A simple and adaptable method to estimate the income distribution in Spain

Pablo García-Guzmán

EBRD

December 1, 2024

Abstract

This methodology note develops a transparent and simple framework to estimate aggregate income distributions in Spain exploiting the availability of granular inequality metrics. Specifically, I leverage administrative tax records and demographic data at the census tract-level to reconstruct local income distributions assuming log-normality. National, regional and municipal income distributions can be subsequently derived as population-weighted mixtures of these tract-level distributions. This approach leverages well-documented regularities in how incomes are distributed within demographically homogeneous groups, and the resulting estimates closely match observed distributional statistics in validation tests. I implement this framework in comparatuingreso.es, a publicly available web platform that enables Spanish households to calculate their relative position within the income distribution. While the focus is on Spain, this approach can be readily adapted to other countries with comparable data available.

Email: garciagp@ebrd.com. I thank Pedro Torres for his valuable feedback and comments. The views expressed in this methodology note are completely my own and do not reflect the views of the EBRD.

1. Introduction

Understanding the income distribution is crucial for economic analysis and policy design, yet its measurement remains challenging. In this paper, I propose a transparent and simple method to estimate national-level income distributions when granular income and inequality measures are available. The approach builds on two key insights: first, that income distributions within demographically homogeneous areas tend to follow log-normal distributions; and second, that aggregate distributions can be effectively modeled as mixtures of these local distributions. Using detailed administrative data from over 36,000 Spanish census tracts, I demonstrate that local income distributions are well-approximated by log-normal distributions, with predicted statistics matching observed values within 5% on average. This method provides a flexible parametric framework for estimating income distributions when individual data is unavailable

The rest of this note is organized as follows. Section 2 describes the data sources, their coverage, and scope. Section 3 outlines the methodology, including the modeling of local income distributions, their aggregation and the calculation of the relevant validation metrics. Section 4 presents the results. Finally, Section 5 concludes the note.

2. Data

The data used in this project is sourced from the Spanish Statistical Office (INE) Household Income Distribution Atlas (*Atlas de Distribución de Renta de los Hogares*, ADRH).[1] This dataset combines administrative tax data with population statistics to provide detailed information about the income distribution and related socioeconomic indicators at granular geographic levels in Spain.

Spain is administratively divided into several territorial levels that structure its governance and statistical reporting. At the highest level are the *Comunidades Autónomas* (autonomous communities), which are regions with significant legislative and executive powers granted by the Spanish Constitution. There are 17 autonomous communities and two autonomous cities (Ceuta and Melilla). These regions are further subdivided into *provincias* (provinces), the primary intermediate level of governance. Each province consists of *municipios* (municipalities), the fundamental local administrative units, which vary widely in size and population.

Below the municipal level, finer geographic divisions provide additional granularity. For statistical purposes, municipalities are divided into *distritos* (districts), which are then further subdivided into *secciones censales* (census tracts). Census tracts, encompassing areas with populations typically ranging from 1,000 to 2,500 residents, provide the finest level of geographic detail available in the ADRH.

Income data in the ADRH is derived from tax declarations submitted to the Spanish tax authorities, including the *Agencia Estatal de Administración Tributaria* (AEAT) and the Foral tax authorities.[2] These tax records provide detailed information on income from various sources, including wages,

[1] Available online at https://www.ine.es/experimental/atlas/experimental_atlas.htm/.

[2] The latter operate in regions with special fiscal regimes, such as the Basque Country and Navarre.

pensions, unemployment benefits, and other forms of revenue subject to the *Impuesto sobre la Renta de las Personas Físicas* (IRPF), Spain's personal income tax system. The dataset excludes non-resident income and focuses exclusively on individuals considered fiscal residents within the territory.

Demographic data in the ADRH is constructed from the *Fichero Precensal de Población* (FPC), a comprehensive register derived from the municipal register (*Padrón*) and other administrative sources. The FPC forms the basis of Spain's population census and ensures alignment between demographic and income data. The temporal reference for income data corresponds to the calendar year, while demographic data is anchored to the population as registered on January 1 of the following year. Individuals residing in collective establishments, such as nursing homes, hospitals, or military barracks, are excluded from the dataset.

Summary statistics. The sample consists of 36,982 census tracts covering all of Spain in 2022. The average tract has 1,303 residents, though size varies considerably (SD = 663). Average net income per equivalent adult is €20,798, about 48% higher than per capita income (€14,030). Within-tract inequality, measured by the Gini coefficient, averages 29 points across tracts. Missing rates are below 2% for most variables, and rise slightly above 5% for equivalized income and the Gini coefficient. Demographically, the average tract has a dependency ratio of 0.63, a mean age of 45.4 years, and 30.5% single-person households.

TABLE 1. Summary statistics

	(1) Min	(2) Max	(3) Mean	(4) SD	(5) % missing
Income distribution					
Net income per capita	4,996.00	34,765.00	14,030.16	4,193.34	1.63
Net income per equivalent adult	8,092.00	55,987.00	20,797.51	6,628.08	5.20
Gini	20.30	44.20	29.02	3.84	5.20
Demographics					
Population	3.00	12,144.00	1,303.10	663.26	1.45
Dependency ratio	0.09	4.00	0.63	0.17	1.45
Mean age	27.40	74.70	45.43	5.52	1.45
Single-person households (%)	0.00	100.00	30.50	9.56	1.45
No. of tracts	36,982				

Notes: all statistics are unweighted and calculated at the census-tract level. Spanish Statistical Office (INE) Household Income Distribution Atlas and author's calculations.

To improve coverage, I conduct two simple adjustments. First, I impute net income per equivalent adult within missing tracts using net income per capita and the average provincial ratio between these two measures, calculated using population weights from non-missing observations. This adjustment reduces missing rates in equivalised income per capita from 5.2% to 1.63%. Second,

I impute missing Gini coefficients (5.2% of tracts) using a machine learning approach. More details can be found in Section 3.

3. Methodology

3.1. Background

The study of income distributions has been a cornerstone of economic research. A long-standing empirical regularity in the literature is that income distributions tend to approximate a log-normal density (Aitchison and Brown, 1957). This stylized fact is supported by both theoretical and empirical foundations. At its core, log-normality arises from the multiplicative interaction of economic factors – such as human capital, local labor market conditions, and productivity – which determine individual incomes (Neal and Rosen, 2000). Gibrat's Law of Proportionate Effect (Gibrat, 1931) formalizes the theoretical basis for log-normality through multiplicative random growth, showing that when the growth rate of a variable is independent of its initial size and the logarithm of the growth rate is independent and identically distributed over time with finite variance, the resulting distribution tends to converge to log-normality.

More concretely, the multiplicative processes through which log-normality arises operate most strongly within demographically homogeneous groups where individuals are subject to similar economic shocks (Weiss, 1972; Aitchison and Brown, 1957; Battistin, Blundell, and Lewbel, 2009). This methodology note relates to a larger body of research providing empirical support for this pattern, showing that a mixture of log-normals is particularly suitable for modeling income in fairly homogeneous sub-populations while at the same time accounting for between-group heterogeneity (Flachaire and Nuñez, 2007; Lubrano and Ndoye, 2016). This finding is particularly relevant at higher levels of aggregation, where finite mixtures of log-normals provide substantial improvements in fit compared to single log-normal specifications (Gardini, Fabrizi, and Trivisano, 2022).

The literature of neighborhood effects offers a complementary explanation for why such homogeneity in income-generating processes exists at local levels. In particular, neighborhood effects reinforce homogeneity within local populations through contextual influences and endogenous spillovers, which emerge from behavioral interactions within the neighborhood, such as peer influences or social norms (Manski, 1993; Durlauf, 1996). These mechanisms create reinforcing feedback loops that generate correlated income trajectories within neighborhoods – for instance, through shared information about job opportunities, similar human capital accumulation patterns, or common responses to local economic shocks. The resulting interaction structures generate strong within-neighborhood homogeneity while maintaining between-neighborhood heterogeneity (Durlauf, 2004). This pattern provides a rationale for estimating aggregate income distributions as mixtures of local-level distributions.

However, while providing a good approximation for much of the income distribution, the log-normal approximation exhibits systematic deviations in the tails. Since Pareto's (1896) seminal work, research has shown that top incomes follow a power law rather than a log-normal decay

(Gabaix, 2016). Similarly, studies using detailed administrative tax data have documented that the upper tail of income distributions across countries and time periods are better described by a Pareto distribution (Atkinson, Piketty, and Saez, 2011). This departure from log-normality at high incomes is important for accurately measuring top income inequality, but modeling the precise behavior of the upper tail is beyond the scope of this note.

3.2. Empirical approach

I estimate the national income distribution by exploiting the granular structure of census tract-level data and the theoretical properties of income distributions. My approach leverages the fact that income distributions within small geographic units tend to follow log-normal distributions more closely than in larger areas, and approximates the aggregate distribution using a mixture of these local-level distributions.

Consider a census tract j with an observed mean income μ_j and Gini coefficient G_j. Under the log-normality assumption, if income Y follows a log-normal distribution with parameters (ν_j, σ_j^2), then:

$$\ln(Y) \sim N(\nu_j, \sigma_j^2) \tag{1}$$

The relationship between these parameters and the observed statistics is thus given by:

$$\mu_j = \exp\left(\nu_j + \frac{\sigma_j^2}{2}\right) \tag{2}$$

$$G_j = 2\Phi\left(\frac{\sigma_j}{\sqrt{2}}\right) - 1 \tag{3}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. From the observed Gini coefficient, I can recover σ_j:

$$\sigma_j = \sqrt{2}\Phi^{-1}\left(\frac{G_j + 1}{2}\right) \tag{4}$$

Given σ_j and μ_j, the distribution for census tract j is fully identified, as ν_j can be obtained analytically. I then estimate the national distribution as a population-weighted mixture of these local log-normal distributions. For a given income level y, the density is given by:

$$f(y) = \sum_{j=1}^{J} w_j f_j(y|\nu_j, \sigma_j^2) \tag{5}$$

where w_j is tract j's population share and $f_j(\cdot|\nu_j, \sigma_j^2)$ is the log-normal density function with parameters (ν_j, σ_j^2).

To calculate percentiles of the national distribution, I solve:

$$p = F(q_p) = \sum_{j=1}^{J} w_j \Phi\left(\frac{\ln(q_p) - \nu_j}{\sigma_j}\right) \tag{6}$$

where q_p is the p^{th} percentile and $F(\cdot)$ is the cumulative distribution function of the mixture. I use numerical root-finding methods to find the value of q_p that satisfies $F(q_p) - p = 0$ within a bounded interval.

Missing outcomes. For census tracts with missing Gini coefficients (approximately 5.2% of the sample), I estimate them using a machine learning approach. Specifically, I train an XGBoost model using the following demographic predictors: dependency ratio (ratio of population under 18 and over 65 to working-age population), mean age, percentage of single-person households, mean logged equivalised income, mean household size, population size, and province fixed effects. Model performance is validated using a 5-fold cross-validation procedure, and is trained using default hyperparameters without grid search or additional tuning. For comparison, I also estimate a baseline OLS model using the same set of predictors. Out-of-sample prediction errors under both approaches are shown in Table 2. Based on the cross-validation results, the XGBoost model demonstrates superior predictive performance compared to a baseline OLS model, as illustrated by a 9.35% lower RMSE.

TABLE 2. Performance comparison between OLS and XGBoost models

Model	RMSE	MAE	MAPE (%)
OLS	3.10	2.32	8.02
XGBoost (CV)	2.81	2.17	7.49
Relative improvement over OLS (%)	9.35%	6.47%	6.61%

Notes: This table compares the out-of-sample performance of two predictive models – OLS and XGBoost – on predicting missing Gini coefficients using demographic predictors. RMSE represents the root mean square error, MAE is the mean absolute error, and MAPE is the mean absolute percentage error. Relative improvements in each metric are calculated as the percentage reduction of the error metric when using XGBoost compared to OLS. Results are based on a 5-fold cross-validation procedure for the XGBoost model.

Validation. To validate the log-normality assumption at the tract level, I compare observed distributional statistics with their theoretical counterparts under log-normality. For each tract j, I first recover the parameters of the theoretical log-normal distribution (ν_j, σ_j) using the observed mean income and Gini coefficient. Then, I calculate two key predicted statistics: the P80/P20 ratio and the median income. Under log-normality, these are given by:

$$\text{P80/P20}_j = \exp(\sigma_j(\Phi^{-1}(0.8) - \Phi^{-1}(0.2))) \tag{7}$$

$$\text{P50}_j = \exp(\nu_j) \tag{8}$$

To assess the fit, I estimate OLS regressions of the form:

$$\hat{y}_j = \alpha + \beta y_j + \epsilon_j \tag{9}$$

where $\hat{y}_j$ is the predicted value under log-normality and y_j is the observed value.

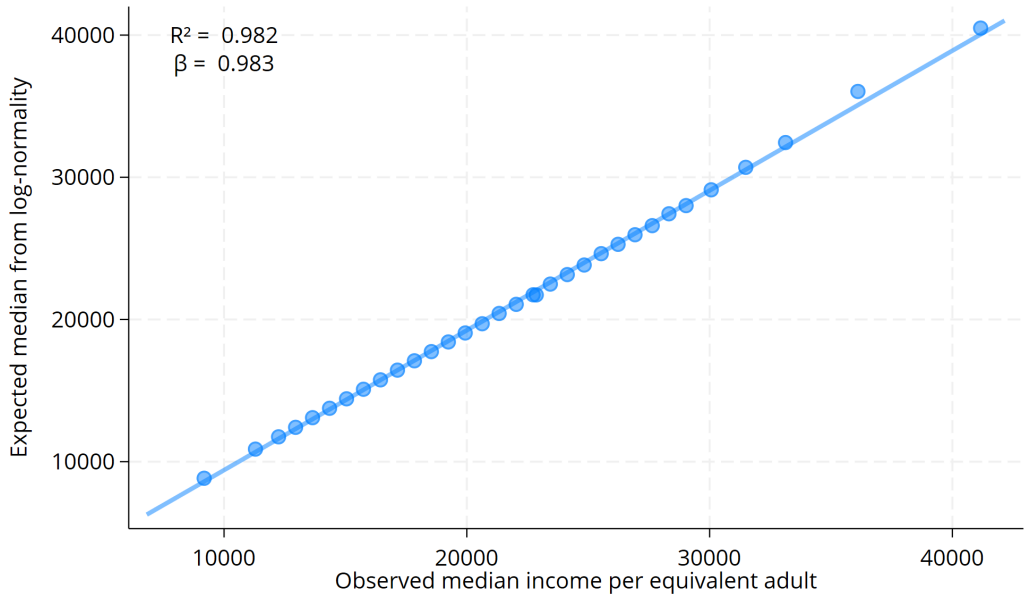
4. Results

I begin by validating the log-normality assumption at the tract level. Figure 1 examines the fit for median income, comparing observed values with those predicted under log-normality for each census tract, where observations are grouped into equal-sized bins using binscatter methods.[3] The relationship is remarkably strong ($\beta = 0.98$, $R^2 = 0.98$), with predicted values differing from observed ones by just 4.6% on average. The near-unit slope coefficient and R^2 indicate that log-normality captures the central tendency of the income distribution particularly well.

Figure 1 shows the relationship for the P80/P20 ratio. The regression yields a slope coefficient of 0.75 ($R^2 = 0.81$), with predicted values deviating from observed ones by 5% on average. Consistent with the fact that real income distributions tend to have heavier tails, the coefficient below unity suggests that the log-normal distribution slightly underestimates inequality.

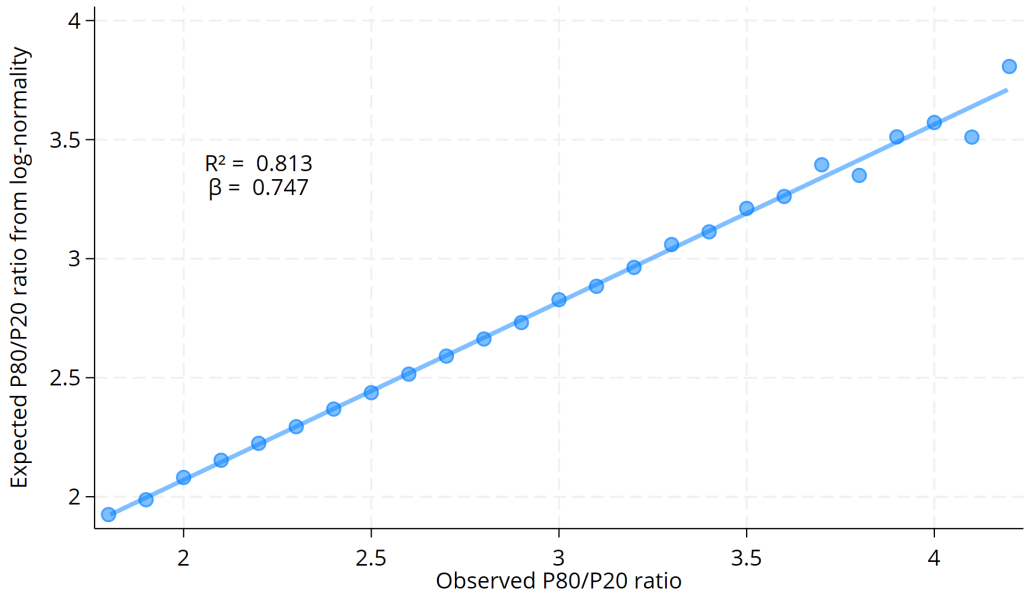
[3] See Cattaneo et al. (2024).

FIGURE 1. Validation of log-normality: median income



Notes: This figure shows a binned scatter plot comparing observed median income with predicted values under the log-normal assumption for each census tract. Source: Spanish Statistical Office and author's calculations.

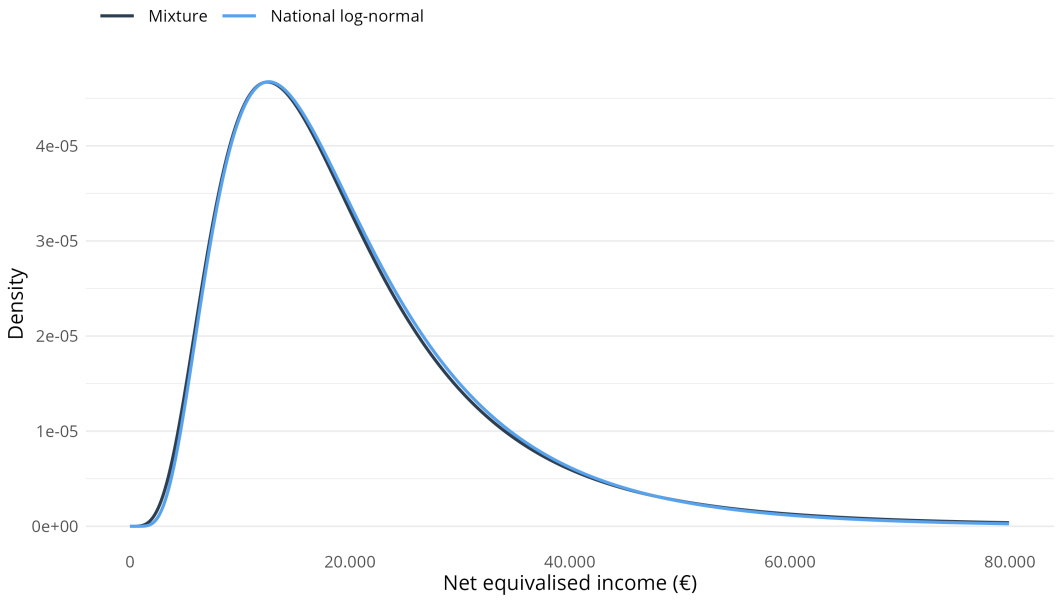
FIGURE 2. Validation of log-normality: P80/P20 ratio



Notes: This figure shows a binned scatter plot comparing observed P80/P20 ratios with predicted values under the log-normal assumption for each census tract. Source: Spanish Statistical Office and author's calculations.

A natural question arises: to what extent does exploiting local-level heterogeneity through a mixture of tract-specific distributions improve our measurement of the national income distribution compared to a simpler log-normal approximation? The comparison shown in Figure 3 reveals a striking finding: despite incorporating rich geographic variation in both average income and inequality metrics, the mixture distribution is remarkably similar to a single log-normal distribution fitted using only the national-level average income and Gini coefficient derived from household survey data (EU-SILC). This pattern aligns with previous evidence showing that log-normal densities approximate very well the empirical distribution of per capita income in large cross-country panels (Lopez and Servén, 2006). In contexts where more granular data is not available, these results suggest that assuming log-normality at the national level might remain an accurate approximation for capturing the overall shape of the income distribution.

FIGURE 3. Estimated mixture vs. national log-normal



Notes: This figure compares the income distribution derived from a mixture of tract-level log-normal distributions (navy) with a single national-level log-normal distribution (light blue). The mixture distribution is constructed using tract-specific log-normal distributions, where the parameters of each component are derived from observed tract means and Gini coefficients, and mixture weights correspond to tract population shares. The parameters for the national log-normal are given by the national-level mean income and Gini coefficient from EU-SILC 2023. Data refers to year 2022. Source: Spanish Statistical Office and author's calculations.

To better understand this result, I examine the spatial structure of income inequality in Spain through a hierarchical variance decomposition exercise. Specifically, I decompose the total variance of log income within each autonomous community c into components corresponding to different geographic levels as follows:

$$\sigma_c^2 = \underbrace{\sum_j w_j \sigma_j^2}_{\text{Within-tract}} + \underbrace{\sum_{k \in c} \sum_{j \in k} w_j \left(\mu_j - \mu_k\right)^2}_{\text{Between-tract}} + \underbrace{\sum_{p \in c} \sum_{k \in p} w_k \left(\mu_k - \mu_p\right)^2}_{\text{Between-municipality}} + \underbrace{\sum_{p \in c} w_p \left(\mu_p - \mu_c\right)^2}_{\text{Between-province}} \tag{10}$$

where σ_j^2 is the variance of log income within each census tract j, μ_j is the mean log income in tract j, μ_k is the mean log income in municipality k, μ_p is the mean log income in province p, and μ_c is the cross-tract mean log income within autonomous community c. Population weights for tracts, municipalities, and provinces, are denoted by w_j, w_k, w_p, respectively. The decomposition separates total income variation into four components: within-tract, between-tract (within municipalities), between-municipality (within provinces), and between-province variation. The within-tract component is calculated assuming log-normality within each tract.

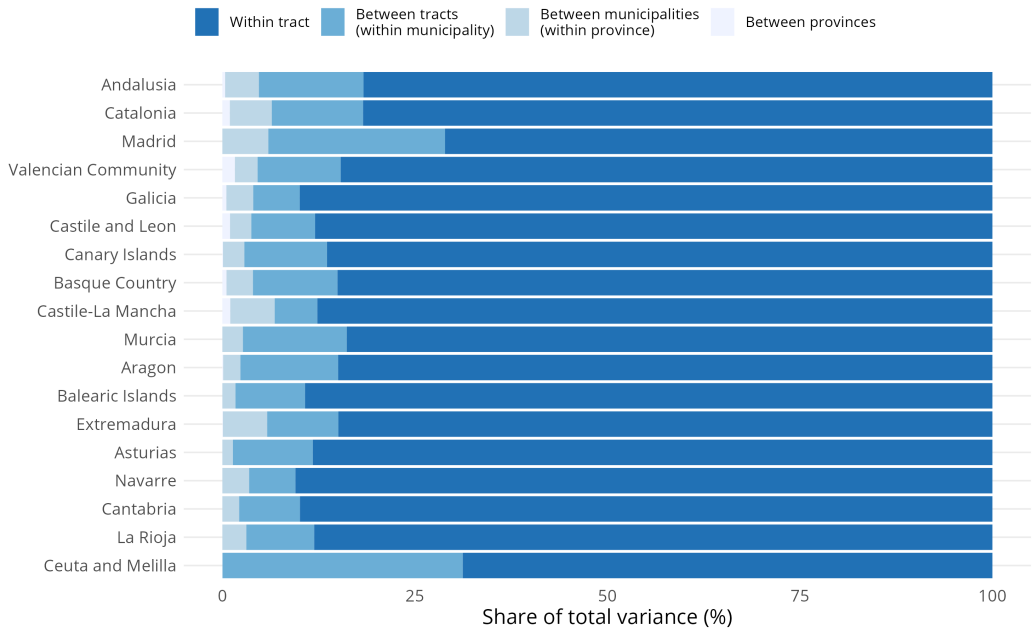
Figure 4 presents the results of this decomposition by autonomous community. The analysis reveals that within-tract variation accounts by far for the largest share of income inequality, representing between 85% and 91% of total variance within autonomous communities. This finding provides additional support for using tract-level distributional estimates as the fundamental building blocks for the aggregate distribution.[4] A further 12% is on average explained by between-tract, within-municipality differences. Income differences between municipalities and provinces contribute only marginally to overall income inequality.[5]

This predominance of within-tract variation provides intuition for why the mixture and single log-normal distributions are so similar. Since between-tract heterogeneity accounts for a relatively small share of total inequality, incorporating tract-level variation through a mixture approach adds limited value beyond a simple national-level approximation in the Spanish context.

[4] Notice that this does not contradict the earlier discussion on homogeneity in income-generating *processes* within neighborhoods. The latter refers to the shared mechanisms (e.g., proportional shocks) influencing income trajectories within neighborhoods, but does not imply homogeneity in outcomes. Individual incomes can still vary significantly due to structural or stochastic factors.

[5] Extending the decomposition analysis at the national level shows that between-community income disparities only account for 5% of total variance.

FIGURE 4. Variance decomposition



Notes: This figure shows the decomposition described in Equation 10 whereby total variance in log income is broken down into four components: within-tract, between-tract (within municipality), between-municipality (within province), and between-province variation. The within-tract component is calculated assuming log-normality within tracts. Regions are ordered by total population. Source: Spanish Statistical Office and author's calculations.

5. Conclusion

This methodology note presents a simple framework for estimating income distributions in Spain, leveraging granular inequality metrics and administrative data at the census-tract level. Specifically, I model local income distributions as log-normal and derive national and regional distributions as population-weighted mixtures of tract-level estimates. Validation results show a high degree of accuracy, with predicted metrics closely matching observed distributional statistics across census tracts. However, results also suggest that the improvement offered by the tract-level mixture relative to assuming log-normality at the national level is very limited, suggesting that the latter can still serve as a reasonable approximation to estimate aggregate distributions in the Spanish context. Results from a hierarchical variance decomposition exercise rationalize this finding by showing that within-tract variation explains the vast majority of income inequality in Spain, and hence accounting for between-tract heterogeneity adds relatively little information to the overall pattern of income inequality at the national level.

Looking ahead, the methodology offers opportunities for extension, such as explicitly addressing the deviations from log-normality at the tails of the distribution or adapting the framework for use in other countries with comparable data available.

References

Aitchison, John, and James AC Brown. 1957. "The Lognormal Distribution." *Cambridge University Press*.

Atkinson, Anthony B, Thomas Piketty, and Emmanuel Saez. 2011. "Top incomes in the long run of history." *Journal of Economic Literature* 49 (1): 3–71.

Battistin, Erich, Richard Blundell, and Arthur Lewbel. 2009. "Why Is Consumption More Log Normal than Income? Gibrat's Law Revisited." *Journal of Political Economy* 117 (6): 1140–1154.

Cattaneo, Matias D., Richard K. Crump, Max H. Farrell, and Yingjie Feng. 2024. "On Binscatter." *American Economic Review* 114 (5): 1488–1514.

Durlauf, Steven N. 1996. "A Theory of Persistent Income Inequality." *Journal of Economic Growth* 1 (1): 75–93.

Durlauf, Steven N. 2004. "Neighborhood effects." In *Handbook of Regional and Urban Economics*, vol. 4, 2173–2242: Elsevier.

Flachaire, Emmanuel, and Olivier Nuñez. 2007. "Estimation of the income distribution and detection of subpopulations: An explanatory model." *Computational Statistics Data Analysis* 51 (7): 3368–3380.

Gabaix, Xavier. 2016. "Power laws in economics: An introduction." *Journal of Economic Perspectives* 30 (1): 185–206.

Gardini, Aldo, Enrico Fabrizi, and Carlo Trivisano. 2022. "Poverty and Inequality Mapping Based on a Unit-Level Log-Normal Mixture Model." *Journal of the Royal Statistical Society Series A: Statistics in Society* 185 (4): 2073–2096.

Gibrat, Robert. 1931. *Les inégalités économiques*.: Librairie du Recueil Sirey.

Lopez, J Humberto, and Luis Servén. 2006. "A Normal Relationship? Poverty, Growth, and Inequality." Policy Research Working Paper 3814, World Bank.

Lubrano, Michel, and Abdoul Aziz Junior Ndoye. 2016. "Income inequality decomposition using a finite mixture of log-normal distributions: A Bayesian approach." *Computational Statistics Data Analysis* 100: 830–846.

Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *The Review of Economic Studies* 60 (3): 531–542.

Neal, Derek, and Sherwin Rosen. 2000. "Theories of the Distribution of Earnings." In *Handbook of Income Distribution*, vol. 1, 379–427: Elsevier.

Pareto, Vilfredo. 1896. *Cours d'économie politique*.: Droz.

Weiss, Yoram. 1972. "The Risk Element in Occupational and Educational Choices." *Journal of Political Economy* 80 (6): 1203–1213.