An Analysis of the Human Development Index and modelling income distributions

A project report submitted in partial fulfilment of the requirements for the degree of

Master of Technology
(Coursework)

by

Prajakta Sanjay Borse
M.tech in Computational and Data Science
Indian Institute of Science

भारतीय विज्ञान संस्थान

Under the supervision of

Prof. Murugesan Venkatapathi
Department of Computational and Data Sciences
Indian Institute of Science
Bengaluru- 560012

Declaration of Originality

I, **Prajakta Sanjay Borse**, with SR No. **06-18-01-10-51-21-1-19706** hereby declare that the material presented in the thesis titled

AN ANALYSIS OF THE HUMAN DEVELOPMENT INDEX AND MODELLING INCOME DISTRIBUTIONS

represents original work carried out by me in the **Department of Computational and Data Sciences** at **Indian Institute of Science** during the years **2022-2023**.

With my signature, I certify that:

- I have not manipulated any of the data or results.
- I have not committed any plagiarism of intellectual property. I have clearly indicated and referenced the contributions of others.
- I have explicitly acknowledged all collaborative research and discussions.
- I have understood that any false claim will result in severe disciplinary action.
- I have understood that the work may be screened for any form of academic misconduct.

Date: 09-06-2023 Prajakta Borse

In my capacity as supervisor of the above-mentioned work, I certify that the above statements are true to the best of my knowledge, and I have carried out due diligence to ensure the originality of the report.

Advisor Name: Prof. Murugesan Venkatapathi Advisor Signature

Acknowledgements

I am grateful to my advisor Prof. Murugesan Venkatapathi for his consistent guidance, insightful advice, and support throughout my project. Working under his guidance has been an enriching experience for me.

I would also like to express my sincere appreciation to the faculty members for their valuable suggestions and insights, which greatly aided me in my research.

I am grateful to my friends and colleagues for their support and valuable input that helped me during the work.

Lastly, I would like to express my deepest gratitude to my parents for encouraging and supporting me always.

Abstract

In the first part of the project, we analyze the Human Development Index, the most used index to measure a country's development. Although human development is a multidimensional quantity, the HDI can be factored into three broad components representing income, education and health. Analysis of the Human Development Index shows that the three indicators used in evaluating these three components of HDI are highly correlated with each other. Thus, we show that two of the three components of the current HDI are mostly redundant due to the indicators used. We also suggest a minor change in the base life expectancy used to evaluate HDI. Further, we propose an updated HDI by adding two new parameters- a fraction of Renewable Energy consumption and a Gender Balance Index. The inclusion of two new parameters makes the HDI index sustainable.

The second part of the report focuses on building statistical models to explain the observed income distributions using a minimum set of assumptions, such as a normal distribution of capacities of individuals. We generate income distributions starting from the initial assumptions using the inverse transform and sequential Monte Carlo methods. This work is motivated by the mismatch between observed data and the early models proposed for a first approximation, which are still in use. The report discusses the evolution of income from a uniform distribution to a Pareto distribution with time and its relationship with the capacity in the two cases of differential rewards and differential opportunity. The former assumes a stochastic increase of income in proportion to the capacity of an individual. In addition, the latter assumes that the probability of an increase in income is non-uniform and proportional to the capacity as well.

Contents

List of Abbreviations

GBI	Gender Balance Index
GDP	Gross Domestic Product
GE	Composition of GNI and Education Index
GNI	Gross National Income
HDI	Human Development Index
IHDI	Inequality-adjusted Human Development Index
LE	Life Expectancy
PPP	Purchasing power parity
UNDP	United Nations Development Programme

List of Figures

List of Tables

Part I

Analysis of Human Development Index

Chapter 1

Introduction and Literature review

The growth of a country lies in enriching the lives of the people of the country. Earlier, the GDP was the basis for measuring human development. However, human development has a multi-dimensional nature. While income has the potential to expand people's choices, it is not sufficient to guide a country's human development achievements. It is evident that higher income levels lead to higher health and educational achievements, but this relationship breaks down in many countries. Some growing countries with high incomes have a low life expectancy and education. High income does not necessarily stop disease spread, criminal activity or pollution in a country. These factors depict that income growth cannot solely measure human development. Hence, an index was developed to measure the human development of a country in terms of education, health, and income.

The primary function of HDI is to rank countries based on the HDI value. The Human Development Report 1990, [1] had been prepared by UNDP under Mahbub ul-Haq. In the report [2], Ul Haq defines human development as "the process of enlarging people's choices." The choice for a long healthy life and education are the most important ones. The Human Development Index is an important measure used by authorities to decide public policies.

In the paper [3], Anand and Sen discuss the concept and measurement of HDI. According to them, the motivation behind HDI was "to focus directly on the lives that people lead – what they succeed in being and doing." They also mention that the core of human development lies in enhancing the capabilities of people to function in elementary ways.

The UNDP used the three indicators- Life expectancy at birth, Adult literacy, and GDP per capita for measuring human development in 1990. The 1990 index calculated the deprivation value for a country. The HDI formulation was modified in 1991, and the HDR published a new report (1991). The minimum and maximum values were observed from the data. The Education index attributed the one-third weight to the years of schooling and the two-thirds weight to the literacy rate [4]. The Income index was calculated by using the Atkinson [5] formula. 'Mean years of schooling' and 'Expected years of schooling' are measured in years while GNI per capita is measured in PPP US$. Combining the indicators into an index requires transformation into a unit-less index lying between 0 and 1. These represented the most basic human capabilities. The data values of these indicators were normalized for easy comparison.

According to the Human Development Report (1994), the construction of the index was done in the following manner:
The aggregation formula for HDI in 1994 was:

$$I_{HDI} = \frac{I_{LE} + I_E + I_{GNI}}{3}$$

where I_{LE} is the Life expectancy index, I_E is the Education index, I_{GNI} is the Gross National Income index.
The Life expectancy index was calculated as,

$$I_{LE} = \frac{\text{life expectancy value in years} - 25}{85 - 25}$$

The Education index was calculated using the same indicators for aggregation as in 1991. The maximum and minimum values were fixed for both indicators. Mean years of schooling had a fixed minimum of 0 years and a maximum of 15 years, and an Adult literacy rate had a minimum of 0% and a maximum of 100%. For the GNI index, the minimum of real GDP PPP was fixed to be $200, and the maximum to be PPP $ 40,000.

The initial development and presentation of the HDI drew substantial interest and criticism from several research domains. Izete Pengo Bagolin and Flavio V. Comim (2008), in their paper [6], give a critical overview of HDI and highlight some of the issues in HDI, such as HDI faces problems of low quality and lagged data. Also, HDI could not reply to much significant criticism at that time. However, they concluded that ”HDI represents advancement indeed, both in terms of the characterisation of the multidimensional nature of development and its refined theoretical basis.”

Although the idea of creating an index for evaluating human development was remarkable, it was also subjected to significant critique regarding the selection of indicators, statistical characteristics of the index, and reliability of the data used for the index. Srinivasan, T. N. (1994), in his paper [7], mentions that GNP data of many developing countries suffer from problems of incomplete coverage, measurement errors, and biases. He also adds that the Human Development Index (HDI) has conceptual weaknesses and empirical shortcomings, which include significant problems of non-comparability across time and space, measurement errors, and biases. According to Hicks, D.A. (1997), [8], no direct measure has been developed thus far to account for inequality in the distribution of education and health/longevity among individuals when calculating the Human Development Index.

The author, Allen K (1991), addresses several issues with HDI. One of the issues mentioned in the paper discusses the sensitivity of HDI to the choice of maximum life expectancy. It effectively assumes that developed countries can make little or no progress in human development [9].

The HDI indicator underwent some major changes in formulation in 2010. The Life expectancy index, which represents long and healthy life, was calculated using the maximum value of 85 years and minimum value of 20 years, respectively. The Education index is calculated using the arithmetic mean of 'Mean years of schooling' and 'Expected years of schooling'. The Income index is calculated using the log of GNI per capita (PPP international dollars), with a maximum of $75,000 and a minimum of $100. The HDI is the geometric mean of the previous three normalized indices.

$$I_{HDI} = (I_{LE} \cdot I_E \cdot I_{GNI})^{1/3}$$

The previous HDI formulation used the arithmetic mean of the three component indices - Life expectancy index, Education index, and GNI index. If one of the three indicators has lower values and others have high values, the HDI index does not change as one low value is substituted for another index's high value. The issue is that one index's low value is not reflected in the HDI index. Desai also described this in his paper [10], in which he supported the geometric mean method. In 2010 modification of HDI, UNDP adopted the geometric mean formulation. Hence, the lower value of any index makes the HDI value lower for that country. So, the current HDI considers the differences in all three indicators.

In the paper of S. Anand 2018, [11], the author critically examines the modifications proposed in HDI in 2010 and verdicts in favour of the old HDI, which was calculated based on the arithmetic mean of the indices. According to the author, the HDI measure suffers from serious defects. The paper suggests how HDI might be recast to overcome the problems identified and better reflect the purpose for which it was devised.

In the 2010 Human Development Report [12], a new Inequality-adjusted Human Development Index (IHDI) was introduced. While the conventional HDI still has its value, the report argues that the IHDI provides a more accurate measure of actual human development by considering inequality. The IHDI index was introduced to address a limitation of the original Human Development Index (HDI). The HDI did not account for inequality within these dimensions, which can have a significant impact on human development [13].

Even though the current HDI has many good properties, it does not take into consideration many factors which are crucial in today's time. According to World Commission on Environment and Development (WCED), [14], development is sustainable if "it meets the needs of the present without compromising the ability of future generations to meet their own needs." This suggests the necessity of radical change in the composition of the current HDI.

The topic of what constitutes a well-designed index and the key factors that must be considered when creating an index is crucial. The construction of an index requires a thorough analysis of potential variables. An index functions as a statistical tool offering valuable information about the field it seeks to gauge. A well-designed index must accurately measure the phenomenon and be easy to understand.

Many authors attempted to modify the HDI index and its components. In the HDR Report of 2010, [15], the author evaluates current measures of political regimes and freedom and whether these indicators are useful in assessing people's political abilities. The author believed that the focus should be on political and civil institutions that affect people's opportunities to achieve their goals. The author suggests adding indicators of political liberties and civil liberties to augment the Human Development Index (HDI). Mehmet et al. [16] suggest adding an additional component to the Human Development Index (HDI), which considers the unemployment rate in a country. OECD countries have been selected for the analysis as the data for other countries is not available. According to the author, the unemployment problem needs to be considered in the long-term development perspective.

In the first part of the report, we analyse the HDI index. We study the correlation between the indicators. Moreover, we propose a new HDI index which incorporates the two new variables- a fraction of Renewable Energy consumption and a Gender Balance Index in its current composition. According to the paper [17], sustainable development within a society demands an effective and efficient utilisation of energy resources, which means that, in the long term, it is readily available at a reasonable cost and can be utilised for all required tasks without causing negative societal impacts. Hence, it is crucial that we include the consumption of Renewable Energy as an indicator in HDI. If a country has very less renewable energy consumption compared to the total energy, the HDI index for that country is penalised. Also, the Gender Balance Index is essential since it is not represented in any of the earlier three indicators of HDI. The Gender Balance Index is calculated as an arithmetic mean of the Gender Parity Indicator and Sex ratio indicator. Furthermore, we also suggest a change in the minimum value used for the Life expectancy index.

Chapter 2

Methodology

In the first part of the project, we analyze the Human Development Index to identify its limitations and propose potential improvements. The goal is to provide suggestions for enhancing the effectiveness and relevance of the HDI.

2.1 Current HDI Indicators

The data for the HDI ranking is sourced from the UNDP website [18]. The three dimensions of HDI are as follows:

1. A decent standard of living:
   This is measured by the GNI per capita (PPP $) indicator.
2. Knowledge:
   This is measured by two indicators of Education. It is the arithmetic mean of the following two indicators [19].
   (a) Mean years of schooling
   (b) Expected years of schooling
3. Long and healthy life:
   This is measured by the Life Expectancy at Birth indicator.

The indicators are normalized by the formula given below,

$$\frac{\text{actual value -min value}}{\text{max value -min value}} \tag{2.1}$$

where maximum and minimum values are the endpoints of the indicators. The GNI index is calculated by applying logarithm to the Gross National Income PPP $ per capita values and then normalizing the indicator by using formula 2.1.

The current HDI aggregation formula is given by,

$$(I_{\text{LE}} \cdot I_{\text{GNI}} \cdot I_{\text{E}})^{1/3} \tag{2.2}$$

It is the geometric mean of the three indicators. GNI is defined as a gross domestic product, plus net receipts from abroad of compensation of employees, property income and net taxes fewer subsidies on production.[20]. The Life expectancy index is calculated by using Life expectancy at birth in years. Life expectancy at birth is a measure that predicts the average lifespan of a newborn based on the current mortality rates [19].

In the next section, we look at the relationship between HDI and IHDI. We also find the variance of the HDI indicators and the correlation between two HDI indicators using the Pearson correlation coefficient.

2.2 Observations

2.2.1 Comparison with IHDI

Inequality-adjusted Human Development Index (IHDI) was proposed to account for the inequality in HDI. To analyse the relationship between IHDI and HDI, we find the Pearson correlation coefficient and plot both data. From Figure 2.1, it can be observed that both indices are highly correlated.

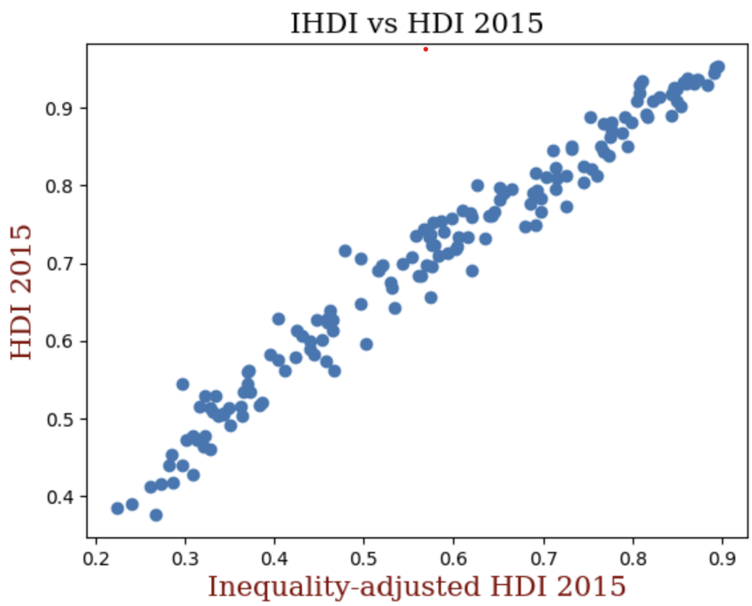


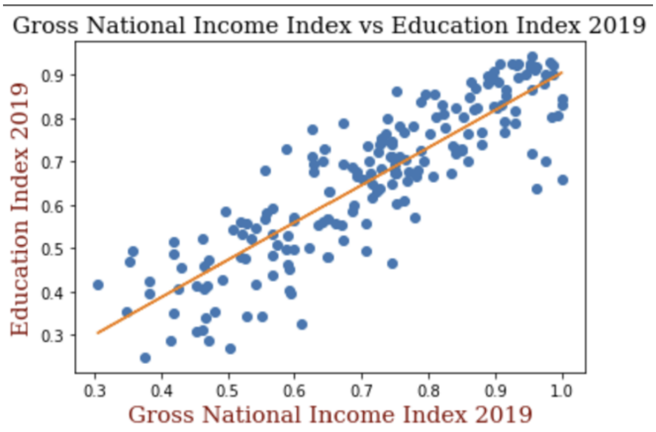
Figure 2.1: This is a scatter plot of the IHDI vs the HDI indicator. The correlation between Inequality-adjusted Human Development Index and the Human Development Index is 0.98.

2.2.2 Variance and correlation of HDI indicators

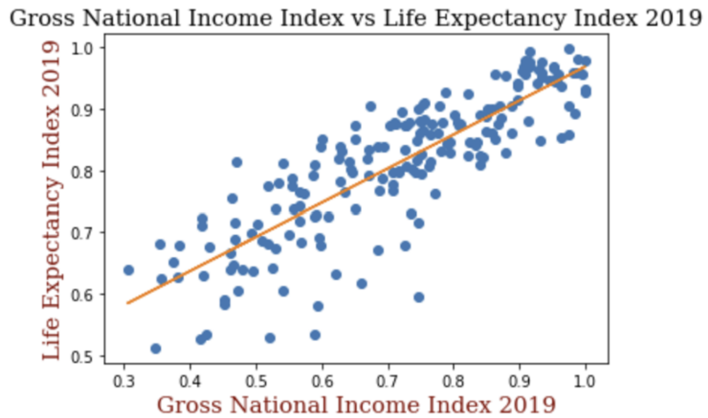
Years	σ_1^2	σ_2^2	σ_3^2	ρ_1	ρ_2	ρ_3
2015	0.031	0.032	0.014	0.96	0.96	0.92
2016	0.03	0.028	0.013	0.956	0.96	0.91
2017	0.029	0.028	0.013	0.957	0.96	0.915
2018	0.029	0.028	0.013	0.96	0.96	0.914

Table 2.1: Variance and Correlation coefficient table

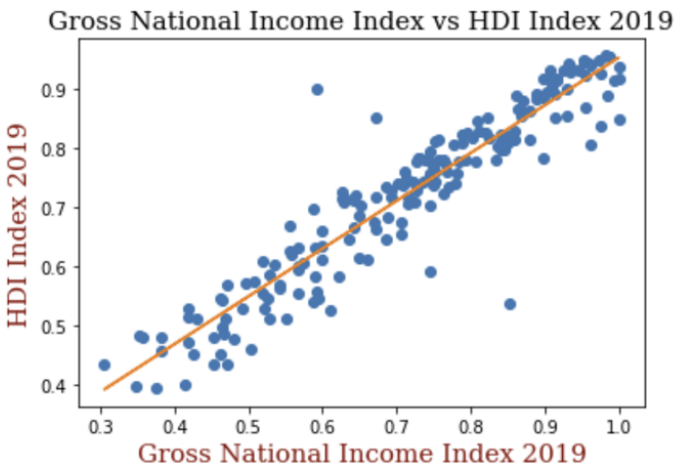
In the Table 2.1, σ_1^2 is the variance of Gross National Index, σ_2^2 variance of Education index, σ_3^2 variance of Life expectancy index and ρ_1 is the correlation between HDI and GNI index, ρ_2 is the correlation between HDI and Education index, ρ_3 is the correlation between HDI and Life expectancy index. The variance of the GNI index is the highest, while the Life expectancy index has the lowest variance.



(a) Scatter plot of GNI vs the Education Index. The Pearson correlation coefficient for this plot is 0.86.



(b) Scatter plot of GNI vs the Life Expectancy Index. The Pearson correlation coefficient for this plot is 0.84.



(c) Scatter plot of GNI vs the HDI Index. The Pearson correlation coefficient for this plot is 0.95.

Figure 2.2: From fig (a), (b), (c), we can observe that Education Index, Life Expectancy Index, and HDI Index are highly correlated to the GNI Index.

From Figure 2.2, it can be inferred that the GNI index is the most important index in measuring HDI.

2.2.3 Analysing the goal posts of indicators

We also study the minimum and maximum values used to normalize the HDI indicators.

Dimension	Indicators	Minimum value	Maximum value
Income	GNI per capita (PPP $)	$100 per capita	$75,000 per capita
Education	Mean years of schooling (years)	0	15
	Expected years of schooling (years)	0	18
Health	Life Expectancy (years)	20	85

Table 2.2: Goal-posts (minimum and maximum values) of the indicators.

2.2.4 Sensitivity analysis of HDI with respect to Life expectancy

To study the change in the HDI ranking by eliminating one indicator, we remove the Life expectancy index from the composition and calculate HDI using the two parameters of GNI and Education. We choose the Life expectancy index because it has the least variance.
So the modified HDI index is given below,

$$I_{GE} = (I_{GNI} * I_E)^{1/2}$$

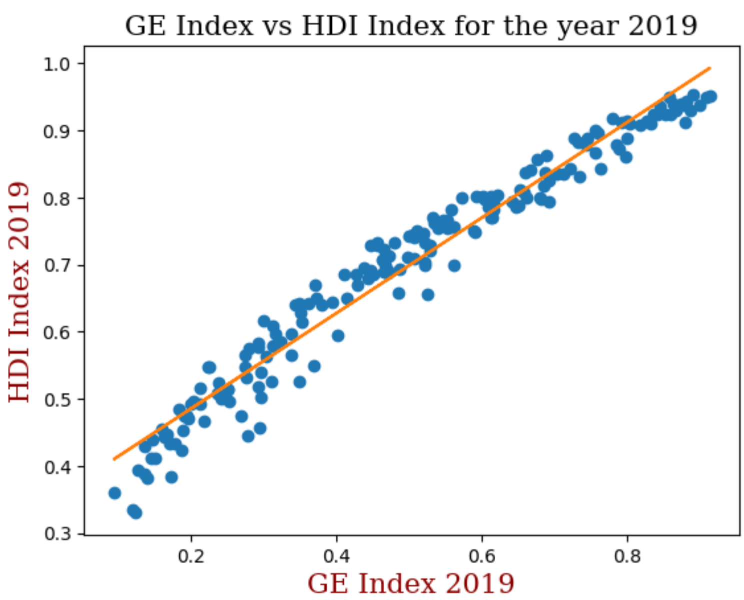


Figure 2.3: The GE index is calculated by dropping the Life expectancy index from the HDI index. A strong correlation between GE scores and the current HDI scores is observed. The Pearson correlation coefficient for the data is 0.98.

Observed limitations of current HDI

From Figure 2.3, we can observe that the HDI indicator has less sensitivity to the Life expectancy indicator. We try to address this issue by incorporating additional parameters in the HDI indicator to make it more comprehensive. From the data, we observed that the sources of data of the indicators still have some errors in measurement. Ideally, the HDI score for a country should be the average of the HDI scores of the entire population. But as it is not possible to get the data of GNI, Life expectancy and education for each individual, we calculate the HDI based on the overall value for a country. Also, the current three indicators of HDI do not include the component of Renewable energy consumption and Gender balance. It has become critical in today's age to think about sustainability and gender balance. From Table 2.2, we observe that the minimum value of the Life expectancy index is set to 20 years, which is lower compared to the life expectancy values of all countries. Hence, we propose to set the minimum value to 40 years.

We enlist the key limitations observed in the current HDI below.

1. $I_{\text{Education}}$, $I_{\text{Life Expectancy}}$, I_{HDI} are highly correlated to $I_{\text{Gross National Income}}$
2. HDI can be predicted by dropping one or two of the three indicators of HDI.
3. The minimum life expectancy is set to 20 years, which is very low.
4. Time-dependent correlation between the parameters.

2.3 Results

2.3.1 Proposed improvements to the current HDI indicator

1. **Re-balancing the Life expectancy index:**
   We replace the minimum value from 20 years to 40 years for re-balancing the index.

   The new Life expectancy index can be calculated in the following way,

$$\frac{\text{actual value} - 40}{85 - 40}$$

   It can be observed from Table 2.3 that the correlation between the Life expectancy index and HDI increased after changing the minimum value of the Life expectancy index. After rebalancing, all three indicators are equally correlated to HDI. This results in all three indicators having equal weightage.

2. **Proposing New HDI Index**

   The current HDI does not give a broader measure of human development. Hence we propose to add two new parameters to the existing HDI.

Years	Correlation coefficient after re-balancing	Correlation coefficient before re-balancing	Variance of LE with minimum set to 40 years	Variance of LE with minimum set to 20 years
2015	0.936	0.92	0.03	0.014
2016	0.936	0.91	0.029	0.014
2017	0.936	0.91	0.028	0.014
2018	0.935	0.914	0.028	0.013

Table 2.3: The correlation coefficient of the Life Expectancy Index with HDI before and after re-balancing for different years is given in the table.

We use the following points while choosing an indicator which is also supported in [21].

(a) Analyzing how the chosen indicator will be useful and related to measuring the phenomenon.

(b) Assessing the quality of data. If there are a large number of missing values or the data is not sufficient in size.

(c) Check, if the indicators to be chosen are different from the other indicators.

(d) Indicators should be normalized to ensure appropriate analysis.

Addition of two new indicators: Per cent Renewable Energy consumption from total energy and Gender Balance Index [19].
Renewable energy consumption is the share of renewable energy in total energy consumption. The Gender Balance Index is the arithmetic mean of two indicators; Birth rate (crude) and Gender Parity Index (GPI)-School enrollment, primary and secondary (gross). We normalize the Per cent Renewable Energy consumption from the total energy indicator by the formula 2.1. While for normalizing Gender Ratio at the birth indicator and the Gender Parity Indicator, we use the formula of $e^{-(\text{value}-1)^2/\sigma^2}$. We choose the σ equal to 0.03 for the Gender Ratio index, while for the Gender Parity index, we set it to 0.04. We also use the existing three indicators of HDI. Despite their high correlation, the three indicators of HDI remain relevant because they exhibit different growth rates over time. The GDP is the most varying index, which would change if the economy has some drastic change, while life expectancy is the most stable. The indices used for the modified HDI index are:

(a) HDI Index

(b) Percent of Renewable Energy Consumption in the Total energy

(c) Gender Balance Index

The modified HDI formula is

$$(I_{REI} * I_{GBI} * I_{HDI}^3)^{1/5} \tag{2.3}$$

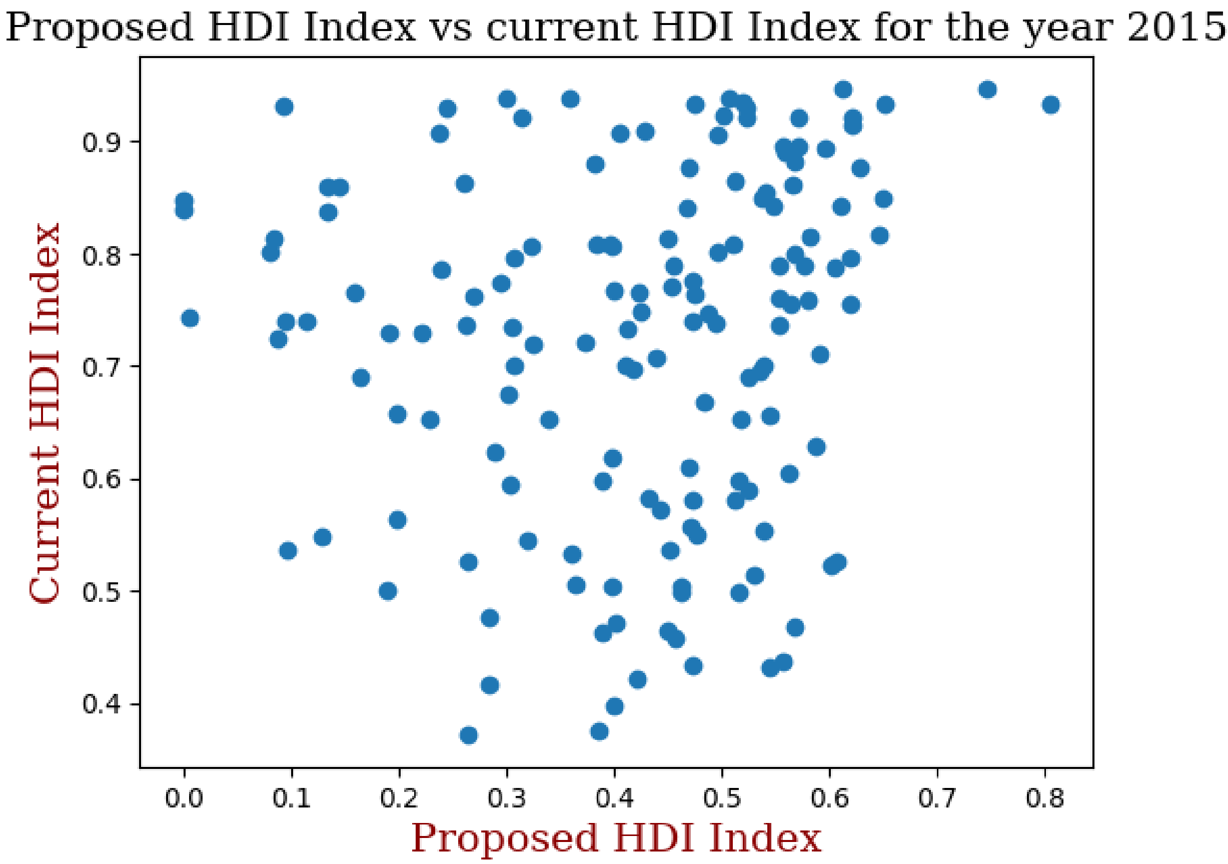
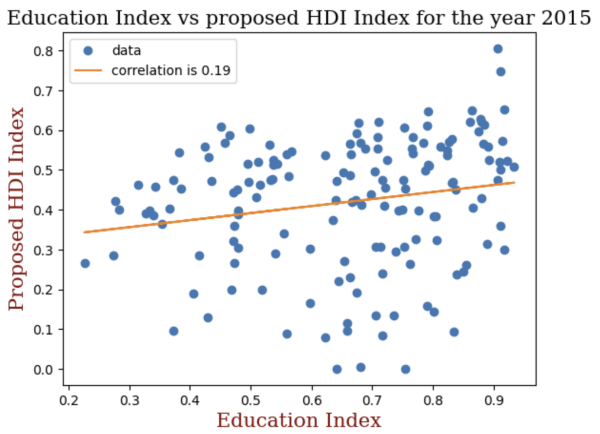
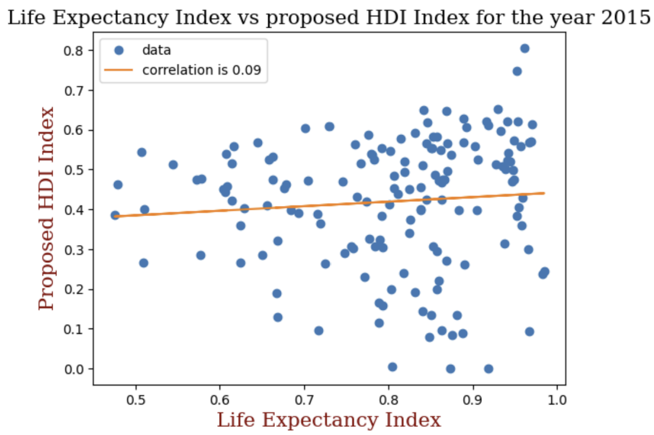


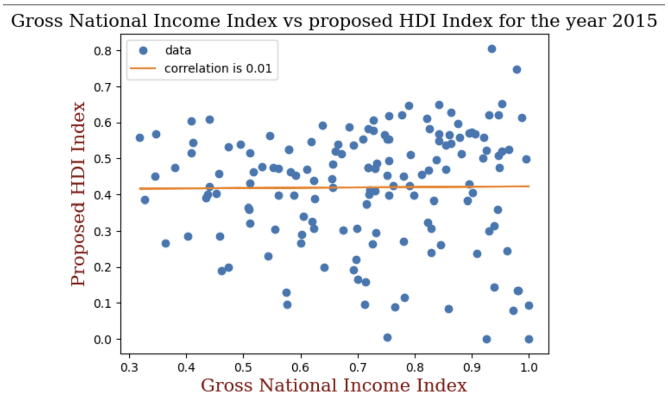
Figure 2.4: Scatter plot of proposed New HDI Index vs the current HDI Index for the year 2015



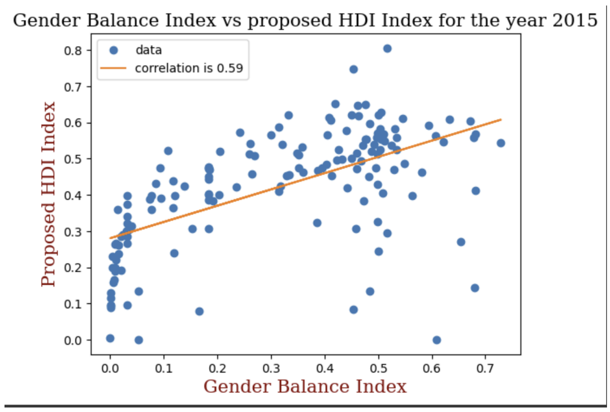
(a)



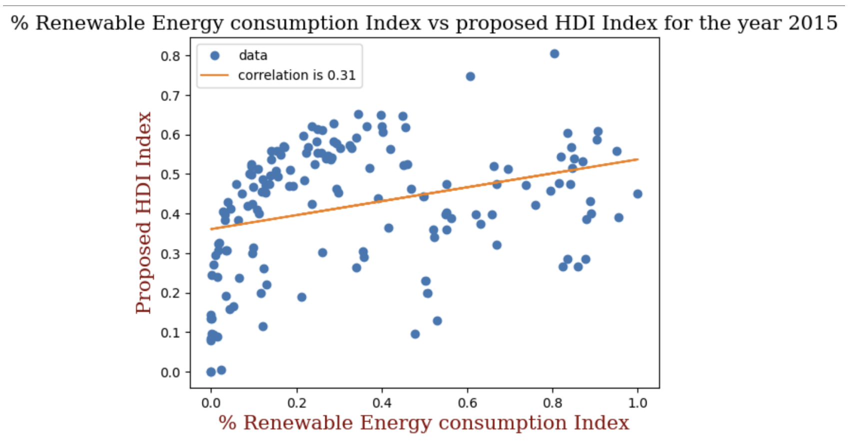
(b)



(c)



(d)



(e)

Figure 2.5: Figures **(a)**, **(b)**, **(c)**, **(d)**, **(e)** represent the scatter plots of five indicators of proposed HDI vs the proposed HDI Indicator. It can be observed that there is a low correlation between each of the indicators and the proposed HDI indicator.

Countries	Proposed HDI Rank	Current HDI Rank
Iceland	1	7
Norway	2	1
Denmark	3	9
Latvia	4	38
Montenegro	5	46
Estonia	6	29
Austria	7	18
Canada	8	15
Brazil	9	77
Costa Rica	10	58

Table 2.4: Ranks of the top 10 countries in the proposed HDI and the current HDI for the year 2015. The most change in rank is of the country Brazil while the least change in rank is of the country Norway.

In Table 2.4, Iceland, Norway, and Denmark do not have significant changes in their rankings as compared to Latvia, Montenegro, Brazil, and Costa Rica. However, Estonia, Austria, and Canada exhibit slightly more changes in their rankings than the top three countries. Brazil has the most change in rank. This is because it has more consumption of renewable energy than most developed nations [22, 23].

2.4 Conclusion

We aimed to analyze the current HDI index. The revised HDI takes a more comprehensive approach to measuring human development. Unlike the existing HDI, the modified HDI incorporates renewable energy consumption, gender parity, and the male-to-female ratio at birth, which are important aspects of human development. Furthermore, from Figure 2.5, we observe that the indicators used in the new HDI are not highly correlated with each other, indicating that each one is equally valuable in assessing human development. We also proposed to set the minimum value of the Life expectancy index to be 40 years, which is a realistic value as all countries have their life expectancy values (years) above 40 years. We also observed that each indicator has a weaker correlation with the proposed HDI, making it more reasonable. We include the three highly correlated indicators of HDI as well because we note that they are useful in capturing both the slow and faster variations in development, i.e. while the national income captures the fast varying component of the development, the life expectancy captures the slowest component of development. Significant room for improvement remains within the HDI, with one particular area being the consideration of inequality within a country into the HDI ranking.

Part II

Introduction to modelling income distributions

Chapter 3

Introduction and Literature review

Income inequality has become a prominent issue worldwide, with increasing attention paid to the disparities in income distribution across different social groups. The study of income distribution is not only of theoretical interest but also has practical implications for public policy, social welfare, and economic growth. Along with the modelling of income, we also model wealth distribution by accumulating the savings of an individual over time. The wealth of a person is often defined as accumulated income or assets over a period of time. Wealth can take many forms, including physical assets such as property and financial assets such as stocks and investments, which can be difficult to value and track. In contrast, income is typically reported more regularly and accurately through tax and other financial records. Furthermore, wealth is often passed down through inheritance, which can further complicate efforts to measure its distribution accurately.

By modelling income distribution, policymakers and researchers can gain a better understanding of the drivers of inequality and identify policy interventions to address it. Modelling can help identify how different variables, such as the capability of an individual or the expenses, can impact income distribution.

Before discussing our work, we provide a brief history of fitting various distributions to the income data and summarize the previous research conducted in this area. Pareto, in 1896 [24], was the first to discover that income follows a Pareto distribution. Pareto distribution is a type of power law probability distribution. It was built on his observation that 80% of the income is accumulated with 20% of people. From the observed tax data, Pareto stated that the incomes at the top are well characterized by a power law.
A power law is given by the equation,

$$N(x) = cx^{\alpha}$$

where c and α are constants, α is called the scaling exponent, N is the number of households with income greater than x.

Early on, there was a debate about whether income distribution has Pareto or lognormal distribution. According to Gibrat, lognormal distributions are generated by the law of proportionate effect [25, 26]. In the paper [27], the author mentions that the distributions-power law and lognormal distribution have similar generative mechanisms. The question of whether lognormal or power law fits the income distribution was addressed by Aitchison and Brown in their paper [28]. Aitchison and Brown proposed that lognormal may fit the income distribution better. As much as it is important to understand the distribution of income, it is also important to know the underlying factors leading to a Pareto distribution. Many researchers worked on finding factors affecting the distribution of wealth and income. Piketty [29] proposed a theoretical framework in his work published in 2014, which explains the underlying factors that influence wealth and inequality. According to him, "If a 30% tax rate is applied on all types of capital, it can lead to a significant reduction in the concentration of wealth." A simple model to explain the evolution of income is described in the paper [30].

Evolution of income

Assume that people are exponentially distributed across age. Hence,

$$P[Age > x] = e^{-\delta x} \tag{3.1}$$

where δ denotes the death rate in the population.

Assume that income y grows exponentially with age at rate δ. Then,

$$y = e^{\mu x} \tag{3.2}$$

Inverting this assumption gives us the age at which an individual earns income y. The expression for age at which an individual earns income y is,

$$x(y) = \frac{1}{\mu} log(y)$$

So, the Pareto inequality can be derived in the following way,

$$P[income > y] = P[Age > x] \tag{3.3}$$

Using equation 3.1 and equation 3.3, we get,

$$
\begin{aligned}
P[income > y] &= e^{-\delta x(y)} \\
&= e^{-\delta(\frac{1}{\mu}log(y))} \\
&= e^{log(y^{\frac{-\delta}{\mu}})} \\
&= y^{\frac{-\delta}{\mu}}
\end{aligned}
\tag{3.4}
$$

This simple derivation gives us the Pareto inequality index, which is the factor $\frac{\mu}{\delta}$. For the wealth inequality index, additional factors come into the picture. Wealth inequality is influenced by the inheritance of wealth, disparities in education and human capital, the concentration of assets, tax policies, expenditure, and social and demographic factors. This makes wealth distribution more heavy-tailed than income distribution. Income distribution primarily reflects income over a specific period, while wealth distribution reflects the accumulation of assets over time. The question of whether or not talent or human capacity is a factor affecting the distribution of income or wealth has been discussed for many years. The papers [31, 32] investigate wealth accumulation by financial investments. It describes that the unequal distribution of wealth at the higher range is by chance and not dependent on the skills of humans. In this project, one of the major goals is to study the evolution of income, assuming that the human capacity is normally distributed.

This report covers the approaches to studying income distributions from analytical and stochastic modelling perspectives using the minimum assumption that the capacities of individuals have a normal distribution [33, 34, 35]. Our research is motivated by the need to understand the drivers of income inequality better and provide insights into potential policy interventions to reduce inequality. It is also of interest to study if human capability has any role in determining one's income.

The first part of our study focuses on the analytical modelling of the income distribution. For analytical modelling, we use the inverse transform method for modelling income distribution under different assumptions. In this part, we are able to interpret the log-normal as well as the Pareto component in the income distribution [27]. Our modelling approach is based on minimal assumptions, making it particularly useful for informing policy decisions. We compare these analytical distributions with empirical data of countries for different years and study the evolution of income.

The second part is motivated by the paper [36], which debates the role of randomness in success and failure. We use the Sequential Monte Carlo technique to simulate the evolution of income distribution over time. We discuss the evolution of income from a uniform distribution to a Pareto distribution with time and its relationship with the capacity in the two cases of differential rewards and differential opportunity. The former assumes a stochastic increase of income in proportion to the capacity of an individual. In addition, the latter assumes that the probability of an increase in income is non-uniform and proportional to the capacity. We develop a dynamic model which studies the relationship between human capability and income. We simulate the model to imitate the behaviour of income and, subsequently, wealth and compare the distributions.

Overall, our research provides a comprehensive analysis of income and wealth distributions using both analytical and stochastic modelling techniques. We aim to offer a better understanding of the evolution of income with time and also study the accumulation of savings to analyse the wealth distribution of a system of population.

Chapter 4

Methodology

4.1 Income modelling

4.1.1 Inverse transform method

Using the inverse transform method, we model income analytically. The inverse transform method is used when there is a need to transform a probability distribution from one random variable to another.

Suppose X is continuous with probability density function $f_X(x)$. Let $y = h(x)$ with h a strictly increasing continuously differentiable function with inverse $x = g(y)$. Then $Y = h(X)$ is continuous with probability density function $f_Y(y)$ given by

$$f_Y(y) = f_X(g(y))g'(y) \tag{4.1}$$

The growth of income is observed to be a multiplicative process. Starting from the initial income I_0, the current income is proportional to the previous income. This represents the case of the differential reward. The multiplicative process is particularly useful in this case, as the growth of income is dependent on the capability of an individual and cannot be described using simple linear models. The proportional random growth creates a Pareto-like distribution. Starting from the initial income I_0, the next income will be $I_0(1+cx)$. The following equations 4.2 and 4.5 represent income growth. We find the density function of Income I in the following two cases.

Case 1:

$$I = I_0 \cdot (1 + cX)^t, \tag{4.2}$$

where I_0 is the initial income, X the capability distribution with $\mathcal{N}(0.5, 0.15)$, I income distribution, c normalizing constant.

X is a continuous random variable with a Gaussian distribution. I is an increasing continuously differentiable function of X, hence, from equation 4.1 inverse exists. In equation 4.2, income in the next time step is proportional to the previous income.

$$X = \frac{1}{c}(\frac{I}{I_0}^{1/t} - 1) \tag{4.3}$$

Equation 4.3 gives the inverse function. By putting the values of I ranging from I_0 to I_{max} (I_{max} is calculated using the mean and standard deviation of capacity distribution), we find the values of X associated with the income I for each t. Then, the probability density function for I is given by

$$f_I(i) = e^{-(X_{up}-\mu)^2/2\sigma^2} \cdot \frac{dX}{dI} \tag{4.4}$$

where,

$$\frac{dX}{dI} = \left(\frac{c \cdot t \cdot I}{(1+cX)}\right)^{-1}$$

and X_{up} are the values obtained by evaluating equation 4.3 at the values of I. In the following equation, the growth is proportional to previous income, as well as the capability distribution.

Case 2:

$$I = I_0 \cdot (1 + cX)^{cXt} \tag{4.5}$$

where I_0 is the initial income, X the capacity distribution with $\mathcal{N}(0.5, 0.15)$, I income distribution, c normalizing constant.

X is a continuous random variable with a Gaussian distribution. The inverse function for equation 4.5 does not have a closed form. Hence, we use a numerical method to estimate values of X.

$$f(X) = c \cdot X \cdot ln(1 + cX) - \frac{1}{t} ln\frac{I}{I_0} \tag{4.6}$$

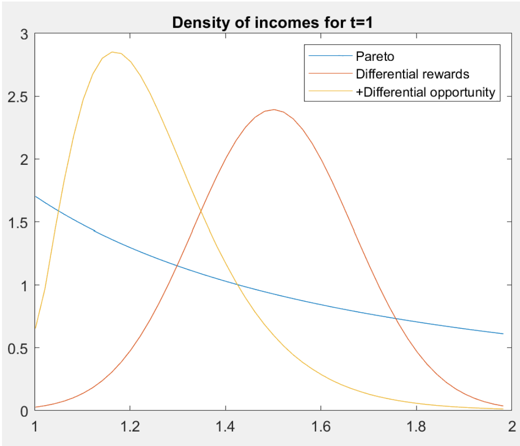
We obtain the values of X by applying the bisection method to equation 4.6. Hence, the density function for income I is

$$f_I(i) = e^{-(X_{up}-\mu)^2/2\sigma^2} \cdot \frac{dX}{dI}, \tag{4.7}$$

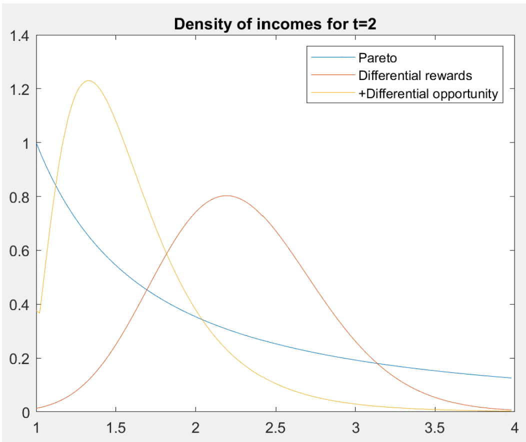
where X_{up} are the values obtained by the bisection rule and

$$\frac{dX}{dI} = \{y[(\frac{c^2Xt}{1+cX}) + c \cdot t \cdot ln(1+cX)]\}^{-1} \tag{4.8}$$

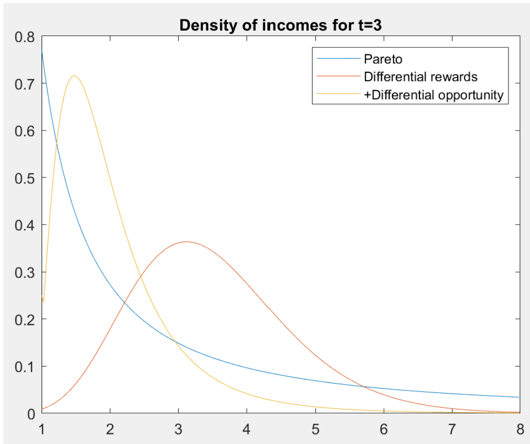
4.1.2 Results for Inverse Transform Method



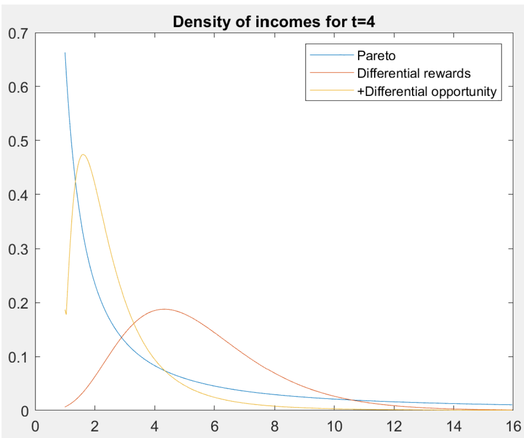
(a)



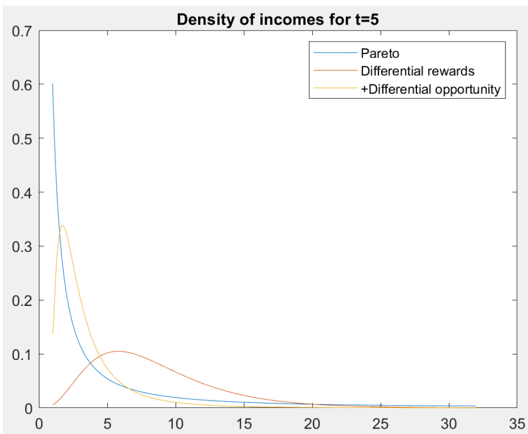
(b)



(c)

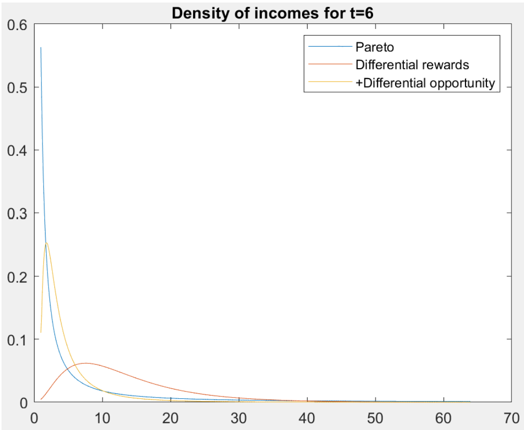


(d)

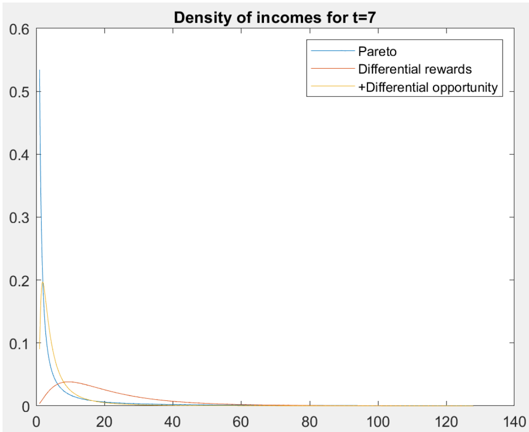


(e)

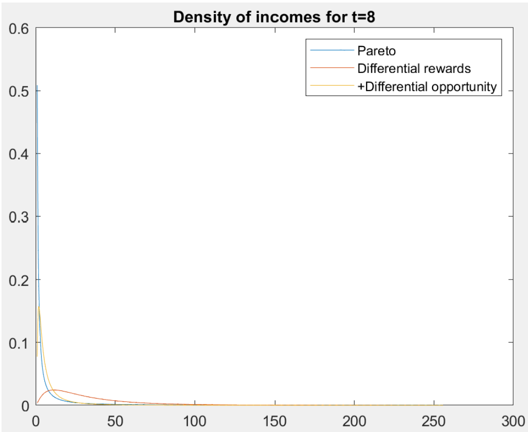
Figure 4.1: Figures 4.1a- 4.1e show the evolution of income as the number of iterations increases from $t = 1$ to $t = 5$.



(f)



(g)



(h)

Figure 4.1: Figures 4.1a-4.1h show the evolution of income for the analytical case. The curves for the differential opportunity and differential rewards appear to converge to the Pareto curve for an increasing value of t.

From Figures 4.1a -4.1h, it can be observed that the curve for the case of the differential reward slowly evolved into a Pareto-type curve [37].

4.1.3 Sequential Monte Carlo simulation

We study the variations in case (2) of the inverse transform model using sequential Monte Carlo because the equations for including both the positive and negative events become complex. Sequential Monte Carlo, also known as particle filtering, is a computational technique that sequentially approximates the posterior distribution. It is especially efficient in complex models.

Case 1: The income changes in proportion to the capability of the individuals.

Algorithm

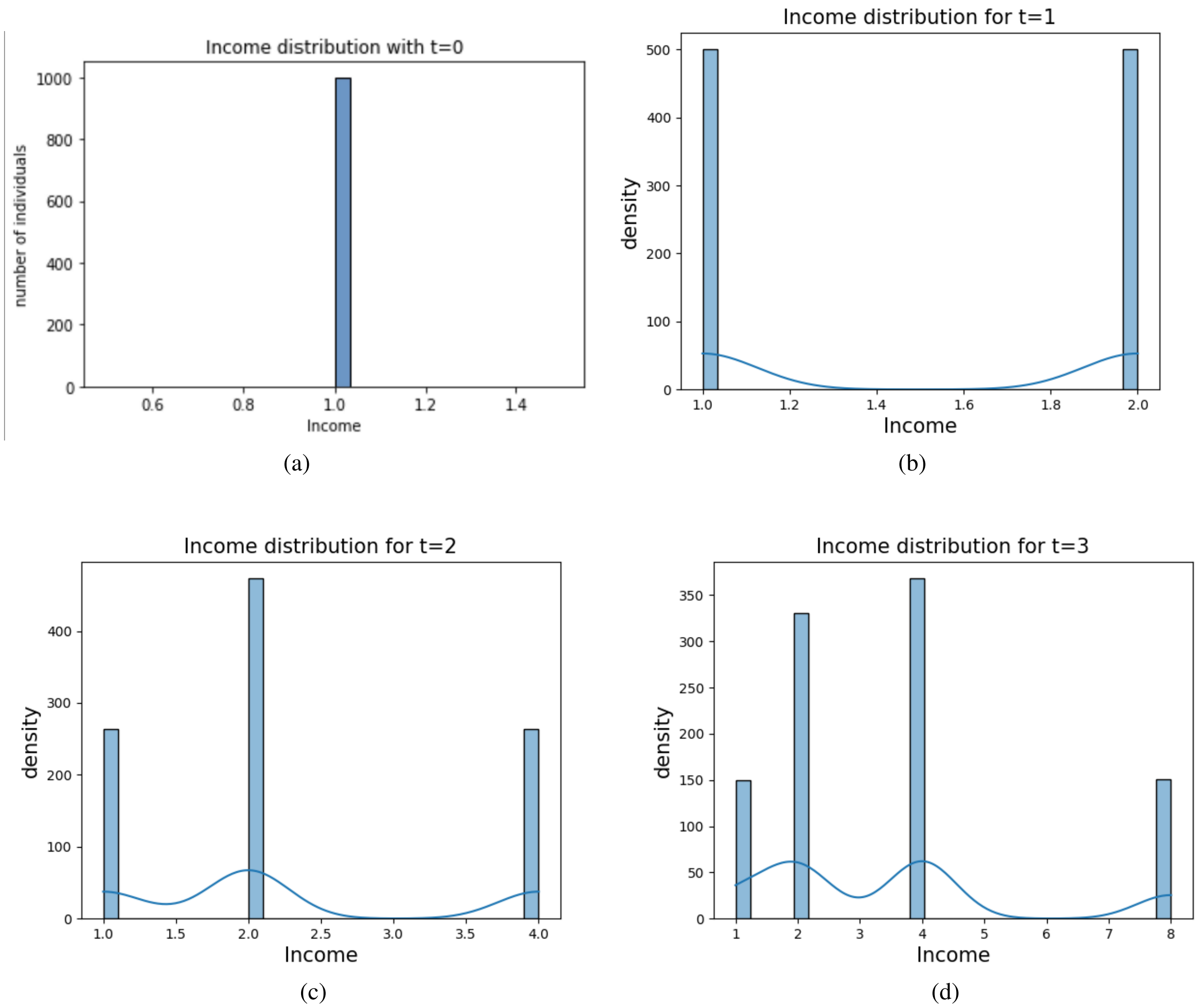
Input:
1. Array of initial income: $I_0 = 1$
2. Capacity distribution array (Normal): $C(i) \sim \mathcal{N}(\mu, \sigma)$, for $i = 1$ to N

Output: Final income array (I)

for k iterations **do**
  for every individual $i = 1$ to N **do**
    if $C[i] > rand(U[0, 1])$ **then** ▷ Similar to Coin toss on an individual
      $I[i] = 2 * I[i - 1]$

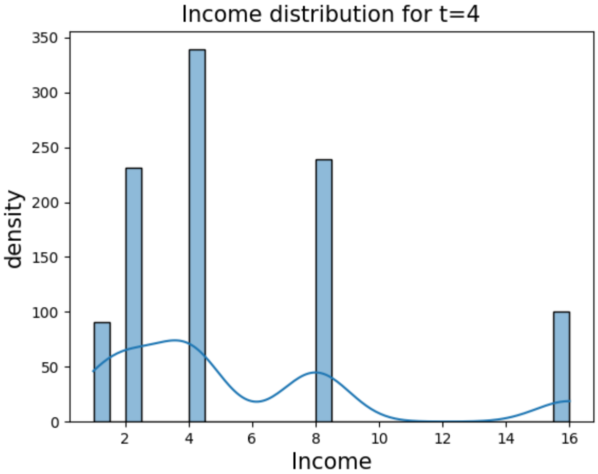
The initial income is 1 unit for all individuals, and every individual is associated with a capacity sampled from a normal distribution with a mean of 0.5 and a standard deviation of 0.15. In each iteration, we check if the capacity of an individual is greater than a random number generated from $U[0, 1]$. This condition signifies that a high-capability individual should have a high probability of income update. In the real world, an individual's potential may not be the only factor in determining their success, as luck also plays a role in whether they are able to seize opportunities or not. This can be referred to as a coin toss. It induces stochastic nature in the model.

4.1.4 Results for Sequential Monte Carlo Case 1

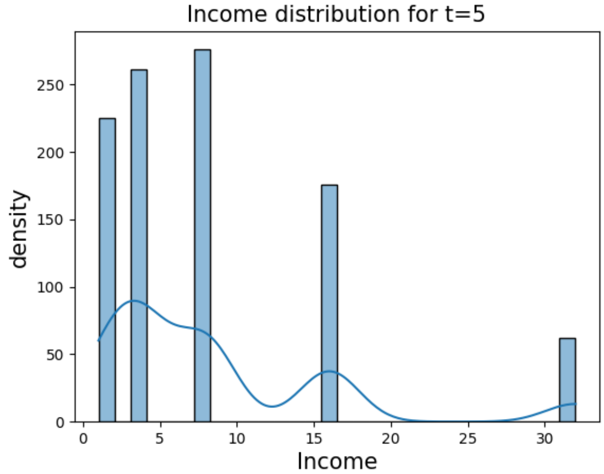


(a) (b) (c) (d)

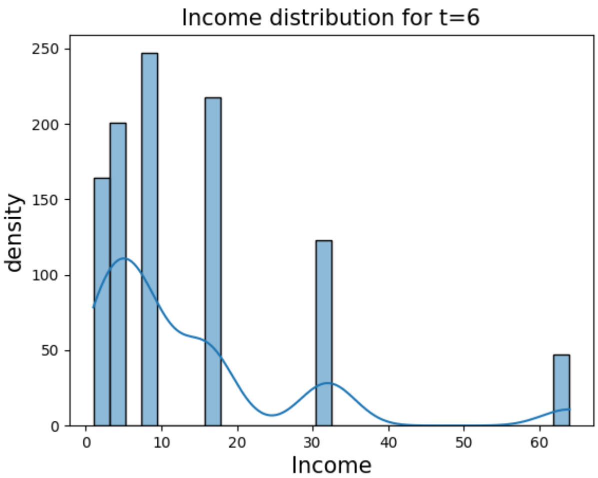
Figure 4.2: Figures 4.2a- 4.2d show the evolution of income as the number of iterations increases from $t = 0$ to $t = 3$.



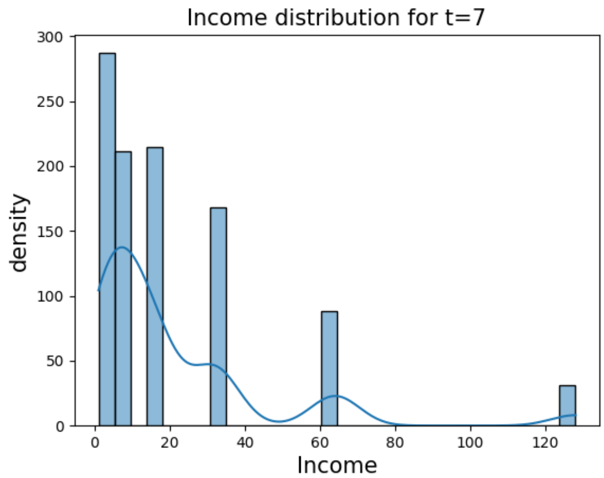
(e)



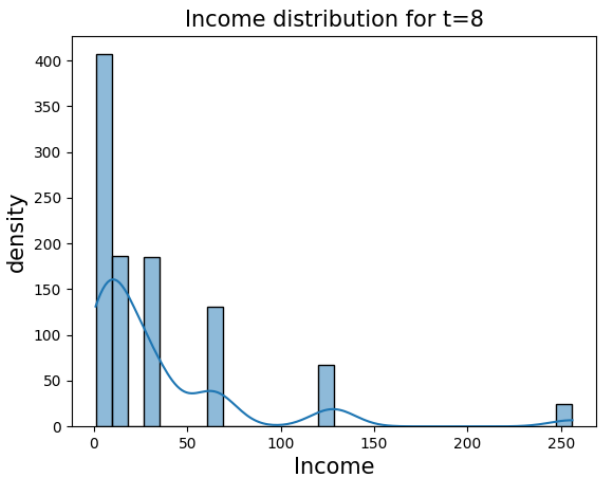
(f)



(g)



(h)



(i)

Figure 4.2: Figures 4.2a-4.2i show the evolution of income for the case where income update is only dependent on the capability distribution.

It can be observed that as the iterations increase from $t = 1$ to $t = 8$, the distribution appears to follow the power law distribution.

Case 2: The income changes in proportion to the capability, positive and negative events.

Algorithm Case 2:

Input: 1. Array of initial income: $I_0 = 1$

2. Capacity distribution (Normal): $C(i) \sim \mathcal{N}(\mu, \sigma)$ for $i = 1$ to N

3. N Individuals associated with position either positive and negative numbers in the ratio 2:1. Position $P(i) \sim rand([-1, 1])$

Output: Final income array (I)

```
for k iterations do
    Shuffle array P
    for every individual i = 1 to N do
        if P[i] > 0 then
            if C[i] > rand(U[0,1]) then          ▷ Similar to Coin toss on an individual
                I[i] = 2 * I[i − 1]
        else
            I[i] = I[i − 1]/2
```

The initial conditions of the model are:

N=1000, where N is the number of individuals. $I_0 = 1$, where I_0 is the initial income. C: Capability distribution with mean $\mu = 0.5$ and $\sigma = 0.15$. The ratio of positive to negative events is 2:1.

N individuals are associated with a position $R(i) \sim rand([-1, 1])$, and initial income $I(i) = I_0 = 1$. Each individual has capability $C(i) \sim \mathcal{N}(\mu, \sigma)$. The capability distribution is fixed at the start of the simulation, but the income is updated at each time step (iteration). At each time step t, for every individual, if $R(i) \in [0, 1]$ for $i \in \{1, \ldots k\}$, i.e. positive event occurs, and if $C(i) > rand(U[0, 1])$ then the income of an individual with position $R(i)$ is double of the previous income. If $R(i) \in [-1, 0)$ for $i \in \{1, \ldots k\}$, i.e. negative event occurs, then the income is half of the previous. At the end of the simulation, we get the final income distribution.

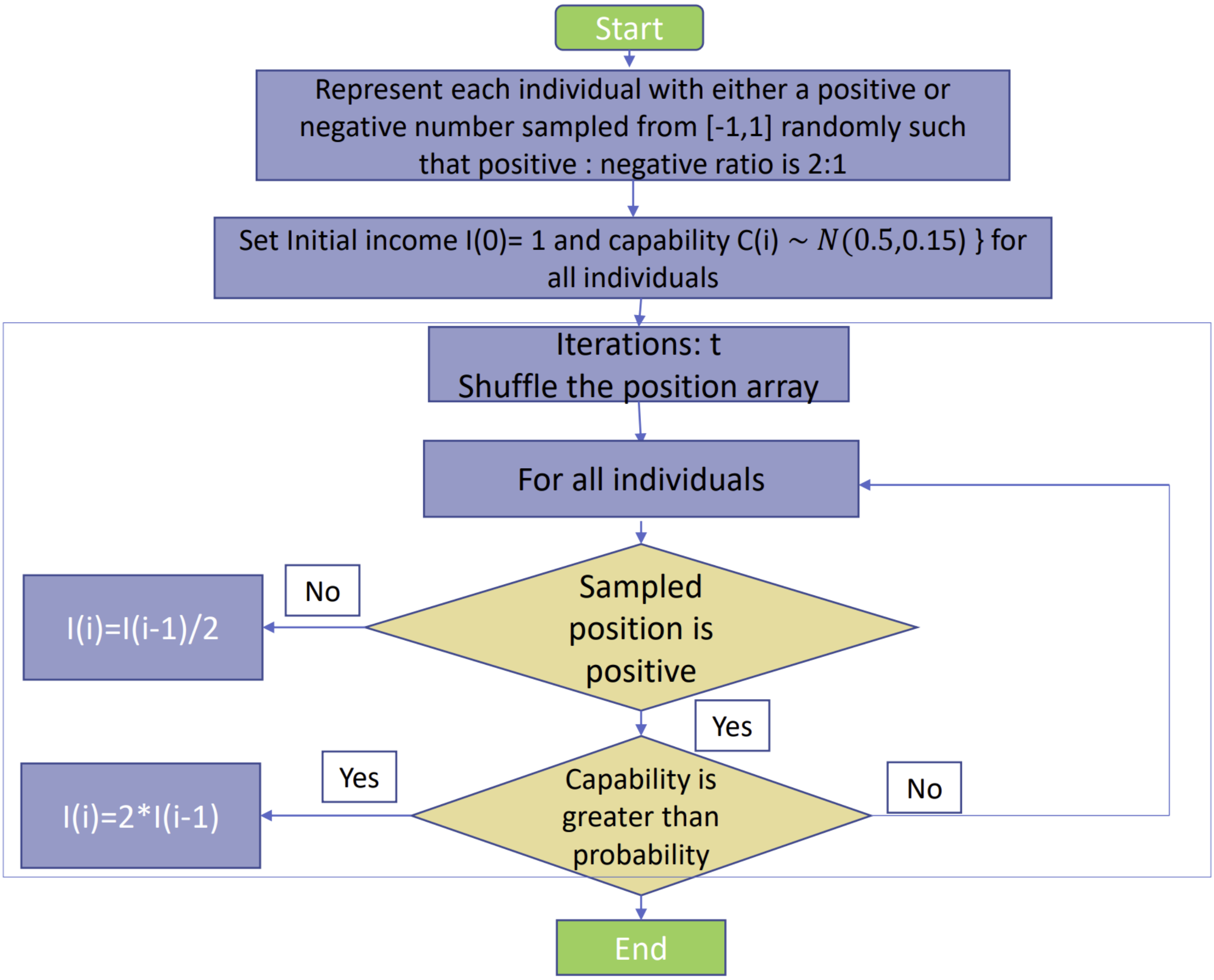
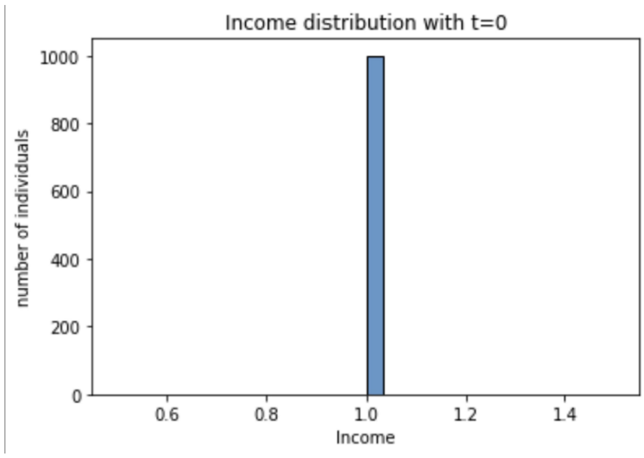


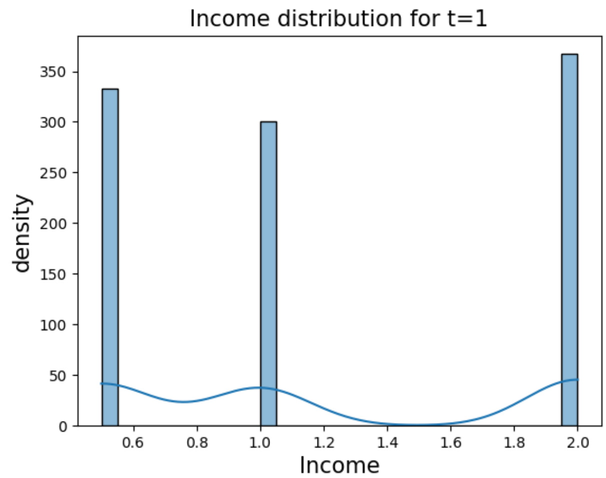
Figure 4.3: Flowchart representing the working model

The flowchart in Figure 4.3 represents the case (2) model.

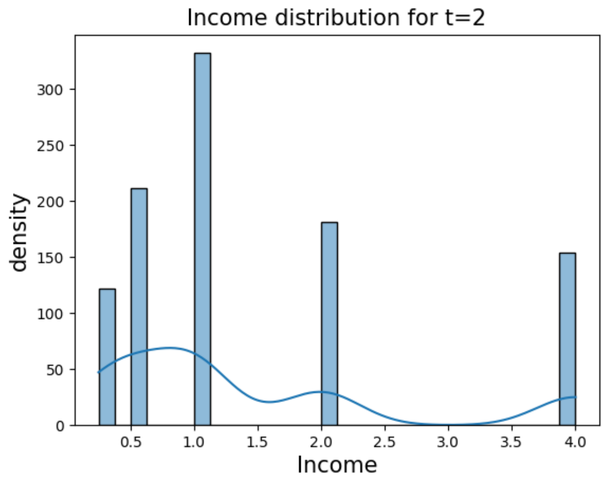
4.1.5 Results for Sequential Monte Carlo Case 2



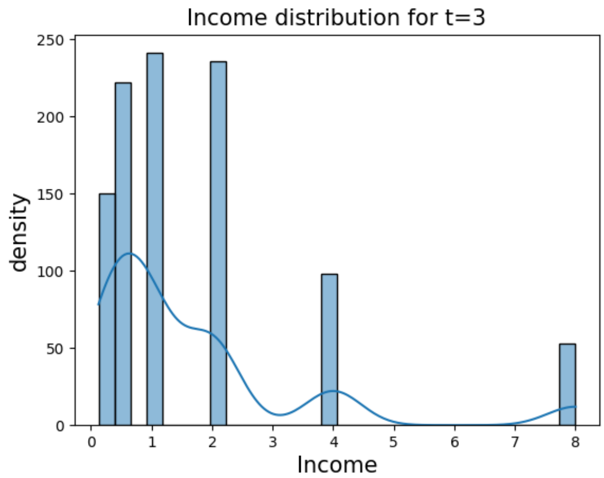
(a)



(b)

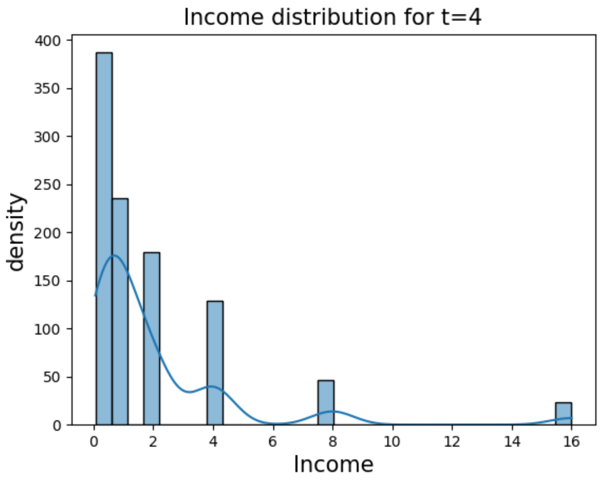


(c)

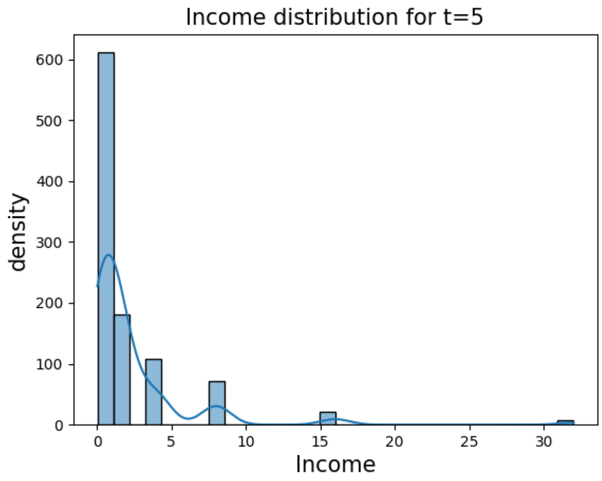


(d)

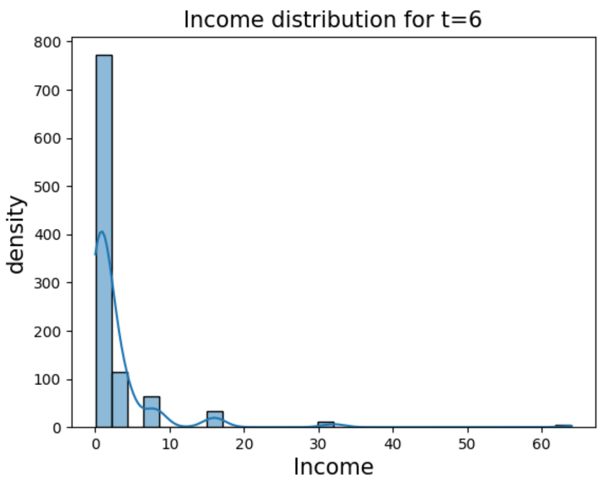
Figure 4.4: Figures 4.4a- 4.4d show the evolution of income as the number of iterations increases from $t = 1$ to $t = 3$.



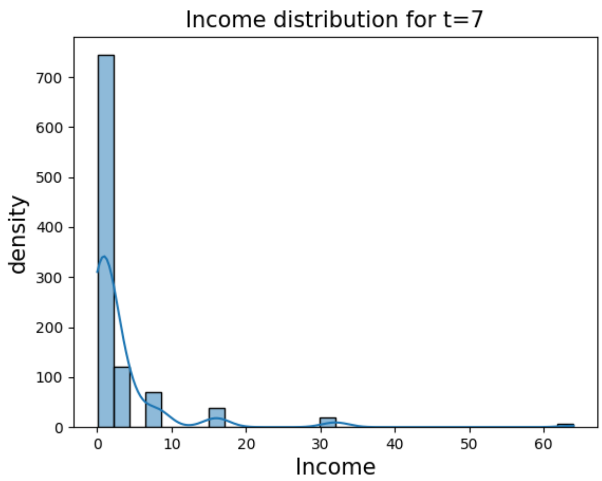
(e)



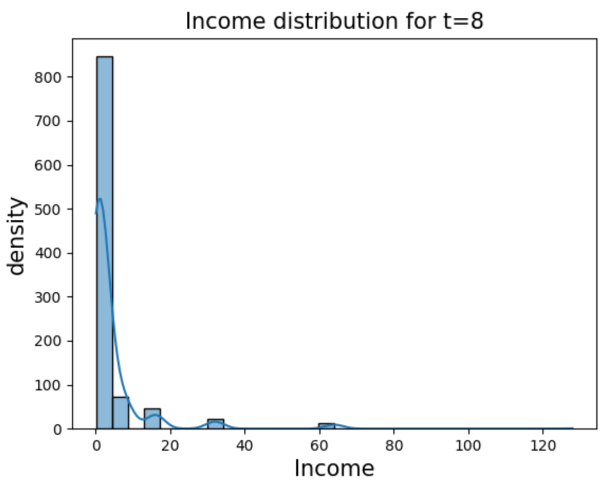
(f)



(g)



(h)



(i)

Figure 4.4: Evolution of income for the case where income update is dependent on the capability distribution as well as positive and negative events.

In Figures 4.4a - 4.4i, it can be observed that as the iterations increase from $t = 1$ to 8, the distribution appears to follow power law distribution.

4.1.6 Household annual income data distribution

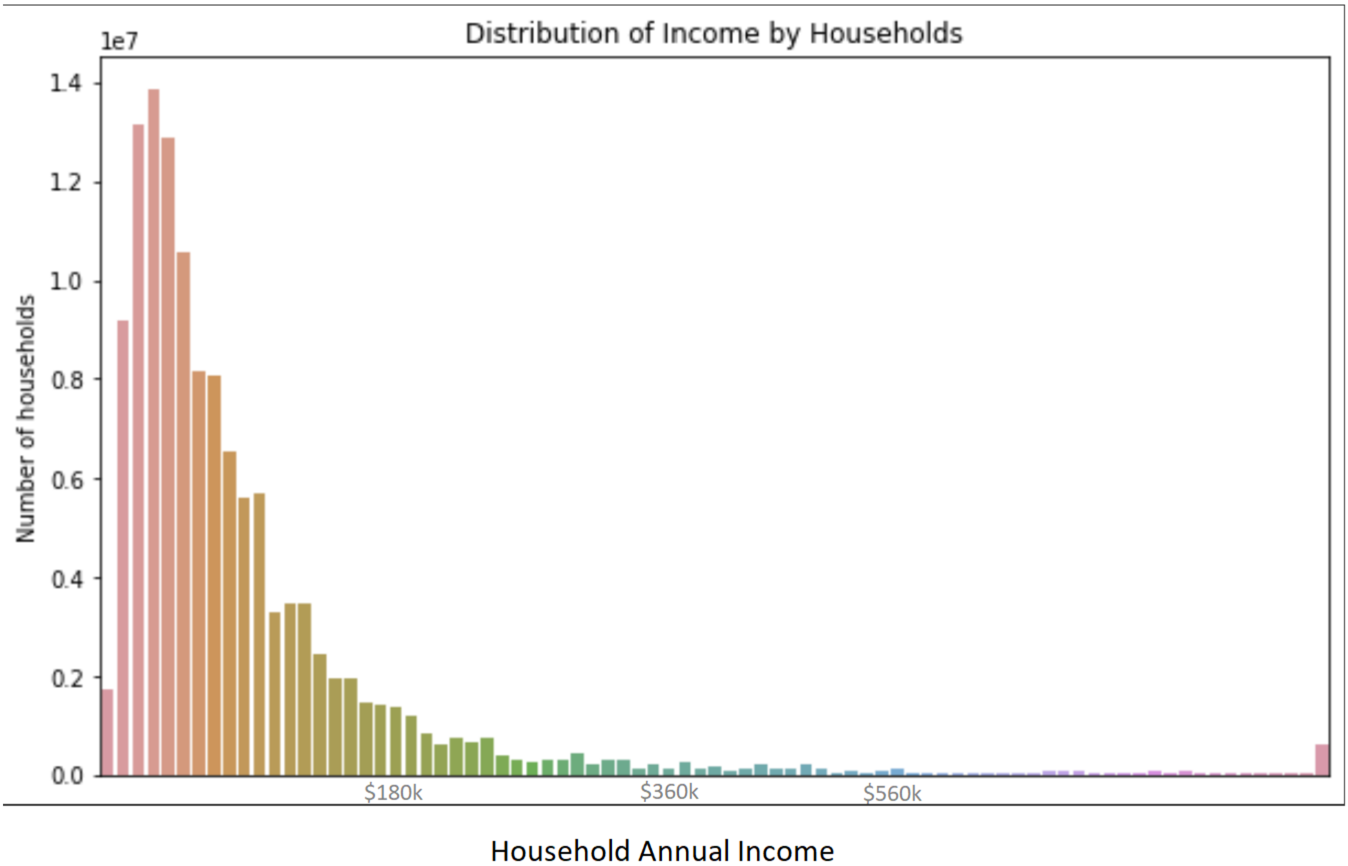


Figure 4.5: This data represents the distribution of income in US annual households.

From Figure 4.5, it can be verified that the income distribution data follows a similar distribution as that of the simulation plots [38]. The tail of the empirical data distribution is heavy-tailed, which suggests it follows the power law distribution.

4.1.7 Observations

Does income rise exponentially with age?

One of the issues with generating the Pareto distribution mechanism is that the derivation is dependent on the fact that income rises exponentially with age. But from Figure 4.6 of US median household income data, we find that the median income rises for the ages 15-54 and then drops for those ages above 55. Hence, the claim that income rises exponentially with age may not be necessarily true.

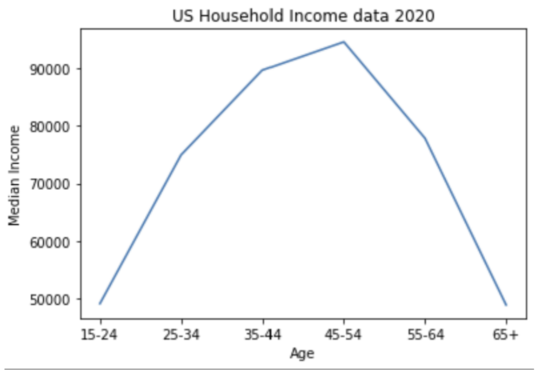


Figure 4.6: US household median income data 2020

4.1.8 Conclusion

Does the highest capability person have the highest income?

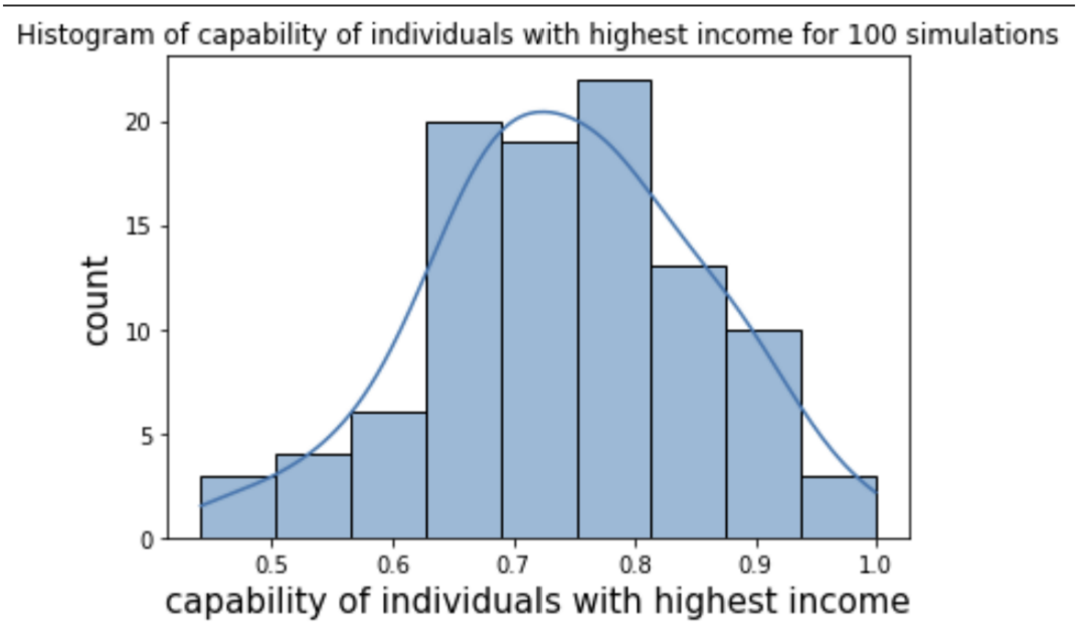


Figure 4.7: Distribution of capabilities of individuals with the highest income over 100 simulations.

As observed from Figure 4.7, The capability of individuals with the highest income for 100 simulations has a mean of around 0.7. This is because the CDF of a uniform random variable is a straight line with an increasing slope. As the capability of individuals is drawn from a normal distribution, most of the values lie around 2 standard deviations from the mean of the normal distribution. Hence, the mean shifts from 0.5 to 0.7. This is because, as the capability value increases, the probability that the value is greater than a number from uniform [0,1] increases. Therefore, the highest capability person is necessarily the person with the highest income.

4.2 Wealth distribution

We extend income modelling and include savings as a variable which will accumulate over time. In this model, we assume all individuals have equal initial income. The stochastic factor comes from the individual's ability to take advantage of the opportunity. Although wealth consists of many factors, such as investments, savings, inheritance, and donations, it is equally difficult to model these factors. We model the wealth distribution by considering different cases for the savings rate.

4.2.1 Constant savings rate

We build a simple wealth model by considering a constant savings rate for all individuals. In this case, we ask the question of how wealth will evolve if each individual allocates a uniform per cent of their income, regardless of the income, in a specific time period.

Algorithm

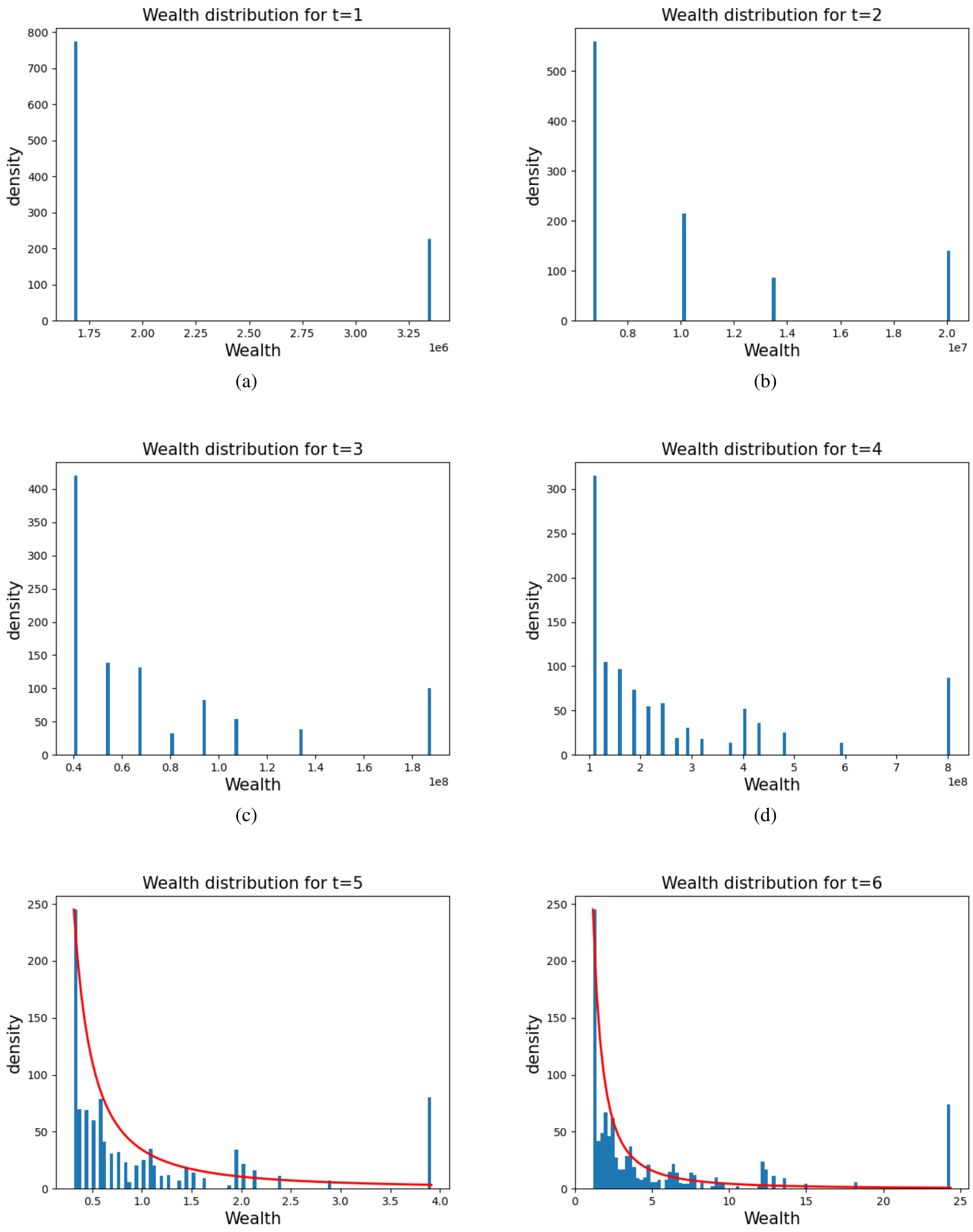
Input:
1. Array of initial income: $I_0 = 1$
2. 2d Capability distribution, with one dimension being the capability of the person, which is drawn from the normal distribution with a mean of 0.5 and standard deviation of 0.15. The second dimension is savings which are initialized to be zero for all individuals.
3. Savings rate: $S_r = 0.2$

Output: Final income array ($\vec{I}$) and final Savings $\vec{S}$

for k iterations **do**
  for every individual $i = 1$ to N **do**
    Create a unit vector of savings
    if 1-norm of 2d capability vector is greater than sum of $rand(U[0,1])$ and mean of unit savings vector **then**
      $\vec{I}[i] = 2 * \vec{I}[i-1]$
    $\vec{S}[i] = \vec{S}[i] + S_r * \vec{I}[i]$

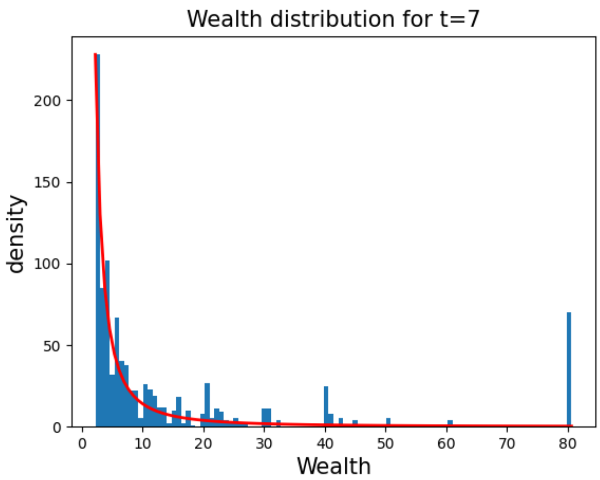
In the wealth distribution model, we assume that the capability distribution is a two-dimensional quantity, with one dimension being the normal distribution and the other dimension of savings. The savings vector updates in each iteration while we fix the normal distribution at the start of the simulation.

Results

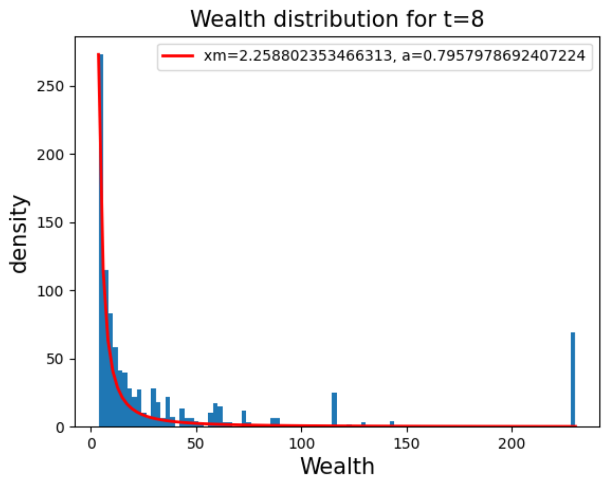


(a) (b) (c) (d) (e) (f)

Figure 4.8: Figures 4.8a- 4.8f show the evolution of wealth as the number of iterations increases from $t = 1$ to $t = 6$



(g)



(h)

Figure 4.8: Figures 4.8g-4.8h show the evolution of wealth as the iterations increase from $t = 7$ to $t = 8$

4.2.2 Linear savings rate

(Dynamic minimum and maximum)

Algorithm

Input:

1. Array of initial income: $I_0 = 1$

2. 2d Capability distribution, with one dimension being the capability of the person, which is drawn from the normal distribution with a mean of 0.5 and standard deviation of 0.15. The second dimension is savings which are initialized to be zero for all individuals.

3. Savings rate vector: $\vec{S_r} = [0.2, 0.2, ..., 0.2]$

Output: Final income ($\vec{I}$) and final Savings $\vec{S}$

---

for k iterations **do**

$I_{min} = 0.05 * max(\vec{I})$

$I_{max} = max(\vec{I})$

for every individual $i = 1$ to N **do**

Create a unit vector of savings

if 1-norm of 2d capability vector is greater than sum of $rand(U[0, 1])$ and mean of unit vector of savings **then**

$\vec{I}[i] = 2 * \vec{I}[i - 1]$

if $\vec{I}[i] < I_{min}$ **then**

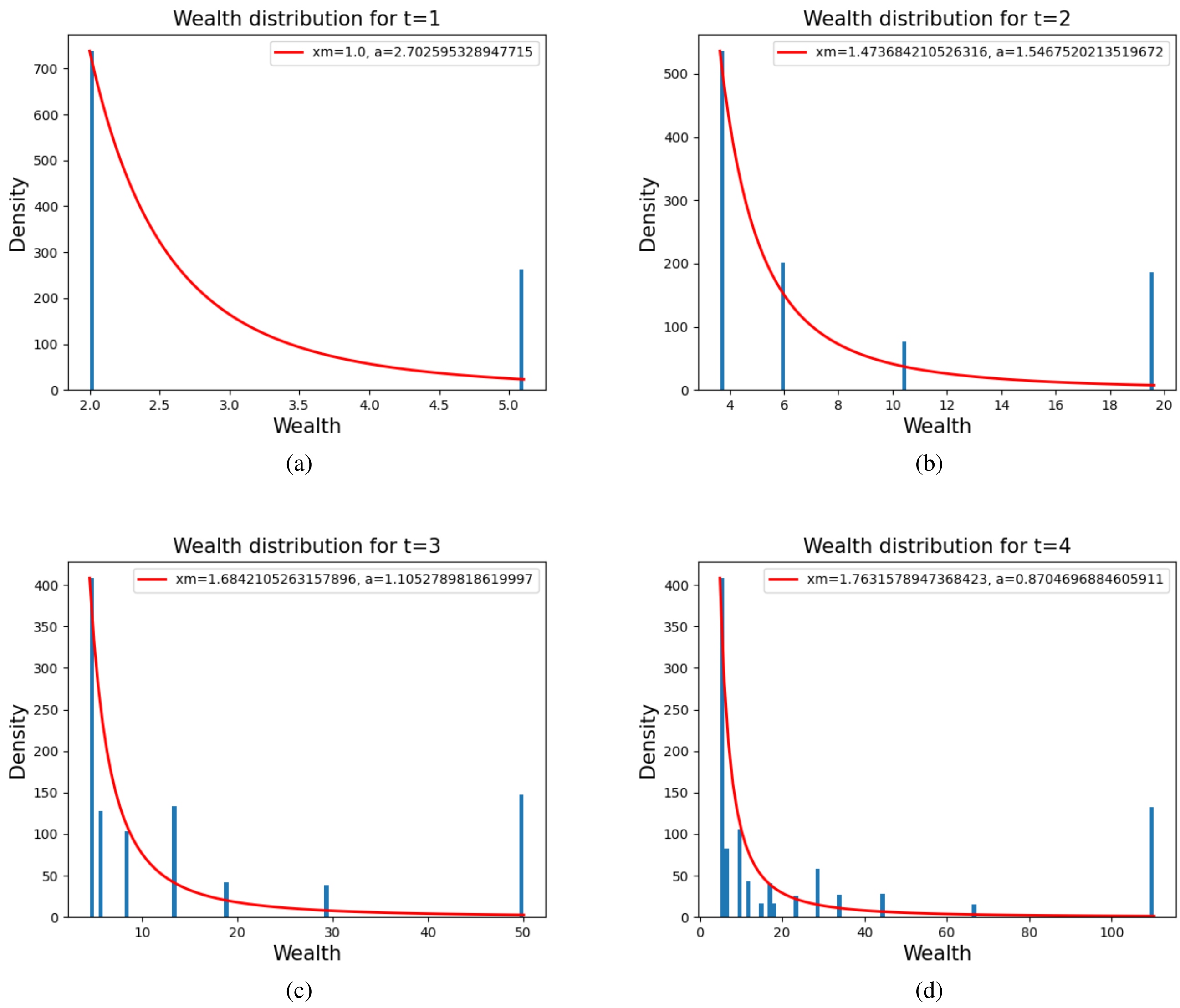
$\vec{S_r}[i] = 0$

else

$\vec{S_r}[i]$ = ($\vec{I}[i] - I_{min}$) / (I_{max} - I_{min})

$$\vec{S}[i] = \vec{S}[i] + \vec{S_r}[i] * \vec{I}[i]$$

Results



(a) (b) (c) (d)

Figure 4.9: Figures 4.9a- 4.9d show the evolution of wealth as the number of iterations increases from t=1 to t=4.

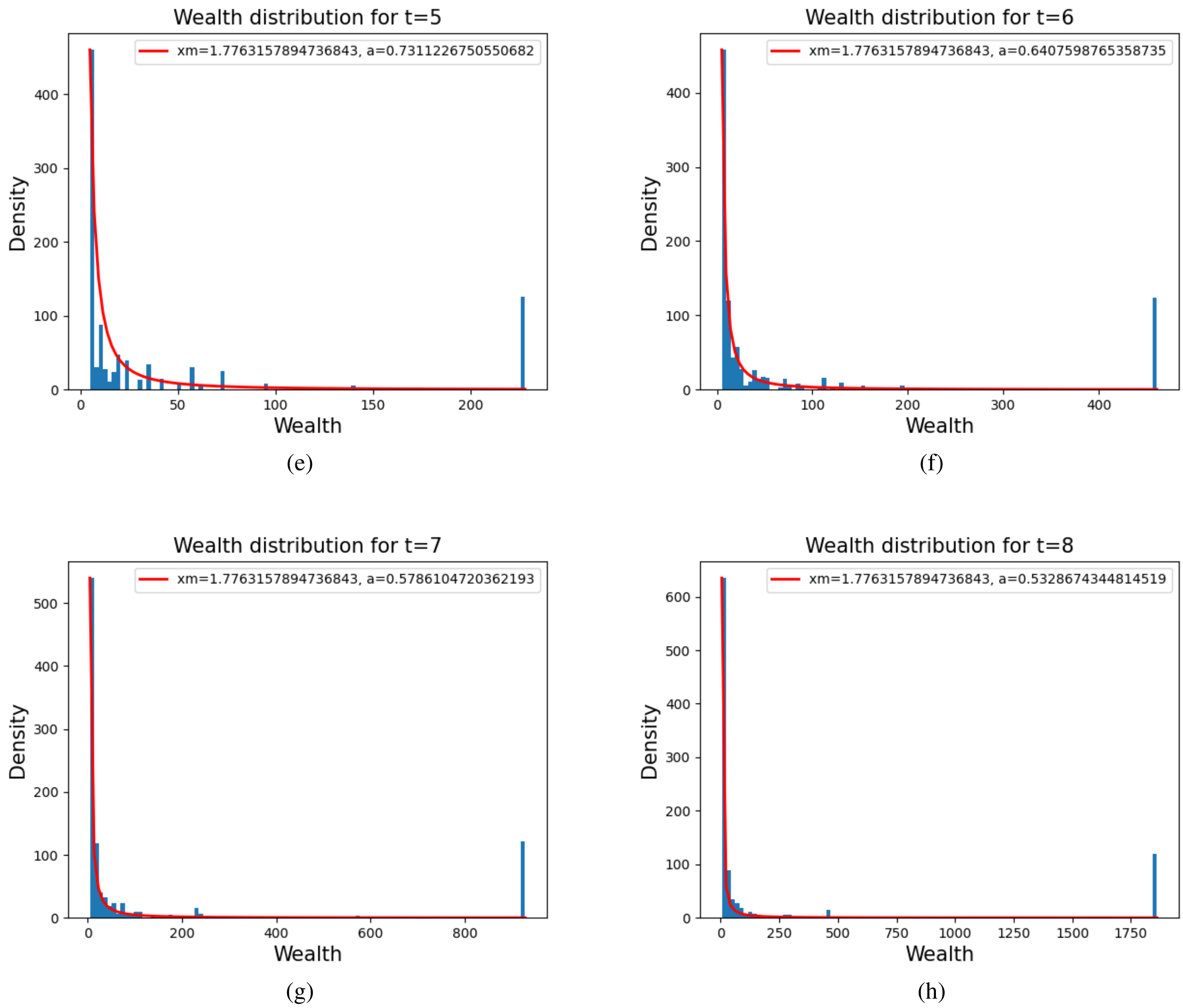


Figure 4.9: Figures 4.9e-4.9h show the evolution of wealth as the iterations increase from t = 5 to t = 8.

In the case of a dynamic linear saving rate, the saving rate changes in every iteration dependent on income. We assume that individuals with incomes at the bottom 5% have zero saving rate. In this case, every individual saves a fraction of their income based on their current income.

4.2.3 Observations

Comparison of convergence of income and wealth distribution to Pareto

	Income model (Case 2)	Wealth Model (Constant saving rate)	Wealth Model (Dynamic saving rate)
α after 8 iterations	1.1	0.69	0.53
x_m after 8 iterations	1.1	1.6	1.77

Table 4.1: Estimation of Pareto parameters for the three models after 8 iterations. As observed from the table, the wealth model with a dynamic saving rate has a small alpha value than the other two models.

We estimate the scale and shape parameters of the Pareto fit for the above models [39]. The Pareto distribution is characterized by the shape parameter α and the scale parameter x_m. The shape parameter α determines the shape of the distribution, while the scale parameter x_m determines the location of the minimum value of the distribution. For the income model, the shape parameter has a value larger than that for the wealth model. The Pareto curve for the wealth model converges faster than that for the income model.

Factors affecting the runway cycle of wealth.

As the income increases in the model, the individual's savings also increases. This is because the savings are proportional to income. However, some factors negatively affect the runway cycle of wealth.

The wealth model presented in this report is a generalized and simple model which showcases the increasing inequality in wealth distribution. As we add more parameters to the model, such as taxes, government laws and policies related to incomes, expenditure, inflation, and so on, the model becomes complex. In real-life scenarios, many such factors affect the change in distribution, which are difficult to model for a large population. These are some of the factors for the increasing Pareto curve.

4.3 Conclusion

This project aimed to model the evolution of incomes with time under a minimal set of assumptions. We discussed the evolution of income using the two stochastic models- differential rewards and differential opportunity. In the case of differential rewards, the probability of an increase in income is proportional to the capacity of an individual. While in the case of differential opportunity, the probability of an increase in income is non-uniform and proportional to the capacity.

We observed that the income and wealth distribution exhibit Pareto distribution. We assumed that the capability of individuals has a normal distribution. In the case of the evolution of wealth distributions, the convergence to Pareto is slow for the constant savings rate compared to the linear savings rate. In the case of a linear saving rate, the individual with higher income has a higher saving rate, and the individual with low income has a lower saving rate, leading to highly unequal wealth distribution.

Many underlying factors impact the distribution of wealth, such as behavioural, geographic, economic and political factors, which can make the model very complex. But some of these factors can indeed be included in the sequential Monte Carlo model used in this work, with little effort.

Bibliography

[1] UNDP (United Nations Development Programme). “Human Development Report 1990”. In: *UNDP (United Nations Development Programme)* (1990).

[2] Mahbub Ul Haq. *Reflections on human development*. oxford university Press, 1995.

[3] Sudhir Anand and Amartya Sen. “Human Development Index: Methodology and Measurement.” In: (1994).

[4] Harald Trabold-Nübler. “The human development index—a new development indicator?” In: *Intereconomics* 26 (1991), pp. 236–243.

[5] Anthony B Atkinson et al. “On the measurement of inequality”. In: *Journal of economic theory* 2.3 (1970), pp. 244–263.

[6] Izete Bagolin and Flavio Comim. “Human Development Index (HDI) and its family of indexes: an evolving critical review”. In: *revista de Economia* 34.2 (2008), pp. 7–28.

[7] Thirukodikaval Nilakanta Srinivasan. “Human development: a new paradigm or reinvention of the wheel?” In: *The American Economic Review* 84.2 (1994), pp. 238–243.

[8] Douglas A Hicks. “The inequality-adjusted human development index: a constructive proposal”. In: *World development* 25.8 (1997), pp. 1283–1298.

[9] Allen C Kelley. “The human development index:” handle with care””. In: *Population and development review* (1991), pp. 315–324.

[10] Meghnad Desai. “Human development: concepts and measurement”. In: *European Economic Review* 35.2-3 (1991), pp. 350–357.

[11] Sudhir Anand. “Recasting human development measures”. In: (2018).

[12] Jeni Klugman, Francisco Rodrıguez, and Hyung-Jin Choi. “The HDI 2010: new controversies, old critiques”. In: *The Journal of Economic Inequality* 9 (2011), pp. 249–288.

[13] Sabina Alkire and James E Foster. “Designing the inequality-adjusted human development index”. In: (2010).

[14] Special Working Session WCED. “World commission on environment and development”. In: *Our common future* 17.1 (1987), pp. 1–91.

[15] José Antonio Cheibub et al. *How to Include Political Capabilities in the HDI?: An Evaluation of Alternatives*. Citeseer, 2010.

[16] Mehmet Taner, Bülent Sezen, and Hakan Mıhcı. "An alternative human development index considering unemployment". In: *South East European Journal of Economics and Business* 6.1 (2011), pp. 45–60.

[17] Ibrahim Dincer. "Renewable energy and sustainable development: a crucial review". In: *Renewable and sustainable energy reviews* 4.2 (2000), pp. 157–175.

[18] *UNDP*. https://hdr.undp.org/data-center/.

[19] *World Bank data*. https://data.worldbank.org/.

[20] *OECD*. https://www.oecd.org/.

[21] Michaela Saisana and Stefano Tarantola. *State-of-the-art report on current methodologies and practices for composite indicator development*. Vol. 214. European Commission, Joint Research Centre, Institute for the Protection and . . ., 2002.

[22] *World Bank data*. https://data.worldbank.org/indicator/EG.FEC.RNEW.ZS?locations=BR.

[23] *Climate Score card*. https://www.climatescorecard.org/2021/01/brazil-sources-45-of-its-energy-from-renewables/.

[24] Vilfredo Pareto. *Cours d'économie politique: professé à l'Université de Lausanne*. Vol. 1. F. Rouge, 1896.

[25] Robert Gibrat. "Les inégalits économiques". In: *Sirey* (1931).

[26] Benoit B Mandelbrot. *Fractals and scaling in finance: Discontinuity, concentration, risk. Selecta volume E*. Springer Science & Business Media, 2013.

[27] Michael Mitzenmacher. "A brief history of generative models for power law and lognormal distributions". In: *Internet mathematics* 1.2 (2004), pp. 226–251.

[28] John Aitchison and James Alan C Brown. "On criteria for descriptions of income distribution". In: *Metroeconomica* 6.3 (1954), pp. 88–107.

[29] Thomas Piketty et al. "Capital in the Twenty-First Century: a multidimensional approach to the history of capital and social classes". In: *The British journal of sociology* 65.4 (2014), pp. 736–747.

[30] Charles I Jones. "Pareto and Piketty: The macroeconomics of top income and wealth inequality". In: *Journal of Economic Perspectives* 29.1 (2015), pp. 29–46.

[31] Moshe Levy. "Are rich people smarter?" In: *Journal of Economic theory* 110.1 (2003), pp. 42–64.

[32] Moshe Levy and Haim Levy. "Investment talent and the Pareto wealth distribution: Theoretical and experimental analysis". In: *Review of Economics and Statistics* 85.3 (2003), pp. 709–725.

[33] James Stewart. "The distribution of talent". In: *Marilyn Zurmuehlen Working Papers in Art Education* 2.1 (1983).

[34] David Wechsler. "The measurement and appraisal of adult intelligence". In: (1958).

[35] Alan S Kaufman. *IQ testing 101*. Springer Publishing Company, 2009.

[36] Alessandro Pluchino, Alessio Emanuele Biondo, and Andrea Rapisarda. "Talent versus luck: The role of randomness in success and failure". In: *Advances in Complex systems* 21.03n04 (2018), p. 1850014.

[37] Wikipedia contributors. *Pareto distribution — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Pareto_distribution&oldid=1150331751. [Online; accessed 7-May-2023]. 2023.

[38] *US household income*. https://www.census.gov/library/visualizations/2015/demo/distribution-of-household-income--2014.html.

[39] *Scipy Pareto*. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pareto.html.