Logbook for project

Sevi Rodriguez Mora
University of Edinburgh

Rafael Lopes de Melo
University of Edinburgh

Raman Singh Chhina
University of Chicago

Contents

Chapter A

Call notes

A.0.1 20th August 2021 Meeting

We discussed the following things

- Reviewed our previous work.
- Sevi read from a new document (add the link here) which sort of confirms that the CUSEC medians are infact all the same; and it's value is the median income of the Spain (apart from basque country and one another region)
- Should we work with Danish Data?
  - Sevi's former student Sarah could get us the access.
  - There might not be enough inequality in Denmark to write a paper on inequality.
- US Data?
  - Getting access is the hard part.
  - The US census with the entire population for 1940 is vailable online. We can start with it.
- Next meeting scheduled for 24th August 2021.

A.0.2 24th August 2021 Meeting

We discussed the following things

- Went through Fogli and Guerrieri (2019) results. The slides are here.
- Data: We didn't discuss.
- Research Question:
  - What is a Neighbourhood? Rather than census tracts shouldn't it be something which is defined as a locus of economic activity? I have opened this issue on Github with relevant details.
  - How linked is the segregation in labour market and segregation in residential locations? Could the labour demand in a particular city explain the segregation patterns? Issue opened on Github.

  - Where is inequality: Have we closed this topic? Don't know.

- Next meeting on 27th August 2021.

Chapter B

Literature Review

B.1 Closest to what we are trying to do

Wheeler and La Jeunesse (2008) is a small paper in the Journal of Regional Science. They assume LogNormal distribution and just do the variance decomposition exercise that we do in E.1.3 with US Census Tracts and Blocks. Tracts have populations between 1200-8000 (most of them are around 4000) and the block is roughly one-third of the tract. There is not much (overstatement) analysis in the paper but they report that 85% of the variance comes from within neighbourhoods in a Metropolitan Area. And only 15% is between neighborhoods. I find this result very surprising though, 85% is a very huge number.

Andreoli, Peluso, et al. (2018) is a paper by two economists at the Luxemburg Institute of Socio-Economic Research. It is not yet published anywhere but one of the author's website shows that it is submitted and under review. They use the data from Census Bureau described in D.2 and also look at neighbourhood inequality. However, they don't use the variance decompistion approach. They define a new measure, called Neighbourhood Inequality which is calculated, for a population of n individuals with an income vector $\mathbf{y} = (y_1, y_2, ...y_n)$ as

$$NI(\mathbf{y}, d) = \frac{1}{2}\sum_{i=1}^{n}\frac{1}{n}\Delta_i(\mathbf{y}, d)$$

where

$$\Delta_i(\mathbf{y}, d) = \frac{1}{\mu_{id}}\sum_{j\in d_i}\frac{|y_i - y_j|}{n_{id}}$$

Here d is the distance that the researcher chooses to define a neighbourhood. Δ_i then measures difference in incomes of the neighbours of individual i scaled by the mean income of i's neighbours. They then take the average of Δ_i's for the individuals in the neighbourhood to constuct the NI index.

They also use the US Census data from 1985 to 2010 and find high levels of inequality within neighborhood. The analysis shows that inequality within neighbourhoods is infact representative of the inequality at the city level and has been rising overtime. The paper also looks at the impact of the rising inequality on inter-generational mobility but that isn't a very strong analysis.

The paper doesn't uses lifetime incomes. Nor is there an effort to concile the reported indices with the US National income distribution. They don't control for race, education, occupation etc. There is no attempt at all to understand what is driving the findings.

Minneapolis Fed's Alessandra Fogli and Chicago Booth's Veronica Guerrieri (Fogli and Guerrieri, 2019) also have a paper on this topic. They document the rising neighbourhood segregation and then develop a general equilibrium model where parents choose the neighborhoods to raise their children. They find that segregation and inequality amplify each other because of local spillovers that affect the education returns.

The work of Yanis Ioannides (Hardman and Ioannides, 2004) is also quite relevant. I haven't looked at it properly yet.

B.1.1 Thoughts

These papers cover almost everything in our initial plans. There are gaps which we should discuss. The results are also contrary to each other, Wheeler and La Jeunesse (2008) and Andreoli, Peluso, et al. (2018) report that neighbourhoods are not very seggregated and Fogli and Guerrieri (2019) argue that they are segregated. They both use the same data too.

Sevi, Rafael: What do you think?

B.2 Neighbourhood Segregation

Apart from income segregation, race and ethnicities have historically been a strong grounds for different neighbourhood formations. I review some important economics papers in this literature

B.2.1 Racial

Card, Mas, and Rothstein (2008) is a 2008 QJE paper that tested the predictions of influential Schelling (1971) theory. The theory argued that if the black population of a neighborhood exceeded a certain tipping point then all the whites from that neighbourhood would leave causing extreme segregation. They use regression discontinuity methods and Census tract data from 1970 through 2000 to test for discontinuities in the dynamics of neighborhood racial composition.

Boustan (2010) researches whether the post-war sub-urbanisation "white flight"? i.e. did the migration of blacks from the South lead the whites, who lived in the cities, to move to the suburbs. She uses an instrument variable approach.

B.3 Urban Consumption inequality

Davis et al. (2019) study the segregation of urban consumption in NYC using Yelp reviews. They focus on ethnic and racial disparities of restaurant consumers and estimate a discrete choice model. Their primary

data sources are a Yelp database of 18,000 reviews and US Census Bureau Data for consumption decisions and demographics respectively.

Chapter C

Research design

C.1 Where is inequality? Proposal: Sevi

I was thinking on thigs to do on the inequality front with you two guys... and while dealing with the Covid paper, I had an idea...

actually I had it some time ago... but I thought of an **easy** way to implement it... at least at the beginning.

If you remember, sometime ago I got very excited with doing the follwing exercise:

- Take the total inequality (variance) in a country
- Decompose it in:
  - Inequality between regions
  - Inequality between provinces within regions
  - Inequality between municipalities within province
  - Inequality between Districts within municipalities
  - Inequality between Blocks within district
  - Inequality within Blocks

This decomposition talks about the importance of things like trade (within country) and convergence (between regions) and whatnot... but more interestingly, to me, is that it talks about what inequality meanse, and (I guess) about its persistence.

The point is that now I know how to do this very easily for Spain.

Raman knows already how to do it... and which data to use... because we have been using that data in the covid project.

This is it:

- In the spanish statistical institute (INE) page they have made public the data on the income (average AND DISTRIBUTION) per what they call "sección censal", that has the acronym CUSEC.

  The data is here
- A CUSEC is a group of 1500 individuals

- it has a very uniform size,
- for each CUSEC we know:
    1. Average Income
    2. Percentage of population getting wages
    3. Percentage of population getting pension
    4. Percentage of population getting unemployment Insurance
    5. Percentage of population getting other transfers
    6. Percentage of population getting other sources of income
    7. Percentage of population getting less than 5000, 7500, and 10000 EUR. Plus per gender, nationality and age
    8. Percentage of population getting less than 40%, 50%, 60%, 140%, 160% and 200% of the *median* income (I do not know if in the CUSEC or Spain, but it can be checked. Plus per gender, nationality and age
    9. Population size
    10. Average age
    11. % youngr than 18
    12. % older than 65
    13. average household size
    14. % of 1 member households

and some other things could be obtained, I think, from the CENSUS (such as education, occupation etc..) but I do not think it is necessary (and it is not cost less, we should start an application... the data above is public and ready to use.

OK. There is another data source for Spain that I think that will be useful fro this: the administrative data on Social security. The whole life life employment of spaniards... of... more on that later, in any case I know this dataset less.

OK, so what to do?

1. calculate the variance of income within each CUSEC... which should be straightforward given 8.
2. aggregate CUSECS into districts (another standard definition), calculate average variance between CUSECS within district.
3. aggregate districts into municipalities... id
4. Aggregate municipalities into provinces
5. Aggregate provinces into regions
6. aggregate regions into whole country

I think that doing this right is something that would have lots of value per se... I have never seen this done... and it talks about what sort of luck it is important...which region to be born, or which neighborhood to be born...

Then I would like to think in several things:

1. How to introduce lifetime income into this: for this we could look at how is the income process of people from the administrative data... and within a moll style model get how would the lifetime income distribution should be.

2. Get a spatial model of trade: regions/areas with less income, areas with more (because agglomeration?)... inequality within

As an extra topping I would like to think about inheritance and intergenerational mobility... it is obviously related... I have not thought of how to do it yet.

One thing... with Crtistina and Juraj I am doing a project with Italin data with the dream of doing this somehow at some point, but I think it is very much in the future... but I just realized that I can do this with public data in Spain... right now...

Would you like to do this together?

It is somehow more concrete a thing that what we were talking about... which frankly, I am a bit lost... and I think that hte first things of this can be done super fast!!!

C.1.1 Variance Decomposition

Variance decomposition of two places... between CUSECs that are within a certain district, and within those CUSECs

$$V_d = \sum_{c \in d} s_{cd} \frac{(\bar{y}_c - \bar{y}_d)^2}{N_{cd}} + \sum_{c \in d} s_{cd} V_{cd}$$

where:

- V_{cd} is the variance within CUSEC c in district d
- N_{cd} is the population of CUSEC c in district d
- s_{cd} is the share of cusec c within district d: $s_{cd} = \frac{N_{cd}}{N_d}$; N_d being the size of the district.

between districts within municipality:

$$V_m = \sum_{d \in m} s_{dm} \frac{(\bar{y}_{dm} - \bar{y}_m)^2}{N_{dm}} + \sum_{d \in m} s_d V_d = \sum_{d \in m} s_{dm} \frac{(\bar{y}_{dm} - \bar{y}_m)^2}{N_{dm}} + \sum_{d \in m} s_{dm} \left[\sum_{c \in dm} s_{cdm} \frac{(\bar{y}_{cdm} - \bar{y}_{dm})^2}{N_{cdm}} + \sum_{cdm} s_{cdm} V_{cdm} \right]$$

And we can expand this iterative to province and country...

$$
\begin{aligned}
V_{spain} = & \sum_{p} \frac{N_p}{N_{spain}} \frac{(\bar{y}_p - \bar{y}_{spain})^2}{N_p} \\
& + \sum_{p} \sum_{m \in p} \frac{N_p}{N_{spain}} \frac{N_{mp}}{N_{mp}} \frac{(\bar{y}_{mp} - \bar{y}_p)^2}{N_{mp}} \\
& + \sum_{p} \sum_{m \in p} \sum_{d \in mp} \frac{N_p}{N_{spain}} \frac{N_{mp}}{N_m} \frac{N_{dmp}}{N_{mp}} \frac{(\bar{y}_{dmp} - \bar{y}_{mp})^2}{N_{dmp}} \\
& + \sum_{p} \sum_{m \in p} \sum_{d \in mp} \sum_{c \in dmp} \frac{N_p}{N_{spain}} \frac{N_{mp}}{N_m} \frac{N_{dmp}}{N_{mp}} \frac{N_{cdmp}}{N_{dmp}} \frac{(\bar{y}_{cdmp} - \bar{y}_{dmp})^2}{N_{cdmp}}
\end{aligned}
$$

$$
\begin{aligned}
V_{spain} = & \sum_{p} \frac{N_p}{N_{spain}} \frac{(\bar{y}_p - \bar{y}_{spain})^2}{N_p} \\
& + \sum_{p} \sum_{m \in p} \frac{N_{mp}}{N_{spain}} \frac{(\bar{y}_{mp} - \bar{y}_p)^2}{N_{mp}} \\
& + \sum_{p} \sum_{m \in p} \sum_{d \in mp} \frac{N_{dmp}}{N_{spain}} \frac{(\bar{y}_{dmp} - \bar{y}_{mp})^2}{N_{dmp}} \\
& + \sum_{p} \sum_{m \in p} \sum_{d \in mp} \sum_{c \in dmp} \frac{N_{cdmp}}{N_{spain}} \frac{(\bar{y}_{cdmp} - \bar{y}_{dmp})^2}{N_{cdmp}}
\end{aligned}
$$

C.2 What is a neighbourhood?

What is a neighbourhood? How do we define it's doundaries? Do these definitions affect the way we think about segregation and neighbourhood effects on resident outcomes?

C.2.1 Measurement Issues

We start with Fogli and Guerrieri (2019). They calculate neighbourhood segregation in a given city with n census tracts using the following formula, where X and Y are the total number of rich and poor in the city and X_i and Y_i are the corresponding quantities in Census Tract i. They define rich as the people in the top 20% of the distribution and the remaing population as poor.

$$D = \frac{1}{2}\sum_{i=1}^{n}\left|\frac{X_i}{X} - \frac{Y_i}{Y}\right|$$

Is this formula robust to the definitions of neighborhood? To test this relationship I run a few simulations. I define a city with N^2 neighbourhood blocks and calculate the above index for various distributions of the rich and poor in these neighbourhoods.

Fig C.2.1 has equal number of rich and poor in every neighbourhood. There is no segregation and we get $D = 0$. In Fig C.2.2 rich and poor are randomly distributed across neighbourhoods, some blocks end up getting high proportions of poor and others get high proportions of rich. Due to random distribution D, however has a low value, equal to 0.031.

This measure however, quickly runs into issues. Consider, Fig C.2.3, for instance, here each block has either 100% rich or 100% poor population and accordingly we get a high $D = 0.167$. Is this city really segregated? Rich and poor all still live next to each other. That is, compare it to Fig C.2.4 which has the same D but we can argue by looking that it has considerably more segragation. Another example is of Fig C.2.5 which has exactly the same distributions as Fig C.2.3, I just define the census tract to be double the existing size. This makes the segragation measure to exactly go to zero!

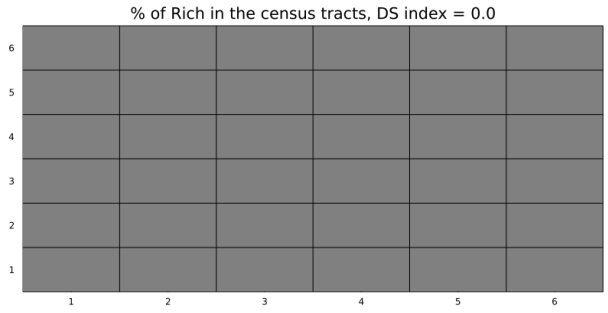

Figure C.2.1: Scenario 1: Equal number of poor and rich in each tract

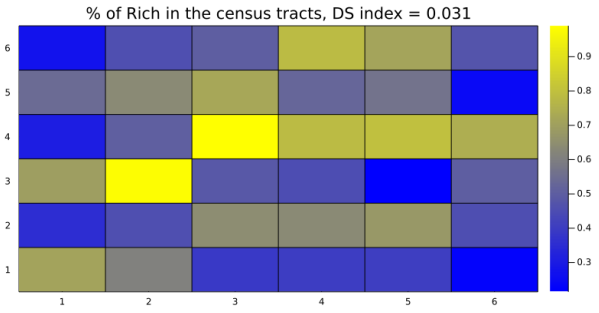

Figure C.2.2: Scenario 2: Random number of poor and rich in each tract

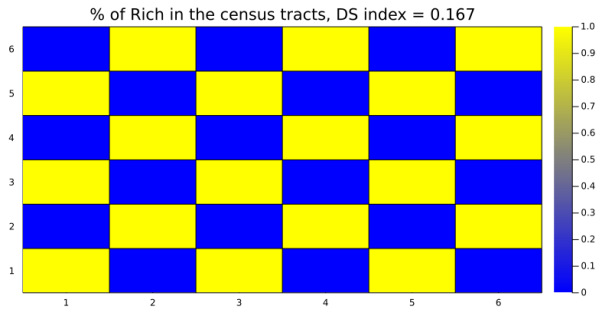

Figure C.2.3: Scenario 3: Rich and Poor perfectly segregated but next to each other

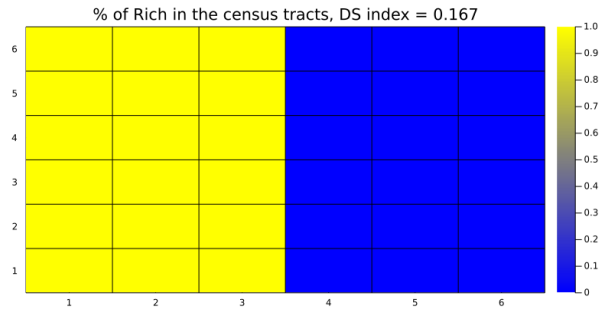

Figure C.2.4: Scenario 3: Rich and Poor perfectly segregated and on the extreme ends

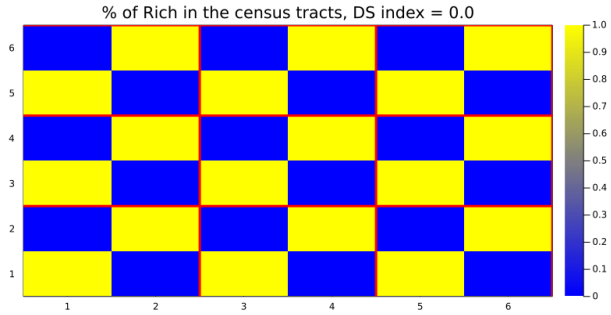

Figure C.2.5: Scenario 5: Rich and Poor perfectly segregated but tracts double the size

Scenario's 4 and 5 highlight the issue well. Segregation measures depend highly on the definition we choose and the census tract approach COULD just be an arbitraty choice among the possible alternatives.

NB - Well, I got carried away working on these, turns out Scenario 3 is called the checkerboard problem and people were researching about it in 1983! (Reardon and O'Sullivan, 2004). Reardon et al. (2018) discuss further problems with segregation measures caused due to sampling issues and above discussed problems. They also provide robust measures which we should keep in mind if we wrok with segregation.

C.2.2 Theory of the Neighbourhood

Chetty et al. (2014) talk about this a bit as well —

"To characterize the variation in children's outcomes across areas, one must first partition the U.S. into a set of geographical areas in which children grow up. One way to conceptualize the choice of a geographical partition is using a hierarchical model in which children's outcomes depend upon conditions in their immediate neighborhood (e.g., peers or resources in their city block), local community (e.g., the quality of schools in their county), and broader metro area (e.g., local labor market conditions). To fully characterize the geography of intergenerational mobility, one would ideally estimate all of the components of such a hierarchical model."

Further they argue that

"We focus on CZ-level variation because mobility statistics in very small neighborhoods are likely to be heavily affected by sorting. Because property prices are typically homogeneous within narrow areas and home values are highly correlated with parent income, comparisons within a small neighborhood affectively condition on a proxy for parent income. As a result, the variation in parent income across individuals in a small area (such as a city block) must be correlated with other latent factors that could affect children's outcomes directly, making it dicult to interpret the resulting mobility estimates. For example, it would be dicult to estimate the degree of intergenerational mobility on Park Avenue in Manhattan because any families with low observed income in such a high-property-value area would have to be latently wealthy to be able to afford to live there."

Chapter D

Data description

D.1 US 1940 Census

The US census data is released after 72 years of the record date. The latest available right now is the 1940 Census Data and the 1950 wave would be released in April 2022. This is the link which tells what questions were asked in the census. The feilds relevant for us are:

- Street, Avenue, Road, etc
- House Number
- Income in 1939
  - Amount of money, wages, or salary received
  - Did this person receive income of $50 or more from sources other than money wages or salary

The data is huge and can be accessed using the Amazon Web Services Registry of Open data. The access process is described here.

D.1.1 Geographic Information

The house number, street is available in the raw images but we can't directly make use of them unless they are 'geo-coded' i.e. we know their exact place on a map. This is not made available by the US Census Authority. The smallest geographic unit which is used in a consistent fashion is the **Enumeration District**. It is exactly like the CUSEC in the Spanish Data and was defined so as the census official could collect all the data in about a week. It ranges from blocks in urban areas to entire counties in the rural areas.

Again the shapefiles (geo-coded locations) are not made available for the Enumeration Districts by the Census Authority. However, there are people who have worked on it

- Steve Morse provides a way to get EDs from street addresses.
- John Logan at Brown University. This is the most promising one. They have a paper in which they perform street level analysis of white/black segregation in 191 cities.

- Mark Fosett at Texas A&M. The shapefiles are available for Charleston, South Carolina, and Buffalo, New York

So, the census along with the Brown University data could in principle get us to the street level. All of it is publicaly available as well.

D.1.2 Income Information

The Census Data records the self reported income for the year 1939. Even with this limitation, we get a pretty microscopic view of the income distribution. However, one major drawback of using the 1940 Census is that we don't know the overall distribution of income very well. Statistics of Income (SOI) Public Use Files for Income Tax data which are used to construct modern estimates in Piketty, Saez, and Zucman (2018) are not available pre 1962. We also can't use IRS Tax statistics as in Piketty and Saez (2003) becuase earlier to 1944 a large population was exempt from income taxes. So, this can be used to get estimates about the very top end.

D.2 Chicago RDC

Chicago Research Data Center is a collaboration between US Census Bureau, Chicago Fed, Northwestern and University of Chicago. They can provide access to census data with geographic information down to the block level. It has a tedious approval process but is nicely outlined here.

This dataset would be the best available to us. It can be accessed directly from Chicago which is a plus. It requires funding to cover the cost though.

Chapter E

Empirical results

E.1 City Facts: Raman

I use the data from the Urban Audit project, joint between INE and EuroStat to get two definitions of cities: Core city and the Functional Urban Area. This data is at the municipality level and for each City and FUA tells the corresponding Municipalities. This definition gives a total of 81 Functional Urban Areas in Spain with 1,218 Municipalities.

E.1.1 Zipf's Law

Working with cities is useful because the satisfy certain regularities and we have some economic theory explaining these results (Gabaix, 1999). The most conspicuous of these being the Zipf's law — stating that city population sizes follow a power law distribution with an exponent $\zeta = 1$, at least approximately. It can be easily explained by a proportional random growth process and it holds quite well for the cities in our data, with $\zeta = 0.98$ and $\zeta = 0.95$ for core cities and FUAs respectively.

In a minimalist model to explain Zipf's law Gabaix (1999) makes use of amenities shocks across cities. Young workers choose which city to migrate to in the beginning of their career and the only variation of wages within a city arises because of the difference between young workers and old workers. There are a

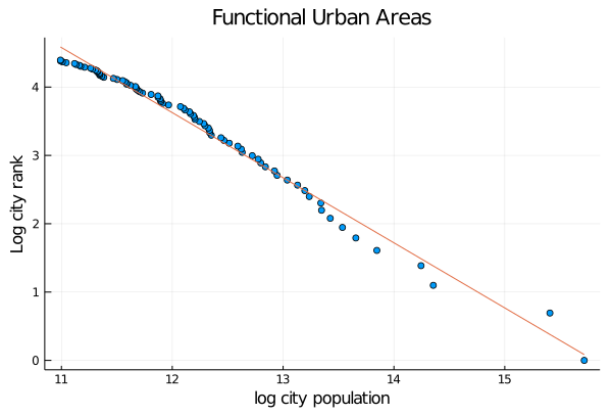

Figure E.1.1: $\zeta = 0.95$

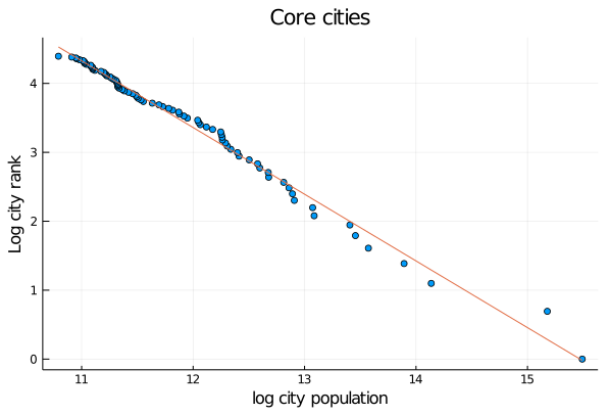

Figure E.1.2: $\zeta = 0.98$

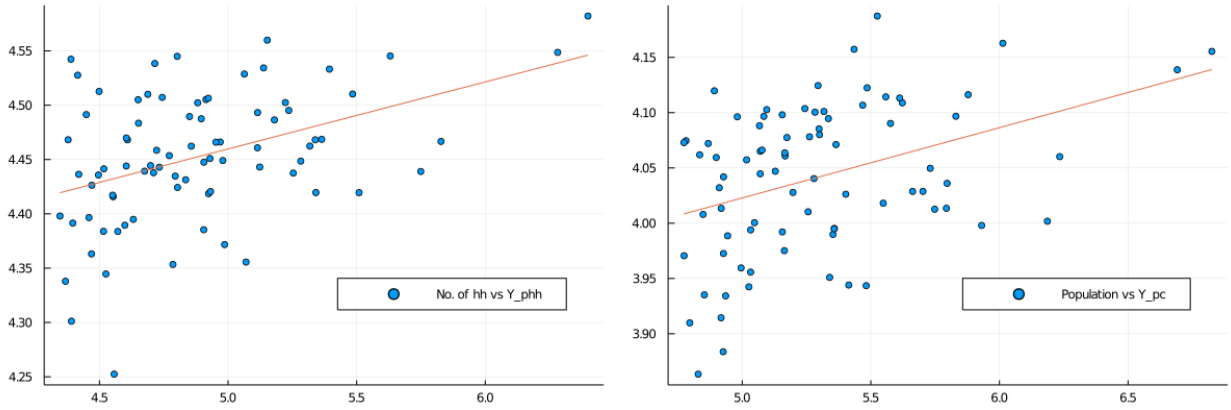

Figure E.1.3: $\beta_{hh} = 0.062$ & $\beta_{pc} = 0.064$

number of other models based on amenities, positive and negative externalties of agglomeration/congestion but I haven't found any which explicitly match facts both on Zipf's law and inequality within cities (two of which should obviously be very connected).

E.1.2 Larger Cities have higher per capita/household income

As it is well established in the literature, larger cities tend to have higher wages. It can be explained by a number of theories such as agglomeration, selection, sorting which either lead to a higher skilled population in larger cities or increases the skill premium.

E.1.3 Variance Decomposition within/between cities

I do the same exercise that we were trying to with Provinces but now rather with Cities. One main fact that should be of interest to match with the predictions of above models would be the variance decomposition between/within cities. As a first pass, I make an extreme assumption that the population living in a particular CUSECs is all homogeneous. This would provide us with an upper bound on the variance between cities.

The incomes are in logs and c denotes city, m municipality, d district and z CUSEC. Here we are looking at only the urban population.

$$V_{spain} = \frac{1}{N_{spain}} \sum_c \sum_{i \in c} (\bar{y}_c - \bar{y}_{spain})^2 + \frac{1}{N_{spain}} \sum_c \sum_{i \in c} (\bar{y}_i - \bar{y}_c)^2 \tag{E.1.1}$$

$$\equiv \frac{1}{N_{spain}} \sum_c N_c (\bar{y}_c - \bar{y}_{spain})^2 + \frac{1}{N_{spain}} \sum_c \sum_{i \in c} (\bar{y}_i - \bar{y}_c)^2 \tag{E.1.2}$$

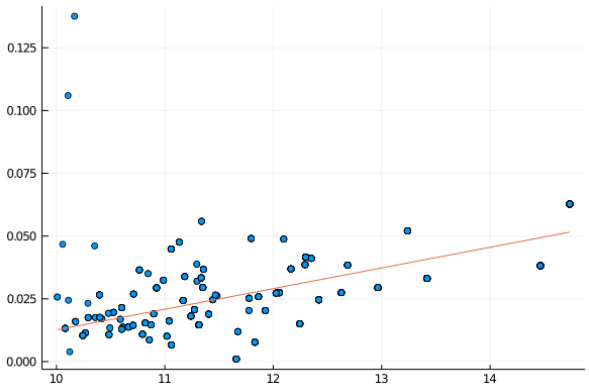

Figure E.1.4: MLD vs City Population, $\beta = 0.008$

$$
\begin{aligned}
V_{spain} = & \frac{1}{N_{spain}} \sum_{c} N_c (\bar{y}_c - \bar{y}_{spain})^2 \\
& + \frac{1}{N_{spain}} \sum_{c} \sum_{m \in c} N_{mc} (\bar{y}_{mc} - \bar{y}_c)^2 \\
& + \frac{1}{N_{spain}} \sum_{c} \sum_{m \in c} \sum_{d \in mc} N_{dmc} (\bar{y}_{dmc} - \bar{y}_{mc})^2 \\
& + \frac{1}{N_{spain}} \sum_{p} \sum_{m \in p} \sum_{d \in mp} \sum_{z \in dmp} N_{zdmp} (\bar{y}_{zdmp} - \bar{y}_{dmp})^2
\end{aligned}
$$

As we are assuming that within a CUSEC everyone has the same income (equal to the mean), the fifth term in the above equation would be zero.

Total Variance	Between Cities	Within Cities
0.097	0.021 (22%)	0.076 (78%)

Variance	Value
Total	0.097 (100%)
Between Cities	0.021 (22%)
Between Municipalities (given a city)	0.015 (15.3%)
Between Districts (given a city, Muncip.)	0.025 (25.5%)
Between CUSECs (given a city, Muncip. & Dist.)	0.036 (37.2%)

The variance however suffers from the scale issue and the variance of the logarithms is also not aggregative (Bourguignon, 1979), causing the estimates to be biased. Hence, I also use another measure of inequality which satisfies all the desirable properties to be used as a decomposable measure of inequality — Mean Log Deviation or the Generalised Entropy Index with parameter 0, GE(0).

$$
MLD_{spain} = \frac{1}{N_{spain}} \sum_{i=1}^{N} \log \left(\frac{\bar{y}}{y_i} \right) \tag{E.1.3}
$$

Decomposing into between/within cities

$$MLD_{spain} = \frac{1}{N_{spain}} \sum_c N_c \log\left(\frac{\bar{y}}{\bar{y}_c}\right) + \frac{1}{N_{spain}} \sum_c N_c MLD_c \tag{E.1.4}$$

$$\equiv \frac{1}{N_{spain}} \sum_c N_c \log\left(\frac{\bar{y}}{\bar{y}_c}\right) + \frac{1}{N_{spain}} \sum_c \sum_{i \in c} \log\left(\frac{\bar{y}_c}{y_{ic}}\right) \tag{E.1.5}$$

Decomposing further

$$\begin{aligned} MLD_{spain} = & \frac{1}{N_{spain}} \sum_c N_c \log\left(\frac{\bar{y}}{\bar{y}_c}\right) + \\ & \frac{1}{N_{spain}} \sum_c \sum_{m \in c} N_{mc} \log\left(\frac{\bar{y}_c}{\bar{y}_{mc}}\right) + \\ & \frac{1}{N_{spain}} \sum_c \sum_{m \in c} \sum_{d \in mc} N_{dmc} \log\left(\frac{\bar{y}_{mc}}{\bar{y}_{dmc}}\right) + \\ & \frac{1}{N_{spain}} \sum_c \sum_{m \in c} \sum_{d \in mc} \sum_{z \in dmc} N_{zdmc} \log\left(\frac{\bar{y}_{dmc}}{\bar{y}_{zdmc}}\right) \end{aligned}$$

Total MLD	Between Cities	Within Cities
0.0498	0.0109 (21.8%)	0.039 (78.2%)

Mean Log Deviation	Value
Total	0.0498 (100%)
Between Cities	0.0109 (21.8%)
Between Municipalities (given a city)	0.0087 (17.5%)
Between Districts (given a city, Muncip.)	0.0125 (25.1%)
Between CUSECs (given a city, Muncip. & Dist.)	0.0177 (35.6%)

Even with this assumption only 22% of the total variance/MLD can be explained by cities. This is an upper limit, if we consider within CUSEC distribution too then this would obviously come down.

E.1.4 CUSEC Composition

Each Functional Urban Area and Core city can divided into its constituent CUSECs. Although CUSECs are quite small units they have significant income variance. It is interesting to look at this variance, in the form of top and bottom inequality.

Rich CUSECs have almost no one earning less than the 40% of the CUSEC median P.C.U income whereas the poor CUSECs have significant populations under this threshold. This suggests that poor CUSECs have a fatter lower tail and similarly looking at E.1.6 we can infer that rich CUSEC's income distribution has a fat upper tail.

This is quite interesting and goes against our initial hypothesis that rich and poor CUSEC's distribution would be almost degenerate. They however have degenerate lower or top halves and have fat tails in the

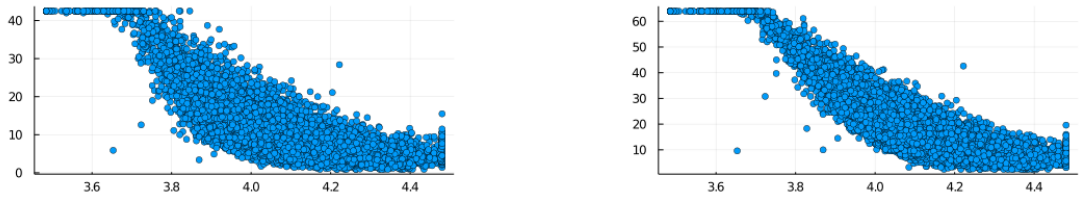

Figure E.1.5: % of CUSEC population earning below 40 and 60% of the CUSEC median vs log mean income

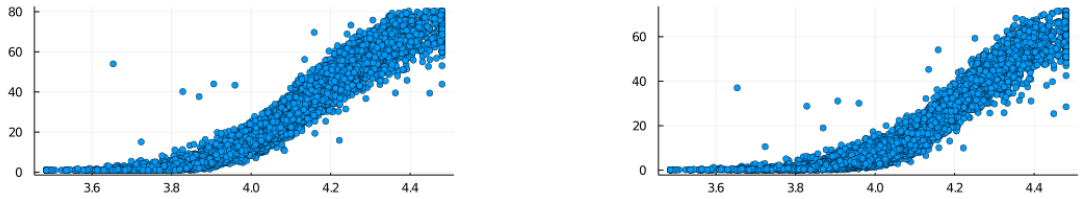

Figure E.1.6: % of CUSEC population earning more than 140 and 160% of the CUSEC median vs log mean income

other end.

The (sort of) variance measures however (E.1.7) then show that rich and poor CUSECs are highly unequal.

E.1.5 Something really cool — Power law for neighbourhoods

Again, municipalities and districts are state defined entities and do not necessarily reflect the socio-economic segregation in reality. However, with CUSECs which have small populations and are constructed to reflected spatial boundaries such as roads, blocks etc, I construct neighbourhoods with particular characteristics by combining CUSECs together which are close to each other both spatially and economically.

To construct these neighbourhoods/localities, I use the following algorithm —

1. Identify the kind of localities to be constructed. For example — poor ones (with say mean income $\leq$ 8,000 Euros), rich ones (mean income $\geq$ 25,000) etc.
2. Keep only the CUSECs which satisfy this condition.

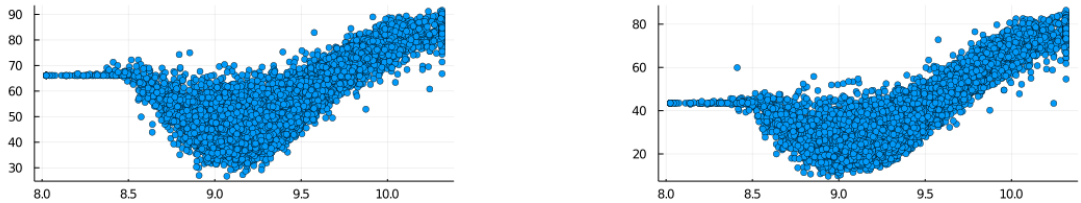

Figure E.1.7: % of CUSEC population outside 0.6-1.4 and 0.4-1.6 band of the CUSEC median vs log mean income

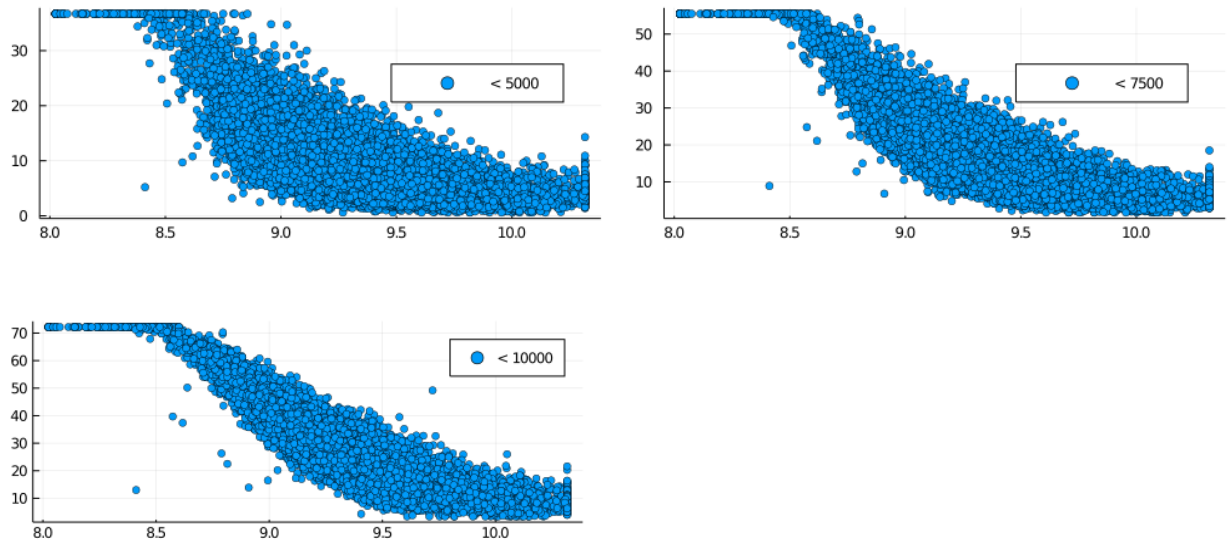

Figure E.1.8: Fraction of people earning less than a threshold vs income

3. Start with the first CUSEC and group it with all its immediate spatial neighbours in the above set.

4. Combine the localities in 3. with their neighbouring localities.

5. Repeat till the time all the resulting localities do not share any boundaries.

Fig. E.1.11 (page 7) shows all the resulting localities with mean p.c. income less than 8000 Euros. Geographically these are spread from the city centre to the subarban areas and have largely varying sizes and populations. However, interestingly (and I haven't seen this done anywhere — I suppose due to the lack of such granular data) the locality populations follow a power law distribution too!

Fig. E.1.9 and Fig. E.1.10 show the plots for poor and rich localities constructed using the above algorithm. The slope for the poor localities is around -1 but varies a bit depending on the cut-off value that we chose to classify a CUSEC as 'poor'. For the rich neighbourhoods the value of the exponent stays around 0.6. These plots are just for the city of Madrid. I am trying to think of a way to verify this for other cities too. I found this very interesting — the process leading to formation of these localities in such

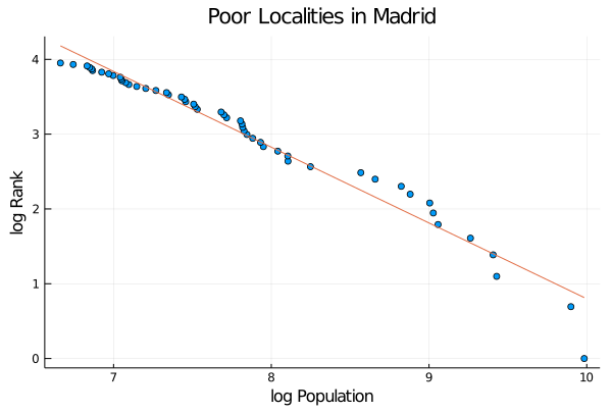

Figure E.1.9: $\zeta = 1$

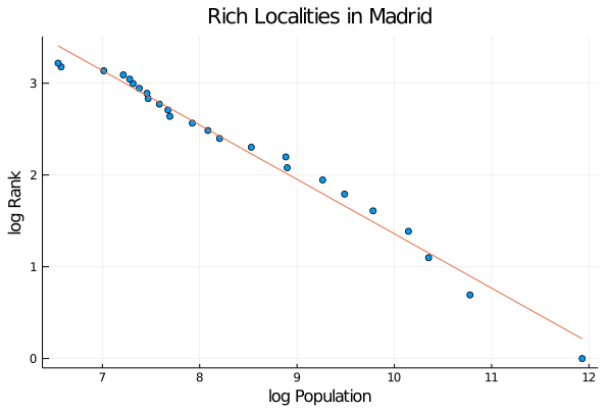

Figure E.1.10: $\zeta = 0.6$

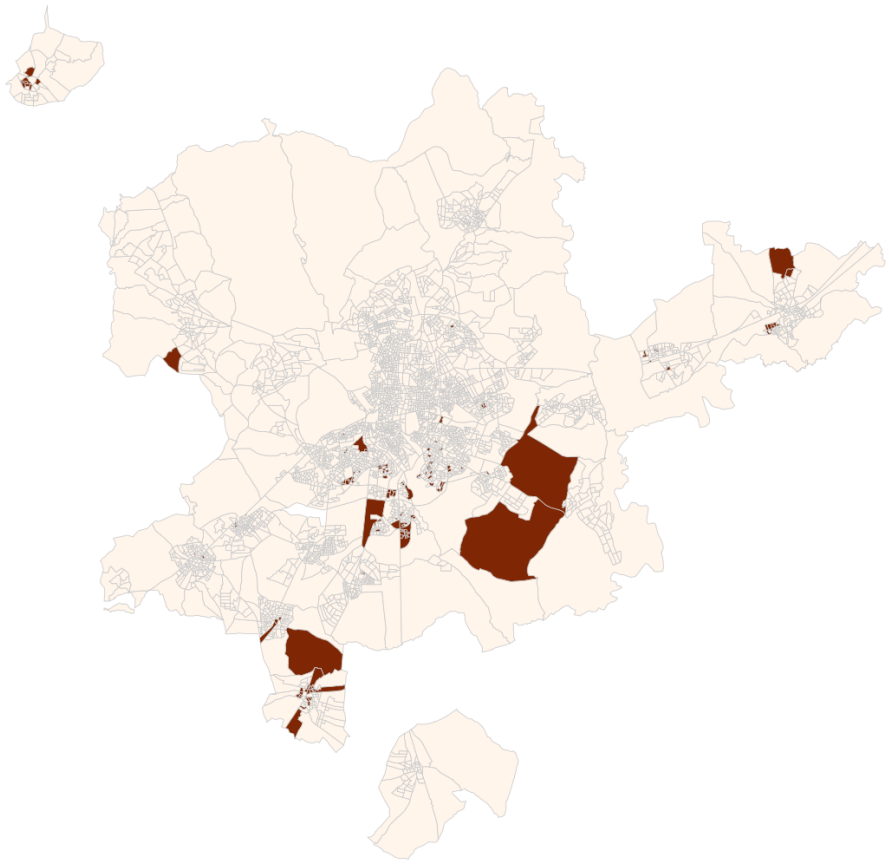

Figure E.1.11: Localities with mean income p.c. less than 8,000 Euros

a fashion gives us a way to think about a lot of things happening within the city from segregation, housing affordability, inter-generational mobility, ghettoizing and obviously income inequality.

E.1.6 Neighbourhood Composition

Chapter F

Theory

F.1 Model

We discussed the implications of baseline Gabaix (1999) for inequality within/between cities while simultaneously explaining the Zipf's law. The wages of all the individuals in a city are same, hence, there is no within city inequality. Moreover, in equilibrium the utility adjusted wages are exactly same across all cities, so there are no systematic differences in the wages between cities either.

F.1.1 Rich and poor individuals

Suppose there are two kinds of agents, high and low skilled. Each city has $N^o_{i,H}$ and $N^o_{i,L}$ number of rich and poor residents who already live in the city at the beginning of the current time period. Each city is characterised by a 3-tuple of amenities and wages for the high and low skilled $(a_{it}, w^H_{it}, w^L_{it})$.

In equilibrium all the individuals would be indifferent between different cities and hence, utility adjusted high and low skilled wages should be same across places.

$$a_{it} w^H_{it} = u^H_t$$
$$a_{it} w^L_{it} = u^L_t$$

If the production technology is CRS and is independent for the low and high skilled sectors, then it is equivalent to the last section and the population of rich and poor converge to a Zipfian distribution.

F.1.2 Heterogeneous agents model

In the above model I was assuming that people differ in skill and was trying to see how they distribute themselves in cities. Another, standard, way of looking at inequality is not via skill but via luck i.e. individuals get idiosyncratic income shocks, which makes some people rich while others poor. I take a measure of such agents but impose a further structure

- There are N cities

- At the beginning of their life, agent decides which city to live in
- Cities receive amenities shocks — same for everyone in the city

I want to see if the resulting distribution of city sizes follows Zipf's law and what is the composition of the resulting cities?

Null Hypothesis - It would generate N identical cities with equal sizes and an equivalent distribution of wealth.

Consumer's problem

The consumer household solves the following problem

$$\max_{\{c_t\}_{t=0}^{\infty}} \mathbb{E} \int_0^{\infty} e^{-\rho t} c_t$$
$$\dot{a_t} = w_t z_t + r_t a_t - c_t$$

where $z_t \in \{z_L, z_H\}$ follows a Poisson process.

$$b_{j,t} =$$

F.1.3 Neighbourhoods: Imagined Communities?

Countries, states, cities have fixed definitions. But what exactly is a neighbourhood? This is very important to understand how they are formed and how they determine outcomes. What are the different ways that the literature descibes a neighbourhood?

- Chetty
- Guerreri
- others

Chapter G

Feedback

Bibliography

Andreoli, Francesco, Eugenio Peluso, et al. 2018. "So close yet so unequal: Neighborhood inequality in American cities."

Bourguignon, Francois. 1979. "Decomposable income inequality measures." *Econometrica: Journal of the Econometric Society*, 901–920.

Boustan, Leah Platt. 2010. "Was postwar suburbanization "white flight"? Evidence from the black migration." *The Quarterly Journal of Economics*, 125(1): 417–443.

Card, David, Alexandre Mas, and Jesse Rothstein. 2008. "Tipping and the Dynamics of Segregation." *The Quarterly Journal of Economics*, 123(1): 177–218.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. "Where is the land of opportunity? The geography of intergenerational mobility in the United States." *The Quarterly Journal of Economics*, 129(4): 1553–1623.

Davis, Donald R, Jonathan I Dingel, Joan Monras, and Eduardo Morales. 2019. "How segregated is urban consumption?" *Journal of Political Economy*, 127(4): 1684–1738.

Fogli, Alessandra, and Veronica Guerrieri. 2019. "The end of the american dream? inequality and segregation in us cities." National Bureau of Economic Research.

Gabaix, Xavier. 1999. "Zipf's law for cities: an explanation." *The Quarterly Journal of Economics*, 114(3): 739–767.

Hardman, Anna, and Yannis M Ioannides. 2004. "Neighbors' income distribution: economic segregation and mixing in US urban neighborhoods." *Journal of Housing Economics*, 13(4): 368–382.

Piketty, Thomas, and Emmanuel Saez. 2003. "Income inequality in the United States, 1913–1998." *The Quarterly journal of economics*, 118(1): 1–41.

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman. 2018. "Distributional national accounts: methods and estimates for the United States." *The Quarterly Journal of Economics*, 133(2): 553–609.

Reardon, Sean F, and David O'Sullivan. 2004. "Measures of spatial segregation." *Sociological methodology*, 34(1): 121–162.

Reardon, Sean F, Kendra Bischoff, Ann Owens, and Joseph B Townsend. 2018. "Has income segregation really increased? Bias and bias correction in sample-based segregation estimates." *Demography*, 55(6): 2129–2160.

Schelling, Thomas C. 1971. "Dynamic models of segregation." *Journal of mathematical sociology*, 1(2): 143–186.

Wheeler, Christopher H, and Elizabeth A La Jeunesse. 2008. "Trends in neighborhood income inequality in the US: 1980–2000." *Journal of Regional Science*, 48(5): 879–891.