Household Income Prediction in Real Time: An analysis with Peruvian Data *

Franco Calle
University of Chicago
Booth School of Business

Angelo Cozzubo
The University of Chicago
NORC

Hernan Winkler†
The World Bank

February 25, 2024

Abstract

This paper presents an innovative approach to predicting household income one year in advance by integrating traditional household survey data with big data from administrative sources, weather data, and nightlight satellite imagery. Utilizing state-of-the-art machine learning methods, we aim to overcome the limitations of conventional income prediction models that often rely solely on survey data. Our methodology combines granular, real-time data from diverse sources, providing a more comprehensive and dynamic understanding of the socioeconomic factors influencing household income. By incorporating weather patterns and nightlight data, we capture environmental and economic activity indicators that are crucial for accurate forecasting. This analysis not only enhances the precision of income predictions but also offers valuable insights for policy-making and economic planning. The outcome of this research is expected to be instrumental for governments and organizations in implementing targeted interventions.

*This is a footnote.

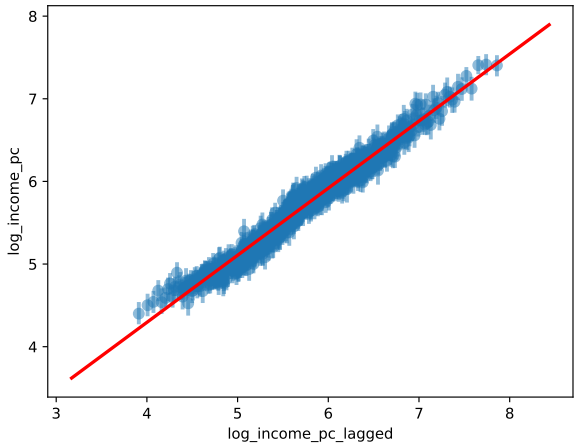
†The authors would like to thank the World Bank for financial support and ...

Email: angelo.cozzubo@pucp.edu.pe, francocalle@chicagobooth.edu, hwinkler@worldbank.org.

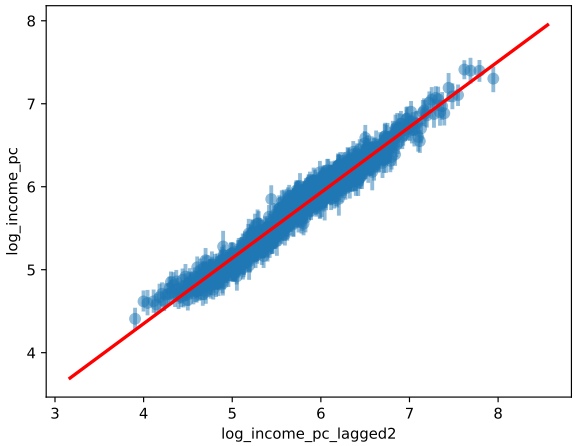
1 Introduction

2 Data and Summary Statistics:

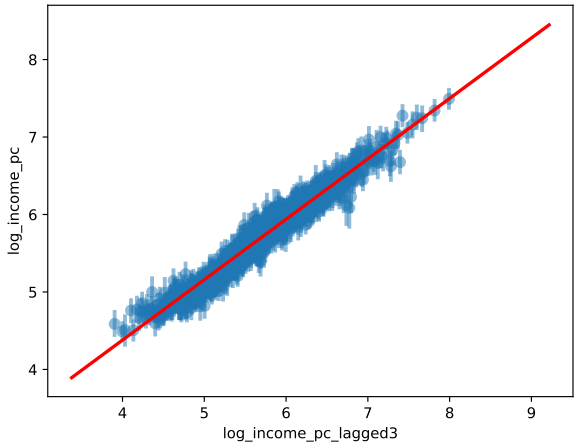
Figure 1: Correlation of log(Income) with Lagged Conglomerate Log Income



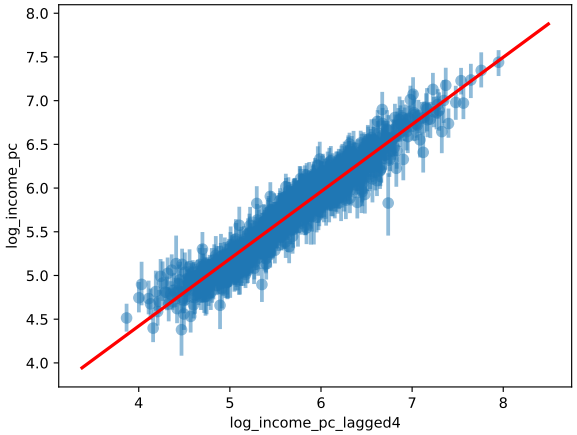
(a) Lag: t-1



(b) Lag: t-2



(c) Lag: t-3



(d) Lag: t-4

Note:

Figure 2: Distribution for each variable

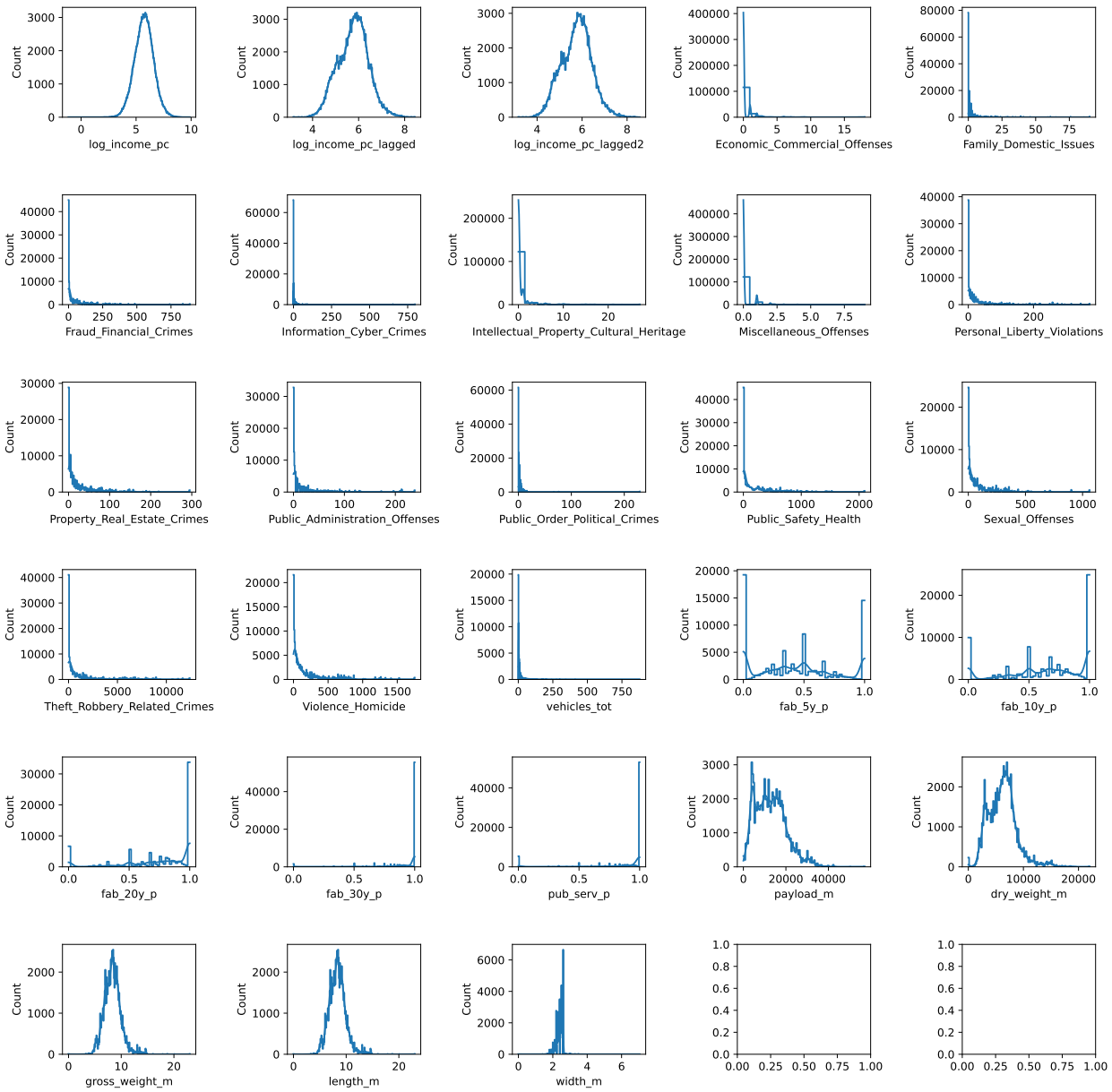


Figure 3: Correlation of log(Income) with variables

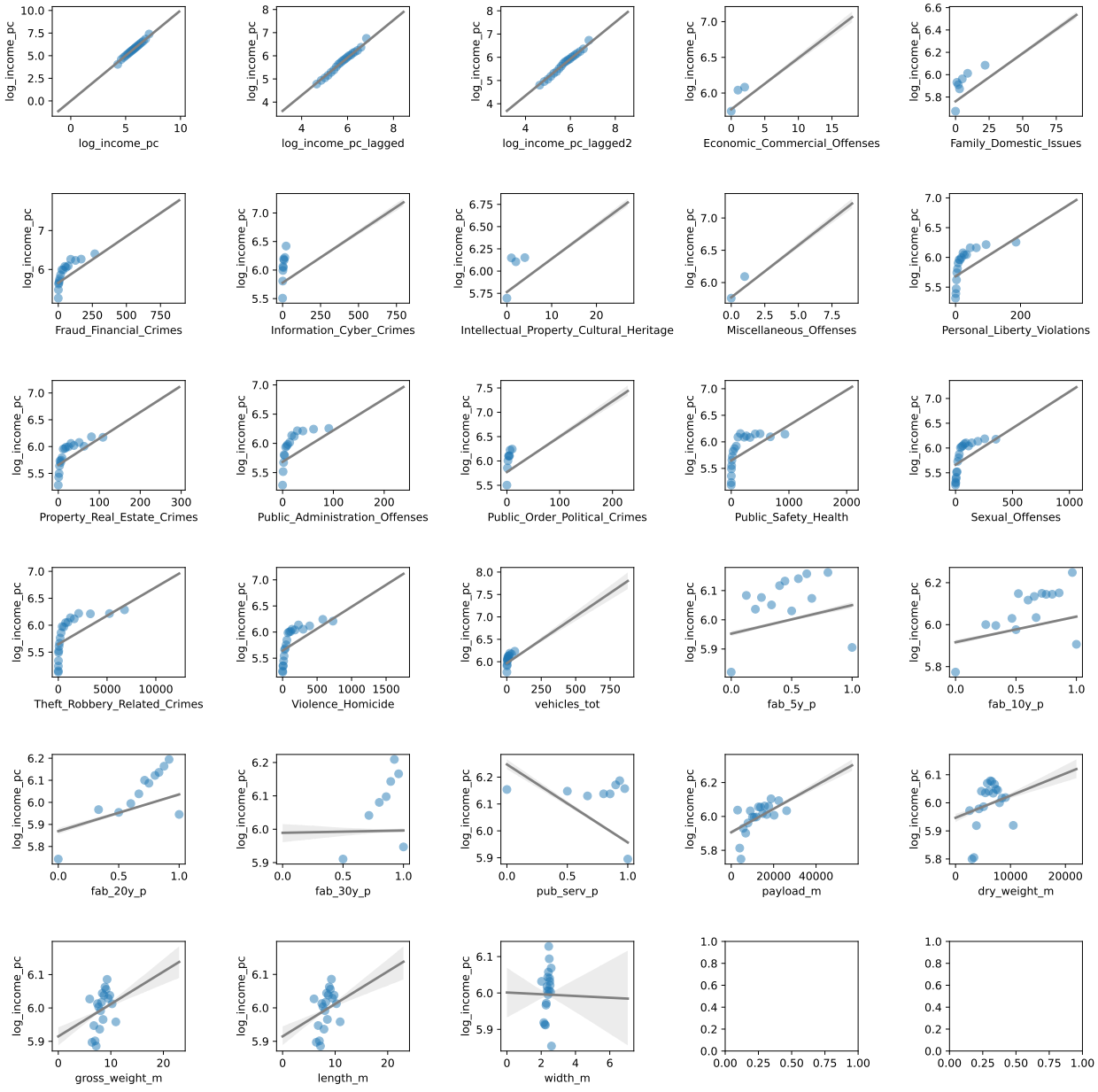


Figure 4: Correlation of log(Income) with Weather Variables

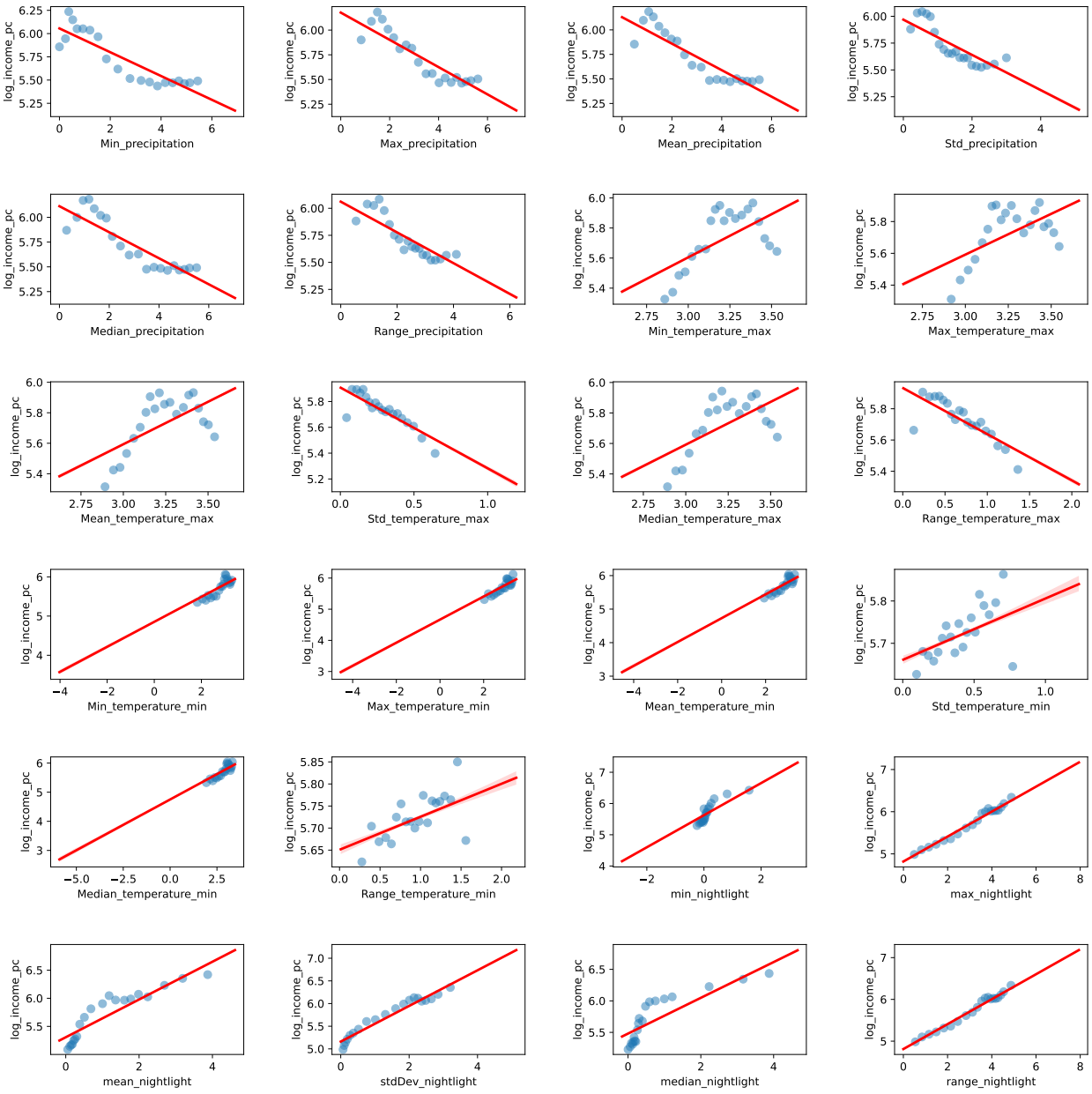


Figure 5: Share of Missing Values

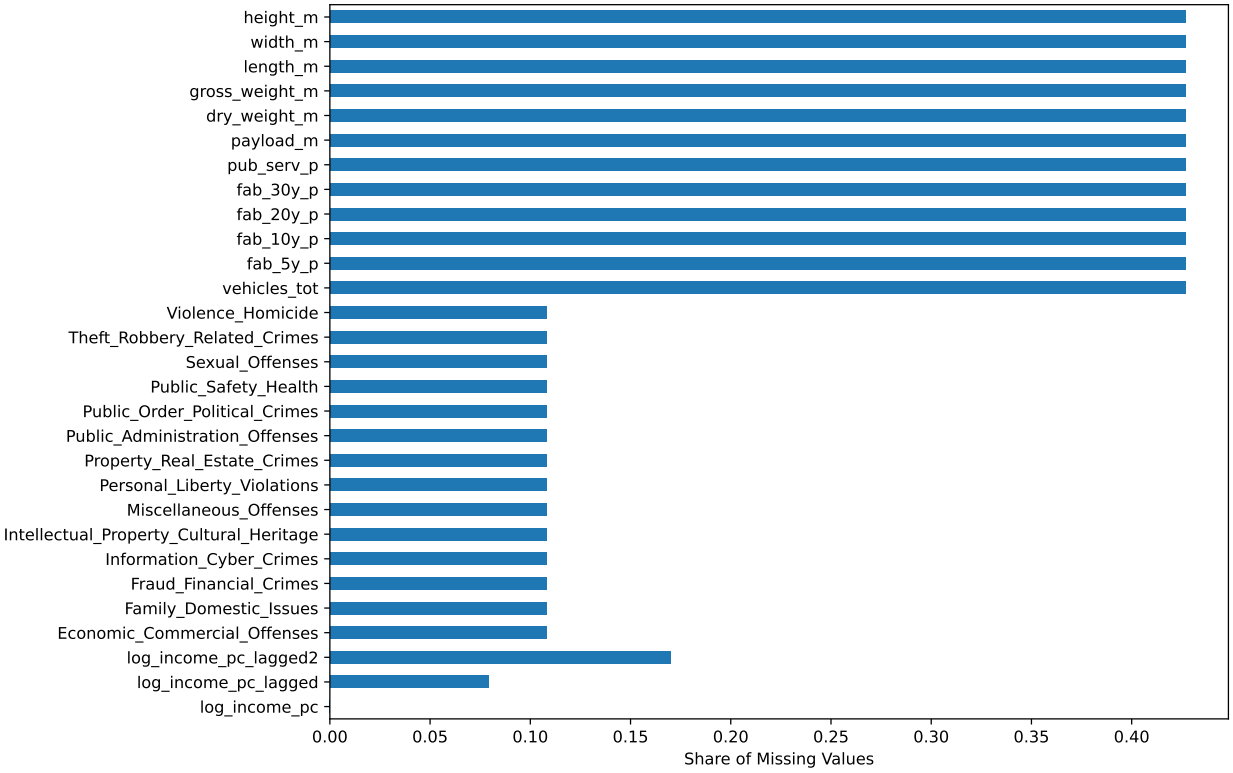


Table 1: Missing lags at CCPP level

year	Lag 1	Lag 2	Lag 3	Lag 4
2008	1	1	1	1
2009	0.003	1	1	1
2010	0.004	0.007	1	1
2011	0.151	0.155	0.157	1
2012	0.006	0.156	0.159	0.162
2013	0.445	0.448	0.535	0.537
2014	0.288	0.724	0.725	0.77
2015	0.191	0.457	0.868	0.868
2016	0.182	0.352	0.588	0.955
2017	0.051	0.195	0.371	0.617
2018	0.074	0.121	0.254	0.418
2019	0.011	0.017	0.067	0.209
2020	0.319	0.327	0.331	0.375
Total	0.199	0.357	0.510	0.659

3 Empirical Evidence

4 Estimation:

Our preferred specification is a model of the following form.

$$y_{h,t} = a_c + bt + \sum_{m=1}^{12} \mathbb{1}\{\text{M}_{h,t} = m\}\gamma_m + \sum_{s=1}^{S} \phi_s \bar{y}_{c,t-s} + \sum_{k}^{K} \sum_{r}^{R} \mathbb{1}\{\text{Reg}_{h,t} = r\} \cdot \beta_r^{[k]} X_{h,t}^{[k]} + \epsilon_{h,t}$$

Where in the left-hand side $y_{h,t}$ is the logarithm of income per-capita of household h at time t. In the right-hand side the explanatory variables are α_c which is a fixed effect for the conglomerate c where the household h belongs. The parameter b corresponds to the trend changes in income and t is just the period of analysis. Parameters γ_m are fixed effect for the month at which the survey was collected for household h that controls for the seasonality of income. In addition to the fixed effects we control for the lagged value of log income per capita of the conglomerate which is denoted as $\bar{y}_{c,t-s}$ and ϕ_s is the parameter associated with lag $s = \{1, ..., S\}$. We also include interaction terms between the region $Reg_{h,t}$ where the household lives and $k = \{1, ..., K\}$ characteristics $X_{h,t}^{[k]}$ from the household that come from administrative data and satelite data. Finally, there is an error term that we do not observe $\epsilon_{h,t}$

5 Results

Figure 6: Variables Contribution to the Model (LASSO)

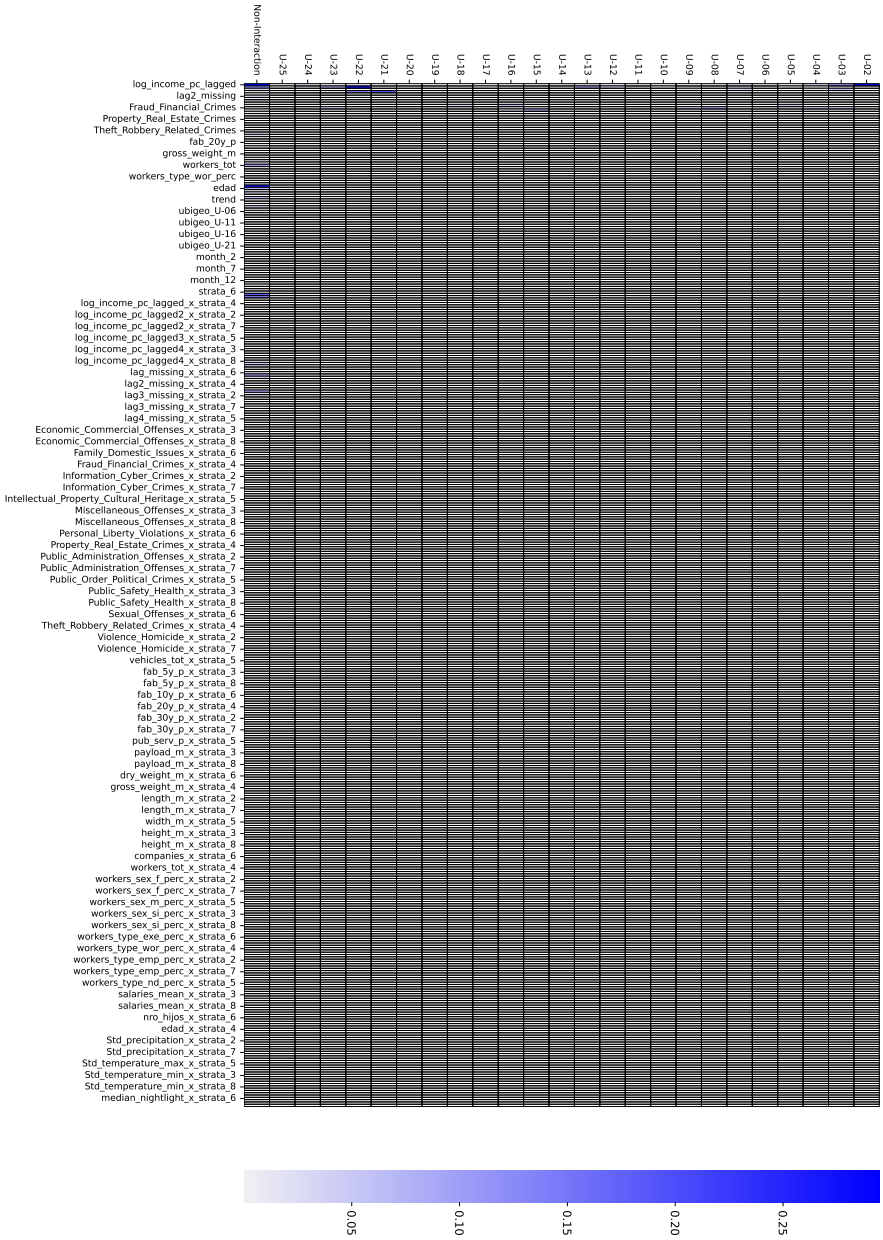


Figure 7: Standard Deviation of Predicted Errors

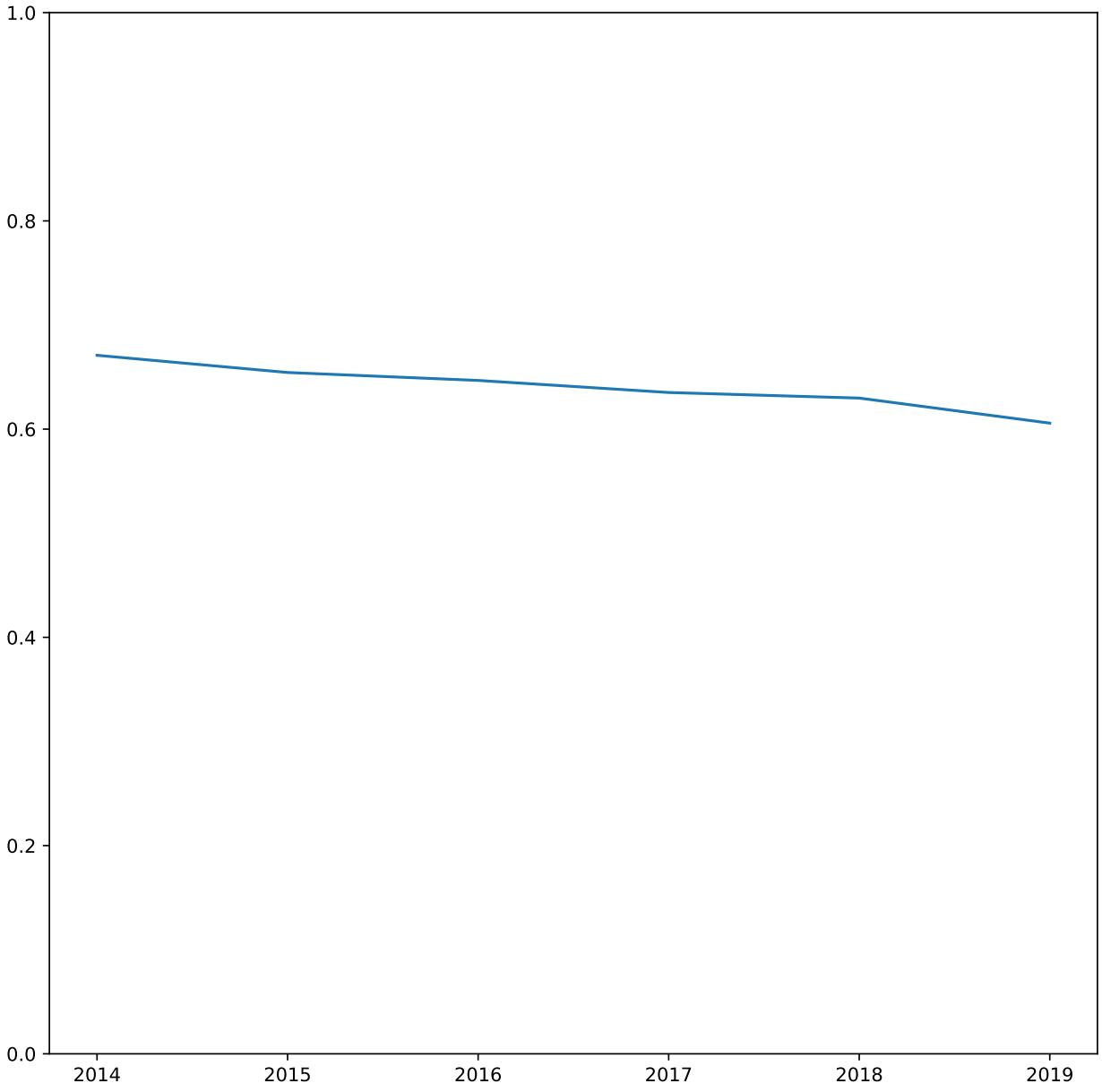


Figure 8: Correlation: Predicted vs True Per-capita Household Income

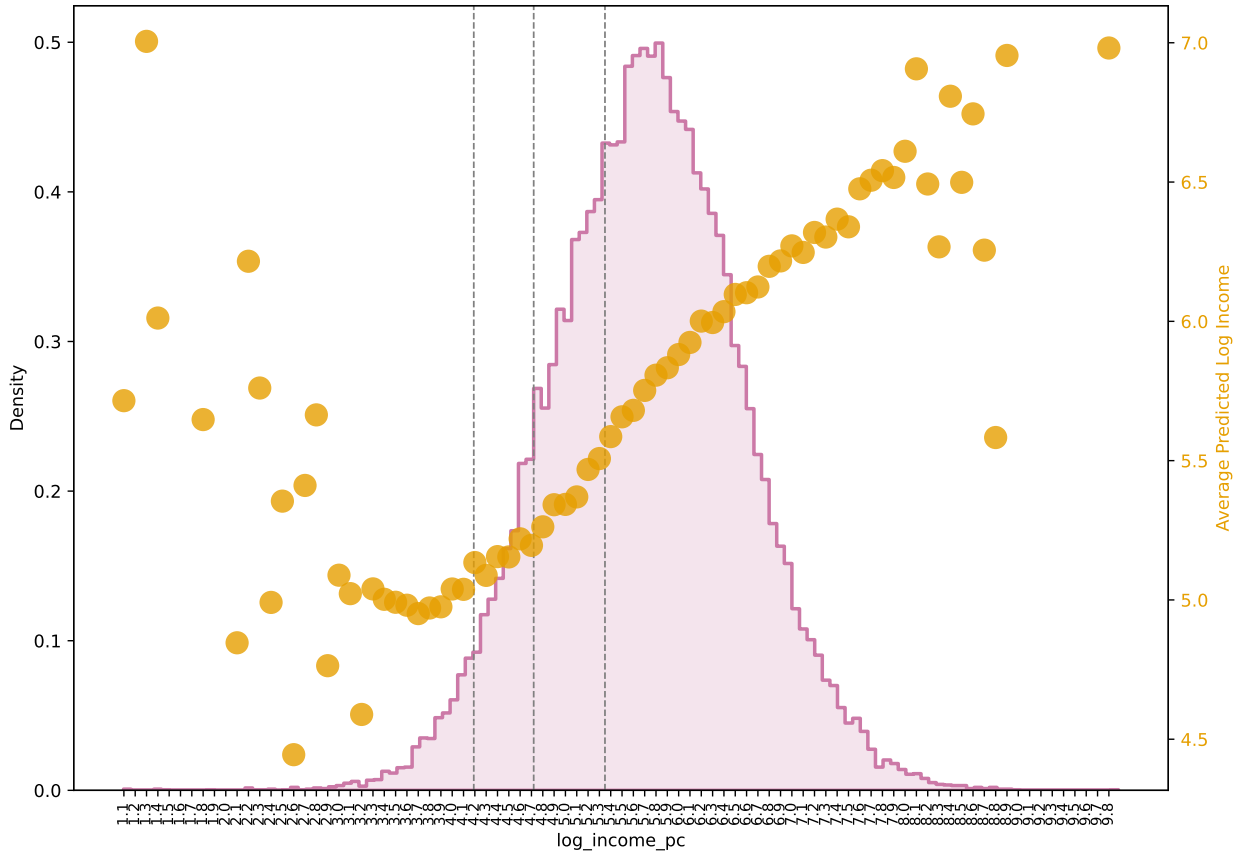


Figure 9: Out of Sample Prediction vs True Income Distribution (Standardized) 2019

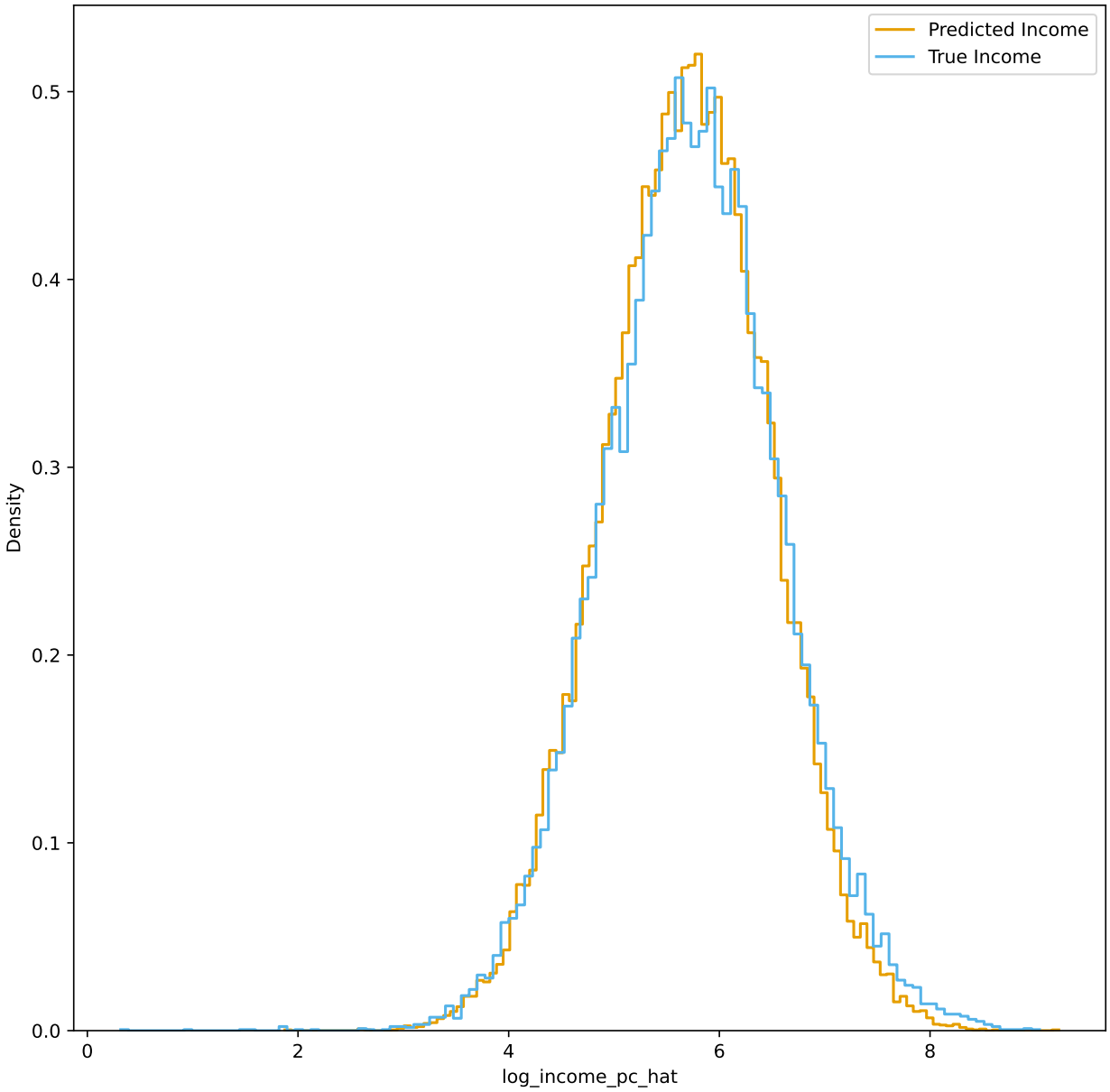


Figure 10: Out of Sample Prediction vs True Income Distribution 2019

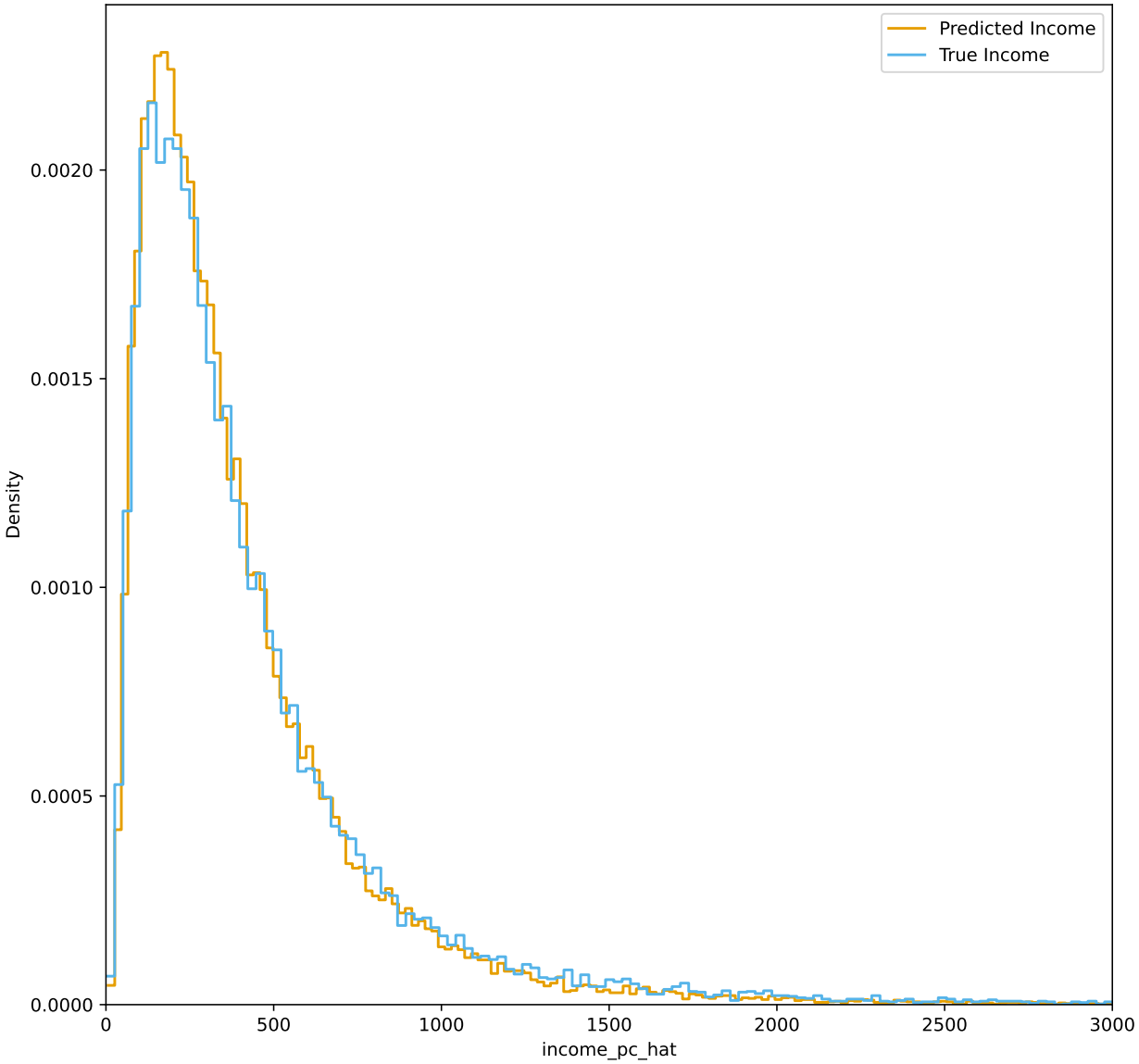


Figure 11: Prediction vs True Per-capita Household Income ECDF 2019

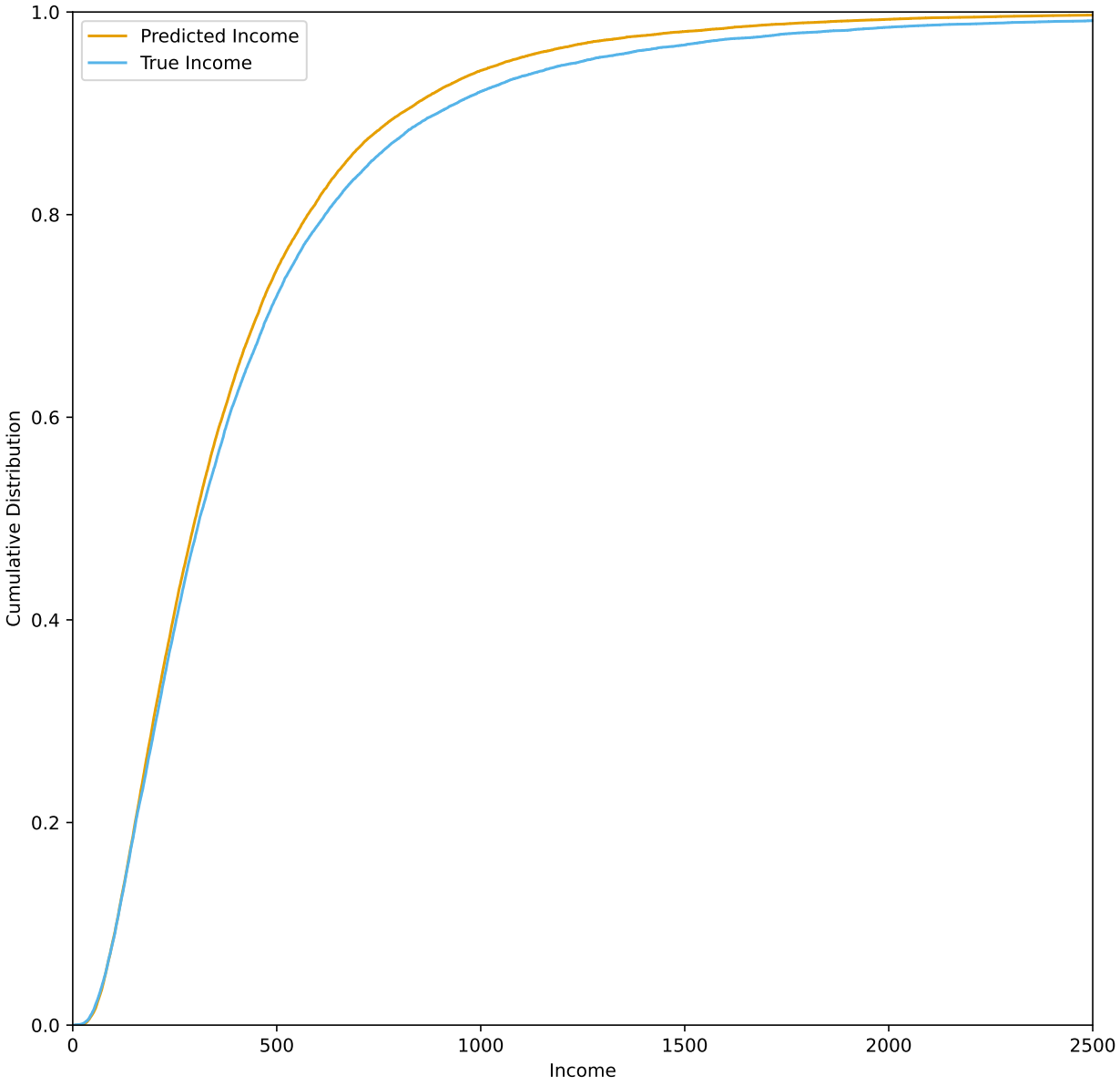


Figure 12: Per-capita Household Income Distribution by Region 2019

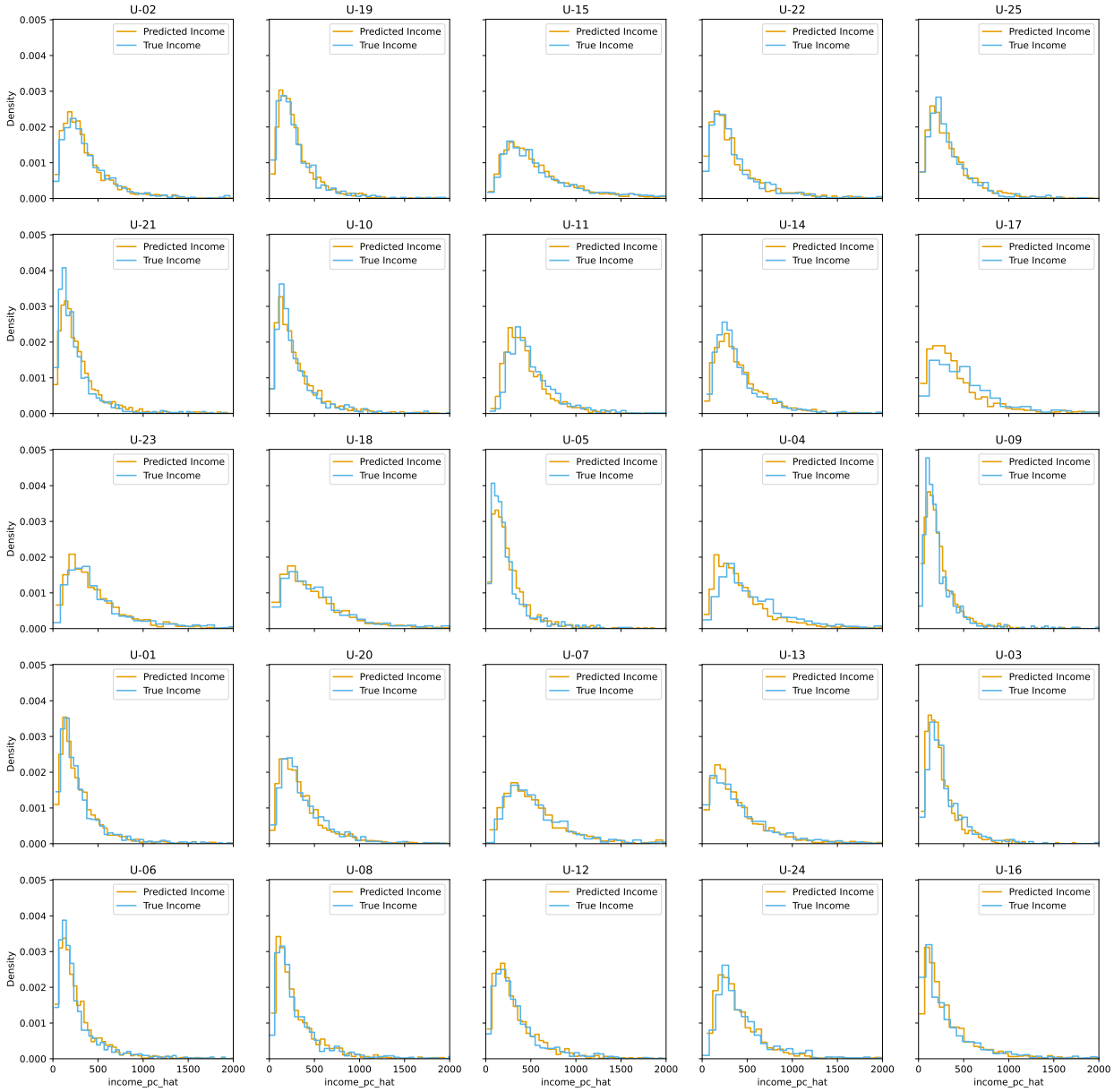


Figure 13: Poverty Rate Predicted vs True National

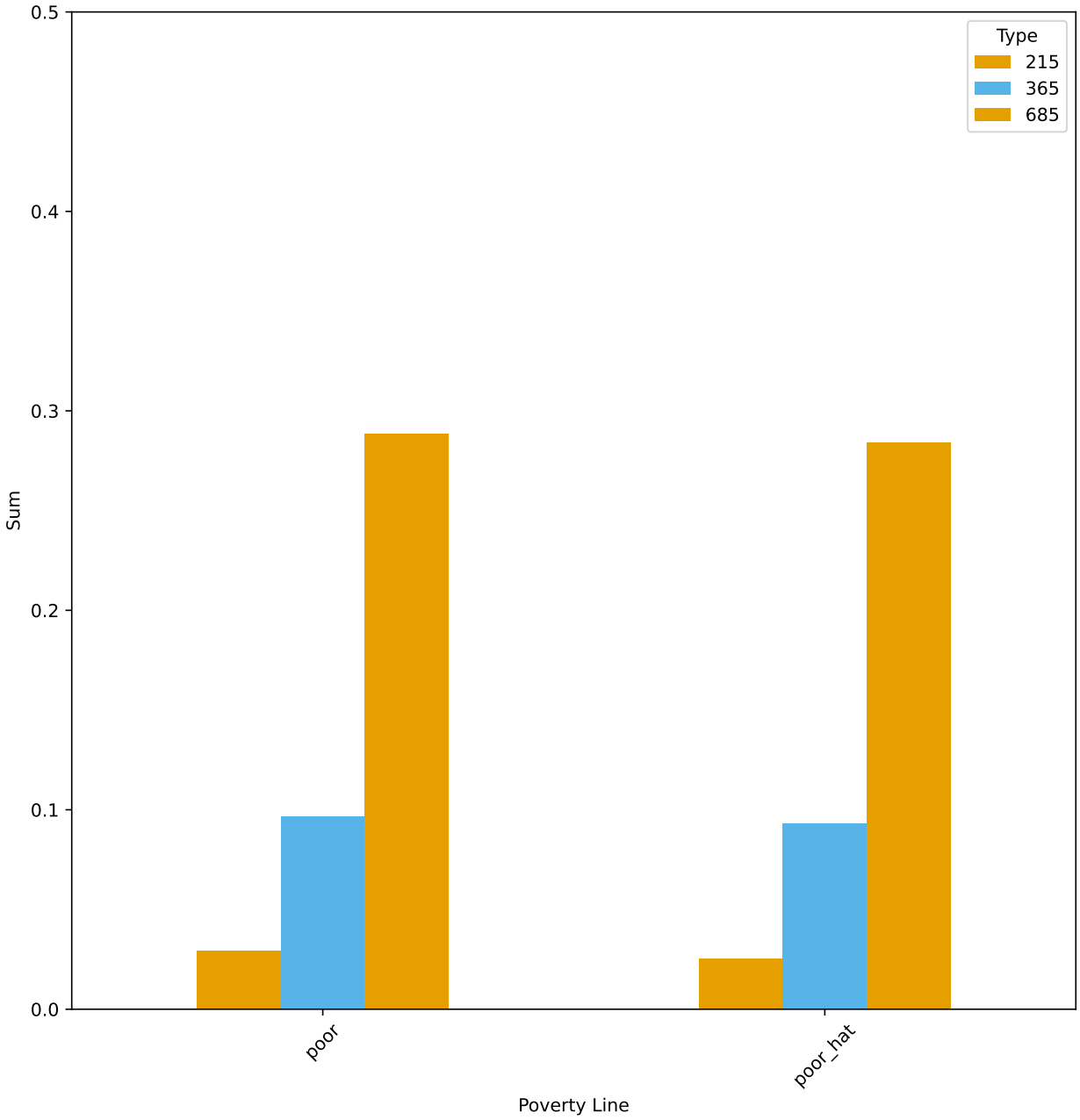
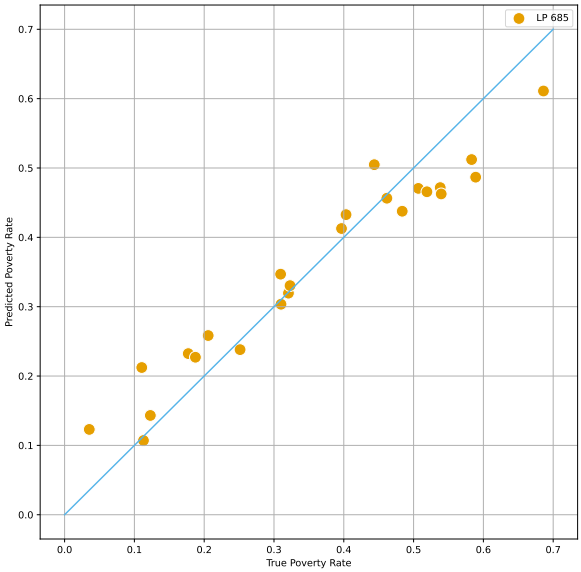
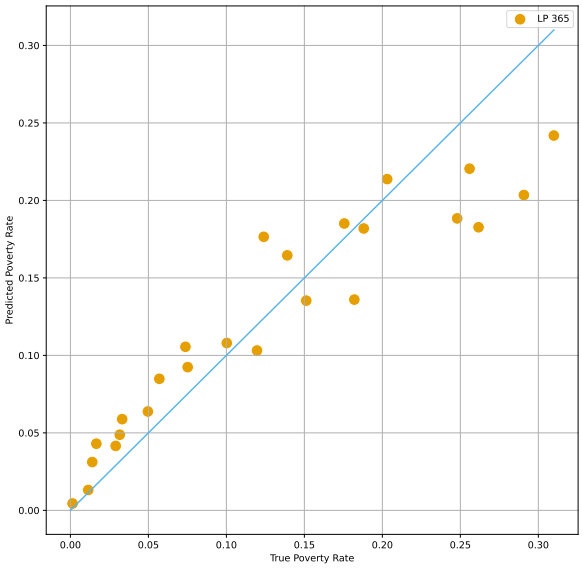


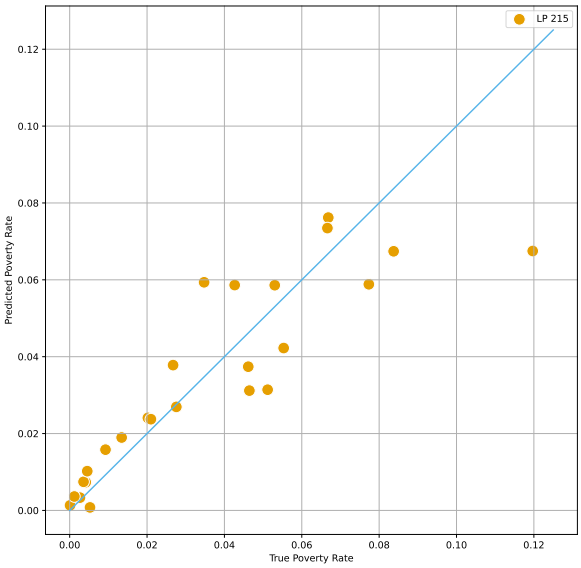
Figure 14: Correlation: Predicted Poverty against true by Region



(a) Poverty Line: 685



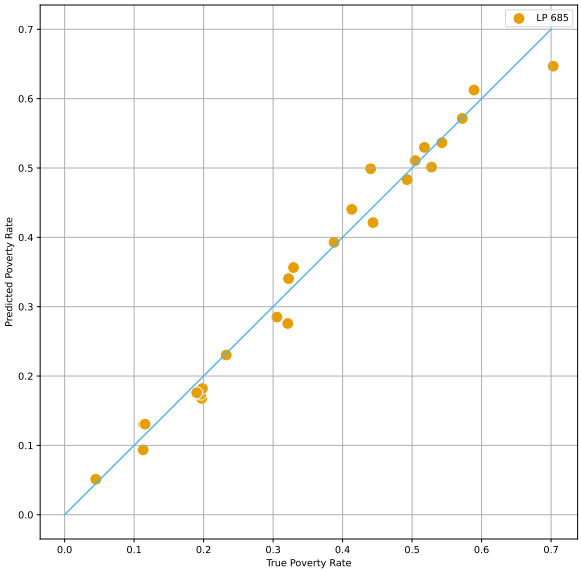
(b) Poverty Line: 365



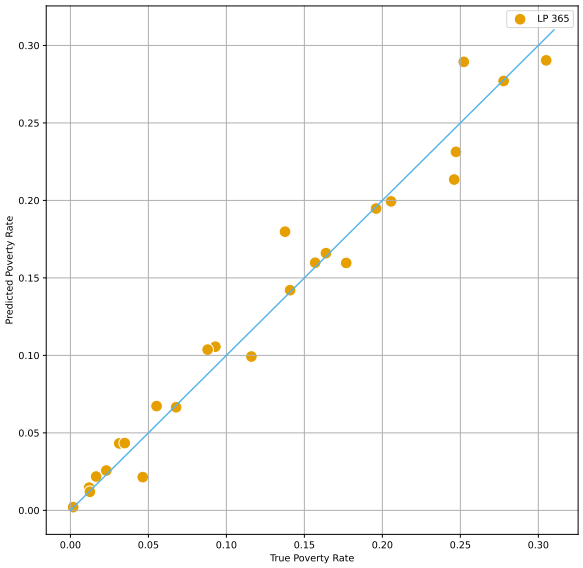
(c) Poverty Line: 215

Note:

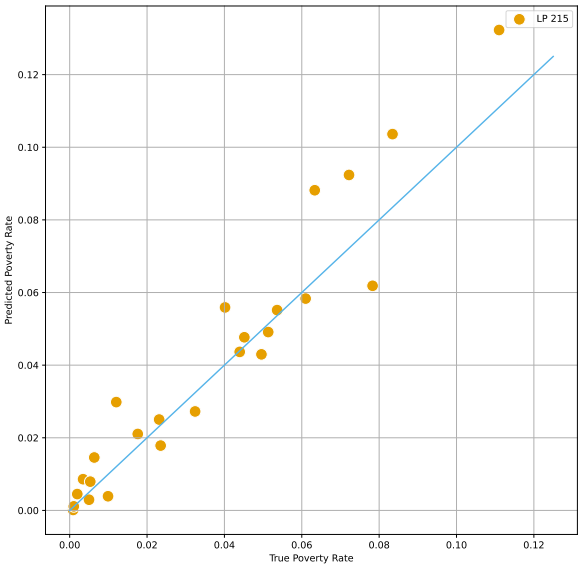
Figure 15: Correlation: Predicted Poverty against true by Region (WB Version)



(a) Poverty Line: 685



(b) Poverty Line: 365



(c) Poverty Line: 215

Note:

Figure 16: Poverty Rate by Region

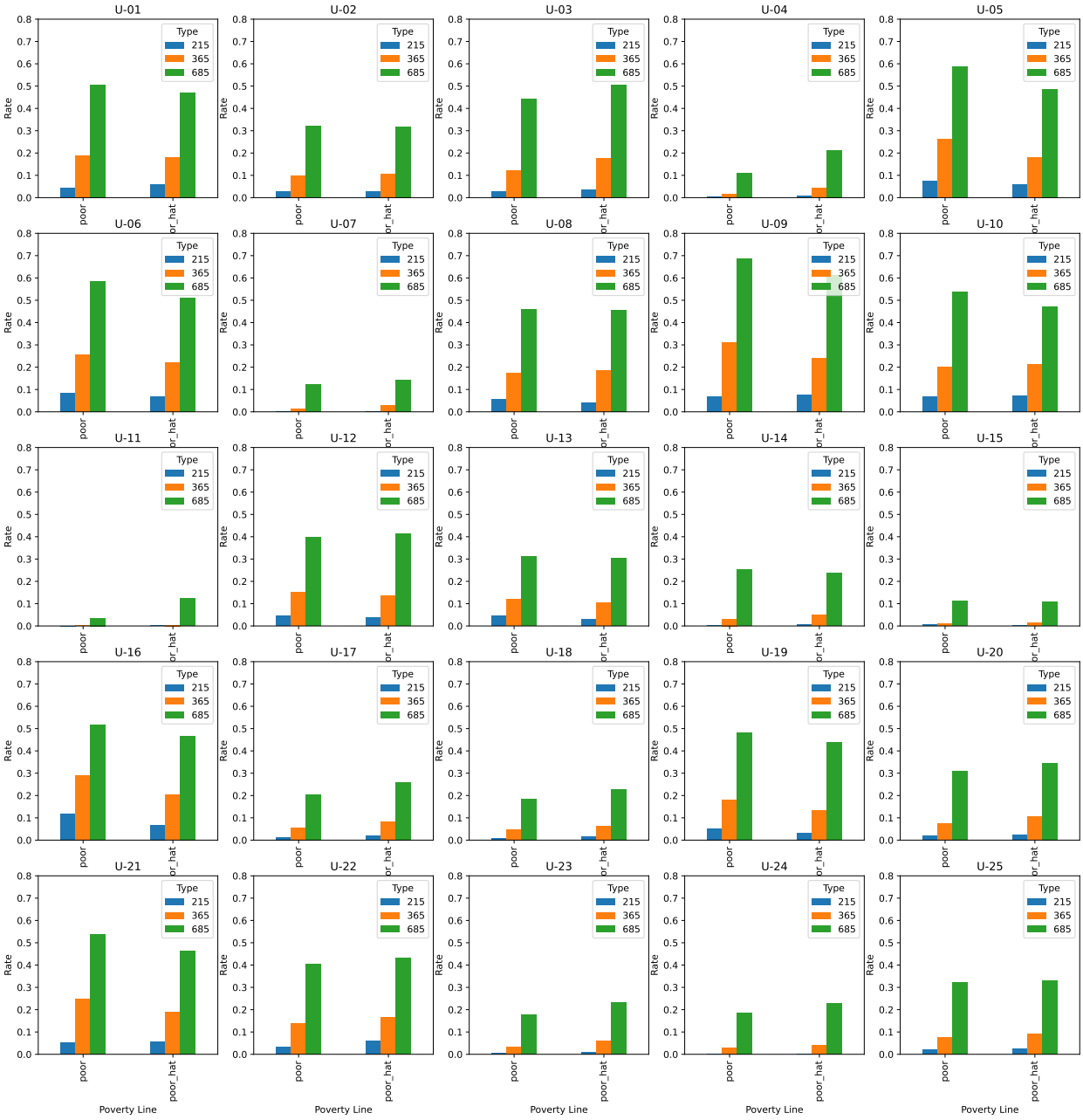


Figure 17: Average Per-Capita Household Income by Year

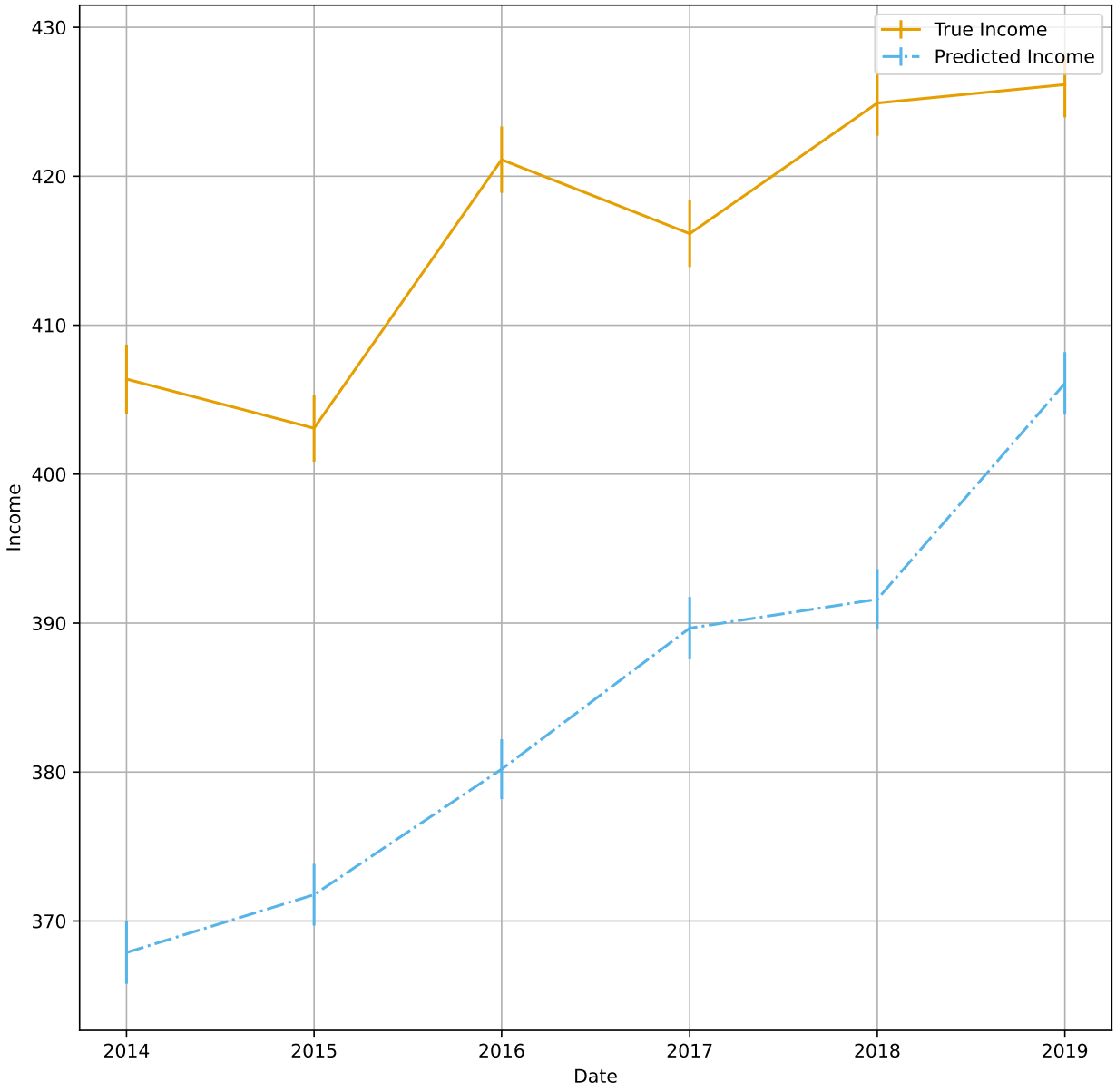


Figure 18: Average Per-Capita Household Income by Year and Area

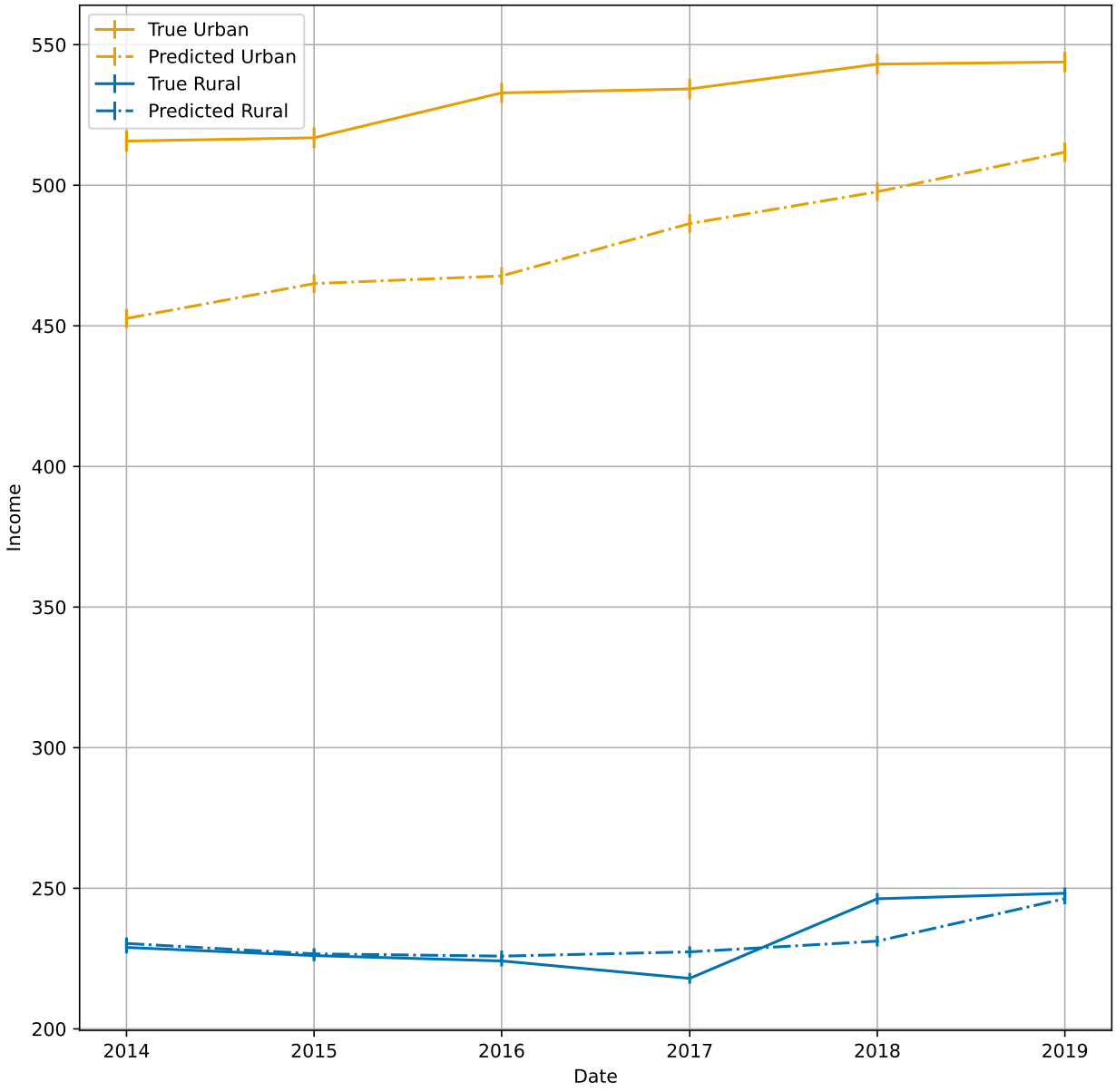


Figure 19: Average Per-Capita Household Income Quarterly

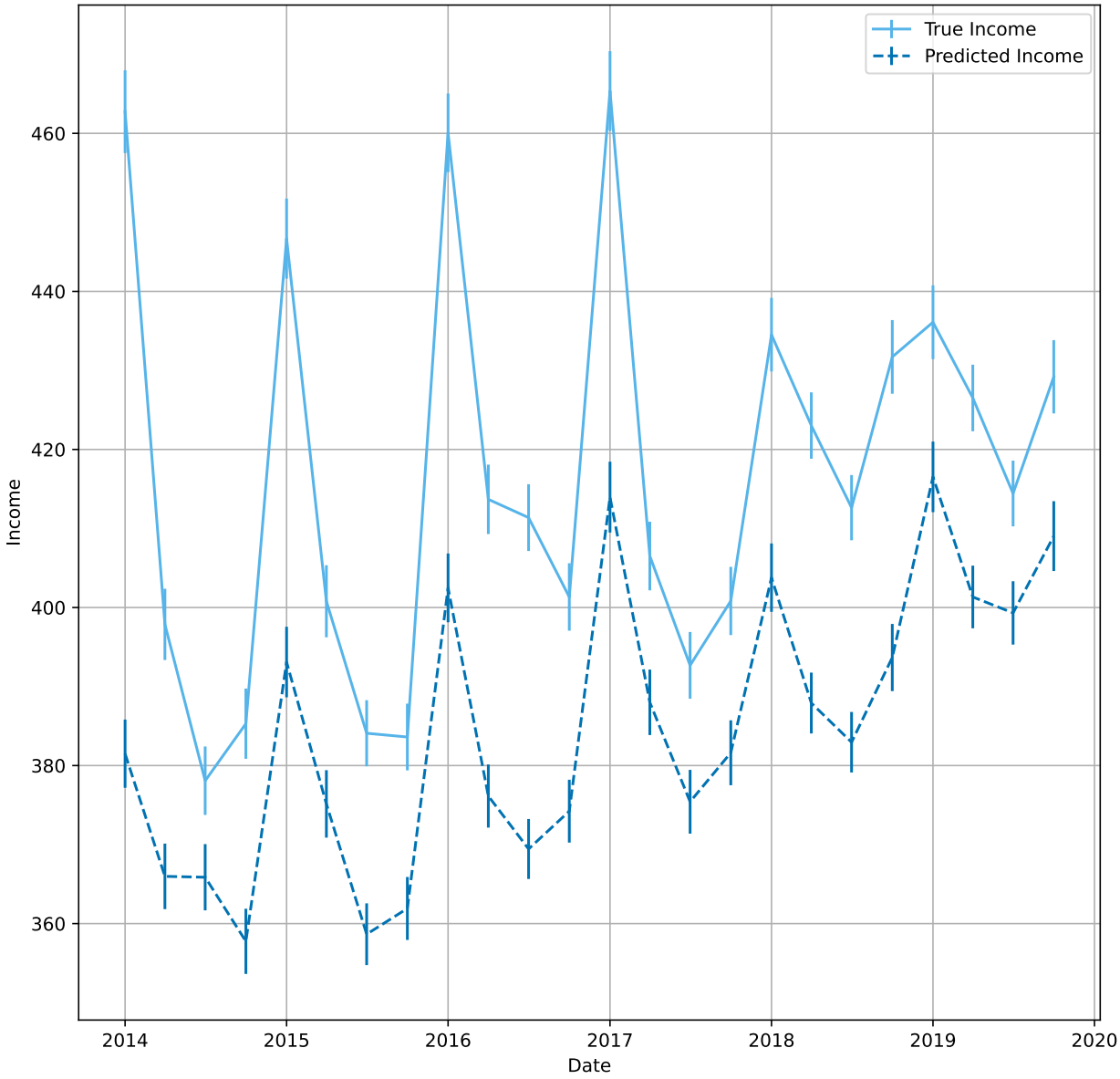


Figure 20: Poverty Rate Yearly

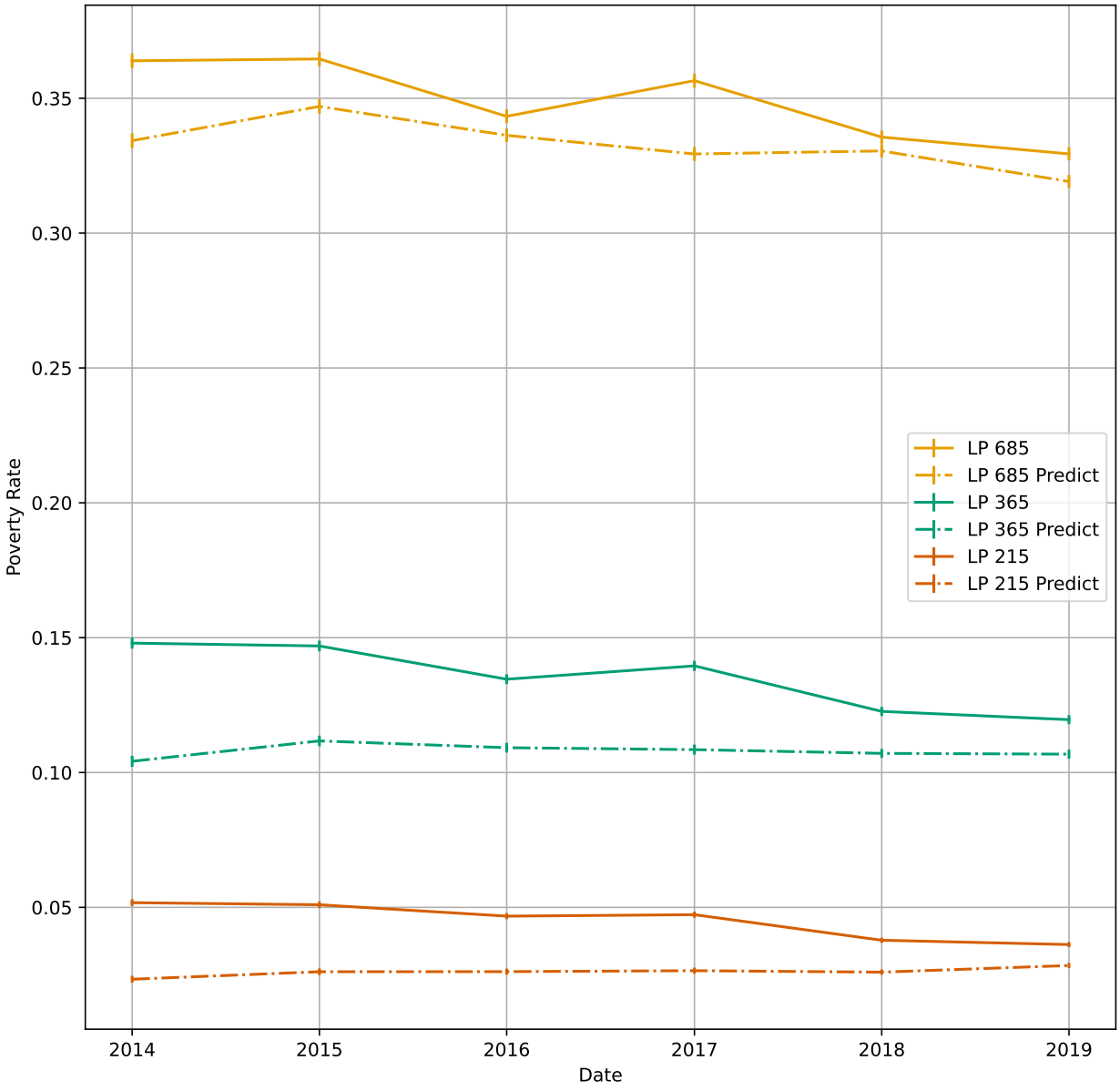


Figure 21: Poverty Rate Yearly Urban

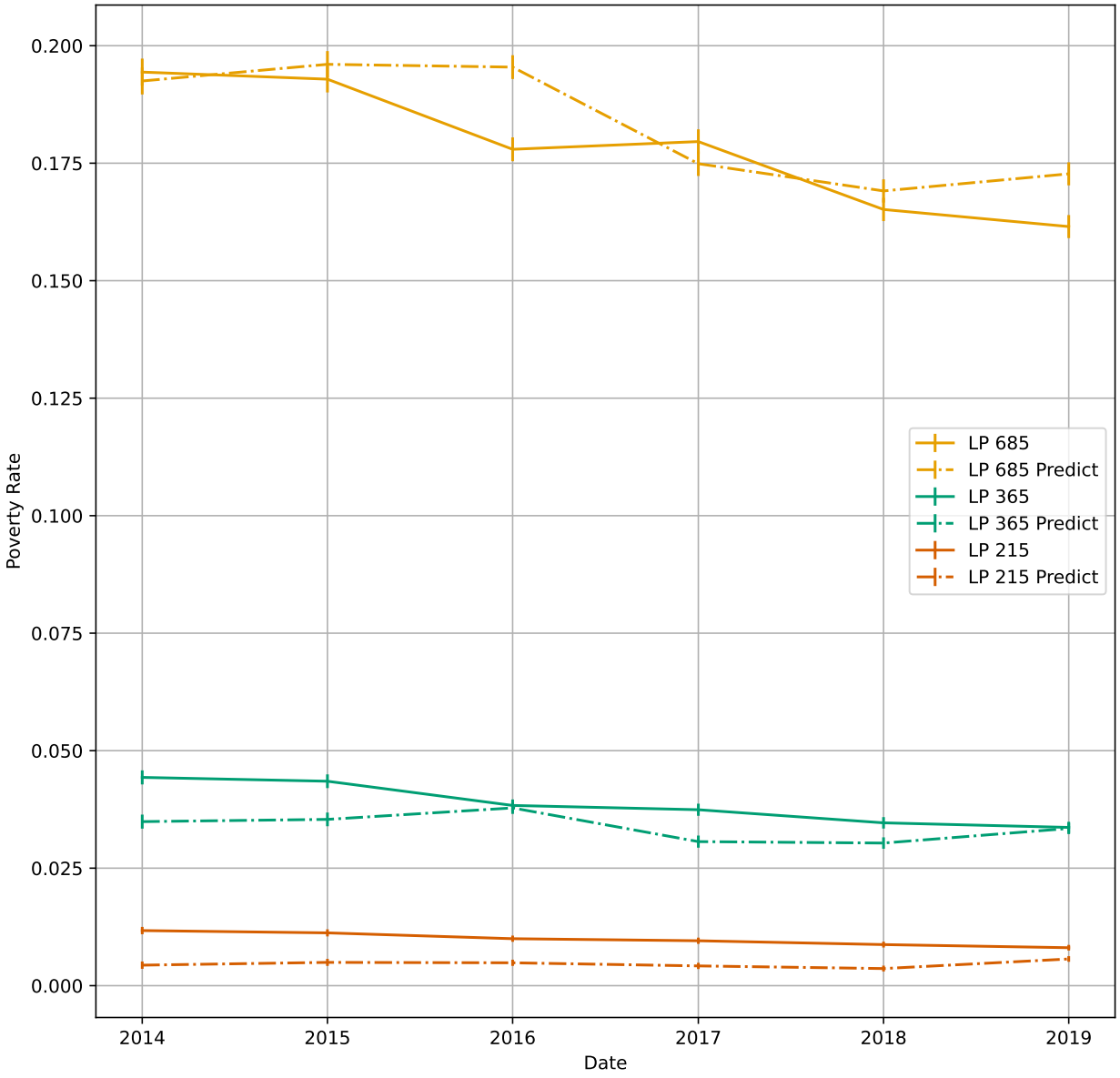


Figure 22: Poverty Rate Yearly Rural

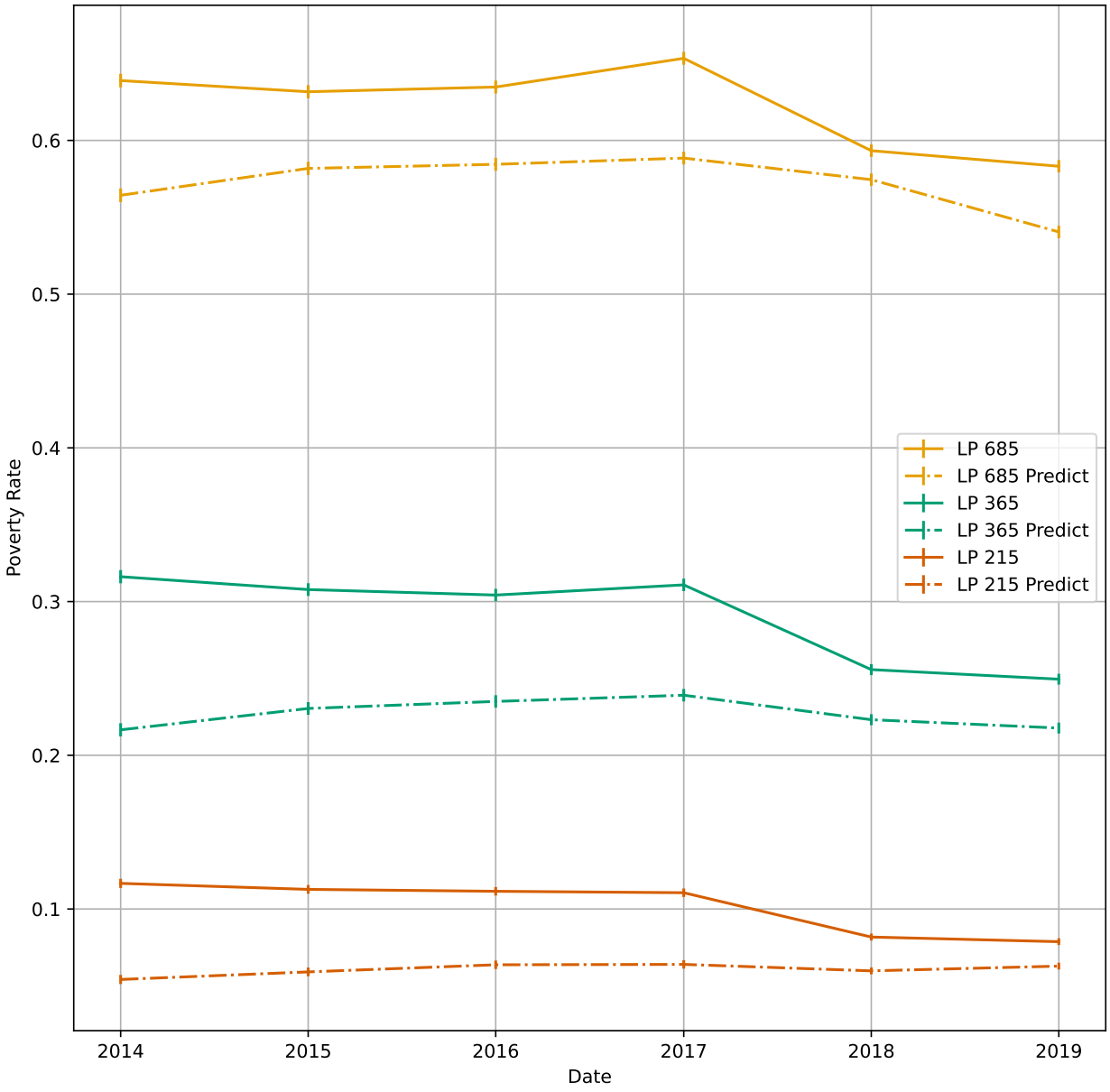


Figure 23: Poverty Rate Yearly Lima

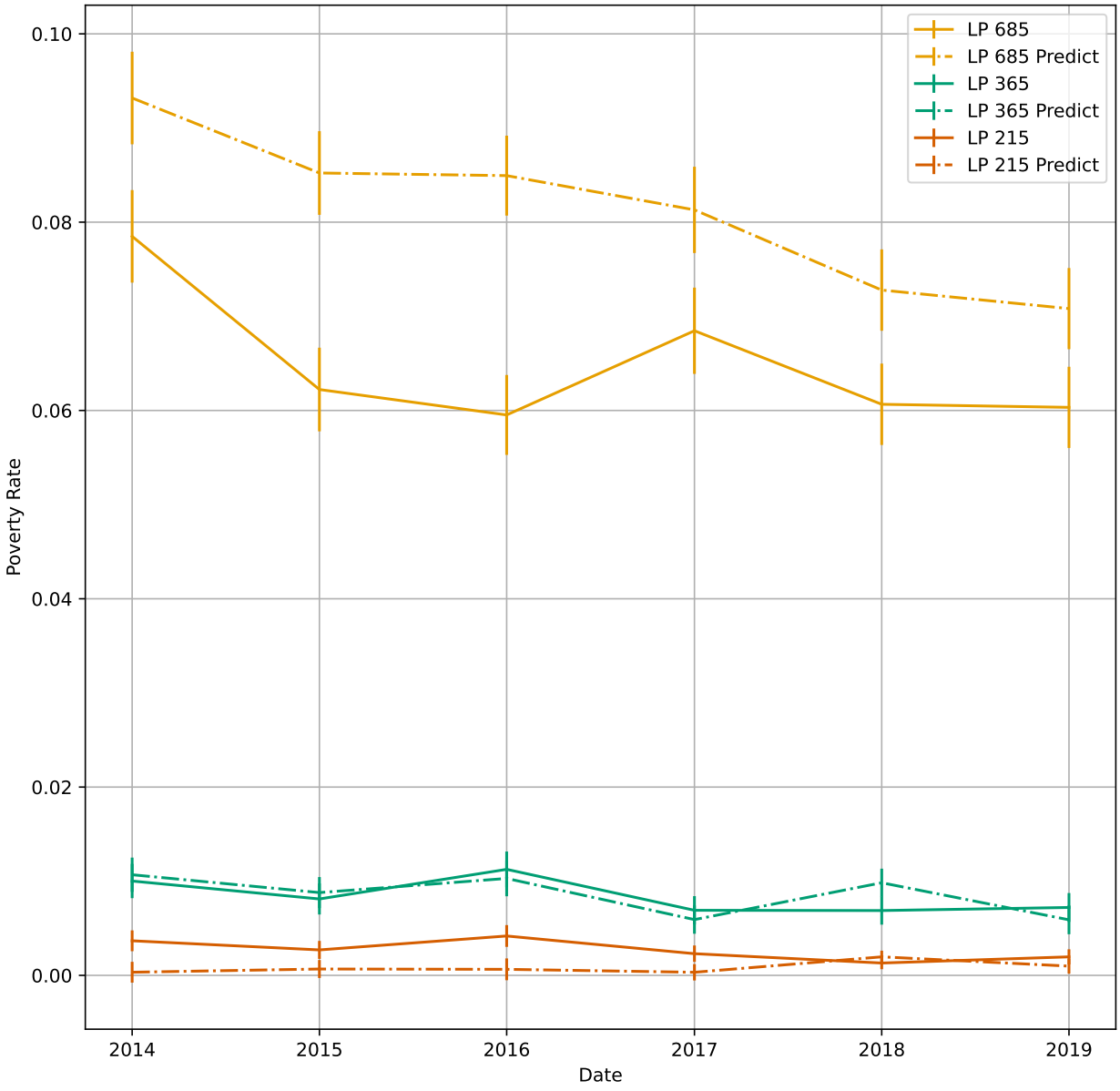
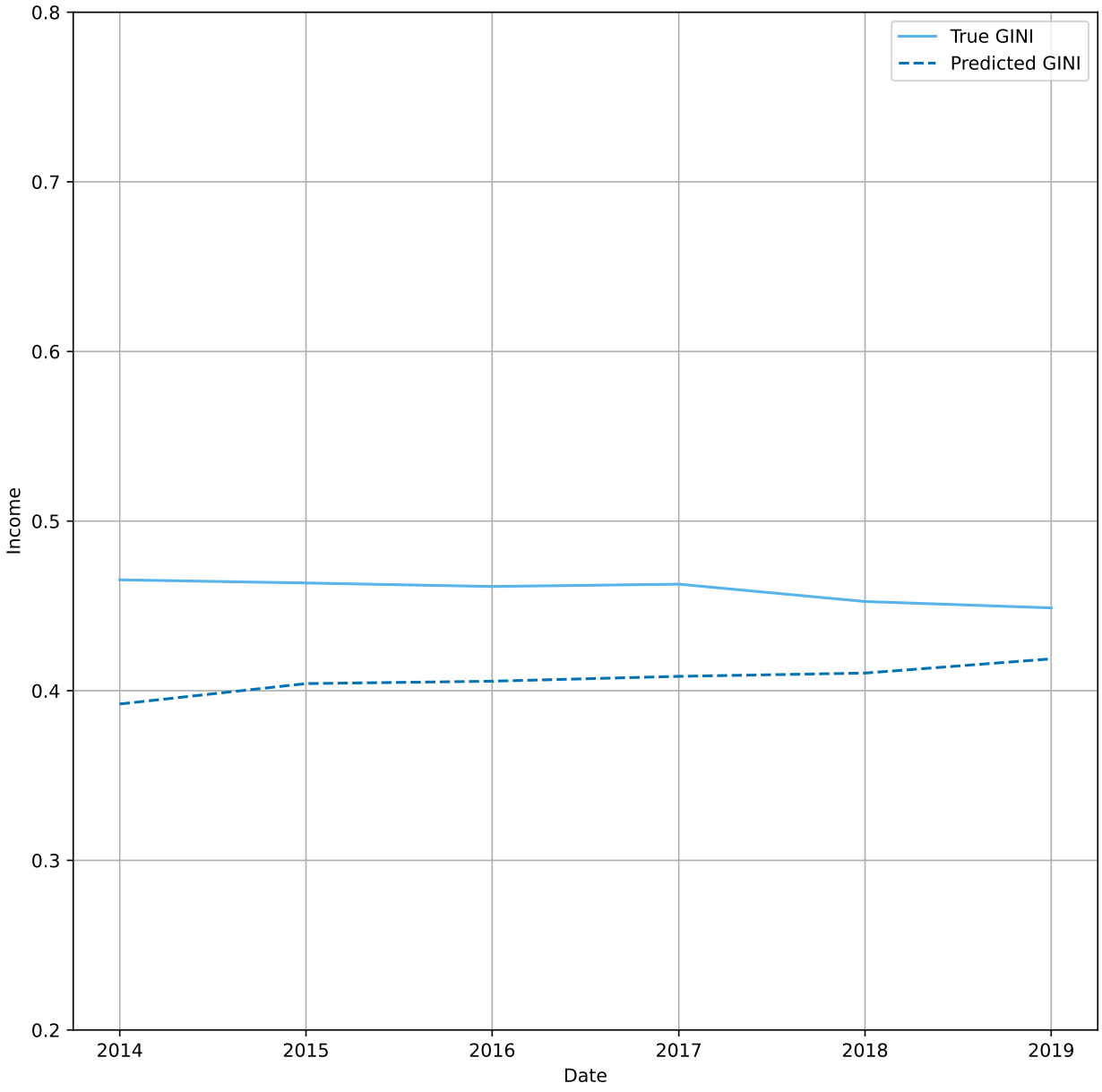


Figure 24: Gini Coefficient Yearly National



6 Conclusions:

The preliminary findings of our research indicate a significant improvement in income prediction accuracy by leveraging a extensive data sources. The integration of household survey data with big data sources, such as administrative records, weather patterns, and nightlight satellite imagery, has provided a more nuanced and comprehensive understanding of the variables affecting household income. Machine learning techniques, especially advanced predictive models, have successfully identified complex, non-linear relationships within the data, leading to more robust forecasting. Preliminary results suggest that environmental factors and economic activity, as inferred from weather and nightlight data, are particularly influential in predicting income changes. These insights underline the potential of using a diverse data ecosystem to inform economic policies and social welfare programs. Our research paves the way for more dynamic, data-driven decision-making processes in economic planning and poverty alleviation strategies. Further analysis and model refinement are ongoing to validate these initial outcomes and to explore the broader applicability of our approach.

7 Sampling Weighting Factor:

Probability of selection of a conglomerate according to INEI

$$Pr(U_{rhi}) = \frac{n_{rh} M_{rhi}}{M_{rh}}$$

- U_{rhi}: i-th conglome in h-th stratum, in r-th region.
- n_{rh}: Number of conglome in h-th stratum in r-th region.
- M_{rhi}: Number of households in i-th conglome, in h-th stratum in r-th region.
- M_{rh} Total number of households in the h-th stratum in r-th region.

Probability of selection of a household within a conglome. Inei clusters households by type (households with children, without children and others) let's call these type k.

$$Pr(H_{krhi}|U_{rhi}) = \frac{m_{krhi}}{M_{rhi}}$$

- H_{krhi}: household type k, in region r, stratum h and conglome i.
- m_{krhi}: Number of households type k, in region r, stratum h and conglome i.

Joint probability of choosing a household and a conglome:

$$\begin{aligned} Pr(H_{krhi}, U_{rhi}) &= Pr(H_{krhi}|U_{rhi}) \times Pr(U_{rhi}) \\ &= \frac{m_{krhi}}{M_{rhi}} \times \frac{n_{rh} M_{rhi}}{M_{rh}} \\ &= \frac{m_{krhi} \times n_{rh}}{M_{rh}} \end{aligned}$$

In ENAHO, *factor_expansion*$_{krhi}$ is the inverse of the joint prob of choosing family type k and conglome i in stratum h from region r.

$$Pr(H_{krhi}, U_{rhi}) = \textit{factor\_expansion}_{krhi}^{-1} = \frac{m_{krhi} \times n_{rh}}{M_{rh}}$$

This is something that we observe. So what we can do is use *law of total probability* to integrate across households type k and we can back out the probability of choosing U_{rhi}.

$$\begin{aligned}\sum_k Pr(H_{krhi}, U_{rhi}) &= \sum_k \textit{factor\_expansion}_{krhi}^{-1} \\ &= \sum_k \frac{m_{krhi} \times n_{rh}}{M_{rh}} \\ &= \frac{n_{rh} \sum_k m_{krhi}}{M_{rh}} \\ &= \frac{n_{rh} \times M_{rhi}}{M_{rh}} = Pr(U_{rhi})\end{aligned}$$

Note that the final step uses the fact that: $M_{rhi} = \sum_k m_{krhi}$. Since we are summing across all household types within rhi, we get the total number of households in rhi which is M_{rhi}. That way we can back out $Pr(U_{rhi})$. QED.