

Modeling Survey Time Series Data with Flow-Observed CARMA Processes

Journal of Official Statistics

2024, Vol. 40(4) 601–632

© The Author(s) 2024

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0282423X241286236

journals.sagepub.com/home/jof



Patrick M. Joyce¹  and Tucker S. McElroy²

Abstract

Published survey data often are delivered as estimates computed over an epoch of time. Customers may desire to obtain survey estimates corresponding to epochs, or time points, that differ from the published estimates. This “change of support” problem can be addressed through the use of continuous-time models of the underlying population process, while taking into account the sampling error that survey data is subject to. The application of a Continuous AutoRegressive Moving Average (CARMA) model is investigated as a tool to provide change of support applications, thereby allowing interpolation for published survey estimates. A simulation study provides comparisons of competing estimation methods, and a synthetically constructed data set is developed in order to elucidate real data applications. The proposed method can be successful for change of support problems, despite modeling challenges with the CARMA framework.

Keywords

interpolation, continuous-time series, American Community Survey

1. Introduction

Published survey data typically correspond to regions and time periods (or epochs) that a statistical office deems sufficient to guard against disclosure, while also providing sufficient quality for policy inferences (Janicki et al. 2022). For example, the American Community Survey (ACS) publishes data for geographies larger than

¹Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC, USA

²Research and Methodology Directorate, U.S. Census Bureau, Washington, DC, USA

Corresponding author:

Patrick M. Joyce, Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233, USA.

Email: patrick.m.joyce@census.gov



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

65,000 persons for one-year epochs; however, for sparsely populated regions and smaller geographies such as Census tracts, only the longer epochs of five-years are available (United States Census Bureau 2014). Demand from governmental stakeholders as well as the general public for population features measured at increasingly granular points in time and space continues to grow. In addition, users of survey data may require customized data publications for non-standard regions or epochs. For example, the Veterans Affairs (VA) utilizes the ACS data for its internal calculations, but requires epochs corresponding to September 30 of each given year (McElroy et al. 2019).

Whereas direct calculation of custom epoch estimates from ACS micro-data is possible, such estimates might fail to meet publication standards on estimation quality and data privacy; moreover, additional resources are required to compute such custom tabulations. The goal of this paper is to develop methodology for custom epoch estimates computed from publicly available data via continuous-time models; these models are straightforward to fit and apply, thereby keeping the resource burden low. Whereas McElroy, Pang, and Sheldon 2019 (henceforth MPS) addressed this problem of custom epoch estimation using the same strategy of continuous-time modeling, in this paper we utilize a richer class of continuous-time stochastic processes toward the end of improving the accuracy of interpolations required for customized estimation. In the same spirit as MPS, we isolate the temporal aspect of the problem—essentially ignoring the spatial features of ACS estimates—on the grounds of model simplicity, leaving spatio-temporal interpolation for future work.

The problem of custom epoch estimation can be framed as a “change of support” problem, as it is described in the spatial-temporal literature (Bradley et al. 2015, 2016). In our paper the change of support problem is approached with continuous-time modeling tools, offering an extension of MPS that generalizes the Fay-Herriot model (henceforth FH; see Fay and Herriot 1979; Rao and Molina 2015). Whereas the methodology of MPS involves an interpolation constraint that removed the impact of sampling error variability on interpolated values, in this paper this restriction is relaxed. Another contrast is that the underlying model of MPS was a Brownian Motion with linear drift, which enforces a simple trend structure on the data; however, a more flexible model class is given by the CARMA (Continuous AutoRegressive Moving Average) processes (with regression effects), which allows for more nuanced dynamics (such as business cycles and seasonality). In FH there is no temporal structure assumed; although there are extensions of FH to discrete time series structure, to our knowledge the only continuous-time analogue of the FH model is the Brownian Motion model of MPS, which reduces (as does our work) to the FH model when the continuous-time process is white noise.

CARMA processes are the stationary solutions to a linear stochastic differential equation; see Doob (1944), Jones (1981), Bergstrom (1983), and Brockwell (1995). The covariance structure for the stationary CARMA class is given by Brockwell (2000), with generalizations in Brockwell (2001, 2004, 2009), and Brockwell and Marquardt (2005). Chambers and Thornton (2012) shows that CARMA under regular sampling can be represented by discrete ARMA (AutoRegressive Moving

Average) models, and Thornton and Chambers (2017) works with a mixture of stock and flow series and their discrete ARMA representations in the multivariate setting; see McElroy and Politis (2019) for background on ARMA processes and characteristic polynomials. These known covariance results are applied to obtain interpolation (or kriging) formulae used to generate this paper's custom epoch flow estimates. However, our implementation of the CARMA model is restricted to first and second order Continuous AutoRegressions (CAR), as the parameterization of higher order CAR (and CARMA) remains a challenging problem.

The methodology of this paper treats the change of support problem for data observed over epochs, or subsets of time, expressed as averages of the population variable over the epoch; whereas in a spatial context such averages are referred to as "areal data," in the econometric time series literature they are referred to as "flow data," or "flows" for short; see Chambers and McGarry (2002) and Harvey and Trimbur (2003) for terminology and definitions. Whereas the flow structure involves an average, and therefore may not be appropriate for some variables, this is a common device for describing areal data (Bradley et al. 2017), and this formulation is adopted into our temporal framework. We remark that the continuous-time formulation allows us to interpolate to arbitrary time points; a discrete-time approach, in contrast, does not permit this fuller flexibility. Finally, the role of sampling error is vital, as the increased uncertainty associated with narrower epochs is important to communicate to users.

The rest of this paper is organized as follows: Section 2 discusses the framework of surveys collected over time. Hierarchical modeling and small domain techniques are also reviewed. The main innovation of the paper is the application of CARMA processes to a small domain setting (and how custom epoch inferences can be obtained); this is described in Section 3. Section 4 summarizes a series of simulation studies that address practical inference issues that arise from CARMA models, and Section 5 provides a specific application upon a non-public monthly test data set of the ACS. Section 6 provides conclusions. Supplemental S-1 contains technical results on covariance forms of points and epochs for Brownian motion and for CARMA models, Supplemental S-2 provides further detail on simulations studying a comparison of estimators, and Supplemental S-3 contains plots of a synthetic data simulation. All references originating from the supplemental data file are denoted by S-#.

2. Survey Framework and Small Domain Review

The modeling of survey data (i.e., characterizing the sampling distribution of estimates) in official statistics is a common practice within the small area literature. The term "small area" refers to a region (temporal or spatial) or categorization whose extent is so small that it contains few sampling units; hence, estimates constructed from such small areas (also called "small domains") have high variability. Typically, the distribution of the survey statistic is modeled with a normal distribution, expressing the observed statistic as a true unknown mean plus a latent error, whose variance is obtained from the survey estimate of the statistic's variance (or

some other data-derived proxy). In the small area (or small domain) literature, small areas correspond to geographical regions; the most common model in this context is the FH model (Fay and Herriot 1979).

Supposing that Y_i (for $1 \leq i \leq m$) denotes a quantity or population parameter of interest (e.g., a population total) for the i -th domain (here, an epoch), its raw survey estimate (usually the estimator of Horvitz and Thompson (1952)) is denoted by \hat{Y}_i ; in the FH model this estimate is assumed to have a normal sampling distribution

$$\hat{Y}_i | Y_i \sim N(Y_i, \sigma_i^2). \quad (1)$$

Note that the mean value is Y_i , and σ_i^2 is the sampling variance. The survey variance estimator $\hat{\sigma}_i^2$ will be used to estimate σ_i^2 ; in the FH formulation the true variance is assumed to be known, but in practice is unknown. Each Y_i is a latent random variable whose mean is expressible in terms of known non-random covariates X_i (of size p):

$$Y_i \sim N(X_i \boldsymbol{\beta}, \tau^2). \quad (2)$$

Here $\tau^2 > 0$ is the variance of the regression errors, and $\boldsymbol{\beta}$ is the vector of regression parameters. This can be rewritten in matrix form as follows. Supposing there are m small domains, we have $\hat{\mathbf{Y}} | \mathbf{Y} \sim N(\mathbf{Y}, \mathbf{D})$ and $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \tau^2 \mathbf{I})$, where $\hat{\mathbf{Y}}$ and \mathbf{Y} are m -dimensional vectors, $\mathbf{D} = \text{Diag}(\{\hat{\sigma}_i^2\}_{i=1}^m)$ the diagonal matrix of survey error variances, and \mathbf{X} is a $m \times p$ covariate matrix. Assuming the independence of sampling error and regression error, the vector normal distribution yields the vector marginal form

$$\hat{\mathbf{Y}} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{D} + \tau^2 \mathbf{I}). \quad (3)$$

Assuming the parameters are known, we can use Bayes Theorem to obtain the conditional distribution for the estimand:

$$\mathbf{Y} | \hat{\mathbf{Y}} \sim N\left(\mathbf{D}^{-1}(\mathbf{D}^{-1} + \tau^{-2} \mathbf{I})^{-1} \hat{\mathbf{Y}} + \tau^{-2}(\mathbf{D}^{-1} + \tau^{-2} \mathbf{I})^{-1} \mathbf{X}\boldsymbol{\beta}, (\mathbf{D}^{-1} + \tau^{-2} \mathbf{I})^{-1}\right). \quad (4)$$

The unknowns $(\tau^2, \boldsymbol{\beta})$ can then be estimated through maximum likelihood estimation (subject to the constraint $\tau^2 \geq 0$) or through Bayesian methods; Gelman (2006) uses the prior $\pi(\boldsymbol{\beta}, \tau^2) = \pi(\boldsymbol{\beta}) \times \pi(\tau^2)$, $\boldsymbol{\beta} \sim N(\mathbf{0}, 10^6 \mathbf{I})$, and $\pi(\tau^2) \propto (1 + (\tau^2)^2)^{-1} I[\tau^2 \geq 0]$. For a multivariate Bayesian extension of the Fay-Herriot model see Arima et al. (2017). The maximum likelihood solution of parameter fits of Equation (3) applied to the posterior mean of Equation (4) constitutes the empirical best linear unbiased predictor. This paper restricts estimation to maximum likelihood (empirical Bayes) estimation; predictions are then made by using the conditional mean of $\mathbf{Y} | \hat{\mathbf{Y}}$. The conditional mean of Equation (4) is an element-wise linear combination of survey observation vector $\hat{\mathbf{Y}}$ with weights $\hat{\sigma}_i^2 / (\tau^2 + \hat{\sigma}_i^2)$ and regression predictor vector $\mathbf{X}\hat{\boldsymbol{\beta}}$ with weight $\tau^2 / (\tau^2 + \hat{\sigma}_i^2)$, which emphasizes

the relationship in the estimator between survey error and model error. The estimator is similar to that of James and Stein (1961).

In the case that the domains are epochs, we can view the Y_i as a time series, and some authors have developed time series models that generalize Equation (2); see Bell and Hillmer (1987) and Bell and Hillmer (1990). Tiller (1992) expresses this basic formulation in the state space framework. Similarly, Rao and Yu (1994) provides an extension of small area models to time series data, formulating a model that includes a temporal component while allowing for cross-sectional modeling. Ghosh et al. (1996) provides a Bayesian structural time series model with small area regression covariates. Pfefferman and Tiller (2006) and Pfefferman et al. (2014) provide state space models for small area estimation while involving benchmarking. Further citations and details can be found in Rao and Molina (2015). However, we emphasize that this literature (including the spatio-temporal work of Bradley et al. (2015)) is focused entirely upon discrete time series frameworks, which do not easily allow interpolation nor allow for varying epoch sizes; the integration of continuous-time processes into a survey framework has received little attention so far.

A few papers examine the uncertainty of the sampling variance's sample-based estimator. You and Chapman (2006) applies a Bayesian small area model, approaching the sampling variance as if the sample variance estimator's sampling distribution follows a χ^2 distribution. Sugawara et al. (2017) follows a similar Bayesian construction, but considers the application of shrinkage toward the variance estimator itself in addition to that of the survey estimator. Consideration of such forms is beyond the reach of this paper.

3. Custom Epoch Estimation

This section first describes the mathematical framework for temporal change of support, and secondly specializes to a discussion of CARMA processes, which is the main innovation of the paper. The third and fourth subsections further specialize to CAR processes of order 1 and 2, which are the only cases that are implemented so far.

3.1. Change of Support Problems for Continuous Time Processes

Cross-sectional surveys (such as the ACS) can represent either single time points (i.e., point domains) or time periods (i.e., epochs). Whereas classical sampling mechanisms (Lohr 2010) do not consider the flow of time, here a discussion of surveys collected over time is provided so that population estimands and their corresponding sample estimates are clearly delineated.

The ACS is a cross-sectional survey conducted by using a rolling monthly panel, and is produced on an annual basis. This survey produces one-year and five-year estimation products for various geographies (see United States Census Bureau 2023b). Single year ACS estimates represent larger geographies, whereas the five-year ACS estimates represent both these larger geographies as well as smaller

geographies. Both products are assumed to represent their related epoch. The ACS is not the only survey that is assumed to represent an epoch; some governmental surveys report activity over an entire month (e.g., the Current Population Survey (United States Census Bureau 2022)), whereas many polls represent weeks or a few days (e.g., Gallup Research (Jones 2020)). In all of these cases there is a common need to express population parameters as an average over an epoch.

Surveys collected over time yield flow estimands computed over epochs that are dictated by the survey construction. A user may instead require an estimate for a different epoch than those being published; this is known as the “change of support” problem (Cressie 1993), and is similar to interpolation, where a continuous time model is needed to describe the data process at times that are not directly observed. As in the previous section, Y denotes a quantity of interest for the population, and this is viewed as a function of time t , denoted as $Y(t)$. Then the population process is denoted by $\{Y(t)\}$, which is assumed to be a stochastic process; a so-called stock at time t is simply $Y(t)$ (by definition), whereas the flow over an epoch $A \subset \mathbb{R}$ is defined as

$$|A|^{-1} \int_A Y(t) dt \quad (5)$$

where $|A|$ denotes the Lebesgue measure of A . (Most epochs of interest are intervals.)

The treatment of stock and flow is unified by recognizing that $Y(t)$ is identical to Y_A as A shrinks to the singleton $\{t\}$; heuristically, $Y_{\{t\}} = Y(t)$. Supposing that $\{Y(t)\}$ has mean function $\mu(t) = E[Y(t)]$ and covariance function $C(t, s) = \text{Cov}[Y(t), Y(s)]$, it follows that the flow mean over A is given by $|A|^{-1} \int_A \mu(t) dt$, and its variance is $|A|^{-2} \int_{A \times A} C(t, s) dt ds$. Likewise, the covariance between two flows is

$$\text{Cov}(Y_A, Y_B) = |A|^{-1} |B|^{-1} \int_A \int_B C(t, s) dt ds, \quad (6)$$

where $A, B \subset \mathbb{R}$. This formula can be generalized to include the case that either or both of A and B are singletons by simply replacing scaled integration by the Dirac functional, that is,

$$\text{Cov}(Y_A, Y(s)) = |A|^{-1} \int_A C(t, s) dt. \quad (7)$$

Whereas the above expressions hold for a generic square integrable continuous time process, it is useful to assume that the population process $\{Y(t)\}$ is Gaussian, which is denoted by $Y \sim GP(\mu, C)$.

Two more practical modifications are made to the Gaussian process. First, note that $C(t, s)$ for parametric covariances can be written as $C(t, s) = \tau^2 C^*(t, s; \theta)$, where $\tau^2 > 0$ and θ is a set of model parameters relating to the covariance function C^* . For instance, C^* can be a correlation function, in which case τ^2 (independent of t) is the variance; alternatively, τ^2 could equal the variance of the innovations of the Y process, when this is stationary. Second, we assume that

$$\mu(t) = \mathbf{x}(t)\boldsymbol{\beta}, \quad (8)$$

where row-vector $\mathbf{x}(t)$ is a non-random vector-valued covariate function known on A —for example, $\mathbf{x}(t) = (1, t)$.

Applying the discussion in Section 2, the sampling distribution of the survey-based statistical estimates $\hat{Y}_{A_i} = Y_{A_i}$ (for any given epoch A_i with $1 \leq i \leq m$) are available to us; utilizing (1) gives

$$\hat{Y}_{A_i} | Y_{A_i} \sim N\left(Y_{A_i}, \sigma_{\hat{Y}_{A_i}}^2\right),$$

where $\sigma_{\hat{Y}_{A_i}}^2$ is the true sampling error variance of \hat{Y}_{A_i} . The provisional assumption is that these true sampling error variances are known. Furthermore, since the multi-year vintages of the ACS produce sample survey estimates from common data, it is plausible that there is correlation between the sampling errors for distinct epochs A_i and A_j . To our knowledge, there is no public empirical measure of such correlations for the ACS, and we adopt a correlation measure based on the proportion of common length shared by two epochs:

$$\rho_{i,j}^* = \frac{|A_i \cap A_j|}{\sqrt{|A_i||A_j|}}. \quad (9)$$

Such a formula for the correlation amounts to an unverifiable assumption to those external to the U.S. Census Bureau, and is indeed restrictive (as it presumes a lack of bunching in the data collection), but is arguably the simplest formulation in this context (it is also adopted in MPS; future research could investigate whether such an assumption is empirically sound). It follows that the covariance matrix of the sampling errors is defined by

$$D_{i,j}^* = \rho_{i,j}^* \sigma_{\hat{Y}_{A_i}} \sigma_{\hat{Y}_{A_j}}$$

for $1 \leq i, j \leq m$. Hence the vector of survey estimates $\hat{\mathbf{Y}} = (\hat{Y}_{A_1}, \dots, \hat{Y}_{A_m})$ has joint sampling distribution

$$\hat{\mathbf{Y}} | \mathbf{Y} \sim N(\mathbf{Y}, \mathbf{D}^*),$$

where $\mathbf{Y} = (Y_{A_1}, \dots, Y_{A_m})$ is the vector of latent epoch averaged population parameters. Note that if all the epochs A_i are disjoint, then $\mathbf{D}^* = \mathbf{D}$, the diagonal matrix of sampling variances defined in Section 2.

From the assumptions on the process $\{Y(t)\}$, which is governed by parameters $(\boldsymbol{\beta}, \tau^2, \theta)$, the joint distribution of $\hat{\mathbf{Y}}$, \mathbf{Y} , and $Y(t)$ can be derived. For a particular epoch A_i , it follows from Equation (5) that $E[Y_{A_i}] = \mathbf{x}_{A_i}\boldsymbol{\beta}$, where

$$\mathbf{x}_{A_i} = |A_i|^{-1} \int_{A_i} \mathbf{x}(t) dt. \text{ Set } \mathbf{X} = (\mathbf{x}_{A_1}, \dots, \mathbf{x}_{A_m}), \text{ so that } E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}.$$

The covariance matrix of \mathbf{Y} is obtained from Equation (6), yielding $\mathbf{C}(\boldsymbol{\theta}) = \tau^2 \mathbf{C}^*(\boldsymbol{\theta})$. Further, define the column-vector $\mathbf{c}_B(\boldsymbol{\theta}) = \tau^2 \mathbf{c}^*(\boldsymbol{\theta})$ to be the covariance between Y_B and the collection of Y_{A_i} , $\mathbf{C}_{B,B}(\boldsymbol{\theta}) = \tau^2 \mathbf{C}_{B,B}^*(\boldsymbol{\theta})$ to be the variance of Y_B , and the column-vector $\mathbf{c}(t; \boldsymbol{\theta}) = \tau^2 \mathbf{c}^*(t; \boldsymbol{\theta})$ to be the covariance between

$Y(t)$ and the collection of Y_{A_i} from Equation (7). Hence the distribution of $(\hat{Y}, Y, Y_B, Y(t)) | (\beta, \tau^2, \theta)$ is given by

$$\begin{pmatrix} \hat{Y} \\ Y \\ Y_B \\ Y(t) \end{pmatrix} | \begin{pmatrix} \beta \\ \tau^2 \\ \theta \end{pmatrix} \sim N \left(\begin{pmatrix} X\beta \\ X\beta \\ x_B \\ x(t) \end{pmatrix}, \tau^2 \begin{bmatrix} \frac{1}{\tau^2} D^* + C^*(\theta) & C^*(\theta) & c_B^*(\theta) & c^*(t; \theta) \\ C^{*'}(\theta) & C^*(\theta) & c_B^*(\theta) & c^*(t; \theta) \\ c_B^{*'}(\theta) & c_B^{*'}(\theta) & C_{B,B}(\theta) & c_B^*(t; \theta) \\ c^{*'}(t; \theta) & c^{*'}(t; \theta) & c_B^*(t; \theta) & C^*(t, t; \theta) \end{bmatrix} \right) \quad (10)$$

for any t . In particular,

$$\hat{Y} \sim N(X\beta, D^* + \tau^2 C^*(\theta)) \quad (11)$$

which will be used to estimate model parameters, and is the composition of survey error and continuous-time process error.

This treatment generalizes the FH model when epochs are disjoint and have equal length, that is, $|A_i| = \alpha$ for all i . This occurs as Equation (3) is recovered from Equation (11) by setting $C(t, s) = \delta(s - t)$, where δ is the Dirac delta function; moreover, the variance in (11) is $D^* + \tau^2 \alpha^{-1} I$, whereas the variance in Equation (3) is $D^* + \tau^2 \alpha^{-1} I$, and these are equivalent if the constant α^{-1} is incorporated into the FH parameter τ^2 .

The problem of custom epoch estimation is to obtain the distribution of $Y_B | \hat{Y}$, where B is any given epoch desired by the user; again, typically B is an interval or a point. From Equation (10) it follows that $Y_B | \hat{Y} \sim N(\mu_B^\dagger, \sigma_B^{2, \dagger})$,

Where

$$\mu_B^\dagger = \mu_B^\dagger(\beta, \tau^2, \theta) = x_B \beta + \tau^2 c_B^*(\theta)' (D^* + \tau^2 C^*(\theta))^{-1} (\hat{Y} - X\beta) \quad (12)$$

and

$$\sigma_B^{2, \dagger} = \sigma_B^{2, \dagger}(\beta, \tau^2, \theta) = \tau^2 C_{B,B}^*(\theta) - \tau^4 c_B^*(\theta)' (D^* + \tau^2 C^*(\theta))^{-1} c_B^*(\theta). \quad (13)$$

μ_B^\dagger is referred to as the custom epoch predictor, which is a small domain estimate when B is some A_i . In the special case that $B = A_i$, then μ_B^\dagger is the BLUP for Y_{A_i} . When B is a singleton $\{t\}$, the conditional distribution $Y(t) | \hat{Y} \sim N(\mu^\dagger(t), \zeta^\dagger(t))$ is obtained, where

$$\mu^\dagger(t) = \mu^\dagger(t; \beta, \tau^2, \theta) = x(t) \beta + \tau^2 c^*(t; \theta)' (D^* + \tau^2 C^*(\theta))^{-1} (\hat{Y} - X\beta) \quad (14)$$

and

$$\zeta^\dagger(t) = \zeta^\dagger(t; \beta, \tau^2, \theta) = \tau^2 C^*(t, t; \theta) - \tau^4 c^*(t; \theta)' (D^* + \tau^2 C^*(\theta))^{-1} c^*(t; \theta). \quad (15)$$

$\mu^\dagger(t)$ is referred to as a point domain predictor.

The estimation of the model form using Equation (11) and its likelihood, denoted $L(\hat{Y}; x, \beta, \tau^2, \theta)$, is carried out via the optimization of the log-likelihood,

and hence $(\hat{\beta}, \hat{\tau}^2, \hat{\theta}) \operatorname{argmax}_{(\hat{\beta}, \hat{\tau}^2, \hat{\theta})} \log L(\hat{Y}; x, \beta, \tau^2, \theta)$. The normal equations for β can be solved in terms of the parameters (τ^2, θ) , and are given by

$$\hat{\beta}(\tau^2, \theta) = \left(X' (D^* + \tau^2 C^*(\theta))^{-1} X \right)^{-1} X' (D^* + \tau^2 C^*(\theta))^{-1} \hat{Y}. \quad (16)$$

Plugging Equation (16) into the log likelihood, the desired solution is obtained by optimizing the result with respect to τ^2 and θ , obtaining $(\hat{\tau}^2, \hat{\theta}) := \operatorname{argmax}_{(\tau^2, \theta)} \log L(\hat{Y}; x, \hat{\beta}(\tau^2, \theta), \tau^2, \theta)$. β is then found from $\hat{\beta}(\hat{\tau}^2, \hat{\theta})$, which yields the custom epoch estimator $\hat{\mu}_B^\dagger(\hat{\beta}, \hat{\tau}^2, \hat{\theta})$ and the point domain estimator $\mu^\dagger(t) = \mu^\dagger(t; \hat{\beta}, \hat{\tau}^2, \hat{\theta})$. The complete optimization of the likelihood for the CARMA model is addressed in Subsection 3.4. We remark that procedures to obtain a maximum likelihood, which will obtain a local maximum, may not find the global maximum.

3.2. Brownian Motion and CARMA as the Population Process

In the previous subsection C^* is a general covariance function, and so is applicable to a wide range of temporal change of support problems. Here we develop two particular choices of the covariance function, corresponding respectively to a Brownian Motion model (as proposed in MPS) and a CARMA model. The CARMA model is chosen because it is a stationary continuous-time model that has been well-studied, having known formulas for covariance easily obtained in terms of parameters that have a natural interpretation (associated with a stochastic differential equation that the CARMA process satisfies); moreover there are appealing connections to the popular discrete-time ARMA model, which forms a dense model class within the space of stationary discrete-time processes. While CARMA offers greater flexibility than Brownian Motion, and hence an improved ability to capture the dynamics of the population process, there are more parameters to estimate, which makes the modeling endeavor more challenging.

The use of continuous-time processes has a long history, especially in the domain of financial mathematics (Steele 2012 as an example). MPS proposed Brownian Motion to model the population process, addressing the problem of estimating $Y(t)$ —a single point in time—within the span of the available data. A simple linear trend function (corresponding to covariate functions $x(t) = (1, t)$) along with scaled Brownian Motion was used for the Gaussian process modeling $\{Y(t)\}$. Formulae for C^* in this case can be found in Supplemental S-1 (as well as the appendix of MPS); there are no parameters θ in this case, and an estimate of β is obtained from

$$\hat{\beta}(\tau^2) \left(X' (D^* + \tau^2 C^*)^{-1} X \right)^{-1} X' (D^* + \tau^2 C^*)^{-1} \hat{Y}. \quad (17)$$

Further, τ^2 can be obtained by optimizing a criterion discussed in MPS. A constraint on the interpolation estimator was also employed in MPS, but here it is preferred to instead use Equations (12) and (13).

For a standard Brownian Motion model, $C^*(t, s) = \min\{t, s\}$ has no covariance parameters to estimate, although a scaling parameter τ^2 can be used. However, the Brownian Motion implies non-stationary trend growth—which might not be realistic for some time series data—and is too simplistic to describe cyclical dynamics, or other more nuanced stationary process dynamics. Although rigid cyclical dynamics could be modeled through the inclusion of sinusoidal regressors, such a device will not capture evolutive dynamics, for which a stochastic approach—such as a CARMA process—is required. Indeed, the CARMA class of processes can flexibly describe many types of stationary dynamics, and has a long history within the stochastic process literature (Doob 1944)—although applications have been somewhat limited due to implementation challenges.

The covariance structure of a mean zero stationary Gaussian process can be summarized through the spectral density f , an integrable function of frequencies $\omega \in \mathbb{R}$. In particular (see Trimbur and McElroy 2017 for details) the covariance function for a stationary process only depends on the lag h , and thus $C(t + h, t)$ can be written as $C(h)$. Similarly, the autocorrelation function is denoted $C^*(h)$. Then the spectral density is defined as the Fourier transform of the autocovariance function:

$$f(\omega) = \int_{-\infty}^{\infty} C(h) e^{-i h \omega} dh, \omega \in \mathbb{R}.$$

Here $i = \sqrt{-1}$. It is known that this function f is non-negative so long as $C(h)$ is the autocovariance function of a stationary process. Conversely, the autocovariance function can be recovered from the spectral density f via Fourier inversion:

$$C(h) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(\omega) e^{i h \omega} d\omega.$$

Hence, we can specify a zero-mean stationary Gaussian process by proposing an integrable non-negative function f for the spectral density; the CARMA class amounts to taking $f(\omega)$ to be a rational function in ω^2 .

Specifically (following Brockwell (2000)), a zero-mean Gaussian process $\{Y(t)\}$ is a CARMA process of order p, q (with $0 \leq q < p$), denoted CARMA(p, q), if its spectral density takes the form

$$f(\omega) = \frac{|b(i\omega)|^2}{|a(i\omega)|^2} \quad (18)$$

for $\omega \in \mathbb{R}$, where $b(z)$ and $a(z)$ are real polynomials of order q and p , respectively. These polynomials are written in a particular format:

$$a(z) = z^p + a_1 z^{p-1} + \dots + a_p$$

(which is called the characteristic polynomial), and

$$b(z) = b_0 + b_1z + \cdots + b_qz^q,$$

where it is assumed that $b_q \neq 0$. While it is clear that the CARMA spectral density (18) is non-negative, we must make additional assumptions on $a(z)$ and $b(z)$ to guarantee its integrability. For this reason, we assume that $q < p$ (which ensures that the tails of f have sufficient rate of decay) and that every root of $a(z)$ has negative real part (which ensures that $a(i\omega)$ is non-zero, enforcing that f is bounded). Specifically, the stability assumption for a CARMA process is that the zeroes of $a(z)$, denoted by $\{\lambda_r\}_{r=1}^p$, are assumed to satisfy $\Re(\lambda_r) < 0$, $\forall r$; then the spectral density is integrable, and the autocovariances can be explicitly computed. If the roots λ_r are distinct—and none of them are roots of $b(z)$ —then the autocovariance can be expressed as a function of the lag h via

$$C(h) = \sum_{r=1}^p \frac{b(\lambda_r)b(-\lambda_r)}{a'(\lambda_r)a(-\lambda_r)} e^{\lambda_r|h|}, \quad (19)$$

where $a'(z)$ denotes the derivative of $a(z)$. Formulae also exist for the case of repeated roots, but are more complicated to state; see Brockwell (2000).

It should be noted that there is no natural order to the roots of the polynomial $a(z)$, which consist of real roots and pairs of complex conjugate roots. It is difficult to characterize the process through the roots. For example, when $p = 2$ we may have either two real roots or two complex conjugate roots, and the parameterization of either case is different. For $p > 2$, attempts to characterize $a(z)$ through the roots requires complicated schemes involving sequential ordering. However, parameterization through the coefficients of $a(z)$ is also difficult, because the stability constraint (i.e., the constraints on the roots) implies complicated nonlinear restrictions upon the polynomial coefficients. In the case that $p = 2$ the stability constraint is equivalent to $a_1, a_2 > 0$, but the constraint of positive coefficients is not sufficient to describe the space of stable polynomials if $p \geq 3$.

A similar issue afflicts discrete-time ARMA processes, where the stability constraint is concerned with whether the polynomial roots are outside the unit circle. Although in this discrete-time case the parameterization issue has been solved, it appears to be an open problem for CARMA; for this reason, we focus on CAR(1) and CAR(2) processes in our implementation. We remark that there are many connections between discrete-time ARMA and CARMA processes: Chan and Tong (1987), He and Wang (1989), and Brockwell and Brockwell (1999) establish that some ARMA models can be embedded into the CARMA framework. Conversely, regularly sampled CARMA processes become discrete-time ARMA processes, although this mapping is not surjective; see Chambers and Thornton (2012) for more details.

3.3. Special Cases of CARMA(1,0) and CARMA(2,0)

Here we develop the cases $p = 1, 2$ with $q = 0$ in greater detail. Beginning with $p = 1$ and $q = 0$, the single root λ must be real and negative, and the autocovariance function has formula

$$C(h) = e^{\lambda|h|}/(-2\lambda). \quad (20)$$

Hence, $C^*(h) = \exp\{\lambda|h|\}$, which is identical to the autocorrelation function of a discrete ARMA(1,0) process with autoregressive parameter $\phi = \exp\{h\}$. Recall that the stationarity condition ensures that λ is real and negative, ensuring that $0 < \exp\{\lambda\} < 1$. (This provides an immediate connection between discrete ARMA and CARMA; however, note that $\exp\{\lambda\} \in (0, 1)$, whereas the discrete ARMA(1,0) process has ϕ ranging over $(-1, 1)$, which indicates that regular-sampling of a CARMA(1,0) cannot yield a discrete ARMA(1,0) with negative autoregressive parameter.) The autocovariance function Equation (20) also corresponds to that of the well-known Ornstein-Uhlenbeck process, of which the CARMA class can be seen as a generalization; see Brockwell (2004) for the explicit connection to CAR(1).

In the case of a CARMA(2,0) process, the roots λ_1 and λ_2 are either both real (and negative) or are complex conjugates (with negative real part). Conveniently, λ_1, λ_2 have negative real part if and only if $a_1, a_2 > 0$, which means that the space of positive coefficient quadratics has a simple parameterization, and the roots can be computed from the quadratic formula. (Such a property does not hold for higher degree polynomials.) Equivalently, the real roots can be written as $-u \pm v$ ($0 < v < u$) and complex conjugate roots as $-u \pm vi$, where $u, v \in \mathbb{R}^+$. (For repeated roots, we must allow $v = 0$.) The autocovariance and autocorrelation function in the case of two distinct real roots $-u \pm v$ is given by

$$C(h) = \frac{1}{8uv(u-v)}e^{(-u+v)|h|} - \frac{1}{8uv(u+v)}e^{(-u-v)|h|}C(0) = \frac{1}{4u(u+v)(u-v)}$$

$$C^*(h) = \frac{u+v}{2v}e^{(-u+v)|h|} - \frac{u-v}{2v}e^{(-u-v)|h|}.$$

In the case of complex conjugate roots $-u \pm vi$ the autocovariances and autocorrelations are given by

$$C(h) = \frac{2u \sin(v|h|) + 2v \cos(v|h|)}{(8uv)(u^2 + v^2)}e^{-u|h|}$$

$$C(0) = \frac{1}{(4u)(u^2 + v^2)}C^*(h) = \left(\frac{u}{v}\sin(v|h|) + \cos(v|h|)\right)e^{-u|h|}.$$

For the double root solution of $-u$, an extension of (19)—see McElroy (2013) and Brockwell (2000), for example—yields

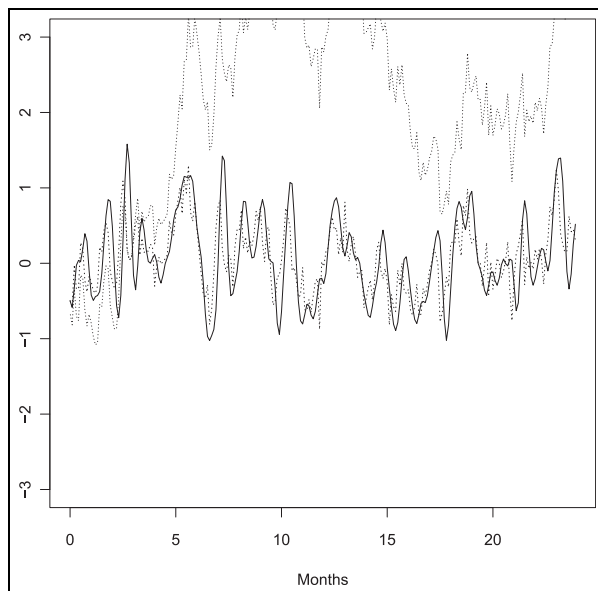


Figure 1. Simulation of a stationary process (solid line), a CAR(1) process (dashed line), and a Brownian Motion (dotted line). The basic time unit is a month, with ten observation times per month.

$$C^*(h) = (u|h| + 1)e^{-u|h|},$$

which is also the limit as $\nu \downarrow 0$ of $C^*(h)$ from the complex conjugate case. This means that the autocovariance function is smooth in ν near zero, implying that statistical estimation for small values of ν will not cause numerical instability in the covariance function.

To illustrate the flexibility of the CARMA class we provide two graphical examples. First, we plot (see Figure 1) a stationary GP (this is the Gaussian kernel process described in Subsection 4.3 below) together with a CAR(1) process (dashed line), which tracks the simulation quite well; for purposes of illustration, we chose the parameters so as to facilitate a visual correspondence, and used the same random seeds. Also plotted is a simulated Brownian Motion (dotted line), whose trajectory reflects linearly growing variability. Clearly, Brownian Motion is not appropriate for the modeling of stationary data; in contrast, data with trend growth can possibly be modeled with a Brownian Motion, or by a CARMA model with linear trend regressor.

A second graphical example is displayed in Figure 2. The solid line is the same Gaussian kernel process added to a sinusoidal signal with a period of twelve months (or 120 observation times); the dash-dotted line is a CAR(2) process with complex conjugate roots $\lambda_1, \lambda_2 = -1 \pm \pi/6$. This setting generates a somewhat mild cyclical effect of period equal to twelve months, and the CAR(2) is able to mimic some of the sinusoidal behavior of the original process. A CAR(1) is not

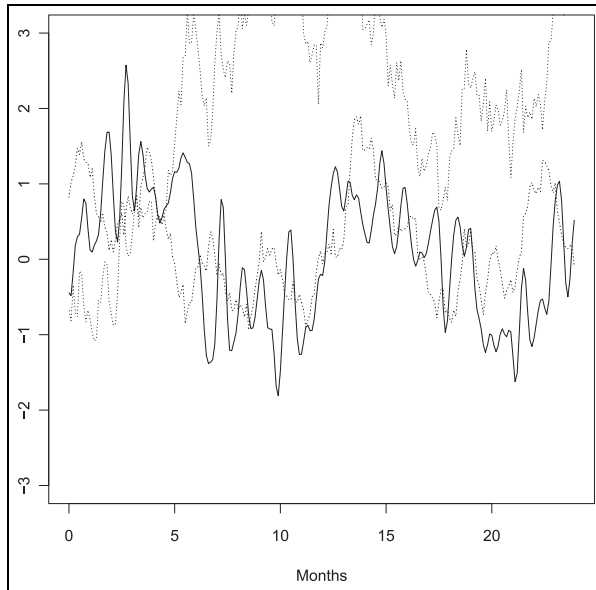


Figure 2. Simulation of a sinusoidal process (solid line), a CAR(2) process (dashed-dotted line), and a Brownian Motion (dotted line). The basic time unit is a month, with ten observation times per month.

capable of exhibiting such dynamics, since its characteristic polynomial only has a single real root; thus we do not display such a sample path. Here again the Brownian Motion (dotted line) is inappropriate due to its non-stationarity.

3.4. Custom Epoch Inference Under the CARMA Model

Higher order CARMA models are capable of capturing much of the autocorrelation structure of an arbitrary stationary process, since the CARMA spectral density is a rational function of squared frequency—this is analogous to the ability of discrete ARMA models to describe a large class of stationary processes. The roots of the autoregressive characteristic polynomial can be associated with particular autocorrelation dynamics (or equivalently, peaks in the spectral density). Due to the challenges of parameterization we restrict ourselves to the $p = 2$, $q = 0$ setting for the remainder of this paper (note that the CAR(1) is a sub-model of the CAR(2)); although this may not be a sufficiently rich class of models to adequately capture all the features of the population process, we are able to obtain improvements over other methods (such as Brownian Motion) while maintaining simplicity of formulation and estimation. The implementation of higher order CARMA will require additional research into a flexible parameterization of the characteristic polynomial $a(z)$ to enable estimation.

To perform inferences for custom epochs using the CAR(2) model for the population process, the modeler must provide the data vector \hat{Y} along with specific regression functions x and sampling variances for \hat{Y} , which allows for the construction of D^* . The method of fitting for the statistical problem is maximum-likelihood. The likelihood presented, denoted L , is obtained from the distribution given in Equation (11). As the likelihood is not convex in the parameters, plausible initial values for the maximum-likelihood are generated by obtaining a variance $\tau_0^2 := \hat{\tau}_{FH}^2$ from the FH model, along with a regression estimate $\beta_0 = \hat{\beta}_{FH}$. The purpose is to first fix as many parameters as possible before formulating an initial estimate for the CAR(2) model terms, $\theta = (a_1, a_2)$, followed by estimating the unconstrained CAR(2) model. The CAR(2) model with τ_0^2 is then fitted to obtain an initial θ_0 . The procedure is summarized as follows:

1. Estimate an initial value of τ^2 by fitting the FH model using Equation (11) such that $C^*(\theta) = I$, and such that variances of the survey estimators σ^2 are replaced by their estimator $\hat{\sigma}^2$. Obtain $\tau_0^2 := \hat{\tau}_{FH}^2$. β_0 can be computed afterward using $\beta_0 = \hat{\beta}_{FH}$.
2. Estimate initial CAR(2) parameters by computing $\theta = \operatorname{argmax}_{\theta} \log(L(\hat{Y}; x, \beta_0, \tau_0^2, \theta))$, using the likelihood from model Equation (11). θ is initialized at $\theta = (a_1 = 1, a_2 = 1)$.
3. With $\beta(\tau^2, \theta)$ given by Equation (16), $(\hat{\tau}^2, \hat{\theta}) = \operatorname{argmax}_{(\tau^2, \theta)} \log(L(\hat{Y}; x, \beta(\tau^2, \theta), \tau^2, \theta))$, followed by $\hat{\beta} = \beta(\hat{\tau}^2, \hat{\theta})$. Computation is initialized at (τ_0^2, θ_0) .

Optimization for obtaining τ_0 was carried out via the Brent optimization method (Brent 1973). Optimizations for θ_0 and for (τ^2, θ) were carried out using the Nelder-Mead algorithm (Nelder and Mead 1965). Joint optimization of (τ^2, θ) was carried out such that τ^2 was on a log-scale, whereas θ required a more complex treatment—as described below. An identical approach is employed to estimate the stock form of the model.

The initializations delineated above are chosen due to the sensitivity of the likelihood to inputs. Because our methodology is designed to accommodate data with diverse epoch lengths, method of moments initializations (which rely upon regularly-sampled flow data) are unlikely to yield good results. Furthermore, this situation is exacerbated by the presence of unequal survey variance errors.

In order to estimate the parameters $\theta = (a_1, a_2)$ of the CAR(2) model, it is convenient to reparameterize the characteristic polynomial $a(z) = z^2 + a_1z + a_2$ with $\tilde{a}(z) = \tilde{a}_2z^2 + \tilde{a}_1z + 1$, which is given by $\tilde{a}(z) = a(z)/a_2$, that is, $\tilde{a}_2 = 1/a_2$ and $\tilde{a}_1 = a_1/a_2$. This reparameterization is motivated by cases where estimates of a_2 arising from successive iterations of the likelihood's nonlinear optimization tended toward ∞ ; since $a_2 = z_1z_2$ and $a_1 = -(z_1 + z_2)$, where z_1, z_2 are the polynomial's roots, it follows that a_1 would also tend to ∞ . This behavior should occur when the CAR(1) model is preferred, as the analog would be one root equal to $-\infty$. Note that $\tilde{a}(z)$ expressed in terms of the original roots is

$$\tilde{a}(z) = \frac{1}{z_1 z_2} z^2 - \frac{(z_1 + z_2)}{z_1 z_2} z + 1;$$

hence if only one root equals $-\infty$, then the model reduces to CAR(1). These parameters are handled through a log transformation (recall that the coefficients are non-negative); specifically, the likelihood is optimized with respect to $a_j^* = \log(\tilde{a}_j)$.

After the estimation procedures, the parameter estimates are inserted into Equations (12) and (14). Of specific interest toward estimation and model evaluation within this paper is the prediction of $\mathbf{Y}|\hat{\mathbf{Y}}$ and $Y(t)|\hat{\mathbf{Y}}$, recalling that \mathbf{Y} is the vector containing the flows Y_{A_i} .

4. Simulation Study

4.1. Simulation Setup

We assess through simulations the effectiveness of the CARMA process for modeling flow-observed survey estimates, and make comparisons with related models, viz. comparisons are made against a model based upon a mid-point observed CARMA with the same sampling error; against the Brownian Motion model of MPS both with strict restriction (as applied in MPS) and without; and against FH using flow CARMA as the true covariance function in Subsection 4.2, and using a Gaussian kernel covariance function as the true covariance function in Subsection 4.3. Although FH was not designed for temporal data, and can be expected to perform poorly, we include it as a baseline of comparison. While the methods (as well as the computer code) allow for overlapping epochs, the simulations focus upon non-overlapping scenarios—this mimics the intended ACS application. The restriction to non-overlapping flows implies that $\mathbf{D}^* = \mathbf{D}$, a diagonal matrix of survey sampling variances. Likewise, the simulation study uses given survey variances; also, survey error is constant across all flow epochs, that is, $\mathbf{D} = \text{Diag}(\{\sigma_i^2 = \sigma^2\})$. This equal sampling variance construction is chosen so that the relationship between sampling variance and process variance can be studied.

Four configurations of parameters for the study are presented here. Four monthly flow epoch estimators and four continuous-time estimators are studied.

- For each simulation three time series are generated, corresponding to ten years of monthly data (the units for the continuous-time process are months) for \mathbf{Y} , $\hat{\mathbf{Y}}$, and $Y(t)$ jointly from Equation (10). The two monthly flow series, \mathbf{Y} and $\hat{\mathbf{Y}}$, are generated at points corresponding to sub-intervals $A_i = \{t \in [i-1, i)\}$, where $i \in \{1, \dots, T\}$. (We set $T = 120$.) A time series corresponding to $Y(t)$ is created at times $t \in \{k/10\}_{k=0}^{10T}$.

This last series is meant to evaluate the continuous-time estimate.

- There are $M = 1000$ simulation iterations for each parameter configuration.

Table 1. CARMA Simulation Study Plan Configurations. Excerpts of Configurations from Table S-2.1.

Configuration	Reg. coeff.	AR polynomial	Roots	Survey error
14	(0, 0)	$\frac{1}{72} + \frac{1}{4}z + z^2$	$z = -\frac{1}{12}, z = -\frac{1}{6}$	0.25^2
15	(0, 0)	$\frac{1}{72} + \frac{1}{4}z + z^2$	$z = -\frac{1}{12}, z = -\frac{1}{6}$	1.00^2
20	(0, 0)	$2 + 3z + z^2$	$z = -1, z = -2$	0.25^2
21	(0, 0)	$2 + 3z + z^2$	$z = -1, z = -2$	1.00^2
32	(0, 0)	$1 + (\frac{\pi}{6})^2 + 2z + z^2$	$z = -1 \pm \frac{\pi}{6}i$	0.25^2
33	(0, 0)	$1 + (\frac{\pi}{6})^2 + 2z + z^2$	$z = -1 \pm \frac{\pi}{6}i$	1.00^2

- Each simulation involves the regressors $x(t) = (1, t)$. There are two configurations for the true parameters: $(\beta_0, \beta_1) = (0, 0)$ and $(\beta_0, \beta_1) = (15, 0.5)$.
- There are three levels of survey error: $\sigma^2 := 0.05^2$, $\sigma^2 := 0.25^2$, and $\sigma^2 := 1^2$, though only the levels $\sigma^2 := 0.05^2$ and $\sigma^2 := 1^2$ are considered for the Gaussian kernel process.
- There are six configurations of the CARMA characteristic polynomial: two representing a CAR(1) process and four representing CAR(2) processes. The latter case breaks into two cases of complex conjugate roots and two cases of a pair of real roots. Each pair of cases mentioned divides into two scenarios, namely a slowly decaying autocovariance function and a rapidly decaying autocovariance function. See Table S-2.1 for the complete configuration, and Table 1 for the shortened version.
- There are three configurations of the Gaussian kernel scale parameter q , viz. $q = 0.5$, $q = 1$, and $q = 2$. See Table S-2.38 for the complete configuration, and Table 8 for the shortened version.
- Four models are used to estimate the small domains (via Equation (12) in the first three cases).
 - M1 uses the flow-sampled CARMA process with survey error model.
 - M2 uses the stock form of the CARMA process with survey error model, such that each survey estimate is assumed to have occurred at the center of the flow interval.
 - M3 uses the flow-sampled Brownian Motion with survey error model.
 - M4 uses the FH model.
- Four models are used to estimate the point domains (via Equation (14)).
 - M1 uses the flow-sampled CARMA process with survey error.
 - M2 uses the stock-sampled CARMA process with survey error model, such that each survey estimate is assumed to have occurred at the center of the flow interval.
 - M3a uses the flow-sampled Brownian Motion with survey error model.
 - M3b uses the flow-sampled Brownian Motion with survey error model, but also constrains each flow estimate to the exact value observed.

Table 2. Squared Error Comparisons for Configuration Number 14 of Table S-2.1.

Regression: $y = 0 + 0x$, Roots of $a \rightarrow z = -\frac{1}{12}$ or $z = -\frac{1}{6}$ Survey error = 0.25^2											
Small domain MSEs						Integrated MSE					
Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$						Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$					
Model	M1	M2	M3	M4	MSE	Model	M1	M2	M3a	M3b	IMSE
M1	—	321	791	1,000	1.5925	M1	—	424	801	1,000	1.6334
M2	679	—	803	1,000	1.5943	M2	576	—	823	1,000	1.6264
M3	209	197	—	1,000	1.7634	M3a	199	177	—	1,000	1.8265
M4	0	0	0	—	6.5491	M3b	0	0	0	—	8.699

Table 3. Squared Error Comparisons for Configuration Number 15 of Table S-2.1.

Regression: $y = 0 + 0x$, Roots of $a \rightarrow z = -\frac{1}{12}$ or $z = -\frac{1}{6}$ Survey error = 1^2											
Small domain MSEs						Integrated MSE					
Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$						Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$					
Model	M1	M2	M3	M4	MSE	Model	M1	M2	M3a	M3b	IMSE
M1	—	561	291	972	14.5866	M1	—	453	291	1,000	14.7218
M2	438	—	295	971	14.7202	M2	547	—	305	1,000	14.7279
M3	709	705	—	972	13.0922	M3a	709	695	—	1,000	13.1833
M4	28	29	28	—	42.0273	M3b	0	0	0	—	138.9304

Table 4. Squared Error Comparisons for Configuration Number 20 of Table S-2.1.

Regression: $y = 0 + 0x$, Roots of $a \rightarrow z = -1$ or $z = -2$, Survey error = 0.25^2											
Small domain MSEs						Integrated MSE					
Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$						Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$					
Model	M1	M2	M3	M4	MSE	Model	M1	M2	M3a	M3b	IMSE
M1	—	469	596	959	6.1233	M1	—	916	680	979	11.6304
M2	531	—	625	966	6.1168	M2	84	—	54	800	12.2191
M3	404	375	—	944	6.1565	M3a	320	946	—	955	11.7484
M4	41	34	56	—	6.9420	M3b	21	200	45	—	13.1170

Small domain estimates are jointly compared by computing the sums of squares $S_{1,\ell}^{(*)} = \sum_{i=1}^T \left(\hat{\mu}_{A_i,\ell}^{\dagger, (*)} - Y_{A_i,\ell} \right)^2 / T$ for each simulation $\ell = 1, \dots, M$ where $(*)$ represents the model form. Each such sum of squares is compared pairwise for each of the models (M1, M2, etc.), for each simulation (i.e., $P_{M_x, M_y} = \sum_{\ell=1}^M I$

Table 5. Squared Error Comparisons for Configuration Number 21 of Table S-2.I.

Regression: $y = 0 + 0x$, Roots of $a \rightarrow z = -1$ or $z = -2$, Survey error = 1^2										
Small domain MSEs						Integrated MSE				
Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$						Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$				
Model	M1	M2	M3	M4	MSE	Model	M1	M2	M3a	IMSE
M1	—	375	821	945	47.4797	M1	—	432	819	1,000
M2	625	—	828	948	47.4492	M2	568	—	856	1,000
M3	179	172	—	660	55.5159	M3a	181	144	—	1,000
M4	55	51	340	—	58.4343	M3b	0	0	0	—
										145.1511

Table 6. Squared Error Comparisons for Configuration Number 32 of Table S-2.I.

Regression: $y = 0 + 0x$, Roots of $a \rightarrow z = -1 \pm \frac{\pi}{6}i$, Survey error = 0.25^2										
Small domain MSEs						Integrated MSE				
Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$						Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$				
Model	M1	M2	M3	M4	MSE	Model	M1	M2	M3a	IMSE
M1	—	502	610	987	5.8196	M1	—	841	670	995
M2	498	—	637	993	5.8139	M2	159	—	200	963
M3	390	363	—	992	5.8616	M3a	330	800	—	995
M4	13	7	8	—	6.9991	M3b	5	37	5	—
										11.0682

Table 7. Squared Error Comparisons for Configuration Number 33 of Table S-2.I.

Regression: $y = 0 + 0x$, Roots of $a \rightarrow z = -1 \pm \frac{\pi}{6}i$, Survey error = 1^2										
Small domain MSEs						Integrated MSE				
Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$						Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$				
Model	M1	M2	M3	M4	MSE	Model	M1	M2	M3a	IMSE
M1	—	367	809	870	45.2884	M1	—	412	804	1,000
M2	633	—	812	971	45.3040	M2	588	—	832	1,000
M3	191	188	—	754	52.4549	M3a	196	168	—	1,000
M4	30	29	246	—	58.5073	M3b	0	0	0	—
										142.2776

$[S_{1,\ell}^{(M_x)} < S_{1,\ell}^{(M_y)}]$, where M_x and M_y are two models), and the mean squared error (MSE), that is, the mean of $S_{1,\ell}^{(*)}$, is recorded across all simulations. The point domain estimates for the true value of the function, $Y(t)$, are jointly compared by

Table 8. Gaussian Simulation Study Plan Configuration. Excerpts of Configurations from Table S-2.38.

Configuration	Reg. coeff.	q	Survey error
41	(0, 0)	1	0.05^2
45	(0, 0)	2	0.05^2
40	(15, 0.5)	0.5	1.00^2
46	(0, 0)	2	1.00^2

computing $S_{2,\ell}^{(*)} = \int_0^T \left(\hat{\mu}_\ell^{\dagger, (*)} - Y_\ell(t) \right)^2 dt / T$ for each simulation $\ell = 1, \dots, M$. Each such sum of squares is compared for each pair of models (M1, M2, etc.), for each simulation (i.e., $P_{Mx, My} = \sum_{\ell=1}^M I[S_{2,\ell}^{(Mx)} < S_{2,\ell}^{(My)}]$, where Mx and My are two models), and the integral mean of the squared error (IMSE), that is, the mean of $S_{2,\ell}^{(*)}$, is recorded across all simulations. The integral calculation is done using Simpson’s integration rule (Zwillinger 2003) with the interval value $\delta = 0.1$. The relative error of integration (the ratio of the approximated error of the numerical integral and the estimate) across all combinations for each estimate and for all iterations was at most 5.79×10^{-4} (or 0.0579%) for the CARMA models, and was at most 3.40×10^{-4} (or 0.0340%) for the Gaussian kernel models. The complete results of the simulation studies are provided in Supplemental S-2.

4.2. Study: CARMA as Data Generating Process

This section examines the performance of the models when the true data process is CARMA. A short list of the parameter configurations is given in Table 1. These excerpts, as given in Tables 2-7, focus upon the pair $\sigma^2 = 0.05^2$ and $\sigma^2 = 1^2$, along with two choices of characteristic polynomials, and the regression coefficients are held fixed at (0, 0) (the simple regression parameters did not appear to be influential).

The main lesson of the simulation study is that the M1 (flow) and M2 (stock) models perform similarly in terms of monthly squared errors, and in some cases the stock model performs better upon flow-generated data. However, there are differences between M1 and M2 when comparing integrals of square errors when σ^2 is smaller and the covariance function decays quickly (as seen by Tables 4 and 6 in comparison to Tables 5 and 7). This appears to be a result of extrapolation on the ends of the observation window. All cases appear to show that the FH model is inadequate in the presence of temporal covariance; this is unsurprising given the non-temporal design of FH, but provides a benchmark for assessing the temporal models. In general, both CARMA estimators perform better than Brownian Motion (M3), except when there is a slowly decaying strictly positive covariance function paired with a high amount of survey error, as seen in Table 3. Model M3b (the estimator of MPS) performs poorly with regards to the displayed

configurations, except for a single configuration (see Supplemental S-2) where M3b is at parity with M1.

4.3. Study: Gaussian Kernel as Data Generating Process

The Gaussian kernel process (see Supplement S-1.3 for the full definition) has covariance function $C(h) = (2\pi)^{-1/2} \exp\{-h^2/2q\}$, where $q > 0$ is a scaling parameter, and correlation function $C^*(h) = \exp\{-h^2/2q\}$.

The Gaussian kernel process has strictly positive autocorrelations that decay more rapidly than the exponential rate of CARMA processes. Hence, Gaussian kernel and CARMA are distinct classes except in the trivial case that both correspond to a white noise (the limiting case of q tending to infinity); in particular, a CARMA model is always mis-specified for a Gaussian kernel process with $q < \infty$.

A short list of the parameter configurations is given in Table 8. A main feature, as demonstrated by Tables 9 and 10, is that M1 gives smaller small domain MSEs for some of the configurations when the survey error is small relative to model error; however, the results are inconclusive when comparing to M2. For cases where survey error and model error are at parity, M2 can exhibit smaller small domain MSEs than M1; moreover, M2 appears to have better integrated MSEs. Examples of this can be seen in Tables 11 and 12. Across all simulations, excepting where M2 has difficulties with extrapolation, the CARMA models perform better than the Brownian Motion (M3, M3a, M3b) and FH (M4) models.

5. Application to ACS Data: Monthly Test Data

The ACS collects many requested and legally mandated questions from the American people, and is published annually at many levels of geographic aggregation. The survey produces one-year (single year) estimates for large geographies and five-year average estimates for smaller geographies. The survey is conducted on a monthly basis with multiple modes of response, along with follow-up for item non-response conducted annually for each calendar year the data is collected. The ACS weighting is calibrated such that totals computed across important demographic categories (such as Age-group, Sex, Race, and Hispanic origin) are enforced to equal independently constructed annual "Population Estimates," with the matching performed using either the data-collection year for one-year ACS estimates or all years for the five-year estimate.

The purpose of this section is to exhibit properties of the model's monthly and yearly estimates on a real data exercise consisting of monthly ACS data. Due to the difficulty of making such data public, the choice was made to create synthetic datasets in place of monthly ACS data. The method of constructing synthetic datasets is provided below, and then applied under different CARMA covariance parameter configurations. These datasets are then estimated under the model to demonstrate monthly and yearly shrinkage in variance.

Table 9. Squared Error Comparisons for Configuration Number 41 of Table S-2.38.

Configuration number 41										
Regression: $y = 0 + 0x$, $q = 1$, Survey error = 0.05^2										
Small domain MSEs						Integrated MSE				
Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$						Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$				
Model	M1	M2	M3	M4	MSE	Model	M1	M2	M3a	IMSE
M1	—	520	541	607	0.2987	M1	—	1,000	929	4.4889
M2	480	—	537	612	0.2987	M2	0	—	0	6.3920
M3	459	463	—	541	0.2989	M3a	71	1,000	—	4.8786
M4	393	388	459	—	0.2995	M3b	80	1,000	986	4.8566

Table 10. Squared Error Comparisons for Configuration Number 45 of Table S-2.38.

Configuration number 45										
Regression: $y = 0 + 0x$, $q = 2$, Survey error = 0.05^2										
Small domain MSEs						Integrated MSE				
Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$						Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$				
Model	M1	M2	M3	M4	MSE	Model	M1	M2	M3a	IMSE
M1	—	739	745	811	0.2953	M1	—	1,000	1,000	0.7402
M2	261	—	709	813	0.2960	M2	0	—	0	1.4863
M3	255	291	—	785	0.2973	M3a	0	1,000	—	0.9956
M4	189	187	215	—	0.3006	M3b	0	1,000	969	0.9824

Table 11. Squared Error Comparisons for Configuration Number 40 of Table S-2.38.

Configuration number 40										
Regression: $y = 15 + 0.5x$, $q = 0.5$, Survey error = 1^2										
Small domain MSEs						Integrated MSE				
Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$						Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$				
Model	M1	M2	M3	M4	MSE	Model	M1	M2	M3a	IMSE
M1	—	371	942	811	53.146	M1	—	459	939	65.8309
M2	629	—	946	833	52.8096	M2	541	—	959	65.7081
M3	58	54	—	122	71.5545	M3a	61	41	—	86.5283
M4	189	167	878	—	56.629	M3b	0	0	22	144.7887

Table 12. Squared Error Comparisons for Configuration Number 46 of Table S-2.38.

Configuration number 46											
Regression: $y = 0 + 0x$, $q = 2$, Survey error = 1^2											
Small domain MSEs						Integrated MSE					
Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$						Pairwise comparisons: $M_x \downarrow, M_y \rightarrow$					
Model	M1	M2	M3	M4	MSE	Model	M1	M2	M3a	M3b	IMSE
M1	—	507	887	964	48.1737	M1	—	451	879	1,000	54.3111
M2	493	—	886	961	48.2160	M2	549	—	918	1,000	53.9684
M3	113	114	—	551	59.4304	M3a	121	82	—	996	66.5650
M4	36	39	449	—	58.9219	M3b	0	0	4	—	140.7279

5.1. Experiment on Missing Data Dealing with a Real Case

Monthly ACS estimates constitute a case of interest for the application of custom epoch estimation, as monthly estimates give a smaller temporal granularity than yearly estimates (which were explored in MPS). To explore the possibility of such a data product, the U.S. Census Bureau created an experimental dataset consisting of the observed household units for forty-nine states (omitting Alaska), along with the various survey weights for computing monthly estimates, as well as the survey responses. The steps taken to create this dataset are described in Albright and Asiala (2016) (additional discussion of sub-annual ACS estimates can be found in King (2010)). U.S. Census Bureau staff studied this dataset and determined that such monthly estimation methods would not meet the various quality standards for production (however, this dataset was made available for further time series research); in this paper we construct a synthetic monthly ACS dataset from published annual data.

One particular item of interest within the dataset is the question of the number of U.S. military veterans. The goal of MPS was to provide a method of sub-annual inference for U.S. military veterans using public ACS one-year, three-year, and/or five-year data for a specific date within the calendar year. This goal was addressed through modeling of the autocovariance structure of the time series data, thereby allowing for a change of domain. The present paper's utilization of CARMA processes shares a similar goal to that of MPS, which was to estimate the number of U.S. military veterans. Here we examine a synthetic monthly dataset instead, whose longer time series sample length makes feasible our continuous-time modeling.

Although our primary interest is in assessing how the flow-sampled CARMA model performs in estimating monthly values, a secondary goal is to infer missing values in the dataset. On October 2013, due to lapse of appropriations the ACS was unable to be collected, and so this value is missing; we are interested in estimating this missing monthly flow as well. (Although the collection and processing of the survey was also disrupted for September and November of 2013, for

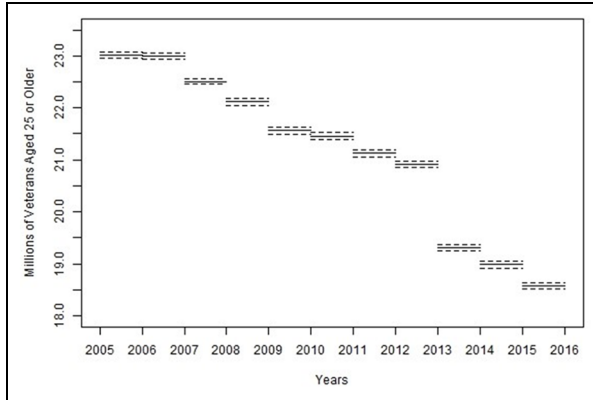


Figure 3. Plot of the one-year ACS estimates with 90% confidence margins of error for veterans aged twenty-five or older from 2005 to 2015.

simplicity it is assumed that the ACS was conducted without disruption from January 2005 to December 2015, excepting only October 2013.)

A drawback of using the monthly test dataset described in Albright and Asiala (2016) is that it is unavailable for public viewing, and hence any derived statistics cannot be readily published. Instead, similar datasets are constructed from annual public data available through the U.S. Census Bureau's Application Programming Interface (API; United States Census Bureau 2023a). In terms of the variable of interest, we focus upon the population of veterans of at least twenty-five years of age, which is obtainable with margin of error from the API. (The U.S. Census Bureau no longer provides Veteran population ACS estimates to the public for specific ACS years.)

It is acknowledged that such a synthetic dataset will not exactly represent all the facets of a real monthly dataset, and therefore this data scenario is a simplification of reality. The constructed datasets explore different parametric values of the CARMA model in order to explore what may happen with real data.

There are two other differences in scope between the synthetic and real datasets: the former does not include Alaska, since the synthetic dataset was created for investigative purposes, and the different sampling rules in Alaska cause additional complication; secondly, the synthetic dataset does not include group quarters. The synthesis will implicitly include both populations, as it is derived from the entire national population.

5.2. Generating a Synthetic Dataset from ACS Public Data

The task is to generate a monthly (synthetic) dataset from ACS public data; however, a few more annotations are necessary before proceeding. The yearly dataset with estimates and 90% confidence intervals is displayed in Figure 3. A box is used to emphasize that this estimate is a flow, corresponding to an epoch; the horizontal

extent of the box corresponds to the epoch (time interval), whereas the vertical extent of the box corresponds to the confidence interval. Note that there is a visible shift downward in the yearly data values as time progresses. The cause of this level-shift is known, owing mainly to a change in the survey measurement instrument—the question regarding veteran status changed in 2013. A level-shift regressor is used in the modeling of the trend. All further analysis is on a timeline shifted to initial time $t = 0$, and re-scaled so that the scale represents a single month.

A synthetic dataset D' is constructed so that there are monthly estimates and related variances for those estimates, for each month from January 2005 to December 2015. The process of data construction starts with obtaining yearly sampling variances from public data; using that yearly public data with the base model Equation (11) to formulate parameter inputs for the monthly model (which are formed by assuming different values of CARMA parameters θ , and then estimating appropriate model variance and regression terms from yearly data), a monthly data series is created by conditioning upon the yearly data series in tandem with assumed monthly survey variances obtained from yearly sampling variances.

There are three notations for time: t measures time continuously (in units of months), whereas m is a discrete index for the month, and j indexes the year. That is, let $t = 0$ correspond to the beginning of January 2005 and $t = 132$ correspond to the end of December 2015. Further, months are individually indexed $m = 1, \dots, 132$ with the relationship $m = \lceil t \rceil$. Years are individually indexed by $j = 1, \dots, 11$, and thus $j = \frac{t}{12} = \frac{m}{12}$.

Construction of D' starts with ACS one-year public data from the API-sourced data from which yearly estimates, \hat{Y}^{year} , and margins of error MOE^{year} are obtained. The first step taken is to reconstruct yearly survey variances. The U.S. Census Bureau uses 90% confidence statements in all public products, so to recover the yearly variance estimates compute $\hat{\sigma}_{\text{year},j}^2 = \left(\text{MOE}_j^{\text{year}} / z_{0.95} \right)^2$, where $z_{0.95}$ is the 95th percentile of a standard normal distribution.

Once yearly variances are obtained, synthetic model parameter values for β , τ^2 , and θ need to be formulated, and the regression functions $x(t)$ need to be specified. The regression function is specified as a simple linear model with a level jump, yielding $x_1(t) = 1[t \in [0, 96]]$, $x_2(t) = 1[t \in [96, 132]]$, and $x_3(t) = t$. Regarding the parameters of the CARMA autocovariance function, the short length (eleven years) of the annual data precludes fitting; rather, several plausible configurations of θ are explored (see Table S-3.1). After setting θ , both τ^2 and β are generated using (11) (under a Bayesian prior of $\pi(\beta, \tau^2) \propto \frac{1}{\sqrt{\tau^2}}$) by computing $\hat{\tau}^2 = E[\tau^2 | \hat{Y}^{\text{year}}, \theta]$ and $\hat{\beta} = E[\beta | \hat{Y}^{\text{year}}, \hat{\tau}^2, \theta]$ in sequence. Such computations arising from (11) account for the absence of October 2013, with corresponding impacts upon the inputs to X and C^* . This hybrid parameter estimation approach is adopted in order to keep computation simple while also avoiding the possibility of obtaining $\tau^2 = 0$.

The next step is to generate monthly variances, followed by monthly observations. The monthly variances are defined by

$$\sigma_m^2 := \begin{cases} 12 \hat{\sigma}_{\text{year, month}}^2 & \text{year} \neq 9 \\ 11 \hat{\sigma}_{\text{year, month}}^2 & \text{year} = 9 \cap \text{month} \neq 10 \\ \emptyset & \text{year} = 9 \cap \text{month} = 10 \end{cases}$$

for years indexed as $\text{year} = 1, \dots, 11$ and months indexed within year by $\text{month} = 1, \dots, 12$, with the relationship that $m = 12(\text{year} - 1) + \text{month}$. As a programming remedy, $\sigma_{106}^2 \gg \tau^2$ (as $m = 106$ corresponds to October of 2013) was utilized in order to mimic the consequence of missing status (as no information is imparted by a normal observation of very large variance), while still permitting model fitting and estimation. Alternatively, one could impute this missing value, perhaps via the expectation-maximization algorithm. These variances are meant to be proxies for a monthly variance estimate, but are assumed to be true variances per model and data assumptions. These monthly variances are constructed on the idea that the yearly estimate ought to be the average of the months, and this variance would be the consequence of such an average.

Having created the synthetic model parameters based on yearly data and a choice of θ , and generated sampling variances for each month, the remaining task is to draw the synthetic monthly dataset from the public set. To do this, a conditional model for 132 monthly data points, $\hat{\mathbf{Y}}^{\text{month}}$, given eleven yearly data points, \mathbf{Y}^{year} , must be drawn. Note that while the parameters are created assuming the model (11), it is assumed here that the true value of the model for the affixed parameters is the ACS one-year public data, and thus $\hat{\mathbf{Y}}^{\text{year}} = \mathbf{Y}^{\text{year}}$. This restriction is imposed so that the monthly simulation matches the yearly estimate from the public data. The distribution draw arises from $\mathbf{Y}^{\text{month}} | \mathbf{Y}^{\text{year}}, \tau^2, \boldsymbol{\beta}, \boldsymbol{\theta}$. The relationship between $\hat{\mathbf{Y}}^{\text{month}}$ and \mathbf{Y}^{year} is defined such that $\mathbf{Y}^{\text{month}} = (Y_1^{\text{month}}, \dots, Y_{132}^{\text{month}})$ and $\mathbf{Y}^{\text{year}} = (Y_1^{\text{year}}, \dots, Y_{11}^{\text{year}})$, where $Y_i^{\text{year}} = \sum_{\ell=1}^{12} Y_{12(i-1)+\ell}^{\text{month}} / 12$. Also, $Y(t)$ is drawn from the Gaussian process: $Y(t) \sim GP(\mu(t) = \mathbf{x}(t)\boldsymbol{\beta}, \tau^2 C(t, s; \theta))$, where $Y_m^{\text{month}} = \int_{m-1}^m Y(t)dt$ and $Y_y^{\text{year}} = \int_{12(y-1)}^{12y} Y(t)dt / 12$. Following this scheme, the simulated data $\hat{\mathbf{Y}}^{\text{month}}$ is obtained by generating $\hat{\mathbf{Y}}_m^{\text{month}} | Y_m^{\text{month}} \sim N(Y_m^{\text{month}}, \sigma_m^2)$.

5.3. Applications of the Model to Veteran Status from Synthetic Data

The purpose of this subsection is to examine the capability of a CAR(2) model for generating custom monthly flow estimates. In particular, we are interested in whether the custom epoch variances (either Equation (13) for monthly epochs or Equation (15) for point domains) are less than the monthly sampling error variances, which would indicate a gain from using CARMA modeling. To examine this the synthetic data was run under six configurations of two root solutions. The first three configurations involve two real roots of a such that the roots of a are the pairs $(-1/12, -1/6)$, $(-1/4, -1/2)$, and $(-2, -4)$. The second three configurations involve two imaginary roots of a such that the roots of a are pairs $(-1/6 \pm i\pi/6)$, $(-1/2 \pm i\pi/6)$, and $(-2 \pm i\pi/6)$. Supplement S-3 contains a fuller discussion of these plots, but the analysis for the cases of $(-1/12, -1/6)$, $(-1/4, -1/2)$, $(-1/6 \pm i\pi/6)$,

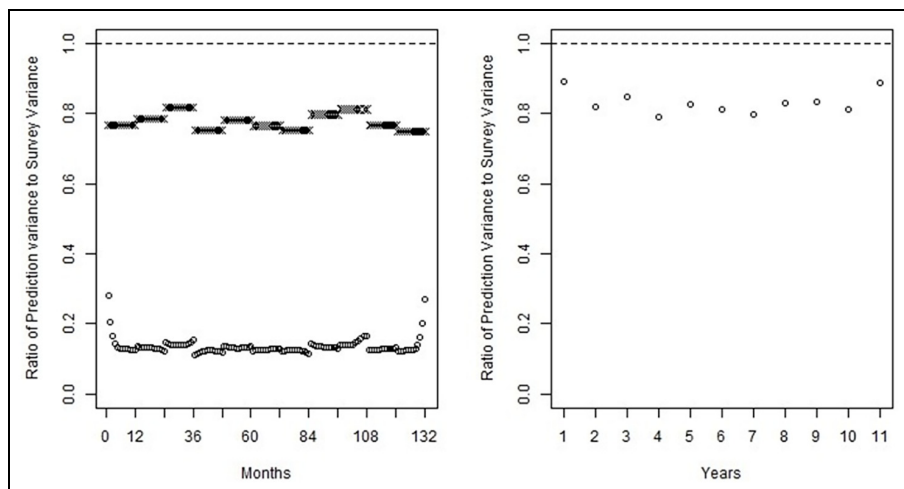


Figure 4. Plot of the ratio of model monthly prediction variances to ACS monthly generated sampling variances in circles, and the ratio of fraction of variance decrease that would be expected from a FH model in hashes (left), and the ratio of model yearly prediction variances to ACS yearly sampling variances (right), in the case where $a(z)$ has roots at $z = -1/6$ and $z = -1/12$.

and $(-1/2 \pm i\pi/6)$ are similar to each other and to the analysis for the cases where $(-2, -4)$ and $(-2 \pm i\pi/6)$. As such, the cases for roots of a are $(-1/12, -1/6)$ and $(-2, -4)$ are given as examples.

Figures 4 and 5 relate in the left plot the decrease in variance from the monthly synthetic sampling variance to the FH model variance in the cross-marked figures and the CARMA model variance in the circle-marked figures. The right plot of those same figures relates the yearly increase/decrease of the CARMA model compared to the input ACS annual sampling variance. The takeaway from Figure 4, corresponding to a case with a slowly decaying covariance function, is that the FH and CARMA models both provide improvements in the monthly synthetic data scenario, with the CARMA model providing a substantial amount of decrease. The annual decrease also exists but it is not as strong as the monthly decrease. For Figure 5, a case with a quickly decaying covariance function, the monthly decreases of FH and CARMA models exist but are not nearly as strong. The annual decreases attributable to the CARMA model does exist for most years, excepting 2013, where there is an increase at the missing month.

Supplement S-3 contains the remaining figures, as well as plots relating the estimator and confidence interval for the estimator with the synthetic sampling distribution data expressed as a 90% confidence interval for each scenario. Subsection S-3.3 provides the inputs from the ACS API, as well as the synthetic data that was generated.

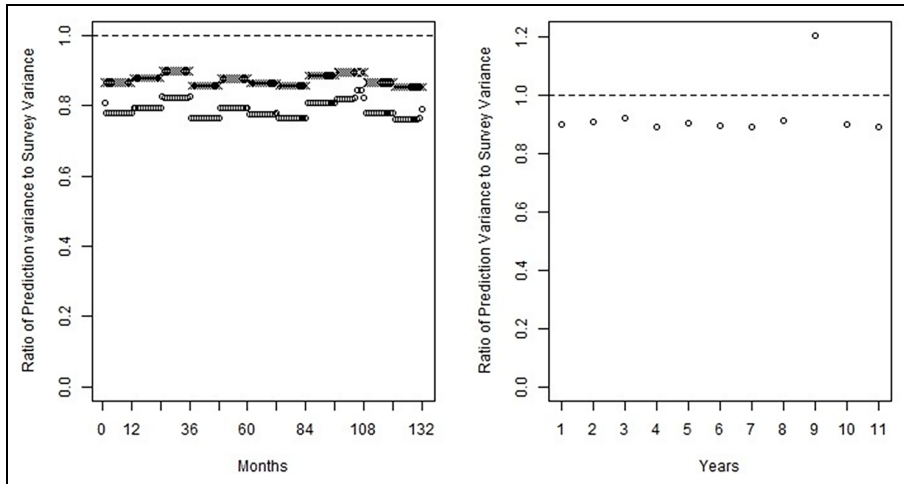


Figure 5. Plot of the ratio of model monthly prediction variances to ACS monthly generated sampling variances in circles, and the ratio of fraction of variance decrease that would be expected from a FH model in hashes (left), and the ratio of model yearly prediction variances to ACS yearly sampling variances (right), in the case where $a(z)$ has roots at $z = -4$ and $z = -2$.

6. Conclusions

This paper studies the problem of customizing flow estimates for survey data, and applies change of support ideas to a time series context, utilizing a continuous-time model for the underlying population process. In particular, we propose using the CARMA class of processes since these have established properties (known formulas for autocovariances, etc.), and outline the general methodology for their application. We focus more specifically on the CAR(1) and CAR(2) processes, noting that some open problems remain on effectively parameterizing higher order CARMA models; the proposed CARMA models can account for some of the temporal dependence in the population process, and the CAR(2) can even accommodate cyclic behavior. Even these fairly simplistic models provide some improvement—based on simulation studies—over more restrictive models, such as that of MPS or FH. Specifically, continuous-time processes appear to have some efficacy for modeling datasets for which a stationarity assumption is tenable; their use appears to outperform the simpler Brownian Motion model, except for cases with more slowly decaying autocovariances, when model and survey error are comparably large.

A remaining research challenge lies in the difficulty of estimating CARMA model parameters. Although the Gaussian likelihood surface is non-convex (such as is commonly the case in discrete time series modeling), the main difficulty is in the parameterization of the space of stable characteristic polynomials. Although it is possible to parameterize the stable polynomials for CAR(1) and CAR(2) processes (the latter case being one of the novel contributions of this paper), for higher

order models such a parameterization remains elusive. Given that the present investigation implies that there is utility in using CARMA models for change of support problems, it would be useful to study alternative parameterizations, or perhaps an alternative class of continuous-time models that are more easily parameterized. We remark that the presence of survey error compounds the estimation task.

As for the utility of stock-estimation versus flow-estimation on data simulated from a survey error flow CARMA process, it appears neither method is universally superior in terms of MSE over small domains; however, in terms of the integrated MSE of the prediction function, the flow-based estimation method tends to perform better—since it appears to suffer less from extrapolation challenges at the epoch boundaries. However, when sampling error and model error are comparably large, the stock model appears to have a better MSE for many (but not all) such cases. As a result, this study cannot establish that flow-estimation is better than stock-estimation on flow-observed data drawn from the CARMA class of processes.

This investigation suggests several avenues for further research. First, how to broaden the class of continuous-time models; for CARMA to be implemented, stable parameterization of the higher order models is necessary. Alternatively, other types of models (e.g., marked point processes) for flow-observed data could be proposed instead. Second, alternative methods for fitting (such as a method-of-moments or Bayesian approaches) should be studied, along with corresponding techniques for model selection and fitting. Third, parameter estimation uncertainty should be incorporated into the variance formulas for the custom epoch estimates, and this could be pursued through either a Bayesian apparatus or a parametric bootstrap. We leave these matters for future work.


Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Disclaimer

This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the authors and not those of the U.S. Census Bureau.

ORCID iD

Patrick M. Joyce  <https://orcid.org/0000-0003-1574-8735>

Supplemental Material

Supplemental material for this article is available online.

References

- Albright, K., and M. E. Asiala. 2016. *Investigating Methods to Support Subannual State-Level Estimates*. American Community Survey Research and Evaluation Program. United States Census Bureau.
- Arima, S., W. R. Bell, G. S. Datta, C. Franco, and B. Liseo. 2017. "Multivariate Fay-Herriot Bayesian Estimation of Small Area Means." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180 (4): 1191–209. DOI: <https://doi.org/10.1111/rssa.12321>.
- Bell, W. R., and S. C. Hillmer. 1987. "Time Series Methods for Survey Estimation." Technical Report RR87-20, U.S. Census Bureau.
- Bell, W. R., and S. C. Hillmer. 1990. "The Time Series Approach to Estimation for Repeated Surveys." *Survey Methodology* 16 (2): 663–85.
- Bergstrom, A. R. 1983. "Gaussian Estimation of Structural Parameters in Higher Order Continuous Time Dynamic Models." *Econometrica* 1: 117–52. DOI: <https://doi.org/10.2307/1912251>.
- Bradley, J. R., C. K. Wikle, and S. H. Holan. 2015. "Spatio-Temporal Change of Support with Application to American Community Survey Multi-Year Period Estimates." *Stat* 4 (1): 255–70. DOI: <https://doi.org/10.1002/sta4.94>.
- Bradley, J. R., C. K. Wikle, and S. H. Holan. 2016. "Bayesian Spatial Change of Support for Count-Valued Survey Data with Application to the American Community Survey." *Journal of the American Statistical Association* 111 (514): 472–87. DOI: <https://doi.org/10.1080/01621459.2015.1117471>.
- Bradley, J. R., C. K. Wikle, and S. H. Holan. 2017. "Regionalization of Multiscale Spatial Processes by Using a Criterion for Spatial Aggregation Error." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79 (3): 815–32. DOI: <https://doi.org/10.1111/rssb.12179>.
- Brent, R. P. 1973. *Algorithms for Minimization Without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall.
- Brockwell, A. E., and P. J. Brockwell. 1999. "A Class of Non-Embeddable ARMA Processes." *Journal of Time Series Analysis* 20 (5): 483–6. DOI: <https://doi.org/10.1111/1467-9892.00151>.
- Brockwell, P. J. 1995. "A Note on the Embedding of Discrete-Time ARMA Processes." *Journal of Time Series Analysis* 16: 451–60. DOI: <https://doi.org/j.1467-9892.1995.tb00246.x>.
- Brockwell, P. J. 2000. "Continuous-Time ARMA Processes." In *Stochastic Processes: Theory and Methods, Handbook of Statistics*, edited by C. R. Rao and D. N. Shanbhag, 457–80. Amsterdam: North-Holland.
- Brockwell, P. J. 2001. "Levy-Driven CARMA Processes." *Annals of the Institute of Statistical Mathematics* 53 (1): 113–24. DOI: <https://doi.org/10.1023/A:1017972605872>.
- Brockwell, P. J. 2004. "Representations of Continuous-Time ARMA Processes." *Journal of Applied Probability* 41: 375–82. DOI: <https://doi.org/10.1239/jap/1082552212>.
- Brockwell, P. J. 2009. "Lévy-Driven Continuous-Time ARMA Processes." In *Handbook of Financial Time Series*, edited by T. Mikosch, J.-P. Kreiss, R. A. Davis, and T. G. Andersen, 457–80. Berlin, Heidelberg: Springer-Verlag. DOI: https://doi.org/10.1007/978-3-540-71297-8_20.
- Brockwell, P. J., and T. Marquardt. 2005. "Lévy-Driven and Fractionally Integrated ARMA Processes with Continuous Time Parameter." *Statistica Sinica* 15: 477–94.

- Chambers, M. J., and J. S. McGarry. 2002. "Modeling Cyclical Behavior with Differential-Difference Equations in an Unobserved Components Framework." *Econometric Theory* 18 (2): 387–419. DOI: <https://doi.org/10.1017/S0266466602182077>.
- Chambers, M. J., and M. A. Thornton. 2012. "Discrete Time Representations of Continuous Time ARMA Process." *Econometric Theory* 38: 219–38. DOI: <https://doi.org/10.1017/S0266466611000181>.
- Chan, K. S., and H. Tong. 1987. "A Note on Embedding a Discrete Parameter ARMA Model in a Continuous Parameter ARMA Model." *Journal of Time Series Analysis* 8 (3): 277–81. DOI: <https://doi.org/10.1111/j.14679892.1987.tb00439.x>.
- Cressie, N. A. C. 1993. *Statistics for Spatial Data, Revised Edition*. Hoboken, NJ: John Wiley & Sons, Inc.
- Doob, J. L. 1944. "The Elementary Gaussian Process." *Annals of Mathematical Statistics* 15 (3): 229–82. DOI: <https://doi.org/10.1214/aoms/1177731234>.
- Fay, R. E., III, and R. A. Herriot. 1979. "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data." *Journal of the American Statistical Association* 74: 269–77. DOI: <https://doi.org/10.2307/2286322>.
- Gelman, A. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis* 1: 515–33. DOI: <https://doi.org/10.1214/06-BA117A>.
- Ghosh, M., N. Nangia, and D. H. Kim. 1996. "Estimation of Median Income of Four Person Families: A Bayesian Time Series Approach." *Journal of the American Statistical Association* 91 (3): 1423–31. DOI: <https://doi.org/10.1080/01621459.1996.10476710>.
- Harvey, A. C., and T. Trimbur. 2003. "Trend Estimation, Signal-Noise Ratios and the Frequency of Observations." Proceedings of the 4th Colloquium on Modern Tools for Business Cycle Analysis, EUROSTAT.
- He, S. W., and J. G. Wang. 1989. "On Embedding a Discrete-Parameter ARMA Model in a Continuous-Parameter ARMA Model." *Journal of Time Series Analysis* 10 (4): 315–23. DOI: <https://doi.org/10.1111/j.14679892.1989.tb00031.x>.
- Horvitz, D. G., and D. J. Thompson. 1952. "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47: 663–85. DOI: <https://doi.org/10.1080/01621459.1952.10483446>.
- James, W., and C. Stein. 1961. "Estimation with Quadratic Loss." In *Berkeley Symposium on Mathematical Statistics and Probability Vol. 4.1*, edited by J. Neyman, 361–79. Berkeley, CA: University of California Press.
- Janicki, R., A. M. Raim, S. H. Holan, and J. J. Maples. 2022. "Bayesian Nonparametric Multivariate Spatial Mixture Mixed Effects Models with Application to American Community Survey Special Tabulations." *The Annals of Applied Statistics* 16 (1): 144–68. DOI: <https://doi.org/10.1214/21-AOAS1494>.
- Jones, J. M. 2020. "National Satisfaction Rises as Election Nears." *Gallup Research*. <https://news.gallup.com/poll/323234/national-satisfaction-rises-election-nears.aspx> (accessed July 23, 2024).
- Jones, R. 1981. "Fitting a Continuous-Time Autoregression to Discrete Data." In *Applied Time Series Analysis*, edited by D. F. Findley, 651–74. New York, NY: Academic Press.
- King, K. E. 2010. *Issues Related to Adding Sub-Annual Estimates to the Data Products Available from the American Community Survey*. American Community Survey Research and Evaluation Program. United States Census Bureau.
- Lohr, S. L. 2010. *Sampling: Design and Analysis*. 2nd ed. Boston, MA: Brooks/Cole.
- McElroy, T., O. Pang, and G. Sheldon. 2019. "Custom Epoch Estimation for Surveys." *Journal of Applied Statistics* 46 (4): 638–63. DOI: <https://doi.org/10.1080/02664763.2018.1508561>.

- McElroy, T. S. 2013. "Forecasting Continuous-Time Processes with Applications to Signal Extraction." *Annals of the Institute of Statistical Mathematics* 65: 439–56. DOI: <https://doi.org/10.1007/s10463-012-0373-x>.
- McElroy, T. S., and D. N. Politis. 2019. *Time Series: A First Course with Bootstrap Starter*. Boca Raton, FL: Chapman and Hall/CRC.
- Nelder, J. A., and R. Mead. 1965. "A Simplex Method for Function Minimization." *Computer Journal* 7: 308–313. DOI: <https://doi.org/10.1093/comjnl/7.4.308>.
- Pfefferman, D., A. Sikov, and R. Tiller. 2014. "Single- and Two-Stage Cross-Sectional and Time Series Benchmarking Procedures for Small Area Estimation." *TEST* 23: 631–66. DOI: <https://doi.org/10.1007/s11749-014-0398-y>.
- Pfefferman, D., and R. Tiller. 2006. "Small-Area Estimation with State-Space Models Subject to Benchmark Constraints." *Journal of the American Statistical Association* 101 (476): 1387–97. DOI: <https://doi.org/10.1198/016214506000000591>.
- Rao, J. N. K., and I. Molina. 2015. *Small Area Estimation*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc.
- Rao, J. N. K., and M. Yu. 1994. "Small-Area Estimation by Combining Time-Series and Cross-Sectional Data." *Canadian Journal of Statistics* 22 (4): 511–28. DOI: <https://doi.org/10.2307/3315407>.
- Steele, J. M. 2012. *Stochastic Calculus and Financial Applications*. New York, NY: Springer New York.
- Sugasawa, S., H. Tamae, and T. Kubokawa. 2017. "Bayesian Estimators for Small Area Models Shrinking Both Means and Variances." *Scandinavian Journal of Statistics* 44 (1): 150–7. DOI: <https://doi.org/10.1111/sjos.12246>.
- Thornton, M. A., and M. J. Chambers. 2017. "Continuous Time ARMA Processes: Discrete Time Representation and Likelihood Evaluation." *Journal of Economic Dynamics and Control* 79: 48–65. DOI: <https://doi.org/10.1016/j.jedc.2017.03.012>.
- Tiller, R. B. 1992. "Time Series Modeling of Sample Survey Data from the U.S. Current Population Survey." *Journal of Official Statistics* 8 (2): 149–66. DOI: <https://doi.org/10.2478/jos-2014-0049>.
- Trimbur, T., and T. McElroy. 2017. "Signal Extraction for Nonstationary Time Series with Diverse Sampling Rules." *Journal of Time Series Econometrics* 9 (1): 20140026. DOI: <https://doi.org/10.1515/jtse-2014-0026>.
- United States Census Bureau. 2014. *American Community Survey Design and Methodology*. Washington, DC.
- United States Census Bureau. 2022. "Current Population Survey (CPS)." <https://www.census.gov/programssurveys/cps.html> (accessed July 23, 2024).
- United States Census Bureau. 2023a. "Available APIs." <https://www.census.gov/data/developers/datasets.html> (accessed July 23, 2024).
- United States Census Bureau. 2023b. "Geography & ACS." <https://www.census.gov/programs-surveys/acs/geography-acs.html> (accessed July 23, 2024).
- You, Y., and B. Chapman. 2006. "Small Area Estimation Using Area Level Models and Estimated Sampling Variances." *Survey Methodology* 16 (1): 97–103.
- Zwillinger, D. 2003. *CRC Standard Mathematical Tables and Formulae*. 31st ed. Boca Raton, FL: Chapman & Hall/CRC.

Received: September 30, 2023

Accepted: August 19, 2024