Article

JOS

Journal of Official Statistics
2024, Vol. 40(2) 317–332

DOI: 10.1177/0282423X241244671
journals.sagepub.com/home/jof
Sage

Nonlinear Fay-Herriot Models for Small Area Estimation Using Random Weight Neural Networks

Paul A. Parker

Abstract

Small area estimation models are critical for dissemination and understanding of important population characteristics within sub-domains that often have limited sample size. The classic Fay-Herriot model is perhaps the most widely used approach to generate such estimates. However, a limiting assumption of this approach is that the latent true population quantity has a linear relationship with the given covariates. Through the use of random weight neural networks, we develop a Bayesian hierarchical extension of this framework that allows for estimation of nonlinear relationships between the true population quantity and the covariates. We illustrate our approach through an empirical simulation study as well as an analysis of median household income for census tracts in the state of California.

Keywords

American Community Survey, Bayesian hierarchical model, household income

1. Introduction

Large surveys conducted by federal agencies provide a rich set of information about the underlying population from which the sample was taken. Classical design-based estimation methods are often used for population level analysis with such surveys, however, when considering various sub-domains within a population (e.g., county, census tract, etc.), sample sizes are often quite small, leading to unreliable direct estimates.

Small area estimation models overcome the issue of small sample sizes by "borrowing strength" from other sub-domains in the population. For example, the Small Area Income and Poverty Estimates (SAIPE) program and the Small Area Health Insurance Estimates

Department of Statistics, University of California Santa Cruz, Santa Cruz, CA, USA

Corresponding author:
Paul A. Parker, Department of Statistics, University of California Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA.
Email: paulparker@ucsc.edu

(SAIHE) program are two important use cases for small area estimation models (Bauder et al. 2018; Bell et al. 2016). These models can be conducted either at the area level, or the unit level, however, unit-level models require access to individual microdata which may be confidential (see Parker et al. (2023a, 2023b) for an overview of unit-level modeling approaches). In contrast to this, area-level models are typically feasible for any analyst, as they only require access to the design-based direct estimates, along with their associated uncertainty. These estimates are often disseminated to the public by various federal statistical agencies. Due to small sample sizes, these direct estimates often have substantial uncertainty, and area-level models reduce the uncertainty by smoothing in some fashion, typically through the use of covariates and/or dependence structure.

The popular Fay-Herriot model (Fay and Herriot 1979) is perhaps the most widely used method for constructing small area estimates. The model assumes that the observed direct estimates are equal to the true but unknown population quantity plus a sample-induced noise term. The true quantity is further modeled as a linear combination of some area-level covariates as well as an independent and identically distributed area-level random effect.

There have been numerous extensions to the Fay-Herriot model that relax the various assumptions. These range from spatial and/or temporal dependence among the random effects (Chandra et al. 2015; Chung and Datta 2020; Marhuenda et al. 2013), to modeling of the sample-induced noise term (Parker et al. 2023; Sugasawa et al. 2017; You and Chapman 2006), to incorporating multivariate structure (Porter et al. 2015), among others. One area that has not seen much attention in the literature is relaxation of the linearity assumption, although Giusti et al. (2012) do consider a semi-parametric approach through the use of penalized splines.

In this work we relax the linearity assumption within the Fay-Herriot model, in order to consider general nonlinear relationships between the population quantities of interest and the covariates. Specifically, we model the unknown population quantity as a nonlinear function of input covariates as well as an area-level random effect. The nonlinear function is estimated using a feed-forward neural network where the hidden layer weights are randomly generated and fixed before model fitting. This allows for flexible estimation of the nonlinear mean function at very little computational cost. Importantly, we show that the nonlinear model has the potential for superior prediction and uncertainty quantification compared to the linear alternative. Beyond neural networks, there are a number of popular nonlinear regression techniques, including random forests (Breiman 2001) and gradient boosting (Friedman 2001), among others. However, it is not immediately clear how to embed these approaches into a hierarchical model, such as the Fay-Herriot model, with appropriate uncertainty quantification. Alternatively, Gaussian Process regression is a powerful nonparametric regression tool (Rasmussen and Williams 2006) that naturally fits into a hierarchical framework. Yet, these approaches do not scale well as the number of observations becomes large. In addition, specification of a valid and efficient covariance function can become difficult as the number of covariates grows. Our use of a random weight neural network both fits nicely into the Fay-Herriot model, and scales easily with the number of data points, as demonstrated in the analysis presented in Section 4.

The remainder of this work is outlined as follows. In Section 2 we provide necessary background information and introduce our proposed methodology. Section 3 provides an empirical simulation study using data from the American Community Survey (ACS). In Section 4 we utilize our proposed approach to generate tract level estimates of median

household income for the state of California. Finally, we provide discussion and concluding remarks in Section 5.

2. Methodology

Before introducing methodology for small area estimation, we briefly establish some notation. We let y_i be an observed direct estimate of some unobserved population quantity θ_i for areas $i = 1,\ldots,d$. For example, the Horvitz-Thompson estimator (Horvitz and Thompson 1952) and Hájek-type estimators (Hájek 1960) are commonly used direct estimators for many survey datasets. Alternatively, when dealing with small sample sizes, one may consider bias corrected direct estimates. Associated with each direct estimate is a design-based variance, σ_i^2, that is assumed to be known and included with the data. Additionally, associated with each area is a length p vector of covariates, $\mathbf{x}_i$.

2.1. Basic Fay-Herriot Model

The standard Fay-Herriot model (Fay and Herriot 1979) is written hierarchically as

$$\begin{aligned} y_i \mid \theta_i, \sigma_i^2 &\overset{ind}{\sim} \text{Normal}\left(\theta_i, \sigma_i^2\right), i = 1,\ldots,d \\ \theta_i &= \mathbf{x}_i'\boldsymbol{\beta} + v_i \\ v_i &\overset{iid}{\sim} N\left(0, \tau^2\right), \end{aligned}$$

where $\text{Normal}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. In other words, each direct estimate $\left(y_i\right)$ is assumed to be normally distributed around the true population quantity of interest $\left(\theta_i\right)$, with known sampling variance $\left(\sigma_i^2\right)$. The true population quantity is then modeled as a linear combination of observed area-level covariates $\left(\mathbf{x}_i\right)$ as well as an area-level random effect $\left(\eta_i\right)$. Finally, this model can be estimated in a Bayesian fashion by placing an appropriate prior distribution over the length p vector of regression coefficients, $\boldsymbol{\beta}$ as well as the random effect variance, τ^2.

2.2. Feed-Forward Neural Networks

Our goal is to replace the linear function assumed by the Fay-Herriot model with a nonlinear function. In other words, rather than assuming $E(\theta_i) = \mathbf{x}_i'\boldsymbol{\beta}$, we will assume $E(\theta_i) = f(\mathbf{x}_i)$ for some nonlinear function $f(\cdot)$. In order to estimate this nonlinear function, we appeal to the literature on neural networks.

A single layer feed-forward neural network can be written as

$$f(\mathbf{x}_i) = \sum_{j=1}^{N} g\left(\mathbf{a}_j'\mathbf{x}_i + b_j\right)\beta_j, \tag{1}$$

where N is considered the number of hidden nodes. In other words, N different affine transformation of the input data are taken, distinguished by their parameters $\mathbf{a}_j$ and b_j. These parameters are typically termed hidden layer weights. Each affine transformation is followed by a nonlinear transformation according to the function $g(\cdot)$. This function is known as an activation function, and common choices are the sigmoid function,

$$g(x) = \frac{1}{1+e^{-x}},$$

or the ReLU function,

$$g(x) = \max(0, x).$$

Finally, the parameters β_j, $j = 1, \ldots, N$ are known as the output layer weights. Thus, the nonlinear function $f(\cdot)$ can be viewed as a weighted average of N different nonlinear transformations of the input data, $\mathbf{x}_i$. Note that the input variables are typically scaled to have mean zero and variance of one.

Ordinarily to fit such a model, a loss function would be chosen relating the input data to the response data, and stochastic gradient descent would be used to estimate the entire set of weights (parameters), $\mathbf{a}_j$, b_j, and β_j for $j = 1, \ldots, N$ (Goodfellow et al. 2016). However, there has been interest recently in alternative approaches that randomly generate and fix the hidden layer weights ($\mathbf{a}_j$ and b_j), only requiring estimation for the output layer weights $\left(\beta_j\right)$ (e.g., see Huang et al. 2006). Because the model is linear conditional on the hidden layer weights, estimation of β_j is straightforward.

The random weight feed-forward neural network may be considered part of the broader class of reservoir computing, where weights are randomly generated rather than estimated. For example, random projection techniques are often used for dimension reduction (Bingham and Mannila 2001) and echo state networks are popular in the context of sequential data (Prokhorov 2005). Still, these random weight methodologies are most often used outside of a statistical context through optimization of a loss function rather than through the use of a likelihood. However, recently the echo state network has been incorporated into a statistical framework under both classical (McDermott and Wikle 2017) and Bayesian (McDermott and Wikle 2019) approaches. Furthermore (Parker and Holan 2023) utilize the random weight feed-forward neural network for unit-level modeling of survey data under informative sampling.

2.3. Nonlinear Fay-Herriot Model

Through the use of a random weight feed-forward neural network, we construct a nonlinear Fay-Herriot model (NFH). We construct the NFH as a Bayesian hierarchical model, allowing for uncertainty quantification as well as regularization of the output layer weights. More specifically, the NFH is written as

$$\begin{aligned}
y_i \mid \theta_i, \sigma_i^2 &\overset{ind}{\sim} \text{Normal}\left(\theta_i, \sigma_i^2\right), i = 1, \ldots, d \\
\theta_i &= \mathbf{g}_i' \boldsymbol{\beta} + \nu_i \\
\mathbf{g}_i' &= \frac{1}{1+e^{-\mathbf{A}\mathbf{x}_i}} \\
\nu_i &\overset{iid}{\sim} \text{N}\left(0, \tau^2\right) \\
\beta_j &\overset{iid}{\sim} \text{N}\left(0, \kappa^2\right), \; j = 1, \ldots, N.
\end{aligned}$$

The NFH model requires specification of prior distributions for τ^2 and κ^2. More detail on these choices is given in the subsequent paragraph. Here, $\mathbf{g}_i$ is the length N vector that results from the hidden layer applied to $\mathbf{x}_i$. Note that the sigmoid activation function (applied elementwise) is used here. Although other activation functions could be explored, we have found that results are usually not sensitive to this choice. The $N \times p$ matrix $\mathbf{A}$ comprises the hidden layer weights. We note that by including a value of one in the first element of $\mathbf{x}_i$ for all i, this matrix implicitly contains both $\mathbf{a}_j$ and b_j for $j = 1, \ldots, N$ as defined in Equation (1). Importantly, we randomly generate the elements of $\mathbf{A}$ independently from a standard normal distribution. This is done a single time before model fitting, and these values are considered fixed. Thus, the hidden layer term, $\mathbf{g}_i$ is completely predetermined before model fitting. Again, other distributions may be used to generate the weights, such as $\text{Uniform}(-1,1)$, however we have found minimal impact on the results from such choices.

The model is completed by placing a prior distribution over the parameters τ^2 and κ^2. For τ^2 we use a vague inverse gamma prior distribution with shape, α_τ, and scale, β_τ, both equal to 0.1. The hierarchical model for $\boldsymbol{\beta}$ is designed to induce regularization, an important component of neural networks and deep learning approaches in general. The normal prior here is analogous to Bayesian ridge regression, with the amount of regularization controlled through κ^2, where smaller values induce more regularization. Thus, with the goal of regularization in mind, we develop a prior that gives more weight to small values. More specifically, we use an inverse gamma prior distribution with shape (α_κ) equal to 20 and scale (β_κ) equal to 8. For illustration purposes, we do not explore alternative values for these hyperparameters. However, it may be possible to use cross-validation or further modeling of these parameters to improve results beyond those presented herein. Furthermore, there is a large literature on Bayesian variable selection and shrinkage priors. For example, as an alternative to the Bayesian ridge regression prior that was used herein, one may use a Bayesian Lasso prior (Park and Casella 2008), or a global-local shrinkage prior (Carvalho et al. 2010; Piironen and Vehtari 2017).

The Bayesian hierarchical model that makes up the NFH is conditionally conjugate. This allows for the posterior distribution of the NFH model to efficiently be sampled using Gibbs sampling. That is, one can iteratively sample from the following full-conditional distributions:

$$1 \quad \boldsymbol{\beta} \mid \cdot \sim \text{Normal}_N \left(\boldsymbol{\mu} = \left(\mathbf{G}'\Sigma_y^{-1}\mathbf{G} + \frac{1}{\kappa^2}\mathbf{I}_N \right)^{-1} \mathbf{G}'\Sigma_y^{-1}\left(\mathbf{y} - \boldsymbol{\nu}\right), \Sigma = \left(\mathbf{G}'\Sigma_y^{-1}\mathbf{G} + \frac{1}{\kappa^2}\mathbf{I}_N \right)^{-1} \right)$$

Here, $\text{Normal}_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes an N-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Also, $\mathbf{y} = (y_1, \ldots, y_d)'$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)'$, $\Sigma_y = \text{Diagonal}(\sigma_1^2, \ldots, \sigma_d^2)$, and the $d \times N$ matrix $\mathbf{G}$ contains $\mathbf{g}_i'$ in the ith row, for rows $i = 1, \ldots, d$.

$$2. \quad \boldsymbol{\nu} \mid \cdot \sim \text{Normal}_d \left(\boldsymbol{\mu} = \left(\Sigma_y^{-1} + \frac{1}{\tau^2}\mathbf{I}_d \right)^{-1} \Sigma_y^{-1}\left(\mathbf{y} - \mathbf{G}\beta\right), \Sigma = \left(\Sigma_y^{-1} + \frac{1}{\tau^2}\mathbf{I}_d \right)^{-1} \right)$$

3. $\tau^2 \mid \cdot \sim \text{IG}\left(\alpha_\tau + \frac{d}{2}, \beta_\tau + \frac{\sum_{i=1}^{d} v_i}{2}\right)$

Here, $\text{IG}(\alpha, \beta)$ denotes an inverse gamma distribution with shape parameter α and scale parameter β.

4. $\kappa^2 \mid \cdot \sim \text{IG}\left(\alpha_\kappa + \frac{N}{2}, \beta_\kappa + \frac{\sum_{j=1}^{N} \beta_j}{2}\right)$

One of the primary advantages of the random weight neural network here is the ability to link the neural network to a likelihood. This is straightforward due to the fact that the hidden layer is completely predetermined before model fitting is done. Thus, linking the neural network to the likelihood is akin to inclusion of extra predictors in the model, where the predictors are the randomly generated outputs of the hidden layer. Furthermore, conditional on the known hidden layer, the model is linear, allowing for a straightforward procedure to sample from the posterior distribution, similar to the linear Fay-Herriot model. Thus, one attains the benefit of a complex nonlinear model with little to no additional computational constraint.

Finally, unlike most machine learning approaches, the random weight neural network used here requires little tuning. Because the hidden layer weights are randomly generated, they do not require tuning parameters related to regularization. We do encourage regularization on the output layer weights, however in our case, this is done naturally through the Bayesian model fitting procedure, again without tuning. Lastly, because we do not use stochastic gradient descent, which is typically required for neural networks, we avoid the need to tune parameters related to optimization. The primary tuning choice to be made with the NFH model is the number of hidden nodes, N. In general, larger values of N can result in more flexible models, but with diminished returns at the cost of increased computation time. We have found that in practice one hundred to two hundred nodes can work quite well across many different datasets. Furthermore, neural networks typically rely on large amounts of data, so we recommend setting N much smaller than the number of data points. As a general strategy, when possible, we recommend fitting a series of models with an increasing number of hidden nodes until diminishing returns for predictive accuracy are observed.

3. Empirical Simulation Study

In order to evaluate the utility of the NFH, we construct a simulation study. Rather than simulating data from a parametric model, which could unnecessarily favor one approach, we build our simulation around an existing dataset. Specifically, we obtain the five-year period estimates of median household income at the census tract level for the state of California and treat these estimates as the true population quantity of interest. Doing so preserves existing structure in the underlying data and potential covariates of interest. From there, we generate noise around the "true" values according to the reported sampling errors associated with the original estimates. This results in simulated data with

similar structure and sampling error variance as the original dataset. The simulated noisy data can be considered as direct estimates that may be used in a Fay-Herriot or NFH model, and the sampling error variances are still known. Note that the California census tracts result in the number of areas $d = 8,927$ for this simulation. We repeat the simulation and estimation procedure one hundred times. We consider five covariates: the tract level poverty rate, and the proportion of the population associated with four different race groups (white, black, hispanic, and asian). All data is collected from the tidycensus package in R (Walker and Herman 2021). Lastly, all models are fit on the log scale, and then estimates are transformed back to the original scale. MCMC was run for two thousand iterations, discarding the first five hundred as burn-in. Convergence was assessed visually through inspection of the trace-plots, where no lack of convergence was detected.

The primary consideration when fitting the NFH is the choice of how many hidden nodes to include. In order to assess the impact of this choice, we fit variations of the NFH considering $N = 30, 50,$ and $100.$ We compare this with the standard Fay-Herriot model (FH) as well as the "direct estimates" (i.e., using the simulated data as the estimator itself). We evaluate the quality of the point estimates and uncertainty estimates for each approach. In terms of point estimates, we value estimators with low mean squared error (MSE),

$$MSE_i = \sum_{k=1}^{K} \frac{\left(\hat{\theta}_{ki} - \theta_i\right)^2}{K}.$$

Here, θ_i represents the true population quantity of interest for area i while $\hat{\theta}_{ki}$ represents an estimate for sample dataset k. In terms of uncertainty, we construct 95% credible intervals, and then evaluate the interval score (IS) (Gneiting and Raftery 2007),

$$IS_i = \frac{1}{K} \sum_{k=1}^{K} \left\{ \left(u_{ki} - \ell_{ki}\right) + \frac{2}{\alpha}\left(\ell_{ki} - \theta_i\right) I\left(\theta_i < \ell_{ki}\right) + \frac{2}{\alpha}\left(\theta_i - u_{ki}\right) I\left(\theta_i > u_{ki}\right) \right\},$$

where $\alpha = 0.05,$ u_{ki} is the upper bound of the interval, and ℓ_{ki} is the lower bound of the interval for sample dataset k and area i. A low interval score is desirable as it strikes a balance between accurate interval coverage as well as interval width, similar to the bias and variance decomposition of MSE. In addition to the MSE and interval score, we report the absolute bias and the interval coverage rate. Table 1 presents a summary of these results, where MSE is presented relative to the direct estimator and all results are averaged across tracts.

The direct estimator exhibits roughly zero bias and perfect coverage by design. However, the bias was more or less negligible in all cases given the scale of the data (median household incomes are in the tens of thousands). Interestingly, there is a slight increase in bias as the number of hidden nodes increases for the NFH model. This was investigated by looking at the bias of the fixed effects component only (i.e., not including the random effects) and the opposite pattern was found. In other words, the bias of the fixed effects only component decreases as the number of hidden nodes increases. This

Table 1. Mean Squared Error (MSE), Absolute Bias, Coverage Rate, and Interval Score for the Empirical Simulation Using Tract Level Data from the American Community Survey Five-Year Period Estimates. MSE is Presented Relative to the Direct Estimator. The 25th and 75th Percentiles Are Also Given in Parentheses

Estimator	*N*	Rel. MSE	Abs. Bias ($\times 10^{-3}$)	Cov. Rate	Int. Score ($\times 10^{-4}$)
Direct	-	1.000	0.961 (0.27, 1.25)	0.950 (0.94, 0.96)	5.608 (3.07, 7.08)
FH	-	0.758 (0.18, 0.84)	3.537 (0.68, 4.21)	0.946 (0.94, 0.97)	5.308 (2.98, 6.57)
NFH	30	0.693 (0.17, 0.74)	3.978 (0.84, 4.95)	0.939 (0.93, 0.97)	5.068 (2.91, 6.12)
NFH	50	0.668 (0.16, 0.70)	4.378 (0.94, 5.43)	0.937 (0.93, 0.97)	4.996 (2.85, 5.91)
NFH	100	**0.659** (0.16, 0.68)	4.507 (1.00, 5.58)	0.937 (0.93, 0.98)	**4.969** (2.82, 5.83)

The model with the lowest MSE and interval score is shown in bold.

leads us to believe that the slight increase in overall bias is due to a slight departure from normality for the random effects as the fixed effects component becomes more complex. If one were strictly interested in decreasing the bias, it might be worth considering more complex models for the random effects. However, in our case, the goal is to reduce the MSE of the point estimates, which is achieved, so we do not pursue more complex random effects models. Unsurprisingly, all model-based approaches outperformed the direct estimator both in terms of MSE as well as interval score. The primary goal of small area estimation is to reduce the uncertainty around the direct estimates, so this result is to be expected. Beyond that, the NFH was able to outperform the standard Fay-Herriot model in all three cases. This indicates that relaxation of the linearity assumption is leading to superior point and uncertainty estimates. As expected, the NFH performs better when a larger number of hidden nodes is selected, although this appears to have diminishing returns, as the gains are minimal between fifty and one hundred nodes. This provides indication that there would be little value to increasing the number of nodes beyond one hundred for this example. Finally, although the NFH has slightly worse interval coverage rate compared to the FH, it results in substantially narrower intervals, leading to an overall preferable interval estimate in terms of the interval score.

Figure 1 examines the MSE for individual tracts, comparing the NFH to FH estimates. Note that the left subplot zooms into a smaller region that contains the majority of the data points for clarity (there are roughly 9,000 total points). The line indicates one-to-one correspondence in MSE, and points falling below the line indicate a reduction in MSE through the use of the NFH when compared to the standard FH. We can see that the majority of points fall below the line, and thus experience reduced MSE through the use of the proposed model. For reference, roughly four out of five tracts exhibited lower MSE from the NFH than the FH. Thus, although the proposed approach did not outperform the FH uniformly, the vast majority of tracts saw an improvement, and there was a substantial improvement on average.

4. California Median Household Income Estimation

Estimation of household income for various geographic areas is an important use case for small area estimation. For example, both poverty and income county level estimates are

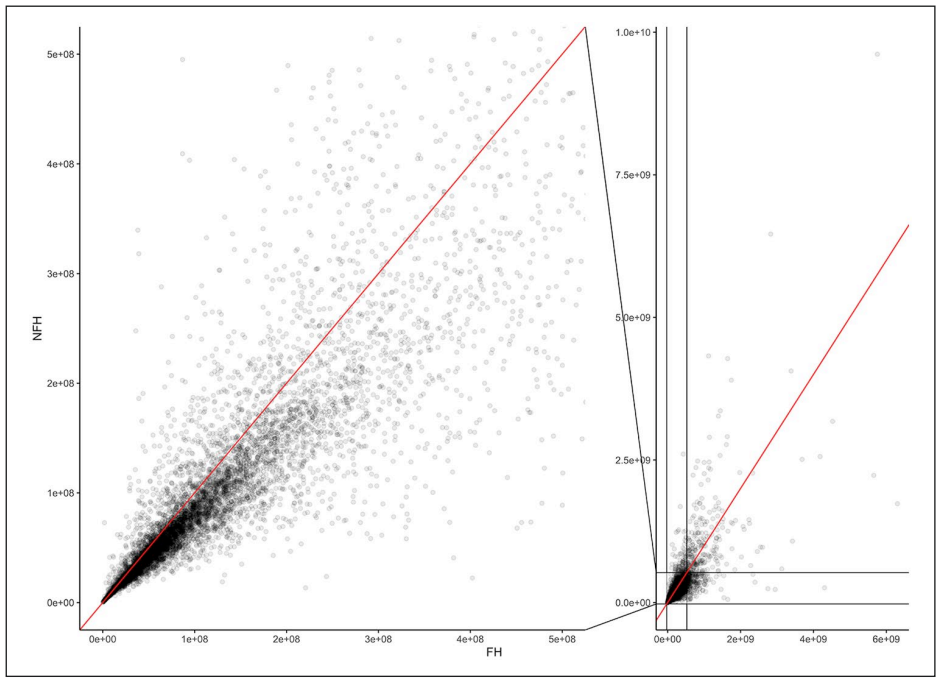

Figure 1. Scatterplot of tract-level MSE of the NFH model versus the FH model for the empirical simulation using tract level data from the American Community Survey five-year period estimates. The left subplot zooms into the region with most of the points.

produced by the Small Area Income and Poverty Estimates program (SAIPE) (Bell et al. 2016). In fact, these estimates are utilized for administration of various federal programs as well as for local allocation of federal funds (https://www.census.gov/programs-surveys/saipe/about.html).

Using the NFH, we construct estimates of median household income by census tract for the state of California. We use the direct estimates of median household income for the 2021 five-year period based on the American Community Survey as our response, as well as the same covariates outlined in Section 3. California contains 9,129 census tracts, although only 8,927 of these have direct estimates available. Tracts without direct estimates could be due to privacy concerns in the case of extremely small sample size, or due to unpopulated areas, such as industrial land. The direct estimates range from USD8,667 to USD249,901 with a mean of USD90,177 and the sampling standard errors range from about USD63 to USD83,000 with a mean of about USD12,000. Figure 2 shows an exploratory plot of the California median income data. Specifically, we show scatterplots of log median income direct estimates against four different covariates. This exploratory analysis gives early indication that some degree of nonlinearity exists in the relationship between log median income and the covariates. Note that the plots presented only show pairwise relationships between the response of interest and a single covariate, however interactions may exist between the covariates that add another degree of nonlinearity.

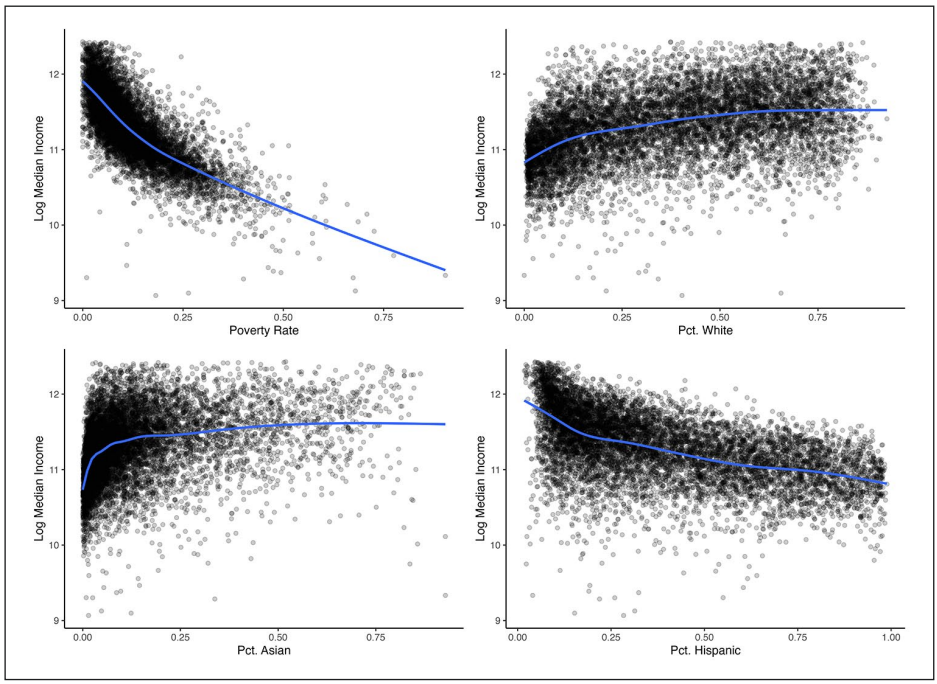

Figure 2. Pairwise scatterplots of log median income versus four covariates for the census tracts in the state of California.

We fit the standard FH model as well as the NFH model using $N = 100$, generating point estimates (posterior means) of household median income for each tract, as well as standard errors (posterior standard deviations) for both models. These are compared to the direct estimates. Figure 3 shows the three point estimates. All three estimators result in the same general spatial trend. That is, median income tends to be higher in densely populated areas such as the San Francisco Bay and Los Angeles, and generally lower in more rural regions of the state. There is little discernible difference across the methods in tracts with low sampling error, which tend to be in more populated areas. However, in certain rural regions with high sampling variability, there is some variation in estimates across the three estimators. Figure 4 presents scatterplots of these estimates, where again we see that all three methods generally result in similar point estimates.

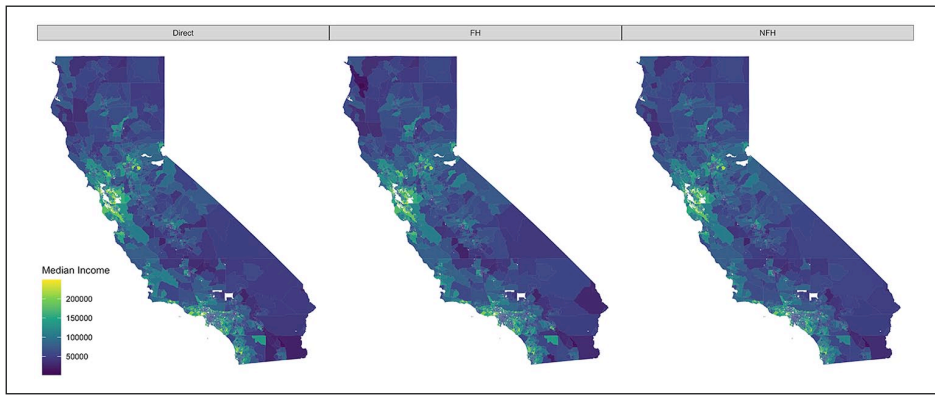

Figure 3. Point estimates of household median income for California census tracts using 2021 American Community Survey five-year period data. Results include direct estimates, Fay-Herriot (FH) estimates, and nonlinear Fay-Herriot (NFH) estimates.

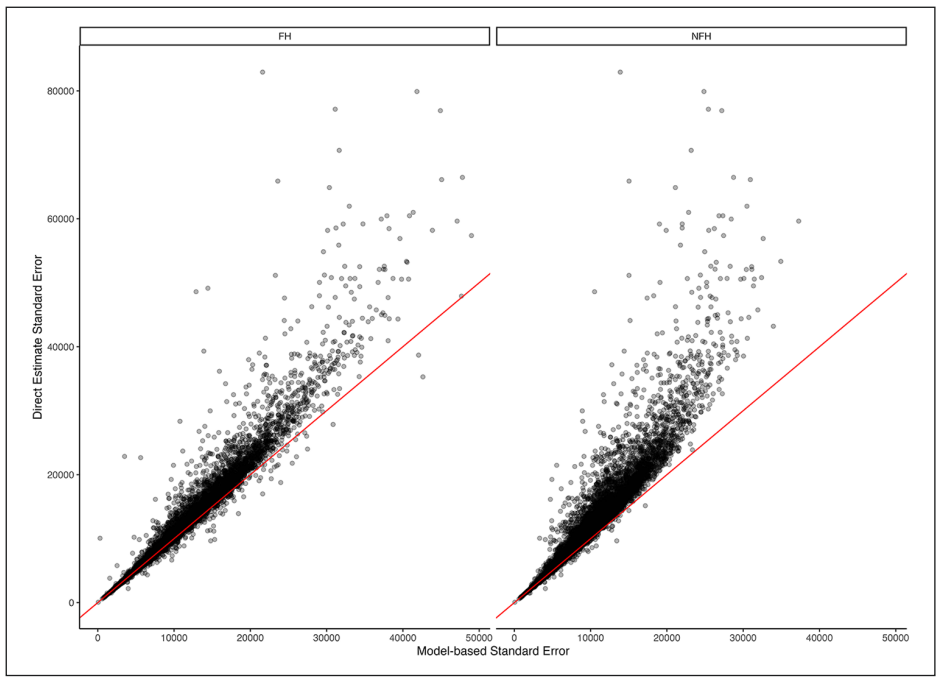

Figure 4. Scatterplots of direct estimates of household median income for California census tracts using 2021 American Community Survey five-year period data versus Fay-Herriot (FH) estimates, and nonlinear Fay-Herriot (NFH) estimates.

Similarly, Figure 5 shows the standard errors for each of the three estimators. As expected, the largest standard errors are associated with the direct estimators, especially in less populated tracts. Both model-based estimators are able to reduce the standard errors substantially in many cases, however the NFH results in the most reduction, yielding lower standard errors than that of the FH. Finally, Figure 6 shows scatterplots of the direct estimate standard errors versus the two model-based standard errors. Note that points falling above the diagonal line corresponse to a reduction in standard error through the model. As expected, most tracts saw a reduction in standard error when using a model-based approach. This is the primary advantage of SAE models. However, the NFH model results in a greater number of tracts that experience a reduction, and generally larger reductions when compared to the standard FH model.

One interesting point concerns the variance of the random effects, τ^2, under both models. For reference, the posterior mean of τ^2 was roughly twice as large under the FH model compared to the NFH model. This gives indication that the nonlinearity modeled by the NFH approach is reducing the reliance on random effects.

Both models were run using a standard 2.3 GHz 8-Core Intel Core i9 processor. The total computation time of the FH was around 17.5 seconds while the total computation time of the NFH was around 76.5 seconds. Thus, although the NFH model took longer in terms of total clock time, both models were extremely quick to run and pose minimal computational burden to the analyst. For reference, the total computation time for the NFH model when using only thirty hidden nodes was roughly thirty-two seconds. Thus, in extremely high-dimensional cases where computation could become a bottle-neck, one may consider using fewer hidden nodes in order to attain computation time on the order of the standard FH model, while still seeing potential gains in precision attributable to the nonlinear model. In total, there is a large potential advantage to the use of the NFH in terms of constructing precise and accurate small area estimates, with little to no trade-off in terms of computational resources that are required.

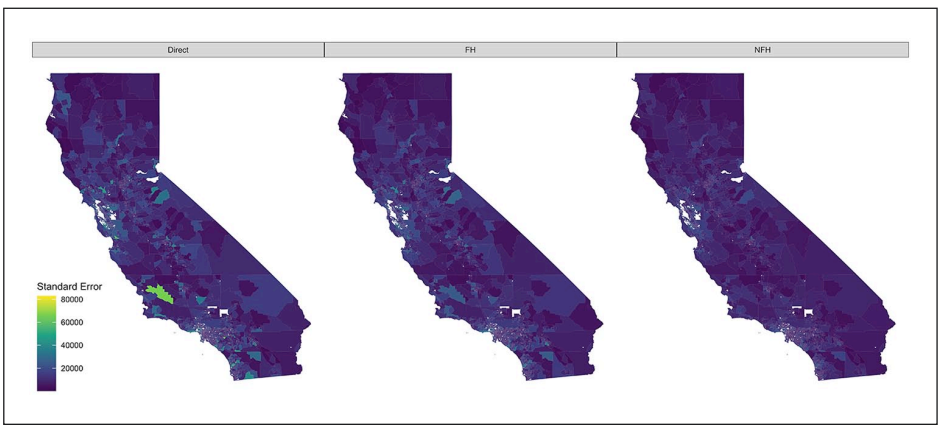

Figure 5. Standard errors of household median income for California census tracts using 2021 American Community Survey five-year period data. Results include direct estimates, Fay-Herriot (FH) estimates, and nonlinear Fay-Herriot (NFH) estimates.

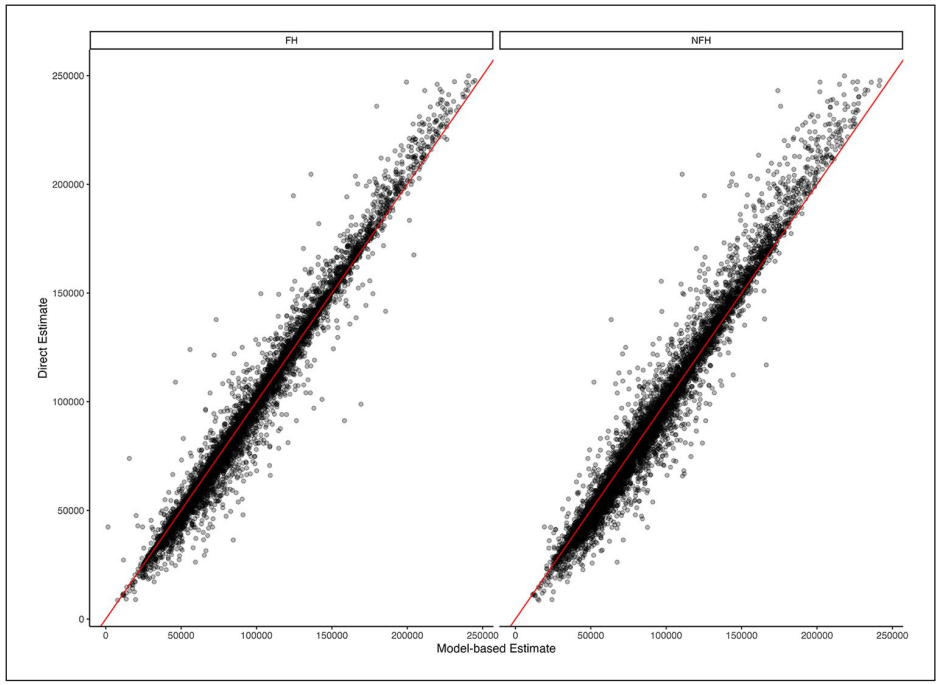

Figure 6. Scatterplots of direct estimate standard errors of household median income for California census tracts using 2021 American Community Survey five-year period data versus Fay-Herriot (FH) standard errors, and nonlinear Fay-Herriot (NFH) standard errors.

5. Discussion

Small area estimation is an important problem that offers tremendous value to federal statistical agencies. Area-level models in particular have become widely used, with a great deal of research taking place on how to incorporate various dependence structures into the model. However, the research on nonlinear modeling of covariates for these area-level models has been quite limited. We contribute to the literature by building a nonlinear Fay-Herriot model. Our approach uses random weight neural networks to flexibly model the mean function. A key point is that due to the nature of the hidden layer weights being random, estimation only takes place for the output layer weights, which is straightforward and computationally efficient.

We assess our proposed approach through an empirical simulation study that builds on data from the American Community Survey. We were able to show that the use of the nonlinear Fay-Herriot model has the potential to generate estimates with substantially lower MSE as well as more desirable interval estimates with reduced uncertainty. Finally, we use the NFH model to generate estimates of median household income at the census tract level for the state of California. The R code used to run both the simulation study and the data analysis can be found at https://github.com/paparker/NFH. Importantly, the

estimates generated by the NFH approach yielded lower standard errors compared to both the direct estimates as well as the standard linear Fay-Herriot model.

Still, there are some unanswered questions that leave opportunity for future work. First, the NFH proposed herein did not explore spatial or other dependence structure among the random effects. In some cases, different spatial patterns have been observed for different subgroups within the population (Janicki et al. 2022). For such situations, an extension of the NFH that considers nonlinear spatial dependence with covariate interactions could be valuable. One challenge when considering spatial dependence will be the scalability as the number of areas in the model becomes large. Another aspect worth considering is measurement error within the covariates. For example, when covariates are themselves estimated from a survey, there is opportunity to improve the model through acknowledgment of the measurement error process. Although these topics are beyond the scope of the current work, they present interesting future directions.

Finally, the results presented here were based on a relatively small number of covariates. In situations where the number of covariates is large, the number of of potential interactions grows quickly. We suspect that in such situations the linearity assumption becomes quite limiting, and thus we may expect even greater accuracy and precision gains through the use of the proposed nonlinear model.

Acknowledgements

We thank the associate editor and anonymous referees for valuable comments that have helped improve this paper.

Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was partially supported by the U.S. National Science Foundation (NSF) under NSF Grant NCSE-2215169. This article is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical issues are those of the author and not those of the NSF.

ORCID iD


Paul A. Parker https://orcid.org/0000-0002-3206-6122

References

Bauder, M., D. Luery, and S. Szelepka. 2018. “Small Area Estimation of Health Insurance Coverage in 2010 – 2016.” *Technical Report*, Small Area Methods Branch, Social, Economic, and Housing Statistics Division, U. S. Census Bureau. Available at: https://www2.census.gov/programs-surveys/sahie/technical-documentation/methodology/2008-2016-methods/sahie-tech-2010-to-2016.pdf

Bell, W. R., W. W. Basel, and J. J. Maples. 2016. “An Overview of the U. S. Census Bureau’s Small Area Income and Poverty Estimates Program.” In *Analysis of Poverty Data by Small Area Estimation*, edited by M. Pratesi, 349–78. New York, NY: Wiley. DOI: https://doi.org/10.1002/9781118814963.ch19.

Bingham, E., and H. Mannila. 2001. “Random Projection in Dimensionality Reduction: Applications to Image and Text Data.” In *Proceedings of the Seventh ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, 245–50. New York, NY: ACM. DOI: https://doi.org/10.1145/502512.502546.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45: 5–32. DOI: https://doi.org/10.1023/a:1010933404324.

Carvalho, C. M., N. G. Polson, and J. G. Scott. 2010. "The Horseshoe Estimator for Sparse Signals." *Biometrika* 97 (2): 465–80. DOI: https://doi.org/10.1093/biomet/asq017.

Chandra, H., N. Salvati, and R. Chambers. 2015. "A Spatially Nonstationary Fay-Herriot Model for Small Area Estimation." *Journal of Survey Statistics and Methodology* 3 (2): 109–35. DOI: https://doi.org/10.1093/jssam/smu026.

Chung, H. C., and G. S. Datta. 2020. "Bayesian Hierarchical Spatial Models for Small Area Estimation." *Research Report Series*, Statistics #2020-07.

Fay, R. E., 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 (366a): 269–77. DOI: https://doi.org/10.2307/2286322.

Friedman, J. H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29: 1189–232. DOI: https://doi.org/10.1214/aos/1013203451.

Giusti, C., S. Marchetti, M. Pratesi, and N. Salvati. 2012. "Semiparametric Fay-Herriot Model Using Penalized Splines." *Journal of the Indian Society of Agricultural Statistics* 66 (1): 1–14.

Gneiting, T., and A. E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477): 359–78. DOI: https://doi.org/10.1198/016214506000001437.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning*. Cambridge, MA: MIT Press. Available at: http://www.deeplearningbook.org.

Hájek, J. 1960. "Limiting Distributions in Simple Random Sampling from a Finite Population." *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5: 361–74.

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 (260): 663–85. DOI: https://doi.org/10.2307/2280784.

Huang, G.-B., Q.-Y. Zhu, and C.-K. Siew. 2006. "Extreme Learning Machine: Theory and Applications." *Neurocomputing* 70 (1–3): 489–501. DOI: https://doi.org/10.1016/j.neucom.2005.12.126.

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.

Marhuenda, Y., I. Molina, and D. Morales. 2013. "Small Area Estimation with Spatiotemporal Fay–Herriot Models." *Computational Statistics & Data Analysis* 58: 308–25. DOI: https://doi.org/10.1016/j.csda.2012.09.002.

McDermott, P. L., and C. K. Wikle. 2017. "An Ensemble Quadratic Echo State Network for Non-Linear Spatio-Temporal Forecasting." *Stat* 6 (1): 315–30. DOI: https://doi.org/10.1002/sta4.160.

McDermott, P. L., and C. K. Wikle. 2019. "Bayesian Recurrent Neural Network Models for Forecasting and Quantifying Uncertainty in Spatial-Temporal Data." *Entropy* 21 (2): 184. DOI: https://doi.org/10.3390/e21020184.

Park, T., and G. Casella. 2008. "The Bayesian Lasso." *Journal of the American Statistical Association* 103 (482): 681–6. DOI: https://doi.org/10.1198/016214508000000337.

Parker, P. A., and S. H. Holan. 2023. "Computationally Efficient Bayesian Unit-Level Random Neural Network Modelling of Survey Data Under Informative Sampling for Small Area

Estimation." *Journal of the Royal Statistical Society Series A: Statistics in Society* 186 (4): 722–37. DOI: https://doi.org/10.1093/jrsssa/qnad033.

Parker, P. A., S. H. Holan, and R. Janicki. 2023. "Conjugate Modeling Approaches for Small Area Estimation with Heteroscedastic Structure." *Journal of Survey Statistics and Methodology*. Published electronically February 25 2023. DOI: https://doi.org/10.1093/jssam/smad002.

Parker, P. A., R. Janicki, and S. H. Holan. 2023a. "Comparison of Unit-Level Small Area Estimation Modeling Approaches for Survey Data Under Informative Sampling." *Journal of Survey Statistics and Methodology* 11 (4): 858–72. DOI: https://doi.org/10.1093/jssam/smad022.

Parker, P. A., R. Janicki, and S. H. Holan. 2023b. "A Comprehensive Overview of Unit-Level Modeling of Survey Data for Small Area Estimation Under Informative Sampling." *Journal of Survey Statistics and Methodology* 11 (4): 829–57. DOI: https://doi.org/10.1093/jssam/smad020.

Piironen, J., and A. Vehtari. 2017. "Sparsity Information and Regularization in the Horseshoe and Other Shrinkage Priors." *Electronic Journal of Statistics* 11 (2): 5018–51. DOI: https://doi.org/10.1214/17-ejs1337si.

Porter, A. T., C. K. Wikle, and S. H. Holan. 2015. "Small Area Estimation via Multivariate Fay–Herriot Models with Latent Spatial Dependence." *Australian & New Zealand Journal of Statistics* 57 (1): 15–29. DOI: https://doi.org/10.1111/anzs.12101.

Prokhorov, D. 2005. "Echo State Networks: Appeal and Challenges." In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks*, Vol. 3, 1463–6. New York, NY: IEEE. DOI: https://doi.org/10.1109/ijcnn.2005.1556091.

Rasmussen, C., and C. Williams. 2006. *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press. DOI: https://doi.org/10.7551/mitpress/3206.001.0001.

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): 150–67. DOI: https://doi.org/10.1111/sjos.12246.

Walker, K., and M. Herman. 2021. "tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames." R package version 1.1. Available at: https://CRAN.R-project.org/package=tidycensus.

You, Y., and B. Chapman. 2006. "Small Area Estimation Using Area Level Models and Estimated Sampling Variances." *Survey Methodology* 32 (1): 97.

Date Received: September 13, 2023
Date Accepted: March 4, 2024