Are the poor constrained to use the car? Commuting to the periphery or the city center. The case of Madrid.*

Guillem Tobías Larrauri

June 19, 2024

Abstract

This paper sets out to understand the public transport ridership differences across income groups in the Madrid metro area. Using survey data from the regional transport authority, it finds a significant decrease in public transport ridership across income groups. The lowest income group commute by public transport 50 percent of the times while the highest income group uses public transport 28 percent of the time. However, these income groups live and work in different locations, higher income groups tend to work in the city center, while lower income groups tend to work in the periphery. This affects their modal choice as public transport works best towards the center. When commuting toward the city center public transport is the dominant choice, while the opposite is true for commutes in the periphery. Then, the center-periphery dimension reduces the aggregate difference between income groups and is key to understanding the effects of transport policies on different groups. Finally, these factors result in differences in commuting times across income groups, lower income groups take 10 minutes more per day to get to their job.

1 Introduction

In the XXI century there have been a number of policies to reduce pollution and congestion in large metropolitan areas. In Madrid, policies to reduce the cars in the city center

*I am deeply grateful to my tutor, Diego Puga, Dmitry Arkhangelsky and Manuel Arellano for their support and advice. This paper has also benefited from Giorgio Pietrabissa's, Jiaxuan Ren's help and from all the participants in the presentations at CEMFI. I am also indebted to Nezih Guner for his support during my time at CEMFI. Finally, I am honored to have received the Bank of Spain postgraduate scholarship during my two years at CEMFI.

and to ban the most polluting cars have been politically contentious, with some people claiming that these policies can be regressive.

This paper analyzes the existing differences in commuting patterns across income groups in the Madrid metro area with the final objective of understanding the differential effects of policies across different income groups.

To do this, this paper uses a transport survey (EDM2018) created by the regional transport authority of Madrid, it leverages its information on travel, workplace and job location. The survey contains basic demographic information from the individuals, however it does not contain income information. The income information is imputed geographically using census tract level administrative data.

This paper focuses its attention on commutes, trips from work to home and vice versa, and it finds significant heterogeneity across income groups. While public transport has a significant share in commutes for all income groups, car is the dominant mode of transport. Moreover, there is a negative correlation between income and public transport use. This is as predicted by [Glaeser et al., 2008].

However, differently from [Glaeser et al., 2008], where the poor live near the city center because of public transport, and more related to [Brueckner et al., 1999], where the amenities in the city center attract high income households, in Madrid the city center is populated largely by high income households, while the lower income households live more towards the periphery. A similar picture emerges when looking at the job distribution, low income households are less likely to work in the city center. The combination of both residential and job differences result in vastly different commuting patterns across income groups.

Then the question emerges of how much do these differences in job and residential location affect the modal choices the different income groups make. More precisely, the public transport network seems to work best when going **to** the city center or when going towards other destinations **from** the city center. The public transport shares are largest in these kinds of trips. Making the distinction between center and periphery trips reveals large differences in modal choices. When travelling to or from the city center, public transport is the dominant mode of transport, reaching shares of up to 75 percent for the lowest income group. When travelling in the periphery this trend is reversed, car is dominant. Furthermore, when combining these large modal choice differences between center and periphery with the fact that low income groups are much more likely to commute to the periphery, one can see that the differences in job and residential location reduce the differences in public transport use that would be observed if these differences did not exist.

These factors compound to create significant raw differences in the durations of these commutes, commutes for the bottom income group are 10 minutes longer than those for the highest income group, per day. When aggregating these differences to the weekly level, the difference becomes 50 minutes, which was the average raw black-white difference in durations of commutes in the US **in 1980**, per [Bunten et al., 2024].

Analyzing the durations of trips made by car, we see that after controlling for distance as well as geographical and sociodemographic characteristics, there is still a negative correlation between income and duration of commute. But these low income individuals are still using the car, then the unobserved counterfactual duration by public transport

would have to be specially high.

This paper contributes to two main strands of literature, one is on the differential effects of transport policy and the other examining how people sort in cities. A very good example of the former literature is [Barwick et al., 2021], in which the authors build a joint framework of modal choice with endogenous sorting and travel times and use it to estimate the implications of different transport policies. Estimating this paper for Madrid could be an interesting avenue for further research. [Akbar, 2022b] looks into who benefits from faster public transport, estimating a similar joint residential-modal choice framework. Finally, [Sleiman, 2024] provides an insightful view of the effects the closure of a road in the center of Paris had on different individuals. Related papers also include, [Tsivanidis, 2019], [Tang, 2021] and [Basso and Silva, 2014].

As for the sorting literature, [Glaeser et al., 2008] argues the poor live in the city center due to public transport and shows that low income individuals tend to live near public transport stations, [Akbar, 2022b] looks into the effects of public transport expansions on income segregation from a theoretical point of view, [Brueckner et al., 1999] builds a theory of location with amenties which is key to explain european cities such as Madrid where the wealthy concentrate in the city center, [Couture et al., 2023] builds a model in the same direction. [Gaigné et al., 2022] is another interesting example.

Finally, it expands a small literature focusing on Madrid, [Pietrabissa, 2023] focuses on the relationship of sorting and schools. [Muñoz Miguel et al., 2014] conducts a small survey investigating the key determinants of public transport use.

This paper emphasizes that differences in the job distribution in conjunction with a public transport network that works best for trips towards the city center has large effects on both the modal choice and residential of different income groups. It finds empirical facts supporting this hypothesis. It also gets quantitatively similar results to [Bunten et al., 2024] when comparing the commutes of low and high income groups. Finally it advances the literature for the particular case of Madrid, and while it is not able to build a general equilibrium model to study counterfactuals, it does much of the preliminary work and sets up the basis for future research.

The rest of the paper proceeds as follows, Section 2 introduces and describes the main data sources for the analysis. Section 3 conducts the main descriptive analysis. Section 4 conducts robustness checks and Section 5 concludes and outlines further steps. The Appendix contains robustness tables as well as figures and the unused part of my work.[1]

2 Data

This section describes the two main data sources for the analysis, the EDM2018 a transport survey conducted in Madrid in 2018 that contains household and trip characteristics and the Atlas2018, a geographic map of summary statistics of income data at a census tract level built from administrative data. Combining the geographical measure of income with the geographical location of the household gives a measure of the income of

[1]This includes a brief discussion of a joint residential and modal choice model by [Akbar, 2022b], a procedure to download counterfactual travel times from the TravelTime API and a download and aggregation of rental prices.

each household.

2.1 Encuesta Domiciliaria de Movilidad (EDM) 2018

The EDM2018[2] constitutes the main data source for this analysis, it is a transport survey conducted by the provider of public transport in the autonomous community of Madrid. It contains both detailed trip (mode of transport), geographic (location) and demographic characteristics for 95000 individuals. This data was collected employing a mix of face-to-face and telephone interviews. The survey strives to be representative of Madrid, in doing so, the surveyors compared the resulting data to other external sources. This process revealed no large biases beyond the imperfections that usually arise when using surveys. In making the survey, the interviewers asked for all trips made by the subject on the day previous to the interview, by doing this they built a picture of a random working day in Madrid. The survey contains detailed trip data, it contains three key components for this analysis, the modal choice for every trip, the origin and destination and the time of origin and arrival.[3] The survey also contains some basic demographic information, age, gender as well as sector of occupation and employment status. The demographic information is defined by what it lacks, that is, some sort of income measure. Below, in 2.2 I impute income geographically. During most of my analysis I restrict my sample in two main ways. First, I focus only on commutes, while adding other trips makes the analysis more interesting it also complicates it and leaves me without a clear plan. Second, the survey contains data for the whole autonomous community of Madrid, I restrict my analysis to individuals who live in the Madrid metro area and to trips that both start and end within the area. This decision was made initially due to the fact that the transport areas for the rural areas are much bigger; reflecting lower population, and that makes imputing counterfactual travel times tougher, see A.4 for a more in-depth discussion. The Metro Area was built using a criterion from the local government, it contains most important municipalities as well as Madrid, for all graphs below I call "center" the area within the M-30, which is the innermost concentric highway in the municipality of Madrid. In A.1 I discuss this classification more in depth.

The survey's design is complex, the part done in person followed a proportionally stratified design and the part done by telephone was done by quota based sampling. Probability weights were provided with the survey data. In the end the weights turned out not to matter[4], to ascertain this I conducted robustness checks and built bootstrapped confidence intervals.

Due to budgetary and representativeness concerns, the in person part of the survey was done in a stratified manner. Each transport zone received a minimum number of interviews and the rest of the interviews were assigned proportionally by population. Within each transport zone, the interviews were stratified by household size, that is if 50 percent

[2]Household mobility survey.

[3]The survey also asks people that commute by car why they didn't use public transport, and why people that commuted by public transport didn't use a car. This is very interesting information that I couldn't figure out how to leverage into my analysis.

[4]My main intuition is that weights did not matter because I aggregated everything to a certain degree, if this analysis was conducted at the original transport zones I would think that the weights would matter a significant amount.

of the households consist of 3 people in the population, then 50 percent of the sample of that stratum should be of families of 3 people.[5] For the telephone part of the survey, they sampled at the individual level and it was done using quota based sampling, from a telephone database. The quotas were based on age and gender. The fact that the two surveys used different strata complicated obtaining valid confidence intervals, however I show in Section 4 that when using bootstrap at the different strata for the different sub-samples, the errors do not change a lot and if anything they become smaller.

Because of the complex survey design the surveyors were thorough in checking that the sample was representative. They checked the demographic characteristics against other data sources from the statistical office (INE), these checks revealed no large differences beyond what one might expect from a survey. I complemented their analysis by doing the same checks without the weights and found not using the weights made no large difference. The surveyors also checked the data against very granular private data from the regional transport authority, at a granular level there were larger differences.[6] Finally, time of origin and destination were reported by the individual, there may be some biases with self reported times, [Barwick et al., 2021] mention they do happen in a similar transport suvey for Beijing. I tried downloading counterfactual travel times from TravelTime API, but there were some issues, as discussed in A.4. However, according to the methodology, [Consejo Regional de Transportes Madrid, 2019a], the times reported were checked using Google maps, so there is a standard of quality. Concluding, the survey design was complex which complicates creating valid confidence intervals, however, according to the numerous checks the surveyors performed, the sample is representative. I now move on to explain how I imputed income data for each household. For a more in-depth discussion of the survey design and representativeness, [Consejo Regional de Transportes Madrid, 2019a] is the official methodology documentation of the survey.[7]

2.2 Income data: Atlas de Distribución de Renta de los Hogares (Atlas) 2018

The EDM2018 does not provide any measure of income for the households in the survey but it does provide the geographical location of the households. Using that information in addition to the Atlas2018, which provides income information at the census tract level, we can impute a measure of household income for each household in our survey. Since the key results stem from this imputation, the procedure and assumptions behind it are explained in detail. The procedure consists of three steps: The first step is to go from the census tract level Atlas data to a transport zone level data, to do that I use areal interpolation. The second step is just to impute the income of each household, knowing their area of residence. As this is an imputation that can be imprecise the households

[5]With this is makes sense that approximately the weights shouldn't matter as from [Arellano, 2014], the weight factor should be close to 1. If there are 60 penguins and 40 lions in my population and I stratify at the animal level and I sample 30 penguins and 20 lions the weight factor is 1. In my more complex case this happens approximately.

[6]I could not replicate this analysis because this data was not available.

[7]It is not particularly well structured or written.

are aggregated into five income groups, that is the third step.[8]

The survey data contains information at a reasonably fine geographical level, it contains 1259 for the whole community of Madrid, with the areas being smaller where the population density is higher. However, the Atlas2018 data is at a census tract level, which is even finer. To go from one geography to another, the Tobler[9] subpackage of the PySal package, [Rey and Anselin, 2010], was used. The procedure used is just a weighted average where the weights come from the area the two geographies share.[10][11]

This procedure would be more problematic if the average income of census tracts that made up a transport area were very different, as it would mean significant income heterogeneity amongst the transport zone. However, this does not seem to be the case in the data. This still does not address the fact that there may be significant income heterogeneity within the census tract.

To address this heterogeneity, I aggregated the measure of income into 5 groups. The distribution and cutoffs are shown in Figure 1. The hope is that while the income signal may be noisy, it falls within the income bucket. Of course, this argument is imperfect, especially for household where the income is close to one of the cutoffs for the different income groups. As a worse case scenario, we know that the individuals in the lowest income quintile are very likely to have lower income than those at the highest income quintile.

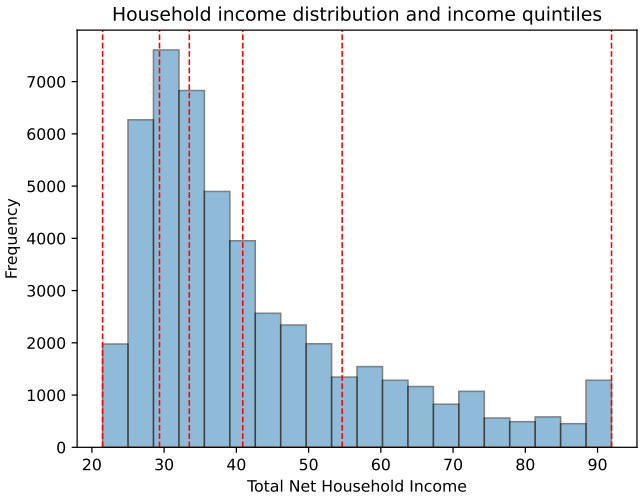


Figure 1: Income distribution and income group cutoffs. *Notes:* This graph shows the imputed income distribution at a household level for the households in the sample who live in the Madrid metro area.

[8] A way to build a more precise measure of income would be to use the age and education data from the individuals as a signal of income in addition of the geographic signal.

[9] https://pysal.org/tobler/

[10] An example, a transport zone (big area) is made of (shares area) with two census tracts (small area). The zone's area is made of 70 percent census tract one and the remaining 30 percent is census tract two. Then my procedure results in the income of the transport area to be 0.7*(income of census tract 1) +0.3*(income of census tract 2).

[11] The package contains more advanced commands like including other population weights, those require more data and I believe the simple areal interpolation provides a good enough measure.

3 Descriptive analysis

This section describes the observed differences in transport mode choices amongst income groups. Firstly, it finds that public transport ridership is negatively correlated with income. Secondly, it finds large differences in both residential and job location, which result in stark differences in commuting patterns. Third, it distinguishes between trips to the periphery and to the city center. When commuting to the city center public transport is the dominant choice, however, low income groups commute to the city center disproportionally less. Fourth, the first three factors result in a raw duration gap of 10 minutes per day between the top and bottom income group. That is, the bottom income quintile commutes an average of 10 more minutes per day. Fifth, for commutes made by car, low income groups take longer even after controlling for geographical and sociodemographic characteristics. The fact that these low individuals still use the car must reflect very poor public transport connection along their routes.

3.1 Modal choice by income group

While commuting by car is the dominant choice in the aggregate distribution, there is a significant share of commutes made using public transport. Moreover, as you can see in Figure 2, the aggregate transport shares hide substantial heterogeneity across income groups. There is a strong negative correlation between income and public transport use.

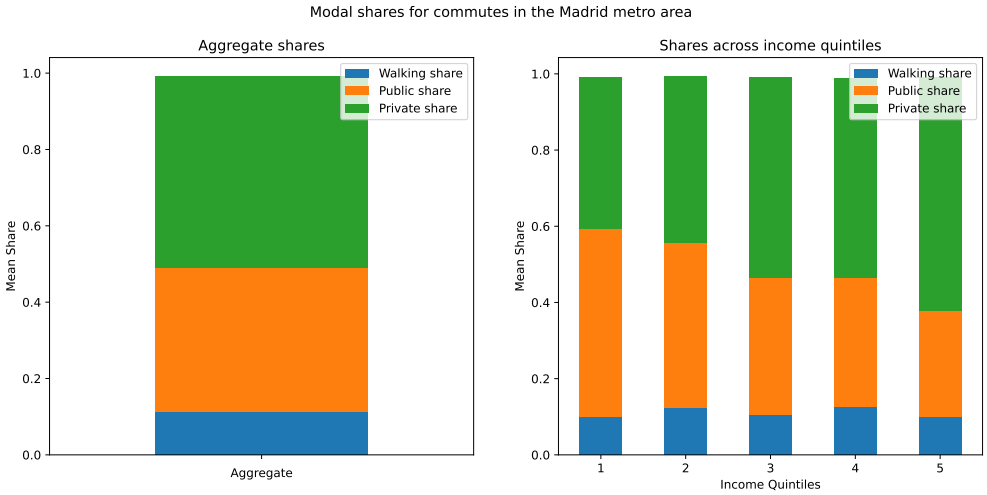


Figure 2: Aggregate transport shares for commutes and transport shares by income groups. *Notes:* This graph shows the aggregate transport shares as well as the income quintile differences for commutes. *Sample:* The sample consists of all commutes from individuals who live in the Madrid metro area and for trips that both start and end within the metro area.

For other kinds of trips, such as trips to study, whereas school high-school or university, or shopping trips, a clear pattern emerges, high income individuals tend to use the car more. While the shares are very different across different reasons for the trip, high income individuals use the car more, the opposite is not always true, low income individuals don't always have higher public transport shares than high income individuals. For example, in

study trips, public transport shares are constant and the increase in share of trips made by car as income increases comes at the expense of a lower share of trips made on foot. This is shown in Figure A2.

Focusing back on commutes, even as this negative correlation between income and public transport use is predicted from Becker's theory of allocation of time[12], it is interesting to see how the location of residence and jobs from different groups affects this choice. This is also connected to the literature, [Glaeser et al., 2008] argues that the low income groups should live in the city center due to public transport, while the poor will use public transport in larger numbers they will not live in the city center, see Table 1.

3.2 Differences in residential and job location across income groups

The different income groups have very different residential and job locations. In particular, low income individuals are very unlikely to reside in the city center and are very likely to live towards the southern outskirts of Madrid and the other southern towns within the metro area. On the other hand, high income individuals are very likely to live in the city center as well as the northern outskirts of Madrid and the western municipalities. The job distribution presents similar geographical patterns, while it still presents significant skewness, that skewness is lower than that exhibited by the residential distribution. In particular, a large proportion of low income individuals work in the city center, while almost none live there. The combined residential and job differences result in vastly different commuting patterns, particularly important is that low income individuals are much less likely to commute from or towards the city center.

Income Quintile Residential Location	1	2	3	4	5
center	210	462	665	1360	1278
east	100	727	772	526	76
madrid down	2332	1197	1081	393	50
madrid up	548	652	483	1008	1568
north	40	186	123	592	373
south	1130	1164	1357	615	19
west	15	9	134	168	1142

Table 1: Residential location for different income groups. *Notes:* The center (of Madrid), madrid up and madrid down are the municipality of Madrid, the city center and the outskirts respectively. The rest are areas from the metro area named geographically. *Sample:* The sample consists of all commutes from individuals who live in the Madrid metro area and whose trips both start and end within the metro area.

[12]High income individuals have higher wages and when commuting to work they are losing those wages, this makes them want to go faster and use a car

Income Quintile Job Location	1	2	3	4	5
center	1318	1262	1347	1607	1840
east	187	498	470	337	130
madrid down	909	622	582	383	209
madrid up	872	920	936	1198	1184
north	173	237	225	476	350
south	718	677	786	407	177
west	194	164	261	246	609

Table 2: Workplace location for different income groups. *Notes:* The center (of Madrid), madrid up and madrid down are the municipality of Madrid, the city center and the outskirts respectively. The rest are areas from the metro area named geographically. *Sample:* The sample consists of all commutes from individuals who live in the Madrid metro area and whose trips both start and end within the metro area.

The fact that income is imputed geographically is specially worrying here, the fact that the income distribution is so skewed may be be caused by the way the income data was imputed, most worryingly of all is how this translates to the differences I observe for the job distribution. It is hard to logically derive in which direction the hypothetical bias might go. Let us suppose that the income measure is imperfect and that in every income group there are some individuals that do not belong in it, would a perfect income measure result in larger or lower differences across the job distribution? Intuitively if the income group is what dominates where a job is, you would think the differences in job distribution would be larger. However I am sure there are counterarguments where the balance would swing in the other direction.

Finally, putting the job and residential distribution together we can observe the resulting matrix of trips, in Figure 3 and Figure 4, you can observe the commutes for the bottom and top income group respectively. Immediately, one can see the large disparities in the the areas where the commutes are happening, due to the job and residential location disparities, and more interestingly, the top income group individuals are much more likely to commute towards and from the city center (middle row and middle column) than low income group individuals. This is due both to the fact that they are more likely to live and more likely to work in the city center. This distinction between trips to/from the center and trips in the periphery is a key distinction in the analysis that follows.

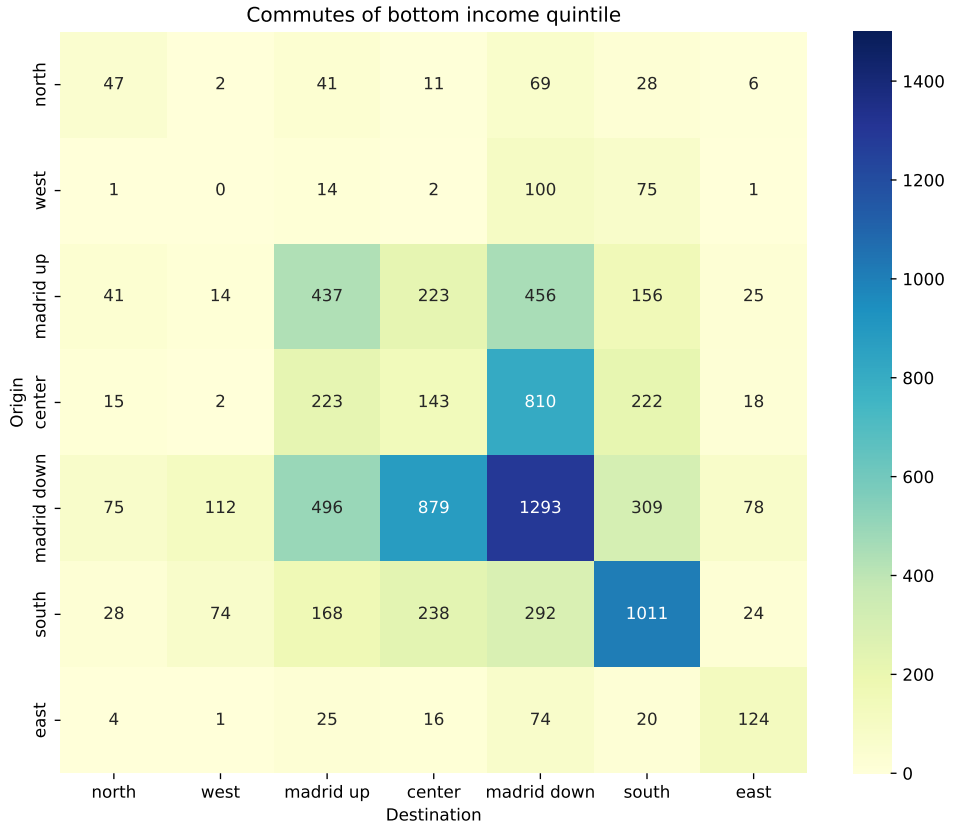


Figure 3: Trip matrix of commutes from the bottom income quintile. *Notes:* This figure shows the trip matrix for the commutes made by the individuals from the lowest income quintile. The rows are the different origins and the columns are the destinations. *Sample*: Individuals who live in the metro area and commutes that start and end in the metro area, the income groups were computed at the household level for households living in the metro area.

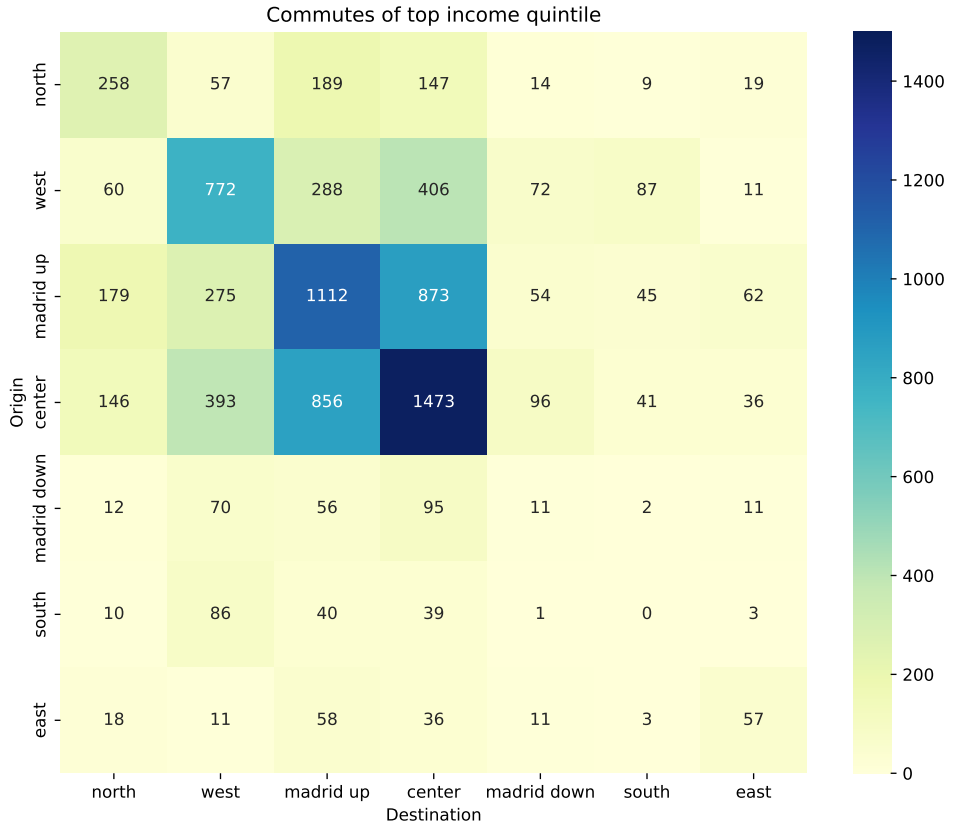
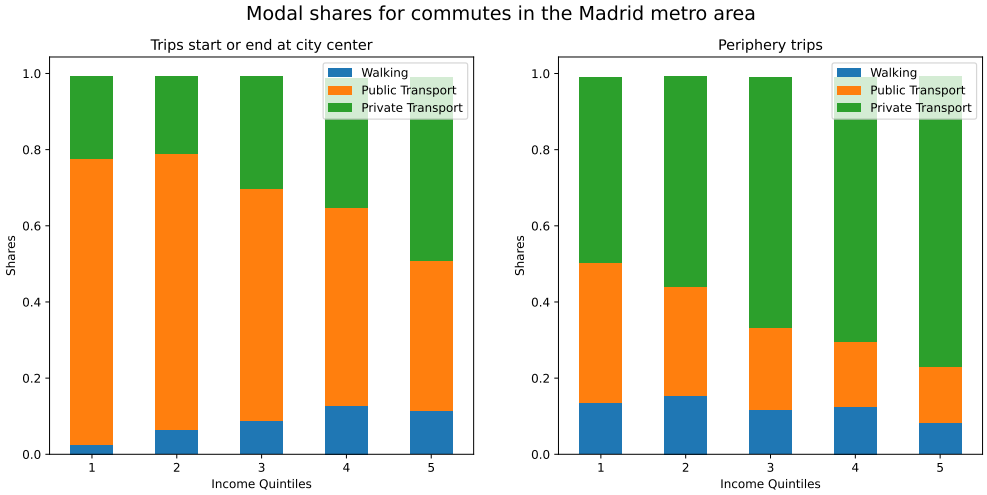


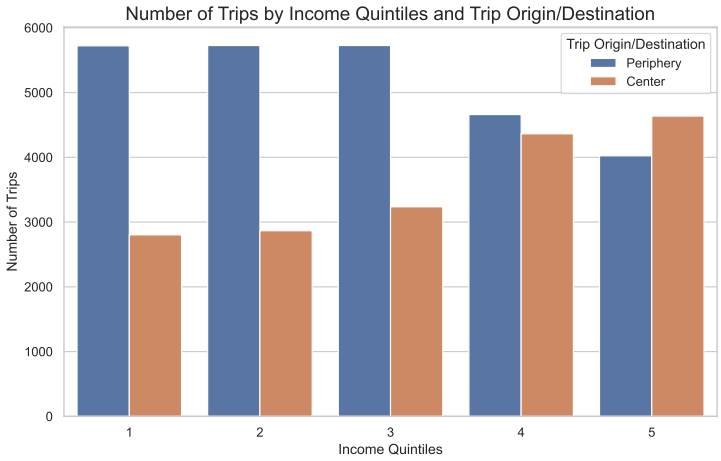
Figure 4: Trip matrix of commutes from the top income quintile. *Notes:* This figure shows the trip matrix for the commutes made by the individuals from the lowest income quintile. The rows are the different origins and the columns are the destinations. *Sample*: Individuals who live in the metro area and commutes that start and end in the metro area, the income groups were computed at the household level for households living in the metro area.

3.3 Differences between center and periphery

There are large differences in public transport shares depending on if the trip is to/from the center (center) or periphery. When the type of trip is of "type" center, public transport shares reach up to 75 percent for the lowest income groups. On the other hand, when commuting in the periphery, car is the dominant mode of transport for all income groups. Interestingly, due to differences in the job and residential distribution, low income groups are much more likely to commute in the periphery. This significantly reduces the observed differences in modal choice across income groups from a hypothetical situation where both the residential and job distribution were equal across income groups.



(a) Modal shares by income group and origin-destination: center or periphery. *Notes:* The left panel shows the shares for trips that either start or end in the city center, by income quintile. The right panel shows the shares trips for all other locations, denoted as periphery trips, by income quintile.



(b) Number of trips by income group and origin-destination: center or periphery. *Notes:* Every income group has two columns the left is trips in the periphery and the right column are trips that either start or end in the city center.

Figure 5: Transport shares and number of trips (commutes) by income and center/periphery dimension. *Sample:* These are commutes that both start and end in the Madrid metro area for individuals who reside in the metro area.

3.3.1 Who drives in the city center: Implications for policy

An immediate implication of Figure 5, is who drives in the city center. In Madrid there have been endless debates about traffic-pacifying policies in the city center, this shows that the individuals who drive in the city center are, to a large degree, from high income groups, see Figure 6.

This is the consequence of two distinct facts. First, low income individuals are less likely to work in the city center and much less likely to live in the city center 1. This combines to make their commute much less likely to begin or end at the city center. Second, when commuting to the city center, the lowest income groups are very unlikely to commute

using the car. The combination of these two facts results in 38 percent of all car trips made in the city center being from individuals from the highest income quintile.

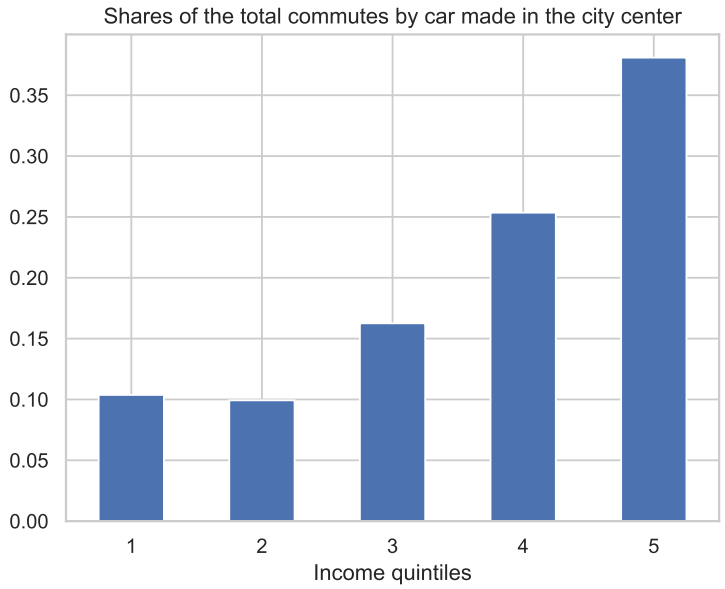


Figure 6: Who drives in the city center? *Notes:* Share of the total commutes made **by car** in the city center (either origin or destination) (by all income groups). *Sample:* These are commutes that both start and end in the Madrid metro area for individuals who reside in the metro area.

While it is naive to think that a policy that affects traffic in the city center will not have spillovers to the rest of the city([Sleiman, 2024]), this provides a picture as to whom a hypothetical policy would affect most directly.

3.3.2 Back of the envelope counterfactual

What if all income groups had an equal number of trips to the periphery and city center? Mechanically, from Figure 5 the public transport share of low income groups would increase substantially, as can be seen from Figure 7. On the other hand, the public transport shares for the high income groups would not change significantly as their trips split close to half and half between center and periphery.

While this counterfactual lacks the validity of a general equilibrium framework, it works perfectly as an excuse to start thinking about deeper questions. If the job distribution were the same across income groups what would happen to the location choices of different households, what about the transport choices? Would a larger portion of low income individuals live in the city center and travel by public transport or would the increase in prices in the city center result in low income groups working further away and using car? It could also be used to think about the implications of transport improvements to the city, again, both in terms of segregation and transport mode shares. [Akbar, 2022a], claims that the equilibrium sorting would depend on whether it was fast public transport, that is metro/train, or slow public transport, like a bus, this could be a nice empirical case to check his theoretical hypothesis.

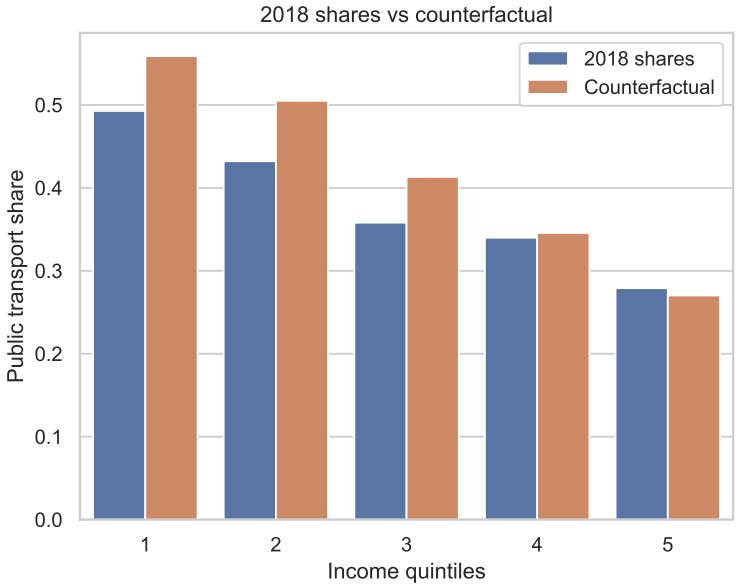


Figure 7: 2018 public transport shares by income quintile versus hypothetical shares if trips to the periphery and the city center were equalized. *Notes:* Each income quintile has two columns, the left column is public transport shares in 2018 and the right column is the counterfactual of equal trips. *Sample:* These are commutes that both start and end in the Madrid metro area for individuals who reside in the metro area.

3.4 Duration analysis

Poorer individuals take significantly longer to get to work, when looking at the raw differences by income, the poorest individuals take, on average, 10 more minutes to work every day. When analyzing trips made by car and public transport separately, after controlling for observable characteristics, low income individuals take longer when commuting by car, when commuting by public transport there are no income differences, the implications are discussed below.

3.4.1 Raw duration differences

The lowest income group take, on average, 35 minutes to get to work, the highest income group takes 30 minutes. While only a 5 minute difference may seem small, this trip happens 2 times per day and 5 days per week, this leads to very significant weekly differences of on average 50 minutes per week. To compare, [Bunten et al., 2024] find that 50 minutes per week was the raw black-white duration, *in 1980*, therefore these differences are large. In the appendix, A10, shows the differences are statistically significant.

3.4.2 Public and private transport trips

Analyzing the trip durations is complex as there are obvious selection issues involved, every individual has a choice between commuting by car and by public transport, and, the inferior choice is discarded, while duration is not the only motive for picking one

transport mode or the other, it is one of the main ones, and so, this will affect our results. However we can work around the selection issue to our advantage.

As you can see in Table 3 , when we focus on trips made by car, after controlling for individual education, job sector, age, gender, as well as geographical indicators; origin-destination dummies[13] and straight line distance, income still has statistically significant power. A higher coefficient relative to the baseline (lowest income individuals) means that wealthier individuals are more likely to stop commuting, which results in shorter commutes. More specifically, after controlling for all these factors, the lowest income individuals still take more time in their commutes. Therefore it seems that low income individuals tend to commute to worse communicated or more congested places.

Dependent Variable:	Private transport duration		
Income group	exp(coef)	confidence (95)	observations
1	-	-	3384
2	1.05	[0.97-1.14]	3742
3	1.11	[1.03-1.19]	4703
4	1.14	[1.06-1.23]	4696
5	1.14	[1.04-1.24]	5249
Controls	YES		

Table 3: Effects of income on private transport durations. *Notes:* These results are the coefficients from a proportional hazards duration model, higher coefficients (larger than 1) relative to the baseline, means shorter commutes. The controls include origin-destination dummies, straight line distance as well as demographic characteristics. *Sample:* These trips are commutes that both start and end within the Madrid metro area from individuals who reside in the Madrid metro area and use private transport.

Moreover, we can rephrase this finding in a more interesting way, these low income individuals take longer than the high income individuals after controlling for a variety of factors, but these low income individuals are still commuting by car, therefore the time these individuals would face if they made their commutes by public transport would need to be substantially worse to justify the use of car. While this is interesting it does not inform us of the relative choice (car and public times) difference between the rich and poor, this is because we expect the rich to use the faster mode of transport even for small time differences between car and public transport, therefore we don't know if the car-public time differential is bigger for rich or poor individuals, just that it must be very big for these low income individuals.

On the other hand, when looking at the public transport times,5, it seems that there is not a big relationship with income, the coefficients are not statistically significant and even become sligthly negative for the highest income group, that is for similar public transport trips, high income individuals take slightly longer. This result may be due to residential sorting, in line with [Glaeser et al., 2008] the rich may locate further away from public transport, however, the question is then why are these high income individuals using public transport in the first place.

[13]The dummies are all origin-destination combinations from A.1

A more direct and complete way to proceed would be to compare the observed travel time the individuals face with the counterfactual travel time from the transport mode they did not pick. I tried to obtain this data but wasn't succesful, see A.4. An imperfect way to go around this missing data issue would be to aggregate the areas a reasonable amount and compare similar trips from different individuals, while there would be some selection issues, this could be an approach worth trying in further research, if downloading counterfactual travel times is impossible.

4 Robustness

I conducted different robustness exercises to see if the center periphery patterns I observed were robust, these exercises included building a valid confidence interval using bootstrap, using the probability weights provided by the survey, as well as going beyond commutes and restricting my sample to individuals from age 30 to 60 that are registered to live in Madrid, in all these cases the center-periphery results do not different significantly from the baseline results. How to improve the income imputation is briefly discussed and left for future research.

4.0.1 Bootstrapped standard errors and sampling weights

As I mentioned in Section 2.1, the survey has a complex survey design, therefore building valid confidence intervals is not trivial. Even though the sample size is large I believed this exercise to be worthwhile, in the end, bootstrapping at the different strata gave small confidence intervals such that all the center-periphery facts I had found were statistically significant. Redoing my center-periphery analysis using weights gave me identical results, as you can see in Figure A7, to the unweighted baseline center of Figure 5.

Income Quintile Job Location	1	2	3	4	5
Periphery	[0.35-0.38]	[0.27-0.30]	[0.20-0.23]	[0.16-0.18]	[0.13-0.16]
Center	[0.73-0.77]	[0.71-0.74]	[0.59-0.63]	[0.50-0.53]	[0.38-0.41]

Table 4: Bootstrapped confidence intervals for public transport shares by center-periphery and income.*Notes:* Bootstrap of the in-person and telephone survey at their corresponding strata, within strata the bootstrap happens at the family level. *Sample:* All possible households in the original sample. 1000 bootstrap samples.

To bootstrap, I followed [Arellano, 2014], and bootstrapped the different strata. As the strata were different for the in person and telephone part of the survey I distinguished between the two parts of the survey. For the in-person I bootstrapped at a transport zone level by size of the family and for the telephone one I bootstrapped at the transport zone level and by gender.[14]

[14]For both samples I was bootstrapping families, I thought that was the most adequate.

4.0.2 Different samples

I conducted the same center-periphery analysis on two different samples, one was for all trips in the metro area, bunching together trips caused by different motives, and the other was to restrict my sample to individuals of age 30 to 60 who are registered to live in Madrid, both samples gave very similar answers as to my baseline. This can be seen in A.2.3.

I wanted to see whether the existing commuting patterns in terms of center and periphery modal choice were similar to the patterns for all other trips.[15] Very similar patterns emerged, although there is a larger share of trips made walking, especially for the lower income groups. In terms of center and periphery, the same relationship holds because a lot of high income individuals live in the city center and for short trips, such as shopping or school trips, they "mechanically" travel where very close to where they live.

The reason to analyse the 30 to 60 sub-sample was for modal choice analysis, while in the end I couldn't conduct this modal choice analysis, the individuals should be in some sort of steady state. Imagine an individual who has just arrived to Madrid and their job is very far away, they go by bus and take a long time but they are saving up to buy a car, that would bias any analysis. A bigger worry is young people that are changing jobs, a very similar story can be told. That is why I restricted the sample to people age 30 to 60 and who are registered in Madrid, this should proxy for a steady state. It is an imperfect measure, as these people may be changing jobs or having children or changing residence, however it is the best I could do given the limited information in the data. The center periphery results did not change.

4.0.3 Augmented income measure

This should have been my immediate focus in any robustness exercise I conducted, as it is the key assumption from where my results follow. While the fact that I used aggregate income groups should reduce the adverse effect of the noise in my income measure, there may still be some individuals that I wrongly assigned to an income group that was not theirs.

Unfortunately I didn't have time, but a first step in improving the income measure would have been to use the age and education status of the individuals to infer their income more precisely. It is interesting to think whether this information should be inferred at the family or the individual level. But regardless of any new potential issues, this improved income measure would have greatly improved my analysis.

5 Conclusion

This paper finds large differences in transport shares across income groups, while higher income groups tend to commute more by car, this difference is dampened by the fact that high income individuals live and work more in the city center, where the public transport network works best relative to car. When analyzing the duration of the commutes, low

[15] This sample has the disadvantage of comparing very different trips in terms of length.

income groups take significantly longer. On the observed commutes by car, low income groups take longer, even after controlling for geography and sociodemographic characteristics. This suggests that the counterfactual public transport durations must be very high for those low income individuals. While I was not able to download counterfactual travel times[16], which would have allowed me to answer deeper questions of city structure as well as to run a proper modal choice analysis, I believe this paper shows interesting facts and sets the basis for future research. An option would be to follow, [Barwick et al., 2021], who build a transport-residential choice model to analyse transport policies. Another direction would be to see how the transport infrastructure affects the growth of the city, [Duranton and Puga, 2023], as well as the welfare of its different inhabitants.

References

[Akbar, 2022a] Akbar, P. A. (2022a). Public transit access and income segregation. Technical report, SSRN.

[Akbar, 2022b] Akbar, P. A. (2022b). Who benefits from faster public transit? Technical report, Unpublished manuscript. Available at Google Drive: `https://drive.google.com/file/d/1zexOoCxNZThfXl9wzk5DGFlvuIb4gJv7/view`.

[Arellano, 2014] Arellano, M. (2014). Econometrics of survey data. CEMFI Slides.

[Barwick et al., 2021] Barwick, P. J., Li, S., Waxman, A. R., Wu, J., and Xia, T. (2021). Efficiency and equity impacts of urban transportation policies with equilibrium sorting. Working Paper 29012, National Bureau of Economic Research. Revision Date: February 2022.

[Basso and Silva, 2014] Basso, L. J. and Silva, H. E. (2014). Efficiency and substitutability of transit subsidies and other urban transport policies. *American Economic Journal: Economic Policy*, 6(4):1–33.

[Berry et al., 1995] Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890.

[Brueckner et al., 1999] Brueckner, J. K., Thisse, J.-F., and Zenou, Y. (1999). Why is central paris rich and downtown detroit poor?: An amenity-based theory. *European Economic Review*, 43:91–107.

[Bunten et al., 2024] Bunten, D. M., Fu, E., Rolheiser, L., and Severen, C. (2024). The problem has existed over endless years: Racialized difference in commuting, 1980–2019. *Journal of Urban Economics*, 141:103542.

[16] I did download them but they were unfortunately not good enough. I also devoted a large part of my time to understand [Berry et al., 1995] style econometrics and while I didn't get the chance to apply it in the main body of the paper, I did learn from it. Looking back, if I had to start again, I would focus my time much more in robustness for my income measure, which is the key from where my results follow, instead of doing non-essential robustness. I would also zero in on a specific question from the beginning and I would also have kept a much tidier code. However, as I feel I have learnt and improved along every step of this process, I am satisfied, even if the final product is very imperfect.

[Consejo Regional de Transportes Madrid, 2019a] Consejo Regional de Transportes Madrid (2019a). Documento i. metodología y trabajo de campo. edm2018. Technical report, Comunidad de Madrid, Pl. del Descubridor Diego de Ordás, 3, Chamberí, 28003 Madrid.

[Consejo Regional de Transportes Madrid, 2019b] Consejo Regional de Transportes Madrid (2019b). Documento síntesis edm2018. Technical report, Comunidad de Madrid, Pl. del Descubridor Diego de Ordás, 3, Chamberí, 28003 Madrid.

[Couture et al., 2023] Couture, V., Gaubert, C., Handbury, J., and Hurst, E. (2023). Income growth and the distributional effects of urban spatial sorting. *The Review of Economic Studies*, 91(2):858–898. Retrieved June 15, 2024.

[Davidson-Pilon, 2019] Davidson-Pilon, C. (2019). lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40):1317.

[Duranton and Puga, 2023] Duranton, G. and Puga, D. (2023). Urban growth and its aggregate implications. *Econometrica*, 91(6):2219–2259.

[Gaigné et al., 2022] Gaigné, C., Koster, H. R., Moizeau, F., and Thisse, J.-F. (2022). Who lives where in the city? amenities, commuting and income sorting. *Journal of Urban Economics*, 128:C.

[Glaeser et al., 2008] Glaeser, E. L., Kahn, M. E., and Rappaport, J. (2008). Why do the poor live in cities? the role of public transportation. *Journal of Urban Economics*, 63:1–24.

[Muñoz Miguel et al., 2014] Muñoz Miguel, J. P., Simón de Blas, C., and Jiménez Barandalla, I. C. (2014). Estudio empírico sobre la utilización del transporte público en la comunidad de madrid como factor clave de movilidad sostenible. *Cuadernos de Economía*, 37:112–124.

[Pietrabissa, 2023] Pietrabissa, G. (2023). School access and city structure. Job Market Paper.

[Rey and Anselin, 2010] Rey, S. J. and Anselin, L. (2010). Pysal: A python library of spatial analytical methods.

[Sleiman, 2024] Sleiman, L. B. (2024). Displacing congestion: Evidence from paris. Working paper, Center for Research in Economics and Statistics.

[Sánchez, 2020] Sánchez, R., C. F. . P. J. (2020). Mapa — la brecha económica de la movilidad en madrid: las zonas ricas se desplazan en coche y las pobres, a pie. *diario.es*. 7 de marzo de 2020, 20:48.

[Tang, 2021] Tang, C. K. (2021). The cost of traffic: Evidence from the london congestion charge. *Journal of Urban Economics*, 121:103302.

[Train, 2009] Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge.

[Tsivanidis, 2019] Tsivanidis, N. (2019). Evaluating the impact of urban transit infrastructure: Evidence from bogotá's transmilenio. Preprint, University of California, Berkeley.

[Villanueva, 2021] Villanueva, J., G. D. . A. B. (2021). ¿cómo ha cambiado la movilidad en madrid? un millón de viajes menos en transporte y coches como antes. *El País*. 12 de diciembre de 2021, 05:00 CET.

A Appendix

This appendix includes:

- My subdivision of the Madrid metro area into 7 areas.
- Tables and figures from robustness checks.
- A brief description of [Akbar, 2022b].
- A procedure to download counterfactual travel times.
- Data sources for rental prices in Madrid.

A.1 7 areas of the Madrid metropolitan area.

The original areas (1259) from the EDM2018 are too many to plot them in any meaningful way. That is why I restricted my analysis to the Madrid metro area[17] and then subdivided the resulting area into 7 large areas.

The area was created following a criterion made by the local government at the start of the 21st century and it includes Madrid and most other important population centers[18]. A valid criticism is to think if my conclusions are robust to another criterion of the Madrid metro area such as a similar but wider definition made by the transport authority.[19] These excluded municipalities are more likely to be populated by low income individuals, however as these municipalities are less populous, they are unlikely to have large effects on my results

[17] In addition of making it easier to make plots, I also restricted my analysis to the Metro area because due to higher population densities, the transport zones in the area were much smaller, this should have helped have a more accurate picture when downloading counterfactual travel times. This is discussed in detail in the Section A.4 of the Appendix.

[18] The name is Atlas de la Comunidad de Madrid en el umbral del siglo XXI

[19] In retrospect this would have been better criterion but I didn't find see it at the time, it can be found in the same Wikipedia page.

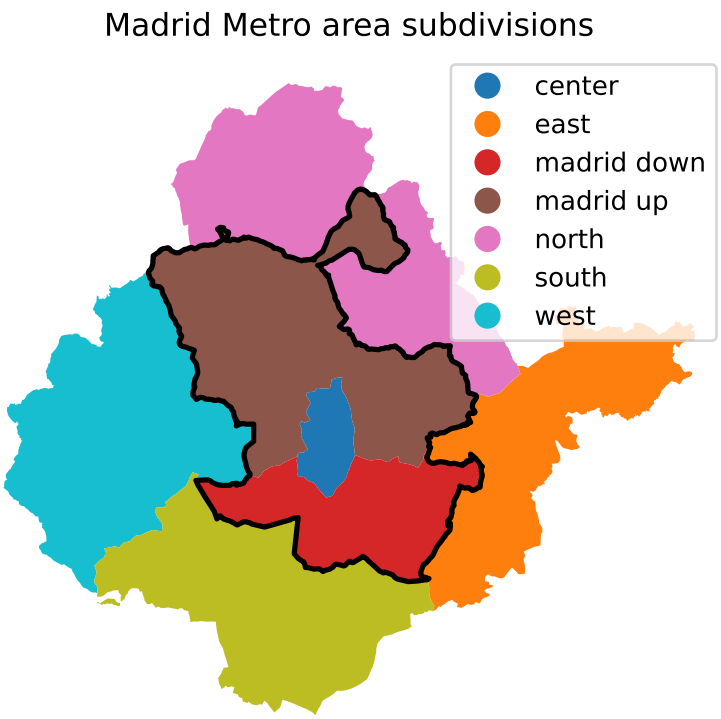


Figure A1: 7 areas of the Madrid Metro area. *Notes:* These are the areas used in the commuting matrix plots. The Madrid municipality is delineated in black. The center is delineated by first concentric highway (M30). The rest of the areas are named geographically.

A.2 Figures and tables

This section contains figures from income differences in other kinds of trips, the commuting matrices for the 3 middle income quintiles, some robustness of the center-periphery result and the duration analysis.

A.2.1 Income differences for other kinds of trips

I show the income differences for study trips and shopping trips.

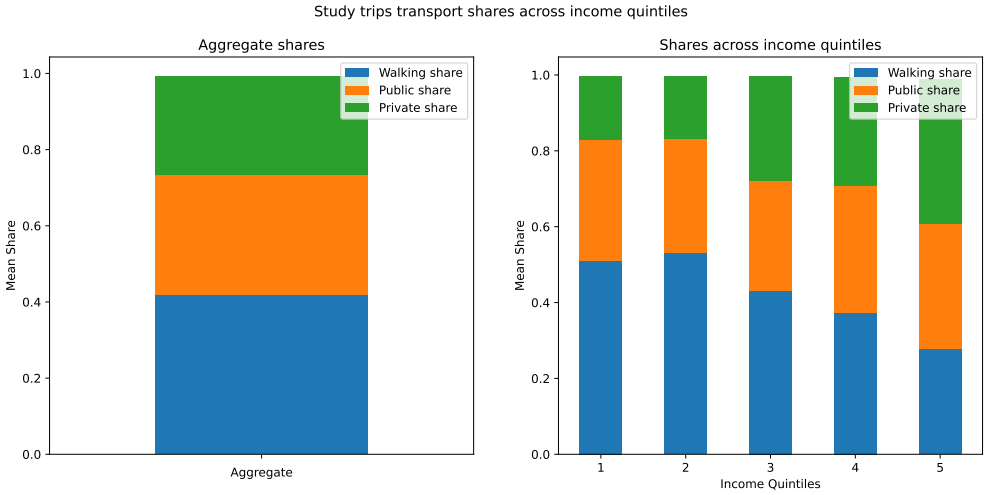


Figure A2: Aggregate transport shares for study trips and transport shares by income groups. *Notes:* This graph shows the aggregate transport shares as well as the income quintile differences for study trips. *Sample:* The sample consists of study trips from individuals who live in the Madrid metro area and for trips that both start and end within the metro area.

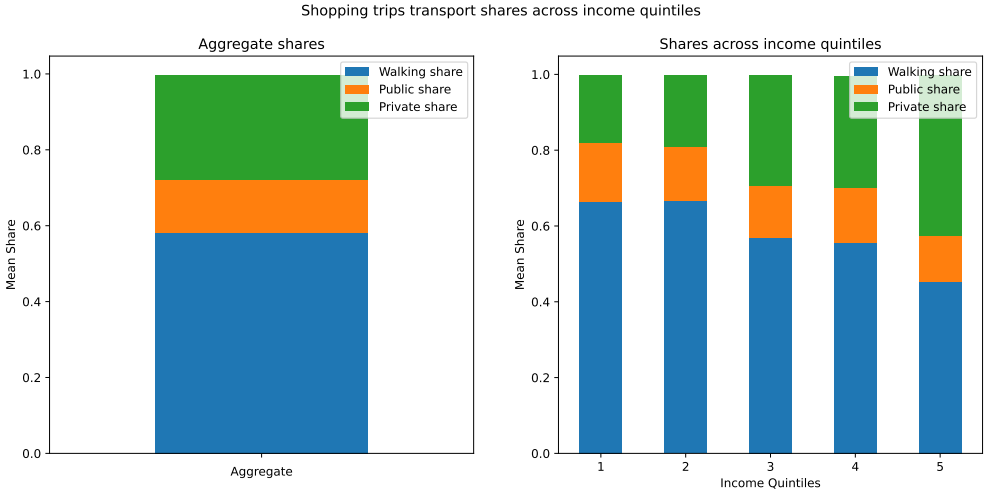


Figure A3: Aggregate transport shares for shopping trips and transport shares by income groups. *Notes:* This graph shows the aggregate transport shares as well as the income quintile differences for shopping trips. *Sample:* The sample consists of shopping trips from individuals who live in the Madrid metro area and for trips that both start and end within the metro area.

A.2.2 Commuting Matrices

I show the commute matrices for the second, third and fourth income quintiles, respectively.

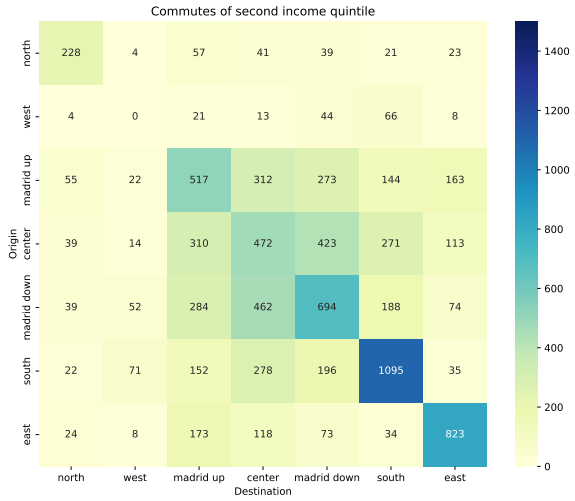


Figure A4: Trip matrix of commutes from the second income quintile. *Notes:* This figure shows the trip matrix for the commutes made by the individuals from the second income quintile. The rows are the different origins and the columns are the destinations.*Sample*: Individuals who live in the metro area and commutes that start and end in the metro area, the income groups were computed at the household level for households living in the metro area.

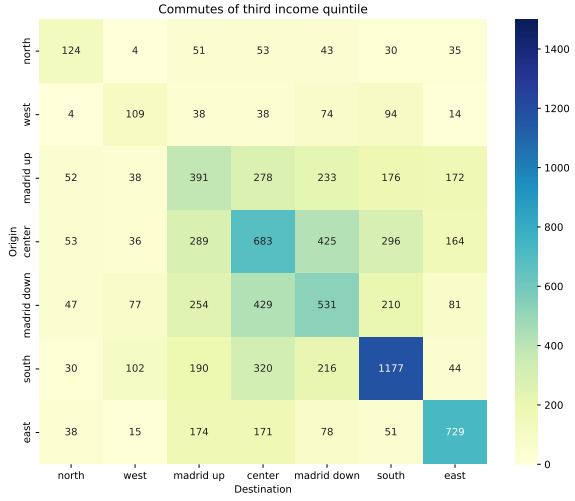


Figure A5: Trip matrix of commutes from the third income quintile. *Notes:* This figure shows the trip matrix for the commutes made by the individuals from the third income quintile. The rows are the different origins and the columns are the destinations.*Sample*: Individuals who live in the metro area and commutes that start and end in the metro area, the income groups were computed at the household level for households living in the metro area.

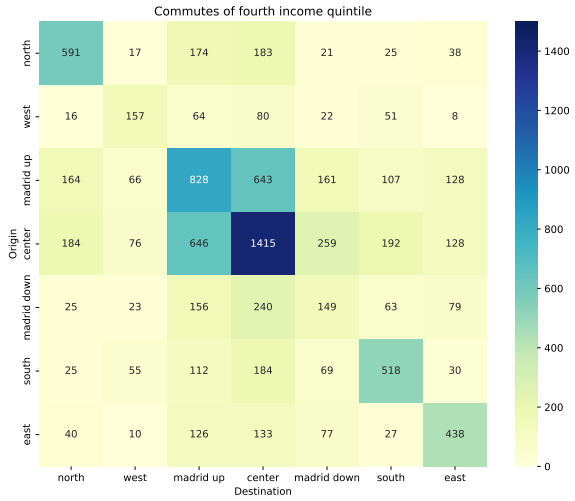
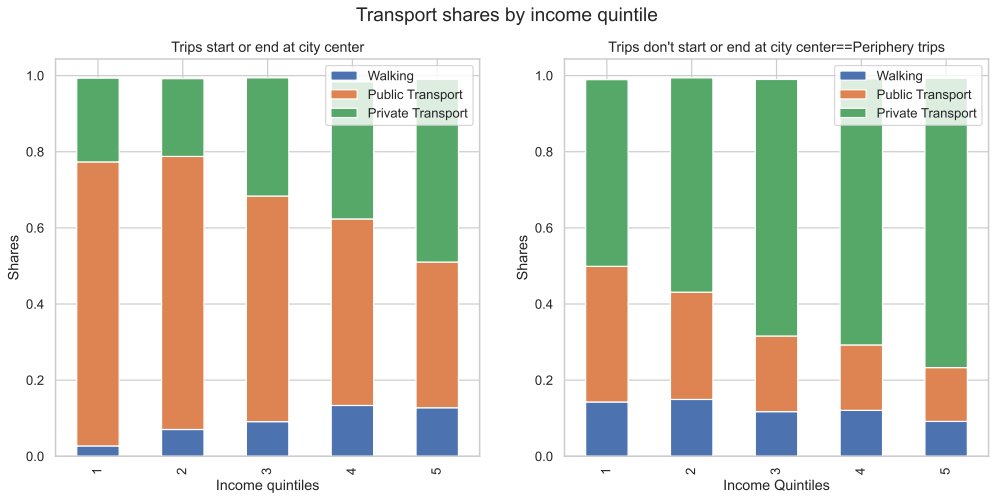


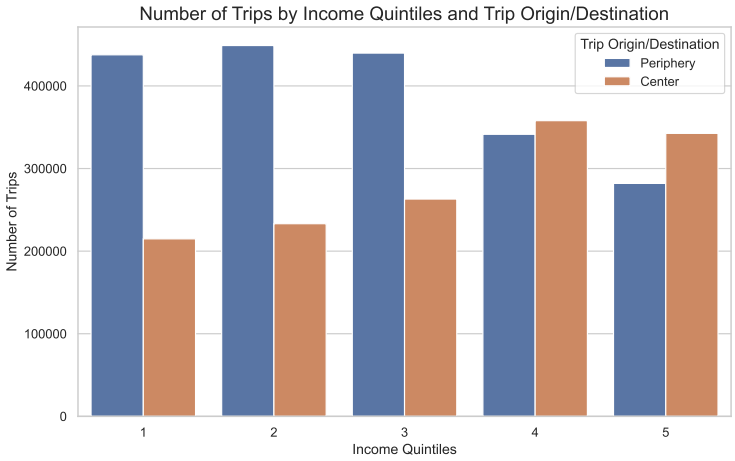
Figure A6: Trip matrix of commutes from the fourth income quintile. *Notes:* This figure shows the trip matrix for the commutes made by the individuals from the fourth income quintile. The rows are the different origins and the columns are the destinations.*Sample*: Individuals who live in the metro area and commutes that start and end in the metro area, the income groups were computed at the household level for households living in the metro area.

A.2.3 Robustness exercises

This shows the main center-periphery figure, Figure 5, for different samples, one is using the weights provided in the sample and the other restricts the sample to individuals aged 30 to 60 who are registered to live in Madrid, the last one is done for all trips (commutes, study, shopping, etc.).

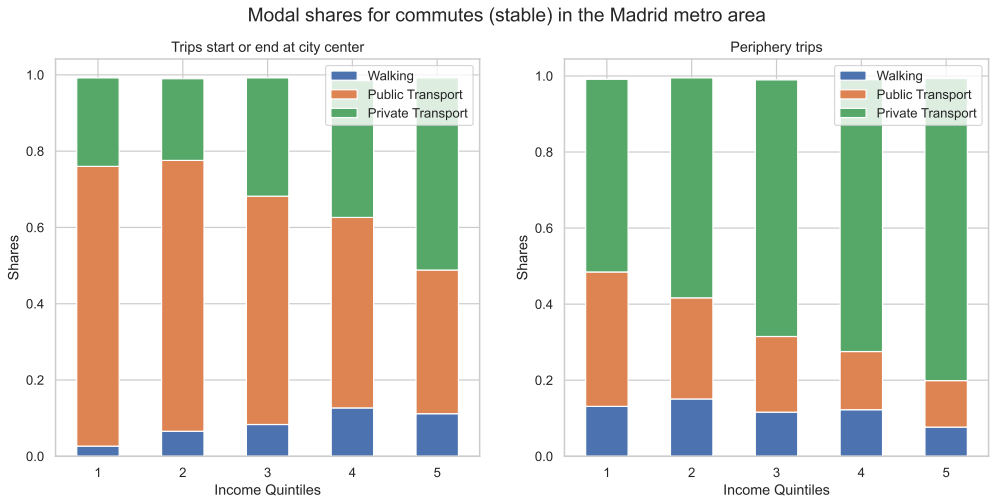


(a) Modal shares by income group and origin-destination: center or periphery. *Notes:* The left panel shows the shares for trips that either start or end in the city center, by income quintile. The right panel shows the shares trips for all other locations, denoted as periphery trips, by income quintile.

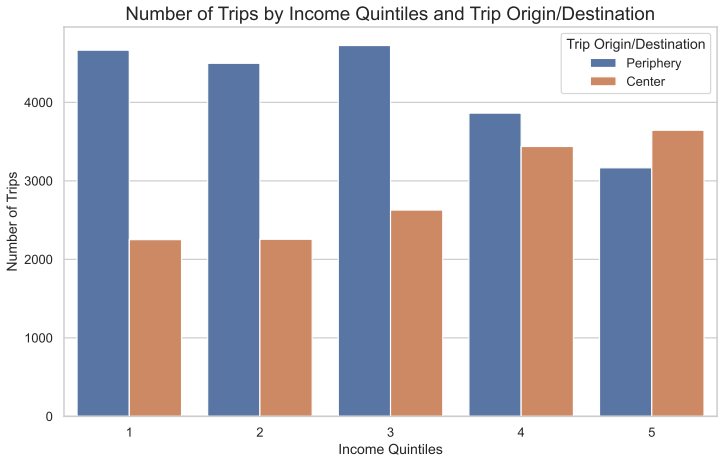


(b) Number of trips by income group and origin-destination: center or periphery. *Notes:* Every income group has two columns the left is trips in the periphery and the right column are trips that either start or end in the city center.

Figure A7: Transport shares and number of trips (commutes) by income and center/periphery dimension. *Sample:* These are commutes that both start and end in the Madrid metro area for individuals who reside in the metro area.

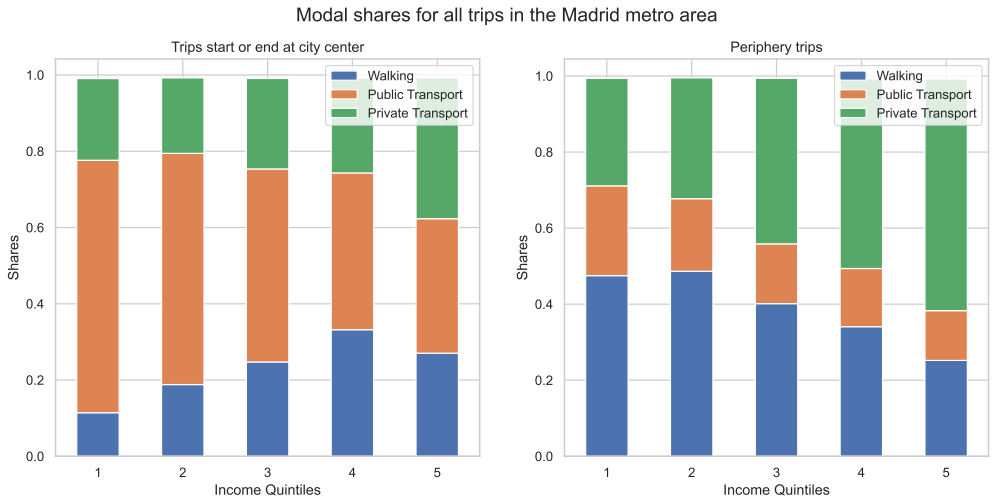


(a) Modal shares by income group and origin-destination: center or periphery. *Notes:* The left panel shows the shares for trips that either start or end in the city center, by income quintile. The right panel shows the shares trips for all other locations, denoted as periphery trips, by income quintile.

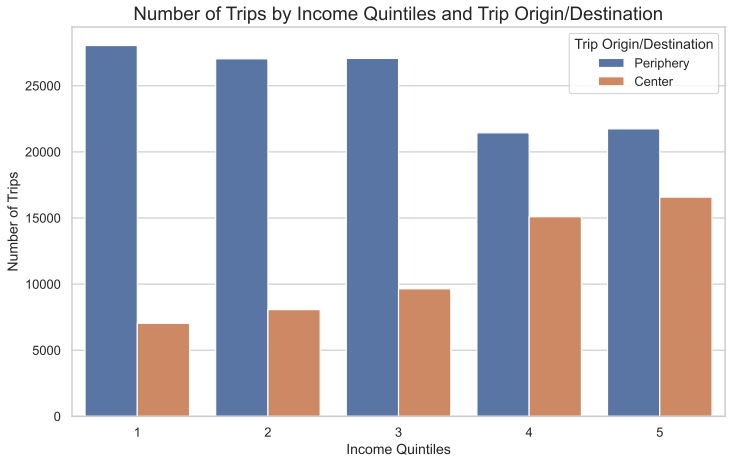


(b) Number of trips by income group and origin-destination: center or periphery. *Notes:* Every income group has two columns the left is trips in the periphery and the right column are trips that either start or end in the city center.

Figure A8: Transport shares and number of trips (stable-commutes) by income and center/periphery dimension. *Sample:* These are commutes that both start and end in the Madrid metro area for individuals who reside in the metro area. The individuals are aged from 30 to 60 years and are registered to live in Madrid.



(a) Modal shares by income group and origin-destination: center or periphery. *Notes:* The left panel shows the shares for trips that either start or end in the city center, by income quintile. The right panel shows the shares trips for all other locations, denoted as periphery trips, by income quintile.



(b) Number of trips by income group and origin-destination: center or periphery. *Notes:* Every income group has two columns the left is trips in the periphery and the right column are trips that either start or end in the city center.

Figure A9: Transport shares and number of trips (all trips) by income and center/periphery dimension. *Sample:* These are all trips that both start and end in the Madrid metro area for individuals who reside in the metro area.

A.2.4 Duration analysis

This contains a bootstrap of the raw duration and the income group effects on public transport durations.

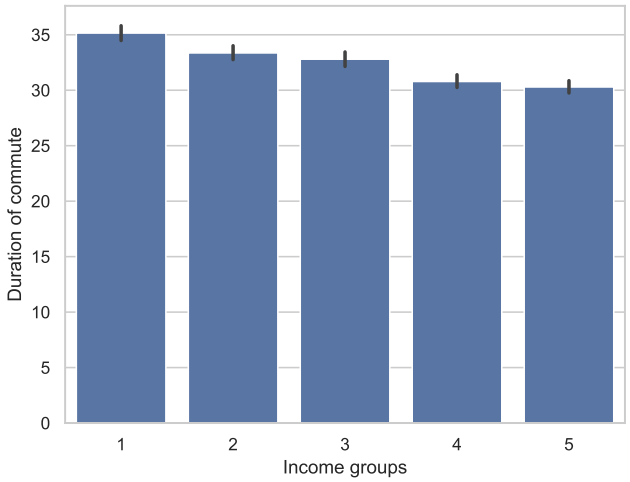


Figure A10: Differences in durations of commutes across income groups. *Notes:* The errors are bootstrapped at the individual level. *Sample:* Individuals who live in the metro area and commutes that start and end in the metro area, the income groups were computed at the household level for households living in the metro area.

Dependent Variable:	Public transport duration		
Income group	exp(coef)	confidence (95)	observations
1	-	-	4176
2	1.01	[0.94-1.09]	3701
3	1.02	[0.95-1.09]	3199
4	1.03	[0.95-1.12]	3043
5	0.94	[0.83-1.06]	2400
Controls	YES		

Table 5: Effects of income on public transport durations.*Notes:* These results are the coefficients from a proportional hazards duration model, higher coefficients (larger than 1) relative to the baseline, means shorter commutes. The controls include origin-destination dummies, straight line distance as well as demographic characteristics. *Sample:*These trips are commutes that both start and end within the Madrid metro area from individuals who reside in the Madrid metro area and use public transport.

A.3 Estimation of joint choice model of residential location and mode of transport

This section follows [Akbar, 2022b]. My intention was to estimate the demand side prior to building the general equilibrium (housing market), I couldn't carry it out in the end, mainly due to data limitations, but I outline the procedure here.

Summarizing, it is a logit model where the choice is the mode and residential location[20], there are also [Berry et al., 1995] style fixed effects. Following the original notation from [Berry et al., 1995], the job locations are the product markets and the products are the (residence x modal choice) given a job location. The estimation by maximum likelihood is not straightforward due to the high number of dummies to estimate.

[Akbar, 2022b]'s approach: N residential neighborhoods, J work locations, M modes of travel. 1 heterogeneity Y: 4 Income groups. Each worker is exogenously assigned a work location j and an income w_i The agent maximizes over residential choice and travel mode given heterogeneous preferences over where to live.

$$U_{mn|ijy} \equiv \alpha^S_{my} S_{jmn} - \alpha^D_y D_{jn} + \frac{w_i^{1-\alpha_w}}{1-\alpha_w} - \frac{p_n^{1+\alpha_h}}{1+\alpha_h} + \delta_{mny} + \epsilon_{imny}$$

Most important things in this utility specification:

- Decomposition of travel time into speed and distance to avoid some biases. Alpha changes across income groups and across modes of transport, this is in line with the literature. The alpha of distance reflects different distributions of jobs in the city for different income groups.
- Price is taken as exogenous for each worker.
- The δ_{mny} are mode x residential location x income group fixed effects, they follow the spirit of [Berry et al., 1995] but they are a bit different due to the inclusion of the income dimension.
- The housing demand is assumed Cobb-Douglas and depends on price of housing and income.

After assuming ϵ_{imny} is i.i.d Type 1 for each individual, we get the nice closed form solution to the probabilities of each choice for different individuals[21]:

$$\pi_{mn|jy} = \frac{exp(V_{mn|jy})}{\sum_{m' \in M} \sum_{n'} exp(V_{m'n'|jy})}$$

where

$$V_{mn|jy} \equiv \alpha^S_{my} S_{jmn} - \alpha^D_y D_{jn} - \frac{p_n^{1+\alpha_h}}{1+\alpha_h} + \delta_{mny} + \epsilon_{imny}$$

[20] Then, location x and commuting by public transport or location x and commuting by car are different choices. The choice set explodes pretty quickly if the locations are geographically small.

[21] The w_i ends up in the error term or absorbed by the fixed effect(?) or there is a typo in the paper or an unexplained assumption. This way we go from individual to aggregate estimation. This follows from the utility specification where individuals do not differ amongst anything other than income.

Then these probabilities can be estimated via MLE with the aggregate shares of each income group that work and live in a particular location.

An initial reason for why one would want to model residential and housing location jointly, is that people who like public/private transport may locate closer to public transport stations/highways, making travel time endogenous and the estimates of value of time biased, however the joint estimation is also complicated. There are a few issues with identification Akbar discusses in the paper[22], I found that the potential sources of endogeneity coming from both the modal choice and residential choice, to be headache inducing but very interesting as well.

From a purely econometric perspective there are also issues. Depending on where you put the dummies you can have issues if some shares are 0, if there is a zero share and a dummy for that choice the coefficient for the dummy will go to infinity and the estimation will fail, you can pretend that then the 0 share choice is not an actual choice but I don't know how this would affect the other estimated coefficients. Also, depending at which level you put the dummies, some important coefficients like price will not be identified directly and you will need an instrumental variable. Complicated.

Chapter 13 of [Train, 2009] provides a very nice explanation of the [Berry et al., 1995] procedure as well as of other ways to deal with endogeneity in discrete choice settings. [Barwick et al., 2021] implement a similar, more complicated but probably more robust, procedure in the same setting of joint modal and residential choice.

A.4 Counterfactual travel times: TravelTime API

To complement the survey data and do modal choice analysis; see the second chapter of [Train, 2009] for the basics, I attempted to download counterfactual travel times using the TravelTime API. While the travel time download was not fully successful, as there were significant differences between the downloaded times and the reported times[23], I showed it can be done.This section outlines the procedure I followed[24] and where I think I went wrong. [25]

The first step, and where I think I made my mistake, is to define the geographical start and end positions. The survey does not provide the exact geographical origin and destination for each trip, instead it provides an area. The finer the area, the better the approximation, however not all the geographical areas are created equal. Towards the city center, all the areas are small and accessible. In those small areas the centroid of the area is a good approximation for where the trips going to that area are going. However that is not the

[22]While I have not read the key transport literature I do not think they are well explained and the estimation and assumptions are a bit obscure. However this working paper provides a simple concept of what needs to be estimated in these models of joint residential and housing location

[23]I downloaded both public transport and private transport times as a failsafe to compare the download to the reported times.

[24]My initial intention was to download counterfactual travel times for the observed trips. Moreover, for an analysis where households can choose their residence one could modify the procedure to download the full matrix of travel times. This would almost be easier, as it is what the API is designed for.

[25]I assume that it was me that did something wrong and that the the Traveltime API model works well.

case for all the areas, in areas towards the periphery or in the rural areas of Madrid, these areas become bigger, as they are less populated, and going to the centroid becomes a worse approximation. A relevant example would be the airport. The airport is a relatively big area and its centroid is in the middle of runway 6, that is unlikely to be the destination of people going to work there. For these extreme cases the Travel time API can't find a time and the code breaks, which is good, as it makes you realise you are making a mistake. However for less extreme cases, like a transport zone where the buildings are skewed to a side, this can be worse. In those cases the code will not break but it will get to the buildings by car and then take a 30 minute hike, which is a big issue.

I realised the importance of accurately defining the start and endpoint a bit too late, however I did fix this issues for the most extreme cases. I did this by using the amenities from OpenStreetMap within a transport area. If there is a restaurant or a gas station it should be reachable by car, and by a combination of public transport and walking. The ideal would be to pick a cluster of amenities, as that is the place with the highest likelihood, I just picked a random one due to time constraints. Even if I fixed the most extreme cases, it is likely that there are a multitude of intermediate cases where the centroid is not providing a good enough approximation but not being bad enough so that the program breaks. Fixing that, either by using the amenity alternative for all areas or some other approach, should be a top priority for any further download. [26]

A.5 Rental prices

Another key ingredient for the model is the price of housing/land, a possible measure of this is this map built by the Spanish housing ministry, it contains quantiles of the distribution of prices of rental units for 2021-2022 at different geographic levels. There are prices for apartments and for single family homes, both the full price and the price per square meter. While the median housing price at a census tract level is an imperfect measure of a housing price, you can assume symmetry and the median becomes the mean[27]. The data also provides the number of of units rented on a given area, then, you can use that to weigh census tracts when moving across different geographies. This was my procedure to construct a measure of housing prices: take the median at a census tract level, assume symmetry, now the median is the mean, then do a weighted average with all other census tracts that belong to the bigger geography using the number of listings as weights. All legal rental listings should appear on the data so the data is as representative as it can possibly be, however housing transaction data would also be very interesting, possibly a superior alternative, [Barwick et al., 2021] have mortgage data for Beijing, this data could be very difficult to obtain for Madrid.

[26] Beyond this, the API has full free access for a couple of weeks and the rest is just a matter of optimizing the download speed, and building an algorithm to keep it running. This is not trivial but it is feasible. I provide a (not so good) Python code that implemented this in my replication package. The public transport times are not available for 2018 but they shouldn't be have changed much since then. For any further details you can check their documentation.

[27] Alternatively, since a few quantiles are given, you could build some sort of empirical distribution with some additional assumptions.