Backcasting Population Data for Children in the 1960s with Supervised Learning*

Esra Kose, UC Merced
Henry Manley, Stanford University
Douglas L. Miller, Cornell University and NBER

July 19, 2024

Abstract

In this paper, we produce population estimates for the US in the 1960s at a newly available level of detail, by age, year, and county, covering ages 1 through 20. To do so, we train an artificial neural network on a rich set of demographic features on data from the 1970s and 1980s. We train our model to predict reported population counts from the Survey of Epidemiology and End Results. The contributions of this paper are twofold. First, we produce and make our new population estimates publicly available. These estimates outperform a linear interpolation benchmark; the median absolute percentage error is about 59% smaller using an out-of-sample testing data set. Our second contribution is to frame population estimation as a prediction problem and to demonstrate that tools from the machine learning literature can give improved estimates for this class of problem. This new approach can be used for other population prediction exercises.

PRELIMINARY: PLEASE DO NOT CITE OR CIRCULATE.

*We thank Colin Cameron, Jamein Cunningham, Ari Decter-Frain, Dave Hacker, Matt Hall, Julia Zhu, and participants in seminars at UC Davis Economics Alumni Conference, Cornell Population Center Pro-seminar, and 2023 Annual Population Association of America Meeting for helpful comments. All errors and omissions are our own. Contact: esrakose@ucmerced.edu, hmanley@stanford.edu, dlm366@cornell.edu.

1 Introduction

In this paper, we use supervised learning to produce new population estimates for the US in the 1960s at a newly available level of detail: by age, year, and county, covering ages 1 through 20.[1]

Estimates at this level of detail are available for 1960 from tabulations of the 1960 Census and beginning in 1969 from Census estimates. For data starting in 1969, researchers often use a version of the Census estimates produced by the Survey of Epidemiology and End Results (SEER). However, there are no available population counts at the county-age-year level for 1961 through 1968. In this paper, we aim to fill in this gap with *predictions* generated by a neural network model, which we then compare to a commonly used linear interpolation benchmark.[2]

There are no readily available data sources that provide direct estimates for the 1961-1968 period, and researchers who study economic occurrences of the 1960s typically construct their own. For example, Bailey (2012) examines general fertility rates by linearly interpolating county-level population estimates from the 1950, 1960, and 1970 Censuses (an approach similar to that demonstrated by Haines & ICPSR (2010)). Thompson (2018) uses county population to scale treatment exposure: Head Start funding is divided by the linearly interpolated count of children aged three to six, using 1960 Census and 1969 SEER data. Ludwig & Miller (2007) address the same challenge slightly differently, using raw counts from the 1960 Census. There is no universal choice for population estimates, which is needed in order to build outcome variables, key explanatory variables, or controls.

To create population predictions, we train an artificial neural network on a rich set of demographic features, using data from the 1970s and 1980s. Our primary predictive features include linear interpolation (which is commonly used to in estimates in current applications), and demographic cohort component estimates. We also include information on broader county demographics and time series-based measures, totaling over 200 predictive features. We train our predictive model on a 75% sample of counties, using data from the 1970s and 1980s. We assess its performance on a "test sample" of 12.5% of counties from the same time period and reserve a further 12.5% as our "hold-out sample." The "hold-out sample" (inspired by Toutanova & Wu (2014) and Kleinberg *et al.* (2018)) is reserved to further check the testing performance once the paper has been conditionally accepted.

The contributions of this paper are twofold. First, we produce and make our new population estimates publicly available. These estimates outperform a linear interpolation benchmark; the median absolute percentage error is about 59% smaller using the out-of-sample testing data set. Proportional improvements are similar across county population sizes. Improvements are relatively larger for years in the middle of the decade. Our second contribution is to frame population estimation as a prediction problem, and to demonstrate that tools from the machine learning literature can improve estimates for this class of problem.

[1] Our population estimates can be downloaded at https://github.com/henrymanley/population_predictions_1960s

[2] The Census/SEER estimates may be imperfect due to under or over-counting. We do not worry about those enumeration issues and treat these estimates as our goal.

2 Data

We compile county-age-year population counts from the 1960 Census (Haines & ICPSR, 2010) and the 1969-1989 SEER. In this paper, we call the population counts from these sources "ground truth," and this forms our outcome variable to be predicted.[3] We use these data to generate predictive features, train our prediction model, and assess its out-of-sample performance.

We additionally use birth records from the Vital Statistics Natality Files (NCHS, 2023a),[4] and death records from the Vital Statistics Multiple Cause of Death Data by age from 1960-1989 (NCHS, 2023b). We discuss how we turn these "raw ingredients" into predictive features in Section 4.

The conceptual geographic unit of our analysis is a county. However, there are many documented instances of counties changing their borders, merging into neighboring counties, or seceding into independent entities. Additionally, different sources and years of raw data encode county identifiers differently. To build a balanced county-year panel, we create a time-invariant identifier that accounts for changes in the underlying status of counties. We refer to our geographic identifier, the county, as "super FIPS," which retrofits the scheme of the 3,083 U.S. counties in 1989 to all preceding years in our sample. Online Appendix C details this process.

Our final dataset encompasses ages 0 to 20, years 1960 to 1989, and super FIPS of 3,083 counties. In total, this sums to 1,942,290 county-age-year observations. Approximately 27% of these observations are from 1961-1968, for which ground truth does not exist.

3 Prediction problem

SEER publishes annual population estimates starting in 1969. For years prior to this, the best data on population counts (at the county and single year of age level) is reported in the decadal Census. This means there are no available population counts at the county-age-year level for 1961 through 1968. In empirical settings where these counts are important, researchers need to estimate them. One popular strategy is to use age-based linear interpolation (henceforth LI). This involves "drawing" a line between the counts in 1960 and 1969 for each single year of age. Alternatively, another reasonable choice might be to interpolate between counts by birth cohort linearly. An additional choice is to use "cohort components" methods. This strategy considers, for example, the 1960's count of a year-olds as a predictor for 1961's count of $a+1$ year-olds, subtracting out deaths. In general, there are several possible methods, each generating slightly different population estimates. It is unclear which to use. As an example, Figure 1 illustrates the performance of LI for Orange County, California.

This paper casts the goal of generating population estimates by county, age, and year from 1961 to 1968 as a *prediction* problem. We propose combining the predictions from different strategies. For example, we see in Figure 1 that LI accurately predicts population in the 1980s but deviates significantly in the 1970s. Is there a way to teach a model to avoid using LI in cases like Orange County in the 1970s?

[3] Our prediction exercise can be thought of as targeting "what would Census/SEER have estimated the population as?" We do not attempt to model discrepancies between Census estimates and actual population counts.

[4] We compile birth records by county and year from county-year aggregate files for 1960 through 1968 and from microdata starting in 1968.

To answer this question, we use 1969-1989 data from SEER to generate two decades of complete population estimates by age and year for each U.S. county (N = 3,083). For each decade, we emulate the missing data problem of the 1960s by defining decadal endpoints (i.e., 1970 and 1979; 1980 and 1989) that we would need to interpolate between in the face of a similar constraint. Because we observe ground truth in these years, we can evaluate the quality of the predictions that each strategy yields. One challenge in this exercise is that the goodness of fit could be calculated using only in-sample data. That is, a model that generates accurate predictions in the 1970s and 1980s might be overfit and produce poor predictions for the 1960s. This is especially problematic because there are no data to validate predictions for the 1960s. To simulate the out-of-sample nature of the 1960s, we construct a supervised learning framework where a randomly selected 75% of counties in our sample are used to *train* our prediction model, 12.5% are used to *test* the accuracy of this model, and the remaining 12.5% are stowed away as *hold-out* counties. Figure 2 illustrates this breakdown. The general idea is to learn a relationship between population counts (Y) and observable predictive features (X) based on data from the training counties and then evaluate their accuracy using data from the testing counties. Data from hold-out counties are not used in either step and instead are a final step in evaluating how well the trained model generalizes to unseen data.[5]

Generally, the goal of this supervised learning exercise is to learn the best mapping of $Y = f(X)$ and apply it to generate population predictions, $\hat{Y} = f(X_{1960s})$. This exposition of the problem is helpful because of its importance of both the X and also the functional form, $f(\cdot)$. Section 4 discusses the features we construct that go into X. Section 5.1 discusses the details of $f(\cdot)$. In this paper, we train an artificial neural network for $f(\cdot)$.

4 Predictive features

There are several ways to produce population estimates by county, age, and year in the 1960s. Instead of proceeding with just one approach, our strategy is to combine them into a rich set of *predictive features* that a machine learning model can learn from. This model then generates out-of-sample predictions of population counts, including for the 1960s. To do so, we build over three thousand features, consistently defined within-decade from 1960 to 1989. However, most of these features are transformations and variations of four basic building block demographic estimates. We discuss these features in the rest of this section.

4.1 Four basic demographic estimates

The most fundamental population estimate we use is linear interpolation (LI).[6] In addition, we construct three "cohort component" features: age-forward (AF), age-backward (AB), and migration-adjusted age-forward (MAAF). These population estimates use a cohort-component model that calculates changes in population by birth cohort. To explain the construction of these estimates, let us consider the 1955 birth cohort who was five years old in 1960. Age-forward works by subtracting the number of five-year-old deaths in 1960 to estimate the count of *six*-year-olds in 1961. This subtraction of deaths by a single year

[5] We will not use the hold-out data until after this paper's conditional acceptance.

[6] In addition to its later-discussed benchmark role, LI is also used to derive many of the predictive features described in this section.

of incrementing age is applied to a birth cohort until the end of the decade, in this case, 14-year-olds in 1969. An analogous method is implemented for the age-backward estimate, with a key difference being the direction in which the recursion occurs— it starts at the end of the decade and moves backward in time, indexing a single birth cohort and adding to it the count of deaths by decrementing age.

These cohort-based methods can incorporate a degree of curvature that linear interpolation cannot. One limitation is that they lack an important component of population flows: migration. To adjust for this, we interpret the difference between the age-forward estimate and the SEER tabulation at the end of a given decade as an estimate of net migration.[7] We then linearly interpolate this migration estimate over the decade and add it to the age-forward estimate to generate its migration-adjusted version.

Several of these baseline demographic estimates require data from years that span other decades and, in some cases, ages and years that extend our sample. For example, AB will need estimates of ages 21-28 to provide mid-decade estimates for 20-year-olds. Online Appendix A shows the mathematical formulas used to arrive at these measures, and the cohorts and years required to calculate them. We apply an inverse hyperbolic sine transformation (IHS) to each measure because population counts are heavily right-skewed and contain zeros. The remainder of this section builds from these baseline estimates to generate the other predictive features our model uses. Broadly, these features span six distinct categories.

4.2 Other features built from the core predictors

Transformations of baseline demographic estimates: While LI is a strong predictor of population counts, it is less effective when within-decade population growth is not linear. This can occur when other baseline demographic estimates – AF, AB, and MAAF – differ from LI. We transform these cohort component estimates into measures of *deviations* from LI. For example, for AF, this is written as $arcsinh(AF) - arcsinh(LI)$, which represents AF's percentage difference from LI. We speculate that instances where a cohort-component-based feature like AF is meaningfully different from LI are ones where LI alone might yield poor predictions.

In addition to the baseline demographic estimates and their deviation-from-LI variants, we include versions that are interacted with the last digit of the year (e.g., 3 for 1973). The motivation for this interaction is that LI should be a better predictor at the beginning and end of the decade compared to the middle of the decade. This is because LI is constructed as an interpolation between decadal endpoints. Similarly, we expect that AF is better at the beginning of the decade than it is at the end, and the opposite is true for AB. Including a version of the baseline demographic estimates that interact with the last digit of the year provides our model with rich information that it can use to learn when one measure is preferred over another.

Variations on linear interpolation: Our core LI predictive feature is based on interpolating over the decade for a given age. It is also based on interpolating the level of population counts, which we then apply the IHS transformation to. We use three additional LI measures based on variations in these choices. The first of these is to interpolate within a birth cohort over the decade. For example, this would interpolate the 1955 birth cohort counts between 1960 and 1969, assigning the interpolated value to the 1955 birth

[7] A similar procedure was implemented by Egan-Robertson *et al.* (n.d.).

cohort for each of the intervening years. We label this *cohort*-based LI, which differs from our baseline *age*-based LI measure. A separate variation is to perform the IHS transformation before interpolation. This has the effect of allocating population changes along a constant-growth-rate path rather than a constant-levels-change path. This "IHS first, then interpolate" approach can be applied to both the age-based and cohort-based approaches. Altogether, this gives us four different candidate LI measures.

Our baseline LI goes from year 0 of a decade through year 9 of that decade (e.g., 1970 to 1979). If we instead interpolate from year 0 to the next decade's year 0 (1970 to 1980), we can compare the interpolated value for the year ending in 9 (1979) to the truth for that year. We speculate that the error from this interpolation could be a useful predictor for the remaining years in that decade, and perhaps especially so for the later years in the decade. We include this error, as well as its interaction with dummy variables for the last digit of the year, as predictor variables.

Measures of variance and curvature: LI will perform relatively poorly when the population counts evolve non-linearly. Using our cohort-based migration-adjusted age-forward measure (as well as its IHS transformation), we compute a measure of curvature over the decade. To do this, for each county-age cell, we first calculate a quadratic in-time regression model over the 1960-1969 time span (and for each subsequent decade): $Y_{cat} = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2$. From this model, we calculate a measure of curvature for each year as $curv_{i,t} = \frac{f''}{(1+f'^2)^{3/2}}$, with $f' = \beta_1 + 2 \cdot \beta_2 \cdot t$ and $f'' = \beta_2$. In rare cases this produces outlier values of measured curvature, and so we replace values with a Windsorization-type procedure. Specifically, we replace values more extreme than $\mu_{curve} \pm 3 \cdot \sigma_{curve}$ with the value $\mu_{curve} \pm 4 \cdot \sigma_{curve}$, with $\mu_{curve}, \sigma_{curve}$ the mean and standard deviation of the curvature measure. We include these measures of curvature, and their absolute value, along with their group-decade average. The idea is to allow for up- or down-weighting of the different core estimates, in response to different amounts of measured curvature.

Separately, we speculate that variation in the population counts within a cell over the decade could be informative. We calculate the within-decade variance of the baseline demographic estimates, births, and deaths, and use these as predictive features.

Demographic features, age & race: A county's demographic characteristics might provide relevant information for population count prediction. For example, consider the college town of Ithaca, New York. Ithaca is home to two large universities – Ithaca College and Cornell University – that together nearly equal the city's population. College towns will have a disproportionately large count of 18-21-year-olds, which especially distorts prediction in small, rural populations like Ithaca. To incorporate this aspect of college-based migration, we use the Census (1960) or SEER (1970, 1980) tabulations at the beginning of each decade to estimate the fraction of a county's college-aged population. Because our main sample includes counts of 0-20 year-olds, we estimate this fraction by dividing the count of 17-20 year-olds by the total count of 0-20-year-olds. Additionally, we produce measures of (i) the share of the total population (among all ages) of 0-20 year-olds, and (ii) the share of the 0-20 population that is non-white. Together, these features characterize potentially meaningful variations in the age and race of populations across counties.

Time series inspired features: Annual lags and leads of the population might be additionally predictive of counts. However, we cannot formulate contemporaneous time series measures for the 1960s, since the population data to create them do not exist. The best that we can do in a mode that respects the missing data in the 1960s is to generate one-decade-ahead averages. To begin, we hold county-age fixed and generate a ten-year lead measure (e.g., for 5-year-olds in 1963, this is the count of 5-year-olds in 1973). One variant of this approach is to hold county-age fixed and average the counts recorded one decade into the future (e.g., for 5-year-olds in 1963, this is the average count of 5-year-olds in the 1970s). We explore two additional dimensions. First, we replicate the aforementioned county-age-based measures but hold county-*cohorts* fixed instead (e.g., the simple 10-year lead measure for 5-year-olds in 1963 would be the count of 15-year-olds in 1973). Second, we swap our various demographic estimates for true population counts to construct the leads. This includes age-based LI, cohort-based LI, and each in their "IHS first" units.

"Utility" features: To conclude, we construct a set of features that, on their own, are likely weak predictors of population counts. However, when interacted with other measures described in this section, these "utility" features might provide additional explanatory power. One previously mentioned example is the encoding of the last digit of the year. Interacting last-digit indicators with, say, LI or MAAF, allows our model to learn to rely on the estimate that provides a more reasonable prediction by year. This is helpful because almost all our features are calculated within-decade and carry advantages and disadvantages based on their imputation method. We also create a measure of the "distance from the nearest decade",[8] generate a single year of age indicator variables, include age directly, and calculate an average county population aged 0-20 in decadal end years.

5 Methods

5.1 Prediction model

The goal of this paper is to predict population counts (Y_{cat}) in the 1960s, indexed by county c, age a, and year t. Because these counts are heavily right skewed and contain zeros[9], we apply the inverse hyperbolic sine (IHS) transformation to Y_{cat} and its population-based predictive features (LI, AF, AB, MAAF). Each feature is normalized to have a mean of zero and a standard deviation of one. Next, we leverage the fact that linear interpolation alone explains 99.8% of the variance in population counts and transform the outcome of interest into a measure of percentage deviation from linear interpolation.[10] This measure is written as:

$$\Delta_{cat} = arcsinh(Y_{cat}) - arcsinh(LI_{cat}) \tag{1}$$

where Δ_{cat}, the difference between "ground truth" and linear interpolation, is the value that our model predicts. Given the series of transformations to Y_{cat}, we apply their inverse to $\hat{\Delta}_{cat}$ to obtain population

[8] Calculated as $min(mod(t,10), 9 - mod(t,10))$.

[9] From 1970 to 1989, there are $3,083$ counties $\times$ 21 ages $\times$ 20 years $= 1,294,860$ county-age-year cells in our data; 19 of these cells contain zero population counts.

[10] This estimate comes from the R^2 in the regression of ground truth on linear interpolation, each IHS transformed and using only counties in the training dataset. The estimated coefficient $\hat{\beta}$ on $arcsinh(LI)$ is .9989 ($t = 1.7 \times 10^4$).

predictions in levels.

To generate predictions, we follow the supervised learning procedure described in Section 3 that maps our predictive feature set to ground truth, given by $\hat{\Delta}_{cat} = f(X_{cat}) + \varepsilon_{cat}$. Using data from the 2,312 counties assigned to the training dataset, we train a densely connected artificial neural network (ANN) to learn the mapping of $f(\cdot)$. Our network consists of an input layer, three hidden layers, and one output layer. The input layer has 171 nodes, which is equal to the number of predictive features that are filtered through a first-stage collinearity check.[11] There are 140 nodes in each hidden layer. Each layer is initialized with Gaussian random weights and biases, and each node in a given hidden layer applies a rectified linear activation function to its inputs. The weights in the network are updated over a series of 200 epochs with batch sizes of 20,000 county-age-years; the magnitude of each update uses the Adam optimization algorithm (Kingma & Ba, 2014). To reduce overfitting, we apply an L1 regularization penalty to the weight vector.[12] The value of each tuning parameter (number of nodes in a hidden layer, epochs, batch size, and L1 penalty) is chosen by a custom five-fold cross-validation algorithm that targets median absolute error. The details of this algorithm are described in Online Appendix B.

5.2 Re-transformation bias

Our prediction model targets Δ_{cat} in Equation (1), but we are ultimately interested in predictions for population counts Y_{cat}. Because the function $sinh(\cdot)$ is nonlinear, we need to account for re-transformation bias. To do so, we generate predicted populations by rearranging Equation 1 and multiplying by an inflation factor like so:

$$\widehat{Y}_{cat} = sinh(\widehat{\Delta}_{cat} + arcsinh(LI_{cat})) \cdot exp(\hat{\sigma}^2_{\varepsilon}/2) \tag{2}$$

with $\hat{\sigma}^2_{\varepsilon}$ a testing sample estimate of the MSE for $\widehat{\Delta}_{cat}$. The inflation factor $exp(\hat{\sigma}^2_{\varepsilon}/2)$ is approximately 0.0009, indicating a trivial adjustment.

5.3 Measuring goodness of fit

Because population counts are right-skewed, we use median absolute percentage error (MAPE) to measure the goodness of fit of the neural network. Since it relies on the *median* error, it is less sensitive to outliers than the alternative root mean squared error. We calculate MAPE as:

$$\text{MAPE}(Y_{cat}, \hat{Y}_{cat}) = 100 \cdot median(\frac{|Y_{cat} - \hat{Y}_{cat}|}{Y_{cat}}) \tag{3}$$

where the magnitude of each residual ($|Y_{cat} - \hat{Y}_{cat}|$) is scaled by ground truth (Y_{cat}) and multiplied by one hundred. Our results aggregate MAPE down to the age, year, and county size levels. MAPE is examined separately for predictions generated by our preferred neural network model, its variants, and LI.

[11] Using OLS, we regress Δ_{cat} onto X_{cat} and drop any collinear features before passing them to the neural network. This helps to shrink the number of parameters the network needs to estimate and update without losing any predictive information.

[12] The L1 regularization penalty $\lambda = \psi/slopes$, with $slopes = (features \cdot width) + 2 * width^2 + width$ and ψ tuned to be 0.03.

6 Results

6.1 Predictive accuracy

County population size is an important feature of how we present our results. Figure 3 shows the distribution of average county population counts summed over ages 0-20, from 1970 to 1989. We use this distribution to create three county-size bins— small, medium, and large— that contain approximately a third of the counties in our sample. When applying the IHS transformation, the average county-level population counts approximate a normal distribution.

We select three case study counties, one from each size bin and all members of our testing set, to show the results of our prediction model graphically. Figure 4 previews the results of our neural network prediction model to document the process we follow to select "case study" counties. We select counties whose predictions are representative of the *average* performance, comparing the neural network against linear interpolation. For each size bin, we choose a county that falls near the intersection of the averages of MAPE generated from each model. This leads us to choose Ellsworth County, Texas; Columbia County, New York; and Orange County, California.

Figure 5 shows the time series for the number of four-year-olds in each of these three counties. "Ground truth" is missing from 1961-1968, which motivates the prediction problem. In cases where the population changes linearly (e.g., Orange County, CA in the 1980s), linear interpolation will be an accurate method to recover the missing data. However, in some periods, population changes are not linear (e.g., Orange County, CA in the 1970s). In these cases, other predictive features that reflect county-decade curvature in population counts could be better. For instance, migration-adjusted age-forward (MAAF, described in Section 4) tracks much closer to the truth in Orange County, CA in the 1970s.

Linear interpolation and migration adjusted age forward are two examples of the 203 predictive features we build and feed to the neural network. Illustrated in Figure 6, the neural network generates predictions that track closer to the true counts of four-year-olds than either of the individual features do. The model succeeds at predicting population counts in decades, both with and without curvature. Each case study county comes from the testing data, meaning the neural network did not use data from these counties during its training.

Until now, our results have been illustrated using county case studies. Figure 7 aggregates the accuracy of our neural network model and linear interpolation by single year of age in panel (a) and by single year in panel (b). The results, in terms of MAPE, are shown separately by county size in the first three columns and aggregated across all counties in the rightmost column. Across all ages, the neural network generates more accurate predictions than linear interpolation. These predictions are also much less sensitive across ages than linear interpolation. Particularly, linear interpolation struggles to predict counts of 5-10-year-olds. The neural network does not have this difficulty, yielding predictions that are $\sim$3 percentage points closer to the truth. Similarly, when considering predictions across time, the neural network dominates. This is especially true in the middle of each decade. In general, the neural network tends to be about 2.39 percent closer to the true population count than linear interpolation. This corresponds to a 59% reduction in MAPE. Both the neural network and linear interpolation do better for larger counties and for years closer to decadal

endpoints.

Table 1 shows additional results on the distribution of the absolute percentage errors for the testing sample. The neural network yields population predictions that are significantly more accurate than linear interpolation across the distribution. The difference in the mean prediction error is more pronounced than the median error.[13]

In summary, our results indicate that the neural network gives meaningful improvements across all county sizes, ages, and calendar years.

6.2 Magnitude of differences: Neural network vs. linear interpolation

Section 6.1 shows that population predictions from the neural network are more accurate than linear interpolation, using data from the 1970s and 1980s. For our main use case of 1960s populations, we do not have "ground truth" to compare against. In this section, we examine the difference between our new estimates and those from linear interpolation.

Figure 8 shows the median absolute percentage difference (MAPD) between linear interpolation and our neural network predictions in the 1960s. It shows that the neural network's predictions differ from linear interpolation by roughly 5%. Panel (a) plots MAPD by single year of age. In the most extreme case, the models differ by $\sim$7% and in the most benign case, a $\sim$2%. The differences are most pronounced for 1-5 year-olds and 15-20 year-olds. Panel (b) plots MAPD by single year; the models differ the most in the middle of the decade. This matches our intuition about prediction errors growing larger in distance from decadal endpoints. In contrast to the testing sample performance in Section 6.1, the percentage differences do not appear to be sensitive to population size. We present more detailed results on the distribution of absolute percentage differences in the testing data in Table 2. We present absolute percentage differences for all counties, pooling testing and training data, in Online Figures Appendix Figure 1 and Appendix Table 2, and show that the results are similar. In addition, as seen in Online Table Appendix Table 3, differences between our new measure and LI are slightly larger when weighting each observation by the underlying population size.[14]

7 Robustness

7.1 Sensitivity to choice of prediction algorithm

In this section, we re-run our prediction exercise described in Section 3 using different choices of prediction algorithm. Each model that we explore – random forest regression (RF), gradient-boosted trees (GB), OLS and LASSO – uses the training sample discussed in Section 3 and predictive features described in Section 4 to learn its parameters and generate population predictions. Then, out-of-sample performance is evaluated using the testing data in comparison to our preferred neural network model.

[13] In addition, Online Appendix Table 1 describes the distribution of absolute percentage errors, *weighted* by population size. We show that the MAPE is 70% lower using the neural network compared to LI.

[14] Because population counts are not observed in the 1960s, we use $\frac{\hat{y}_{ANN}+\hat{y}_{LI}}{2}$ as an estimate of population to weight the distribution of absolute percentage difference.

A perennial goal in supervised learning is tuning your model to include the empirically optimal set of parameters. In our context, determining this set of tuning parameters is computationally taxing. As a result, we follow an ad-hoc tuning procedure for each model.

For the RF model, we tune one parameter – the number of regression *trees* to average over in the forest. Indeed, there are other parameters one can tune – e.g., minimum leaf size, maximum tree depth, number of features to sample in each tree – but we proceed with the understanding that these parameters are meaningfully subsumed by a conservative choice in the number of trees to include or are related to reducing training computational resources, which we are agnostic to. Our preferred RF model has a minimum leaf size of two observations, samples from a share of all 203 predictive features equal to $\frac{\sqrt{203}}{203}$ in each tree, and is tuned to create and average over, 150 trees. We explored 10, 25, 50, 75, 100, 150, 200, and 300 as candidates for the number of trees.

For the GB model, we tune two parameters – the number of trees and the maximum tree depth. We hold the learning rate fixed at 0.075 and the maximum number of features equal to $\frac{\sqrt{203}}{203}$. The maximum tree depth parameter implies the complexity of the interactions that can occur in a single tree: the deeper the tree, the more combinations of features that can occur, and the more conditions that get applied to those features. We consider this parameter somewhat balanced by the tree count parameter insofar as more complex interactions could lead to overfitting. However, averaging over many individually overfit trees leads to a less biased prediction. Of course, this is not exactly precise, and so we choose to tune both. Our preferred model has a maximum tree depth of eight and creates a total of 400 trees.

OLS does not have parameters to tune. LASSO has one – the penalty weight applied to the sum of the magnitudes of the regression coefficients – which we identify as 1.9×10^{-5} via five-fold cross-validation. This selection results in the choice of 130 non-zero coefficients (64% of all features).

We show the MAPE results of this horse race by single year of age and year in Online Appendix Figure 2. The ANN dominates across almost all ages and all years. Using OLS or LASSO leads to a marginal improvement over using LI alone. However, there are large reductions in MAPE when employing RF or GB. Absolute percentage differences in predictions generated by each model and LI for the 1960s are shown in Online Appendix Figure 3.

7.2 Sensitivity to choice of predictive features

In this subsection, we consider how the quality of the prediction model is impacted by the choice of predictive features (covariates). Our main model includes 203 features. We consider ten alternative sets of predictive features, each one based on eliminating a subset of the main model's features.

Because the ANN model relies on a set of tuning parameters, and the optimal set of tuning parameters can depend on the feature set, for each of our alternative feature sets, we engage in a modest re-tuning of the ANN model before generating predictions. For computational reasons, we do not fully explore the space of tuning parameters for each model. Instead, we do the following. First, we set the "minibatch size" to 1-million, based on our impression that it was the least important parameter and that setting it to a large value could improve computational time. Second, starting from the main model's tuning parameters, we profile over a range of XX values of the L1 penalty parameter, using 5-fold cross-validation (CV) as our measure

of goodness of fit. Starting from the best-obtained value, we then profile over XX choices for the number of hidden nodes per layer. Following that, we undertake one more profile over the L1 penalty parameter. We use the best CV goodness of fit to choose the tuning parameters, re-train the ANN based on these, and then consider the out-of-sample goodness of fit using the testing sample.

Results from this exercise are in Table 3. We see that [results in progress; to be updated later].

7.3 Sensitivity to choices of testing and training samples

Our main training sample consists of a 75% sample of counties observed in the years 1971-1978 and 1981-1988. Our main testing sample consists of a further 12.5% sample of counties observed in the same years. We chose this as our primary testing sample because it offers a clean out-of-sample interpretation for measuring goodness of fit. However, it relies on within-time-period data, while our use case for the model is predictions for the 1960s. Here, we explore sensitivity to cross-decade measures of goodness of fit. Table XX illustrates our alternative models.

These models are based on three different training samples and a variety of samples used to measure goodness of fit. In each case, the goodness of fit samples are "out-of-sample" in that their data points were not used to train the prediction model. We have chosen these specifications with the goal of learning more about the importance of the cross-county versus cross-decade nature of out-of-sample. We choose to train only on our training data, and to exclude any data from the hold-out sample in this subsection, until after we are given a "conditional accept' decision from the editor.

For each of the training samples, we re-tune the ANN parameters for the L1 penalty and the number of nodes per hidden layer. We do so in a fashion analogous to that described in Section 7.2.[15]

[Results in progress and discussion still to come.]

8 Conclusion

In this paper, we explore the possibility of recasting intercensal population estimation in the 1960s as a prediction exercise. Our neural network (ANN) model outperforms a LI benchmark by 59%. Improvements in predictive accuracy compared to LI are primarily driven by the flexibility of our model and the construction of a rich set of predictive features. We publish population counts – for ages 1-20 from 1961-1968 for almost all U.S. counties– predicted by our model.

Separate from providing new and improved predictions, this paper demonstrates the scope for supervised learning to improve population estimation strategies.

[15] The main exception is that for each candidate tuning parameter value, we train the ANN model eight times using different random seeds, and use the average of the CV goodness of fit to choose our tuning parameters.

References

AUTOR, DAVID H., & DORN, DAVID. 2013. The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market. *American Economic Review*, **103**(5), 1553–1597.

BAILEY, MARTHA J. 2012. Reexamining the Impact of Family Planning Programs on US Fertility: Evidence from the War on Poverty and the Early Years of Title X. *American Economic Journal: Applied Economics*, **4**(2), 62–97.

EGAN-ROBERTSON, DAVID, CURTIS, KATHERINE J., WINKLER, RICHELLE L., JOHNSON, KENNETH M., & BOURBEAU, CAITLIN. *Age-Specific Net Migration Estimates for US Counties, 1950-2020*. Applied Population Laboratory, University of Wisconsin - Madison, 2023 (Beta Release). Web. https://netmigration.wisc.edu/.

HAINES, MICHAEL R., & ICPSR. 2010. *Historical, Demographic, Economic, and Social Data: The United States, 1790-2002: Version 3*.

KINGMA, DIEDERIK P, & BA, JIMMY. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

KLEINBERG, JON, LAKKARAJU, HIMABINDU, LESKOVEC, JURE, LUDWIG, JENS, & MULLAINATHAN, SENDHIL. 2018. Human decisions and machine predictions. *The quarterly journal of economics*, **133**(1), 237–293.

LUDWIG, JENS, & MILLER, DOUGLAS L. 2007. Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design*. *The Quarterly Journal of Economics*, **122**(1), 159–208.

NCHS. 2023a. *Birth Data - Vital Statistics Natality Data*. https://www.nber.org/research/data/vital-statistics-natality-birth-data.

NCHS. 2023b. *Mortality Data - Vital Statistics NCHS Multiple Cause of Death Data*. https://www.nber.org/research/data/mortality-data-vital-statistics-nchs-multiple-cause-death-data.

THOMPSON, OWEN. 2018. Head Start's Long-Run Impact: Evidence from the Program's Introduction. *Journal of Human Resources*, **53**(4), 1100–1139.

TOUTANOVA, KRISTINA, & WU, HUA. 2014. Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers). *In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Tables & Figures

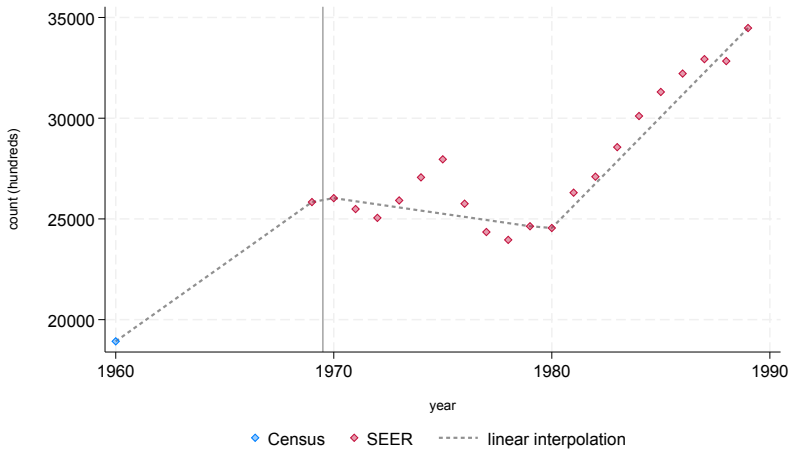


Figure 1: A motivating example of the missing data problem: Orange County, CA

Notes: This figure illustrates the missing data problem that arises in 1961-1968 as a result of a gap in reporting between the 1960 census and the inaugural SEER estimates in 1969. Plotted is the estimate for the count of four year-olds in Orange County, California. A dashed grey line shows the prediction given by age-based linear interpolation. A vertical grey line divides the 1960s from the other two decades.

1961-1968	1971-1978	1981-1988
	Hold-out data (385 counties)	Hold-out data (385 counties)
	Testing data (385 counties)	Testing data (385 counties)
Target data (3,083 counties)	Training data (2,313 counties)	Training data (2,313 counties)

Figure 2: Setting up the supervised learning exercise

Notes: This chart shows the breakdown of our county-year panel into training, testing, and hold-out datasets. There are a total of 3,083 counties in our sample. Our prediction framework only incorporates data between and not including decadal endpoints. Population counts in 1960s represent the target data.

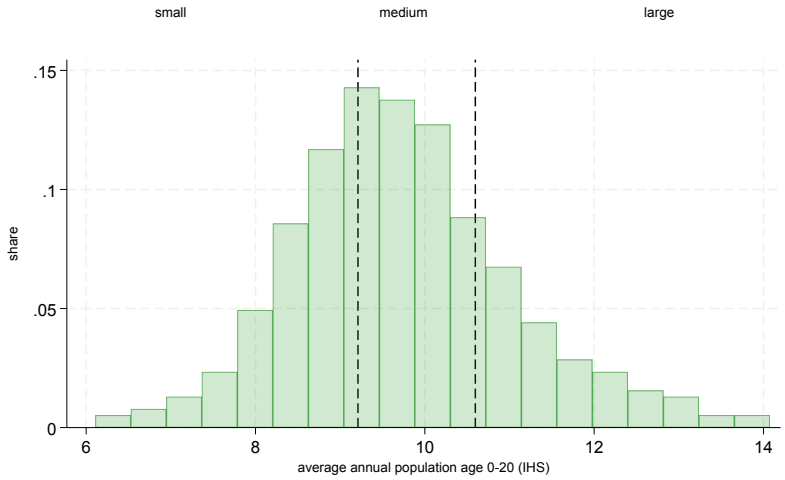


Figure 3: Distribution of county population counts

Notes: This figure shows the distribution of mean county-level population counts for those aged 0-21 from 1970-1989. This mean is calculated by summing counts of 0-20 year-olds by year, and then averaging across years within counties. Only counties in our testing sample are used (N = 385). For each county, this mean is reported in IHS units. Two dashed vertical lines display the county size bins described in Section 6.1, which we use to present our results in terms of baseline county population size. As a result, there are three population size "regions" – small, medium, and large.

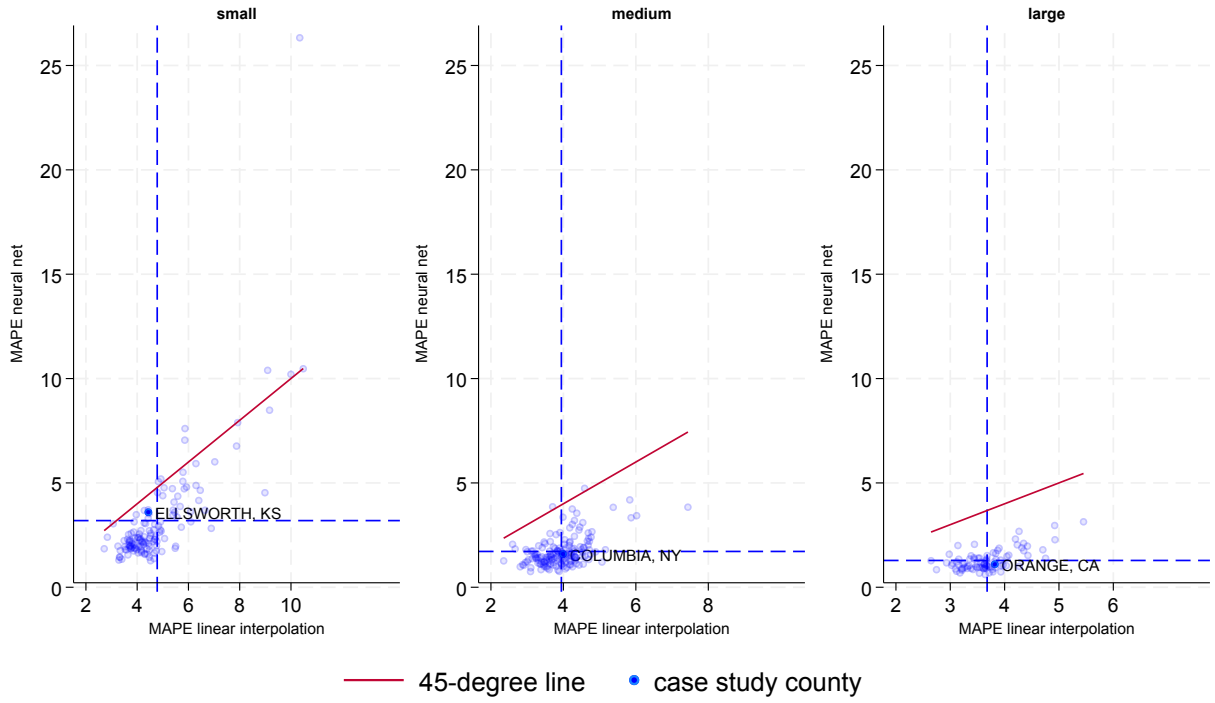


Figure 4: Selecting case study counties

Notes: This figure illustrates the procedure by which we select case study counties. Each panel shows results of our neural network prediction model separately by county-level population size. Each light, unshaded blue dot is a single testing county (N = 385). The median absolute percentage error of the neural network is plotted on the y-axis and the median absolute percentage error of linear interpolation is plotted on the x-axis. These two statistics are calculated across all ages and out-of-sample years. Each plot includes a red 45-degree line, which represents the point where the median absolute error of each model is theoretically equal. The mean value of each axis is shown via a dashed blue line. The general motivation for this exercise is to choose counties to case-study that are representative of the "average" performance. Thus, we selectively label (in dark, shaded blue) a single county near in the intersection of the dashed blue lines.

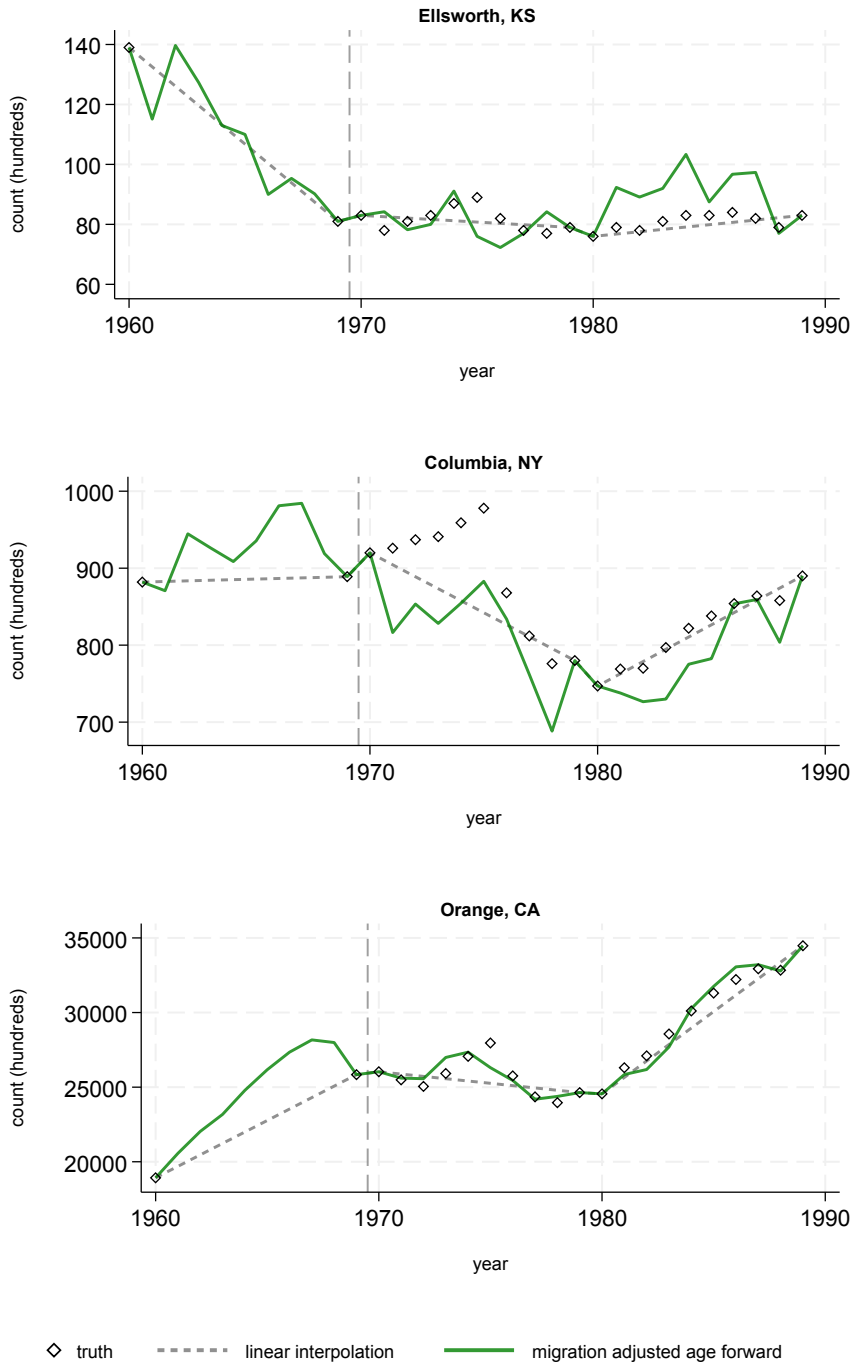


Figure 5: County case studies: prediction problem

Notes: Using the counties identified in Figure 4, these plots show the time series for the number of four-year-olds (in hundreds) from 1960-1989. Hollow diamonds show the "ground truth" count of four-year-olds per the 1960 census and 1969-1989 SEER files. Indeed, the series for 1961-1968 is incomplete. The dashed gray line shows the count of four-year-olds predicted by linear interpolation. A solid green line shows the count of four-year-olds predicted by the cohort component migration adjusted age forward method, as described in Section 4.

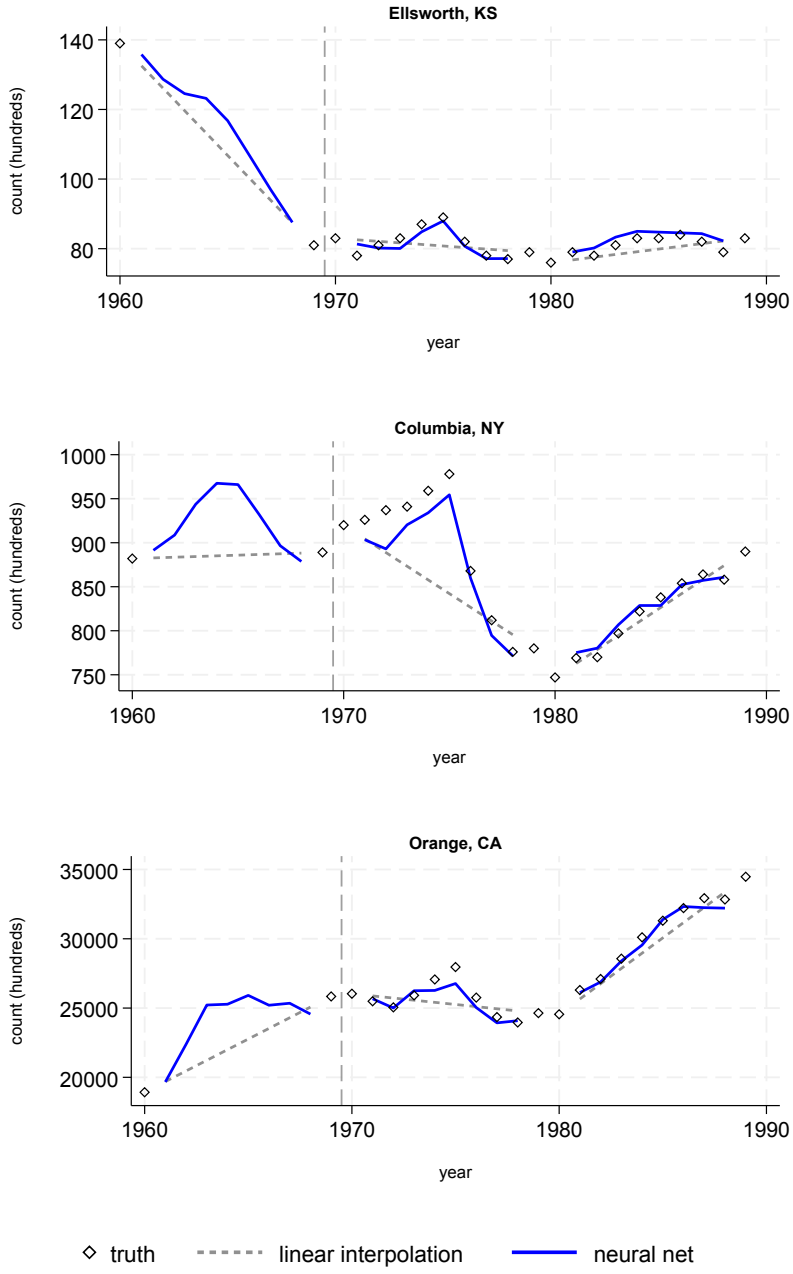
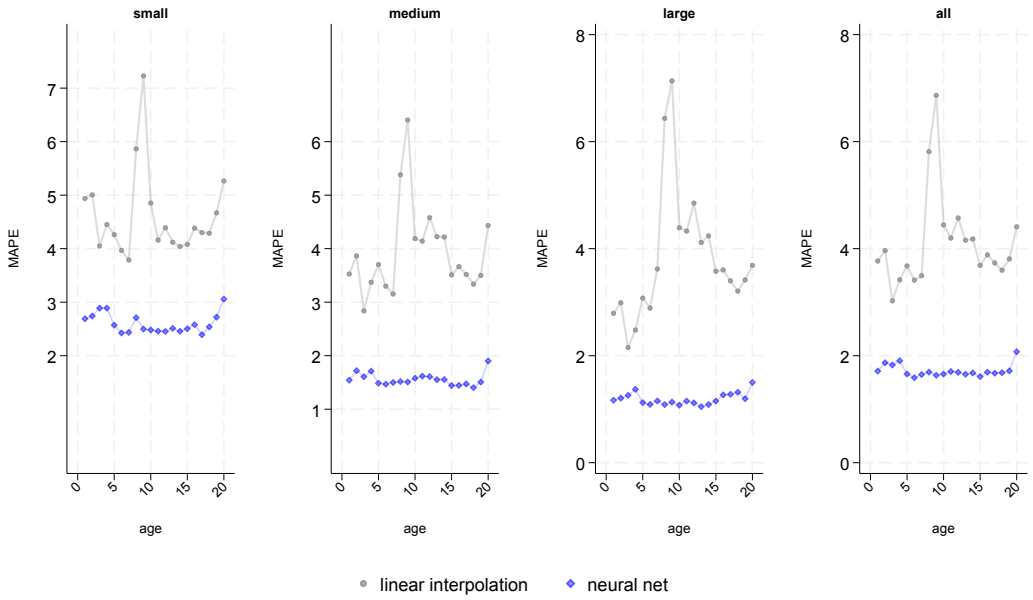
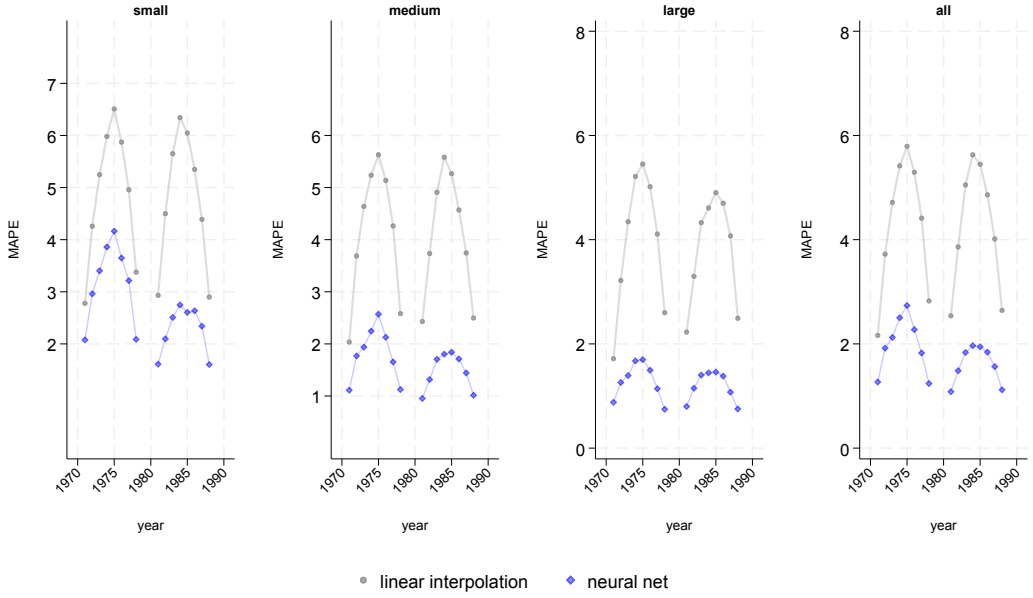


Figure 6: County case studies: horserace

Notes: Using the counties identified in Figure 4, these plots display predictions of the count of four-year-olds (in hundreds) from 1960-1989. Predictions generated from our neural network model are shown by a solid blue line. Predictions drawn from linear interpolation are shown by a dashed gray line. The "ground truth", as represented by the 1960 Census and 1969-1989 SEER tabulations, is plotted as hollow diamonds. Because our model does not predict population counts in years ending in zero or nine (or what we denote as decadal end-caps), the time series for each prediction is missing in such years. This is because we already have available to us the value of ground truth in 1960 and 1969 and our prediction problem is to interpolate values *between* these years. This distinction ports directly to how we calculate out-of-sample goodness of fit, with the added caveat that we do not include zero-year-olds. A vertical grey line distinguishes where the predictions become out-of-sample: the 1960s.



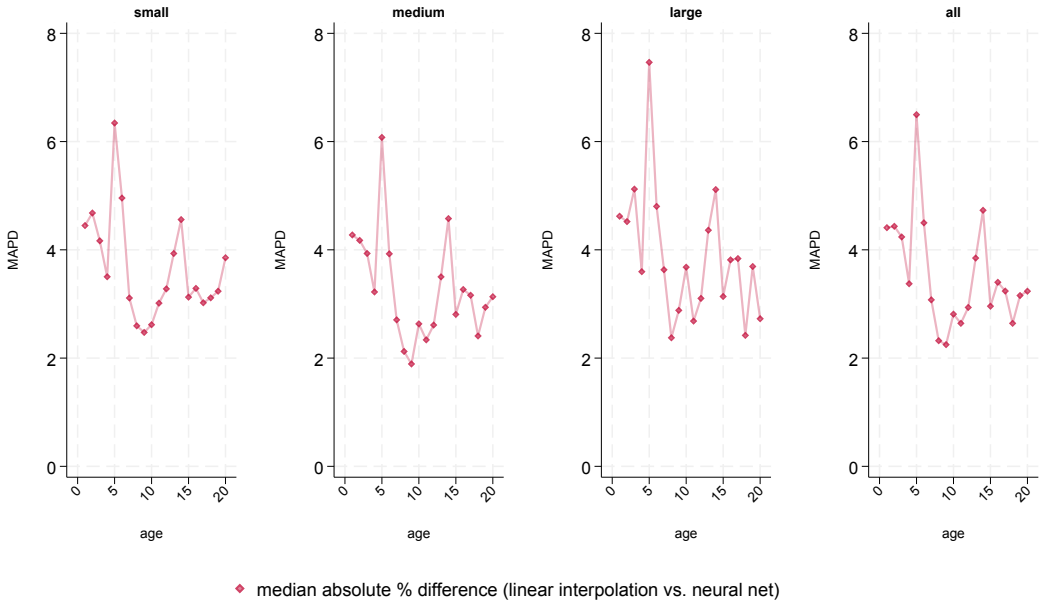
(a) MAPE by age



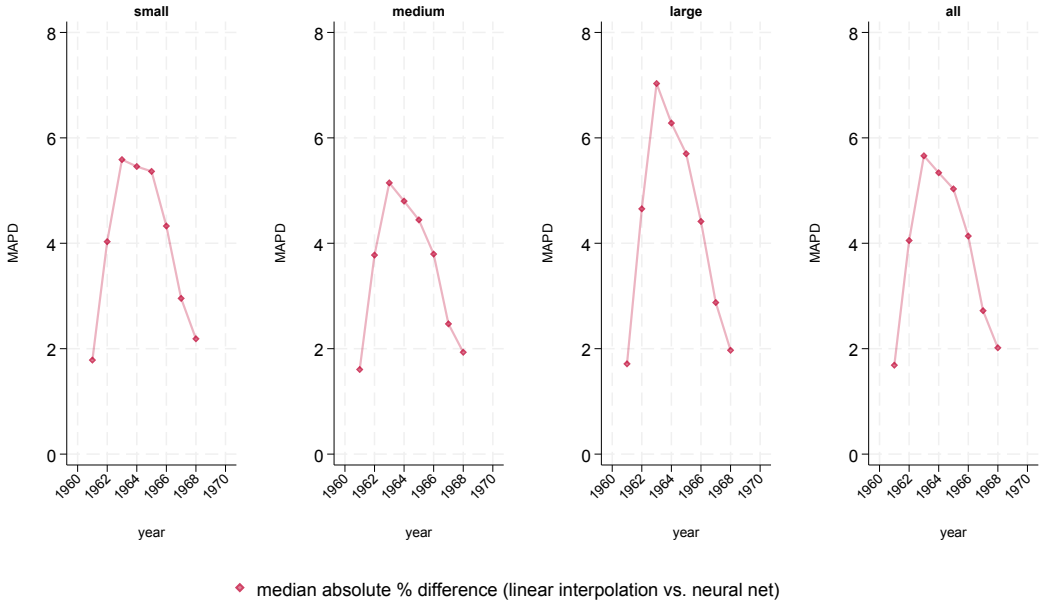
(b) MAPE by year

Figure 7: Out-of-sample predictive accuracy: testing counties

Notes: Panel (a) shows median absolute percentage error (MAPE) by single year of age (1-20), calculated using linear interpolation (gray line connecting dots) and our neural network predictions (blue line connecting diamonds). MAPE by age is calculated for only testing counties, separately by size (small, medium, and large). It represents the median error as a percentage of "ground truth". The rightmost panel pools across all testing counties (N = 385). This corresponds to 123,200 county-age-years. Panel (b) shows MAPE by single calendar year. For each panel, only years between decadal end-caps are used (so those ending with one or eight). This can be seen visually in Panel (b), where the series are missing for 1970, 1979, 1980, and 1989. Because we do not observe "ground truth" population counts in the 1960s, we cannot directly assess the accuracy of predictions over this period.



(a) MAPD by age



(b) MAPD by year

Figure 8: Differences across predictions: testing counties

Notes: Panel (a) shows the absolute percentage difference (APD) between our neural network predictions and linear interpolation by single year of age (1-20) for the 1960s. APD is calculated separately by county size (small, medium, and large), and the rightmost panel pools across all testing counties (N = 385). This corresponds to 61,600 county-age-years. Panel (b) illustrates APD by calendar year. Like before, only prediction data from 1961-1968 is used since these are the years for which there exists a gap in "ground truth" population counts.

Prediction Model	County Size	Mean	1st ptile	10th ptile	50th ptile	90th ptile	99th ptile
Neural Net	Small	3.91	0.04	0.44	2.54	8.55	23.54
	Medium	2.15	0.03	0.27	1.50	4.57	10.54
	Large	1.51	0.02	0.19	1.07	3.29	7.43
	All	2.63	0.03	0.28	1.65	5.75	16.84
LI	Small	5.83	0.05	0.84	4.54	12.33	23.64
	Medium	4.78	0.07	0.74	3.89	9.79	18.22
	Large	4.50	0.07	0.69	3.65	9.37	16.78
	All	5.09	0.07	0.76	4.04	10.64	20.23

Table 1: Distribution of absolute percentage errors: testing counties

Notes: This table shows summary statistics on the distribution of absolute percentage errors generated by predictions from linear interpolation and from our neural network model. This distribution is summarized by county size, along several percentiles. The 50th percentile corresponds to the median absolute percentage error (MAPE), which is disaggregated in Figure 8. Values shown are computed using only testing counties (N = 385), ages 1-20. This corresponds to 123,200 county-age-years.

Prediction Model	County Size	Mean	1st ptile	10th ptile	50th ptile	90th ptile	99th ptile
Neural Net vs. LI	Small	4.48	0.06	0.64	3.47	9.34	18.12
	Medium	4.07	0.06	0.56	3.18	8.84	15.72
	Large	5.36	0.08	0.69	3.63	10.08	26.92
	All	4.51	0.06	0.61	3.37	9.33	17.70

Table 2: Distribution of absolute percentage differences: testing counties

Notes: This table shows summary statistics on the distribution of absolute percentage differences in population counts generated from predictions from linear interpolation and from our neural network model. This distribution is summarized by county size, along several percentiles. The 50th percentile corresponds to the median absolute percentage difference (MAPD), which is disaggregated in Figure 8. Values shown are computed using only testing counties (N = 385), ages 1-20. This corresponds to 61,600 county-age-years.

	# of features	*minibatch size* ($\times 10^{-7}$)	*penalty weight* (ψ)	*MAPE*
Baseline	203	1	3.30	1.75
LI only	1	1	4.50	4.03
No LI	151	0.0016	6.59	3.40
No Cohort Components	91	1	6.75	1.75
No LI year 9 error	195	1	7.43	1.87
No Curvature	191	1	7.43	1.74
No Std. Dev.	197	1	4.50	1.75
No (c,y) demog	104	1	5.45	1.77
No F10. lead	191	1	0.539	1.80
No Utility Features	147	1	0.717	2.21
Change LHS	203	1	2.93	2.37

Table 3: Comparing predictive feature sets

This table shows the selected tuning parameters and MAPE for each neural network model we construct, using a different set of predictive features. Each row corresponds to a different model with different predictive feautures. The number of features per model is shown in the second column. MAPE values are calculated using only testing counties (N = 385), ages 1-20. This corresponds to 61,600 county-age-years.

Online Appendix for *"Backcasting Population Data in the 1960s with Supervised Learning"*

By Esra Kose, Henry Manley, Douglas L. Miller

A Formulating baseline estimates

Linear interpolation (age-based) To construct LI_{cat}, we begin by letting $r(t) = mod(t, 10)$, where r indexes *relative* years from the beginning of the decade. Then, for each decade, we let $t^{max} = \underset{t}{\text{argmax}}\, r(t)$ and $t^{min} = \underset{t}{\arg\min}\, r(t)$ and substitute for t such that:

$$LI_{cat} = Y_{cat^{min}} + \frac{(Y_{cat^{max}} - Y_{cat^{min}})(t - t^{min})}{t^{max} - t^{min}} = Y_{cat^{min}} + \frac{(Y_{cat^{max}} - Y_{cat^{min}}) \cdot r(t)}{9} \tag{4}$$

where $Y_{cat^{min}}$ is the count at the beginning of the decade, $Y_{cat^{max}}$ is the count at the end, and their difference, $(Y_{cat^{max}} - Y_{cat^{min}})$ is apportioned linearly[1] over the decade. We know that for every decade, $t^{max} - t^{min} = 9 - 0 = 9$. As described in Section 4.2, there is additional consideration for the units of $(Y_{cat^{max}} - Y_{cat^{min}})$. Written above, the most parsimonious version calculates the difference in levels. However, we also implement a version with an IHS transformation that so that each year observes a constant growth rate. Additionally, we build a version of linear interpolation indexed by birth cohort and not age.

Age forward & age backward Drawn from a cohort-component model, age forward (AF_{cat}) works by holding a birth cohort $b(a,t) = t - a$ fixed through a decade and recursively subtracting deaths. This follows from the intuition that the count of a year old's this year is equal to the count of $a+1$ year-olds next year, minus the count of deaths to a year-olds this year.

One important detail of the age-forward approach is that not all birth cohorts appear in every year. In particular, it necessitates additional consideration for cohorts born mid-decade (e.g., the 1965 birth cohort). As a result, we consider the availability of relative years to vary by cohort, such that $t_b^{max} = \underset{t}{\text{argmax}}\, r_b(t)$ and $t_b^{min} = \underset{t}{\text{argmin}}\, r_b(t)$. Though, we also leverage prior definitions of t^{min} and t^{max} as cohort-invariant, within-decade indices. Taken together, and by substituting in birth cohort for age, we arrive at the following piecewise equation:

$$AF_{cat} = \left\{ \begin{array}{ll} Y_{cbt,} & t = t_b^{min} = t^{min} \\ B_{ct,} & t = t_b^{min} \neq t_b^{min} \\ AF_{cb,t-1} - D_{cb,t-1}, & t_b^{min} < t \leq t^{max} \end{array} \right\} \tag{5}$$

where the "base case" for a birth cohort that appears in every year of the decade is its count at the beginning of the decade $Y_{cbt^{min}}$. For birth cohorts born mid-decade, the count of births recorded in the cohort-defining year, B_{ct} serves as the base case. For every subsequent year, regardless of the number of times it appears in a decade, each cohort is decremented by the count of deaths to those in b reported during the preceding year, $D_{cb,t-1}$. This process recurses until the end of the decade is reached, where $t = t_b^{max} = t^{max}$.

As the name implies, age backward (AB_{cat}) is calculated by tracking a birth cohort *backwards* through a decade. Like with age-forward, age-backward also has to handle the case of birth cohorts beginning mid-decade. In practice, this means a cohort will "disappear" in the calculation of age-backward. Using the same notation as above that substitutes age for birth cohort, age-backward is written, piecewise, as:

[1] The difference $(Y_{cat^{max}} - Y_{cat^{min}})$ is apportioned, annually, according to the $\frac{t - t^{min}}{t^{max} - t^{min}}$ term.

$$AB_{cat} = \left\{ \begin{array}{ll} Y_{cbt}, & t_b = t^{max} \\ AF_{cb,t+1} + D_{cbt}, & t_b^{min} \leq t_b < t^{max} \end{array} \right\} \tag{6}$$

where the population count at the end of the decade (Y_{cbt}) is incremented each year, moving back in time, by the count of deaths to those in b reported occurring in t, D_{cbt}. Note, "disappearing" cohorts are those for which $t_b^{min} > t^{min}$. To construct consistent measures of age-backward for each age, we rely on data that extend our sample described in Section 2. One such example is how age backward might calculate, say, the count of 18-year-olds in 1961. Though our sample spans ages 0-21, we need to observe the 1943 birth cohort in 1962-1969 to calculate age-backward. And if we need to the count of 21-year-olds in 1961, we will need to recurse on the count of 30 year-olds in 1969 to get there. In practice, this means we need data on up to $max(a) + max(r(t)) = 21 + 9 = 30$ year-olds, which is available in the SEER tabulations.

Migration-adjusted age forward Migration-adjusted age-forward is akin to the age-forward measure, barring one additional consideration: it incorporates an estimate of net migration. This estimate comes from the difference in the AF prediction and ground truth in the year t_b^{max}. For the majority of cohorts, this is the difference in predictions at the *end* of the decade. Net migration is linearly interpolated for the count of years the cohort appears in the decade. When combined and calculated separately by decade, this measure is denoted as:

$$MAAF_{cat} = AF_{cat} + \frac{(AF_{cbt_b^{max}} - Y_{cbt^{max}})(t - t_b^{min})}{t_b^{max} - t_b^{min}} \tag{7}$$

where $AF_{cbt_b^{min}} - Y_{cbt^{min}}$ is the estimate of net migration that is scaled by the number of years from the base cohort-decade year. Note, if $AF_{cbt_b^{min}} > Y_{cbt^{min}}$, then this measure adds an estimate of net out-migration. If the converse is true, we adjust AF according to an estimate of net in-migration.

B Custom cross validation algorithm

For our main estimates, we use the following tuning parameters:

- 3 hidden layers
- 120 nodes per hidden layer
- L1 penalization on weights, with penalty parameter given by $\lambda = 0.03/slopes$, where $slopes = (features \cdot width) + 2 * width^2 + width$
- 20,000 observations per minibatch
- 200 maximum epochs
- Rectified linear (ReLu) activation function in each node
- Adam optimization routine

The optimization algorithm (Adam) (Kingma & Ba, 2014), activation function (ReLu), and the use of an L1 penalization on the weights are fixed choices, not chosen through tuning based on model performance. The choice of three layers was based on informal comparisons we made during the process of tuning the other parameters. The main parameters that we select through tuning are: the number of nodes per layer (width), the L1 penalty parameter λ, the minibatch size, and the maximum number of epochs used to train the model, $\phi = \{nodes, \lambda, batchsize, epochs\}$. We use 5-fold cross validation average Median Absolute Error as a measure of goodness of fit, with the folds defined over the training data, and observations randomized into each fold based on their county.

With four model parameters to tune, the choice space is large. For a given set of candidate parameters, the computer we used took an average of 30 minutes to compute a cross-validation goodness of fit measure.

An additional complication is that due to the random initialization of the model parameters, the same set of tuning parameters can produce slightly different models (and hence slightly different goodness-of-fit measures). This led us to use the following algorithm to choose our tuning parameters:

1. Choose a set of grid points $\phi_j \in \Phi$ to consider as candidate tuning parameters.
2. For each grid point, compute a cross-validation measure of RMSE. Add this point and RMSE value to the set of measurements, $\{\phi_i, RMSE_i\}$
3. Estimate a locally weighted polynomial using all completed measurements, $\widehat{RMSE_i} = f(\phi_i)$
4. Review the parameter values with the best predicted loss, using human judgement
5. Go back to step (1), with a new choice of grid points

In practice, for step (1) we often search over only two or three of the tuning parameters, holding the other one or two parameters fixed. In the end, we performed a total of 273 estimates. This resulted in best values of the tuning parameters as: number of nodes = 55, number of epochs = 80, L1 Penalty $= 0.6/slopes$ (with $slopes = (features \cdot width) + 2 * width^2 + width$), and minibatch size = 25,000. We then rounded the number of nodes up to 60, reflecting the fact that several of the "runner-up" best predicted choices came from this value.

C Harmonizing county identifiers over time and across datasets

The conceptual geographic unit of analysis in our paper is the county. Considering different sources of data and different years, our paper combines several different datasets. Among these, county ID is recorded in many different ways – including changes in recording within a data source across years. The table below summarizes some of these differences. Separate from recording differences, there have been changes to some county compositions over the years 1960-1990. Some counties have split or merged.

Source	Years	County ID encoding
Natality tabulations	1959-1967	text (County name)
Natality microdata	1968-1969	NCHS codes vintage 1962-1969
Natality microdata	1970-1981	NCHS codes vintage 1970-1981
Natality microdata	1982-1988	NCHS codes vintage 1982-1988
Mortality microdata	1960-1961	NCHS codes vintage 1960-1961
Mortality microdata	1962-1969	NCHS codes vintage 1962-1969
Mortality microdata	1970-1981	NCHS codes vintage 1970-1981
Mortality microdata	1982-1989	NCHS codes vintage 1982-1989
1960 Census population tabulations	1960	FIPS (1960 vintage)
SEER	1969-1999	SEER adjustments to FIPS

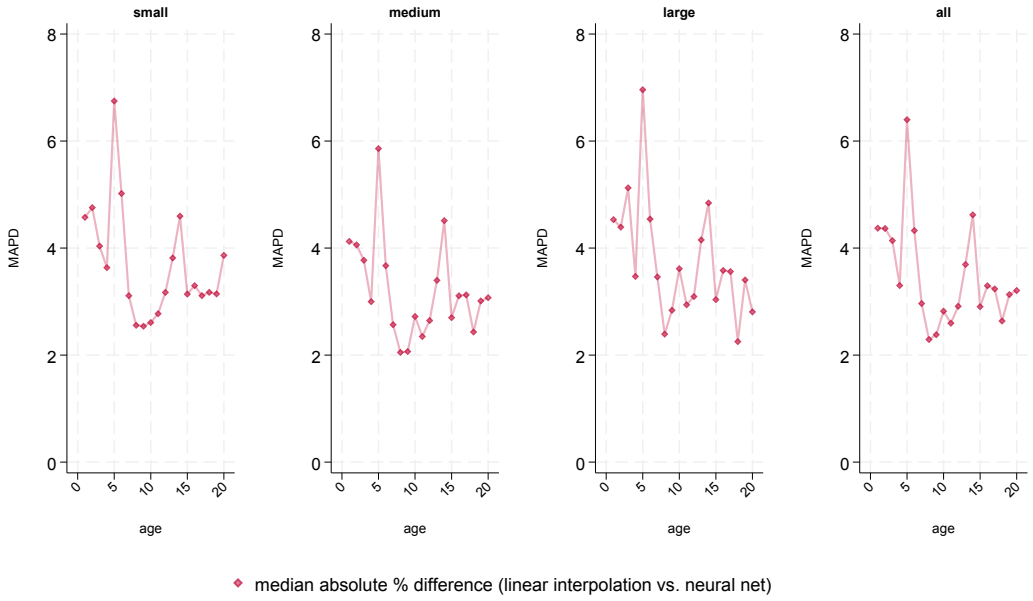
To construct our data set, we account for these two challenges by creating a series of cross-walks from each data source into a "super FIPS" county measure. This is intended to be a geographic unit that is consistent over time, which is as close as possible to the county, and into which each data source from each year can be merged.

The details of the construction of these cross-walks are embedded in the Stata code among the supplementary materials for this paper. Prior work by Autor & Dorn (2013) was greatly helpful as we built these crosswalks. Here we note a few specific considerations and challenges:

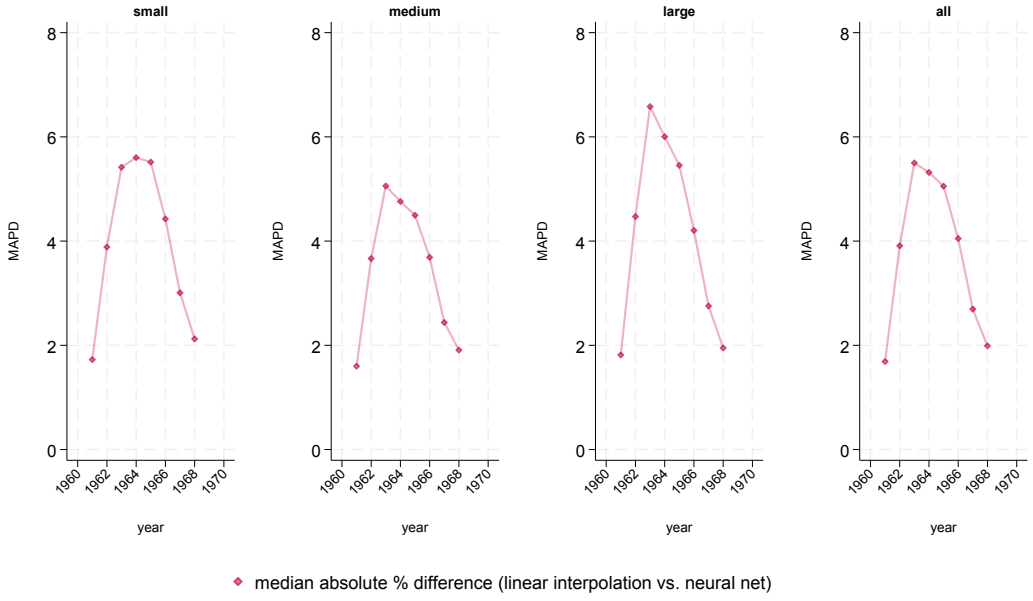
- SEER pools all of Alaska and Hawaii into one geographic identifier per state. So we lose sub-state geographic information for these two states.
- Virginia was a real challenge. Based on lots of web searching, we built a spreadsheet to encode the information needed to handle the many special cases for Virginia across different years and datasets.

- For SEER data for 1969-1979, all five NYC boroughs are given only one code. As a result, we combine all of NYC into one super FIPS.
- Among the datasets that record counties by the text of their names, there are inconsistent spellings and abbreviation conventions. Our crosswalks correct for these differences.

D Figures and Tables



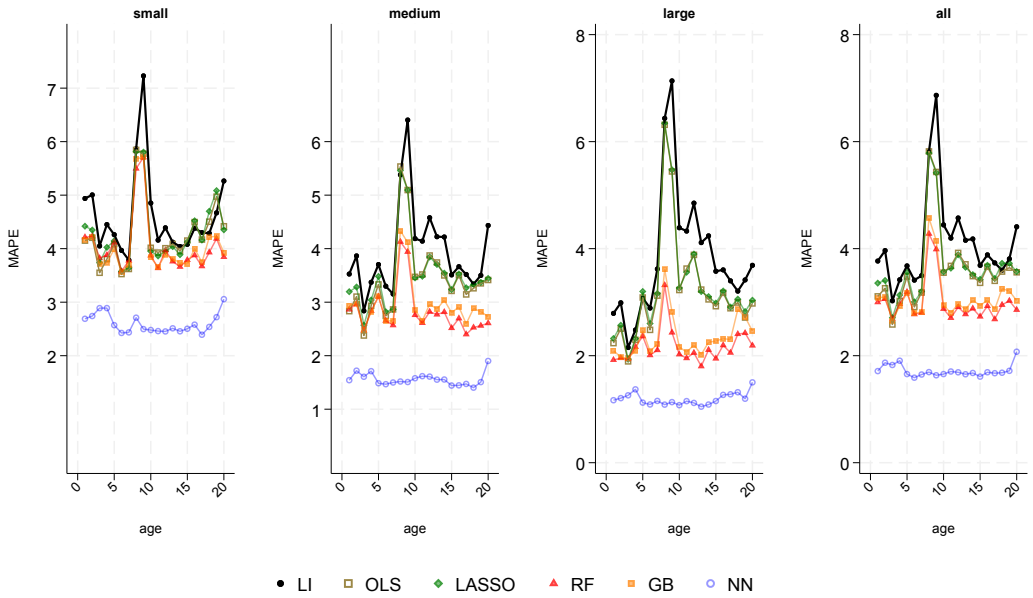
(a) MAPD by age



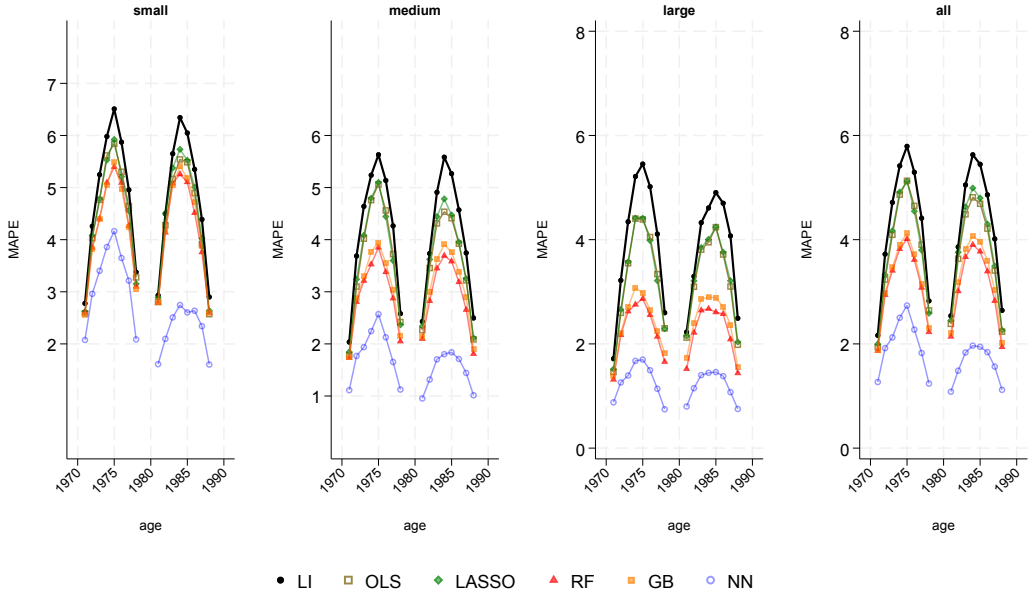
(b) MAPD by year

Appendix Figure 1: Differences across predictions: all counties

Notes: Panel (a) shows the absolute percentage difference (APD) between our neural network predictions and linear interpolation by single year of age (1-20) for the 1960s. APD is calculated separately by county size (small, medium, and large), and the rightmost panel pools across all counties (training + testing) (N = 2,697). This corresponds to 863,040 county-age-years. Panel (b) illustrates APD by calendar year. Like before, only prediction data from 1961-1968 is used since these are the years for which there exists a gap in "ground truth" population counts.



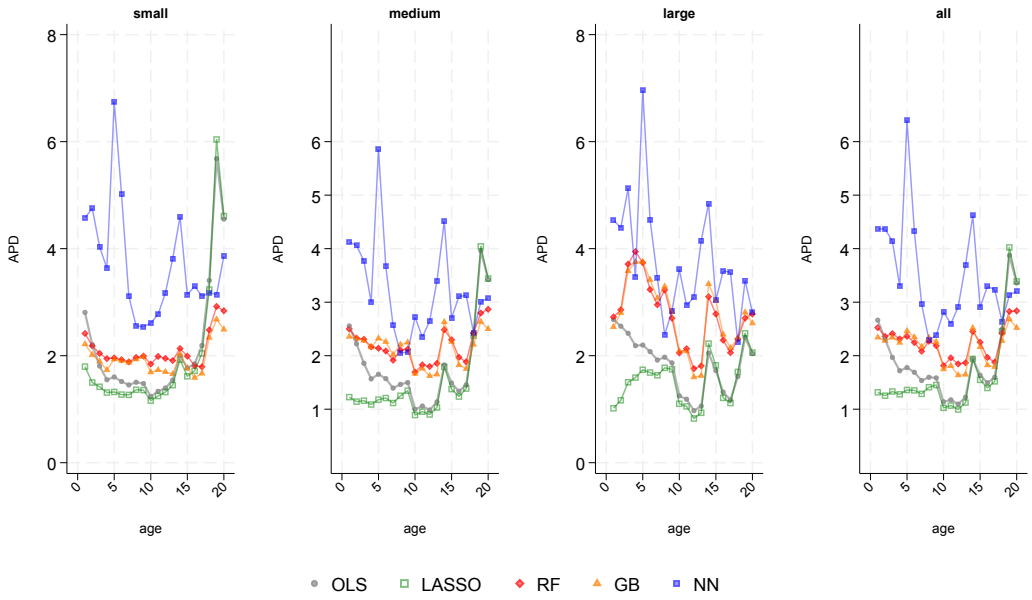
(a) MAPE by age



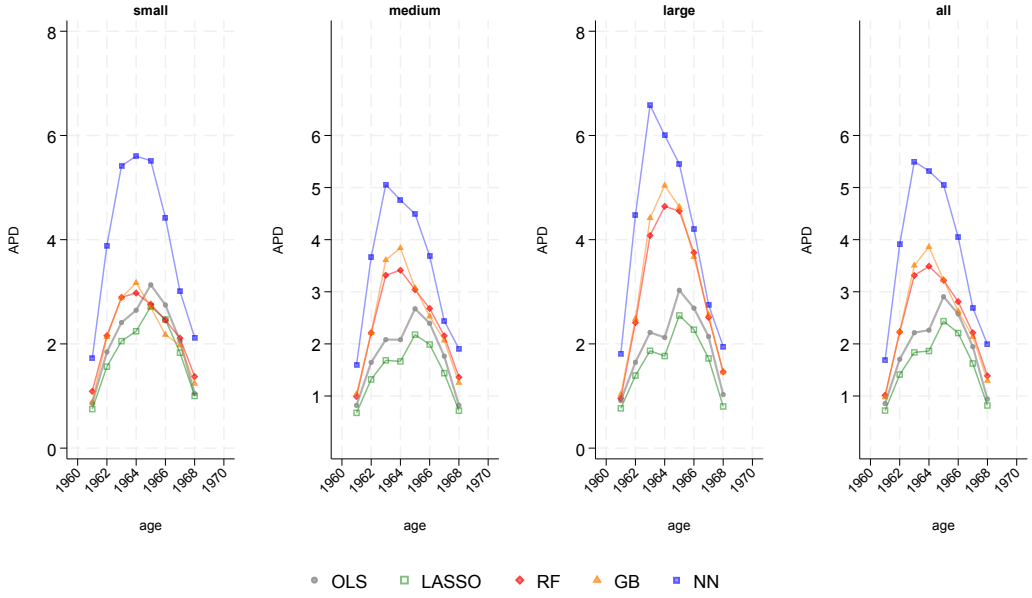
(b) MAPE by year

Appendix Figure 2: Performance by choice of prediction model: testing counties

Notes: Panel (a) shows median absolute percentage error (MAPE) by single year of age (1-20), calculated using linear interpolation (black line connecting solid circles), our neural network predictions (blue line connecting hollow circles), and a variety of other prediction models. MAPE by age is calculated for only testing counties, separately by size (small, medium, and large). It represents the median error as a percentage of "ground truth". The rightmost panel pools across all testing counties (N = 385). This corresponds to 61,600 county-age-years. Panel (b) shows MAPE by single calendar year. For each panel, only years between decadal end-caps are used (so those ending with one or eight). This can be seen visually in Panel (b), where the series are missing for 1970, 1979, 1980, and 1989. Because we do not observe "ground truth" population counts in the 1960s, we cannot directly assess the accuracy of predictions over this period.



(a) APD by age



(b) APD by year

Appendix Figure 3: Level differences in predictions by choice of model: testing counties

Notes: Panel (a) shows the absolute percentage difference (APD) between a series of prediction models and linear interpolation by single year of age (1-20) for the 1960s. APD is calculated separately by county size (small, medium, and large), and the rightmost panel pools across all testing counties (N = 385). This corresponds to 61,600 county-age-years. Panel (b) illustrates APD by calendar year. Like before, only prediction data from 1961-1968 is used since these are the years for which there exists a gap in "ground truth" population counts.

Prediction Model	County Size	Mean	1st ptile	10th ptile	50th ptile	90th ptile	99th ptile
Neural Net	Small	3.21	0.04	0.40	2.23	7.02	16.35
	Medium	2.07	0.02	0.26	1.44	4.39	10.38
	Large	1.41	0.02	0.18	1.03	3.08	6.54
	All	1.60	0.02	0.20	1.12	3.45	8.07
LI	Small	5.40	0.08	0.80	4.27	11.34	21.28
	Medium	4.70	0.07	0.74	3.86	9.59	17.95
	Large	4.47	0.08	0.70	3.63	9.23	16.96
	All	4.55	0.08	0.71	3.69	9.40	17.30

Appendix Table 1: Distribution of absolute percentage errors weighted by population size: testing counties

Notes: This table shows summary statistics on the distribution of absolute percentage errors generated by predictions from linear interpolation and from our neural network model. These absolute percentage errors are weighted by population. This weighted distribution is summarized by county size, along several percentiles. The 50th percentile corresponds to the median absolute percentage error (MAPE), which is disaggregated in Figure 7. Values shown are computed using only testing counties (N = 385), ages 1-20. This corresponds to 61,600 county-age-years.

Prediction Model	County Size	Mean	1st ptile	10th ptile	50th ptile	90th ptile	99th ptile
Neural Net vs. LI	Small	4.46	0.06	0.63	3.45	9.41	18.40
	Medium	4.00	0.05	0.56	3.11	8.49	15.16
	Large	4.61	0.06	0.65	3.52	9.42	17.32
	All	4.30	0.06	0.60	3.32	9.04	16.79

Appendix Table 2: Distribution of absolute percentage differences: all counties

Notes: This table shows summary statistics on the distribution of absolute percentage differences in population counts generated from predictions from linear interpolation and from our neural network model. This distribution is summarized by county size, along several percentiles. The 50th percentile corresponds to the median absolute percentage difference (MAPD), which is disaggregated in Figure 8. Values shown are computed using all (training + testing) 2,697 U.S. counties, ages 1-20. This corresponds to 863,040 county-age-years.

Prediction Model	County Size	Mean	1st ptile	10th ptile	50th ptile	90th ptile	99th ptile
Neural Net vs. LI	Small	4.24	0.06	0.61	3.39	8.96	16.28
	Medium	4.04	0.06	0.55	3.17	8.75	15.61
	Large	5.87	0.10	0.87	4.49	12.18	22.52
	All	5.45	0.09	0.77	4.09	11.40	21.78

Appendix Table 3: Weighted distribution of absolute percentage differences: all counties

Notes: This table shows summary statistics on the distribution of absolute percentage differences in population counts generated from predictions from linear interpolation and from our neural network model. These absolute percentage differences are weighted by an estimate of population. Because population counts are unobserved in the 1960s, an average of predictions from the neural network and linear interpolation are used. This weighted distribution is summarized by county size, along several percentiles. The 50th percentile corresponds to the median absolute percentage difference (MAPD), which is disaggregated in Figure 8. Values shown are computed using all (training + testing) 2,697 U.S. counties, ages 1-20. This corresponds to 863,040 county-age-years.