Nowcasting Distributional National Accounts for the United States: A Machine Learning Approach

Authors Gary Cornwall and Marina Gindelsky, U.S. Bureau of Economic Analysis[1]

Contact Marina.Gindelsky@bea.gov

Date September 2024

Abstract Inequality statistics are usually calculated from high-quality, comprehensive survey or administrative microdata. Naturally, this data is typically available with a lag of at least 9 months from the reference period. In turbulent times, there is interest in knowing the distributional impacts of observable aggregate business cycle and policy changes sooner. In this paper, we use an elastic net, a generalized model that incorporates lasso and ridge regressions as special cases, to nowcast the overall Gini coefficient and quintile-level income shares. We use national accounts data starting in 2000, published by the Bureau of Economic Analysis, as features instead of the underlying microdata to produce a series of distributional nowcasts for 2020–2022. We find that we can create advance inequality estimates approximately one month after the end of the calendar year, reducing the present lag by almost a year.

Keywords Inequality, income distribution, national accounts, nowcasting, machine learning

JEL Code C52, C53, D31, E01

[1]The authors would like to thank participants at the Federal Forecasters Conference and BEA Seminar Series for their helpful comments and suggestions.

Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE

1. Introduction

Income inequality is usually calculated by comparing individuals or households using available microdata, but do we *need* to wait for this data in order to obtain a reasonably accurate estimate? Currently, the U.S. Bureau of Economic Analysis (BEA) publishes a provisional (using partial data) distribution of personal income (PI) about one year after the close of a calendar year and official estimates one year after that. More timely estimates could provide valuable information to data users and policy makers. This paper leverages macro totals, from the National Income and Product Accounts (NIPAs), to produce advance distributional estimates of PI through a nowcasting process.[2] This estimate could be provided in conjunction with advance NIPA estimates of the previous calendar year—an improvement of nearly eleven months.

The nowcasting methodology exploits variation in the NIPA totals primarily from table 2.1, which disaggregates PI into its main components. Since BEA methodology distributes these subtotals to households according to their responses in microdata in order to calculate inequality (see 2000–2022 estimates and methodology in Gindelsky (2023)), the relationships between aggregated income sources are highly informative of inequality levels and trends.[3] These totals (e.g., labor earnings, income receipts on assets, government social benefits etc.), in addition to historical values of the distributional metrics, are used as features in an elastic net model, a generalization of penalized regression which includes both ridge and lasso as special cases, to construct nowcasts of the equivalized Gini coefficient and quintile-level shares of PI.

Arguably, the purpose of nowcasting is to provide information as quickly as possible during times of uncertainty and economic volatility, which typically result in policy implementation with distributional impacts. Therefore, we prioritize model performance during periods of economic turbulence, rather than maximizing fit during stable periods. Our approach not only outperforms a retrospective approach when nowcasting the COVID–19 period (2000–2022: initial shock, response, and recovery) both in turning point accuracy and error, but also fits the sample well overall, including the Great Recession period. We correctly predict 95 percent of turning points at the 0.1 reporting level for the full sample, and 89 percent for out-of-sample estimates. Furthermore, we show that these nowcasts are not particularly sensitive to standard revisions in the NIPA tables, making them suitable for use as an advance estimate.

[2]When we speak of "nowcasting", we speak of providing annual estimates of inequality alongside estimates of macroeconomic totals. We note that BEA produces macroeconomic totals for the preceding period (e.g., annual totals in 2024 for calendar year 2023) and thus our "nowcast" would be the first available inequality estimate, but for the previous year.

[3]The distribution is constructed by disaggregating PI into approximately 75 components, which are distributed separately according to the most pertinent micro information available (see Gindelsky (2023) for more detail). Each variable is estimated at the individual or household level and scaled such that the total matches the NIPA totals for that item in tables 2.1 and 2.9 (distinguishes between households and nonprofits). The time series is updated (and revised backwards) every year.

There are several key advantages from the approach we employ over current approaches which can be classified into three categories: (1) alternative microdata, (2) forecasting, or (3) nowcasting based on microsimulation. First, unlike much of the existing work, we are able to accurately estimate the effects of the COVID–19 pandemic and related policy responses on the income distribution accurately (significantly more so than with traditional forecasting techniques), without first having observed the economic impacts. We will demonstrate that the revisions are small on average, even in the out-of-sample period, which is crucial to the success of an advance distributional estimate. Second, the only contemporaneous data used by the model consists of several aggregates, which represent approximately 97 percent of PI. This simple approach is effective, while significantly less resource intensive than obtaining alternative microdata or building microsimulation models. Third, our estimates can be constructed and updated within minutes of a BEA release, as compared to an average time of 24 days to produce a nowcast (ranging from 8 to 90) found by O'Donoghue and Loughrey (2014) in a review of European methodologies. Finally, we anticipate that this approach can be generalizable to other datasets, wherein researchers may be able to analyze the relationships of the pertinent macro totals for their income distributions, and customize accordingly.

The remainder of this paper is organized as follows. In Section 2 we broadly cover existing approaches to nowcasting distributional statistics. In Sections 3 and 4 we outline the elastic net structure and the data we use. In Sections 5 and 6 we cover the main results of methodology, their generalizability, and sensitivity to revisions. Section 7 concludes.

2. Current Nowcasting Approaches

As discussed above, previous nowcasting efforts have focused on alternative microdata and/or microsimulation. For the BEA exercise, the primary underlying annual survey microdata (supplemented by other sources) and revised NIPA totals are available 9–10 months after the reference period, for the previous calendar year. Alternative microdata needs to have (1) a consistent time series with a regular release schedule, (2) broad coverage of well-measured income sources such as wages, self-employment, capital income, and transfers consistent with national account concepts, (3) a national sample which is representative of the population, and (4) a sample size which is sufficiently large for calculating granular income distribution statistics robustly. There are few microdata sources which meet these criteria in the United States.[4]

In the United States, there are two prominent approaches to nowcasting inequality: Blanchet et al. (2022) and the Distributional Financial Accounts of the Federal Reserve Board (Batty et al., 2021). Blanchet et al. (2022) produce a monthly distribution of national income (in addition to wealth estimates) for the bottom 50 percent, 50–90 percent, top 10 percent, top 1 percent, top 0.1 percent,

[4] Fixler et al. (2021) conducted an exercise in which the BEA distribution of PI, lagged by a year, was applied to quarterly NIPA totals and then averaged to assess performance of the nowcast. While the strategy worked well during growth periods, it underestimated inequality during turbulent periods.

and top 0.01 percent by predicting income growth in levels for each quantile and then subsequently calculating income shares. They nowcast income by using information from the Quarterly Census of Employment and Wages and monthly Current Population Survey to impute a monthly distribution of wages and also model unemployment insurance and COVID-specific transfers for each unit. Though this creative model performs reasonably well during periods of economic stability, there are some significant errors in levels and turning points in more volatile periods.

The Distributional Financial Accounts provide measures of the wealth distribution and are produced by the Federal Reserve on a quarterly basis. This measure primarily leverages the triennial Survey of Consumer Finances and macro financial accounts data, using the method from Fernández (1981) to interpolate between surveys.[5] Given the difficulty of interpolating measures for 2020 and 2021 using data for 2019 and 2022, there are significant revisions to the distribution of net worth by income share.[6] Additionally, attempts to extend the distributional methodology to income have run into significant challenges and prediction errors (Batty et al., 2021).

There have been many prominent attempts to nowcast inequality globally, most of which are focused on microsimulation. If the predictions from these models are accurate, the models have the advantage of flexibility in assessing the shocks and their impact, making them particularly useful for policy analysis. Accordingly, this approach has been the focus of the burgeoning literature in nowcasting distributional national accounts (Blanchet et al., 2022; Batty et al., 2021; Statistics Canada, 2023). Levy (2023) reviews these methods, as well as other recent approaches, and shows results for the Organisation for Economic Co-operation and Development Income Distribution Database for 2010–2019. The nowcasted estimates perform reasonably well with regards to most historical turning points, though the uncertainty intervals are often large relative to series variation.

Microsimulation models have been used by Eurostat and statistical offices in the European Union (France, Sweden), in addition to Canada, the United Kingdom, Australia, and others (Levy, 2023).[7] Generally, these methods perform very well during stable economic periods and rather less well (turning point errors) during the recent COVID–19 pandemic, during which there were large changes to the data generating process (DGP) stemming from both the underlying income distribution and in response to multiple policies. Many countries (and the Eurostat "flash estimates" (Leulescu et al., 2023)) base their imputations on EUROMOD, a microsimulation model developed by the European Union, combined with

[5]See Federal Reserve for details.

[6]Also see Batty et al. (2020)), chart 1 for the effect of revisions on the top 1 percent share. Like many inequality series which are modified with new data and methods periodically, previous versions are not archived and available to data users. Results from previous data vintages were generously provided to the authors upon request. This allowed us to identify the revisions and assess their magnitude and relevance.

[7]For the United States, a sophisticated microsimulation was developed at the Urban Institute, the Transfer Income Model (TRIM3), to simulate various cash and in-kind transfer programs in addition to health insurance and taxes. Additionally, the new Dynamic Simulation of Income Model (DYNASIM) projects income and health status for the U.S. population. These well-developed models are frequently used by researchers in a similar way to understand the impact of policy changes on the economy. See Urban Institute for more information.

survey data from the European Union Statistics on Income and Living Conditions (EU-SILC) (Almeida et al., 2021; Cantó et al., 2022; Christl et al., 2021; Leulescu et al., 2021, 2023).[8] In addition to these studies, individual countries use either the EU-SILC or another survey, along with other techniques, including but not limited to outside data sources (e.g., payroll data), additional modeling techniques, and interpolation (Aspachs et al., 2020; Helgeson et al., 2018; Junqueira, 2015; Navicke et al., 2013; Bönke et al., 2023; Blanchet et al., 2022; Li et al., 2022; Statistics Canada, 2023).[9] Recently, O'Donoguhe and Sologon (2023) formulated a new approach using a calibrated dynamic microsimulation to generate counterfactual income distributions which are a function of more timely data than survey or administrative records. Their approach consists broadly of three levels: a calibration to growth, a calibration to the labor market, and then the use of microdata for parameter estimation and subsequent simulation. However, they find that determinants of employment changed pre-and-post great recession and the model will not capture this change if the time period is too long.

3. Elastic Nets

In considering the merits of various nowcasting approaches for this exercise, we return to our objective: to improve the timeliness of BEA distributional accounts with minimal revisions. We note that our objective is not to make claims about the marginal impact of changes in NIPA totals on measures of inequality or income shares. This focus, combined with the limited nature of the data, means that traditional time series model structures use for forecasting such as the autoregressive integrated moving average (ARIMA) and vector autoregression (VAR) are not necessarily the right fit.

There are very few published attempts to forecast inequality given the difficulty of the task. It is well known that the best fitting model is often not the best forecasting model and this holds true in the context of inequality measures as outlined by Gindelsky (2018) and Castle et al. (2024). Gindelsky (2018) shows that naïve approaches outperform complex models for U.S. forecasts, but that there is a high degree of forecast error regardless, which detracts from its utility. In a new paper forecasting the top 1 percent income share in the U.K., Castle et al. (2024), found that the best performing forecasts for their pseudo-out-of-sample period which contained the start of the pandemic (2017–2021) were in fact naïve approaches: a random walk and smoothed random walk. However, the results of those forecasts were equivalent to the average top 1 percent income share of the previous 25 years. Moreover, they were accompanied by very wide 95 percent confidence intervals (+/- 2 standard deviations) which encompassed the minimum and maximum of the share during that time period. Accordingly, we agree with the conclusion of Castle et al. (2024) who wrote, "Comprehensive conditional models are needed to

[8]The latest Eurostat "flash estimates" (Leulescu et al., 2023) are produced by modeling the labor market (including transitions in and out of the labor force) together with simulating social benefits and taxes using EUROMOD.

[9]The analysis of the "signature" method of the U.K. incorporates many indicators for labor market conditions in addition to housing, credit market, and other macro indicators. However, this model is not analyzed or tested for the COVID–19 period because "such an event may alter the relationships learnt by the models during the estimation sample." (Cohen et al., 2023).

understand the evolution of income inequality and inform policy, but for forecasting, an alternative set of models is warranted" (p. 4).

Accordingly, we are not looking to *forecast* future periods, but rather *nowcast* to create an advance estimate of the distributional indicators using contemporaneous information. We emphasize that we are not focused on interpretability, as defined in the ceteris paribus and conditional expectation sense, as a necessary condition for an accurate nowcast.[10] Our target is not x_{t+1} conditional upon x_{t-j} with $j \in \{0, 1, \ldots\}$, rather it is x_t conditional upon x_{t-j} and a set of contemporaneous indicators derived from the NIPA tables and common macroeconomic measures.[11]

Consequently, we rely on the elastic net procedure, first outlined by Zou and Hastie (2005), to create nowcasting models of both the (equivalized) Gini and quintile-level income shares. The elastic net is a generalization of two shrinkage-based estimators, the least absolute shrinkage and selection operator (lasso) (Tibshirani, 1996) and ridge (Hoerl and Kennard, 1988) regression methods. The lasso minimizes the negative log-likelihood subject to an L1 penalty on the coefficients. This leads to a shrinkage estimator which improves predictive performance through the bias-variance trade-off and promotes parsimony since the L1 penalty allows parameters to shrink to exactly zero and thus be removed from consideration. This balance between predictive accuracy and parsimony does come at a cost. In many cases where correlation between two predictors is high, lasso tends to select only one variable and ignore the other. Moreover, if the number of predictors is greater than the number of observations the model quickly becomes saturated. Similarly, ridge regression also minimizes the residual sum-of-squares, though it imposes an L2 penalty on the coefficients. This again improves predictive performance through the bias-variance trade-off, but leads to models that are not parsimonious since a coefficient cannot be shrunk to "exactly zero." It is because of this that ridge regressions cannot be parsimonious by construction. While neither method strictly dominates the other (Fu, 1998), there is evidence to suggest that if the number of predictors is greater than the number of observations and the predictors are highly correlated then lasso is dominated by ridge (Tibshirani, 1996).

The elastic net is a generalization of the penalized regression framework which nests lasso and ridge as special cases. To see this, suppose we have some response vector $Y = [y_1, \ldots, y_N]'$, indexed by $i = (1, \ldots, N)$ and information set $X = [x_1', x_2', \ldots, x_K']$ where $x_1 = [x_1, \ldots, x_n]$ is an $N \times 1$ vector and X is an $N \times K$ matrix with $j = (1, \ldots, K)$ indexing the predictors. The elastic net estimator can be expressed as:

[10] There are differences in the literature in the use of "interpretability." Economists tend prefer a definition of interpretability to mean evaluation of a marginal change in an independent variable and its subsequent impact on the dependent variable holding all other information constant. Interpretability in the context of penalized regression models is often used to reflect parsimony. Going forward we will use interpretability in the latter sense.

[11] We recognize that a VAR structure can include both exogenous variables and their lags as appropriate. However, due to data limitations—the total observed data period is only 23 annual observations—the degrees of freedom are highly limited in a VAR estimation.

$$\hat{\beta} = \arg\min_{\beta} |Y - X\beta|^2$$
$$\text{s.t.} \quad (1-\alpha)|\beta|_1 + \alpha|\beta|_2 \leq t \text{ for some } t,$$

with,

$$\alpha = \frac{\lambda_2}{\lambda_2 + \lambda_1}, \text{ and } \alpha \in [0, 1].$$

The L1 penalty, λ_1, promotes sparsity by allowing coefficients to shrink to a true zero while the L2 penalty, λ_2, promotes model expansion and stabilizes the L1 regularization path (Zou and Hastie, 2005). The elastic net penalty, $(1-\alpha)|\beta|_1 + \alpha|\beta|^2$, is a convex combination of lasso and ridge where $\alpha = 1$ is a ridge regression and $\alpha = 0$ is a lasso. The elastic net is the penalized regression equivalent of having your bacon and eating it, too.[12]

To estimate the elastic net, Zou and Hastie (2005) propose an algorithm called the LARS-EN, or least angle regression – elastic net, which is based on work by Efron et al. (2004).[13] This algorithm has the same order of computations as that in a single OLS fit, meaning that implementation of the elastic net is computationally feasible, even in large dimensional settings. To select the tuning parameters, α and λ_2, we use leave-one-out cross-validation and choose the pair (α, λ_2) from an expansive grid of values that minimizes the cross-validation error.[14]

Similarly, in a follow-up to the elastic net estimator, Friedman et al. (2010) extended algorithms for generalized linear models with convex penalty structures (e.g., the elastic net). Germane to our use is the regularization of parameters in the context of matrix responses. If we observe our targets, Y, as a $T \times P$ matrix with P indexed as $j = (1, \ldots, P)$, and information set X as outlined previously then, with a slight abuse of notation which will be made contextually clearer in Section 4, we get:

$$\min_{(\beta_0, \beta_k) \in \mathbb{R}^{p+1}} \{-\mathcal{L}_j(\beta_{0,j}, \beta_{K,j}) + \lambda P_\alpha(\beta_j)\},$$

where $P_\alpha(\cdot)$ is the penalty term and $\mathcal{L}_j$ is the log-likelihood of the response vector. This allows for a penalized regression to be fit to multiple target vectors at the same time with the caveat that for each j the same subset of K predictors is used, though with different final coefficient values.

[12] The traditional saying is "have your cake and eat it, too," though it is the opinion of the authors that bacon strictly dominates cake in any setting, hence the change.

[13] There is an important difference between implementations of this algorithm between Stata (17) and the 'glmnet' package in R. Specifically, the R package—written by the authors of the original paper—uses an objective function which minimizes the negative log-likelihood while the Stata implementation uses the OLS objective function. We have run results using both methods and while they are qualitatively similar, they differ quantitatively in both parameter estimates and selected regularization parameters.

[14] We used several different cross-validation mechanisms given the data limitations outlined earlier. While the presented results are based on a repeated fourfold cross-validation scheme others explored included a leave-one-out, fixed-window rolling, and other k-fold cross-validation. In most cases the predictions and estimates are largely similar.

Our use of this will be in estimating the quintile-level income shares, a paradoxically convenient application since Friedman et al. (2010) point out that the "working response" is actually a row-stochastic version of the original target matrix; a transformation which is embedded in the concept of income shares.[15]

Recently, the elastic net has been incorporated into the economic forecasting literature. Cepni et al. (2019) uses the elastic net, in addition to other shrinkage and principal component methods, to nowcast Gross Domestic Product (GDP) in emerging markets. They find that shrinkage and other dimensionality reduction techniques perform comparably to "expert judgement" on predictors as proxied by the Bloomberg Relevance Index. Hofmarcher et al. (2015) argue that Bayesian lasso, ridge, and elastic nets provide better out-of-sample prediction properties than standard model averaging methods. They evaluate data from Sala-i Martin et al. (2004), a benchmarking dataset for model uncertainty and averaging.

Finally, Kim and Swanson (2014) compare principal component type models, shrinkage techniques, and model averaging performance in forecasting macroeconomic variables, though in a much longer time series (1960–2009). While they don't use the standard elastic net estimator we employ, eschewing it for a modified version put forth by Bai and Ng (2008), they find that shrinkage-based forecasts are comparable to the best alternatives.[16]

Elastic nets are not new to the nowcasting literature.[17] It is important to note that the elastic net is not modeled directly on the predictors themselves, rather as a preprocessing step the authors calculate a path signature. This preprocessing step is similar to other transformational approaches (i.e., principal-component analysis, independent component analysis, etc.) in that it is leveraging unique elements of the transformed space to better predict the target, often in a reduced dimensional setting from the original feature set.[18] While the transformation of features is different, in our case we use a standardization of both the features and target, the elastic net method is a common element of our nowcasting efforts and that of Cohen et al. (2023).

Nowcasting macroeconomic variables is a much larger literature beyond the confines of shrinkage models and model averaging. For example, in the most recent update for *GDP Now*, a nowcast of GDP growth provided by the Federal Reserve Bank of Atlanta, the construction of the nowcast consists of six steps

[15] The application of the improved algorithms found in Friedman et al. (2010) is found in the package "glmnet" in R. For response matrices the family $mgaussian$ can be chosen, an implementation we will leverage for results found in Section 5.

[16] This is by no means an exclusive literature review regarding the implementation of elastic net estimators in recent economic literature. Rather, our objective is to illustrate that the framework we use to predict inequality measures is gaining traction in the field more broadly.

[17] In an unpublished exercise, Murtin (2020) attempts a lasso regression to nowcast household income by decile, using more than 30 macro predictors for a group of countries for 2014. However, the errors are significant, particularly for the top and bottom of the distribution, likely given the heterogeneity of the sample.

[18] An additional alternative approach we investigated in the area of general-to-specific modelling was Autometrics, which employs an algorithm for automatic model selection (see Doornik (2009) for details). This approach was employed by Castle et al. (2024); Gindelsky (2018) in their inequality forecasting exercises. In this exercise, we used the similar *gets* package in R for the same purpose, but the algorithm produced lower quality out-of-sample results. However, we consider this exploration of this approach to be very valuable and worthwhile for nowcasting generally.

with several underlying model structures including a Bayesian VAR and autoregressive bridging equations constructed using weighted least squares (Higgins, 2014). More recently, and to an extent closer to the topic of this paper, Chen et al. (2019) use a variety of machine learning models—including lasso and ridge—to help produce more timely and accurate macroeconomic estimates in the face of data limitations. They find that non-parametric techniques offer considerable gains in prediction accuracy and are well-suited to conducting implicit variable selection.

4. Data and Methods

The data for this analysis come from the distributional accounts published by BEA annually in December of each calendar year.[19] The distribution is constructed by disaggregating PI, as presented in NIPA tables 2.1 and 2.9, into approximately 75 components that are distributed separately according to the most pertinent micro information available. Each variable is estimated at the individual or household level, and scaled such that the total matches the NIPA total for that item. Once PI (and subsequently disposable income by subtracting taxes) is calculated at a household level, it is equivalized by dividing by the square root of household size, so that households of different sizes can be directly compared. Finally, inequality statistics are constructed from the resulting household-level distribution. The full methodology for the construction of those accounts is available in Gindelsky (2023).

The core microdata used is the Annual Socioeconomic Supplement of the Current Population Survey (CPS). It contains the most relevant set of disaggregated income variables, with the largest available sample size, that is annually produced by the U.S. Census Bureau. It is crucial to this exercise that the core components of PI are well-measured. PI consists of many items including labor income (wages, self-employment), income receipts on assets (interest, dividends), and transfers (Social Security, Medicare, Medicaid, and other in-kind transfers) which require a large amount of information at the household level. The last two decades have seen significant policy and eligibility changes in many programs.[20] As the CPS is focused on such income sources and costly to administer, we are restricted to calculations tied to its annual publication. For our purposes, we focus on the variables which make up the majority of PI: labor income (wages and proprietor's income), asset income (interest and dividends), and transfer income (both in aggregate form, and unemployment insurance and tax credits specifically). These variables are not included directly in levels, but rather as *shares* of PI.

While there are other components of PI[21], such as employer contributions to employee pensions for example, they are omitted from the current analysis due to their limited correlation with household incomes. We also investigate other "external" measures of the economy (e.g., unemployment rate, labor force participation, or financial market indices) as available from Federal Reserve Economic Data (FRED), though we find that their inclusion does little to improve nowcasts. Table 1 details the calculation and

[19] Note that prior to the current version (begun in March 2020), BEA published a distribution of income in the 1970s.

[20] See timeline of policy changes here from the Center for Medicare and Medicaid Services.

[21] This exercise was repeated for disposable PI as well, with consistent results. We chose to focus on PI in this paper for brevity.

sources for the explanatory variables by category (NIPA vs. External) and provides recent values for context on their relative magnitudes.[22]

We begin by using the elastic net outlined above to estimate an equation for the Gini coefficient. Although there is considerable disagreement regarding the utility of the Gini coefficient in measuring inequality (see Ferreira (2020) for a discussion), it remains a seminal measure, most often referenced when the "net" change in inequality is of interest. Trends in the Gini coefficient are not necessarily aligned with trends in income shares. For example, the income share at the top of the distribution may increase, simultaneously with the bottom, with a net loss of share in the middle of the distribution, depending on the driving forces. By first attempting to forecast overall inequality, we were able to assess whether a model which can correctly predict the Gini (especially with respect to turning points) can provide important signals for the more granular quintile-level income shares.

To obtain advance estimates of the Gini and corresponding distributional measures we produce nowcasts using contemporaneous and lagged data. Since the frequency of the target variables is annual, the sample size is rather limited, particularly when compared to other macroeconomic estimation applications which employ either machine learning or non-parametric modeling methods. Moreover, many of the features are highly correlated, meaning that in standard regression analysis our parameters are unlikely to be identified in the traditional sense. The pseudo out-of-sample set is three annual observations from 2020–2022, which also happen to coincide with (perhaps) the most volatile economic periods in recent history due to the COVID–19 pandemic.

Our estimating equation for the Gini is:

$$\begin{aligned}\text{Gini}_t =& f(\text{Wages}_t,\ \text{Proprietor's Income}_t,\ \text{Income Receipts on Assets}_t,\\ &\text{Unemployment}_t, \text{Medicare}_t, \text{Medicaid}_t, \text{Tax Credits}_t, \text{I}_{2019}),\end{aligned} \tag{1}$$

where I_{2019} is an indicator for the 2019.[23]

As mentioned earlier, we use a leave-one-out cross-validation framework to identify the optimal values of α and λ. Since the goal is to nowcast both the Gini coefficient and quintile-level income shares we save the predicted values of the Gini, $\tilde{\text{Gini}}$, produced by this equation for use in our quintile-level estimation. For each quintile income share, $is_{t,j}$, with $j \in \{1, \ldots, 5\}$ we estimate the following:

$$\begin{aligned}is_{t,j} =& g_j(is_{t-1,j}, is_{t-2,j}, \text{Wages}_t, \text{Proprietor's Income}_t, \text{Income Receipts on Assets}_t,\\ &\text{Unemployment}_t, \text{Medicare}_t, \text{Medicaid}_t, \text{Tax Credits}_t, \text{I}_{2019}, \tilde{\text{Gini}}_t)\end{aligned} \tag{2}$$

[22] Castle et al. (2024) included such variables (GDP growth, real wage growth, inflation etc.) in their complex models, which were found to have substantially worse forecast performance than the random walk.

[23] We have added an indicator for 2019 to reflect the distributional anomaly. While the census CPS series for money income does not show a decrease in inequality for 2018–2019, calculations using tax data (CBO, Auten and Splinter (2024)) do show a significant drop in inequality, likely due to the fall in capital gains. As our exercise uses both kinds of data, we end up with a drop in top income shares and the Gini as a result, though our income measure does not include capital gains. Thus, we consider it a deviation and add an indicator for it. This is also consistent with the results of Castle et al. (2024), who found that saturation was key to a well-performing forecast for the U.S. top 1 percent pre-tax national income series.

The full set of $g(\cdot)$ is estimated using an elastic net with a multiple response matrix as the target. Included features are kept consistent across all quintile-level equations with variation allowed in the parameter of each. For example, $\tilde{\text{Gini}}_t$ will appear in every equation (or not, if it is shrunk to true zero) but the parameter value will vary across the five equations. In some cases the parameter may be either arbitrarily close to zero or farther from zero for other quintiles.[24]

We also compare our results with a naïve model, an AR(2) for the Gini and a reduced-form VAR(2) for the quintile income shares:

$$\text{Gini}_t = \alpha + \phi_1 \text{Gini}_{t-1} + \phi_2 \text{Gini}_{t-2} + \varepsilon_t \tag{3}$$

$$\text{Q}_t = \alpha + \Phi_1 \text{Q}_{t-1} + \Phi_1 \text{Q}_{t-2} + \varepsilon_t. \tag{4}$$

We will provide graphical evidence that these functions (produced by our elastic net models) fit well, not only to the in-sample period, but also to the pseudo out-of-sample period. We also provide revision and turning point analysis, as appropriate to an evaluation of BEA advance estimates, to further illustrate the predictive efficacy of our approach.

To evaluate the impact of revisions to NIPA totals, a process that takes place numerous times over the course of the estimate life cycle, we conduct an exercise to approximate the effect of the 2023 comprehensive update to our inequality metrics, as a reasonable upper bound to the potential impact of future NIPA revisions *alone* to the results. Results we provide in Section 5 indicate that the most important source of uncertainty stems from the DGP, rather than the revision process itself.

5. Main Results

We begin our analysis using a VAR specification in order to assess prediction quality without external information. As described above, we estimate three models corresponding to three nowcasts separately for the Gini and the quintiles. Although more than two lags were tried, they were never chosen by the elastic net models. Inequality is typically mean reverting (even if within the context of a long-run trend); an increase in one year is often followed by a decrease the following year. Since we are predicting the level of inequality rather than the change, the two-year lag of the level can encompass sufficient information for these processes.

Figure 1 shows the results for each model as follows: purple squares for 2000–2019 (nowcast=2020), red circles for 2000–2020 (nowcast=2021), and blue triangles for 2000–2021 (nowcast=2022). As expected, the VAR has very similar in-sample fit for each metric. However, out-of-sample performance (nowcast) is of mixed quality. For all metrics, the initial COVID–19 shock (2020) is missed completely. If 2020 is included in the "observed" sample for the models, the quintile models do fairly well overall (less well for quintile 4), but still have a significant error for the Gini (0.5). If 2021 is observed as well, however, most

[24] In untabulated results we have estimated these equations independently with a variety of structures including no cross-quintile lags, neighboring quintile lags, and a variety of other functional forms and lag structures. These results are available upon request of the authors but ultimately have little impact on the qualitative nature of our results.

still significantly underestimate the rebound in 2022, i.e., collectively predicting inequality to be lower than the actual data.

It is important to note, however, that the 2022 data used for this version of the exercise are still categorized as provisional. This designation reflects the fact that they are based on incomplete tax information. Therefore, it is difficult to evaluate model performance against an incomplete "truth" which is likely to change for the third estimate. Thus, while presenting the results of all periods, we will focus our evaluation on 2020 and 2021 data.

Traditionally, the advantage of nowcasting, as compared to a VAR, relies on the exploitation of contemporaneous data. This is particularly relevant for distributional national accounts, whose inequality metrics rely on scaling microdata to this contemporaneous macrodata. Figure 2 presents the models using the same VAR structure as in figure 1, but now supplemented with contemporaneous NIPA totals, as outlined in table 1. It is immediately clear that contemporaneous microdata significantly enhances both in-sample fit, and more importantly, out-of-sample accuracy for most quintiles (performance in quintile 4 is comparable to the VAR). In fact, we are able to nowcast overall inequality (the Gini) with a very high degree of accuracy for all three years. We exploited this high degree of accuracy in the Gini prediction by including the predicted Gini in our models of the quintiles. The predicted value of Gini contributes about 40 percent to the total level of each quintile. The quintile nowcasts are very accurate for the 2020 and 2021 models across the board. There is still some underestimating of the 2022 recovery at the bottom and top, but we keep in that that those are provisional estimates likely to change slightly.

Finally, we explore the possibility that additional external data series could further improve model accuracy in figure 3. Here we include all of the explanatory variables found in figure 2, and include "external" variables listed in Table 1 as well. Interestingly, these external variables do not meaningfully improve out-of-sample fit for most metrics, and in fact actually lead to greater errors for several. Additionally, they rely on the construction and availability of outside data series, which could potentially impact the release of advance inequality estimates.

We can also visually compare model performance for each metric.[25] For this comparison, we chose 2000–2019 (with 2020 nowcast) though the results are fairly stable for the other periods as well. Figure 4 shows this panel of comparisons. For the Gini, it is clear that visually the best model, in both fit and prediction, is the main specification. The main + external specification fits slightly less well, and the VAR substantially so, with a big miss in 2020 as discussed above. This is also the case for most of the quintiles, where either the main fits best, or there is little difference between it and the main + external model. However, for quintile 4, all models miss the small decline in 2020, and generally fit less well in the 2014–2020 period.

Another important metric for model performance is success predicting turning points. Although the graphical analysis allows us to eyeball fit around volatile periods, we also make an attempt to quantify

[25]We note that the model fit is somewhat worse during the early 2000s for some metrics. However, as our objective is to produce an accurate nowcast, rather than to fit previous periods well, we focus on more recent years.

these results for the whole period for each model in table 2. Each year-over-year predicted change is classified as having a "correct" sign if either (1) the model predicts an increase (decrease) in that year corresponding to an observed increase (decrease) in the data of at least 0.1 (i.e. 44.2 to 44.3), (2) the model predicts an increase (decrease) in that year corresponding to no increase (< 0.1) in the observed data, or (3) the model predicts no change (< 0.1), corresponding to an increase (decrease) in the observed data. An example of an "incorrect" sign would be in 2018 for the Gini coefficient, where the 2000–2019 model predicts an increase of 0.1 from 44.5 to 44.6, but the observed series decreases by 0.1 from 44.7 to 44.6. Table 2 sums the share of correct predicted signs for each model and metric, with the final columns summing over all metrics collectively for that model and time period.

While this is a very conservative turning point analysis, from the standpoint that arguably a Gini of 44.2 is not qualitatively different from 44.1 or 44.3, we err on the side of caution when evaluating our models, given the usually small year-over-year changes in inequality series. Naturally, we aim to reduce turning point errors overall, but we are particularly interested in reducing these errors during volatile periods which see more than a 0.1 percentage point change in inequality shares. One way to look at this directly during the COVID–19 period is in table 3, which provides the potential "revisions" for each nowcast, by metric. Since the purpose of this exercise is to construct an "advance" estimate of inequality based on macrodata, the actual estimate based on microdata scaled to subsequently revised macrodata would quality as the "observed," and thus is analogous to a revision. For example, the main 2000–2019 model predicts a Gini of 42.00 in 2020, with an observed Gini of 42.06, thus leading to a revision of 0.06. Similarly, the main 2000–2020 model predicts a Gini of 42.26, with an observed Gini of 42.21, leading to a revision of -0.05. These are then averaged both in raw and absolute values in the subsequent columns. This method of revision analysis is consistent with BEA's current revision analysis (Fixler et al., 2024).

The revision size is remarkably small for the very turbulent period. It does not significantly differ between the main specification and the main + external for most inequality metrics. The root mean square error (RMSE) is provided for both the main analysis period (2000–2021) and also including 2022, which does lead to an increase in error, though as mentioned above this is hard to interpret as the 2022 estimates are still provisional. Importantly here, we can see that the lowest RMSE is not always associated with the best out-of-sample prediction. Overall, the models fit very well for most series but are somewhat different in feature selection between the quintile functions. As an example, figures 5a and 5b show the relative importance of the features for each metric both for the main specification (figure 5a) and main + external (figure 5b) for the 2000–2019 models, as applied to 2020. Among only NIPA features, labor income (wages and proprietor's income) dominates the Gini model (about 2/3) with smaller contributions from the other components of PI. Once the predicted Gini is added to the quintile models, it contributes about 40 percent to each quintile, with lags of the income shares contributing another 50 percent. When external variables are added (figure 5b) we see that the greatest contribution comes from labor force participation (21 percent for the Gini and about 10 percent for the quintiles), but overall, external variables only comprise 30 percent for the Gini, and 3–15 percent for the quintiles. These relationships are salient regardless of the period chosen. Although it is not possible to interpret the magnitudes of the coefficients in such models, by aggregating variables in this manner we are able to assess their relative importance.

Generally speaking, in penalized regression contexts such as the elastic net, features are treated uniformly, unless there is reason to differentiate them. Historically, responses to recessions and other macroeconomic fluctuations have been largely contained to traditional monetary and fiscal policy channels. However, more recently, direct cash payments have been provided to individuals in response to economic activity consistent with recessionary periods.[26] Refundable tax credits (including stimulus payments) are a small share of PI in most periods, but have an outsize effect on inequality metrics during turbulent periods, when they increase. Mechanically, we account for this relationship via an unpenalized coefficient on tax credits.

We must also consider the impact of revisions to either the macrodata, the microdata, or the distribution methodology, particular to the question of distributing national accounts.[27] The advance estimate of PI in Q4, usually published in January with respect to the period of the previous quarter, contains a first annual estimate for the previous calendar year based on partial and preliminary data (Fixler et al., 2024); that is our first opportunity to produce an inequality estimate. The second and third estimates of Q4 that typically follow in February and March as source data is updated for some series also update those annual totals.

Given the relatively high level of aggregation in our data, we examined whether our results were sensitive to using the advance Q4 estimate vs. the third Q4 estimates. We found no difference in the model predictions between the two and thus assessed that it is appropriate to use the advance Q4 series. However, there are additional subsequent revisions which occur. First, while this analysis found that there is virtually no impact of revisions on model estimates from the advance estimate of Q4 to the third estimate (i.e., typically January to March), BEA also receives significant additional information which is incorporated for the September release every year (see 2024 update). From there, past years are also revised backwards on a revision schedule, with a comprehensive update taking place every five years.[28]

While the change in annual inequality is likely minimally impacted by the national accounts revision process, since the annual series are subsequently revised backwards, this is less likely to be true for the level. In the event of a methodology revision which differentially affects the relationship of the income components (i.e., wages increase relative to interest etc.), the final inequality estimate could deviate significantly from the prediction in a given year. For a recent example of this, consider the impact of the 2023 comprehensive update at BEA, which led to substantial decreases in proprietors' income and asset

[26] We acknowledge that, in 2020, when BEA might have created an advance estimate of these inequality metrics it may not have been clear that this channel of response was continual. However, over the pandemic period multiple direct cash payments were issued and it seems likely that this trend will continue with future responses to recessionary periods.

[27] It is usually impossible to anticipate such a revision or include it ex ante in a time series model. This is a particularly relevant concern for distributional national accounts, since national accounts totals are subject to frequent revisions. BEA has incorporated several consequential methodological updates since publishing the series in March 2020.

[28] See "Comparisons of Revisions to Real GDP" for a description.

income in some years. Since most of this income accrues to the top of the distribution, top shares were revised downwards by an average of 0.3 percentage points (pp), and inequality fell overall (downward revision of about 0.3 to the Gini in some years). In table 4, we show the impact of these NIPA revisions to our model, as an example.

In table 4a, we show what the model would predict for January 2023 (first Q4 vintage) and March 2023 (third Q4 vintage) (2020–2022 data) before the comprehensive update contrasted with the September 2023 vintage for the same years. As in the published series, we see a significant revision for 2020 to 2021, and a minor one from 2021 to 2022, with smaller differences year-over-year. This would be true even if the model is 100 percent accurate at the time of the advance estimate. However, this limitation affects any methodology which distributes national accounts.

6. Discussion

The previous section shows clearly that machine learning techniques should be explored in detail when looking to nowcast inequality, particularly in turbulent periods. Moreover, the results suggest researchers would benefit by being flexible by using the elastic net (rather than limiting themselves to certain α) and in the variable selection process. One group of variables might predict very well in combination, despite the small coefficients on some, while another may not.

As countries continue their efforts to produce distributional national accounts, they work to expand available time series. Although a short time series is often considered a significant limitation in more traditional modeling approaches (ARMAs, VARs, etc.) where degrees of freedom are a concern, elastic net models are more flexible in this regard. Additionally, it is not clear that additional years of data, if they were available, would improve the model fit.[29] It is highly likely that the nuanced relationships between the NIPA variables for each quantile are likely to be more consistent in the "short-run" (the exact time constraint will vary by data series).[30] However, this is difficult to prove without a longer series available for testing.

While the flexibility, speed and feasibility are no doubt key advantages of this method, there is still a significant disadvantage: inability to make causal inference or develop counterfactuals. Nowcasting models used by EU member countries, the U.K., Canada, Australia and others rely on microsimulation techniques allowing statisticians to estimate or simulate impacts of changing economic conditions or policies (whether ex ante for proposed, or ex post for enacted). While data is still being gathered on the performance of these models for the recent period, they have performed well for the pre–2020 period. Moreover, if they are found to perform poorly for COVID–19, statisticians can adjust the models to improve future performance (Statistics Canada, 2023).

[29] In the distributional national accounts used by Blanchet et al. (2022) and Castle et al. (2024), the trend is significantly different post–2000 (growing an avg 0.1pp per year) vs. pre-2000 (an avg 0.3pp per year).

[30] The sensitivity of these relationships is highlighted by Castle et al. (2024) as well, "forecasting rarely pays unless variables have closely interacting feedbacks" (p.4).

Accordingly, we view these nowcasting approaches as complementary to microsimulation, rather than as a direct substitute. Indeed, once a successful machine learning approach is found for a given inequality series, initial estimates of inequality deriving from these models may assist researchers in calibrating their simulations and reducing error, particularly when sign is an issue. For example, there was significant speculation in the United States as to whether inequality would rise or fall in 2020, based on the simultaneous effects of job loss (especially in the service sector), increase in unemployment benefits, and economic impact payments. If such a model had been available at that time, we would have been able to answer that question much more quickly.

Our experience constructing the distributional national accounts at BEA provides us with expertise in understanding the relationships between aggregate components of income and associated inequality measures. Through the iterations of these accounts and accompanying sensitivity analysis, we have gained insights into the underlying DGP. Although perhaps surprising to others who expect financial indicators such as stock market performance to perform well in predictive models, we were unsurprised to find that they did not improve accuracy (and often detracted), given their at-best indirect relationship to PI. In fact, including external information may negatively impact both the quality of the nowcast and its timeliness, as these indicators are not released alongside BEA NIPA data.

Our reliance on the NIPAs as the key features of our model does mean that we are subject to revisions to NIPAs as a potential source of error. As discussed in the previous section, though such revisions can be consequential to income sources overall, especially in a comprehensive update, they are less influential on our nowcasting model predictions. However, microdata sources are also subject to frequent revisions. In some cases, the survey instrument is revised (i.e., the 2014–2018 CPS survey change which led to increased inequality at the top), or else methodology is revised to account for issues such as increasing non-response. Series are not usually not adjusted backwards, leading to structural breaks, which are uncorrelated with business cycle changes or national accounts revisions.

Moreover, distributional methodology (the arduous process of assigning macro totals to micro households) continues to evolve. Researchers look for ways to improve accuracy of imputations, incorporate new sources of microdata, or even replace existing sources which have been discontinued. Globally, there have been efforts to incorporate new sources of data (e.g., credit card data, bank account information, real estate transactions) to improve accuracy. Methodological updates to macro- or microdata can significantly affect both levels and trends of resulting inequality series (or the disconnect between them, as in BEA estimates for 2019).

The impacts of revisions to the macrodata, microdata, and distributional methodology are a concern ubiquitous for all modeling endeavors. There is no way to anticipate all such changes ex ante, just as there is no single data source which covers all national accounts concepts, regardless of its source (i.e., survey, administrative, or private sector). Accordingly, researchers do their best to accommodate these changes as they appear and address their impacts to produce an accurate estimate. We view this exercise as another tool in pursuit of this objective.

7. Conclusion

Our application of the elastic net produces nowcasts of both overall inequality (Gini coefficient) and quintile-level income shares with a high level of accuracy, equal to or exceeding a naïve specification. As the United States does not have an official microsimulation model comprehensive enough for use in national accounts, we cannot compare our nowcasting models directly to such estimates. However, these models are flexible and capture variation very well during turbulent periods, including the recent COVID–19 period 2020–2022. Advance distributional estimates can be produced concurrently with advance NIPA totals for the previous calendar year, prior to the availability of microdata. The improvement in timeliness could significantly benefit data users and policy makers.

Our approach exploits the relationship between macro totals, appropriate for an exercise where household allocations must sum to aggregates. We find that both the largest components of income (e.g., labor and capital) and components which are smaller in magnitude but lead to volatility in income shares (e.g., tax credits, unemployment insurance, and health insurance) have the highest predictive power for the Gini coefficient. In turn, the predicted Gini coefficient and lags of the income shares directly make up at least 88 percent of the prediction for the quintiles. The main specification also correctly identifies turning points at least 85 percent of the time for each series.

In addition to the strong model performance, our approach demonstrates that a very simple modelling technique can be used to nowcast inequality within minutes of a macrodata release. We show that our nowcasts are robust to regularly occurring BEA revisions in national accounts totals. This robustness, combined with the nowcast accuracy, could facilitate its use as an "advance estimate" of the distributional accounts. As an advance estimate, the nowcast would provide information to data users about the state of inequality in the one-year period prior to the current release schedule. This would give the distributional accounts a similar structure to the quarterly release of the NIPA total by BEA.

Currently an advance estimate of (quarterly) GDP is provided the month after the close of a quarter with a second estimate, containing more data, released the following month. A third estimate is then released in the third month of the following quarter. The distributional accounts would follow a similar pattern with the nowcast taking on the role of the advance estimate, the provisional, available eleven months after the nowcasts, and official estimates being second and third estimates respectively. Although this approach is inappropriate for policy analysis, the timely and generalizable nature makes it applicable to other datasets and time periods, and a complement to existing nowcasting techniques primarily based on microsimulation.

References

Almeida, V., S. Barrios, and M. Christl (2021). The impact of COVID-19 on households' income in the EU. *The Journal of Economic Inequality 19*(3), 413–431.

Aspachs, O., R. Durante, A. Graziano, J. Mesters, J. Molina, and M. Reynal-Querol (2020). Real-time inequality and the welfare state in motion: Evidence from COVID-19 in spain. Economic Working Paper Series 1741, Universitat Pompeu Fabra.

Auten, G. and D. Splinter (2024). Income inequality in the united states: Using tax data to measure long-term trends. *Journal of Political Economy*. forthcoming.

Bai, J. and S. Ng (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics 146*(2), 304–317.

Batty, M., J. Bricker, J. Briggs, E. Deeken, S. Friedman, E. Nielsen, S. Reber, K. Sommer, and A. Volz (2020, November). Updating the distributional financial accounts. FEDS Notes.

Batty, M., J. Bricker, J. Briggs, S. Friedman, D. Nemschoff, E. Nielsen, K. Sommer, and A. Volz (2021). Measuring distribution and mobility of income and wealth. In *Measuring Distribution and Mobility of Income and Wealth*, pp. 641–677. University of Chicago Press.

Batty, M., S. Friedman, A. Volz, and S. Reber (2021, May). Joint distribution of wealth and income in the u.s.: Evidence from the distributional financial accounts.

Blanchet, T., E. Saez, and G. Zucman (2022). Real-time inequality. Working Paper 30229, National Bureau of Economic Research.

Bönke, T., G. Dany-Knedlik, and L. Pagenhardt (2023). New DIW Berlin model can nowcast the current income distribution; inequality likely to slightly increase in 2023. *DIW Weekly Report 43/44*, 283–289.

Cantó, O., F. Figari, C. Fiorio, S. Kuypers, S. Marchal, M. Romaguera-de-la Cruz, I. Tasseva, and G. Verbist (2022, June). Welfare resilience at the onset of the COVID-19 pandemic in a selection of european countries: Impact on public finance and household incomes. *Review of Income and Wealth 68*(2), 293–322.

Castle, J., J. Doornik, and D. Hendry (2024). Forecasting the UK top 1% income share in a shifting world. *Economica*. Accepted, May 6, 2024.

Cepni, O., I. E. Güney, and N. R. Swanson (2019). Nowcasting and forecasting gdp in emerging markets using global financial and macroeconomic diffusion indexes. *International Journal of Forecasting 35*(2), 555–572.

Chen, J. C., A. Dunn, K. Hood, A. Driessen, and A. Batch (2019). Off to the races: A comparison of machine learning and alternative data for predicting economic indicators. In *Big Data for 21st Century Economic Statistics*. University of Chicago Press.

Christl, M., S. De Poli, F. Figari, T. Hufkens, C. Leventi, A. Papini, and A. Tumino (2021). Quarterly report on the euro area. *European Commission, Directorate-General for Economic and Financial Affairs 20*(4).

Cohen, S. N., S. Lui, W. Malpass, G. Mantoan, L. Nesheim, A. de Paula, A. Reeves, C. Scott, E. Small, and L. Yang (2023). Nowcasting with signature methods. Working Paper 2305.10256, arXiv.org.

Doornik, J. A. (2009). Autometrics. In *in: Castle, J.L. and Shepard, N. (eds.) The Methodology and Practice of Econometrics: Festschrift in Honour of David F. Hendry*, pp. 88–121. Oxford University Press.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407 – 499.

Fernández, R. B. (1981). A methodological note on the estimation of time series. *Review of Economics and Statistics 63*(3), 471–476.

Ferreira, F. (2020, February). In defense of the Gini coefficient. World Bank Blog.

Fixler, D., M. Gindelsky, and R. Kornfeld (2021). The feasibility of a quarterly distribution of personal income. Technical Report 8, BEA Working Paper Series.

Fixler, D. J., E. de Francisco, and I. Schaaf (2024). Revisions to gross domestic product, gross domestic income, and their major components. *Survey of Current Business August*.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software 33*(1), 1.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics 7*(3), 397–416.

Gindelsky, M. (2018). Modeling and forecasting income inequality in the united states. Technical Report 7, BEA Working Paper Series.

Gindelsky, M. (2023). Technical document: A methodology for distributing personal income. Technical report, BEA Technical Document.

Helgeson, T., K. Lindstrom, and H. Hofsten (2018). Nowcasting household income in sweden. Paper prepared for IARIW 2018.

Higgins, P. (2014). GDPNow: A model for GDP "nowcasting". Technical Report 1407, Federal Reserve Bank of Atlanta Working Paper Series.

Hoerl, A. E. and R. W. Kennard (1988). Ridge regression. In S. Kotz and N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, Volume 8, pp. 129–136. New York: John Wiley & Sons.

Hofmarcher, P., J. C. Cuaresma, B. Grün, and K. Hornik (2015). Last night a shrinkage saved my life: Economic growth, model uncertainty and correlated regressors. *Journal of Forecasting 34*(2), 133–144.

Junqueira, V. (2015). Using micro-simulation in EU-SILC for early estimates of income: Strengths and limitations. *INE Statistics Portugal*.

Kim, H. H. and N. R. Swanson (2014). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics 178*, 352–367.

Kornfeld, R. (2023). Improved measures of regulated investment companies and real estate investment trusts for u.s. national economic accounts. *Survey of Current Business May*.

Leulescu, A., M. Agafitei, A. Gallelli, A. Di Guglielmo, and F. Demontis (2021, December). Nowcasting income indicators during the pandemic (FE 2020) - methodological developments. Technical report, Eurostat.

Leulescu, A., A. Di Guglielmo, and M. Agafitei (2023, June). Flash estimates of income inequalities and poverty indicators for 2022 (FE 2022) methodological note. Technical report, Eurostat.

Levy, H. (2023). *Nowcasting and Provisional Estimates of Income Inequality Using Microsimulation Techniques*. Number 12 in OECD Papers on Well-being and Inequalities. Paris: OECD Publishing.

Li, J., Y. Vidyattama, H. A. La, R. Miranti, and D. M. Sologon (2022). Estimating the impact of covid-19 and policy responses on australian income distribution using incomplete data. *Social Indicators Research 162*(1), 1–31.

Mataloni, L., D. Turner, and S. McCulla (2023). Improved estimates of national economic accounts. *Survey of Current Business November*.

Murtin, F. (2020). Nowcasting the household income distribution. Unpublished Working Paper.

Navicke, J., O. Rastrigina, and H. Sutherland (2013). Nowcasting indicators of poverty risk in the european union: A microsimulation approach. *Social Indicators Research 119*(1), 101–119.

O'Donoghue, C. and J. Loughrey (2014). Nowcasting in microsimulation models: A methodological survey. *Journal of Artificial Societies and Social Simulation 17*(4).

O'Donoguhe, C. and D. M. Sologon (2023). The transformation of public policy analysis in time of crisis – a microsimulation-nowcasting method using big data. Technical Report 15937, IZA Discussion Papers.

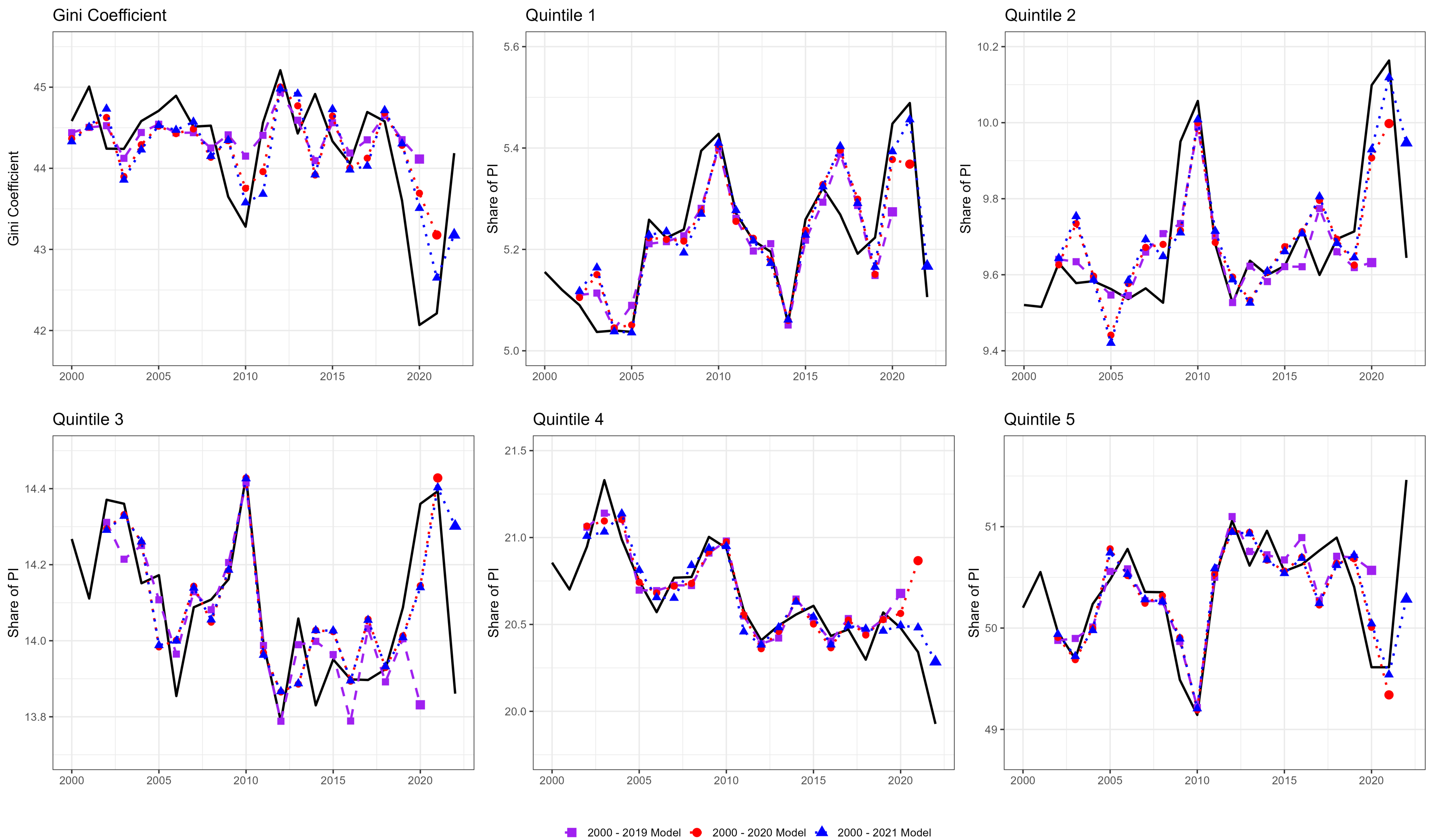
Sala-i Martin, X., G. Doppelhofer, and R. I. Miller (2004). Determinants of long-term growth: A bayesian averaging of classical estimates (BACE) approach. *American Economic Review 94*(4), 813–835.

Statistics Canada (2023). Distributions of household economic accounts, estimate of asset, liability and net worth distributions, 2010 to 2022, technical methodology and quality report. Technical report, Statistics Canada, Ottawa.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

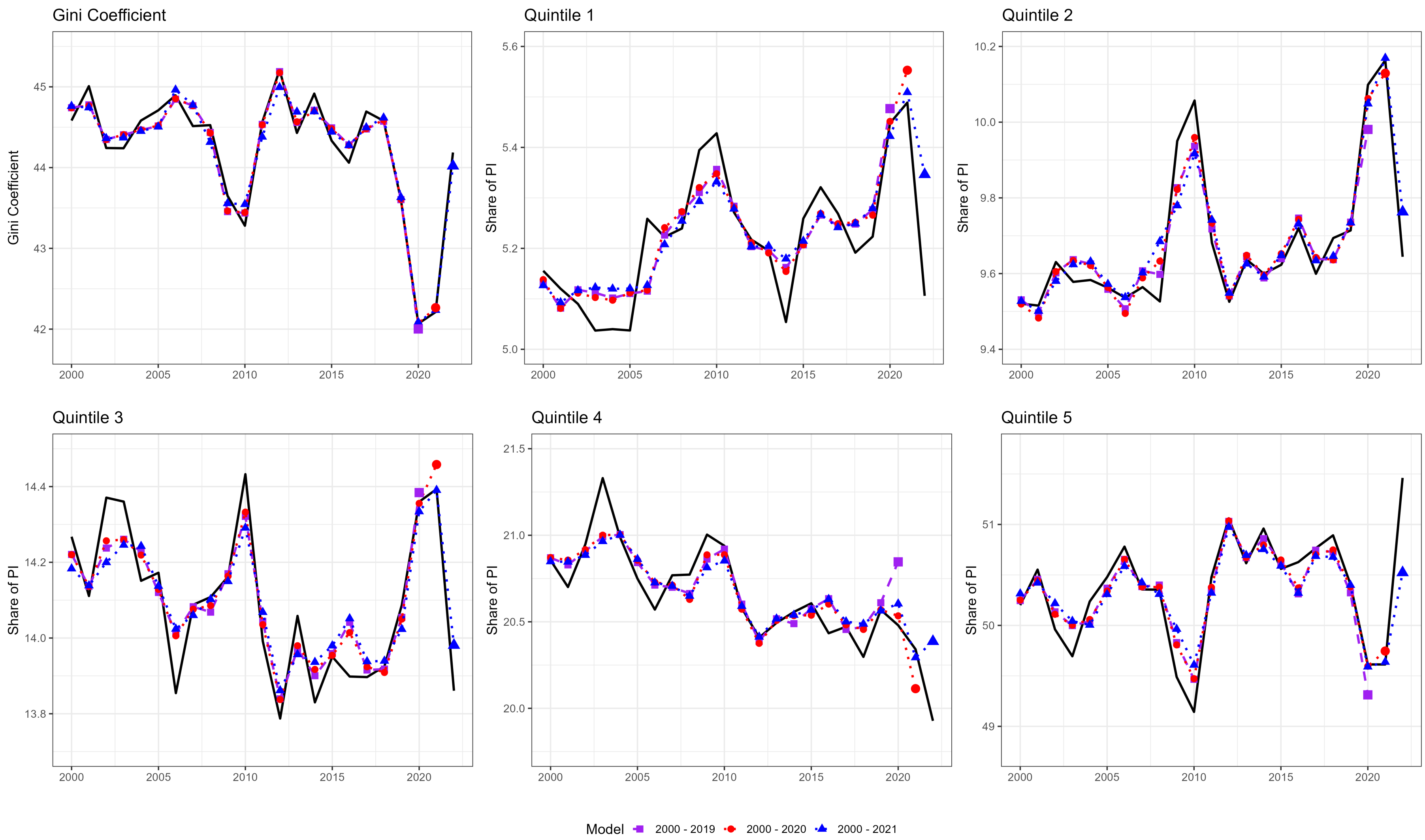
Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.

Figure 1. Naïve Specification



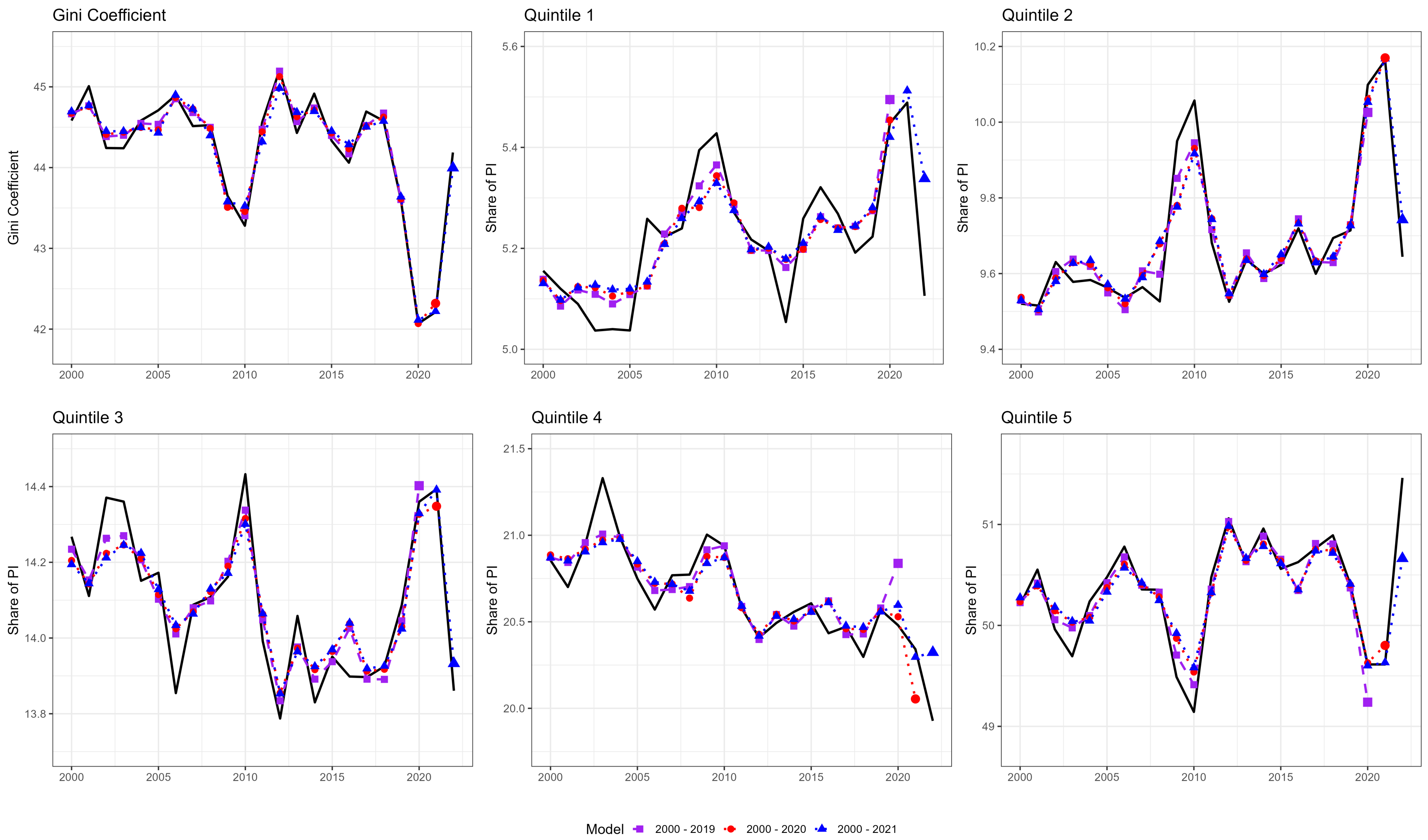
Note: Thi figures shows three models estimated for each metric with 1-year nowcasts from observed series for 2000–2019 (purple), 2000–2020 (red), and 2000–2021 (blue) lines. The black line denotes the observed series of metrics for personal income, as published by BEA in December 2023. The Gini was estimated with equation 3. The quintiles were estimated with equation 4, and subsequently normalized.

Figure 2. Main Specification



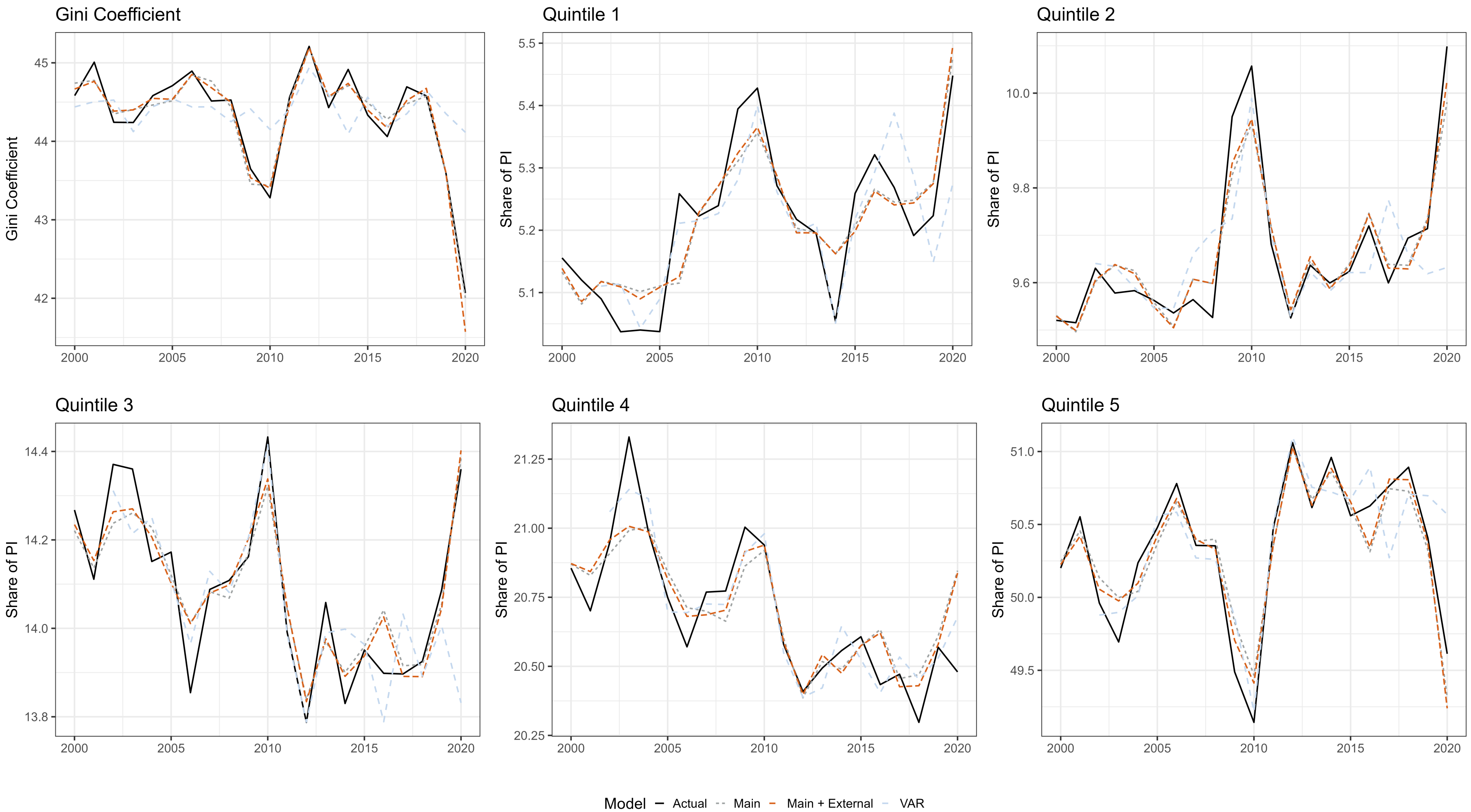
Note: This figure shows three models estimated for each metric with 1-year nowcasts from observed series for 2000–2019 (purple), 2000–2020 (red), and 2000–2021 (blue) lines. The black line denotes the observed series of metrics for personal income, as published by BEA in December 2023. The Gini was estimated with equation 1. The quintiles were estimated with equation 2.

Figure 3. Main Specification + External Variables



Note: This figure shows three models estimated for each metric with 1-year nowcasts from observed series for 2000–2019 (purple), 2000–2020 (red), and 2000–2021 (blue) lines. The black line denotes the observed series of metrics for personal income, as published by BEA in December 2023. The Gini was estimated with Equation 1, with the external variables as outlined in table 1. The quintiles were estimated with equation 2, with the external variables as outlined in table 1.

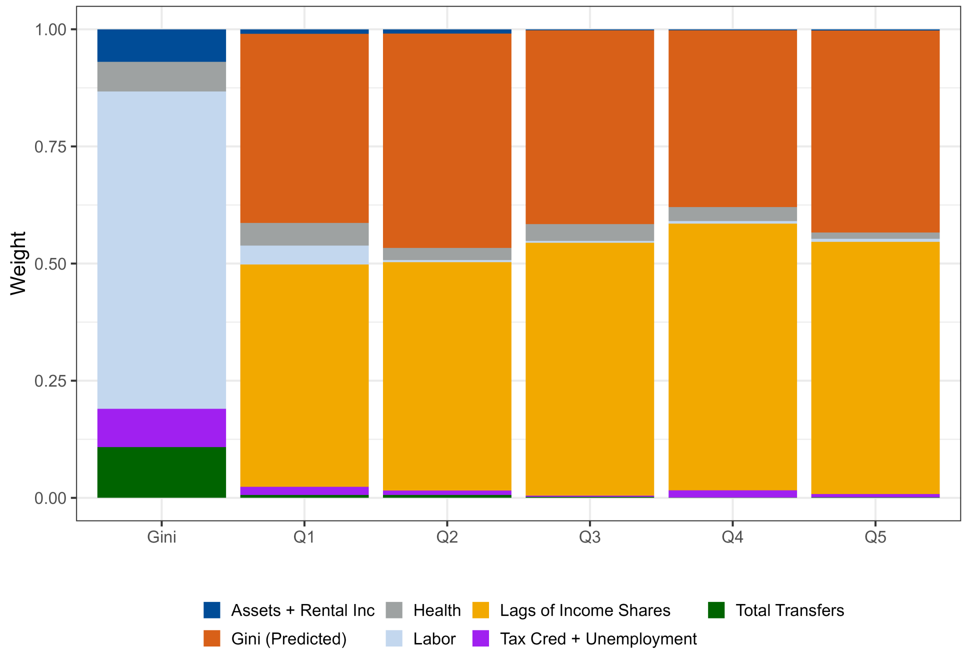
Figure 4. Model Comparison (2000-2019) with 2020 Nowcast



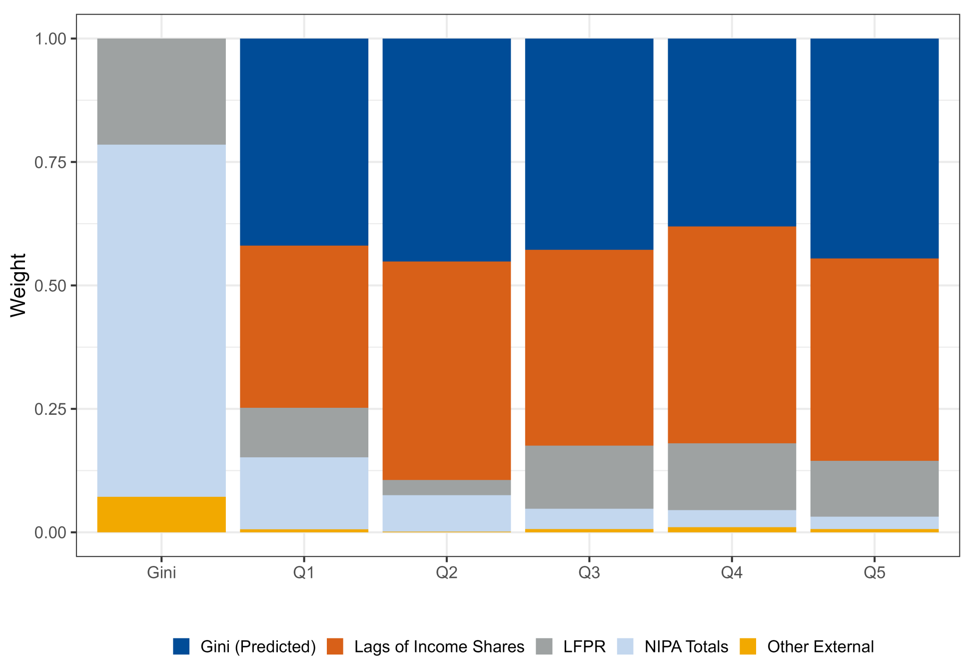
Note: Thi figures shows the models from figures 1 (blue), 2 (gray), and 3 (orange) directly compared for the 2000–2019 period, with 2020 nowcast, for each inequality metric. The observed series is in black).

Figure 5. Features for 2000–2019 Model

a. Relative Importance of Features (2020): Main Model



b. Relative Importance of Features (2020): Main + External



Note: This figure shows the numerical contribution of the explanatory variables to the total level of the variables for the main specification (5a) and main + external specification (5b). The relative contribution is calculated by summing the coefficients on each variable * variable value, and then aggregated as labeled. Health represents the shares of Medicare and Medicaid in personal income. Labor represents the shares of wages and proprietor's income. LFPR is labor force participation rate. "Other External" corresponds to the variables listed under "external explanatory variables" in Table 1 Gini (predicted) is the Gini nowcast, as included in the modelling equation for the income quintiles.

Table 1. Model Variables and 2022 Values

Variable	Definition [2022 value]	Source
Target Variables		
Gini	Gini [44.3]	Distributional Page
IS_q1	PI Share of 0-20% quintile [5.1%]	Distributional Page
IS_q2	PI Share of 20-40% quintile [9.6%]	Distributional Page
IS_q3	PI Share of 40-60% quintile [13.8%]	Distributional Page
IS_q4	PI Share of 60-80% quintile [19.9%]	Distributional Page
IS_q5	PI Share of 80-100% quintile [51.6%]	Distributional Page
Explanatory Variables: NIPA		
Assets	Assets/PI [15.4%]	Table 2.1, line 13/Table 2.1, line 1
Wages	Wages/PI [50.9%]	Table 2.1, line 3/Table 2.1, line 1
Medicare	Medicare/PI [3.7%]	Table 2.1, line 19/Table 2.1, line 1
Medicaid	Medicaid/PI [4.2%]	Table 2.1, line 20/Table 2.1, line 1
Proprietor's Income	Proprietor's Income/PI [8.2%]	Table 2.1, line 9/Table 2.1, line 1
Rental Income	Rental Income/PI [4.0%]	Table 2.1, line 12/Table 2.1, line 1
Unemployment	Unemployment/PI [0.1%]	Table 2.1, line 21/Table 2.1, line 1
Tax Credits	Tax Credits/PI [1.2%]	Table 3.12, line 25/Table 2.1, line 1
Transfers	Transfers/PI [18.5%]	Table 2.1, line 16/Table 2.1, line 1
Explanatory Variables: External		
Case-Shiller Index	Case-Shiller Housing Index [298.5]	FRED:CSUSHPINSA
CPI	CPI Index [292.6]	FRED:CPIAUCSL
Federal Funds Rate	Federal Runds Rate [1.7]	FRED:FEDFUNDS
LFPR	Labor Force Participation Rate [62.2]	FRED:CIVPART
Mortgage Rate	Mortgage Rate [5.3]	FRED:MORTGAGE30US
Prime Rate	Prime Rate [4.9]	FRED:DPRIME
Treasury 5yr	Treasury 5-year rate [3.0]	FRED:DGS5
Unemployment Rate	Unemployment rate [3.6]	FRED:UNRATE
Explanatory Variables: Other		
i2019	Indicator for 2019 [N/A]	N/A
$\tilde{\text{Gini}}_t$	Prediction of Gini [44.0]	Calculated

Note: These variables comprise the choice set for the elastic net models, though not all are used for every model. All NIPA explanatory variables are components of personal income as decomposed in Table 2.1. External variables were obtained via Federal Reserve Economic Data using the series identifier provided and accessed on June 21, 2024.

Table 2. Turning Point Analysis

		% Correct Sign						
		Gini	Q1	Q2	Q3	Q4	Q5	All
2000-2019	Main	95%	95%	100%	95%	90%	95%	95%
	Main + External	95%	95%	100%	100%	95%	100%	98%
	VAR	70%	100%	95%	95%	100%	100%	93%
2000-2020	Main	95%	90%	100%	100%	90%	100%	96%
	Main + External	95%	95%	100%	100%	100%	100%	98%
	VAR	70%	95%	95%	95%	100%	100%	92%
2000-2021	Main	90%	100%	100%	95%	85%	90%	93%
	Main + External	90%	95%	90%	80%	90%	80%	88%
	VAR	70%	95%	65%	60%	65%	65%	70%

Note: This table shows the rate at which the model correctly predicts the sign of the year-over-year change compared with the observed series for each metric. Each year-over-year predicted change is classified as having a "correct" sign if either (1) the model predicts an increase (decrease) in that year corresponding to an observed increase (decrease) in the data of at least 0.1 (i.e. 44.2 to 44.3), (2) the model predicts an increase (decrease) in that year corresponding to no increase (<0.1) in the observed data, or (3) the model predicts no change (<0.1), corresponding to an increase (decrease) in the observed data. The % correct is the number of correct predictions out of the total observations (including the 1-year nowcast for each).

Table 3. Model Performance

	Revision			Mean Rev	Mean Abs Rev	RMSE	
	2020	2021	2022	2020-2022		2000-2021	2000-2022
				Gini, Min: 42.07; Max: 45.21			
Main	0.06	-0.05	0.17	0.06	0.09	0.150	0.176
Main + External	0.49	-0.11	0.19	0.19	0.26	0.150	0.178
VAR	-2.62	0.5	0.8	-0.44	1.31	0.445	0.456
				Q1, Min: 5.04; Max: 5.49			
Main	-0.01	-0.07	-0.23	-0.1	0.1	0.058	0.080
Main + External	-0.05	-0.14	-0.23	-0.14	0.14	0.070	0.079
VAR	0.17	0.12	-0.06	0.08	0.12	0.065	0.060
				Q2, Min: 9.52; Max: 10.16			
Main	0.13	0.04	-0.1	0.02	0.09	0.052	0.070
Main + External	0.07	-0.01	-0.1	-0.01	0.06	0.063	0.067
VAR	0.47	0.17	-0.3	0.11	0.31	0.115	0.126
				Q3, Min: 13.79; Max: 14.43			
Main	-0.01	-0.02	-0.12	-0.05	0.05	0.070	0.094
Main + External	-0.04	0.05	-0.07	-0.02	0.05	0.080	0.084
VAR	0.53	-0.03	-0.44	0.02	0.33	0.110	0.144
				Q4, Min: 19.93; Max: 21.33			
Main	-0.36	0.26	-0.46	-0.19	0.36	0.124	0.159
Main + External	-0.36	0.29	-0.4	-0.16	0.35	0.131	0.145
VAR	-0.2	-0.53	-0.36	-0.36	0.36	0.148	0.132
				Q5, Min: 49.14; Max: 51.46			
Main	0.24	-0.2	0.91	0.32	0.45	0.165	0.291
Main + External	0.37	-0.19	0.8	0.33	0.45	0.184	0.256
VAR	-0.96	0.27	1.18	0.16	0.8	0.253	0.352

Note: The first three columns show the size of the revision (actual—1-year nowcast) for each model and metric. The 2020 nowcast is based on the 2000–2019 model; the 2021 nowcast is based on the 2000–2020 model; and the 2022 nowcast is based on the 2000–2021 model. The following columns are the raw average of the revisions for 2020–2022 and the average of the absolute value of each revision. The final two columns are the root mean squared error (RMSE) calculated for each model for the 2000–2021 and 2000–2022 periods respectively.

Table 4. Modeled Impact of 2023 Comprehensive Update to NIPA

(a) Levels

	January 2023			March 2023			September 2023		
	2020	2021	2022	2020	2021	2022	2020	2021	2022
Gini	42.7	42.4	43.9	42.7	42.4	43.8	42.2	42.5	44.0
Q1	5.3	5.5	5.3	5.3	5.5	5.3	5.4	5.5	5.3
Q2	9.9	10.1	9.8	9.9	10.1	9.8	10.1	10.1	9.8
Q3	14.3	14.4	14.0	14.3	14.4	14.0	14.4	14.4	14.0
Q4	20.4	20.1	20.4	20.4	20.1	20.4	20.5	20.0	20.4
Q5	50.0	49.9	50.5	50.0	49.9	50.4	49.7	50.0	50.5

(b) Year-Over-Year

	2020 - 2021			2021 - 2022		
	January	March	September	January	March	September
Gini	-0.3	-0.3	0.2	1.5	1.4	1.5
Q1	0.1	0.1	0.1	-0.2	-0.2	-0.2
Q2	0.2	0.2	0.1	-0.3	-0.3	-0.3
Q3	0.1	0.1	0.0	-0.4	-0.4	-0.4
Q4	-0.4	-0.4	-0.5	0.3	0.3	0.3
Q5	-0.1	-0.1	0.2	0.5	0.5	0.6

Note: This table shows the results of an exercise where the coefficients on the main specification for 2000–2020 are multiplied by the NIPA values for 2000–2022 for the first advance estimate of Q4 (January 2023, pre comprehensive update), third estimate of Q4 (March 2023, pre comprehensive update) and September 2023 (post comprehensive update). The 2023 comprehensive update significantly changed inequality levels and trends (also see: Kornfeld (2023) and Mataloni et al. (2023)). Panel A shows the levels and panel B shows the differences year-over-year. These are estimates corresponding to a specific model for example purposes, and are not final estimates for these years and metrics.