RESEARCH REPORT SERIES
(Statistics #2017-06)

Custom Epoch Estimation for Surveys

Tucker McElroy
Osbert Pang
George Sheldon [1]

[1] Department of Veterans Affairs

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: September 29, 2017

Custom Epoch Estimation for Surveys

Tucker McElroy[*], Osbert Pang[†], and George Sheldon[‡]

September 26, 2017

Abstract

The paper provides a method for generating epoch estimates for time series survey data, allowing for different periods of time (or even point estimates) according to user demand. The method uses a modified kriging estimator, which suppresses the contribution of sampling error variability in order to guarantee that custom epoch estimates have an interpolation property. For the veteran population variable of the American Community Survey, we utilize a simple Brownian Motion model of the population process and derive the modified kriging estimator for this case. The tuning parameters of this population model can be calibrated to the data via simple formulas. We illustrate the application of this method to the generation of point estimates of veteran population, an important objective for Veterans Affairs.

1 Introduction

Many surveys published by statistical agencies, such as the U.S. Census Bureau (USCB), offer estimates for small regions, and in order to decrease sampling error variability, methodologists resort to pooling. Pooling necessarily results in distortions of features – essentially due to smoothing – that may be of interest to data users, and these entities (persons, businesses, and local governments) have expressed demand for estimates corresponding to "custom epochs." Whereas pooling over time corresponds to an epoch-estimate (sometimes called a "period-estimate") for a variable of interest, there is demand for alternative epoch-estimates based on smaller temporal intervals, or based on temporal intervals situated differently within the calendar year. This paper describes

[*]U.S. Census Bureau, Center for Statistical Research and Methodology, tucker.s.mcelroy@census.gov
[†]U.S. Census Bureau, Center for Statistical Research and Methodology, osbert.c.pang@census.gov
[‡]U.S. Department of Veterans Affairs, george.sheldon@va.gov

a methodology – in a context of limited resources – for generating custom epoch-estimates from publicly available survey estimates, utilizing a mechanistic superpopulation stochastic process.

This work is motivated by a consulting project with Veterans Affairs (VA), whose use of the American Community Survey (ACS) data requires custom epoch-estimates that differ in calendar orientation from the published Multi-Year Estimates (MYEs). (An overview of the ACS can be found in U.S. Census Bureau (2006) and Torrieri (2007).) There is no current methodology or software available to produce such custom epoch-estimates, even though users in government, industry, and academia have often expressed their desire for such a product. Users wish to know not only the present value of certain local variables, but how these compare to the recent past, i.e., they require a knowledge of trends, turning points, and cycles in economic and demographic quantities.

Recent funding changes at USCB have brought about the demise of the 3-year MYE as a data product; data users in state and local governments have expressed alarm (through conference calls to the first author) at this situation, and have considered generating their own estimates utilizing the Research Data Centers – at great cost to their own resources. So long as the USCB continues publication of 5-year MYEs (the most reliable period estimates), the possibility of generating 3-year (and 2-year) MYEs from the published 5-year MYEs has some appeal; moreover, 1-year MYEs for low population counties could also be generated in the same manner. The VA application requires a point estimate (i.e., an epoch of length zero) of veteran population oriented at September 30.

Each of these applications requires a "change of support," which refers to scenarios where the epoch of publication differs from the custom epoch-estimate that the data-user desires. A Bayesian spatio-temporal model was proposed by Bradley, Wikle, and Holan (2015), whereby basis functions are specified from given covariates to permit interpolation to different support regions. One might also consider a longitudinal analysis, if it is believed that population structures are common across adjacent geographical regions. However, for the interpolation problem studied in this paper none of these other methods are appropriate or relevant. While the use of covariates, or spatial modeling, or longitudinal analysis can give more information about the population process, they are not helpful for understanding finer time scales. In particular, if we need to produce estimates at weekly or daily epochs, then covariates available on an annual basis are unhelpful.

Moreover, such parametric approaches introduce subjectivity through the modeler's choices; these choices will be valid in the best of circumstances, but can be misleading when careful model checking is prohibitive. Our problem requires a large-scale analysis (tens of thousands of time series) for which individual scrutiny and analysis is infeasible. In other contexts, where budgets allow a case-by-case analysis, we would not advocate such an approach, but in a climate of constrained resources our methodology may be useful to practitioners.

Proceeding with the simple structure of Bell and Hillmer (1990), which decomposes estimated time series into the sum of uncorrelated superpopulation and sampling error processes, we propose a continuous-time aggregated growth process that is corrupted by sampling errors. Although the

VA application involves an integer-valued variable, the counts are high enough for many regions to permit a Gaussian model, and we suppose that the superpopulation process is a Brownian Motion with linear drift. Given that we require a continuous-time process that allows for trend growth and is compatible with symmetry and low kurtosis, the Brownian Motion assumption is the most agnostic possible. The sampling errors have a correlation structure directly inherited from an idealistic view of the sampling procedure. Once the model parameters are calibrated to the input data, custom epoch-estimates can be computed from kriging formulas.

Although careful inference for parameters is of even greater importance in small samples (such as MYEs), given the resource constraints a meticulous approach is not possible. Hence we view the selection of parameters as a calibration to a series' particular features, and utilize estimates derived from the Generalized Least Squares method wherein the sampling error variability is ignored. This removes the impact of sampling error variability on the linear drift parameters, which otherwise suffer severe distortion. Our custom epoch-estimates are likewise generated from a kriging estimator that minimizes quadratic loss subject to interpolation constraints; this has the desired effect of generating interpolations of the data as the customized epochs are varied. Were a higher frequency time series available, or were a longitudinal analysis (presuming a similar trend structure across sub-populations or geographical regions) empirically plausible, our efforts would be obviated. Such not being the case, the paper at hand can be viewed as an effort to satisfy data-users with a technically defensible methodology given a context of data paucity.

2 Pooling in Surveys

A fuller discussion of pooling and its consequences can be found in Nagaraja and McElroy (2015), but here we offer a few comments. When the sample size is small, it may be desirable to borrow information from statistically similar data. Surveys that utilize temporal pooling implicitly utilize the assumption of stationarity to infer that the flow of time will generate statistically similar data, ignoring any non-stationary effects such as trend growth or seasonality. For example, in the ACS this assumption of temporal similarity yielded the MYE, which are epoch-estimates corresponding to either three or five years (whereas the more timely estimate, based on a single year, is not considered an MYE, as it involves pooling over only a single year), with the survey responses collected together into a so-called "rolling sample."

By pooling survey respondents over time (or space or community), the methodologist can increase sample size; but the utility of this condensation is questionable, because the temporal stationarity assumption is false. If it were true, why would there be a demand for more timely data, available at narrower epochs? Indeed, trends and cycles in economic and demographic variables imply that pooling is potentially misleading – it amounts to smoothing the time series, and hence induces phase delay and suppression of higher frequencies in the spectral representation (McElroy,

2009).

To mitigate misinterpretations of epoch-estimates, USCB has labored aggressively to convey the concept of MYEs as estimating an entire period of activity. For example, a hypothetical 10-year MYE would really assess aggregate activity over the decade, and has nothing definitive to say about the final year (let alone, month or day) of that decade. In the same way, 3-year and 5-year MYEs are estimates of population activity over a 3-year or 5-year stretch of time, and do not correspond to an estimate of a single month or day of activity. This bears analogy to the concept of a flow in retail economics, where the measurement corresponds to total sales over the given epoch (say a month or a quarter), as opposed to a stock time series applicable for inventories. However, actually making the case that pooled estimates accurately reflect a corresponding pooled estimand in the population is difficult, given the intricate system for handling non-response in surveys. A casualty of this mechanism for pursuing non-respondents is a loss of understanding of what quantity we are finally estimating. See Nagaraja and McElroy (2015) for documentation of these issues.

The following discussion defines a variable over some epoch in a way that extends the definition of a variable at an instant of time. Let $X(t)$ denote a variable of interest (such as total veteran population) over some specified region (the country, a county, or a tract) A on a map at time $t \in \mathbb{R}$, measured in annual units. For the jth unit of the finite population, at any time t the unit has a coordinate $\underline{s}_j(t) \in \mathbb{R}^2$, and we can measure membership in region A at time t via the indicator function $1_A(\underline{s}_j(t))$. (If t falls outside the lifetime of the unit, we can set $\underline{s}_j(t)$ to be (∞, ∞), and the indicator will be zero.) Tallying up such indicators over all units j in a subpopulation J yields a count variable

$$X(t) = \sum_{j \in J} 1_A(\underline{s}_j(t)). \tag{1}$$

If J is the collection of all U.S. veterans, then (1) provides an abstract formulation of the veteran population for region A at instant t. If other variables are of interest, then the indicator function in (1) could be modified, but we will focus on population counts. The quantity $X(t)$ may even be observable, if units can be continuously tracked (e.g., through a smartphone app or tracking chip).

$X(t)$ represents the population count for the regional pool A. We can consider pooling over time as well, but the mathematical formulation is not straightforward. Given an epoch of $(t-\delta, t]$ of length δ, do we mean to count a unit if $\underline{s}_j(u) \in A$ for all $u \in (t-\delta, t]$? What if there was a short excursion from the region (for vacation or business travel)? By taking the coordinate to be place of residence rather than existential location we can eliminate the issue of travel, but for persons of a nomadic mentality the question remains of how to count their inclusion in an epoch. We wish to define an epoch tally $X_{(t-\delta,t]}$ that is coherent with the instantaneous tally, in the sense that

$$\lim_{\delta \to 0} X_{(t-\delta,t]} = X(t) \tag{2}$$

for any t. To assess the proportion of time spent in a region, we calculate

$$\delta^{-1} \int_{t-\delta}^{t} 1_A(\underline{s}_j(u))\, du.$$

Thus, if unit j spends the entire epoch in region A, then they contribute full weight, but otherwise are counted as a fraction of a person according to their occupancy time. Note that this has a consistency property, in that summing over all regions A forming a partition of the country yields the value one: the unit j spends all their time in epoch $(t-\delta, t]$ fully among the various regions. As a consequence, the epoch count for region A is

$$\sum_{j \in J} \delta^{-1} \int_{t-\delta}^{t} 1_A(\underline{s}_j(u))\, du = \delta^{-1} \int_{t-\delta}^{t} X(u)\, du.$$

This is a pool over both the region A and the epoch $(t-\delta, t]$. (Bradley, Wikle, and Holan (2015) have a similar formulation for the aggregated population process.) Then this epoch quantity, denoted $X_{(t-\delta,t]}$, satisfies (2).

This framework constitutes our formulation of the estimands. The estimators are constructed from a sampling mechanism, which is statistically independent of the superpopulation stochastic process $\{X(t)\}$. However, the epoch estimands given above suggest a method of sampling that differs from the ACS methodology; the ACS instead samples residences from the Master Address File.

3 Population versus Sample

This section reviews and discusses the two sources of random variation – the superpopulation process and the sampling mechanism.

3.1 Background Framework

In describing two sources of uncertainty – the superpopulation and the sample – we follow the approach delineated in Bell and Hillmer (1990). We utilize lower case letters for realizations of random variables on a particular ω in the probability space. For example, the sampling mechanism (which governs the selection of sampling units from the population) is denoted S, with $s = S(\omega)$ the particular set of unit indices used in our implementation. Suppressing the time index for now, X denotes the epoch estimand – a random variable with realization $x = X(\omega)$ in our particular reality. In general, this estimand is some function of a characteristic $\underline{y}$ of the entire realized population. That is, for each member of the population a measurement is taken, and the whole random vector of such is denoted $\underline{Y}$, with realization $\underline{y} = \underline{Y}(\omega)$. In terms of these measurements, the estimand is written $X = f(\underline{Y})$ for some statistic f, and $x = f(\underline{y})$. If we restrict to a sub-population (a sub-vector of $\underline{y}$) for the estimand, we will still write $f(\underline{y})$ by a small abuse of notation.

We therefore have the following quantities of interest:

- $\widehat{X}_S = f(Y_k : k \in S)$ is the sample-based estimator, which is random with respect to sampling mechanism and population. The corresponding sampling error is $E_S = \widehat{X}_S - X$.
- $\widehat{x}_S = f(y_k : k \in S)$ is the sample-based estimate, which is random with respect to sampling mechanism, and is expressed in terms of a realized population. The corresponding sampling error is $e_S = \widehat{x}_S - x$.
- $\widehat{X}_s = f(Y_k : k \in s)$ is the sampled estimator, which is random with respect to population, and is expressed in terms of a realized sample. The corresponding sampling error is $E_s = \widehat{X}_s - X$.
- $\widehat{x}_s = f(y_k : k \in s)$ is the sampled estimate, which is not random, being a realization of both population and sample. The corresponding sampling error is $e_s = \widehat{x}_s - x$.

It is generally assumed that sampling mechanism S and superpopulation $\underline{Y}$ are independent. Also, we assume there is no non-response, which would enter another source of random error to the entire framework. Referring to Särndal, Swensson, and Wretman (1992), a few common properties and concepts are the following:

1. **Model Unbiased**: $\mathbb{E}[E_S|S = s] = 0$ for all s.
2. **Model MSE**: we have $\text{MSE}[\widehat{X}_S|S = s] = \mathbb{E}[E_S^2|S = s]$ for each s.
3. **Large Sample Normality**: if the sample size is large, we can justify a central limit theory approximation to the error of the sampled estimator: $E_S|S = s \sim \mathcal{N}(0, \text{MSE}[\widehat{X}_S|S = s])$.
4. **Design Unbiased**: $\mathbb{E}[\widehat{X}_S|\underline{Y} = \underline{y}] = x$, or (because $\underline{Y}$ and S are independent) $\mathbb{E}[\widehat{x}_S] = x$.
5. **Design MSE**: we have $\text{MSE}[\widehat{X}_S|\underline{Y} = \underline{y}]$.

The first property states that sampling does not distort the estimand's expectation. The second item can sometimes be computed, or estimated. The third property can be verified by construction of the sample. The fourth property may or may not be true, but sample designs are sometimes constructed to ensure that the estimator is design unbiased. Because $\widehat{X}_S = X + E_S$, this fourth property implies that

$$\mathbb{E}[E_S|\underline{Y} = \underline{y}] = \mathbb{E}[\widehat{X}_S - X|\underline{Y} = \underline{y}] = x - x = 0, \tag{3}$$

or $\mathbb{E}[e_S] = 0$. When the survey is design unbiased, the design MSE can be written

$$\mathbb{E}[E_S^2|\underline{Y} = \underline{y}] = \mathbb{E}[e_S^2],$$

because $\underline{Y}$ and S are independent. The design MSE is often referred to as the sampling error variance, and is estimable from the survey. The following result was proved by Bell and Hillmer (1990) in the context of time series survey data. (We provide a simple proof.) Crucially, it provides a justification for the common assumption that superpopulation and sampling error are independent.

Proposition 1 *If a survey is design unbiased, then X is uncorrelated with E_S.*

Proof of Proposition 1. It follows from the property of nested expectations that

$$\mathbb{E}[E_S] = \mathbb{E}[\mathbb{E}[E_S|\underline{Y}]] = 0$$

if the survey is design unbiased, using (3). Therefore $\mathrm{Cov}[X, E_S] = \mathbb{E}[X\, E_S]$, which equals

$$\mathbb{E}[\mathbb{E}[X\, E_S|\underline{Y}]] = \mathbb{E}[X\, \mathbb{E}[E_S|\underline{Y}]] = 0$$

again by (3), also using the fact that $X = f(\underline{Y})$. $\square$

By the same conditional expectation arguments used in the proof of Proposition 1, the variance of E_S equals the design MSE, if the sample is design unbiased:

$$\mathrm{Var}[E_S] = \mathbb{E}[E_S^2] = \mathbb{E}[e_S^2].$$

Then extending the third property above, we may assume that

$$E_S \sim \mathcal{N}(0, \mathbb{E}[e_S^2]) \tag{4}$$

when the sample is large, and the design is unbiased. We will write $\mathrm{Var}[E_S]$ for the sampling error variance $\mathbb{E}[e_S^2]$. Furthermore, if

$$X \sim \mathcal{N}(\mathbb{E}[X], \mathrm{Var}[X]) \tag{5}$$

for some mean and variance, then Proposition 1 indicates that X and E_S are independent so long as they are jointly normal. Finally, we obtain

$$\widehat{X}_S \sim \mathcal{N}(\mathbb{E}[X], \mathrm{Var}[X] + \mathrm{Var}[E_S]). \tag{6}$$

In summary, by taking the survey to be design unbiased and the sample size large, we can justify assumptions (4) and (5), where X and E_S are independent and sum up to $\widehat{X}_S$ given by (6).

3.2 Estimand Redundancy and Interpolation

A superpopulation model involves specifying the mean and variance in (5). When X represents a vector of epoch-estimands, the integral form means there are possible redundancies in the vector, which will yield a singular covariance matrix. We now use the notation $\underline{X}$ to denote this random

vector of n epoch-estimands, each component of which takes the form $X_{(t-\delta,t]}$. for various t and δ. Possibly, some of these components can be expressed as a linear combination of other components, indicating that the vector

$$\underline{X} = J\,\underline{\mathcal{X}} \tag{7}$$

for some reduced rank matrix J and some $\underline{\mathcal{X}}$.

Example 1 Suppose the estimands of interest are three consecutive 1-year epochs, and a 3-year epoch covering the same time span:

$$\underline{X} = \begin{bmatrix} X_{(t-3,t]} \\ X_{(t-3,t-2]} \\ X_{(t-2,t-1]} \\ X_{(t-1,t]} \end{bmatrix} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_{(t-3,t-2]} \\ X_{(t-2,t-1]} \\ X_{(t-1,t]} \end{bmatrix}.$$

The matrix J here has rank 3, and the estimand has redundancy. If we eliminate one of the components of $\underline{X}$, the corresponding row of J will be deleted, resulting in a full rank matrix.

Example 2 When the ACS was first published, the MYEs came in three species: 5-year, 3-year, and 1-year. Let $\underline{\mathcal{X}}$ consist of all 1-year epoch estimands needed to describe all of the MYEs, such that $\underline{X} = J\,\underline{\mathcal{X}}$. Then if we order the MYEs in $\underline{X}$ such that the 5-year MYEs occur first, followed by 3-year and 1-year MYEs, the matrix J has the following structure:

$$J = \begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & \cdots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \cdots & 0 & 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & \cdots & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 1 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \cdots & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

For regions with more than 60,000 people all three types of MYE were originally available, although now the 3-year MYE has been discontinued due to budget constraints. If the population is less than 20,000, only 5-year MYEs are published and only the upper block of rows in J would be available.

The existence of redundancy makes $\mathrm{Var}[\underline{X}]$ non-invertible, which interferes with calibration and kriging. By employing a generalized inverse we can still proceed, but at the cost of losing the

interpolation property of kriging estimators (discussed below). The interpolation property is stated as follows: given data $\underline{\widehat{X}}_S$ we wish to construct estimators of

$$Z = X_{(t-\delta,t]} \tag{8}$$

for arbitrary epochs $(t-\delta, t]$, such that when an epoch is a component of the corresponding $\underline{X}$ our estimator equals the same component of $\underline{\widehat{X}}_S$. Stated mathematically, if $\widehat{Z}$ is our custom-epoch estimator of Z and the epoch is chosen such that it equals the jth component of the population estimand vector $\underline{X}$ (i.e., $Z = u_j' \underline{X}$ where u_j is the jth unit vector), then the interpolation property requires that

$$\widehat{Z} = u_j' \underline{\widehat{X}}_S. \tag{9}$$

One method for generating interpolations is kriging (Zimmerman and Stein, 2010), which utilizes a Gaussian conditional expectation calculation to obtain $\widehat{Z}$. However, the presence of sampling error and estimand redundancies each interfere with the interpolation property (9).

Note that redundancies only pertain to the unobserved estimands, and not to the data vector itself. Therefore, eliminating redundancies by discarding some data may be undesirable; moreover, the formula for the kriging estimator actually depends on which redundancies are eliminated. For this reason, it is preferable to retain all the available data, unless it is absolutely vital for an application to possess the interpolation property.

In the case that redundancies are allowed to remain, the matrix $\mathrm{Var}[\underline{X}]$ is non-invertible, and it is necessary to obtain its generalized inverse for kriging applications. We can obtain the Q-R decomposition (Golub and Van Loan, 1996) of the rank r matrix J in the decomposition (7):

$$J = Q \begin{bmatrix} U \\ 0 \end{bmatrix} \Pi,$$

where Q is orthogonal, U is upper triangular, and Π is a permutation matrix. Assuming that $\mathrm{Var}[\underline{\mathcal{X}}]$ is invertible, we obtain the following expressions involving the generalized inverse:

$$\mathrm{Var}[\underline{X}] = Q \begin{bmatrix} U\,\Pi\,\mathrm{Var}[\underline{\mathcal{X}}]\,\Pi'\,U' & 0 \\ 0 & 0 \end{bmatrix} Q'$$

$$\mathrm{Var}[\underline{X}]^- = Q \begin{bmatrix} (U\,\Pi\,\mathrm{Var}[\underline{\mathcal{X}}]\,\Pi'\,U')^{-1} & 0 \\ 0 & 0 \end{bmatrix} Q'.$$

Some calculations involving these matrices are used below, and the following properties are useful (I_r denotes the r-dimensional identity matrix):

$$\mathrm{Var}[\underline{X}]\mathrm{Var}[\underline{X}]^- = Q \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} Q' \tag{10}$$

$$\mathrm{Var}[\underline{X}]\mathrm{Var}[\underline{X}]^- J = J. \tag{11}$$

Furthermore, the determinant of $\mathrm{Var}[\underline{X}]$ is re-defined to apply unto the upper left block, so that

$$\det \mathrm{Var}[\underline{X}] = \det U \, \Pi \, \mathrm{Var}[\underline{\mathcal{X}}] \, \Pi' \, U'. \tag{12}$$

3.3 The Interpolating Kriging Estimator

In this section we seek an estimator of Z, defined via (8), which is distributed $\mathcal{N}(\mathbb{E}Z, \mathrm{Var}[Z])$. The available data is $\widehat{\underline{X}}_S$, which is jointly normal with the estimand; set $\underline{\gamma}' = \mathrm{Cov}[Z, \underline{X}]$. For normal variables, the minimum MSE estimators are linear, so we consider estimators of the form

$$\widehat{Z} = a + \underline{w}' \, \widehat{\underline{X}}_S.$$

This will be unbiased so long as $a = \mathbb{E}Z - \underline{w}' \, \mathbb{E}[\underline{X}]$. Imposing this property gives

$$\widehat{Z} = \mathbb{E}Z + \underline{w}' \, (\widehat{\underline{X}}_S - \mathbb{E}[\underline{X}]),$$

and direct minimization of $\mathbb{E}[(\widehat{Z} - Z)^2]$ yields the solution $\underline{w} = (\mathrm{Var}[\underline{X}] + \mathrm{Var}[\underline{E}_S])^{-1} \, \underline{\gamma}$; however, we wish to impose the interpolation constraints as well, and this unconstrained solution will violate the interpolation property. To proceed, set $\underline{Z} = [\underline{X}', Z]'$, which is normal with mean $\underline{b} = [\mathbb{E}[\underline{X}]', \mathbb{E}Z]'$ and covariance matrix satisfying $[I_n, 0] \, \mathrm{Var}[\underline{Z}] = \Gamma'$, where $\Gamma = \mathrm{Cov}[\underline{Z}, \underline{X}]$. The minimal MSE unbiased estimator of $\underline{Z}$ is

$$\widehat{\underline{Z}} = \underline{b} + \Gamma \, (\mathrm{Var}[\underline{X}] + \mathrm{Var}[\underline{E}_S])^{-1} \, (\widehat{\underline{X}}_S - \mathbb{E}[\underline{X}]).$$

This same solution is obtained by minimizing the quadratic function

$$.5 \, \underline{y}' \, H \, \underline{y} - \underline{y}' \, G \, \widehat{\underline{X}}_S \tag{13}$$

with respect to $\underline{y}$, and setting $\widehat{\underline{Z}} = \underline{b} + \widehat{\underline{y}} - \mathbb{E}\widehat{\underline{y}}$, where $\widehat{\underline{y}}$ is the minimizer and

$$\begin{aligned} H &= \left(\mathrm{Var}[\underline{Z}] - \Gamma \, (\mathrm{Var}[\underline{X}] + \mathrm{Var}[\underline{E}_S])^{-1} \, \Gamma' \right)^{-1} \\ G' &= \mathrm{Var}[\underline{E}_S]^{-1} \, [I_n, 0]. \end{aligned}$$

This assertion follows from Lemma 1 below, because

$$H^{-1} \, G = \left(\Gamma - \Gamma \, (\mathrm{Var}[\underline{X}] + \mathrm{Var}[\underline{E}_S])^{-1} \, \mathrm{Var}[\underline{X}] \right) \, \mathrm{Var}[\underline{E}_S]^{-1} = \Gamma \, (\mathrm{Var}[\underline{X}] + \mathrm{Var}[\underline{E}_S])^{-1},$$

using $[I_n, 0] \, \Gamma = \mathrm{Var}[\underline{X}]$. The next result is easily proved using the method of Lagrangian multipliers.

Lemma 1 *The unconstrained solution to the problem of minimizing the quadratic function (13) is given by $\widehat{\underline{y}} = H^{-1} \, G \, \widehat{\underline{X}}_S$, if H is invertible. If there are constraints of the form $R' \, \underline{y} = \underline{q}$, then the*

constrained solution is

$$\begin{aligned}\widetilde{\underline{y}} &= M\,\widehat{\underline{y}} + \underline{m} \\ M &= I - H^{-1}\,R\,[R'\,H^{-1}\,R]^{-1}\,R' \\ \underline{m} &= H^{-1}\,R\,[R'\,H^{-1}\,R]^{-1}\,\underline{q},\end{aligned}$$

where I is an identity matrix of dimension equal to the length of $\underline{y}$.

Now we define the *interpolating Kriging estimator* be the unbiased minimizer of (13) subject to the constraints that $[I_n, 0]\,\widetilde{\underline{Z}} = \widehat{\underline{X}}_S$, because this exactly corresponds to imposition of the interpolation property. As this corresponds to $R' = [I_n, 0]$ and $\underline{q} = \widehat{\underline{X}}_S$ in the notation of Lemma 1, we obtain the solution

$$\widetilde{\underline{Z}} = \underline{b} + \widetilde{\underline{y}} - \mathbb{E}\widetilde{\underline{y}} = \underline{b} + \Gamma\,\mathrm{Var}[\underline{X}]^{-1}\,(\widehat{\underline{X}}_S - \mathbb{E}[\underline{X}]),$$

using $R'\,H^{-1}\,R = \mathrm{Var}[\underline{E}_S]\,(\mathrm{Var}[\underline{X}] + \mathrm{Var}[\underline{E}_S])^{-1}\,\mathrm{Var}[\underline{X}]$, $M = I - \Gamma\,\mathrm{Var}[\underline{X}]^{-1}\,[I_n, 0]$, $M\,\Gamma = 0$, and $\underline{m} = \Gamma\,\mathrm{Var}[\underline{X}]^{-1}\,\widehat{\underline{X}}_S$. Taking the first component of $\underline{Z}$, we obtain the simple expression

$$\widehat{Z} = \mathbb{E}Z + \underline{\gamma}'\,\mathrm{Var}[\underline{X}]^{-1}\,(\widehat{\underline{X}}_S - \mathbb{E}[\underline{X}]). \tag{14}$$

The MSE of this estimator is

$$\mathrm{MSE}[\widehat{Z}] = \mathrm{Var}[Z] - \underline{\gamma}'\,\mathrm{Var}[\underline{X}]^{-1}\,\underline{\gamma} + \underline{\gamma}'\,\mathrm{Var}[\underline{X}]^{-1}\,\mathrm{Var}[\underline{E}_S]\,\mathrm{Var}[\underline{X}]^{-1}\,\underline{\gamma}, \tag{15}$$

and the extra variability due to the interpolation constraint is

$$\underline{\gamma}'\,\mathrm{Var}[\underline{X}]^{-1}\,\mathrm{Var}[\underline{E}_S]\,(\mathrm{Var}[\underline{X}] + \mathrm{Var}[\underline{E}_S])^{-1}\,\mathrm{Var}[\underline{E}_S]\mathrm{Var}[\underline{X}]^{-1}\,\underline{\gamma}.$$

This non-negative quantity is zero if Z is uncorrelated with $\underline{X}$, or if there is no sampling error variability; otherwise, it is positive, and is the efficiency loss due to imposing the interpolation property.

4 Superpopulation and Sampling Error Models

We now discuss a particular model for population and sampling error, which may be appropriate for variables that exhibit linear trending behavior from year to year. Consider a population process $\{X(t)\}$ for $t \in [0, \infty)$ modeled as a Brownian Motion with drift of increment with variance σ^2, such that its initial value $X(0) = 0$. (A variant of this model allows for $X(0)$ to be a mean-zero Gaussian random variable, whose variance is to be estimated; this adds an additional tuning parameter to the model, and in applications gave no appreciable benefit.) The drift is linear:

$$\mathbb{E}[X(t)] = \mu_0 + \mu_1\,t.$$

The de-meaned version of the process is denoted with a tilde, so that $\widetilde{X}(t) = X(t) - \mathbb{E}X(t)$, which is expressible as the sum of $\widetilde{X}(0)$ with a mean-zero Brownian Motion. Then the epoch-estimand is written

$$Z = X_{(t-\delta,t]} = \mu_0 + \mu_1(t - \delta/2) + \delta^{-1} \int_{t-\delta}^{t} \widetilde{X}(s)\, ds.$$

From this formulation we can now express $\mathbb{E}[\underline{X}]$ and $\mathrm{Var}[\underline{X}]$ in terms of $\theta = [\mu_0, \mu_1, \sigma^2]'$, the parameters of the population process. Each component of $\underline{X}$ has the form $X_{[t-\delta,t]}$ for some t and δ, and so the mean is simply $\mu_0 + \mu_1(t - \delta/2)$; the covariances can be computed from the following Lemma, which provides the covariance formulas for flow-cumulated Brownian Motion. (The proof is by direct calculation, and is omitted.)

Lemma 2 *If $\widetilde{X}$ is a mean-zero Brownian Motion of increment variance σ^2, and $0 \le a_1 < a_2$ and $0 \le b_1 < b_2$, then $Cov(\int_{a_1}^{a_2} \widetilde{X}(s)\, ds, \int_{b_1}^{b_2} \widetilde{X}(s)\, ds)$ is given by σ^2 multiplied by*

$$\begin{cases} a_2^2(3b_2 - a_2)/6 - a_1^2(b_2 - b_1)/2 - b_1^2(3a_2 - b_1)/6 & \text{if } a_1 < b_1 < a_2 < b_2 \\ a_2^2(3b_2 - a_2)/6 - b_1^2(a_2 - a_1)/2 - a_1^2(3b_2 - a_1)/6 & \text{if } b_1 < a_1 < a_2 < b_2 \\ (a_2^2 - a_1^2)(b_2 - b_1)/2 & \text{if } a_1 < a_2 < b_1 < b_2 \\ (b_2^2 - b_1^2)(a_2 - a_1)/2 & \text{if } b_1 < b_2 < a_1 < a_2 \\ b_2^2(3a_2 - b_2)/6 - a_1^2(b_2 - b_1)/2 - b_1^2(3a_2 - b_1)/6 & \text{if } a_1 < b_1 < b_2 < a_2 \\ b_2^2(3a_2 - b_2)/6 - b_1^2(a_2 - a_1)/2 - a_1^2(3b_2 - a_1)/6 & \text{if } b_1 < a_1 < b_2 < a_2. \end{cases}$$

For concreteness, we henceforth focus upon the case of Example 2, which supposes the existence of $\underline{\mathcal{X}}$ of length T such that $\underline{X} = J\,\underline{\mathcal{X}}$ for known J, which is $n \times T$ dimensional. Letting ι_T denote a T-vector of ones and $\tau_T = [1, 2, \cdots, T]'$,

$$\mathbb{E}[\underline{X}] = \mu_0\, J\, \iota_T + \mu_1\, J\, [\tau_T - \iota_T/2]. \tag{16}$$

We shall use the shorthand $\underline{W}_0 = J\,\iota_T$ and $\underline{W}_1 = J\,[\tau_T - \iota_T/2]$. For the covariance matrix, we can apply Lemma 2 with $\delta = 1$ and $t \ge 1$ to obtain

$$\mathrm{Cov}(X_{(t-1,t]}, X_{(t+h-1,t+h]}) = \sigma^2\,(t - 1/2)$$

for $h > 0$, while the variance of $X_{(t-1,t]}$ is $\sigma^2\,(t - 2/3)$. We can put this result in a matrix form, by letting Ω denote a $T \times T$ matrix with ones on all the lower triangular portion, with its inverse matrix denoted by $\Delta = \Omega^{-1}$ – this Δ corresponds to the action of first differencing. Then

$$\mathrm{Var}[\underline{\mathcal{X}}] = \sigma^2\, A \tag{17}$$

$$A = \Omega\Omega' - \frac{1}{6}\, I_T - \frac{1}{2}\, \iota_T \iota_T', \tag{18}$$

with I_T denoting the T-dimensional identity matrix. The matrix A does not depend on parameters, and is easily calculated. Furthermore, define

$$B = J\,A\,J'$$

so that $\mathrm{Var}[\underline{X}] = \sigma^2\,B$. Thus, given θ we can use (16), (17), and (18) to compute the mean and variance of $\underline{X}$, utilizing the Q-R decomposition of J.

Remark 1 Our discussion allows for missing values in the time series of MYEs – note that the structure of $\widehat{\underline{x}}_s$ is quite flexible. When working with ACS data that combines published estimates with the trial period of the MYES, there are potential gaps in the time series (see discussion in McElroy (2009)), which can be handled as a missing at random problem, wherein the matrix J is modified accordingly.

A model is also required for the sampling error variances. We assume that the sampling errors are aggregates of a continuous time Gaussian white noise process $\{\epsilon(t)\}$ for $t \in [0, \infty)$, whose autocovariance function is given by the Dirac delta function times a scaling parameter (its value will be irrelevant). Hence (suppressing the sampling mechanism in the notation) we define

$$E_{(t-\delta,t]} = \sqrt{\mathrm{Var}[E_{(t-\delta,t]}]}\,\delta^{-1/2} \int_{t-\delta}^{t} \epsilon(s)\,ds,$$

where the variance $\mathrm{Var}[E_{(t-\delta,t]}]$ will be supplied from the survey estimates. It follows that the correlation between sampling errors is

$$\mathrm{Corr}_S(E_{(t-\delta,t]}, E_{(\ell-\eta,\ell]}) = \frac{\min\{\delta, \max\{0, (t-\ell+\eta)\}\}}{\sqrt{\delta\eta}},$$

for some epoch lengths δ and η, and $t \leq \ell$. The covariance is

$$\mathrm{Cov}_S(E_{(t-\delta,t]}, E_{(\ell-\eta,\ell]}) = \sqrt{\mathrm{Var}[E_{(t-\delta,t]}]}\,\sqrt{\mathrm{Var}[E_{(\ell-\eta,\ell]}]}\,\mathrm{Corr}_S(E_{(t-\delta,t]}, E_{(\ell-\eta,\ell]})$$

In this way we can easily construct the matrix $\mathrm{Var}[\underline{E}_S]$. Note that, because the sampling errors for different periods are all a linear function of the same underlying process, there can be redundancies yielding a variance matrix of reduced rank. However, we do not require the matrix' inversion.

We require estimators of θ. For the regression parameters, we propose Generalized Least Squares (GLS). Setting $\mathbf{X} = [\underline{W}_0\,\underline{W}_1]$ and $\mu' = [\mu_0, \mu_1]$, we obtain the GLS estimator

$$\widehat{\mu} = C^{-1}\,\mathbf{X}'\,B^{-}\,\widehat{\underline{X}}_S \tag{19}$$

with $C = \mathbf{X}'\,B^{-}\,\mathbf{X}$. To obtain an estimator of σ^2, observe that with $\underline{R} = \widehat{\underline{X}}_S - \mathbf{X}\,\mu \sim \mathcal{N}(0, \sigma^2\,B + \mathrm{Var}[\underline{E}_S])$, it follows that

$$\widehat{\underline{X}}_S - \mathbf{X}\,\widehat{\mu} = B\,G\,\underline{R},$$

where $G = B^- - B^- \mathbf{X}\, C^{-1}\, \mathbf{X}'\, B^-$. We can construct the GLS estimator of σ^2 based on these residuals, and once we correct for the bias we obtain:

$$\widehat{\sigma}^2 = (n-2)^{-1}\, (\underline{\widehat{X}}_S - \mathbf{X}\, \widehat{\mu})'\, B^-\, (\underline{\widehat{X}}_S - \mathbf{X}\, \widehat{\mu}) - (n-2)^{-1}\, \mathrm{tr}\{G\, \mathrm{Var}[\underline{E}_S]\}. \tag{20}$$

Here tr denotes the trace. These estimators have small sample properties reviewed in the following result.

Proposition 2 *If $\underline{\widehat{X}}_S \sim \mathcal{N}(\mathbf{X}\,\mu, \sigma^2\, B + Var[\underline{E}_S])$ and B is invertible, then the estimators $\widehat{\mu}$ and $\widehat{\sigma}^2$ given by (19) and (20) are unbiased, with variances given by*

$$\begin{aligned} Var[\widehat{\mu}] &= \sigma^2\, C^{-1} + C^{-1}\, \mathbf{X}'\, B^{-1}\, Var[\underline{E}_S]\, B^{-1}\, \mathbf{X}\, C^{-1} \\ Var[\widehat{\sigma}^2] &= (n-2)^{-2}\, \left(2\,(n-2)\,\sigma^4 + 4\,\sigma^2\, tr\{G\, Var[\underline{E}_S]\} + 2\, tr\{G\, Var[\underline{E}_S]\, G\, Var[\underline{E}_S]\}\right). \end{aligned}$$

Proof of Proposition 2. Observe that

$$\widehat{\mu} = \mu + C^{-1}\, \mathbf{X}'\, B^{-1}\, \underline{R},$$

showing that the estimator is unbiased. The formula for the variance of $\widehat{\mu}$ also follows. From (20) we obtain the expression

$$\widehat{\sigma}^2 = (n-2)^{-1}\, \mathrm{tr}\left\{G\, \left(B\, G\, \underline{R}\, \underline{R}' - \mathrm{Var}[\underline{E}_S]\right)\right\},$$

whose expectation is σ^2, using $\mathrm{tr}(GB) = n-2$ and $G\, B\, G = G$. Hence

$$\mathrm{Var}[\widehat{\sigma}^2] = (n-2)^{-2}\, \mathrm{Var}[\underline{R}'\, G\, \underline{R}] = \frac{2}{(n-2)^2}\, \mathrm{tr}\{G\, \mathrm{Var}[\underline{R}]\, G\, \mathrm{Var}[\underline{R}]\},$$

which simplifies to the stated expression. □

Now we describe the interpolated kriging estimator for this model. Using (7), and setting $\sigma^2\, \underline{c}(t) = \mathrm{Cov}[\underline{\mathcal{X}}, Z]$, we have $\gamma = J\, \underline{c}(t)\, \sigma^2$; applying (14), we obtain

$$\widehat{Z} = \mu_0 + \mu_1(t - \delta/2) + \underline{c}(t)'\, J'\, B^-\, \left(\underline{\widehat{X}}_S - \mu_0\, \underline{W}_0 - \mu_1\, \underline{W}_1\right).$$

In the special case that $\underline{\widehat{x}}_s$ follows an exact linear pattern the kriging estimate reduces to $\widehat{Z} = \mu_0 + \mu_1(t - \delta/2)$, the mean value of the population estimand. The kth component of $\underline{c}(t)$ is determined from Lemma 2 with $a_2 = t$, $a_1 = t - \delta$, $b_2 = k$, $b_1 = k - 1$ for $k = 1, 2, \cdots, T$. Supposing that $t \geq \delta$ and $\delta \geq 1$, we find that $\underline{c}_k(t)$ is given by δ^{-1} times

$$\begin{cases} [t^2 - (t-\delta)^2]/2 & \text{if } t < k-1 \\ [t^2 - (t-\delta)^2]/2 - (t-k+1)^3/6 & \text{if } k-1 < t < k \\ [t^2 - (t-\delta)^2]/2 - (t-k+1)^3/6 + (t-k)^3/6 & \text{if } k < t < k-1+\delta \\ \delta[k^2 - (k-1)^2]/2 + (t-\delta-k)^3/6 & \text{if } k-1+\delta < t < k+\delta \\ \delta[k^2 - (k-1)^2]/2 & \text{if } t > k+\delta. \end{cases}$$

When $\delta < 1$, the expression is instead δ^{-1} times

$$\begin{cases} [t^2 - (t-\delta)^2]/2 & \text{if } t < k-1 \\ [t^2 - (t-\delta)^2]/2 - (t-k+1)^3/6 & \text{if } k-1 < t < k-1+\delta \\ [t^2 - (t-\delta)^2]/2 - (t-k+1)^3/6 + (t-k+1-\delta)^3/6 & \text{if } k-1+\delta < t < k \\ \delta[k^2 - (k-1)^2]/2 + (t-\delta-k)^3/6 & \text{if } k < t < k+\delta \\ \delta[k^2 - (k-1)^2]/2 & \text{if } t > k+\delta. \end{cases}$$

An instantaneous estimate can be obtained, by taking the limit as $\delta \to 0$ of $c_k(t)/\delta$. In this situation, there are only three cases of interest, and $\underline{c}_k(t)$ equals

$$\begin{cases} t & \text{if } t < k-1 \\ t - (t-k+1)^2/2 & \text{if } k-1 < t < k \\ [k^2 - (k-1)^2]/2 & \text{if } t > k. \end{cases}$$

These expressions for $\underline{c}(t)$ are also featured in the formulas for the kriging MSE. Applying Lemma 2 to determine $\mathrm{Var}[Z]$, and using $J' B^- J = A^{-1}$, we obtain from formula (15) that

$$\mathrm{MSE}[\widehat{Z}] = \sigma^2 (t - 2\delta/3) - \sigma^2 \underline{c}(t)' A^{-1} \underline{c}(t) + \underline{c}(t)' J' B^- \mathrm{Var}[\underline{E}_S] B^- J \underline{c}(t).$$

This formula also applies with $\delta = 0$, if point estimates are preferred over epoch-estimates.

5 Application to Veteran Population

The methodology of the previous section, using the framework of Example 1, is here applied to the variable "Total Veteran Population" stratified by veteran status, gender, and age. The values, along with sampling error confidence intervals, were generated by a custom tabulation – for the smaller counties, the 1-year estimate is not available on American Factfinder. The confidence intervals are constructed using implicit Gaussian quantiles for 95% coverage, and thus the sampling error variances can be deduced. For our analysis, the 1-year MYEs are available from 2006 through 2012. We begin by connecting the nomenclatures for these estimates in Table 1; for reference, see `http://www.census.gov/programs-surveys/acs/guidance/estimates.html`.

In this paper, time t is in units of years, so that the fraction .25 corresponds to a quarter of a year. Because $X_{(0,1]}$ denotes the first 1-year MYE available (for 2006), time $t = 0$ corresponds to the instant preceding the beginning of January 1, 2006. The ACS defines an MYE as covering a collection of years, where the actual times covered include January 1 of the start year through December 31 of the final year. For example, the 2006-2008 MYE actually covers times up until (but not including) the beginning of 2009. Such estimates are commonly (and fallaciously) viewed

Interval of Dates	ACS Designation	Notation	Midpoint
Jan.1, 2006 through Dec. 31, 2008	2006-2008 MYE	$X_{(0,3]}$	1.5
Jan.1, 2007 through Dec. 31, 2009	2007-2009 MYE	$X_{(1,4]}$	2.5
Jan.1, 2008 through Dec. 31, 2010	2008-2010 MYE	$X_{(2,5]}$	3.5
Jan.1, 2009 through Dec. 31, 2011	2009-2011 MYE	$X_{(3,6]}$	4.5
Jan.1, 2010 through Dec. 31, 2012	2010-2012 MYE	$X_{(4,7]}$	5.5
Jan.1, 2006 through Dec. 31, 2006	2006 MYE	$X_{(0,1]}$	0.5
Jan.1, 2007 through Dec. 31, 2007	2007 MYE	$X_{(1,2]}$	1.5
Jan.1, 2008 through Dec. 31, 2008	2008 MYE	$X_{(2,3]}$	2.5
Jan.1, 2009 through Dec. 31, 2009	2009 MYE	$X_{(3,4]}$	3.5
Jan.1, 2010 through Dec. 31, 2010	2010 MYE	$X_{(4,5]}$	4.5
Jan.1, 2011 through Dec. 31, 2011	2011 MYE	$X_{(5,6]}$	5.5
Jan.1, 2012 through Dec. 31, 2012	2012 MYE	$X_{(6,7]}$	6.5
Sept. 30, 2008	custom epoch MYE	$X_{\{2.75\}}$	2.75
Sept. 30, 2009	custom epoch MYE	$X_{\{3.75\}}$	3.75
Sept. 30, 2010	custom epoch MYE	$X_{\{4.75\}}$	4.75
Sept. 30, 2011	custom epoch MYE	$X_{\{5.75\}}$	5.75
Sept. 30, 2012	custom epoch MYE	$X_{\{6.75\}}$	6.75

Table 1: Correspondence of nomenclatures for various epochs, pertaining to the Total Veteran Population variable.

by users external to USCB not as period estimates, but as estimates of the middle time period, which would be noon of July 2 (or midnight July 2 for a leap year) of the middle year.

The desired application for the VA is to generate point estimates of September 30 of the final year in a given 3-year MYE. Whereas the 2006-2008 MYE is an estimate of $X_{(0,3]}$ we ultimately need to estimate $Z = X_{\{2.75\}}$. In order to make our estimate be more comparable to the published 2006-2008 MYE, we might restrict our data to just $X_{(0,1]}$, $X_{(1,2]}$, and $X_{(2,3]}$ for the purposes of computing $\widehat{X}_{\{2.75\}}$; while not optimal from the standpoint of throwing away information, such an estimate mimics the construction of the 2006-2008 MYE. However, we can also study the impact of including $X_{(0,3]}$ on $\widehat{X}_{\{2.75\}}$, and we consider both scenarios below. Table 1 displays the nomenclature for the three custom epoch quantities, along with five possible 3-year MYEs and seven possible 1-year MYEs corresponding to the entire time span.

Although the methodology has been tested on all counties and stratifications, for illustrative purposes we focus on the national level, aggregating across gender and age groups. Using the three 1-year MYEs and one 3-year MYE corresponding to each three year span yields five possible spans. Table 2 gives information about the data used in each span, and the parameter estimates obtained from (19) and (20). Initially we consider just the single year estimates of a span, and later we

Span	Data Used	$\widehat{\mu}_0$	$\widehat{\mu}_1$	$\log \widehat{\sigma}^2$
2006-2008	$X_{(0,1]}$, $X_{(1,2]}$, $X_{(2,3]}$, $X_{(0,3]}$	23.82	-.50	-5.50
2007-2009	$X_{(1,2]}$, $X_{(2,3]}$, $X_{(3,4]}$, $X_{(1,4]}$	23.26	-.52	-4.69
2008-2010	$X_{(2,3]}$, $X_{(3,4]}$, $X_{(4,5]}$, $X_{(2,5]}$	22.78	-.32	-1.36
2009-2011	$X_{(3,4]}$, $X_{(4,5]}$, $X_{(5,6]}$, $X_{(3,6]}$	22.03	-.20	-2.53
2010-2012	$X_{(4,5]}$, $X_{(5,6]}$, $X_{(6,7]}$, $X_{(4,7]}$	22.08	-.29	-4.31

Table 2: Spans utilized to calibrate parameter estimates (expressed in units of millions).

add the 3-year MYE of the span to see how the kriging estimates change. (But we use the same parameter calibrations, based on just using the single year estimates.)

For each of the five spans of Table 2, the three 1-year MYEs are used to calibrate the model. Clearly, determining three parameters from three data points cannot be conceived as inference, but these rough parameter calibrations are adequate for our application of generating custom epoch-estimates. By setting the epoch length δ to either 3, 1, or 0, and considering all times t pertaining to the total time span (so $t \in (0, 3]$ for the 2006-2008 span, and $t \in (1, 4]$ for the 2007-2009 span, etc.) we can generate a sequence of epoch estimates. The uncertainty can also be generated, which is defined as the square root of 1.96 times the kriging MSE (this generates a 95% confidence interval).

In Figure 1 the results are presented for the first span, with $\delta = 1$ in the left panels. The kriging estimates (red) intersect the right endpoints of the three single year estimates of the span in the upper left panel; however, the interpolation property is not guaranteed once the 3-year MYE is introduced, and hence in the lower left panel the kriging estimates deviate slightly from the right endpoints. This deviation is more evident in the spans of Figures 3 and 4, where the data does not adhere to a linear pattern. The left panels of Figures 2 and 5 are similar to those of the first span, because of the linear structure of the estimates.

Results for $\delta = 0$ are presented in the right panels of Figures 1 through 5. As expected, the kriging estimates pass below the single-year estimates, there being an expected shift of $\mu_1/2$. When the data are roughly linear, as in Figures 1, 2, and 5, the custom point estimates follow a linear trend line and have low uncertainty; but in Figures 3 and 4 there is more curvature in the kriging estimates to accommodate the change in the downward trend of Veteran Population, and the uncertainty is concomitantly greater. These are also the spans with a larger innovation variance σ^2 (Table 2).

The method can be utilized for other values of δ, including non-integer custom epoch lengths. We verified that when using two single year estimates and a 3-year MYE and setting $\delta = 3$, the method exactly replicates the 3-year MYE. We also computed custom estimates for all the available data (all single year estimates and 3-year MYEs exhibited in Table 2) for a variety of δ values, from $\delta = 5$ continuously down to $\delta = 0$. The general feature is that for larger epochs the custom

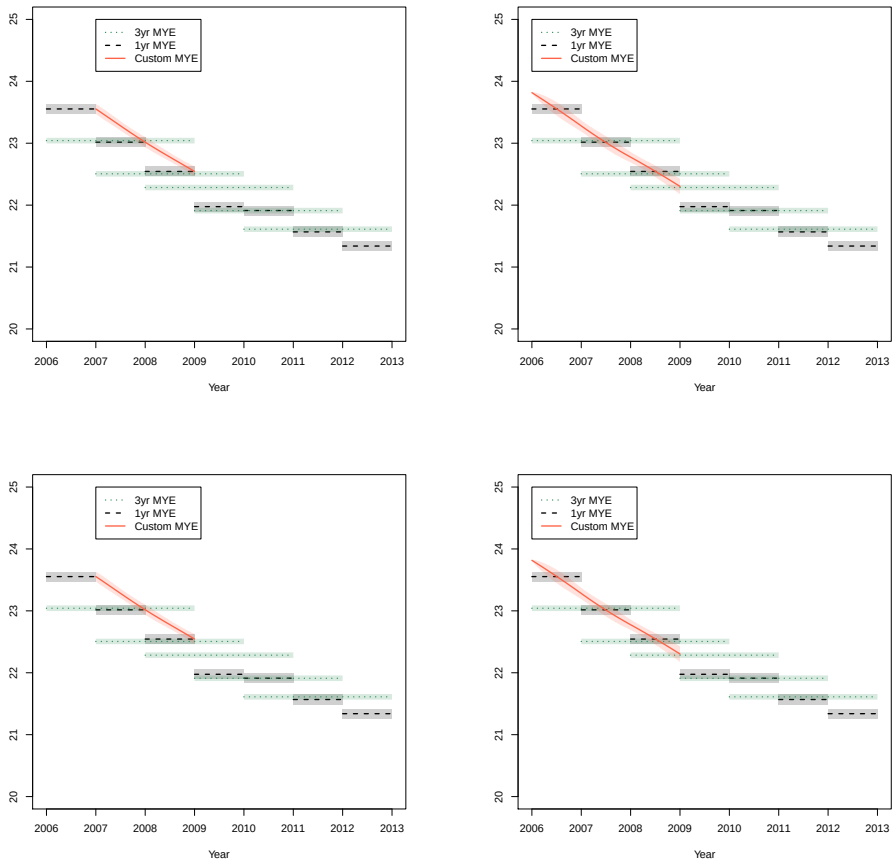

Figure 1: Custom epoch-estimates (red) for 2006-2008 Span of ACS (black), for $\delta = 1$ (left panels) and $\delta = 0$ (right panels). Upper panels correspond to using just single year estimates, while the lower panels add the 3-year MYE of the span. Shading corresponds to sampling uncertainty in the MYEs, and kriging uncertainty in the epoch-estimates.

estimates have a more linear structure with low kriging uncertainty, whereas the curvature and uncertainty increase (and the interpolants shift downwards) as δ decreases to zero.

6 Conclusion

This paper addresses an important applied concern with published MYEs of the ACS: how to construct customized epoch estimates? We present a methodology – applicable in a context of limited resources – to generate custom estimates with arbitrary length and arbitrary start date. The methodology avoids utilizing covariates or spatial correlation structure to build a finer model, as this would require a level of attention that is impractical for the publication goals. The method does smoothing and interpolation, being similar to a cubic spline; the parameters of the smoother are calibrated by applying simple formulas derived from a GLS methodology.

Our approach explicitly recognizes the presence of sampling error, but assumes it is indepen-

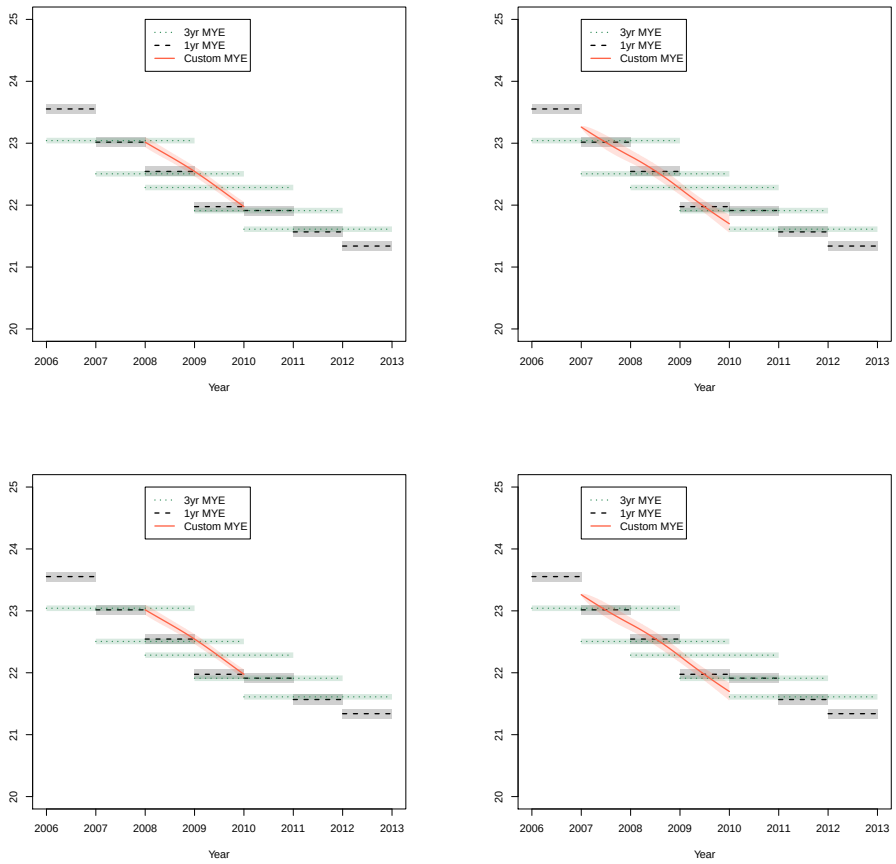

Figure 2: Custom epoch-estimates (red) for 2007-2009 Span of ACS (black), for $\delta = 1$ (left panels) and $\delta = 0$ (right panels). Upper panels correspond to using just single year estimates, while the lower panels add the 3-year MYE of the span. Shading corresponds to sampling uncertainty in the MYEs, and kriging uncertainty in the epoch-estimates.

dent of the population process. It is shown that when using Gaussian conditional expectation formulas the interpolation property is typically not satisfied, but by using a modification and by removing estimand redundancies the interpolation property can be guaranteed. This modification essentially eliminates the contribution of sampling error variability to the kriging formulas, and as a consequence the variability is increased.

These general results are further specified to the case where the population process is a Brownian Motion with linear drift, and the sampling errors are driven by a continuous time white noise. The resulting covariance functions are derived for applications. This method is applied to several spans of the veteran population variable of the ACS, demonstrating both the interpolation property and the facility to generate custom point estimates. An interesting extension would consider an integer-valued process, such as the Poisson, for the population estimand, with alternative derivations of parameter calibration and interpolation.

A criticism of this work states that custom epochs should not be estimated with a δ less

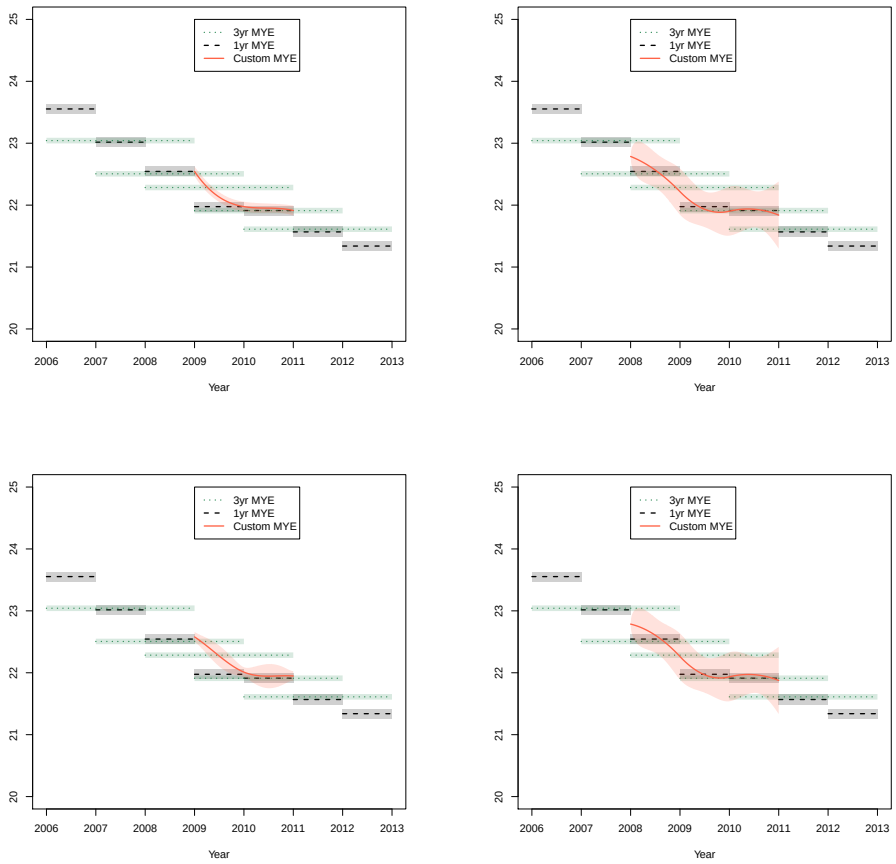

Figure 3: Custom epoch-estimates (red) for 2008-2010 Span of ACS (black), for $\delta = 1$ (left panels) and $\delta = 0$ (right panels). Upper panels correspond to using just single year estimates, while the lower panels add the 3-year MYE of the span. Shading corresponds to sampling uncertainty in the MYEs, and kriging uncertainty in the epoch-estimates.

than the minimal observed epoch length, because statistically we have no information about such short epochs. This is also a criticism of the model, which does not attempt to provide a nuanced description of such short epochs – these features are not identifiable given the data. However, such a critique can be offered against all applications of temporal or spatial interpolation. Thus, our response is that the results should be absorbed with caution: we can provide custom point estimates, contingent on a continuous-time model that cannot be verified at such small time scales.

More nuanced models of the superpopulation process could be entertained, say by considering more flexible serial dependence structures and/or non-Gaussian marginal distributions. Being unable to really validate our model choices anyways, due to the small sample sizes, we have not pursued these generalizations. It might also be of interest to utilize information from neighboring regions in a longitudinal approach to the problem, although the caution here is that such additional data will not furnish a direct improvement to the interpolation problem, because none of the available regions are observed at finer time scales. A functional data analysis could alternatively be

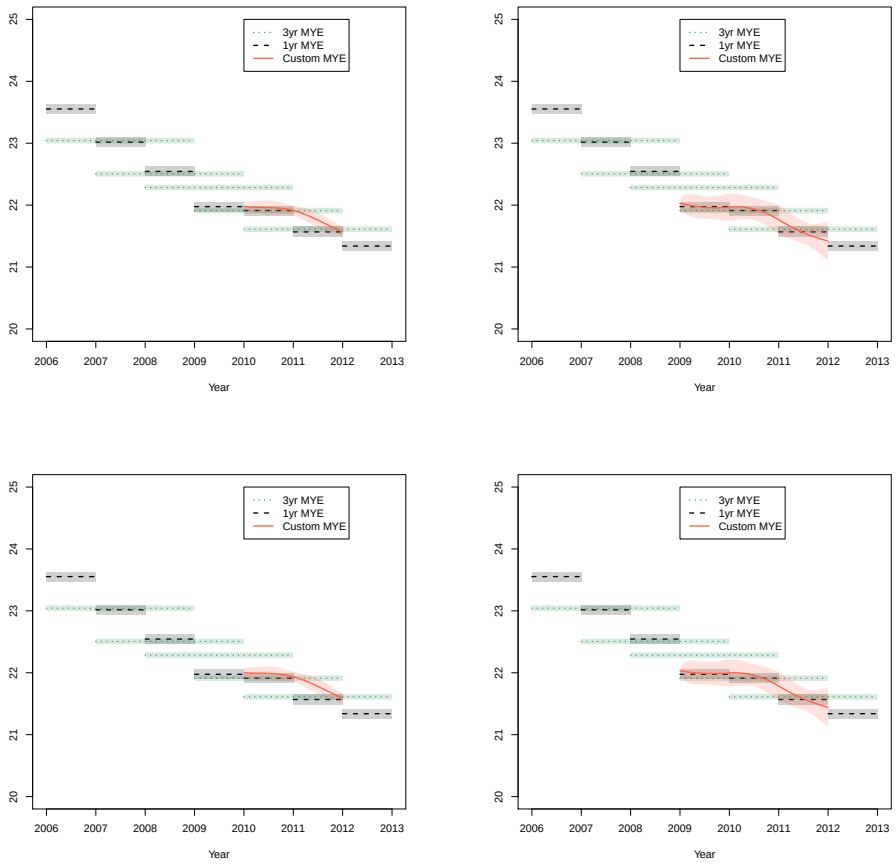

Figure 4: Custom epoch-estimates (red) for 2009-2011 Span of ACS (black), for $\delta = 1$ (left panels) and $\delta = 0$ (right panels). Upper panels correspond to using just single year estimates, while the lower panels add the 3-year MYE of the span. Shading corresponds to sampling uncertainty in the MYEs, and kriging uncertainty in the epoch-estimates.

applied, although the inability to indefinitely refine the temporal sampling scale limits the appeal.

Another suggestion that we have received, is that with access to the micro-data a user might directly compile monthly (or potentially higher frequency) estimates, which could then be smoothed to reduce the higher sampling error variability. Such a solution is viable for USCB employees with access, but all others must either beg for a custom tabulation (unlikely to be granted, due to resource constraints), obtain special sworn status themselves with access to a Census Data Center, or work with the published MYEs. This paper develops a methodology for users belonging to this third category.

Acknowledgements The authors are grateful for helpful criticism from Mark Asiala, Jonathan Bradley, Scott Holan, Patrick Joyce, Eric Slud.

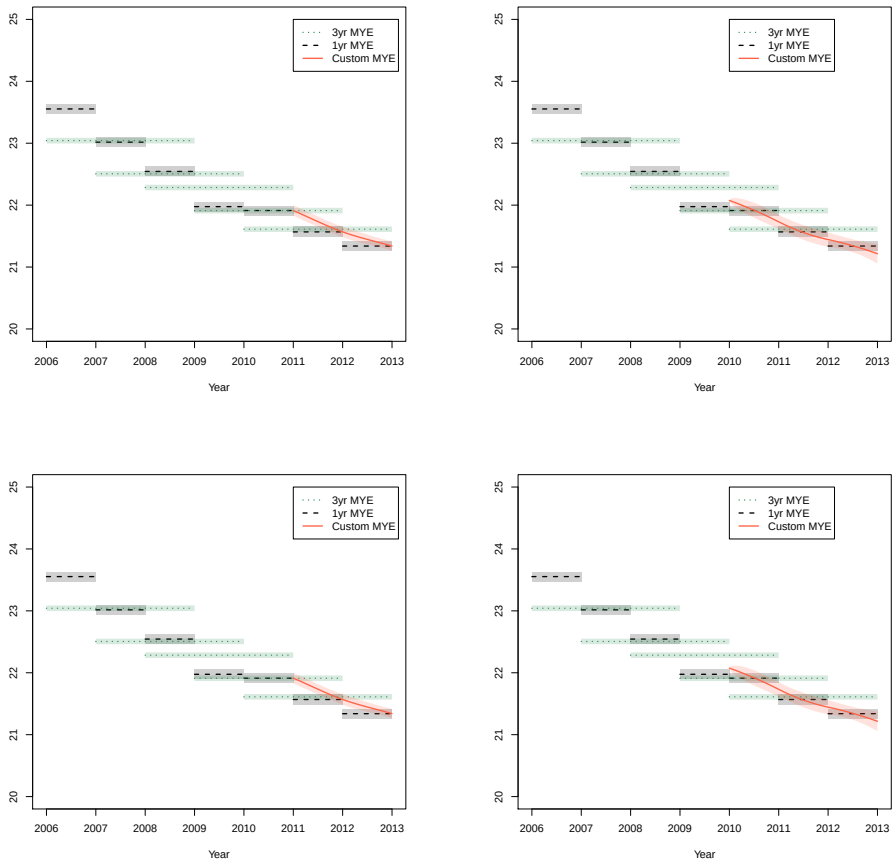

Figure 5: Custom epoch-estimates (red) for 2010-2012 Span of ACS (black), for $\delta = 1$ (left panels) and $\delta = 0$ (right panels). Upper panels correspond to using just single year estimates, while the lower panels add the 3-year MYE of the span. Shading corresponds to sampling uncertainty in the MYEs, and kriging uncertainty in the epoch-estimates.

References

[1] Bell, W.R. and Hillmer, S. (1990). The time series approach to estimation for repeated surveys. *Surv. Methodol.* **16**: 195–215.

[2] Bradley, J.R., Holan, S.H., and Wikle, C.K. (2015) Multivariate Spatio-Temporal Models for High-Dimensional Areal Data with Application to Longitudinal Employer-Household Dynamics. *The Annals of Applied Statistics* **9**: 1761–1791.

[3] Bradley, J.R., Wikle, C.K., and Holan, S.H. (2015) Spatio-Temporal Change of Support with Application to American Community Survey Multi-Year Period Estimates. *Stat.* **4**: 255–270.

[4] Fay, R. (2007) "Imbedding model-assisted estimation into ACS estimation," *Proceedings of the 2007 Joint Statistical Meetings on CD-ROM,* American Statistical Association, pp. 2946–2953.

[5] Golub, G.H. and Van Loan, C.F. (1996) *Matrix Computations*. Johns Hopkins University Press: Baltimore.

[6] McElroy, T. (2009) Incompatibility of Trends in Multi-Year Estimates from the American Community Survey. *The Annals of Applied Statistics* **3**, 1493–1504.

[7] Nagaraja, C. and McElroy, T. (2015) On the interpretation of Multi-Year Estimates of the American Community Survey as period estimates. *Journal of the International Association of Official Statistics* **31**, 661–676.

[8] Särndal, C.-E., Swensson, B., and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag: New York.

[9] Torrieri, N. (2007) America is Changing, and so is the Census: The American Community Survey," *The American Statistician* **61**, 16–21.

[10] U.S. Census Bureau (2006) Technical Paper 67. *Design and Methodology, American Community Survey.*
http://www.census.gov/acs/www/Downloads/tp67.pdf

[11] Zimmerman, D.L. and Stein, M. (2010) Classical geostatistical methods, in *Handbook of Spatial Statistics*, eds. Gelfand, A.E., Diggle, P.J., Fuentes, M., and Guttorp, P. Chapman and Hall: Boca Raton.