AEA Papers and Proceedings 2025, 115: 79–84
https://doi.org/10.1257/pandp.20251105

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

By Gary Cornwall and Marina Gindelsky*

Income inequality is typically measured using detailed microdata, which are available with a lag. The US Bureau of Economic Analysis (BEA) publishes a preliminary distribution of personal income (PI) with a one-year lag using partial data, referred to as a provisional estimate, followed by official estimates a year later using more-complete data sources. However, recent economic turbulence has highlighted the need for policymakers and data users to obtain more timely estimates. In this study, we apply a machine-learning method to nowcast distributional estimates, by leveraging macro totals for PI components such as wages, dividends, and transfers, as part of the National Income and Product Accounts (NIPA) (for a detailed previous draft, see Cornwall and Gindelsky 2024).[1]

Presently, BEA uses microdata from the Annual Social and Economic Supplement of the Current Population Survey (henceforth, "CPS"), augmented with other survey and administrative sources—scaled to decomposed PI components—to construct distributions of PI (and disposable PI = DPI) that sum to NIPA totals (see detailed methodology in Gindelsky 2024). The resulting quintile series are produced at a granular level and fully updated every year,[2] reflecting NIPA revisions and methodological updates.[3] The purpose of the BEA exercise is to distribute macro totals to micro households in order to connect aggregate growth with household experience. Accordingly, distributional results reflect both relationships between aggregate income sources and changes in the population and composition of income. We use the relationships between these macro components as inputs in an elastic net model, a type of penalized regression that combines both ridge and lasso methods, to generate nowcasts of the equivalized Gini coefficient and income shares at the quintile level.

Given that the objective of nowcasting is typically to provide more-timely information during volatile economic periods, our approach prioritizes accurate prediction of turning points and trends instead of minimizing error during stable periods. Thus, our focus is on performance during the COVID-19 period (2000–2022: initial shock, response, and recovery), iteratively treating each year as "out of sample" to mimic real-world nowcasting conditions. We correctly predict at least 90 percent of turning points across all models and time periods, and 100 percent for the COVID years, with a mean revision of at most 0.2 percentage points across all measures and years.

In addition to an improvement in timeliness, our approach has three key advantages. First, this method stands in contrast to traditional nowcasting approaches reliant on microsimulation, which require complex models and significant resources. Second, we do not need to obtain current-period microdata—survey, administrative, or private sector—further reducing costs. Finally, this method is generalizable for those seeking to construct timely distributional national accounts internationally.

*Cornwall: Bureau of Economic Analysis (email: gary.cornwall@bea.gov); Gindelsky: Bureau of Economic Analysis (email: marina.gindelsky@bea.gov). The authors would like to thank David Johnson, participants at the AEA, GW H. O. Stekler Forecasting Program, OECD EGDNA, Federal Forecasters Conference, and BEA Seminar Series for their helpful comments and suggestions. This paper summarizes and updates Cornwall and Gindelsky (2024) (BEA Working Paper 2024-06). The views expressed in this paper are those of the authors and do not necessarily represent the US Bureau of Economic Analysis or the US Department of Commerce.

†Go to https://doi.org/10.1257/pandp.20251105 to visit the article page for additional materials and author disclosure statement(s).

[1] Here, "nowcasting" refers to providing the distributional estimates alongside macro totals that reflect activity in the preceding period, consistent with BEA procedure (i.e., estimates for calendar year 2023 in the first quarter of 2024).

[2] See 2000–2023 estimates (https://www.bea.gov/data/special-topics/distribution-of-personal-income).

[3] All data used for this exercise are publicly available on the BEA website. They are described in the Supplemental Appendix.

Since the approach only uses national accounts totals, researchers in other countries can identify key components of their income concepts and apply them accordingly to produce nowcasts.

I. Methods

There have been limited attempts to forecast income inequality in the United States given the combined difficulty of predicting shocks, responses, and their impacts on the distribution.[4] Traditional time series econometric approaches such as vector autoregressions (VARs) are generally unsuccessful during shocks because of the differential impact of economic shocks on the income distribution. A recent attempt at nowcasting distributional national accounts in the United States by Blanchet, Saez, and Zucman (2022) performed fairly well during stable periods and less well during turbulent times. Internationally, attempts to predict inequality are therefore focused on nowcasting using microsimulation, where contemporary macro information can be used to guide models (see Levy 2023 for a review). However, these estimates are costly to construct (an average of 24 days in a survey by O'Donoghue and Loughrey 2014) and often rely on established models to impute changes in labor markets and transfers. Therefore, a new strategy to create timely inequality estimates is needed.

With prediction as the primary goal rather than statistical inference, penalized regression allows us to better leverage contemporaneous information by forcing coefficients toward zero in a principled manner. Exploiting the bias-variance trade-off, the elastic net (Zou and Hastie 2005) is a penalized regression framework nesting lasso and ridge as special cases. Given a response vector $\mathbf{Y} = \left(y_1, \ldots, y_N\right)'$, indexed by $i = \{1, \ldots, N\}$ and information set $\mathbf{X} = \left[\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_K'\right]$, the elastic net estimator can be expressed as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2,$$

subject to

$$(1 - \alpha)\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|_2 \leq \delta \text{ for some } \delta,$$

with

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

and $\|\cdot\|_1$ and $\|\cdot\|_2$ representing ℓ^1 and ℓ^2 norms, respectively. The two penalties, λ_1 and λ_2, compete; the latter promotes model expansion, while the former shrinks coefficients to a true zero and thus leans toward sparsity. Overall, the elastic net penalty is a convex combination of the ridge and lasso penalties, with $\alpha = 1$ corresponding to a ridge regression and $\alpha = 0$ a lasso.

In a follow-up, Friedman, Hastie, and Tibshirani (2010) revisited models with convex penalty structures. Of particular interest to this work, the regularization of parameters when the response vector is a $N \times P$ matrix rather than a $N \times 1$ vector. This extension, with some notational liberty, can be expressed as

$$\min_{(\beta_0, \boldsymbol{\beta}_K) \in \mathbb{R}^{P+1}} \left\{ -\mathcal{L}_j\left(\beta_{0,j}, \boldsymbol{\beta}_{K,j}\right) + \lambda P_\alpha\left(\boldsymbol{\beta}_j\right) \right\},$$

[4] See Gindelsky (2018) for an exercise using US data and a recent attempt by Castle, Doornik, and Hendry (2024) using dynamic linear regressions with automated variable selection in OxMetrics.

where $P_\alpha(\cdot)$ is the penalty term and $\mathcal{L}_j$ is the log-likelihood of a response vector. The result is a penalized regression that can fit multiple response vectors at the same time using a fixed set of regressors for each vector with different coefficient values.[5]

Using the elastic net, we estimate the following functions, $f(\cdot)$ and $g_j(\cdot)$, to produce nowcasts of the relevant inequality measures:

$$\text{Gini}_t = f\big(\mathbf{x}, \mathbf{1}\{2019\}\big), \tag{1}$$

$$is_{t,j} = g_j\Big(is_{t-i,j}, \mathbf{x}, \mathbf{1}\{2019\}, \widetilde{\text{Gini}_t}\Big), \tag{2}$$

where $is_{t,j}$ refers to the income share of the jth quintile in period t, $\mathbf{x}$ is a $T \times K$ matrix of information obtained from NIPA and $i \in \{1,2\}$.[6] Given their prominence in the distributional composition of income, coefficients on the four main components of PI—assets, wages, proprietor's income, and tax credits—are left unpenalized in the estimating equation.

II. Results

The results of our main specification are presented in Figure 1 for the Gini and the quintiles as follows: purple ovals for 2000–2019 (nowcast = 2020), red triangles for 2000–2020 (nowcast = 2021), blue circles for 2000–2021 (nowcast = 2022), and yellow diamonds for 2000–2022 (nowcast = 2023). Both the model fit and out-of-sample predictions are very accurate, particularly for the Gini coefficient. This accuracy is exploited by using the predicted Gini as an input to the quintile models; the predicted Gini represents 40–70 percent of the predicted estimate for each quintile, with the next most important contributors being labor income and income share lags (see the Supplemental Appendix for a detailed decomposition of features).

In addition to a visual inspection of the data, we can quantify model performance by comparing this specification with a VAR(2) in Table 1. The nowcast improves on the VAR(2) as measured by the root mean square error for every model, with a minimum improvement of 47 percent per nowcast (average across quintiles). Table 1 also presents results for the turning point analysis, conducted at the reporting level (i.e., three-digit shares and Gini). Each annual-change comparison (predicted versus actual) is considered to have a "correct" sign if (i) the direction of the change is the same for both or (ii) the prediction (actual) is "no change," while the actual (prediction) is a positive or negative change.[7]

We can also conduct a revision analysis, similar to that of GDP itself, to understand how new data would impact the distributional estimates. Table 1 provides the mean annual revision (and mean absolute revision) for the nowcast years, which are all less than 0.2. While forecasts are typically evaluated on such "average" metrics of accuracy, we also note that there could be some over/underprediction for adjacent quintiles for an individual year. The most stark example is in the 2020 forecast, wherein the revision is +0.6 percentage points for Q5, and −0.5 for Q4. Nevertheless, we note the exceptional accuracy of the model in capturing the direction of change and trend at the peak of the pandemic without any available microdata or prior (comparable) shock in the training data, results that would surely have been useful to have at the time.

[5] While originally designed for use in a cross section, the elastic net and other penalized regression frameworks have been explored in a time series setting with promising theoretical and empirical results. See Masini, Medeiros, and Mendes (2023) for an informed discussion. Due to the limited length of the published time series, we use the basic elastic net framework, though future work may incorporate procedures more suitable to autoregressive processes.

[6] The income share estimation includes the predicted $\widetilde{\text{Gini}}$, as in the previous equation to improve fit, as explained in the next section. Given that inequality is mean reverting (i.e., deviations in one year correct the following year), at most two lags were included. We have also added an indicator for 2019 to reflect the distributional anomaly stemming from a definitional mismatch between the CPS and PI. See Cornwall and Gindelsky (2024) for more details.

[7] This is a fairly conservative analysis, which relies primarily on numerical accuracy rather than on economic significance; it could be easily argued that a Gini of 44.2 is not qualitatively different from 44.1 or 44.3, yet this would count as a turning point error in our table if the predicted annual change was 44.2 to 44.1, and the actual was 44.2 to 44.3.

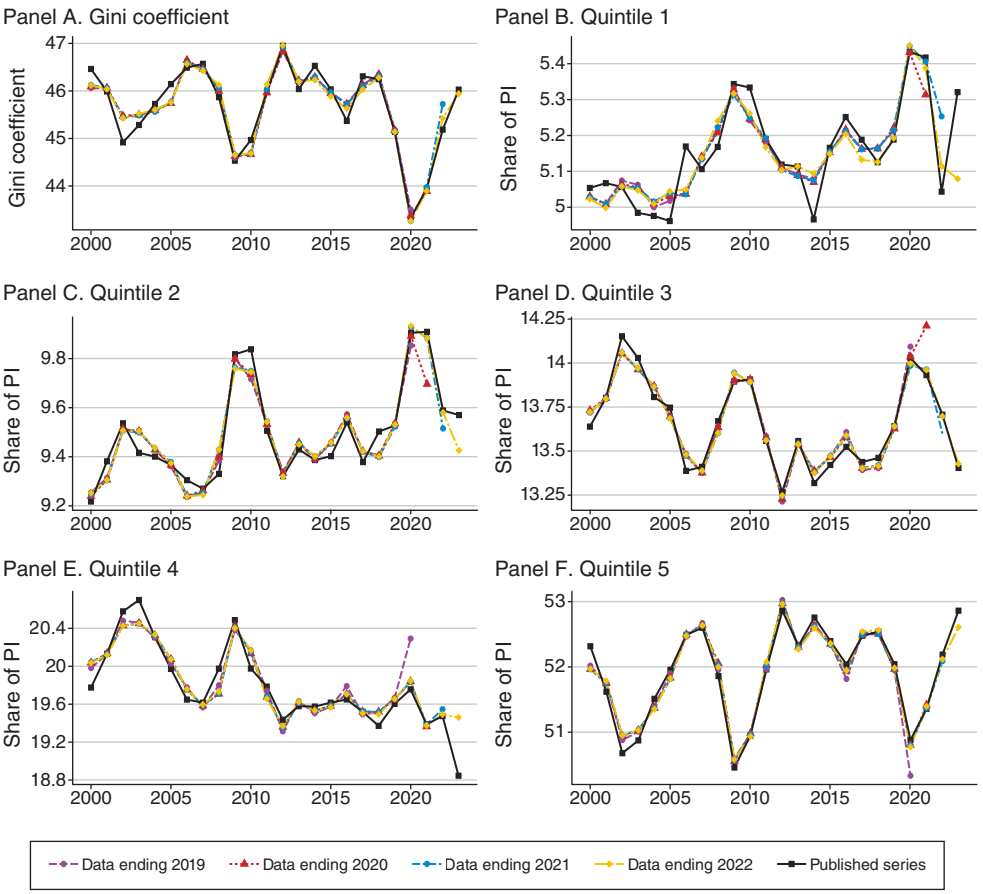

FIGURE 1. MODEL FIT AND NOWCAST PERFORMANCE

Notes: This figure shows four models estimated for each metric with one-year nowcasts from observed series for 2000–2019 (purple lines), 2000–2020 (red), 2000–2021 (blue), and 2000–2022 (yellow). The black line denotes the observed series of metrics for PI, as published by BEA in December 2024.

III. Conclusion

Our successful construction of a nowcast for the BEA distributional national accounts demonstrates that we can produce highly accurate inequality measures overall (Gini coefficient) and at the quintile level shortly after the calendar year without contemporary microdata. The previous version of this analysis (Cornwall and Gindelsky 2024) added external series tied to business cycle volatility such as inflation, unemployment, and mortgage rates, but found that they did not improve model performance and often detracted from it.

As there is no official microsimulation model for the US (though some academic and agency models exist for various components of income or GDP), we cannot compare our results to such approaches. However, the stand-alone performance of these models is impressive, with both average revisions under 0.2 percentage points and at least 90 percent correct sign prediction (100 percent during pandemic years). It is robust to changes in NIPA totals, including annual revisions and larger comprehensive updates (see Cornwall and Gindelsky 2024 for an analysis of the impact of the latest comprehensive update). Although this approach is inappropriate for policy analysis, as we cannot

Table 1—Nowcast Performance

	Q1	Q2	Q3	Q4	Q5	Gini
RMSE improvement over VAR(2)						
Data ending 2019	49%	64%	64%	2%	67%	67%
Data ending 2020	37%	42%	44%	40%	77%	66%
Data ending 2021	27%	54%	49%	42%	73%	57%
Data ending 2022	24%	48%	60%	32%	74%	61%
COVID (2020–2022)	41%	51%	44%	34%	61%	39%
Correct sign						
Data ending 2019	95%	100%	100%	100%	100%	90%
Data ending 2020	95%	100%	95%	100%	100%	90%
Data ending 2021	95%	100%	100%	100%	100%	91%
Data ending 2022	96%	100%	100%	100%	100%	91%
COVID (2020–2022)	100%	100%	100%	100%	100%	100%
Nowcast revisions (*percentage points*)						
Mean revision	0.0	0.1	0.0	−0.2	0.2	−0.2
Mean absolute revision	0.2	0.2	0.1	0.1	0.2	0.2

Notes: The top panel of the table provides the root mean square error (RMSE) "improvement," calculated as follows: (RMSE of Main − RMSE of VAR(2))/RMSE of VAR(2). The second panel gives the portion of observations for which the main specification predicts the "correct sign." A sign is considered correct if (i) the direction of the change is the same for both, or (ii) the prediction (actual) is "no change," while the actual (prediction) is a positive or negative change. The final panel calculates the mean of the revisions (actual − predicted) and the mean of the absolute value of the revisions, across all nowcast observations for each measure.

identify causal impacts of contributing variables, the timely and generalizable nature makes it applicable to other datasets and time periods, and a complement to microsimulation-based nowcasting techniques.

The application of our approach will enable us to produce inequality series one month after the end of the calendar year, eight months prior to the availability of the CPS, from which the first official inequality numbers for the United States derive, providing policymakers and data users with a significant improvement in timeliness from a more comprehensive income measure. By adding this "advance" estimate of the distributional accounts, we would be following a similar structure to other NIPA releases with second (the current "preliminary") and third (with complete data) published estimates. Moreover, this method is parsimonious and can be calculated in a number of minutes, as compared with more costly nowcasting approaches in Europe deriving from microsimulations, significantly reducing resource costs. Finally, this approach is generalizable for other countries and datasets. By determining which components of national accounts are driving distributional changes, we hope that other countries can also produce successful nowcasts with their datasets well ahead of available microdata.

REFERENCES

Blanchet, Thomas, Emmanuel Saez, and Gabriel Zucman. 2022. "Real-Time Inequality." NBER Working Paper 30229.

Castle, Jennifer L., Jurgen A. Doornik, and David F. Hendry. 2024. "Forecasting the UK Top 1% Income Share in a Shifting World." *Economica* 91 (363): 1047–74.

Cornwall, Gary, and Marina Gindelsky. 2024. "Nowcasting Distributional National Accounts for the United States: A Machine Learning Approach." BEA Working Paper 2024-6.

Friedman, Jerome H., Trevor Hastie, and Rob Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22.

Gindelsky, Marina. 2018. "Modeling and Forecasting Income Inequality in the United States." BEA Working Paper 2018-07.

Gindelsky, Marina. 2024. *Technical Document: A Methodology for Distributing Personal Income*. Bureau of Economic Analysis.

Levy, Horacio. 2023. "Nowcasting and Provisional Estimates of Income Inequality Using Microsimulation Techniques." OECD Papers on Well-Being and Inequalities Working Paper 12.

Masini, Ricardo P., Marcelo C. Medeiros, and Eduardo F. Mendes. 2023. "Machine Learning Advances for Time Series Forecasting." *Journal of Economic Surveys* 37 (1): 76–111.

O'Donoghue, Cathal, and Jason Loughrey. 2014. "Nowcasting in Microsimulation Models: A Methodological Survey." *Journal of Artificial Societies and Social Simulation* 17 (4): 12.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67 (2): 301–20.