n_{rep}} \frac{(\hat{\nu}_i - \nu_i)^2}{n_{rep}}. \tag{3.1}$$

Relying on the central limit theorem, and for a sufficiently large number of repetitions:

$$\widehat{MSE} \sim \text{Normal}(MSE, \sum_{i=1}^{n_{rep}} \frac{((\hat{\nu}_i - \nu_i)^2 - \widehat{MSE})^2}{n_{rep}(n_{rep} - 1)}), \qquad (3.2)$$

where the variance parameter is the squared MCSE (Monte Carlo Standard Error) given by Morris et al. (2019). Using (3.2) we can easily construct 95% confidence intervals for the Monte Carlo mean squared error.

3.3 Notes on the simulation

As mentioned before, the non-central hyper-geometric distribution(s) (2.5) & (2.4) introduces computational difficulties. Therefore the models are fit using the binomial approximation (2.6), but the data are still generated from Fisher's Hyper-geometric distribution (2.5), as we believe this is the true data generating mechanism. With $N = 10,000,000$ and the largest n being $n_{kt} = 1000$, the binomial approximation is adequate. Random non-central hyper-geometric values are generated using the MCMCpack library (Martin et al. 2011), which has an implementation based on work by Liao & Rosen (2001). Random values from the truncated normal are generated via the truncnorm library (Mersmann et al. 2018), and the rest of values are generated using R (R Core Team 2023). Moreover, for simulations with a large number of time-points, ϕ_{kt} can grow quickly for large values of t (if ϕ_{kt} is linear, (2.10)) and lead to computational issues. To fix this, the time-points, when the models are fitted, are centered in the $T = 10$ and 15 simulations. Meaning: $\phi_{kt} = \exp(\gamma_{k0} + \gamma_{kt}t')$, where $t' = t - \frac{T}{2}$.

3.4 Simulation Results

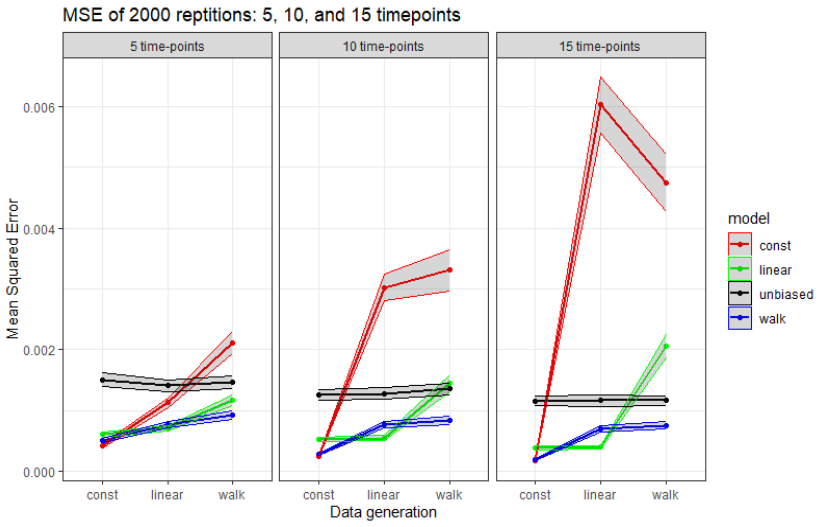

Figure 3.1: Results of simulation study for $T = 5$, 10, and 15 time-points. Simulations run using the model fit in JAGS 4.3.x. The MSE is calculated as the estimated averaged squared difference between $\text{logit}^{-1}(\theta_T)$ and the posterior median. Error bars represent 95% confidence intervals for the estimated MSE (i.e. $\widehat{\text{MSE}} \pm 1.96\text{MCSE}$).

Figure Figure 3.1 shows the results of the simulation study. As expected, when the model for ϕ_{kt} is correctly specified, the MSE is the lowest. Moreover, incorporating the additional 2 biased surveys generally leads to a lower MSE, showing the information from the biased surveys is useful. However, if the model for ϕ_{kt} is drastically under-parameterized, the synthesis model performs worse than using the unbiased survey in isolation. Specifically, fitting a constant ϕ_{kt}, when ϕ_{kt} actually changes significantly, leads to a large increase in MSE. These simulations also suggest that the added complexity of fitting a random walk ϕ_{kt} does not

lead to a large increase in estimator variance. Surprisingly, for scenarios with even a few time-points, a random walk model for ϕ_{kt} appears to be a relatively safe assumption. By increasing the number of time-points, additional complexity can be used to better adjust for the bias without gaining excessive estimator variance.

Why does the linear model perform the worst when the data generating mechanism is constant? This result is surprising as the constant model is nested within the linear model. One reason why this may occur is that the slope parameter γ_{1k} is estimated to be close to zero, and for a sufficiently large number of time-points the linear model diverges significantly from the constant model. This simulation study measures the *now-cast* performance, and thus a linear model may perform worse at the last time-point where the estimate for ϕ_{kt} may be most inaccurate.

Chapter 4

Application: The large vaccine survey

Monitoring the COVID-19 vaccination rates over time was an essential part of assessing the public health response to the pandemic. As discussed in the introduction, Bradley et al. (2021) explains how two survey were found to drastically overestimate the true vaccination rate. Out of the three surveys analyzed, only the Axios-Ipsos was found to track the CDC's benchmark well (acknowledging imprecision in the benchmark). The Axios-Ipsos poll primarily focused on quality probabilistic sampling and maintained a high response rate (approx 50%). As a result, its confidence intervals were found to contain the CDC's historically updated benchmark 10/11 times (up to June 2021), despite its small sample size. Thus, we propose using the method described in this thesis to further improve the estimates of the Axios-Ipsos poll. We graciously use the extended data provided by Bradley et al. (2021) to assess the method on real-world data. We aim to compare the performance of the synthesis method to the Axios-Ipsos Poll and the CDC's historically updated benchmark. We follow Bradley et al. (2021) in adding $\pm 5\%$ error to the CDC's benchmark estimates. Let $\phi_{\text{AI}} = 1$ for all t, and consider the three models for ϕ_{DF} and ϕ_{HP},

where AI, DF, and HP refer to Axios-Ipsos, Delphi-Facebook, and Household-Pulse respectively. We consider two adjustments to the default prior settings suggested in Chapter 2. Firstly, $\theta_0 \sim \text{Normal}(-2, 1)$. Which means: roughly 90% of the mass of $\text{logit}^{-1}(\theta_0)$ lies between 0.036 and 0.327. This reflects the prior belief in a low vaccination rate at time zero before seeing any data. Secondly, $\theta_t|\theta_{t-1} \sim \text{Normal}(\theta_{t-1}, \sigma^2)\text{T}(\theta_{t-1}, \infty)$ which represents the knowledge that the vaccination rate can only increase.

4.1 Unequal Spacing

The surveys are unequally spaced, that is, there is a survey measurement every week for the Delphi-Facebook survey, and much less frequent measurements for the other two surveys. This issue presents a challenge because a random walk depends on the most recent time-point. We select the dates of the Delphi-Facebook survey as the benchmark, and if any other survey has a measurement within the next 6 days of the benchmark, they are considered to have a measurement at the same time. Otherwise the number of vaccinated individuals for that survey are set as missing.

4.1.1 Inference

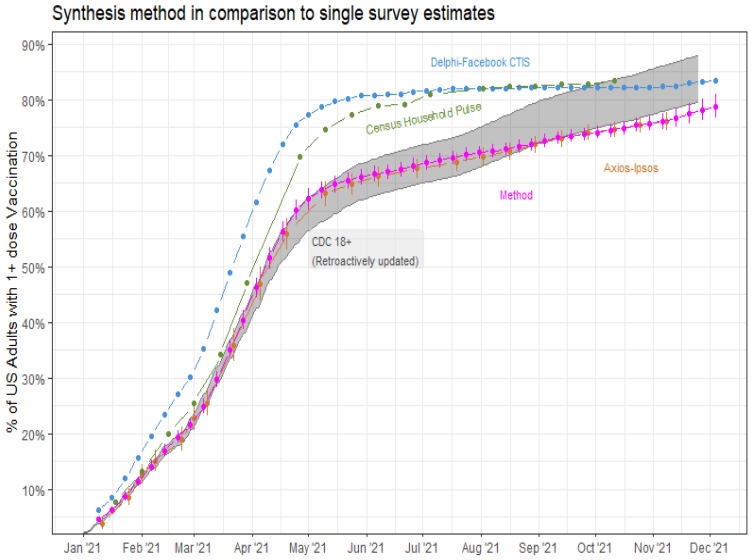

Figure 4.1: Plot of the result for assuming a random walk ϕ_{kt} (2.11), with data extended from Bradley et al. (2021). Point estimates are posterior medians of the positive rate, and intervals are 95% equal-tailed credible intervals. The CDC's historical benchmark has an assumed 5% imprecision, see Bradley et al. (2021) for details.

Figure 4.1 shows estimates obtained from applying the inference from our method in magenta, which shows a strong tendency to closely track the unbiased poll (Axios-Ipsos), with the added advantage of reducing the uncertainty in the estimation.

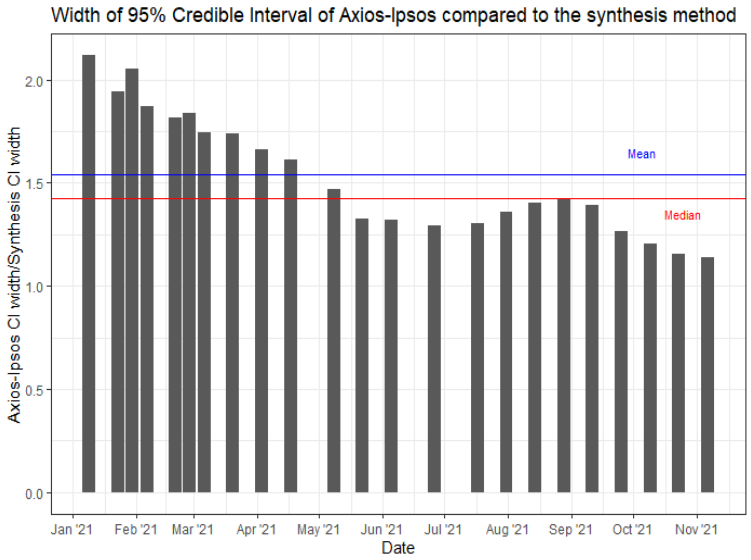

Figure 4.2: Ratio of 95% equal-tailed credible interval widths of Axios-Ipsos surveys and the synthesis method.

In Figure 4.2 we see that the Axios-Ipsos survey alone has a larger 95% credible intervals for all time-points. The average and median Axios-Ipsos credible interval widths are 1.54 and 1.42 times larger respectively (where the Axios-Ipsos survey is not missing). This demonstrates the clear advantage of synthesising the information from multiple surveys, even if we acknowledge that the bias may be changing over time. The unbiasedness property of the Axios-Ipsos survey also appears to be perserved. The synthesis method's 95% credible intervals are within the CDC's benchmark (including the assumed CDC's 5% margin of error) 43/46 (93.5%) of the time (this excludes the last two time-points where no CDC data is present).

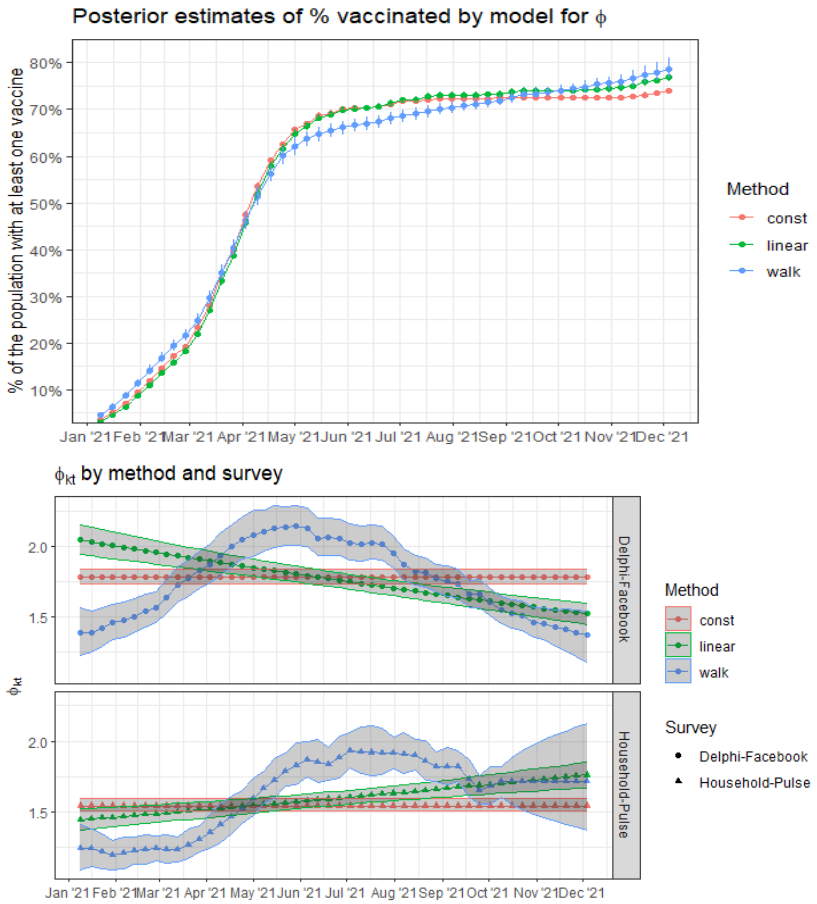

Figure 4.3: Plot of the estimates of vaccination rate by assumption on ϕ_{kt}. Top shows the estimates of ϕ_{kt} for the biased surveys by method (posterior medians + equal tailed 95% credible intervals).

Figure 4.3 shows that changing the specification for ϕ_{kt} changes the estimates of the model significantly, as near the tail end of dates, posterior medians of the vaccination rate can differ by around 10%. Moreover, if the model for ϕ_{kt} is specified as constant, this model predicts a drastically different curve, especially

for the later time points. This aligns with the findings of Bradley et al. (2021), in which the estimated bias is shown to change drastically over time. The MCMC chains for the constant ϕ_k model also fails to converge, with $\hat{R} > 1.1$ for all parameters, which may give unreliable estimates (See Table 4.2). This is more evidently seen with the plot on the right: the random walk ϕ_{kt} moves both up and down, suggesting neither a linear nor constant ϕ_{kt} is an appropriate simplification. We can assess the validity of the random-walk model by comparing the posterior estimates of ϕ_{kt} to the bias estimated by Bradley et al. (2021). In Figure 1b of Bradley et al. (2021), they show the total error by surveys: the survey estimate minus the truth. These estimates have a similar direction as posterior bias estimates presented (up to time-point 20 or June 2021). However, the estimates presented here are not entirely comparable up to June 2021 because information from later time-points affects the estimates of the earlier time-points.

4.2 Now-cast performance

As well as making inference for each time-point after collecting the data on all time-points we also consider the now-casting performance. That is we only consider the information up to and including each time-point. In this section, we compare the now-cast point estimate and 95% credible interval for each of the 48 time-points. Each of the Axios-Ipsos, Houshold-Pulse, and Delphi-Facebook surveys are fit individually using the specified models in the methods section with $\phi_{\cdot t} = 1$, for all t, and our method is fit using information from the three surveys up to and including each point with a random walk assumption for ϕ_{DF} and ϕ_{HP}.

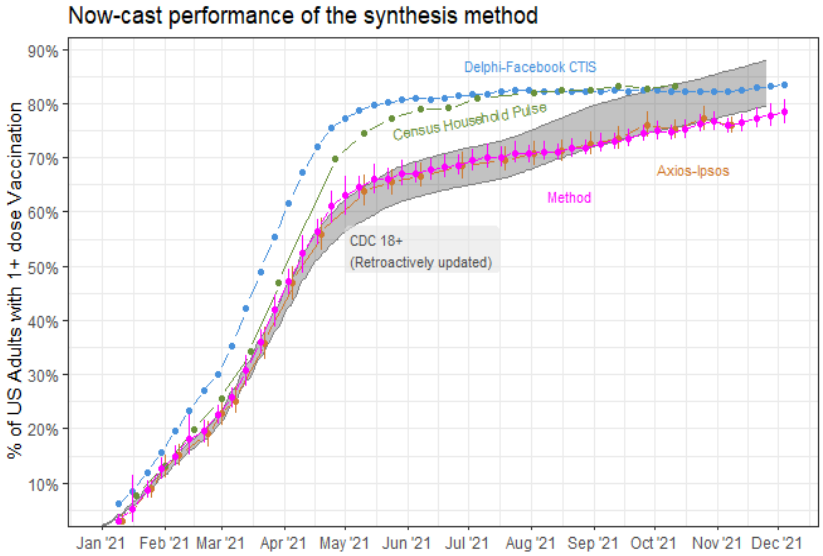

Figure 4.4: Plot of now-cast performance of synthesis method. Estimates are posterior medians and equal-tailed 95% credible intervals. Figure extends data from Bradley et al. (2021).

In Figure 4.4 we can see the point estimates are similar the inference case (Figure 4.1), but with a considerable amount of added uncertainty. Again, the uncertainty of the synthesis method is much lower than that of the Axios-Ipsos survey used in isolation. Precisely, over the time-points where the Axios-Ipsos poll is not missing, the mean and median ratio of the Axios-Ipsos credible interval widths to that of the proposed method are both 1.24 times larger. This decrease in uncertainty does not appear to reduce the empirical credible interval coverage. The synthesis method's 95% credible intervals intersect with the CDC's benchmark range 44/46 (95.7%) of the time. Moreover, there are a few time-points at which the method yields extremely large credible intervals, such as time-points 2 and 6. These are a result of the missing values in the unbiased data. Especially at the beginning of the time-points when few Axios-Ipsos surveys are seen,

a missing value leads to a relatively wide estimation of $P(Y_{kt}|.)$, which ultimately increases the uncertainty about $\text{logit}^{-1}(\theta_t)$.

The now-casted estimates for ϕ_{kt} are much more representative of the uncertainty about the selection bias for each given time-point.

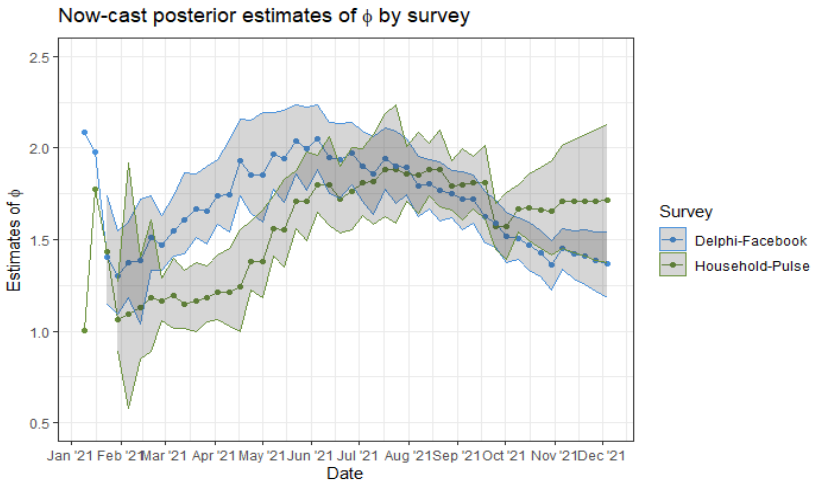

Figure 4.5: Now-casted estimates (posterior medians + equal-tailed 95% credible intervals) for ϕ_{kt} for the Delphi-Facebook and Household-Pulse surveys, assuming a random walk model (2.11). Credible intervals removed for the first 2 to 3 time-points as they are extremely large due presence of missing values.

In that respect, Figure 4.5 better matches Figure 1b in Bradley et al. (2021) (up to June 2021) as compared to Figure 4.4.

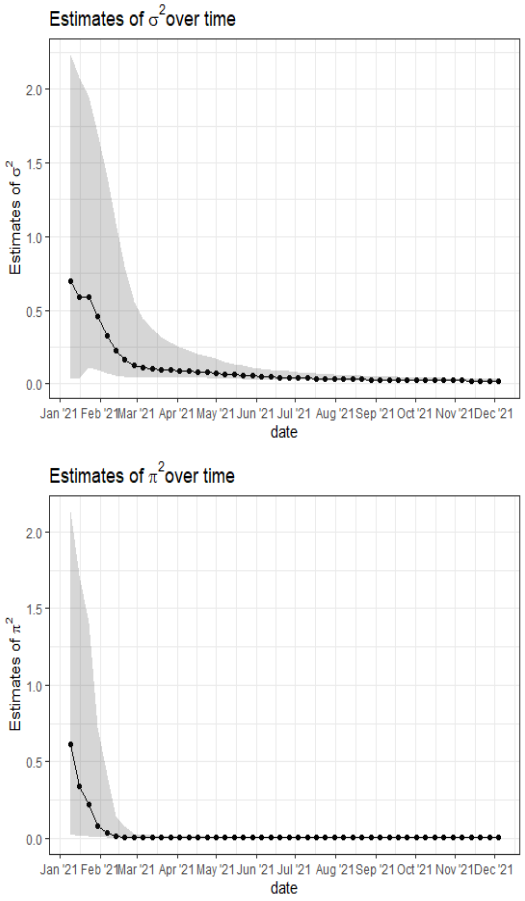

Figure 4.6: Now-casted estimates (posterior medians + equal-tailed 95% credible intervals) for σ^2 and π^2, the jumping variance of the positive rate and variance of γ.

Figure 4.6 shows the posterior median and 95% credible intervals for σ^2 and π^2 across all 48 time-points. Note that for time-point 1, the estimate of σ^2 is essentially equal to that given by the prior. Around mid September there is a slight drop in the estimate for σ^2. This is because as the number of individuals who have been vaccinated starts to level off for both

the Delphi-Facebook and Household-Pulse surveys. The uncertainty around π^2 also reduces quite quickly, and after only about 7-8 time-points most of the uncertainty in the magnitude of the movement in ϕ_{kt} as been reduced.

4.2.1 Inclusion of biased surveys quantified in terms of increased unbiased sample size

One way to measure the reduction in uncertainty, is to measure the increase in sample size required to obtain a reduction in uncertainty that matches the synthesis method. To do this, we fit the method with only the Axios-Ipsos data, and then compare the width of the credible intervals of the Axios-Ipsos survey against the width obtained when of combining the Axios-Ipsos and one of the other surveys, and then ultimately both surveys. The comparison is only made for time-points where an Axios-Ipsos survey is taken. We proceed by calculating n based on the classical frequentist confidence intervals.

- Let $\hat{p}_t$ be the positive rate estimate for time point t of the Axios-Ipsos poll;
- Let R_t be the ratio of credible interval widths of the synthesis method to the baseline Axios-Ipsos poll;
- Let MOE_t be the margin of error of the credible interval of the Axios-Ipsos at time t .

Then we can easily get the number of iid samples required for a $(1 - \frac{\alpha}{2})100\%$ confidence interval:

$$n_{\text{iid samples } t} = \frac{Z^2_{1-\alpha/2}\hat{p}_t(1-\hat{p}_t)}{R_t \text{MOE}_t}. \tag{4.1}$$

Equation (4.1) relies on the assumption that the point-estimates of the synthesis method and Axios-Ipsos polls are equivalent.

This is not the case; however, they should be very similar. Along with (4.1) assuming a completely flat prior, we can only take the following results as a satisfactory approximation. Equivalently, a another method could allow calculate n_{iid} for each survey, ignoring the R_t term. This would also allow $\hat{p}_t$ to differ but also measures changes in point estimates rather purely on reduction in uncertainty. Regardless, both methods should produce very similar results.

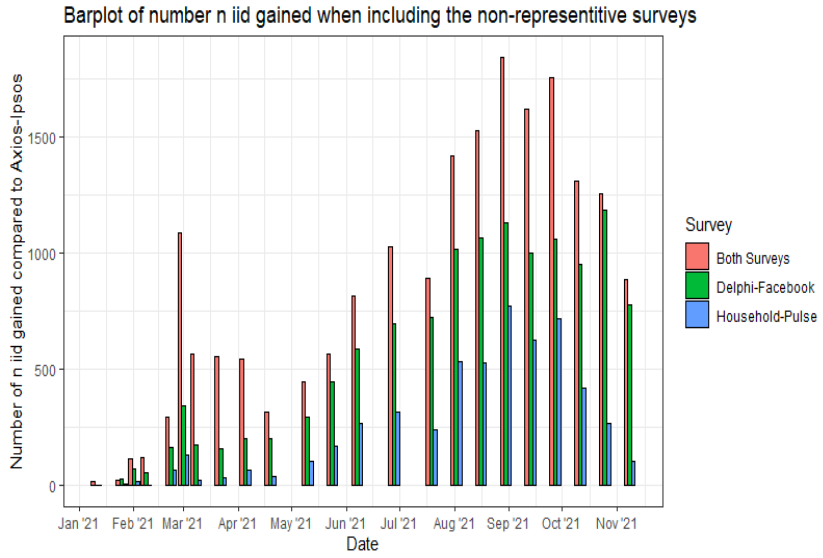

Figure 4.7: Plot of number of iid samples gained by date, for each survey. The bars represent $n_{\text{iid Synthesis}} - n_{\text{iid Axios-Ipsos}}$ at the 95% confidence level.

Surveys included	mean gain	median gain
All surveys	825	817
Delphi-Facebook	534	443
Household-Pulse	235	131

Table 4.1: Summary statistics on the effective number of gained iid samples from the inclusion of the biased surveys, rounded to the nearest integer.

Figure 4.7 shows the plots of the gain in n_{iid} by time-point. We can see that there is considerable information gained for later time-points. The first and fourth time-point have small negative gains when the Household-Pulse survey is included. This is likely due to there being two missing values in four time-points for the Household-Pulse survey (and one missing Axios-Ipsos time-point in between) leading to large posterior variances for parameters. Thus many of these bars are not completely accurate given many of the surveys are missing for some time-points, and we would expect the gain to be larger if all three surveys were present for all time-points. Ultimately, synthesizing a larger survey (in terms of sample size) translates into a larger gain in n_{iid} for the synthesis method. This means that we could conceptually achieve the properties of an unbiased large sample size survey by combing a small unbiased survey with an extremely large biased survey.

In Table 4.1 we can see that despite the large sample size of the Delphi-Facebook (approximately 100,000/time-point), the sample size gained is only approximately 534, which means a loss factor of roughly 200 despite there being a Delphi-Facebook observation for each time-point. Much of this loss is due to the estimating the complex nature of the selection bias and the sparse observations of the Axios-Ipsos poll. Table 4.1 also suggests that the n_{iid} gained using both surveys is greater than the sum of each n_{iid} for combining one survey at a time. This im-

plies that combining more surveys has more than an additive approach to reducing the uncertainty in the estimates. Given the vast number of online surveys (potentially even made at different time-points), the combination of all surveys could drastically improve estimates even if each survey independently only reduces uncertainty a small amount relative to its own sample size.

4.3 MCMC Specifics

This section briefly reviews the specific MCMC details passed through JAGS to ensure replicability. All MCMC models are fit with parallel chains using the dclone package (Sólymos 2010) which links to JAGS. We specify that a chain has sufficiently converged by following the guidelines of $\hat{R} \leq 1.1$ suggested by Gelman et al. (2013). However, due to the large number of models fit and large number of parameters, it is not realistic to calculate $\hat{R}$ for all parameters and all models. For 48 time-points we have about 149 parameters to estimate, each with their own trace-plot and respective $\hat{R}$. In-general $\hat{R}$ is computed where feasible for most of the parameters and trace plots are investigated for satisfactory mixing.

Surveys	chains	burn-in	thin	draws	$\hat{R} \leq 1.1$
AI (Alone)	4	200,000	5	500,000	Yes
DF(Alone)	4	200,000	5	500,000	Yes
HP(Alone)	4	200,000	5	500,000	Yes
S (const ϕ_{kt})	4	250,000	5	500,000	No
S (linear ϕ_{kt})	4	250,000	5	500,000	Yes
S (walk ϕ_{kt})	4	250,000	5	500,000	Yes

Table 4.2: Models fit including objective and subjective diagnostics (Inference). S = "synthesis".

In the now-casting case this problem becomes much more

complex. Each model is fit 48 times, and for some of the early time-points there is so little data available, with multiple missing surveys, which can make the sampling and estimates slightly unstable. The number of independent chains are increased and a large number of burn-in and draws are taken for all synthesis models to account for this. Given there are so many models fit, it's not realistic to check $\hat{R}$ or trace-plots for all parameters and time-points. Thus, to minimize computational cost, $\hat{R}$ is only calculated for the synthesis model of all three surveys, and it is assured that it is less than 1.1 for the positive rate parameters and for all 48 models. If the synthesis model converges it is assumed that the simpler models for a single survey (or only synthesizing two surveys) would also converge and satisfy the condition.

Surveys	chains	burn-in	thin	draws
AI (Alone)	10	50,000	5	100,000
DF (Alone)	10	50,000	5	100,000
HP (Alone)	10	50,000	5	100,000
AI + DF	10	250,000	5	400,000
AI + HP	10	250,000	5	400,000
All surveys	10	400,000	5	400,000

Table 4.3: Models fit for now-casting results, all synthesis methods use a random walk ϕ_{kt}

Chapter 5

Conclusion and limitations

5.1 Limitations and unbiased anchor surveys

We have demonstrated that it is not only possible to synthesize many potentially biased surveys with an unbiased representative survey, but it is advantageous to do so. In aggregate, the proposed synthesis method closely tracks the unbiased survey and reduces posterior variance. However, the proposed method relies on the rather strong assumption that the unbiased anchor survey is truly unbiased (or that the bias is completely known for all t). In reality, this assumption can be either difficult or impossible to meet. It may be feasible if the anchor survey overwhelmingly focuses on the quality of sampling design to achieve properties of a probability sample, rather than simply focusing on sample size. If this is the case, we can achieve the properties of a large sample unbiased survey by synthesizing a small unbiased survey with one or more large online surveys, which are easily available. In the event that an unbiased survey is not present, we may use a well-weighted and carefully designed sur-

vey as the "unbiased" or anchor survey. However, as discussed in the introduction, we may again run into the issue of being unable to meet conditional ignorability. Picking a well weighted anchor survey can still provide substantial benefits by reducing posterior uncertainty about the positive rate if the belief in the weighted survey is justified. Admittedly, if the reference or anchor survey is not properly weighted, using the provided method could worsen survey estimates as we may increase bias significantly, relative to the true positive rate, with a reduction of uncertainty around said biased estimate. This can be easily understood if in the large survey vaccination example we chose the Delphi-Facebook survey as the anchor survey. Therefore, the weighting procedure of the anchor survey should be rigorous, and other factors such as response rate and reputation of the survey practitioner should be considered.

5.2 Selecting a model for ϕ

In this thesis we have provided three possible models for how ϕ may evolve over time: the constant model, the linear model, and the random walk model. The choice of selecting the model for ϕ depends on the prior belief about the evolution of the bias. In most applications the bias would neither be constant nor linear, but interpolating the crude error of a random walk with a linear model for ϕ when there are few time-points is a reasonable choice. However, as shown in Figure 3.1, choosing a random walk model is a conservative and surprisingly efficient choice. There is also the possibility of extending the complexity of the model for ϕ if the number of time-points is large enough. We could consider employing a moving average model or adding other higher order auto-regressive terms. Yet, we must also be careful about adding too much complexity to a latent parameter, but the possibility exists. Furthermore, we have only considered fitting the same model for both of the non-representative sur-

veys. There is an argument to be made to either specify different models for ϕ_k for each k, or to force the non-representative surveys to share the same jump or slope. In the latter case, this assumes the bias ϕ is not necessarily a parameter of the survey, but a parameter of the population that surveys sample from. In either case, the most conservative model (random walk) shown in this thesis can be relied upon, having been shown to have a lower MSE than using the unbiased survey, even when over-parameterized.

5.2.1 Problems with identifying ϕ

We briefly mentioned in the simulation section that one can run into an issue if $Y_{kt} = n_{kt}$ for at least one k or t in the biased surveys. Suppose that $Y_{kt} = n_{kt}$ for at least one k, t in the biased surveys, and $Y_{kt} \leq n_{kt}$ for the unbiased survey(s). Consequently, ϕ_{kt} may lie anywhere in $[1, \infty)$, and so the posterior estimate of ϕ_{kt} would revert to something resembling the prior. The non-central hyper-geometric model (2.5) generally cannot be used if ϕ_{kt} is unidentifiable for any k, t, as JAGS will generally not run. The binomial approximation (2.6) retains the odds ratio interpretation and runs as expected with reasonable results. The question remains: does the unidentifiable ϕ produce excessive problems in inference or now-casting? By exclusively analyzing the simulation results where ϕ is unidentifiable, there is no obvious change in the bias, and the large reduction in variance is still present. Yet this may be due to the correct specification of the priors. In practice, we would hope that there is a rarely a case where surveys have such a large selection bias where the values of Y_{kt} are so extreme.

5.3 Extensions to more general data

The proposed method in this theis is only suited for a binary response. However, extensions to a continuous response would be straightforward. The key idea is that we track how the survey's response bias changes over time. We need not restrict ourselves to any specific type response type, but instead rely on the central limit theorem. For instance, suppose we are interested in surveying salaries and it is presumed that individuals may over-represent their salaries. In this case, a given survey's estimate may be log normal. Let a given survey k estimate at time t be $\bar{x}_{kt}$, then the sampling distribution of survey estimates could be $(\bar{X}_{kt}|\mu_t, \sigma^2, \phi_{kt}) \sim N(\mu_t + \phi_{kt}, \sigma^2(1 - \frac{n_{kt}}{N}))$. Here ϕ_{kt} no longer represents an odds ratio, but rather an additive bias that a survey receives from selection bias and/or biased responses.

5.4 Avenues for future work

In this thesis we've provided a model to synthesize multiple surveys over time. We've shown that it can be robust to over-parametrizing ϕ_{kt}, and that this method can effectively reduce uncertainty in posterior estimates. A natural next step would be to more rigorously analyze the sensitivity of the primary assumption that the unbiased survey must be unbiased for all t. As we've discussed, it is nearly impossible to meet this condition. Thus, practitioners may want some insurance that suggests that this synthesis method does provide a lower MSE, even if the anchor survey is mildly unrepresentative for some t, for all t, or is on-average unbiased for all t. This could be represented in repeating the simulation study in Chapter 3, but while including noise for ϕ_{kt} in parameter generation while still assuming $\phi_{kt} = 1$ for the anchor survey.

Another avenue is to analyze the properties of the proposed method in terms of forecasting metrics. We would hope to un-

derstand the reduction in uncertainty achievable in forecasting results, and whether further model assumptions on the nature of the movement of the positive rate and selection bias are required.

5.5 Concluding remarks

Ultimately, we provide a framework for combining non-representative and representative surveys to reduce uncertainty in survey estimates provided an unbiased survey is present. We can achieve the desirable properties of an unbiased sample, while simultaneously taking advantage of the numerous online surveys that have a large sample size, but suffer from selection bias. These conceptual properties are shown to hold if the movement of selection bias over time is complex. We've also provided a model for ϕ_{kt} that reliably reduces uncertainty in estimates. The framework can be extended to incorporate multiple types of data, and it makes no assumptions on weighting procedures which leaves the door open to practitioners who prefer to weight before making inferences.

Bibliography

Ballerini, V. & Liseo, B. (2022), 'Fisher's noncentral hypergeometric distribution for population size estimation'.

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L. & Flaxman, S. (2021), 'Unrepresentative big surveys significantly overestimated us vaccine uptake', *Nature* **600**(7890), 695–700.

DiSogra, C., Cobb, C., Chan, E. & Dennis, J. M. (2011), Calibrating non-probability internet samples with probability samples using early adopter characteristics, *in* 'Joint Statistical Meetings (JSM), Survey Research Methods', pp. 4501–4515.

Elliot, M. R. (2009), 'Combining data from probability and non-probability samples using pseudo-weights', *Survey Practice* **2**(6), 2982.

Fog, A. (2008), 'Calculation methods for wallenius' noncentral hypergeometric distribution', *Communications in Statistics—Simulation and Computation* **37**(2), 258–273.

Fog, A. (2023), 'Biased urn theory'. `https://cran.r-project.org/web/packages/BiasedUrn/vignettes/UrnTheory.pdf`.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Stern, H. S., Vehtari, A. & Rubin, D. B. (2013), *Bayesian Data Analysis (3rd ed.)*, Chapman and Hall/CRC.

Gelman, A. et al. (2006), 'Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)', *Bayesian Analysis* **1**(3), 515–534.

Harkness, W. L. (1965), 'Properties of the extended hypergeometric distribution', *The Annals of Mathematical Statistics* **36**(3), 938–945.

Heidemanns, M., Gelman, A. & Morris, G. E. (2020), 'An updated dynamic bayesian forecasting model for the us presidential election', *Harvard Data Science Review* **2**(4).

Liao, J. G. & Rosen, O. (2001), 'Fast and stable algorithms for computing and sampling from the noncentral hypergeometric distribution', *The American Statistician* **55**(4), 366–369.

Lohr, S. L. (2019), *Sampling: design and analysis*, CRC press.

Martin, A. D., Quinn, K. M. & Park, J. H. (2011), 'MCMCpack: Markov chain monte carlo in R'.

Mercer, A. W., Kreuter, F., Keeter, S. & Stuart, E. A. (2017), 'Theory and practice in nonprobability surveys: parallels between causal inference and survey inference', *Public Opinion Quarterly* **81**(S1), 250–271.

Mersmann, O., Trautmann, H., Steuer, D., Bornkamp, B. & Mersmann, M. O. (2018), 'Package 'truncnorm'', *R package version* pp. 1–0.

Morris, T. P., White, I. R. & Crowther, M. J. (2019), 'Using simulation studies to evaluate statistical methods', *Statistics in Medicine* **38**(11), 2074–2102. `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8086`.

Neal, R. M. (2003), 'Slice sampling', *The annals of statistics* **31**(3), 705–767.

Plummer, M. (2017), 'JAGS version 4.3.0 user manual'.

Plummer, M. et al. (2003), JAGS: A program for analysis of bayesian graphical models using gibbs sampling, *in* 'Proceedings of the 3rd international workshop on distributed statistical computing', Vol. 124, Vienna, Austria, pp. 1–10.

R Core Team (2023), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

Schuessler, J. & Selb, P. (2023), 'Graphical causal models for survey inference', *Sociological Methods & Research* .

Sólymos, P. (2010), 'dclone: Data cloning in R.', *R Journal* **2**(2).

Wallenius, K. T. (1963), Biased sampling: the noncentral hypergeometric probability distribution, PhD thesis, Department of Statistics, Stanford University, Stanford, CA. Also Published in Technical report no. 70.

Wang, W., Rothschild, D., Goel, S. & Gelman, A. (2015), 'Forecasting elections with non-representative polls', *International Journal of Forecasting* **31**(3), 980–991.

Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A. & Blom, A. G. (2020), 'Integrating probability and nonprobability samples for survey inference', *Journal of Survey Statistics and Methodology* **8**(1), 120–147.