KADIR HAS UNIVERSITY
SCHOOL OF GRADUATE STUDIES
PROGRAM OF MASTER OF ARTS IN ECONOMICS

CORRECTING DOWNWARD BIAS IN INEQUALITY ESTIMATES FOR TURKEY WITH HOUSE PRICE DATA

FIRAT ÇAĞLAR KARABULUT

MASTER OF ECONOMICS THESIS

ISTANBUL, JULY, 2024

CORRECTING DOWNWARD BIAS IN INEQUALITY ESTIMATES FOR TURKEY WITH HOUSE PRICE DATA

FIRAT ÇAĞLAR KARABULUT
SUPERVISOR: PROF. DR. HASAN TEKGÜÇ

A thesis submitted to
the School of Graduate Studies of Kadir Has University
in partial fulfilment of the requirements for the degree of
Master's Degree in
Economics

Istanbul, July, 2024

APPROVAL

This thesis titled CORRECTING DOWNWARD BIAS IN INEQUALITY ESTIMATES FOR TURKEY WITH HOUSE PRICE DATA submitted by FIRAT ÇAĞLAR KARABULUT, in partial fulfillment of the requirements for the degree of Master of Arts in Economics is approved by

Prof. Dr. Hasan Tekgüç (Advisor) ……..
Kadir Has University

Asst. Prof. Ulaş Karakoç ……..
Kadir Has University

Assoc. Prof. Murat Koyuncu ……..
Boğaziçi University

I confirm that the signatures above belong to the aforementioned faculty members.

_____

Prof. Dr., Mehmet Timur Aydemir
Director of the School of Graduate Studies
Date of Approval: 08.07.2024

DECLARATION ON RESEARCH ETHICS AND PUBLISHING METHODS

I, FIRAT ÇAĞLAR KARABULUT; hereby declare

- that this Master of Arts that I have submitted is entirely my own work and I have cited and referenced all material and results that are not my own in accordance with the rules;
- that this Master of Arts does not contain any material from any research submitted or accepted to obtain a degree or diploma at another educational institution;
- and that I commit and undertake to follow the "Kadir Has University Academic Codes of Conduct" prepared in accordance with the "Higher Education Council Codes of Conduct".

In addition, I acknowledge that any claim of irregularity that may arise in relation to this work will result in a disciplinary action in accordance with the university legislation.

Fırat Çağlar Karabulut

_____

Date (08/07/2024)

To My Dearest Ceyda,

For your unwavering support and love.

ACKNOWLEDGEMENT

I would like to acknowledge my esteemed supervisor Hasan Tekgüç, who adds value to my intellectual knowledge day by day, and my esteemed friends Halit Güzelsoy, Hüseyin Emre Almaz, Yunus Kısa, who helped me collect data for this study and kept monitoring my study and constantly sharing their ideas.

I also thank to ***The Scientific and Technological Research Council of Turkey*** *(TÜBİTAK, Grant ID: 122G241) for supporting me during my studies*.

CORRECTING DOWNWARD BIAS IN INEQUALITY ESTIMATES FOR TURKEY WITH HOUSE PRICE DATA

ABSTRACT

Household surveys often fail to capture incomes of top earners. Top earners are less likely to respond to surveys and more likely to not answer questions concerning entrepreneurial income, i.e. the "missing rich" problem. The most common method in the literature to solve these problems is the use of data obtained from income tax records. Using tax records for developing countries is problematic in two respects: (i) even summary measures of tax records are not publicly available in most of these countries and (ii) tax evasion is rampant and official tax records are probably not reliable even if they were available. Therefore, in this study, the "missing rich" problem is corrected by using house price data obtained from www.sahibinden.com to estimate top incomes. We estimate household incomes using house prices and append these estimates to the right tail of survey data. We estimate that the Household Budget Survey undercounted approximately 5 percent of total households in 2019. When the missing rich households are included, top 5% (10%) income shares increase from 17% (27%) to 28% (40%). As a result, the Gini index of household income in Turkey has increased from 35.3 to 50.

Keywords: inequality, Turkey, house prices, top incomes, income inequality, income distribution

TÜRKİYE İÇİN EŞİTSİZLİK TAHMİNLERİNDEKİ AŞAĞI YÖNLÜ YANLILIĞIN KONUT FİYATLARI VERİLERİYLE DÜZELTİLMESİ

ÖZET

Hanehalkı anketleri genellikle en çok kazananların gelirlerini yakalamakta başarısız olmaktadır. En çok kazananların anketlere yanıt verme olasılığı daha düşüktür ve müteşebbis gelirine ilişkin soruları yanıtlamama olasılıkları daha yüksektir, yani "kayıp zengin" sorunu. Bu sorunları çözmek için literatürdeki en yaygın yöntem, gelir vergisi kayıtlarından elde edilen verilerin kullanılmasıdır. Gelişmekte olan ülkeler için vergi kayıtlarının kullanılması iki açıdan sorunludur: (i) bu ülkelerin çoğunda vergi kayıtlarının özet ölçümleri bile kamuya açık değildir ve (ii) vergi kaçakçılığı yaygındır ve resmi vergi kayıtları mevcut olsa bile muhtemelen güvenilir değildir. Bu nedenle, bu çalışmada "kayıp zengin" sorunu, üst gelirleri tahmin etmek için www.sahibinden.com adresinden elde edilen konut fiyatı verileri kullanılarak düzeltilmiştir. Hanehalkı gelirlerini konut fiyatlarını kullanarak tahmin ediyoruz ve bu tahminleri anket verilerinin sağ kuyruğuna ekliyoruz. Hanehalkı Bütçe Anketi'nin 2019 yılında toplam hanelerin yaklaşık yüzde 5'ini eksik saydığını tahmin ediyoruz. Kayıp zengin haneler dahil edildiğinde, en üst %5'lik (%10) gelir payları %17'den (%27) %28'e (%40) yükselmektedir. Sonuç olarak, Türkiye'de hanehalkı gelirinin Gini endeksi 35,3'ten 50'ye yükselmiştir.

Anahtar Sözcükler: eşitsizlik, Türkiye, konut fiyatları, en yüksek gelirler, gelir eşitsizliği, gelir dağılımı

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

y	Household Income
τ	Income threshold where the right tail (Pareto Distribution) starts (Lower bound of upper income group)
λ	the ratio of households with income above threshold to the population
s	the income share of upper income group in total income
x	predictor (proxy) of household income
β_0	constant term of income prediction equation
β_1	slope coefficient of income prediction equation
$n(\tau, y)$	the number of households with income between τ and y
$n(\tau)$	the number of households with income exceeding τ
n	the total number of households in the population.
α	Pareto coefficient of house price distribution
θ	Pareto coefficient of household income distribution
γ	Inverse Pareto coefficient
λ_{prop7}	λ calculated by the Proposition 7 in Weide et. Al (2018)

LIST OF ACRONMYMS AND ABBREVIATIONS

CPI	Consumer Price Index
DB - [1/2]	Database 1/2
HBS	Household Budget Survey
NA	National Accounts
OLS	Ordinary Least Squares
pdf	Probability density function
PPI	Producer Price Index
PS	Piketty and Saez (2003)
TURKSTAT	Turkish Statistical Institute

1. INDRODUCTION

Household surveys are the main data source of inequality estimates. However, due to a critical problem referred to in the literature as the "missing rich" problem, it becomes almost impossible for the upper tail to provide a realistic representation of the income distribution with the income directly obtained from the surveys (Lustig, 2020). The main indicator that makes us think that the surveys suffer from this problem is that when national accounts data, which capture the total income of a country's household sector, are compared with the total income reported in the surveys, there is a serious understatement in total income to the detriment of the survey. The fact that the difference in total income is largely due to income types such as mixed income (obtained from business ownership) and capital income (interest, dividends, rent) generally obtained by upper income groups confirms our suspicions. Rich households are more likely to earn income through these channels, meaning that they are probably underrepresented in surveys.

There are many reasons for the “Missing rich” problem related to survey structure. The first of these is undercoverage. This means that wealthy individuals or households are systematically excluded from the survey sample, perhaps due to difficulties in reaching them or their low propensity to participate. The second is sparseness. Sparseness means that the number of high earners in the sample is not sufficient to provide statistically reliable estimates for income brackets. The third is non-response bias, which occurs when rich households do not fully respond to surveys (unit non-response) or do not answer some questions about income (item non-response). Fourth, underreporting of income may be intentional, and households may strategically conceal all or part of their wealth or the income they receive through their wealth (underreporting). Finally, top-coding practices, which limit survey responses that exceed a certain threshold, further distort the upper tail of the income distribution (top-coding) (Lustig, 2020: 6-10). The cumulative effect of these problems is a downward bias in all inequality estimates from household surveys. This undermines the effectiveness of policy analyzes and simulations based on accurate data on income distribution. For example, a researcher measuring the true impact of

taxation on high-income earners will arrive at biased results due to the "missing rich" problem. This will cause policy makers and policy analysts to implement wrong policies and not achieve the expected results from the implemented policies. As a result, correcting the problem of the "missing rich" is necessary to ensure the reliability of data on income inequality and to develop effective policies aimed at reducing economic inequalities.

Two methods are especially preferred by researchers to correct the "missing rich" problem. These include reweighting the sample and combining survey data with tax data.

Sample reweighting suggests adjusting the weights assigned to individual survey responses within the sample. The goal is to increase the weight of data points representing high-income households, effectively increasing their representation in the final analysis. However, this approach has many limitations. It cannot solve problems such as item non-response, which occurs when wealthy individuals choose not to disclose certain sources of income and underreporting. Moreover, the effectiveness of reweighting depends on the availability of reliable ancillary information about the population. For reweighting to be statistically sound, statistical agencies must make publicly available response rates by primary sampling units (geographical clusters used in survey design). This transparency allows researchers to evaluate potential biases introduced during the sampling process.

Another strategy is to integrate information from household surveys with tax data. Tax records ideally provide a more comprehensive picture of individual income, including capital gains and business profits that may be underreported in surveys. However, the success of this approach largely depends on the quality and accessibility of tax data. Tax authorities must be willing to publish detailed breakdowns of income distribution within tax brackets. Additionally, the tax data itself must accurately reflect the real income of rich households. In many developing countries, there are concerns about the reliability of tax records due to widespread tax avoidance and tax evasion practices facilitated by a phenomenon often described as "forbearance"[1] – the deliberate tolerance of governments for tax violations (Holland, 2016: 233). Forbearance can limit the effectiveness of this

[1] Holland (2016; 233) defines forbearance as 'intentional and revocable government leniency toward violations of law'.

approach by creating situations where tax records understate the real incomes of rich households.

In Turkey, there has been an increase in public interest in income inequality, especially in recent years. As of the last quarter of 2021, unbalanced growth models based on monetary expansion despite high inflation have been implemented. There is a decrease in the wage share from 2020 onward. The decline in wage shares has accelerated due to the monetary expansion policies since the last quarter of 2021. According to TURKSTAT (2024), the annual consumer inflation in September was 85% in 2022 and 64% in 2023. This followed the start of the monetary expansion policy in September 2021. Before the policy was initiated, the year-on-year increase in CPI in September was 20%. The inflation figure for the year before was 12%. Therefore, very high inflation rates were reached after the policy was implemented (TURKSTAT, 2024). Since the annual PPI was twice the annual CPI in the same period and the public felt the cost of living at a higher rate than the announced inflation, different researchers have calculated much higher inflation rates. The accuracy of these studies is debatable, but price increases in basic goods such as food, housing, furniture and transportation were higher than the general inflation rate. Although minimum wage increases have been realized at the annual rate of general inflation, the higher cost of living in basic living goods makes it difficult for lower income groups to keep pace with rising living costs. In addition, wages above the minimum wage were not increased by the minimum wage. This means that middle- and upper-middle income households, whose incomes include a high share of wages, have seen their share of total income decline. Since the negative real interest rate facilitates borrowing, it has led to large increases in the prices of financial products whose prices are negatively correlated with interest rates. Households with large amounts of wealth and the ability to borrow at negative real interest rates experienced a significant increase in wealth. Wealth is therefore thought to have become increasingly concentrated among an elite minority.

In Turkey, while the government has not made a direct statement that it fully acknowledges concerns about income inequality, policy proposals on the government's agenda suggest that this problem has been acknowledged, albeit implicitly. For example, the Motor Vehicle Tax, which can be categorized as a wealth tax, has been doubled for a

one-off increase in 2023. In early 2024, there were discussions on a property tax covering citizens with more than one property. These actions lead us to believe that the government tacitly acknowledges the problem. Paradoxically, even though traditional inequality estimates based on household surveys show a slight increase, these do not reflect the magnitude as expected. This discrepancy can be attributed, at least in part, to the problem of the “missing rich” and the limitations of household survey data, which fail to capture the full spectrum of income sources, especially for rich households. Therefore, it is seen as a necessity to eliminate the "missing rich" problem in official data as much as possible. Unfortunately, attempts to correct these problems using traditional methods such as sample reweighting or combining survey and tax data are not very feasible in Turkey. Relevant authorities do not publish important data sets such as non-response rates for household surveys and detailed breakdowns of taxable income distribution.

This study aims to solve the "missing rich" problem for Turkey to improve the accuracy of further research on income inequality. For example, our method allows for estimating income taxes payable by missing rich and estimate total of avoided income taxes. In this context, house prices will be used as a proxy to estimate the income of the upper income group, which is suspected to be missing in the 2019 Household Budget Survey (HBS). The dataset we will obtain as a result of this study will also allow us to calculate the level of income tax that households should pay. Therefore, one of the objectives of this study is to calculate the income tax evaded by top income earners. In order to make this calculation accurately, it is necessary to know how much tax is paid by those who are not in the top income group, which is missing in the surveys. Tekgüç & Eryar (2023) calculated this data using the HBS. The availability of this dataset, which includes these estimates, makes the HBS attractive for this study. In addition, the study of Alverado et al. (2019), which also aims to address the missing rich problem in Turkey, was also conducted using HBS. This will make our results more easily and reliably comparable to the existing literature. The year 2019 is preferred because it is the period just before the COVID-19 pandemic, when the relative stability of both the real estate market and the income distribution deteriorated. Thus, the income distribution we construct for 2019 will be an accurate dataset for that year that can be compared with Ceritoglu et al. (2023) which reports the findings of first of its kind wealth and inequality study which over-samples more expensive neighborhoods to catch more rich households.

This research proposes a new approach to address the limitations of existing data and provide a more nuanced picture of income inequality in Turkey. We leverage the power of big data by leveraging Turkey's developed market economy, particularly the presence of an extensive online real estate market. Following the methodology established by Weide et al. (2018), we collect house price data from the largest online real estate platform in Turkey (sahibinden.com). Weide et al. (2018) show that there is a correlation between housing prices and household income and that housing prices (rent) can be used as a predictor of household income. This allows us to estimate the income levels of the richest segments of the population who are underrepresented in traditional surveys.

Our contributions are threefold. First, we use house price data from Turkey's leading online marketplace to produce income estimates specifically for the richest households in the country. Second, we establish a link between our income estimates and the observed discrepancy between national accounts data and survey totals. Third, we verify the robustness of our findings through comprehensive sensitivity analyzes that explore the impact of different cut-offs on the house price distribution.

Our approach provides significant corrections to inequality measures in Turkey. By including units from house price data, the Gini coefficient, a standard measure of income inequality, increases from 0.353 to 0.5. This significant increase indicates a more unequal income distribution after accounting for the previously underrepresented “missing rich.” Similarly, the income share of the top 5% (10%) with the highest income increases from 17% (27%) to 28% (40%). These findings highlight the potential magnitude of the "missing rich" problem in Turkey and underline the importance of our methodological approach in providing a more comprehensive picture of income inequality.

The rest of the thesis is organized as follows. Section 2 reviews the literature on approaches to income inequality correction of Gini estimates and discuss their shortcomings in detail. Section 3 of this thesis will describe the methodology used in this thesis. Section 4 introduces the datasets used and Section 5 presents the empirical results. Section 6 discusses the results and Section 7 concludes with a summary of the thesis.

2. LITERATURE REVIEW

2.1 Factors Determining Economic Inequality

There are many factors that cause economic inequality. The existence of a market economy can be seen as one of the direct causes of inequality (Milanovic, 2016; Piketty, 2014). The fact that all households have different formations on access to different income sources such as labor income, capital income, and rent income, which we can define as sources of income in a market economy, can be seen as the cause of economic inequality (Piketty, 2014; Corak, 2013). It should be considered that the homogeneous distribution of these income sources on a micro scale is impossible within the nature of the market economy (Piketty, 2014; Milanovic, 2016). However, the market economy and its own dynamics alone are not enough to explain inequality in today's societies (Stiglitz, 2012; Piketty, 2014).

There are also different economic factors used to explain current economic inequality. The first of these is technological change. The high rate of technological development creates deep differences in the quality of labor in society that can and cannot access to these technologies (Piketty, 2014). Those who have the ability to use high technology can earn higher labor income than those who are less skilled (Piketty, 2014; Fujita, 2023). Secondly, the fact that the return on capital (r) exceeds the economic growth rate (g) is one of the most important factors because it deepens wealth inequality in favor of capital owners (Piketty, 2014). Considering that wealth is also one of the most important sources of income, this situation also increases income inequality (Piketty, 2014; Milanovic, 2016).

There are other factors that cause economic inequality. The first of these is social and demographic factors. The fact that children from high-income households have access to more qualified educational opportunities, have better networks and job opportunities, and marriages are usually between individuals with the same socioeconomic status are examples of social and demographic factors that affect economic inequality (Stiglitz, 2012; Milanovic, 2019). Cultural and ideological beliefs stand out as contributing to the

maintenance of inequality. For example, the income gap between the rich and the poor is associated with the rich being more hard-working and more productive (Benabou & Tirole, 2006; Hacker & Pierson, 2010). The establishment of this belief system in society leads to the belief that restricting the accumulation of wealth in various ways will harm the economy (Hacker & Pierson, 2010). This is used as a tool to legitimize the non-implementation of policies that restrict the accumulation of wealth and to prevent such policies (Milanovic, 2016).

At this point, along with economic factors, political factors, which are among the most important factors, emerge. As Stiglitz (2012) points out, inequality is not just the result of economic forces but significantly driven by policy choices. Tax policies are generally designed in favor of wealthy households (Milanovic, 2016; Stiglitz, 2012). Wealth and capital income tend not to be taxed or to be taxed at very low rates (Stiglitz, 2012; Piketty, 2014). This situation causes households with more wealth to have more income and the increase in wealth concentration at the top (Fujita, 2023; Hacker & Pierson, 2010). Thanks to these policies, the rich can accumulate more wealth, preserve it and have the advantage of transferring it across generations (Milanovic, 2016; Benabou & Tirole, 2006). In this way, inequality becomes more permanent (Hacker & Pierson, 2010). The state's increasing withdrawal from public services such as education and health, which mostly benefit the middle and lower segments, also causes economic inequality to become worse (Stiglitz, 2012). The effect of money on politics complicates the problem even more (Milanovic, 2016; Stiglitz, 2012). In other words, economic inequality triggers political inequality, and political inequality triggers economic inequality, creating a cycle of increasing inequality (Hacker & Pierson, 2010; Stiglitz, 2012). For this reason, there is a general belief in the literature that inequality is a policy-driven issue (Stiglitz, 2012; Piketty, 2014).

2.2 Understanding Economic Inequality in a Global Context

There are different ways to observe economic inequality globally. While China's high growth rates since the 1980s, followed by India, Vietnam, Thailand, Indonesia and other highly populated countries in Asia, seem to have reduced inequality on a global scale,

inequality has increased in many countries since the adoption of neoliberal policies (Milanovic, 2016; Milanovic, 2019).

Inequality varies considerably across both developed and developing countries. While income inequality is generally less dramatic in developed countries, significant increases in inequality have been observed in recent years in countries such as the United States and the United Kingdom (Piketty, 2014). For example, the fact that the top 1% in the United States control an increasingly larger share of total wealth is an important indicator of increasing economic inequality (Saez & Zucman, 2016). Conversely, in some developed economies, such as the Scandinavian countries, which have very low inequality due to social welfare policies, inequality is also increasing due to the gradual deterioration of these policies (Pareliussen, Hermansen, André, & Causa, 2018).

The situation in developing countries is slightly different and shows more variation. Historically, higher Gini coefficients have been observed in underdeveloped and developing countries. In particular, Latin American countries such as Brazil, Colombia and Mexico have had the highest levels of income inequality globally (Makhlouf 2023). At the same time, there is no common trend for developing countries over the past 60 years (Makhlouf 2023). The reasons for this difference are related to almost all of the factors mentioned earlier, but the main difference is due to differences in economic and political preferences. In countries with more progressive taxation and redistributive policies, inequality is reduced (Achcar, 2020). Inequality in developing countries, like in developed countries, has been increasing on average. However, there is more variation in trends in developing countries than in developed countries (Makhlouf 2023). For example, in Eastern European countries, high growth rates are accompanied by rising inequality indicators, while in Latin American countries, inequality has shown declining trends thanks to inequality-reducing policies (Makhlouf 2023).

2.3 Understanding Economic Inequality in a Historical Context

Simon Kuznets addressed inequality empirically for the first time in history. Using US tax records, Kuznets (1953) argues that inequality initially increased due to the shift from agricultural to industrial production, the concentration of savings among the high-income

population and the increase in migration from rural to urban areas. However, according to him, after a certain stage of development, as the industrial sector matures and education becomes more widespread, all segments of the population will begin to benefit more equally from economic growth and inequality will decrease gradually. In other words, there is an inverted U-shaped relationship between inequality and growth. This relationship is called the Kuznets Curve. Unlike Keynes, Kuznets describes this process as a more spontaneous process and does not emphasize the importance of policies.

Efforts to understand income inequality in its historical context and to discuss and analyze its causes and consequences in detail have been reignited by Piketty & Saez (2003)'s attempt to extend Kuznetz's work. The authors are surprised why this seminal work has not been attempted to be extended earlier in these 50 years. This paper differs from Kuznetz (1953) in that it also compares the income shares of other countries. They find significant similarities in the trends of top income shares over the years in France, the United States, and the United Kingdom. In all three countries, the income shares of the top 0.5% fell significantly between 1914 and 1945 and did not return to the very high levels observed on the eve of World War I until 1998 (Piketty & Saez, 2003). Piketty & Saez (2003) interpreted this as the effect of the progressive tax system following the war and depression. However, their findings have faced substantial critique. Geloso et al. (2022), sharing the updated version of Piketty and Saez's (PS) 2003 study, criticized this study with a new methodology and created a new inequality series for US. Geloso et al. (2022) argue that PS exaggerates the levels of income inequality and the extent of the decline during the Second World War. They argue that the Great Depression was more important than the war in reducing inequality. Focusing more on the 1920-1960 estimates, they argue that the decline in inequality between 1930 and 1942 was smoother, not as sharp as in PS's estimates. In addition, despite lower income inequality and smoother trend changes in inequality, there is not much difference in the direction of the trends. A similar criticism of PS's methodology comes from Auten & Splinter (2023). In contrast to the series of Geloso et al. (2022), which ends in 1960, Auten & Splinter (2023) starts from 1960 when presenting their findings in a single graph with Geloso et al. (2022), it makes Auten & Splinter (2023) look like a continuation of Geloso et al. (2022). Auten & Splinter (2023) find an increase in inequality after 1960, like PS. There is no disagreement on the direction of the current inequality trend. However, Auten & Splinter (2023) claim

that "missing income" should be distributed more evenly to lower income groups, unlike the PS methodology. As a result, they argue, the share of upper income groups in total income increased after 1960, but this increase was much smaller than PS claims. When considering these critiques, the post-1920 U-shaped income inequality trend proposed by PS seems to be correct when the critical calculations of Geloso et al. (2022) and Auten & Splinter (2023) are considered. However, the U-shaped trend looks much smoother in these two critical calculations. Moreover, in Auten & Splinter (2023)'s calculations for after-tax, the inequality trend becomes much flatter between 1962 and 2014. The main reason for the difference in the calculations is related to how "missing income" is distributed. Auten & Splinter (2023) argue that unobserved income is much more evenly distributed over time. Piketty et al. (2023) argues that Auten & Splinter's (2023) (AS) assumptions about the distribution of untaxed or unobserved income are unrealistic and lead to wrong conclusions. They emphasize that all observable evidence shows that income and wealth inequality are increasing. For example, the share of financial income (excluding capital gains) earned by the top 1% rose from 8.4% in 1960 to 17.6% in 2019 (Piketty et al., 2023). AS assumes that while 55% of taxed business income is concentrated in the top 1%, only about 15% of untaxed business income is concentrated in the top 1% (Piketty et al., 2023). Piketty et al. (2023) claim that a similar picture for capital income. These assumptions do not seem to be consistent with the observable evidence. Considering the growing wealth inequality that almost everyone has acknowledged since the 1980s, these assumptions are highly unlikely to be realistic. Thus, Piketty et al. (2023) conclude that the assumption made by AS about the unobservable distribution of business and capital income is both inconsistent with observable evidence and logic. The take-away message from this US centric literature is that the assumptions about the distribution of missing incomes to overall income distribution is crucial even when tax data is available.

2.4 Income Inequality and Correction Methods for Measurement Problems

2.4.1 Reweighting higher income households

One of the main focal points of economic inequality is the measurement of inequality in income distribution. For this purpose, the most used indicator is the Gini coefficient. Although other indicators are also used, the share of top income earners in total income is one of the most used indicators. Income distribution and income inequality are usually measured using household surveys. These surveys are the best tools for measuring the income of households at the micro level. However, for the reasons explained above, the surveys do not represent the upper income groups well enough, which affects the accuracy of the results in inequality measurements. To solve this problem, researchers have had to develop many different techniques. Many techniques have different strengths and weaknesses. Which of these techniques is preferred depends on the purpose and data availability.

There are two different main approaches to solving missing rich problem as mentioned above. The first of these approaches is to reweight household surveys with some intra-survey calculations. The aim of the reweighting method is to increase the representation of top income earners in the survey by increasing the weights of high-income households among the surveyed units. In general, this method proposes to find a relationship between income and non-response rate through various demographic and geographic data and to reweight the survey data according to this relationship. For example, if the non-response rate is higher in wealthier neighborhoods (districts) - location data is needed - the weight of households residing in these neighborhoods can be increased relative to poorer neighborhoods. Although such a method has the potential to find a solution to the "missing rich" problem, it has some limitations. The biggest limitation of this method is that it assumes that the problem of underreporting and item non-response bias does not exist. It ignores the possibility that a reweighted high-income household may underreport its total income even though its weight in the population has been increased. The underreporting problem is likely to be one of the most important reasons for the underestimation of survey incomes of top income earners. For this reason, the results obtained with this approach can only be considered as a lower bound benchmark. One of

the most important problems of this method is data availability. It requires as much demographic data as possible, as well as response (or non-response) data by primary sampling units (geographical clusters used in the survey design). The statistical agencies of most countries, including Turkey, do not publish this data. This makes the applicability of this method limited. Another problem is that the method makes various assumptions about the relationship between auxiliary variables and income levels. Although these assumptions seem to eliminate the "missing rich" problem, they may lead to different biases. Finally, since the observations in the survey are reweighted, it cannot solve the problem of sparseness of the extremely rich households that are not included in the survey because they are extremely unlikely to be included in the survey. To summarize, trying to solve the "missing rich" problem in household surveys with the re-weighting method does not solve the problem completely, but it creates a good potential for solving part of the problem. However, the effectiveness of the method depends on the availability and quality of auxiliary data and the transparency with which statistical agencies provide data on survey design. Moreover, it cannot solve all the problems of under-representation of top income earners in surveys.

2.4.2 Combining surveys and tax records

The second approach is to use a different data set in addition to the survey data, which is thought to better capture the upper tail. Income tax data consists of detailed reports of individuals to tax collectors. These data are a much more efficient source of income data, especially for top-earners, because in developed countries the tax reports of high-income individuals are scrutinized in detail and penalties are imposed for underreporting. Assuming that tax records and household surveys can be used to identify well where household surveys are likely to be underreporting, various mathematical and statistical methods can be used to address the right tail problem. In the literature, it is generally assumed that the right tail is distributed according to the Pareto distribution. Under this assumption, correcting the right tail with the Pareto interpolation method in the light of the data obtained from tax records stands out as the most common method. Alvaredo (2011) suggests that when the total income share of the 1% with the highest income is corrected according to tax records data, the Gini Coefficient increases from 59 to 62 for the USA. Yonzan et al. (2022) showed that surveys have been able to capture the income

of the richest 1% according to tax records in recent years, by 50-60% in the USA and 57-59% in Germany. According to Yonzan et al. (2022), although survey data and tax records are sometimes compatible in France, in some years the survey data was not successful enough to determine the income of the richest 1%. Jenkins (2016) shows that upper tail income, calculated based on household survey and tax records data and household survey data for the United Kingdom, reflects 77% of the income calculated on the tax records. However, in less developed and developing countries, tax records data is less reliable. With 2010 data, Alvaredo and Londoño-Velez (2013) found that the average income of the top 1% in Colombia was 50% higher in tax data than in surveys. In other words, surveys estimate the income of the highest income group by approximately 67% compared to their tax records. On the other hand, the Brazilian case is more striking, Morgan (2018) found that in 2015, the income share of the top 1% with fiscal records was 22.5%, while the share of the top 1% reported in the survey was only 10.2%. Again, Surveys show that the income of this upper income group reflects approximately 45% of what appears in tax records. In underdeveloped and developing countries, surveys capture the income of the upper income group at a much lower rate than in developed countries, according to tax records. However, in some developing countries, despite adequate technological resources and staff for tax collection, citizens with entrepreneur income, rental income, etc. are shown leniency when they declare their incomes too low. Holland conceptualizes this leniency as “forbearance”. Tekgüç and Eryar (2023: 19-20) discuss the implications of forbearance in Turkish context in detail. Additionally, in many developing countries, including Egypt and Turkey, tax records are not shared with researchers. For this reason, it is impossible to use this method even under the assumption that it is the most effective method for developing and less developed countries.

2.4.3 Combining surveys with alternative data sources where tax data is unreliable/unavailable

The problematic and in some cases impossible use of tax records in developing and underdeveloped countries has encouraged researchers to develop new methods. The most prominent of these is to re-estimate income inequality by using house price data as an estimator of the income of the upper income group to solve the right tail problem in household surveys (Weide et al. 2018). Using house price data as an income predictor

seems to be a good solution, as high-income households usually reside in houses with higher rental values. Using real estate prices as a proxy for income may provide a better picture of income distribution than that reported in top-tail household surveys and even better than estimates of income distribution in less developed countries adjusted for tax records data. In countries where the tax system is not well-established and tax evasion is widespread for various reasons, this method seems to be the most effective alternative. The results obtained by applying this methodology to Egypt dispel our doubts about its reliability and usefulness. In Egypt, the Gini Inequality Index, which was calculated as 0.385 with survey data, increased to 0.518 when calculated with this method (Weide et al. 2018).

Two recent studies, Alvaredo et al. (2019) and Cerioğlu et al. (2023), provide valuable insights for the Turkish case. While these two studies address the problem of the "missing rich" in the Turkish context, they are also important to see the complexities involved in the problem. Alvaredo et al. (2019) use tax registration data from Lebanon to estimate income inequality in Turkey (and the rest of the Middle East) by correcting for the "missing rich" problem of household surveys. Using the Pareto interpolation technique, this study finds a higher Gini coefficient for Turkey (between 0.56 and 0.61) compared to the official figures (around 0.40). This implies that income inequality in Turkey (and in other countries as well) is much higher than the household surveys suggest. While this study provides some support for our suspicions, it is problematic in several aspects. As mentioned earlier, it is very difficult to rely on the accuracy of tax registration data in developing and less developed countries. In addition, the external data used is a very inadequate source to reveal Turkey's income distribution. Using tax registry data from a different country to overcome the downward bias in income inequality due to the survey design leads to different biases. It also makes it difficult to interpret the direction of the bias. This may lead to certain inconsistencies in the interpretation of this study. Correcting all years using a single year's tax records data also implies assuming that the income distribution of the upper tail does not change over the years, resulting in a less accurate identification of income inequality trends. On the other hand, an important problem arises when we follow their method. The coefficients we derive from the Lebanese tax registry data change the ranking of households in the top quintile, leading to internal inconsistency (Tekgüç et al. 2024). Ceritoğlu et al. (2023), on the other hand, comes up with convincing

more direct method that have been encountered so far that can be applied not only for Turkey. Their main concerns are wealth ownership and household finances. They design a survey where they deliberately over-sample rich neighborhoods and they also use the clout of their affiliation, Turkey's Central Bank, to increase response rate. Their income inequality estimate is 0.517 for 2019. They also estimate that the top 5% (10%) wealth share is 42% (55.3%), and they estimate the Gini coefficient for wealth distribution as 0.773. One of the reasons why the Gini coefficients in these two studies are so different might be that Alvaredo et al. (2019) estimates are for per adult, whereas Cerioğlu et al. (2023) estimates are for total households. The main problem with this methodology is that it is probably very costly and cannot be applied to past years. However, the Gini coefficient for total household disposable income will be a good benchmark for understanding how accurate the methodology used in this study is. In conclusion, these two studies offer valuable contributions to the solution of the "missing rich" problem in Turkey and support our suspicions. At the same time, they also show how complicated the issue is.

Accurately measuring income inequality remains a complicated task in many countries, including Turkey, especially when it comes to capturing the incomes of the richest individuals. For Turkey, the popular and well-accepted methods as reweighting and correction with tax records cannot be used due to lack of data and poor data quality. In addition, in this study, even if the method of combining survey data with tax records is feasible, it is not preferred due to the forbearance problem. Besides, reweighting is not a well-performing top income estimation method due to its limitations. Thus, even if reweighting is possible, it cannot create more than a lower bound for researchers. In Turkey, concerns about the reliability of official data and the underrepresentation of high-income earners in surveys have led us to explore alternative approaches. Given all these issues, a more innovative approach to estimating the incomes of the richest households is the methodology of Weide et al. (2018), where house prices are used as a proxy for the income of the upper income group. This method is based on the strong correlation between the rent (price) of the house where households reside and their income. Households with higher incomes are more likely to prefer to live in more valuable houses. Therefore, house price data can potentially provide insights into the income distribution of the "missing rich" who are underrepresented in household surveys (Weide et al. 2018).

3. METHODOLOGY

3.1 Combining Income Survey with Top Income Data

The aim of this study is to address the upper tail problem of the income distribution (DB-1 in theoretical expressions) in the 2019 Turkey Household Budget Survey as mentioned above. It is assumed that the Household Budget Survey captures the income distribution well except for the upper tail but does not capture the upper tail of the income distribution sufficiently. To overcome this problem, an external dataset will be used. This data set is the house price data set collected from sahibinden.com (DB-2). Let us assume that $F(y)$ represents the cumulative distribution function where y represents household income, as in Weide et. al (2018). Where τ is the lower threshold of the income group that we will classify as the upper income group, λ will represent the ratio of households with income above τ to the population. Consequently, if we call this distribution $F_1(y)$ under the assumption that DB-1, i.e. HBS, correctly estimates the income distribution and hence $F(y)$ for the distribution below τ, then this distribution can be defined as $F_2(y)$ under the assumption that income from the house price dataset correctly captures the income distribution for the income group above τ. Assuming that λ can also be estimated correctly, we can write $F(y)$, the cumulative distribution function of the entire income distribution, as follows:

$$F(y) = \begin{cases} (1-\lambda)F_1(y), & y \leq \tau \\ (1-\lambda) + \lambda F_2(y), & y > \tau \end{cases} \qquad (3.1)$$

If the assumptions in equation 3.1 hold for the available data sets, the Gini coefficient for the entire distribution can be decomposed as follows. In this decomposition P_1 represents the proportion of the population with incomes below the threshold τ (also 1-λ), S_1 represents the share of those with incomes below the τ threshold in total income, and $Gini_1$ represents Gini coefficient within the first database. Similarly, P_2 represents the proportion of the population with incomes above the τ (also λ), S_2 represents the share of those with incomes above the τ in total income, and $Gini_2$ represents Gini coefficient within the second database. Equation 3.2 can be reduced to equation 3.3 if we define S_2 as s.

$$Gini = P_1 S_1 Gini_1 + P_2 S_2 Gini_2 + S_2 - P_2 \quad (3.2)$$

$$Gini = (1-\lambda)(1-s)Gini_1 + \lambda s Gini_2 + s - \lambda \quad (3.3)$$

This decomposition reveals the minimum number of parameters necessary to calculate a realistic Gini coefficient. $Gini_1$ is the Gini coefficient of income below τ from household survey. Similarly, $Gini_2$ is the Gini coefficient of income above τ from DB-2. The sum of income missing in the surveys can be estimated using house price data. Only the problem of estimating the parameters λ and s remains. How to estimate λ is described below. Once λ is estimated, to estimate s it is sufficient to estimate the mean of household income below τ in DB-1 and the mean of household income above τ in DB-2.

3.2 Calculating the Relationship Between House Prices and Household Income Using the Household Survey

The HBS does not include the price of the houses in which households reside. However, it does include data on how much rent households pay if they live in rented accommodation, and what their imputed rent is if they do not live in rented accommodation. In other words, the income of each household in the survey and the (imputed) rental value of the house they reside in are known. Using this data, the following relationship can be calculated:

$$\log(Y_h) = m(x_h;\beta) + \varepsilon_h = \beta_0 + \beta_1 \log(x_h) + \varepsilon_h \quad (3.4)$$

The index h in Equation 3.4 denotes households; Y_h denotes household income; and the variable x_h is the estimator of household income, i.e. house price. Similar to Weide et al. (2018), we estimate Equation 3.4 using (imputed) rent data, assuming that rent is proportional to the price of the house. Assuming a constant capitalization rate between house price and rent, the rental value of the house can be calculated using the house price in DB-2. Once the parameters in Equation 3.4 are estimated using HBS, the rent variable calculated by using DB-2 is converted into household income using these parameters.

3.3 Estimating the Proportion (Λ) of the Highest-Income Group (X%) of the Population That Surveys Fail to Identify

Assume household income can be defined as follows:
if

$$F_2(y) = Pr[Y < y | Y > \tau] \text{ and } \lambda = Pr[Y > \tau] \qquad (3.5)$$

$$F_2(y) = \frac{n(\tau,y)}{n(\tau)}, \qquad (3.6)$$

then,

$$\lambda = \frac{n(\tau)}{n} \qquad (3.7)$$

where $n(\tau, y)$ denotes the number of households with income between τ denotes the number of households with income between y; $n(\tau)$ denotes the number of households with income greater than τ; n denotes the total number of households.

The economic definition of λ , which is mathematically defined in Equation 3.7, can be made as the X% of the society with the highest income whose income is not captured well enough by the surveys. Estimating λ in this study is important both because it will be used in the Gini correction and to determine the ratio of the income group missing in the surveys to the total population of the society. Weide et al. (2018) show that the λ can be estimated by combining two data sets in a district with the following equation (Prop. 7):

$$\widehat{\lambda_d} = \frac{\widehat{f_{1,d}}(\tau)}{\widehat{f_{1,d}}(\tau) + \widehat{f_{2,d}}(\tau)} \qquad (3.8)$$

where the index d stands for the district.
Since the estimation of λ for Turkey is done for the whole population not for a district, the estimator becomes the following:

$$\hat{\lambda} = \frac{\widehat{f_1}(\tau)}{\widehat{f_1}(\tau) + \widehat{f_2}(\tau)} \qquad (3.9)$$

3.4 Estimating Income Inequality for Upper Income Group from House Price Dataset

Assuming that house prices also follow the Pareto distribution, we can write Equation 3.10:

$$G_2(x) = 1 - \left(\frac{x}{x_0}\right)^{-\alpha} \quad (3.10)$$

In Equation 3.10, x is the house price and x_0 is the house price threshold that satisfies the condition $Y > \tau$. α is the Pareto coefficient for house prices. If we assume that the above assumptions are correct, we can write Equation 3.11 for the income distribution $F_2(y)$ of the upper income group missing in the household surveys:

$$F_2(y) = Pr[Y \leq y | Y > \tau] = 1 - \left(\frac{y}{\tau}\right)^{-\theta} \quad (3.11)$$

Assuming that the income threshold τ is set high enough so that $Y > \tau$ satisfies $X > x_0$, we can assume that incomes greater than τ are also distributed according to the Pareto law.

In Equation 3.11, y is the household income, τ is the lower bound of the upper income group as in Equation 3.1, and θ is the Pareto coefficient for the income of the upper income group. Assuming the assumptions behind Equations 3.4, 3.10 and 3.11 are correct, $\theta = \alpha/\beta_1$. The parameter α will be obtained from Equation 3.10 and β_1 from Equation 3.4. Finally, we can estimate the average income of $F_2(y)$ by the inverted Pareto coefficient, γ.

$$E[Y|Y > \tau] = \gamma\tau = \left(\frac{\theta}{(\theta-1)}\right)\tau \quad (3.12)$$

4. DATA

As mentioned before, two different datasets are used in this study. These are the 2019 HBS dataset and the house price dataset obtained from sahibinden.com. HBS 2019 has 11546 observations. Using the weights provided by TURKSTAT with the data, this dataset corresponds to 24,270,586 households. When we weigh these observations using the household size, it can be said that the dataset represents a population of 80,868,501 people. The summary statistics for household disposable incomes are given in Table 4.1.

Table 4.1. Descriptive Statistics of Survey Disposable Incomes

Income Distribution	Group	Population	Mean	Total Inc.	Income Share
Household Total Disposable Income (HBS; 2019)	*Whole Population*	24,270,586	69,987 ₺	1,699 bn. ₺	100%
	top 1%	244,231	402,186 ₺	98 bn. ₺	6%
	top 5%	1,216,020	233,924 ₺	284 bn. ₺	17%
	top 10%	2,428,801	187,118 ₺	45 bn. ₺	27%
Household Disposable Income per capita (HBS; 2019)	*Whole Population*	80,868,501	21,004 ₺	1,699 bn. ₺	100%
	top 1%	808,683	148,002 ₺	120 bn. ₺	7%
	top 5%	4,051,773	85,276 ₺	345 bn. ₺	20%
	top 10%	8,087,072	66,288 ₺	536 bn. ₺	32%

4.1 Evidence on Household Survey Omitting the Rich

In Turkey, as in other countries, there are serious concerns that the household survey does not adequately cover the highest income households. Table 4.2 shows the proportional share of different income types in the national accounts. Labor Income and Government Transfers, which are mostly received by the lower income group, are represented by an

average of 91% and 94% respectively in the surveys, while mixed income and capital income, which we expect to be mostly received by the upper income group, are represented by an average of 37% and 11% respectively in the national accounts. This table therefore suggests that the upper income group is also underestimated in TURKSTAT's Household Budget Surveys.

Table 4.2. Comparison of Income Types in HBS and National Accounts in Different Years

Year	Income Type	HBS	National Accounts (NA)	HBS / NA	Overall Coverage
2011	Labor Income	349 bn.	372 bn.	94%	91%
2015		618 bn.	682 bn.	91%	
2019		1,188 bn.	1,342 bn.	88%	
2011	Government Transfers	115 bn.	130 bn.	88%	94%
2015		196 bn.	207 bn.	95%	
2019		423 bn.	425 bn.	100%	
2011	Mixed Income	107 bn.	296 bn.	36%	37%
2015		180 bn.	450 bn.	40%	
2019		284 bn.	785 bn.	36%	
2011	Capital Income	25 bn.	241 bn.	11%	11%
2015		41 bn.	443 bn.	9%	
2019		73 bn.	560 bn.	13%	

Source: Tekgüç and Eryar (2023).

Also estimated Pareto tail index from the survey is 1.4 which is very low (see section 5.2 for details).

4.2 Real Estate (House Price) Data

The house price dataset to be used for estimating the income of the upper income group was collected from sahibinden.com. In order to ensure that this dataset represents the upper income group well enough, it was first necessary to identify the wealthiest districts in Turkey. Since the Household Budget Surveys in Turkey are conducted over the course

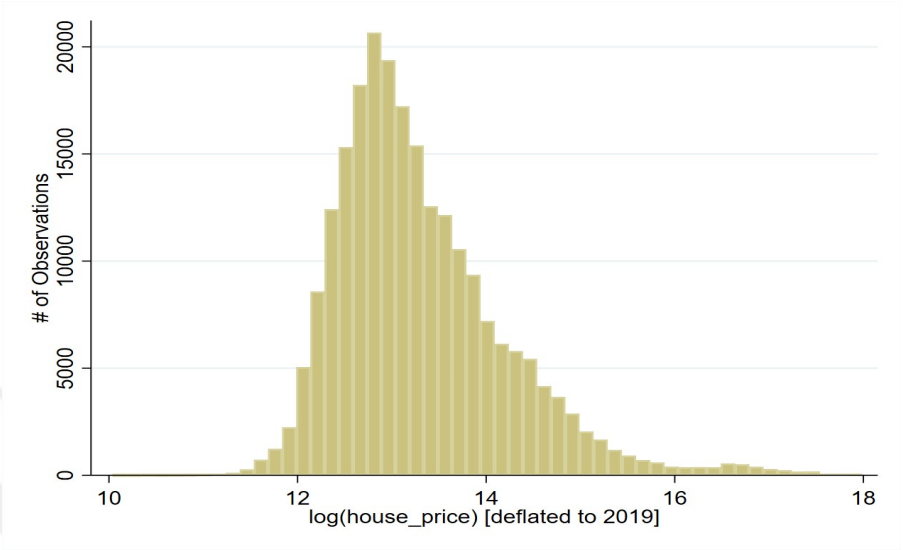
of a year, the data on m2 prices per district are collected from July of each year to represent mid-year, including January and February of 2023. These data consist of the Average House Price per Square Meter Index created by Emlak360 (Sahibinden.com). The m^2 price index was collected for each district with five years of data (2019 - 2023). As a result of this process, 73 districts with the highest average m2 prices of houses for sale were identified. In addition to these districts, the remaining 6 districts of Istanbul, the richest city in Turkey, were included, resulting in 79 districts, representing 25% of the total population (21.4 million).

Within these 79 districts, 228,132 house advertisements for sale data were collected, including the variables "Property Type, Advertisement Title, m² (Gross), Number of Rooms, Price, Advertisement Date, Province / District", which includes various information of all houses advertised for sale on Sahibinden.com. Using the average m2 price dataset for each district, a separate deflator was calculated for each district to deflate the house price data from 2023 to 2019. Using the average m^2 prices for sale and for rent in the average m2 price dataset, a separate capitalization rate was calculated for each district and the annual rent for that house was calculated by dividing the house price data by this coefficient The summary statistics of real estate database with deflated prices (rents) are given in the Table 4.3. In addition, the distribution of the logarithm of house prices in 2019 prices are given in Figure 4.1.

Table 4.3. Descriptive Statistics of House Prices in Real Estate Dataset

Variable	Observation	Mean
House Prices (2023)	228,132	9,193,339
House Prices (deflated to 2019)	228,132	1,224,698
Annual Rent (in 2019 Prices)	228,132	61,190

Figure 4.1. Frequency Distribution of Logarithm of House Prices (deflated to 2019)



5. EMPRICAL APPLICATION

In this section, we take the distribution of the income group excluding the upper income group from HBS (2019) (DB-1) and combine the upper income group with the income estimated from the real estate database (DB-2), which consists of the data we collect ourselves. In this study, we stick to the methodology of Weide et al. (2018) as much as possible. The assumptions of the model are described in detail in Weide et al. (2018)[2]. However, we had to make some additional assumptions to solve some of the problems we encountered or because we thought they would produce more accurate estimates. First additional assumption that is done for this research is the probability density function (pdf) of y, which is a continuous function of y $[f(y)]$, and, pdf's of $F_1(y)$ and $F_2(y)$, $[f_1(y)$ and $f_2(y)]$, can be estimated by Kernel Density Estimation. We create deflators of house prices for each district by average m^2 prices for 2019 and 2023 provided by sahibinden.com. The second additional assumption that we make is that deflated house prices by district from 2023 to 2019 represent the house price distribution in 2019. On the other hand, Weide et al. (2018) assumed in the house price dataset, each house represents one household (the correction will still be underestimated as upper-income households usually have more than one house for their own use). For robustness, the coefficients will be reconstructed under the assumptions that every 1.5 and 2 houses represent one household in this research.[3]

5.1. Pareto Tail Index Estimated on Income Survey Data

First, we did for Turkey what Weide et al. (2018) did for Egypt using the HBS. To calculate the Pareto Tail Index, we use Equation 3.9 to arrive at the following equation:

$$1 - F_2(y) = \left(\frac{y}{\tau}\right)^{-\theta} \qquad (5.1)$$

If we take the natural logarithm of both sides, the equation looks like this:

$$\log(y) = \log(\tau) - \frac{1}{\theta}\log(1 - F_2(y)) \qquad (5.2)$$

[2] A short list of the assumptions we use can be found in the Appendix A.
[3] Please see Appendix B for robustness check.

If the upper tail of the income distribution is Pareto distributed, then $\log(y)$ should be linearly distributed with slope $\frac{1}{\theta}$ against $(1 - F_2(y))$. Figure 5.1. shows the plot of the distribution of the right tail of $\log(y)$ on $\log(1 - F_2(y))$, calculated using total household disposable income and per capita disposable household income from the HBS.

Figure 5.1. The Pareto Quantile Plot for Household Disposable Incomes (Household Survey)

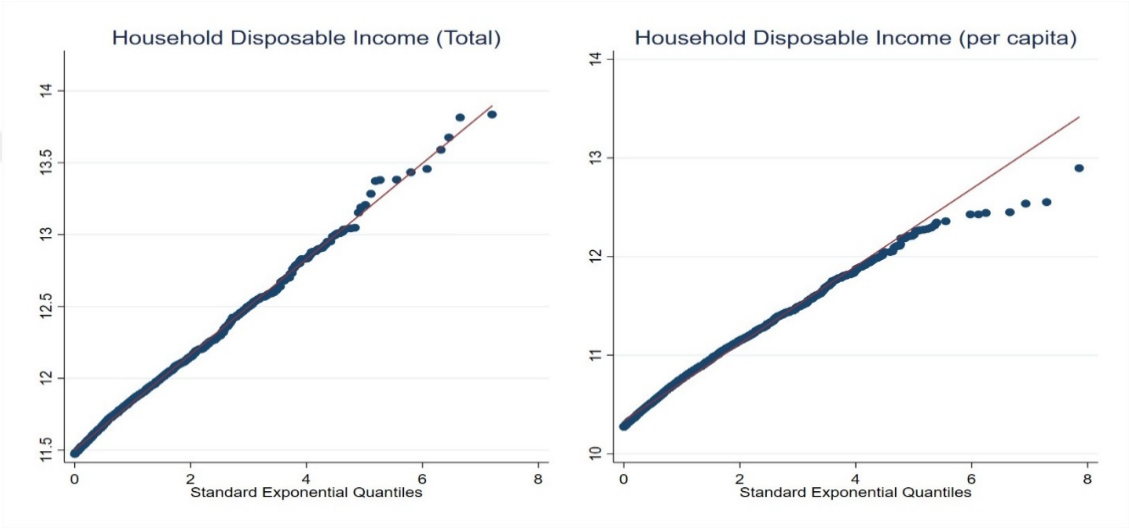
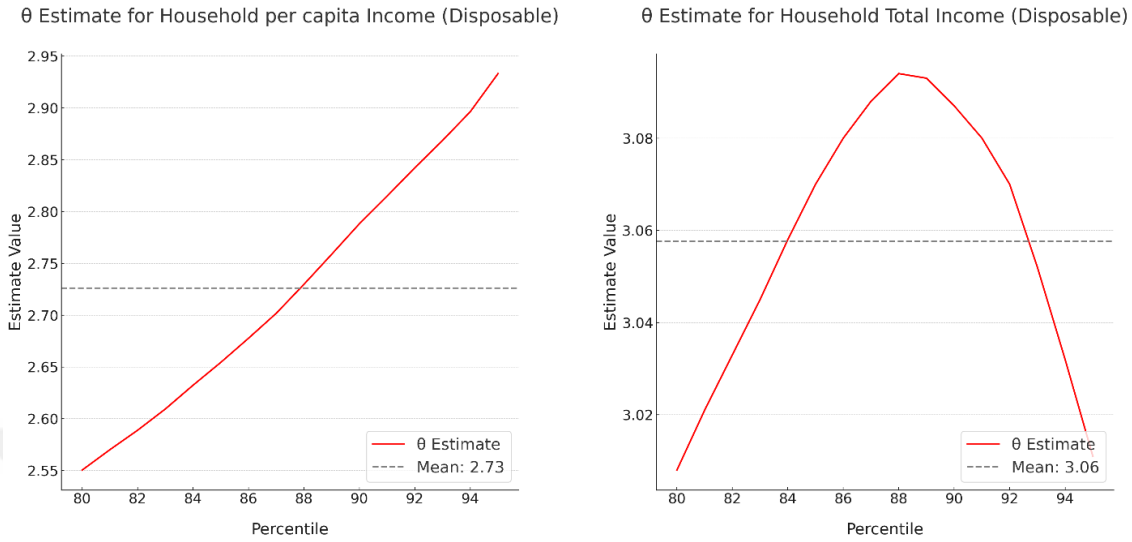


Figure 5.2. shows the OLS-estimated values of θ for each percentile (up to the top 5%) starting from the top 20% of the income distribution. The values 3.06 and 2.73 are the means of all values estimated for θ for different incomes. The means of all estimates (3.06 and 2.73) will be taken as the estimates of θ from now on. In the next section, the same procedures will be performed for the house prices dataset.

Figure 5.2. Pareto Tail Index Estimates for Household Disposable Incomes (Household Survey)



The constant term of this linear equation can be calculated in the same way. The constant terms calculated in this way will give the logarithmic values of the threshold of the missing upper income group [$\log(\tau)$]. We have shown in Table 5.1. the thresholds, i.e. τ , that we calculated for different income types.[4]

Table 5.1. Estimates of Income Thresholds (τ)

Disposable Income Type	Income Threshold (τ)
Household Income per capita (2019)	47,703 ₺
Total Household Income Turkey (2019)	120,610 ₺

5.2. Estimating the Tail Index Using Both Income and House Price Data

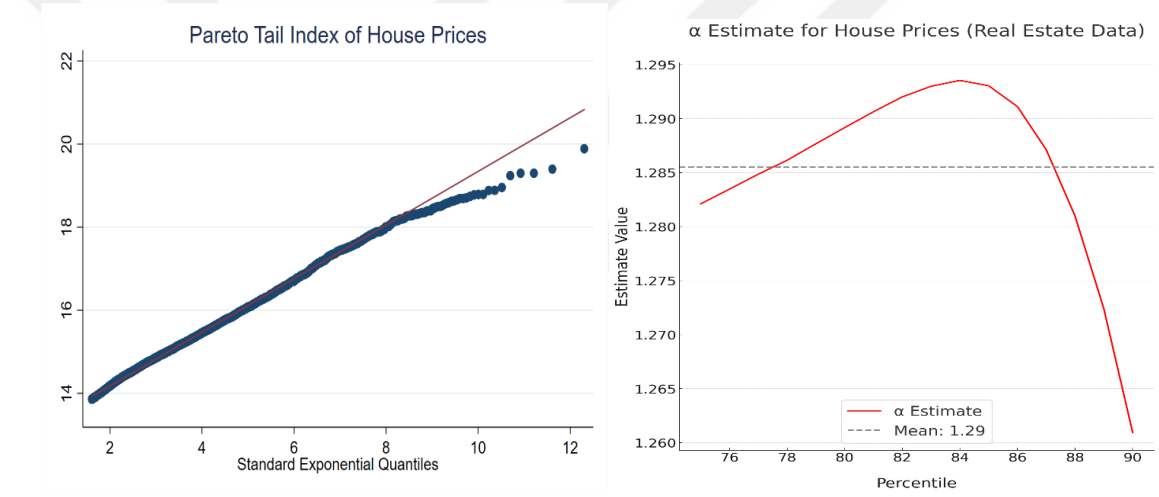
We will apply the same Pareto tail index estimation to the house prices dataset as we did for HBS, both to check whether the HBS is indeed biased downwards for the upper income group and to determine what the inverted pareto coefficient should be if we combine these two datasets if we accept that it is biased downwards. We will call the Pareto coefficient, which we use the notation θ for HBS, α for the house prices dataset.

[4] Please see Appendix D for manually fitting two distributions and determining τ.

In this case, the scatter plot of $\log(price_{19})$ over $\log(1 - F_2(price_{19}))$ for α is as shown in Figure 5.3.

When α is estimated separately for each percentile (up to the top 5%) from the top 25% of the distribution, we obtain the estimated values shown in Figure 5.3. The value of 1.29 in the graph is the average of the estimated values for α and will be considered as the estimate of α from now on.

Figure 5.3. The Pareto Quantile Plot and Pareto Tail Index Estimates for House Prices



Real Estate Data: sahibinden.com

In order to calculate the pareto coefficient of income distribution from the combination of the two datasets, we only need to estimate $\beta_1$in equation 3.4. Because as stated in Weide et al. (2018), $\widehat{\theta_{mix}} = \frac{\widehat{\alpha}}{\widehat{\beta_1}}$. To calculate β_1, we ran a non−parametric Kernel regression with the logarithm of annual (imputed) rent provided in the HBS as the dependent variable and the logarithm of per capita household income as the independent variable.

In Figure 5.4. the blue lines show the fitted values of the non−parametric Kernel regressions. The positive relationship between (imputed) rent and (per capita) household income becomes more pronounced as the income level increases. In addition, Figure 5.4.

also shows that this relationship is quite linear for the top 30%. As a result, the slope of the linear fit in the graph can represent $\widehat{\beta_1}$. However, for a more appropriate calculation of $\widehat{\beta_1}$ as introduced in the Figure 5.5., $\widehat{\beta_1}$ was calculated by OLS for each percentile starting from the top 20% and average of these $\widehat{\beta_1}$ were 0.75 for total household incomes and 0.72 for per capita incomes.

Figure 5.4. Household Income per capita versus (Imputed) Rent (Log-Log, Household Survey)

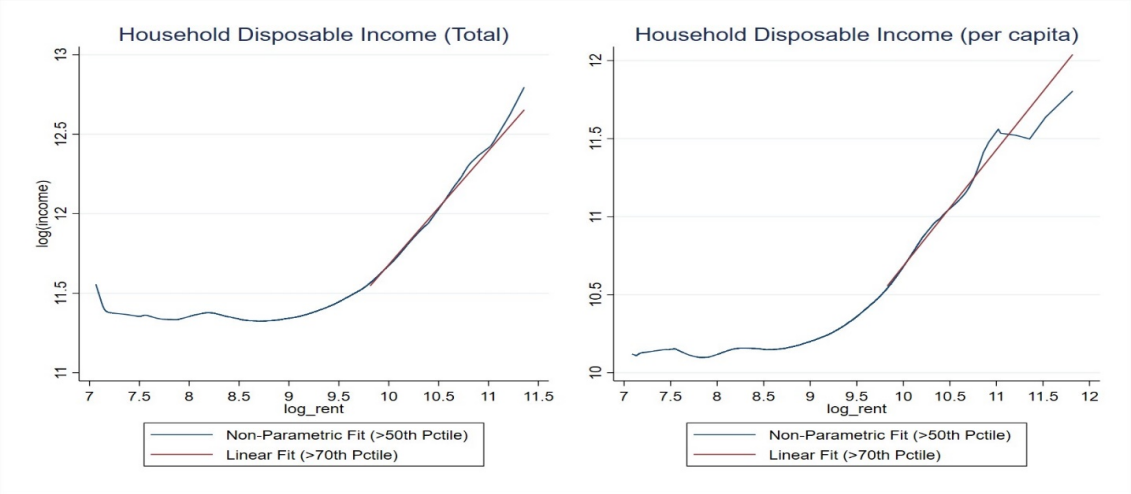
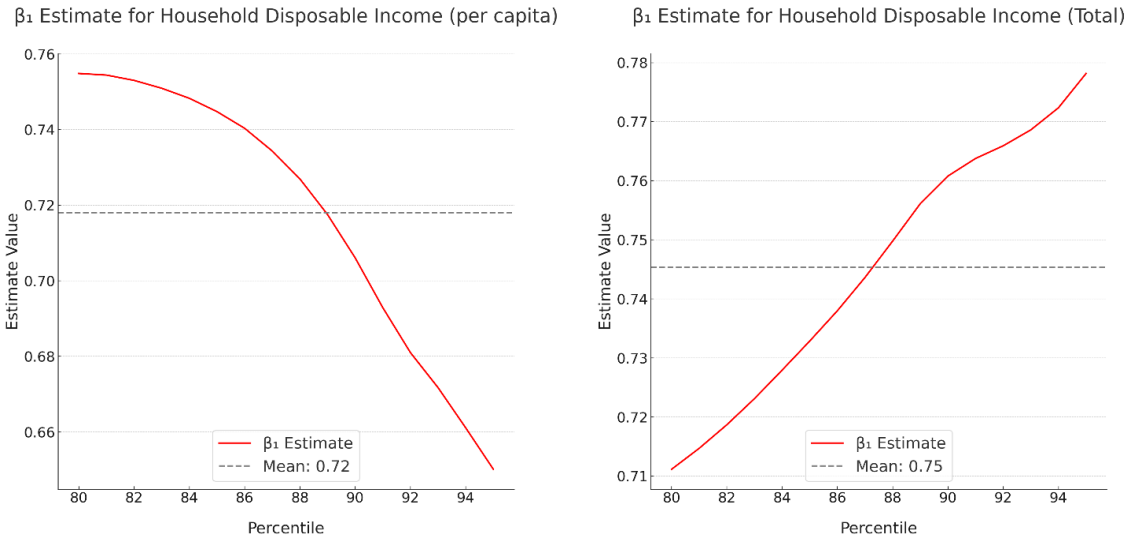


Figure 5.5. Estimates of β_1 Using Increasingly Smaller Numbers of Top Observations (Household Survey)



Everything we have estimated so far is summarized in Table 5.2. $\hat{\gamma}$ is used instead of the inverted Pareto coefficient. The inverted Pareto coefficient shows how many times the average income of the upper tail corresponds to the lower threshold of the tail. $\hat{\gamma}$ is calculated by the formula $\hat{\gamma} = \frac{\hat{\theta}}{\hat{\theta}-1}$. For example, if the inverted Pareto coefficient above 100 thousand liras is 3, this means that the average income above 100 thousand liras is 300 thousand liras. Table 5.2. shows that, $\hat{\gamma}_{svy}$ is smaller than $\hat{\gamma}_{mix}$, which means that while the average income above the threshold (τ), which we calculate as missing in the HBS, is 1.588 times the threshold, when we combine the HBS with house price data, the same average income is 2.442 times the τ. In short, Table 5.2. also shows that HBS underestimates the upper income group.

Table 5.2. Estimates of Main Variables of Interest ($\boldsymbol{\beta_1}$, $\boldsymbol{\alpha}$, $\boldsymbol{\theta_{mix}}$, $\boldsymbol{\theta_{svy}}$, $\boldsymbol{\gamma_{mix}}$, $\boldsymbol{\gamma_{svy}}$)

Disposable Income Type	$\widehat{\beta_1}$	$\hat{\alpha}$	$\hat{\theta}_{mix}$	$\hat{\theta}_{svy}$	$\hat{\gamma}_{mix}$	$\hat{\gamma}_{svy}$
Total Household Income Turkey (2019)	0.745	1.29	1.725	3.06	2.380	1.486
Household Income per capita (2019)	0.718	1.29	1.790	2.73	2.265	1.579

5.3. Main Results: Re-estimating Inequality for Turkey

In Equation 3.4, we have defined the relationship between household income and the (imputed) rent of the houses where households reside. We have the coefficients $\widehat{\beta_0}$ and $\widehat{\beta_1}$estimated using the HBS. In addition, we have the parameters of the Pareto Tail Indices of both the disposable income variable in the HBS and the house price variable in the real estate database ($\hat{\theta}_{svy}$ and $\hat{\alpha}$, respectively), and the Pareto Tail Index ($\widehat{\theta_{mix}} = \frac{\hat{\alpha}}{\widehat{\beta_1}}$) of the distribution that would result if the two databases were combined. Table 5.2. also shows the inverted Pareto coefficients, which are calculated using only the survey data ($\hat{\gamma}_{svy}$) and which would be obtained by combining the two data sets $\hat{\gamma}_{mix}$. However, these variables are insufficient to calculate both the income distribution and Gini coefficient of income distribution result from combination of two datasets. We need the parameter $\hat{\lambda}_{prop7}$ to calculate the income distribution and Gini coefficients that result from combination of the two datasets. We can easily obtain $\hat{\lambda}_{svy}$ by estimating τ from Equation 5.1 and then calculating the ratio of the number of households (population)

with income above τ to the total number of households (total population) in the HBS. Equation 3.9 will be used to estimate $\hat{\lambda}_{prop7}$, which is the ratio of the number of households (population) with income above τ to the total number of households (total population) as a result of combining the two datasets. The distributions of $\widehat{f_1}(y)$ and $\widehat{f_2}(y)$, needed to implement Equation 3.9, will be estimated by Kernel Density Estimation method and the values of $\widehat{f_1}(\tau)$ and $\widehat{f_2}(\tau)$ will be calculated. After calculating all these values, we show the λ values calculated using only the survey data for different income groups and the λ values obtained by combining the two datasets in Table 5.3.

We created Table 5.4. to compare the Gini coefficients and income share of top 5% (10%) calculated using the $\hat{\lambda}_{prop7}$ values obtained by combining the two datasets with the Gini coefficients and income share of top 5% (10%) calculated using only the survey data for the same income group. As a result, while the Gini coefficient calculated with total household income using the HBS in Turkey in 2019 was 0.354, this coefficient increases to 0.501 when the two data sets are combined. When the Gini coefficient is calculated with per capita household income, it increases from 0.42 to 0.542. Similarly, when income share of top 5% (10%) is calculated with total household income, it is 16.75% (26.76%) in the surveys and 28.15% (40.01%) in reality. Income share of top 5% (10%) increases from 20.34% (31.56%) to 30.00% (42.84%) per capita household income. All these calculations are based on the implicit assumption that all observations in the real estate dataset have equal weight.

Table 5.3. Estimates of λ_{svy} and λ_{prop7}

Disposable Income Type	τ	$\hat{\lambda}_{svy}$	$\hat{\lambda}_{prop7}$
Household Income per capita (2019)	₺ 47,703	7.19%	17.96%
Total Household Income Turkey (2019)	₺ 120,610	11.52%	26.00%

Our estimate for total household income Gini coefficient (0.507) is reasonably close to Ceritoğlu et al. (2023) where they find income inequality Gini coefficient for 2019 as 0.517. Even though the estimated Gini coefficient increased from 0.354 to 0.507, the

corrected Gini coefficient is still one percentage point lower than a household survey which over-sampled rich neighborhoods.

Table 5.4. Estimates of Top Shares and GINI

Disposable Income Type	S_5^{svy}	S_5^{mix}	S_{10}^{svy}	S_{10}^{mix}	$Gini^{svy}$	$Gini^{mix}$
Total Household Income Turkey (2019)	16.75%	28.15%	26.76%	40.01%	0.353	0.501
Household Income per capita (2019)	20.34%	30.01%	31.56%	42.79%	0.42	0.542

6. DISCUSSION: CORROBORATE HOUSE PRICE CORRECTION WITH WEALTH CORRECTION AND PRACTICAL USE OF DATASET

6.1. Corroborate House Price Correction with Wealth Correction

The evidence presented so far in this paper is quite convincing that measuring income inequality in Turkey based on household data alone is a highly incomplete measure. In particular, the adaptation of Weide et al. (2018)'s methodology for estimating the upper income group with house prices for Egypt to Turkey shows that the upper income group in Turkey is not captured well enough in the surveys. Nevertheless, testing the consistency of this adjustment with the existing literature will be important for other studies that build on the results of this research. In this section, the results of this study and the broad implications of the new income distribution data will be analyzed and discussed with the existing literature. It will also suggest some recommendations for future research and policy analysis on the use of the new dataset.

The comparison of the HBS with the National Accounts (NA) shows that there is indeed a "missing income" in the surveys. Table 4.2. shows the extent to which different income items calculated by Tekgüç & Eryar (2023) cover the NA in the HBS for 2019. Surveys cover 61% of total disposable income. Except for mixed income (which is mostly entrepreneurial income) and capital income, the coverage rate of the surveys is in good match with the macro data. In other words, most of the 39% of total income not covered by the survey comes from mixed income and capital income, which are generally considered to belong to the upper income group. Therefore, it is safe to argue that the current "missing income" actually points to a "missing rich problem".

Examining how much of the "missing income" in the national accounts is covered by the correction with house price data will also show how reliable this methodology is. Table 6.1. shows that for household income per capita, the house price methodology predicts almost exactly. Although total income is slightly overestimated, the amount of overshooting is negligible compared to total income. The fact that such accurate results

can be obtained by distributing the "missing income" in the national accounts in a way that is closer to the distribution in this survey with a more easily practical method can lead to many interpretations on income inequality. With a more practical methodology, even without a data set as processed and corrected as in this methodology, the change in indicators of inequality over the years, trends and the magnitude of the change can be measured more easily and more accurately than official data. Therefore, the incomes of the missing rich are corrected by imputing the discrepancy between the total income of the household sector in the national accounts and the total income estimated from the survey.

Table 6.1. Comparison of Combined Income with Missing Income in HBS

Disposable Income Type	Threshold	Missing Income (NA)	Added Income	Deleted Income	Income Imputed (Net)	Cover Rate
Household Income per capita (2019)	₺47,703	₺ 1,071,951 mn.	₺ 1,471,194 mn.	₺ 384,943 mn.	₺ 1,086,250 mn.	101%
Total Household Income Turkey (2019)	₺120,610	₺ 1,071,951 mn.	₺ 1,695,929 mn.	₺ 418,362 mn.	₺ 1,277,567 mn.	119%

Intuitively, we perform a simplified version of Alvaredo et al. (2019): We ditch the correction with respect to Lebanese taxable income distribution data. We obtain the arithmetic average of wealth distribution data for the U.S., France, and China from World Income Database (WID) for each ventil and allocate the discrepancy between national accounts and survey to income ventils according to their wealth share. U.S., France and China wealth distribution data are regarded as best quality. In 2019, Household sector total income (according to National Accounts) was 57% of GDP. HBS total household income was 70% of household sector (or 40% of total GDP) and discrepancy was 720

billion₺ (or 17% of GDP). According to WID, average wealth shares of top 5% and top 10% were 48.6% and 63.4%, respectively. These WID estimates are only a bit larger than Cerioğlu et al. (2023) and all the difference is due to top 5% wealth share.[5] We estimate corrected Gini coefficient as 0.491 with this method.

This finding corroborates Weide et al. (2018)'s methodology: Complementing household surveys with income estimates from house prices produces estimates similar to imputing missing incomes with wealth respect to wealth distribution. This finding is intuitive since housing is the largest asset of the great majority of households. Distributing wealth according to the arithmetic mean of the wealth distribution of the US, France and China gives an approximate result. This method is based on the assumption that the arithmetic average of the wealth distribution of the US, France and China can be adapted to Turkey. The underlying claim is that capitalism in the 21st century has a transnational character, and that the distribution of wealth shows similar characteristics in each country. Wealth distribution is a better proxy for global characteristic of Capitalism than income because while income taxation varies significantly across countries, wealth is not taxed properly in almost any country.

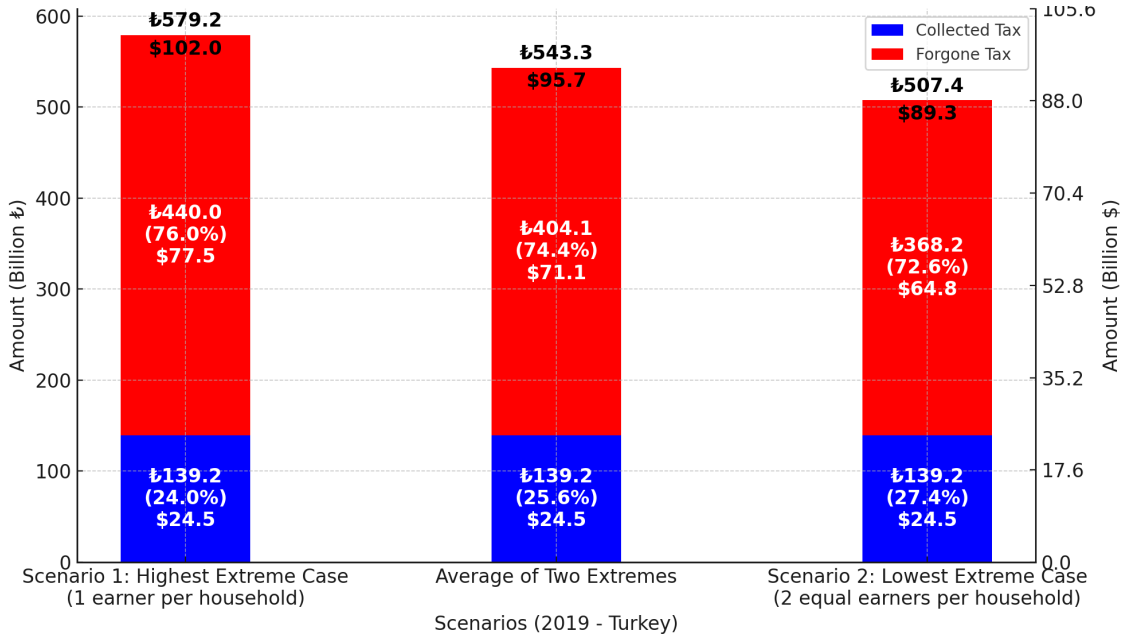
6.2. Practical Use of Dataset: Tax Forbearance Estimation

In many developing countries, the government tolerates tax evasion (Holland, 2016; 233). Tekgüç & Eryar (2023) also state that the government tolerates tax evasion in Turkey. In this study, the dataset produced by estimating the distribution of missing income makes it possible to calculate how much the governments tolerate tax evasion in the upper income group. Using total disposable household income data, we calculate the income tax paid and not paid by households above the threshold using the legal income tax brackets. Figure D.1. shows the result of these calculations. In Turkey there is no joint filing of taxes by couples. In other words, all direct taxes are levied at the individual level. Estimating missing taxes by the rich requires information on each tax payer in the household. We only estimate the total disposable income of the household and have no demographics estimates. So we produce to alternative extreme estimates for the missing

[5] Ceritoğlu et al. (2023) do not report the wealth distribution for whole sample, so we use WID data.

income taxes: (i) assume that all the income is earned by one person in the household; (ii) assume household has two adults earning one-half of the estimated total. The fact that all income is earned by one income earner means that this income earner enters a higher tax bracket faster. Therefore, this extreme scenario will yield the highest extreme results of all possible scenarios. The other extreme scenario is that the total household income is earned by two equal income earners. The more the income is distributed equally to more people in the household, the more unlikely it will be for income earners to enter a higher income tax bracket. There are no significant differences between the two extreme scenarios. Therefore, although a definite and final conclusion cannot be reached, one can be sure that an approximate conclusion has been reached. Assuming that the average of the two extremes is the most realistic scenario, it can be said that the tax that the government does not collect from the upper income group is about three times the total income tax collected. While collecting the income tax that the government does not collect from the upper income group is very important both in terms of the resources it will create for government expenditures and the effect of this policy in eliminating inequality in income distribution should be taken into consideration.

Figure 6.1. Comparison of Collected and Forgone Tax Amounts



(The annual average exchange rate is calculated based on $1 = ₺5.68.)

While the blue bar shows the tax collected from all population, red bar shows only tax evasion of top earners.

7. CONCLUSION

This thesis addresses the problem of income inequality, which has become an increasingly critical issue in Turkey in recent years, by focusing on the problem of the "missing rich" who cannot be accurately captured in household surveys. Aside from the fact that household surveys do not capture the richest people well enough, we had to use innovative methods to solve this problem considering the fact that tax records data, which is the most widely used method in the literature to solve this problem, is problematic in developing countries and this data is not published in Turkey. In this thesis Weide et al. (2018) is followed and the income inequality estimates for Turkey for 2019 is corrected. Weide et al. (2018) intuition is that there is a close relationship between households' income and the house prices. A new income distribution data for Turkey for 2019 is produced by combining house price data from the real estate website "Sahibinden.com" and income distribution data from the Household Budget Survey (HBS). Sahibinden.com is such a website for Turkey containing around half a million house sale listings on an average day. Roughly 220 thousand advertisements from this website are downloaded and household incomes are generated corresponding to those. Combining the HBS with real estate data, we find that the share of total income received by the top of the income distribution is significantly underestimated. This study argues that the top 5% (10%) income share increases from 17% (27%) to 28% (40%) by correcting the missing rich problem in the surveys. As a result, the Gini index of household income in Turkey has increased from 35.3 to 50, which implies a substantial difference. Such a sizeable correction means that disposable income inequality in Turkey is not at the U.S. levels (around 0.4) but closer to Latin American levels (around 0.5).

Our findings are supported by the fact that total household income in household surveys corresponds to only 61% of total household income calculated using national accounts. Inspired by the methodology of Alvaredo et al. (2019), we tested the robustness of our findings. Alvaredo et al. (2019) intuition is that (i) the amount of missing income in the surveys can be obtained by comparing the household survey totals to household sector total income in National Accounts. (ii) The discrepancy between survey and National Accounts can be imputed to surveys with respect to income distribution derived from

wealth distribution. Capital income and mixed income, which are underestimated by surveys, are more likely to be distributed in proportion to wealth share. The resulting Gini coefficient by this method is 0.491. When the missing income is distributed according to wealth distribution derived from the arithmetic average of the US, China and France with the claim that capitalism in the 21st century has a transnational character, and that the distribution of wealth has similar characteristics in each country, the resulting income distribution agrees quite well with the findings using house price data. This comparison confirms our methodology of eliminating the right tail problem in the income distribution corrected using house price data. The use of alternative data sources to obtain a more accurate representation of the income distribution is quite reasonable and consistent with different methodologies. As a result, it is concluded that Weide et al. (2018) methodology has a remarkably good performance of approximating the missing incomes due to missing wealthiest households in surveys.

The official Gini measures, which are historically corrected using the distribution of wealth, also illustrate the extent to which the use of official data in the study of income inequality can lead to misleading conclusions. It is recommended that some studies investigating the evolution of income distribution over time should reconsider their findings. Moreover, there are many questions about the reliability of cross-country comparisons, as household surveys in each country may be inaccurate at different rates. However, the elimination of the "missing rich" problem is a promising area for cross-country comparisons and for understanding the international dynamics of income distribution.

To conclude, this study contributes to the literature on income inequality in Turkey by addressing the "missing rich" problem. Our findings underline the importance of measuring income distribution accurately for effective policy making and evaluation. A new dataset on income distribution is produced by combining the Household Budget Survey and house price dataset in a decent way. This dataset is expected to produce much more realistic results than the survey data in analyzing income distribution. At a time when income inequality is considered a critical global agenda, this is a promising avenue for further research.

BIBLIOGRAPHY

Achcar, Gilbert. 2020. "On the 'Arab Inequality Puzzle': The Case of Egypt." *Development and Change* 51 (3): 746-770.

Alvaredo, Facundo. 2011. "A Note on the Relationship Between Top Income Shares and the Gini Coefficient." *Economics Letters* 110 (3): 274-277.

Alvaredo, Facundo, and Juliána Londoño. 2013. "High Incomes and Personal Taxation in a Developing Economy: Colombia 1993-2010." CEQ Working Paper 12.

Alvaredo, Facundo, Lydia Assouad, and Thomas Piketty. 2019. "Measuring Inequality in the Middle East 1990–2016: The World's Most Unequal Region?" *Review of Income and Wealth* 65 (4): 685-711.

Auten, Gerald, and David Splinter. 2023. "Income Inequality in the United States: Using Tax Data to Measure Long-term Trends." *Journal of Political Economy*. Published in 2024.

Benabou, Roland, and Jean Tirole. 2006. "Belief in a Just World and Redistributive Politics." *The Quarterly Journal of Economics* 121 (2): 699-746.

Blanchet, Thomas, Ignacio Flores, and Marc Morgan. 2022. "The Weight of the Rich: Improving Surveys Using Tax Data." *The Journal of Economic Inequality* 20, no. 1: 119-150.

Ceritoğlu, Evren, Seyit M. Cılasun, Müşerref Küçükbayrak, and Özlem Sevinç. 2023. "Household Portfolios in Türkiye: Results from the Household Finance and Consumption Survey." *Central Bank Review* 23 (4).

Clark, John Bates. 1908. *The Distribution of Wealth: A Theory of Wages, Interest and Profits*. New York: Macmillan.

Corak, Miles. 2013. "Income Inequality, Equality of Opportunity, and Intergenerational Mobility." *Journal of Economic Perspectives* 27 (3): 79-102.

Danish Economic Council. 2016. "Demographic Trends, Such as the Age Structure of the Population, Immigration, Educational Attainment and Household Structure Can Significantly Impact Inequality."

Fujita, Shinya. 2023. "Income Inequality in Terms of a Gini Coefficient: A Kaleckian Perspective." *Cambridge Journal of Economics* 47 (6): 1087-1106.

Geloso, Vincent J., Phillip W. Magness, John Moore, and Phillip Schlosser. 2022. "How Pronounced Is the U-curve? Revisiting Income Inequality in the United States, 1917–60." *The Economic Journal* 132 (647): 2366-2391.

Hacker, Jacob S., and Paul Pierson. 2010. *Winner-Take-All Politics: How Washington Made the Rich Richer—and Turned Its Back on the Middle Class*. New York: Simon & Schuster.

Heckman, James J. 2006. "Skill Formation and the Economics of Investing in Disadvantaged Children." *Science* 312 (5782): 1900-1902. https://doi.org/10.1126/science.1128898.

Holland, Alisha C. 2016. "Forbearance." *American Political Science Review* 110 (2): 232-246.

Işık, Enes, Özgür Orhangazi, and Hasan Tekgüç. 2020. "Heterogeneous Effects of Minimum Wage on Labor Market Outcomes: A Case Study from Turkey." *IZA Journal of Labor Policy* 10 (1): 1-41.

Jenkins, Stephen P. 2016. "Pareto Models, Top Incomes and Recent Trends in UK Income Inequality." *Economica* 84 (334): 261-289.

Kuznets, Simon. 1953. *Shares of Upper Income Groups in Income and Savings*. New York: National Bureau of Economic Research.

Makhlouf, Yousef. 2023. "Trends in Income Inequality: Evidence from Developing and Developed Countries." *Social Indicators Research* 165 (1): 213-243.

Marmot, Michael. 2015. "The Health Gap: The Challenge of an Unequal World." *The Lancet* 386 (10011): 2442-2444. https://doi.org/10.1016/s0140-6736(15)00150-6.

Marshall, Alfred. 1890. *Principles of Economics*. London: Macmillan. 8th ed. published in 1920.

Milanovic, Branko. 2016. *Global Inequality: A New Approach for the Age of Globalization*. Cambridge: Harvard University Press.

Milanovic, Branko. 2019. *Capitalism, Alone: The Future of the System that Rules the World*. Cambridge: Harvard University Press.

Mila, Morgan M. 2018. *Essays on Income Distribution: Methodological, Historical and Institutional Perspectives with Applications to the Case of Brazil (1926-2016)*. Doctoral dissertation, Paris Sciences et Lettres (ComUE).

Pareliussen, Jon K., Mikkel Hermansen, Christophe André, and Orsetta Causa. 2018. "Income Inequality in the Nordics from an OECD Perspective." *Nordic Economic Policy Review* 2018: 17-57.

Piketty, Thomas. 2014. *Capital in the Twenty-First Century*. Cambridge: Harvard University Press.

Piketty, Thomas, and Emmanuel Saez. 2003. "Income Inequality in the United States, 1913–1998." *Quarterly Journal of Economics* 118 (1): 1-39.

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman. 2023. "Comment on Auten and Splinter (2023)."

Reich, Robert B. 2015. *Saving Capitalism: For the Many, Not the Few*. New York: Knopf.

Saez, Emmanuel, and Gabriel Zucman. 2016. "Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data." *The Quarterly Journal of Economics* 131 (2): 519-578.

Saez, Emmanuel, and Gabriel Zucman. 2019. "Progressive Wealth Taxation." *Brookings Papers on Economic Activity* 2019 (2): 437-533.

Stiglitz, Joseph E. 2012. *The Price of Inequality: How Today's Divided Society Endangers Our Future*. New York: W.W. Norton & Company.

Tekgüç, Hasan, and Değer Eryar. 2023. "Redistribution Trends in Turkey: Regressive Taxes, Structural Change, and Demographics." Commitment to Equity Working Paper 134. https://commitmentoequity.org/wp-content/uploads/2024/01/ceq134.pdf.

Tekgüç, Hasan, Halit Güzelsoy, Ayşe Göç, Çağlar Karabulut, and Hüseyin E. Almaz. 2024. "Program Kodu: 3005 Proje No: 122G241." https://point.khas.edu.tr/wp-content/uploads/2024/05/122G241_Tekguc_PGR1_web.pdf.

Van Der Weide, Roy, Christoph Lakner, and Elena Ianchovichina. 2018. "Is Inequality Underestimated in Egypt? Evidence from House Prices." *Review of Income and Wealth* 64: S55-S79.

Yonzan, Nishant, Branko Milanovic, Salvatore Morelli, and Janet Gornick. 2022. "Drawing a Line: Comparing the Estimation of Top Incomes Between Tax Data and Household Survey Data." *The Journal of Economic Inequality* 20 (1): 67-95.

APPENDIX A: SHORT LIST OF ASSUMPTIONS

(a) High earners are largely absent from HBS, and DB-2 does not sufficiently cover the lower-income segments. Individually, each dataset fails to capture the complete picture of income distribution. DB-1 allows for a reliable estimate of $F_1(y) = Pr[Y < y | Y \leq \tau]$, while DB-2 provides a reliable estimate for $F_2(y) = Pr[Y < y | Y > \tau]$.

(b) DB-2 encompasses the entire number of units, such as households or tax units, with incomes exceeding the above τ.

(c) There is a relationship between the households where the upper income group resides and the income of these households.

(d) The upper range of the x_h distribution can be characterized by a Pareto distribution.

(e) House prices and (imputed) rents are proportional.

(f) DB-2 records real household incomes and not predictors of income, allowing us to concentrate solely on the problems outlined in introduction.

(g) $f(y)$, the probability density function (pdf) of y, is a continuous function of y. Additionally, $f_1(y)$ and $f_2(y)$, representing the pdf's of $F_1(y)$ and $F_2(y)$ respectively, can be estimated by Kernel Density Estimation.

(h) Deflated house prices by district from 2023 to 2019 represent the Pareto tail index of house price distribution in 2019.

(i) (i) In the house price dataset, each house represents one household (the correction will still be underestimated as upper-income households usually have more than one house for their own use). For robustness, the coefficients will be reconstructed under the assumptions that every 1.5 and 2 houses represent one household.

APPENDIX B: ROBUSTNESS ANALYSIS

Table B.1. shows the sensitivity analysis for different τ values. As long as it is meaningful, taking different τ values does not significantly change the final $Gini^{mix}$ value. This means that the two distributions behave closely around the intersection of the income distributions obtained from the two datasets.

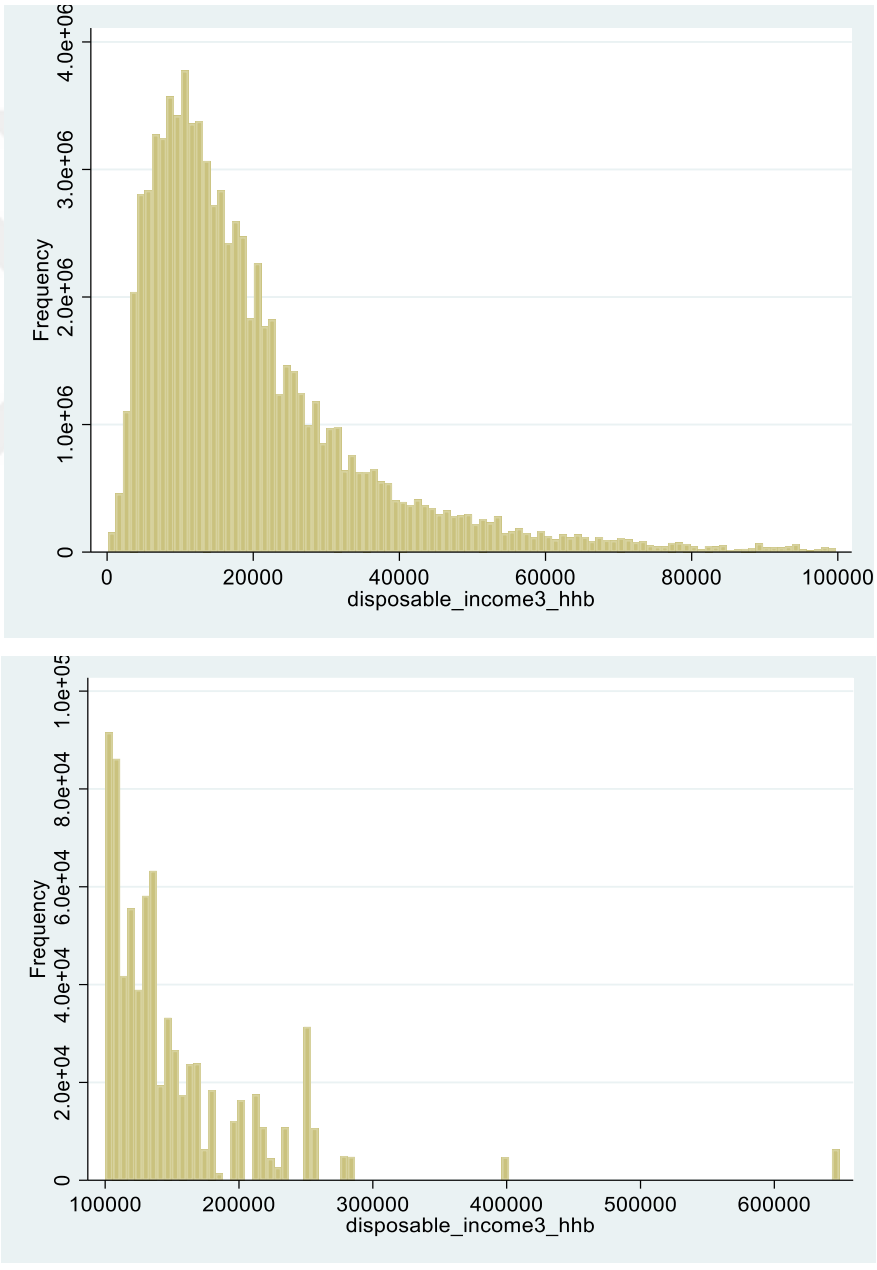
Table B.1. Sensitivity Analysis for τ

$\hat{\tau}$	$\hat{\lambda}_{svy}$	$\hat{\lambda}_{prop7}$	$Gini^{mix}$
96.488	19.88%	35.13%	0.497
108.549	14.91%	29.33%	0.497
114.580	13.02%	26.96%	0.497
120.610	11.52%	26.00%	0.501
126.640	10.02%	26.27%	0.506
132.671	8.73%	25.47%	0.510
150.762	5.77%	22.32%	0.518
180.915	3.17%	18.10%	0.531
211.068	1.90%	13.44%	0.530
241.220	1.36%	8.10%	0.503

APPENDIX C: SPARSENESS

More Evidence for Missing Rich Problem in HBS (2019): Income distribution for households with a per capita household income of less than 100 thousand TL is as follows. Distribution of incomes over 100 thousand TL is the right panel.

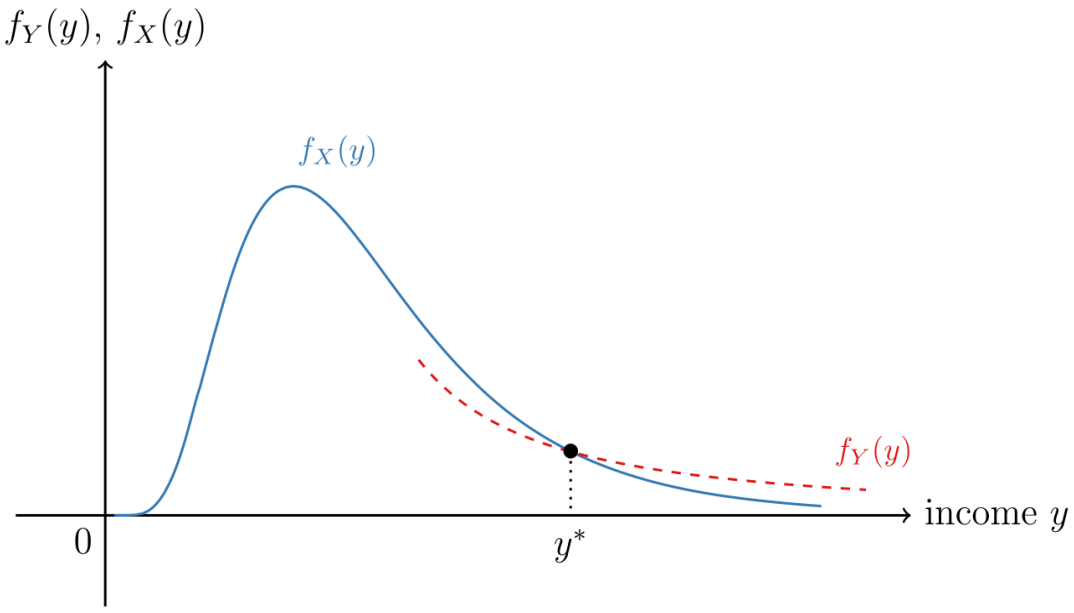
Figıre C.1. Sparseness



APPENDIX D: THEORETICAL INTUITION AND PRACTICAL RESULTS OF COMBINING TWO DATASETS

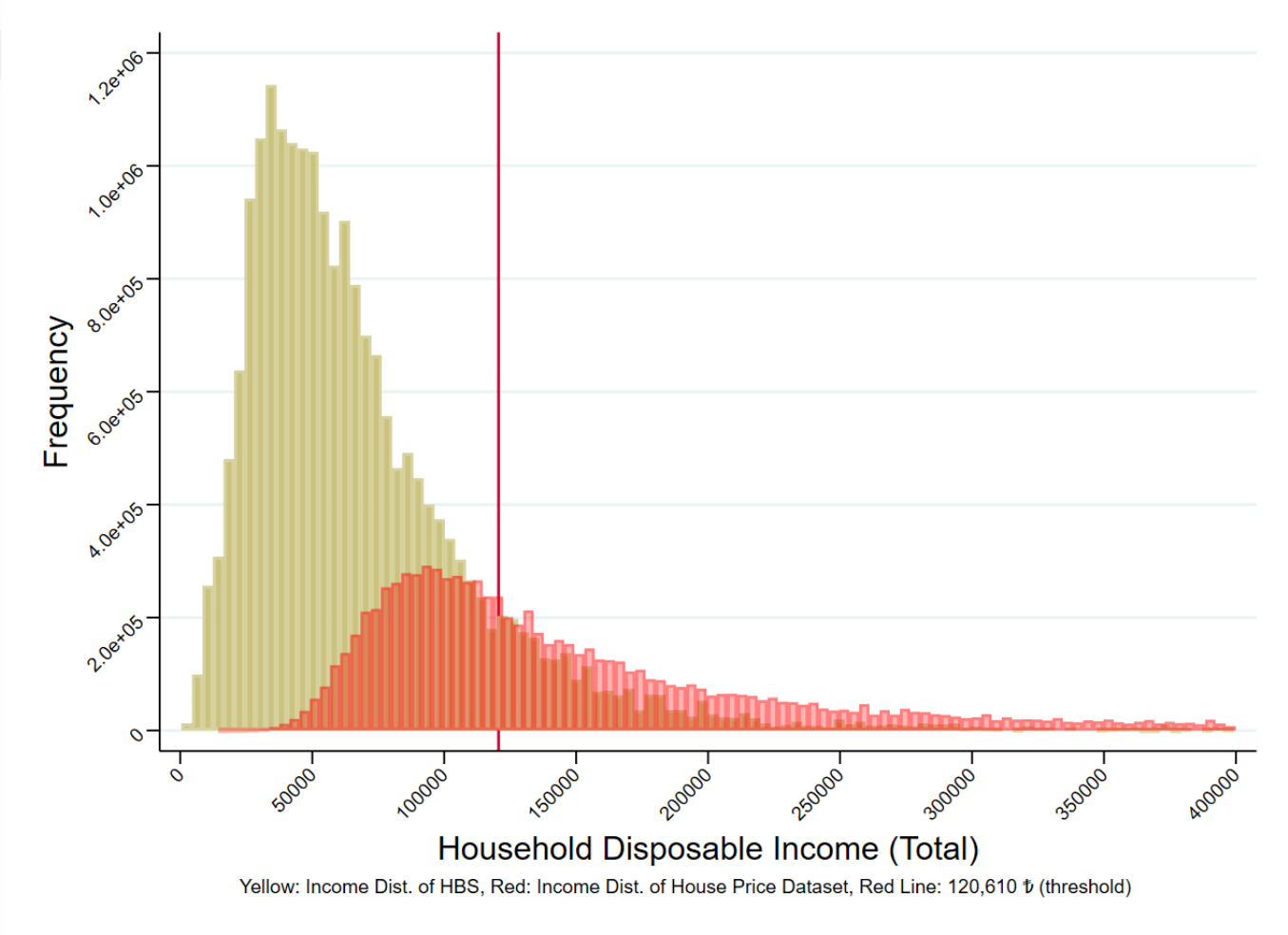
Blanchet et al. (2022) visualize as shown in Figure D.1. the combination of distributions from two different datasets with reference to where they both intersect. To see how this theoretical combination shown in Figure D.1. would look like in practice, we chose to use the histogram in Figure D.2. The income distribution estimated using the house price dataset is synchronized with the income distribution in the HBS at threshold using prop. 7 from Weide et al. (2018) as in Figure D.2.

Figıre D.1. Theoretical Intuition of Combining Two Datasets



Source: Blanchet et al. (2022).

Figıre D.2. Histogram of Two Distributions



CURRICULUM VITAE

Personal Information
Name and surname: Fırat Çağlar Karabulut

Academic Background
Bachelor's Degree Education: Middle East Technical University - Economics
Post Graduate Education: Kadir Has University - Economics
Foreign Languages: English