Economic Data (EFIM 10016) - Final Project

Part 1: Data and Visualization Basics

Note: The code used to produce all tables and graphs in Part 1 can be found in the attached R file called "Part 1 Code". Slight modifications to table design and structure were performed in LaTeX.

A) Download your individual dataset

The downloaded data contains information about the income of 1000 individuals from the years 2000-2010, across gender and occupation. The dataset has 5 variables which can be summarized as follows: Categorical nominal variables include gender, occupation and person ID. Discrete numerical variables include year. Continuous numerical variables include income. Table 1 shows the first 10 rows and columns one, two, 13 and 14 of the downloaded data. I chose to include these columns in order to provide an example of all the column types in the data. The columns "person_id" and "gender" are the only columns in this data that already follow the tidy data principles.

Table 1: Subset of raw dataset

person_id	income_year2000	gender	occupation_busdriver
1	23.93	Female	0
2	24.32	Female	0
3	32.20	Female	1
4	20.17	Female	0
5	26.16	Female	0
6	35.14	Female	0
7	18.59	Female	0
8	20.92	Female	0
9	18.25	Female	0
10	25.82	Female	0

B) Modify the dataset such that it satisfies the tidy data principles.

The data disobeys the three principles of tidy data, outlined by Hadley Wickham [1], for several reasons. There are 16 columns, whilst there are only five variables and the observations for years and occupation are stored within column headers. These issued were amended by storing the observations for years in a new column called "year" and storing the observations for occupation

[1] Wickham, Hadley . "Tidy Data." Journal of Statistical Software [Online], 59.10 (2014): 1 - 23.

in a new column called "occupation". Furthermore, all the income values were gathered in a single column denoted "income". A subset of the first 10 observations of the tidy data can be seen in Table 2. In its tidy format, the data has 11,000 observations of 5 variables.

Table 2: Subset of tidy dataset

person_id	gender	year	income	occupation
1	Female	2000	23.93	Cashier
1	Female	2001	25.16	Cashier
1	Female	2002	26.22	Cashier
1	Female	2003	28.71	Cashier
1	Female	2004	31.10	Cashier
1	Female	2005	31.88	Cashier
1	Female	2006	33.04	Cashier
1	Female	2007	34.79	Cashier
1	Female	2008	36.65	Cashier
1	Female	2009	38.55	Cashier

C) Create a table that shows the overall average annual growth in income, as well as the average annual growth in income by gender and occupation.

I obtained the growth rates by calculating first the overall growth rates by year, then by year and gender and finally by year gender and occupation. Additionally, I calculated the total average growth rates for each category. The results are stored in Table 3.

Table 3: Average annual growth rates

				Bus Driver		Cashier		Nurse	
Year	All	Female	Male	Female	Male	Female	Male	Female	Male
2000	NA	NA	NA	NA	NA	NA	NA	NA	NA
2001	5.02	5.53	4.62	4.40	3.59	5.95	4.68	5.45	4.98
2002	5.14	5.70	4.71	3.77	3.58	5.98	5.22	6.08	4.66
2003	4.61	4.95	4.35	4.29	3.22	4.94	4.66	5.21	4.48
2004	4.48	4.93	4.13	3.86	2.86	5.24	4.19	4.94	4.58
2005	4.33	4.67	4.07	2.98	2.43	5.12	4.35	4.71	4.43
2006	4.36	4.76	4.03	4.13	2.62	5.00	4.30	4.69	4.30
2007	4.14	4.31	4.01	4.32	3.48	4.15	3.78	4.50	4.44
2008	4.14	4.65	3.72	2.65	2.63	5.22	3.88	4.65	3.97
2009	4.04	4.20	3.91	3.07	2.94	4.46	3.88	4.27	4.28
2010	3.91	4.08	3.78	3.39	3.12	4.22	3.56	4.14	4.23
Total Average	4.42	4.78	4.13	3.69	3.05	5.03	4.25	4.86	4.44

D) Describe the table

Table 3 presents the average annual growth rates in income by gender and occupation. The first observation to make is that the first row contains "NA" values. This is due to the fact that growth rates are calculated with reference to the previous year. Since 2000 is the first year, these values are not calculable. Overall, it can be said that income is growing annually for all categories. Comparing the growth rates for males and females, it is clear that females experience a higher growth rate than males by 0.65 percentage points. To further investigate this, these values are broken down by occupation. Male and female nurses experience the lowest discrepancy in growth, whilst cashiers experience the greatest gap in growth.

E) Create a chart that shows the development in income over years across gender and occupation, using an index with the year 2000 as the base year.

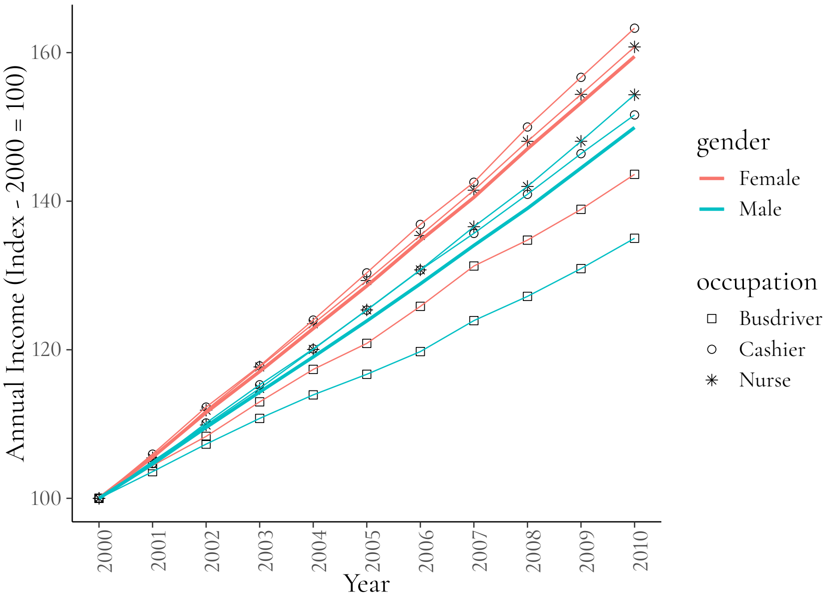


Figure 1: Income by occupation and gender 2000 to 2010, indexed using 2000 as the base year.

F) Describe the graph created in E.

Figure 1 uses an index with base year 2000 to show the development of income across gender and occupation. Red lines indicate females, while blue lines indicate males. Shapes correspond to the occupations. The use of an index implies that values in subsequent years indicate the percentage change from the base year. For example, looking at the graph, it seems that in 2003, the income for male bus drivers has risen by slightly more than 10%, compared to 2000. This can be confirmed by referring back to table 3 and adding the growth rates from years 2001-2003. This graph mainly reaffirms the observations made from Table 3. Overall, we observe that females experience higher growth rates than males, male bus drivers experience the lowest overall growth rates and female nurses experience the highest overall growth rates. To further explore this data, the overall income between men and women across occupation could be observed using a box plot, for instance.

Part 2: Programming Exercise

Note: The code for this program can be found in the attached R file named "Part 2 Code".

Program for seasonal adjustment

Program: Seasonally adjust a time series based on a simplified version of the X-11 algorithm.

Purpose: This function returns a seasonally adjusted series based on input determined by the user. It is adaptable to various time series lengths and types.

Background: A time series can be either additive or multiplicative. An additive time series can be modelled using the equation: $Y_t = S_t + T_t + E_t$, whereas a multiplicative time series can be modelled using the equation $Y_t = S_t * T_t * E_t$. The elements of these equation are the seasonal component S_t, the trend component T_t and the random component E_t. Some time series will also contain a cyclical component C_t. Seasonal adjustment removes the S_t component. This adjustment allows for observation of long-term trends and short-run fluctuations, absent of the effects of seasonality[2]. This Algorithm is based on the X11 algorithm, which was developed by the U.S. Bureau of Census in 1965.

Warning: A time series should only be seasonally adjusted, if there is clear evidence of seasonality[3].

[2] Lee, K. (2018). "7. Seasonal Adjustment". In Quarterly National Accounts Manual (2017 Edition). USA: IMF

[3] Lee, K. (2018). "7. Seasonal Adjustment". In Quarterly National Accounts Manual (2017 Edition). USA: IMF.

Pseudocode

1. Initiate a function with the following five arguments:
    - "series" for the data which should be seasonally adjusted
    - "iter" for the number of times the series should be seasonally adjusted. More iterations will provide a more accurate estimate, at the cost of more data loss.
    - "freq" for the frequency of seasonality, i.e for a monthly time series, freq = 12.
    - "trend_ord" for the order of moving average that should be applied to extract the trend level.
    - "period_ord" for the order of moving average that should be applied to extract the seasonal component.
2. If "series" is stored as a character string, convert it to numeric.
3. Duplicate "series". Store it in a new variable called "series_dup".
4. Start a loop "i" from 1 to the value determined by the input for "iter".
5. Calculate the moving average of "series". The order of the moving average is determined by the input for "trend_ord". Store the output in a new variable "trend".
6. Subtract the content of the "trend" variable from the content of the "series_dup" variable. Store the output in a new variable "season_and_error".
7. Start a loop "j" from 1 to the user determined input for "freq"
8. Store a series of "NA" values of the same length as "series" in a variable called "empty_vec".
9. Extract the seasonal components by extracting each value from the series "season_and_error", starting with "j" and incrementing by "freq" until all values are extracted. Store the output in a new variable called "seasonal".
10. Calculate the moving average of "seasonal". The order of the moving average is determined by the input for "seas_ord". Store the output in "empty_vec" along the sequence from "j" to the full length of "series", by the increment "freq".

11. End loop "j"

12. Subtract "empty_vec" from "series_dup" and store the output in a new variable called "seasonal_series".

13. Store the "seasonal_series" in the variable "series".

14. Return the result of "series"

15. End loop "i"

Part 3 Presenting economic data

A: The labour market

Note: The code used to produce all tables and graphs in section 3A can be found in the attached R files called "Part 3A Task B", "Part 3A Task D" or "Part 3A Task F".

A) Determine your individual set of four countries

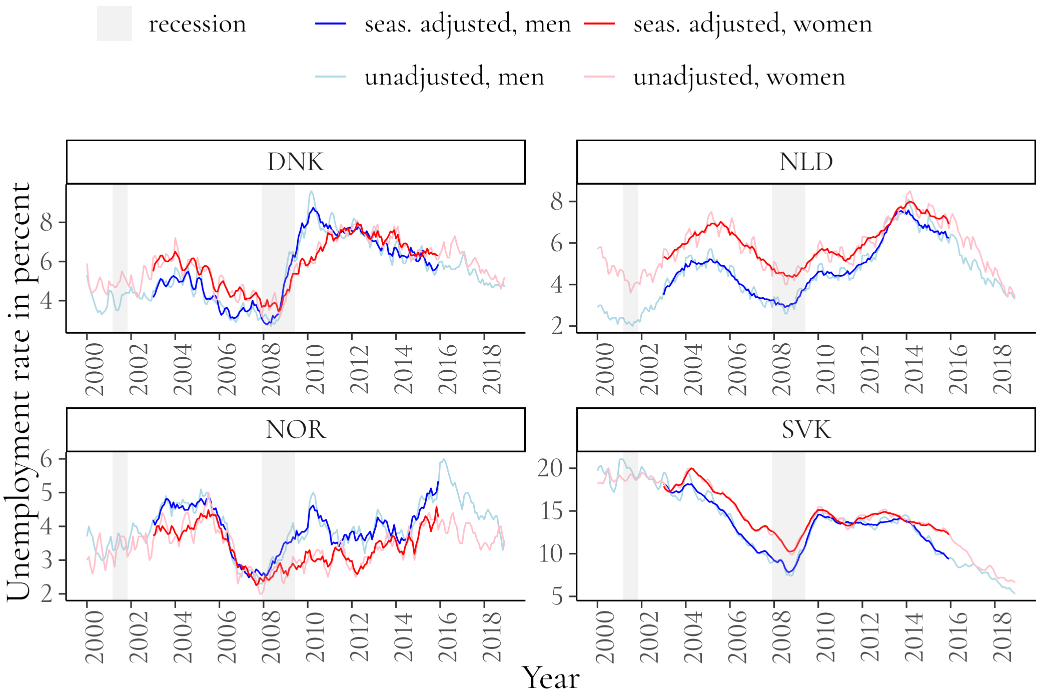
My countries: Denmark, Netherlands, Norway and Slovakia.

B) Create tables or graphs showing raw and seasonally adjusted monthly unemployment rates for the four countries (from A) over the period 2000 to 2018, for men and women.

C) Describe the graphs created in B and describe the development in unemployment rates in the four countries, for men and women.

The unemployment rate is defined as the number of unemployed people as a percentage of the number of people in the labour force[4]. Figure 2 shows the development of unemployment rates (both raw and seasonally adjusted) from January 2000 - December 2018, for Denmark, Netherlands, Norway and Slovakia, for men and women. I chose to vary the y-axis scales in order to make the fluctuations in unemployment within countries more visible and hence, highlight the variation between men and women. It is important to recognize however, that this makes visual comparisons

[4] Sievertsen, H. (2019). "Lecture Note 12". Economic Data 10016



Sources: OECD Labour Force Statistics, FRED

Figure 2: Monthly unemployment rates for Denmark (DNK), Netherlands (NLD), Norway (NOR) and Slovakia (SVK) for men and women.

between countries more difficult. The series were seasonally adjusted using my own function for seasonal adjustment using the default settings (13 and 3 period centered moving averages to identify the trend and seasonal components, respectively). I chose to present the data using line graphs for each country; blue lines for males and red for females. Grey rectangles indicate NBER recessions. For all calculations, I used the data from the seasonally adjusted series. This meant observations for the first and last two years were lost. From initial observations, based on the y-axis, it is clear that Slovakia has the highest unemployment rates and Norway has the lowest. In Slovakia, from 2003-2007, men and women experienced rates of an average of 14.6% and 16.4%, respectively. From 2008-2015, these averaged around 2.3 and 3 percentage points lower, for men and women respectively, which is also evident from the graph, given that unemployment rates seemed to be more stable in the second period and did not reach levels as high as in the early 2000s. Norway experienced the least variability, with an average of 3.9% for men and 3.3% for women over the full time period. From 2008 to 2009, unemployment rates increased for men by 25% and 9% for women. Clearly, men in Norway were hit more severely by the Great Recession, than women, seen on the graph as the increasing gap between the blue and red lines. Denmark had stable unemployment rates of an average of 5.2% and 4.2% for men and women, 2003-2007. However, Denmark's labour market was vastly affected by the Great Recession, with unemployment rates rising by 100 percent for men and 40 percent for women from 2008 to 2009. This can be observed on the graph as the steep upwards trend of the lines at that time period. However, following 2010, the lines start decreasing again, suggesting possibly high job turnover rates in Denmark, allowing unemployed people to quickly find jobs again. In the Netherlands, unemployment rates were on average 6.1% and 4.9% for men and women respectively, over the full time period. As visible on the graph, there were several increases and decreases of these values but the variations were generally experienced fairly equally by both groups. Notably, in 2014 the unemployment rates were practically the same for both men and women. Overall, it can be said that men and women experience similar developments of unemployment rates over business cycles, but men were hit harder by the Great Recession, hinting that male-dominated industries were more severely affected.

D) Create tables or graphs comparing register based and survey based unemployment rates over the period January 2000 to December 2018 for the four countries (from A)

E) Describe the tables or graphs created in D and explain the difference between register and survey based unemployment rate and describe the development of this difference.

Survey based unemployment, otherwise known as ILO unemployment refers to data collected through random sub samples of the population, through labour force surveys. People are considered to be unemployed under the following criteria: They are actively searching for work for at least four weeks, and could start within two weeks, or they have a job and are waiting to start in the next two weeks. Registered unemployment, on the other hand, is based on administrative data recording how many people get unemployment benefits[5]. Since the collection methods differ, as well as the definition for unemployment, the unemployment rates differ based on which measure is used. Survey based data tends to report higher figures than administrative data, due to the fact that not all those who are unemployed receive benefits.

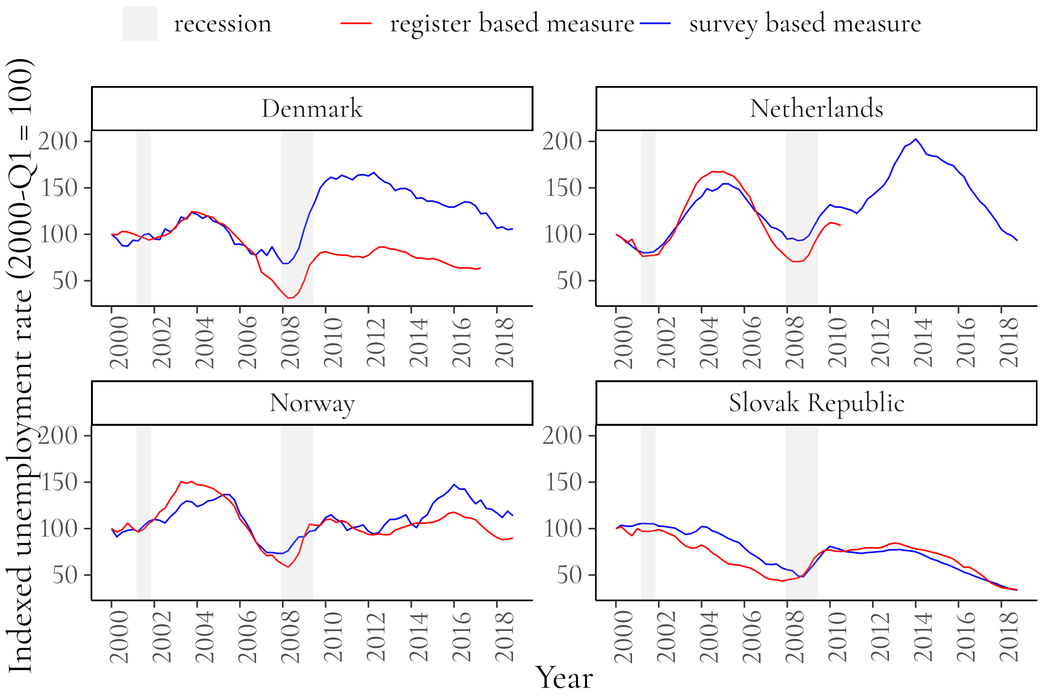
Table 4 shows the ratio of survey based unemployment rates to register based unemployment rates. A ratio below one means, that survey based measures are higher and vice versa[6]. I obtained

[5] Sievertsen, Hans. "Lecture Note 12". Economic Data 10016

[6] Konle-Seidl, R., Lüdeke, B. (2017), What Harmonized and Registered Unemployment Rates Do Not Tell, IAB

Table 4: Ratios of survey based to register based unemployment rates 2000-2016

Year	Denmark	Netherlands	Norway	Slovak Republic
2000	0.80	1.40	1.22	1.04
2002	0.88	1.49	1.15	1.06
2004	0.86	1.26	1.11	1.29
2006	0.86	1.40	1.31	1.29
2008	1.89	1.87	1.60	1.10
2010	1.77	1.66	1.29	1.02
2012	1.70	NA	1.33	0.93
2014	1.64	NA	1.30	0.92
2016	1.81	NA	1.58	0.88
Total Average	1.43	1.54	1.37	1.05



Source: OECD Labour Force Statistics, FRED

Figure 3: Indexed register and survey based quarterly seasonally adjusted unemployment rates for Denmark (DNK), Netherlands (NLD), Norway (NOR) and Slovakia (SVK)

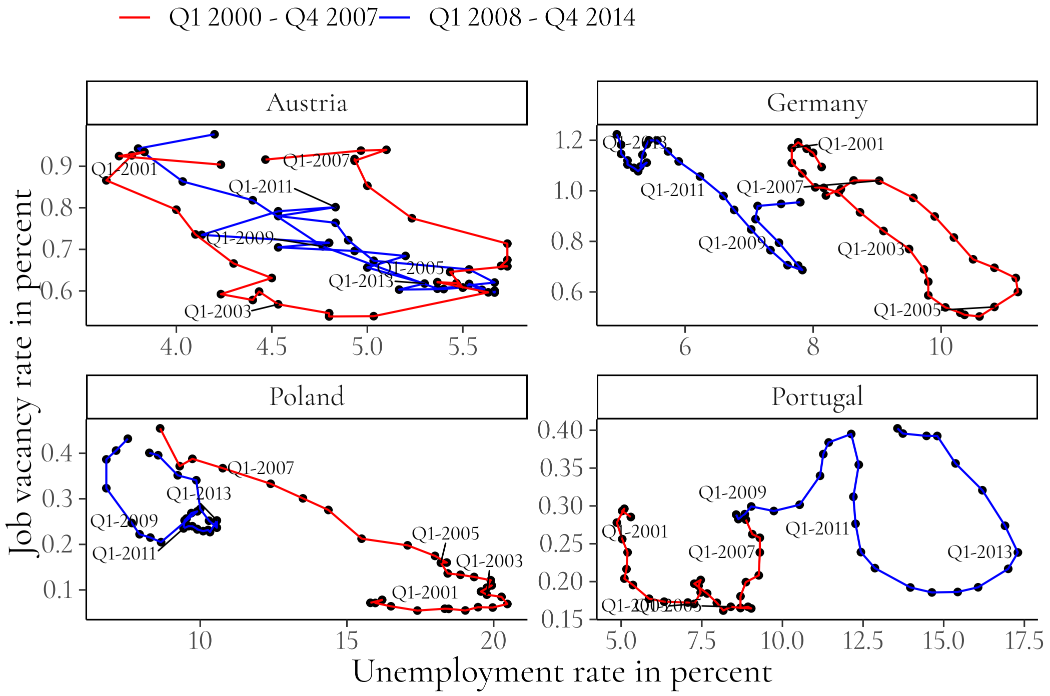
these values by first calculating the average unemployment rates per year and then dividing the survey based measure by the register based measure. I decided to include values in intervals of two years, to preserve readability of the table. However, this does invite the chance of missing significant anomalies in the data. From the table, we confirm that survey based measures are generally higher than register based measures. In Netherlands and Norway the ratios are always positive.

Research Reports, No. 2017/6.

Furthermore, in all countries, the variation in ratios over time is quite consistent, except in Denmark, where the ratio more than doubles from 2008 to 2010.

Figure 3 shows seasonally adjusted register and survey based indexed unemployment rates for the four countries. The index was calculated with base period equal to 2000 Q1. The purpose of using an index is to observe the development of the two types of measures across time when starting at a common value. We observe that the development of these measures are quite similar across time, suggesting that the collection methods remain consistent, in these countries. The only outlier is Denmark, where survey based measure sore above register based measure starting in 2008. This is in line with Table 4, where we saw the ratio of survey to register based values drastically increase. One final observation I made is that in the Netherlands, register based data was only available until 2010. possibly because the Netherlands stopped producing administrative data on unemployment rates.

F) Create Beveridge curves for four countries of your choice. Comment on whether there is evidence of "emergence of a structural mismatch" after the Great Recession in the four countries.



Source: OECD

Figure 4: Beveridge curves for Austria, Germany, Poland and Portugal from 2000-2014

The Beveridge curve is a commonly used indication of the performance of the labour market. It shows the relationship between the supply side (unemployment rates) and the demand side (vacancy rates) of the labour market.[7]. The data I used for figure 4 is from the OECD, however the job vacancy rates are not readily available so they had to be calculated first. In figure 4 we observe the Beveridge curves for Austria, Germany, Poland and Portugal during the period Q1 2000- Q4 2014. Red lines indicate the period Q1 2000 to Q4 2007 (before recession) and blue lines indicate the period Q1 2008 to Q4 2014 (after recession). Important to note is that the axis labels differ between the four graphs, so making direct comparisons between Beveridge curves is not very effective. Austria seems to have experienced an outward shift of the Beveridge curve around 2004 and a slight inward shift in 2008, but the magnitude of changes in unemployment rates and vacancy rates are quite small, compared to the other countries and the behaviour of Austria's Beveridge curve seems quite unusual, with a lot of clustering of data and no clear pattern is visible. Germany experienced an inward shift of the Beveridge curve going from Q1 in 2007 to Q1 in 2009. Slight decreases in the vacancy rates were accompanied by large decreases in unemployment, and hence improved matching efficiency as the curve shifted inwards. However, the shift in the German Beveridge curve is unlikely a result of the crisis, and rather linked to new labour market reforms, which were introduced in the early 2000s[8]. Portugal, on the other hand, experienced a significant outward shift of the Beveridge curve in 2007 during the crisis. Unemployment rates increased drastically, while vacancy rates, saw only very slight changes. Furthermore, Portugal's labor market has not recovered since the Great Recession and the Beveridge curve is continuously shifting outwards, implying a lasting long-term effect. Finally, Poland's Beveridge curve also shifted inwards during the crisis. Based on the Beveridge curves from these four countries, it is difficult to make the conclusion that there is an "emergence of a structural mismatch". Germany and Poland both experienced improving matching efficiencies after the crisis, meaning more people are finding jobs, than before the crisis. Portugal is the only country for which this statement is undoubtedly true. To appropriately evaluate this statement, all countries within the Euro area would need to be examined and perhaps, these four countries do not provide a good sample of the general trend within the euro area.

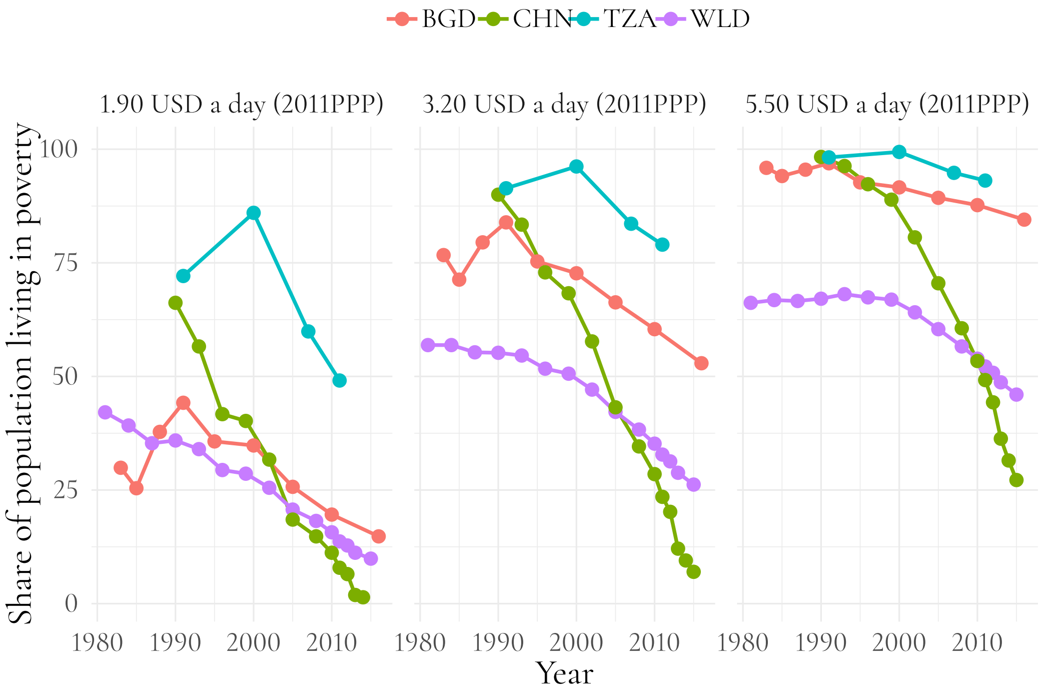
[7] Bova, Jalles, Kolerus, (2017). "Shifting the beveridge curve: what affects labor market matching?".Int. Labour Rev (2017).

[8] Bart Hobijn & Aysegul Sahin (2013). "Beveridge Curve Shifts across Countries since the Great Recession," IMF Economic Review, Palgrave Macmillan;International Monetary Fund, vol. 61(4), pages 566-600

B: Poverty and Inequality

Note: The code used to produce all tables and graphs in section 3B can be found in the attached R files called "Part 3B Task A" or "Part 3B Task D".

A) Create a chart or table showing the development in the share of the population living in poverty over the period 1981 to 2016, for: The World, China, Tanzania and Bangladesh.



Source: World Bank

Figure 6: Share of population living in poverty in The World, China, Tanzania and Bangladesh at three different poverty threshholds. The y-axis is in percent.

B) Based on the graph or table created in A, comment on the following statement:

"The celebrated trend in declining global shares of people in extreme poverty is misleading as this trend is sensitive to using the 1.90USD poverty threshold. Moreover the trend is only driven by a few countries."

Extreme poverty is currently defined by the World Bank as living on less than 1.90 international dollars per day (in 2011 PPP prices)[9]. International dollars (PPP) are used because this makes prices comparable across countries and adjusts for inflation. There are however, other thresholds used to

[9] M. Roser, E. Ortiz-Espina, "Global extreme poverty" (Our World in Data, 2017)

measure poverty such as the \$3.20 and \$5.50 poverty lines. Figure 6 shows the development of the share of the population living in extreme poverty in The World, China, Tanzania and Bangladesh across these different poverty thresholds. Contrary to the statement, extreme poverty is clearly declining in the world (the purple line) under all three poverty thresholds. It is evident, that the decline in world poverty is strongly affected by the vast decline in poverty in China (the green line). Poverty rates in China when from 66.2% in 1990 to 0.7% in 2015, using the \$1.90 a day threshold and using the \$3.20 threshold, poverty rates declined from 90% to 7% in the period 1990-2016. China makes up about 1/5th of the world population and thus the decrease in world poverty is largely driven by China's decrease in poverty, which comes as a result of improvements in the standards of living[10]. The statement that the trend is driven by only a few countries is only partially true, based on these four regions. At the \$1.90 a day threshold it would seem poverty is significantly declining in Tanzania (the blue line) after 2000, however at the \$3.20 and especially the \$5.50 USD a day threshold, it is evident that poverty rates in Tanzania are still very extreme and decreased by 5.1 percentage points from 98.2% to 93.1%, from 1991 to 2011. In Bangladesh both the \$1.90 and \$3.20 USD thresholds indicate a large drop in poverty rates, however this is stunted when using the \$5.50 USD day threshold, where poverty rates only decrease by slightly more than 10 percentage points from 1990 to 2015. Overall, it can be said, that world poverty rates are declining, largely due to China but some countries are still experiencing very high poverty rates, when using higher poverty thresholds.

C) Comment on Figure 1. What is the chart showing? What is the key message? Is the chart type appropriate?

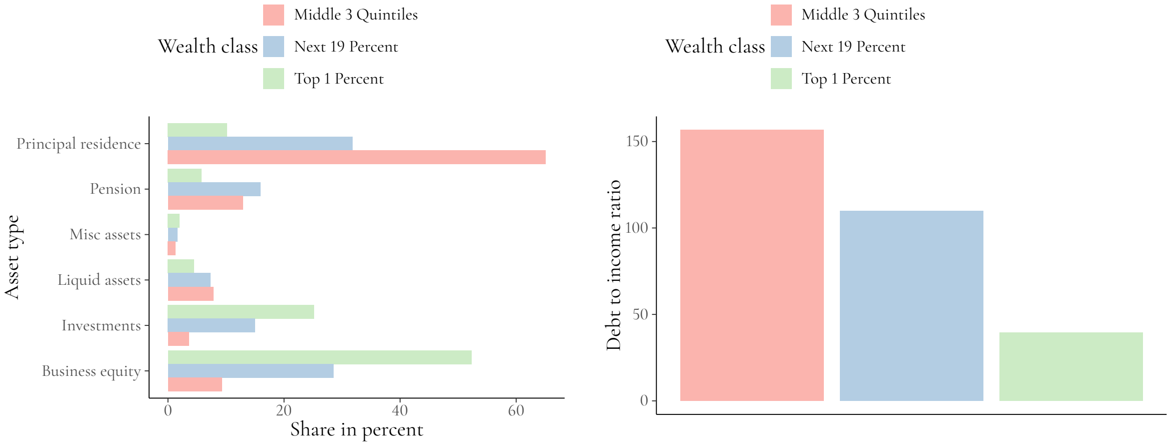
Figure 1 is trying to convey two messages, with the same graph. First of all, it utilizes the red line, corresponding to the right y-axis, to show that debt/income ratios are vastly different between the Top 1%, the next 19% and the middle three quintiles. While, this is true, it is difficult to initially observe this from the graph. My initial intuition, based on the graph was that the red line was connecting the green or yellow areas of the stacked bar chart. Furthermore, the other message that this graph is trying to convey is which assets each group invests their wealth into. The Top 1%

[10] M. Roser, E. Ortiz-Espina, "Global extreme poverty" (Our World in Data, 2017)

invest most of their wealth in business equity and investments, with only 10% going towards their principal residence. The next 19% invest their wealth rather evenly amongst all asset types. The middle three quintiles invest more than 60% in their principal residence. I actually quite like this representation as overall message is quite clear, however it is difficult to identify more exact values when using stacked bar charts and therefore there are better ways to show this.

D) Appendix 1 shows a table with the data that was used to create Figure 1. Create your own visualization of the data with the same key message. Carefully select a visualization method and explain your choice.

The figure in C was ineffective due to being overfilled. Therefore, I decided to separate the two main messages of the graph and attempted to convey them in two individual graphs.



(a) Assets holdings by wealth class

(b) Debt to income ratio by wealth class

Figure 6: Distribution of wealth among different wealth classes (Wealth class is determined by net worth). The source of this data is from Wolff [11]

Figure 6a utilizes a bar chart to identify asset holdings by the three wealth classes. Figure 6a allows straightforward comparisons between the holdings of different types of assets, by comparing the length of the bars. For example, from this graph it immediately stands out that the middle 3 quintiles have 60% of their wealth invested in their principal residence and that the top 1 percent invest more than half their wealth in business equity. These observations were possible but slightly more difficult to make, in the figure in part C. Furthermore, Figure 6b utilizes a bar chart to compare the debt to income ratios between the wealth classes. It is fairly obvious that the top 1% have a very low debt to income ratio, the next 19% hold a debt to income ratio of about 100 and the middle 3 quintiles have a debt to income ratio of more than 150%. I believe this representation of the data is more capable of conveying the key messages, compared to the figure in part C.

References

1. Bart Hobijn & Aysegul Sahin (2013). "Beveridge Curve Shifts across Countries since the Great Recession," IMF Economic Review, Palgrave Macmillan;International Monetary Fund, vol. 61(4).

2. Bova, Jalles, Kolerus, (2017). "Shifting the beveridge curve: what affects labor market matching?".Int. Labour Rev (2017).

3. Lee, K. (2018). "7. Seasonal Adjustment". Quarterly National Accounts Manual (2017 Edition). USA: IMF.

4. Roser, M., Ortiz-Espina, E. (2017).“Global extreme poverty” (Our World in Data, 2017); https://ourworldindata.org/extreme\-poverty/

5. Konle-Seidl, R. Lüdeke, B. (2017). "What Harmonized and Registered Unemployment Rates Do Not Tell", IAB Research Reports http://doku.iab.de/forschungsbericht/2017/fb0617.pdf

6. Sievertsen, H. (2019). "Lecture Note 12". Economic Data 10016

7. Sievertsen, H. (2019). "Lecture Note 17". Economic Data 10016

8. Wickham, H. (2014). "Tidy Data." Journal of Statistical Software, 59.10.

9. Wolff, E. N. (2007). "Recent trends in household wealth in the United States: Rising debt and the middle-class squeeze." Working Paper No. 502. Annandale-on-Hudson, NY: The Levy Economics Institute of Bard College.