Study of characteristics of Medellín City’s areas based on travel habits of different socio-economic classes

Final Report

Team:
Andrea Matteazzi (2010655) Pietro Soldà (2027653) Matteo Visonà (2006675)

Introduction

In this project we analyzed the mobility graph of a Colombian city in order to understand the differences and similarities in travel habits of various socio-economics classes. To reach our results, we extracted different nodes’ and graph’s features from the graphs obtained considering only the trips of a certain class. Then, we tried to explain the obtained measures by doing some consideration about the real world scenario.

In the final part of the project we also considered the temporal information we had about each trip to draw some conclusions about travel habits regarding duration of the trips and “time of the day” distribution of the trips.

In the first section of the report we explain the dataset we used and the elaboration we did on it to obtain the graphs used during the experiments. In the second section we explain the various experiments and methods that we used to extract the data from the graphs. In the last section we analyze the obtained measures and we give a tabular or graphic representation for some of them, expressing also which conclusions can be derived from them.

Dataset

The dataset we used is the “Colombian city inter-zone mobility (2005) - Medellin” dataset, that was obtained from the interviews of people of different socio-economic groups which reported the origin and the destination of the trips they usually take during the day.

The dataset can be found in [1].

It represents a directed multiplex multigraph composed of 413 nodes and 128635 edges. Each node represents an urban zone of Medellín city (Colombia), whereas each edge represents a travel between two urban zones. There are 6 layers in this network, each one representing a different socioeconomic class (where 1 corresponds to the poorest class and 6 corresponds to the richest class). Each edge is also labeled with some additional information such as departure and arrival times, mean travel time and the purpose of each trip (work, school, entertainment etc).

Starting from the original dataset with 7 features for each edge/interview (starting node, arriving node, social class of the interviewed, departure time, arrival time, motivation of the trip and expansion factor of that particular interview), we’ve added a “time” feature for each interview, which consists of the duration of the travel, computed by using the data about departure and arrival time; this feature is useful since it can give some approximate information about the real distance between two city zones. The expansion factors are used to expand the individual responses up to an estimate for the entire population.

Then we implemented a function to create the graphs associated with the different economic classes.

In particular, for each class we wanted to obtain a graph with at most one edge between each pair of nodes, so we merged together each edge with the same origin and destination. Every time we merged two or more edges, we did some operation on the edges' features in order to preserve some of the useful information of the various merged edges: we summed up the expansion factor associated with these edges and then we divided this partial sum with the sum of all the expansion factors of the edges in that class (so to obtain a normalized expansion factor for each edge) and we called this features 'weight', we computed also an unnormalized version of this feature, in order to perform some numerical comparison about the density of the mobility graphs of the various classes, we called this feature "weight_nn" that stands for "weight not normalized"; then we computed the mean of the "time" feature of the merged edges; we also kept count of how many edges was merged. In this way, at the end we obtained a graph for each one of the 6 economic classes, where every edge has an associated weight (i.e the normalized expansion factor) representing the fraction of people of that class that cover that journey, and also a "time" and a "count" feature.

We decided to normalize the expansion factor of each edge because the number of people belonging to each social class vary a lot between different classes and by using the normalized weights we can obtain metrics that can be compared between different classes. In particular, the "weight" of an edge can be interpreted as the fraction of people in the given class, who traverse such an edge.

Experiments

Since the graphs of the various classes were of reasonable dimension, we were able to use the exact implementation of the methods available in the NetworkX library to compute the analytics which presented a dedicated method. [2]

In-degree, out-degree and total degree We used the dedicated methods to compute the in-degree, out-degree and total degree for each node in the graph, and we repeated for each of the 6 class-associated graphs. The in-degree is useful because it allows us to understand for each social class which are the most targeted nodes/urban zones. The out-degree allows us to understand for each social class which are the places from where people spread the most across the rest of the city. The original dataset could have multiple edges between the same couple of nodes (each edge was corresponding to a single trip that differs from the others for the starting hour and/or ending hour and/or motivation of the trip), we collapsed all these edges in a single edge taking the normalized sum of their FEV (expansion factor). In the calculation of the degrees we consider only those single edges. Therefore these statistics show the nodes that are more 'connected' to the others (not in the sense of number of trips but number of other locations).

Strength of nodes Similar to the previous point we computed the degree of all the nodes, but this time we considered the normalized (or unnormalized, we computed 2 versions of the node's strengths) expansion factor as weight for each edge, so from this we obtain an estimation of the fraction of the population of a certain class that passes through a node during the day. From the in-strength metric we can derive which is the node visited by the highest number of people for each class, while from the out-strength we can find the urban zones from which we have a high number of people of that class departing, and consequently where the population of a certain class tends to aggregate. The unnormalized version is useful to compare numerically the density of movements across the various mobility networks given by the social classes.

Closeness centralities We calculated several versions of the closeness centrality.
The first one is without considering weights in the edges, so the standard version of the closeness centrality to find how much each zone is "connected" to the rest of the graph.
The second version is made by considering weighted edges, in particular considering as weights the duration of each trip (the "time" feature associated with each edge/trip in the graph). We used the duration because it is the closest measure to the actual distance between nodes that we have, therefore by considering also the travel time we might discover with a higher precision which are the most 'central/closer' nodes.
The third and last version of closeness centrality considers as edges' weights the inverse of the FEV (expansion factor). It is important to use the inverse of the FEV instead of the FEV because in the closeness centrality the edges with the lowest weight are considered more convenient during the computation of the shortest paths (because the weight is interpreted as distance, so smaller weight means shortest distance) so we wanted to associate the most common trips (those with highest FEV) with shortest edges, in order to central zones with respect to traffic information.

Betweenness centrality Similarly to what we did with the closeness centrality, we considered the unweighted version of the betweenness centrality and also other two weighted versions: one with edges' weight equal to the travel time and one with edges' weight equal to the inverse of the normalized expansion factor.

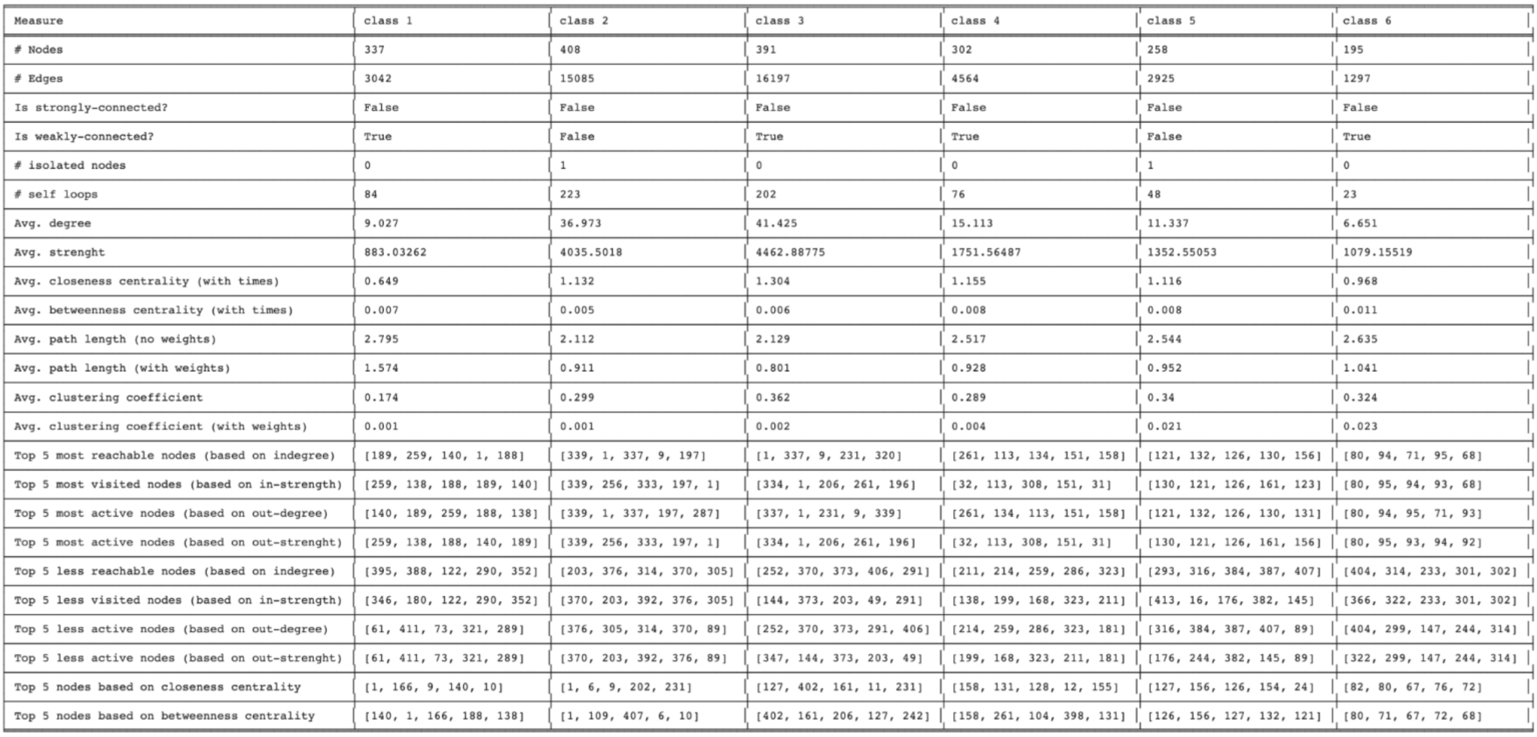
Average shortest path length We used the method for computing the shortest path between two nodes, and we applied it to every pair of nodes in the graph of each class. We summed up the length of all the paths and then we divided them by the number of possible node pairs, so as to obtain the average length of a shortest path for the graph of that class.
Also for this metric we have considered two versions. The first version does not consider weights (i.e. each edge has unitary weight). The second version instead considers weights. As weights we used the average travel time of all trips among the two nodes that the edge connects. Again in this case we have used this kind of weight in order to have an estimate proportional to the actual distance between two nodes. This second version is the slowest part of the code, taking about 30 minutes to complete. The execution time however wasn't so high to justify the implementation of an approximated computation of the average path length.

Clustering coefficient and Average clustering coefficient Using the dedicated method we computed the clustering coefficient for each node in the six graphs. We used the weighted version (normalized FEV) of the clustering coefficient. In such metric, each triangle is not longer counted as 1 but as the geometric mean of the weights in the corresponding edges, so that the more people traverse those edges, the more important is the triangle. We decided to use the weighted version of the clustering coefficient to obtain a more fairly comparable metric, since the middle classes include more people and thus the node of these classes are more likely to have high unweighted clustering coefficient. We also computed the average clustering coefficient across all the nodes of a graph, so as to obtain a mean value for each class. This is useful to determine how much a class tends to move around in the same zones and which class instead tends to spread more across the whole city.

Temporal analysis on data We divided the day into 48 half-hour time slots. For each time slot and for each class, we calculated how many trips were "active" during that time slot. A trip is considered active during a time slot if the departure time is before the end of the time slot and if the arrival time is subsequent to the start of the time slot . After finding the active trips, we summed up their normalized expansion factor. IN this way we can estimate the fraction of the population of that class that is traveling during that time slot.
We also calculated the total normalized sum of all FEVs of all journeys of all social classes in each time slot, so as to obtain a general distribution of trips, without considering the distinction between different socio-economic groups.

Length of trip analysis We wanted to understand if there was some significant difference between the length of the trips between the various economic classes. To do so, we created a distribution of the trips over all the possible trips' durations. That is, we divided all the trips into different time-classes based on their duration; each time-class covered a 15 minutes interval. Then for each time-class we summed the normalized expansion factor of all the trips in it, obtaining an estimation of the fraction of people in that economic class that performs a trip of that duration during the day.
We also computed the average duration of a trip for every economic class, to have a more immediate estimate of the duration of the trips in that class.

Results and Considerations



Measure	class 1	class 2	class 3	class 4	class 5	class 6
# Nodes	337	408	391	302	258	195
# Edges	3042	15085	16197	4564	2925	1297
Is strongly-connected?	False	False	False	False	False	False
Is weakly-connected?	True	False	True	True	False	True
# isolated nodes	0	1	0	0	1	0
# self loops	84	223	202	76	48	23
Avg. degree	9.027	36.973	41.425	15.113	11.337	6.651
Avg. strenght	883.03262	4035.5018	4462.88775	1751.56487	1352.55053	1079.15519
Avg. closeness centrality (with times)	0.649	1.132	1.304	1.155	1.116	0.968
Avg. betweenness centrality (with times)	0.007	0.005	0.006	0.008	0.008	0.011
Avg. path length (no weights)	2.795	2.112	2.129	2.517	2.544	2.635
Avg. path length (with weights)	1.574	0.911	0.801	0.928	0.952	1.041
Avg. clustering coefficient	0.174	0.299	0.362	0.289	0.34	0.324
Avg. clustering coefficient (with weights)	0.001	0.001	0.002	0.004	0.021	0.023
Top 5 most reachable nodes (based on indegree)	[189, 259, 140, 1, 188]	[339, 1, 337, 9, 197]	[1, 337, 9, 231, 320]	[261, 113, 134, 151, 158]	[121, 132, 126, 130, 156]	[80, 94, 71, 95, 68]
Top 5 most visited nodes (based on in-strength)	[259, 138, 188, 189, 140]	[339, 256, 333, 197, 1]	[334, 1, 206, 261, 196]	[32, 113, 308, 151, 31]	[130, 121, 126, 161, 123]	[80, 95, 94, 93, 68]
Top 5 most active nodes (based on out-degree)	[140, 189, 259, 188, 138]	[339, 1, 337, 197, 287]	[337, 1, 231, 9, 339]	[261, 134, 113, 151, 158]	[121, 132, 126, 130, 131]	[80, 94, 95, 71, 93]
Top 5 most active nodes (based on out-strenght)	[259, 138, 188, 140, 189]	[339, 256, 333, 197, 1]	[334, 1, 206, 261, 196]	[32, 113, 308, 151, 31]	[130, 121, 126, 161, 156]	[80, 95, 93, 94, 92]
Top 5 less reachable nodes (based on indegree)	[395, 388, 122, 290, 352]	[203, 376, 314, 370, 305]	[252, 370, 373, 406, 291]	[211, 214, 259, 286, 323]	[293, 316, 384, 387, 407]	[404, 314, 233, 301, 302]
Top 5 less visited nodes (based on in-strength)	[346, 180, 122, 290, 352]	[370, 203, 392, 376, 305]	[144, 373, 203, 49, 291]	[138, 199, 168, 323, 211]	[413, 16, 176, 382, 145]	[366, 322, 233, 301, 302]
Top 5 less active nodes (based on out-degree)	[61, 411, 73, 321, 289]	[376, 305, 314, 370, 89]	[252, 370, 373, 291, 406]	[214, 259, 286, 323, 181]	[316, 384, 387, 407, 89]	[404, 299, 147, 244, 314]
Top 5 less active nodes (based on out-strenght)	[61, 411, 73, 321, 289]	[370, 203, 392, 376, 89]	[347, 144, 373, 203, 49]	[199, 168, 323, 211, 181]	[176, 244, 382, 145, 89]	[322, 299, 147, 244, 314]
Top 5 nodes based on closeness centrality	[1, 166, 9, 140, 10]	[1, 6, 9, 202, 231]	[127, 402, 161, 11, 231]	[158, 131, 128, 12, 155]	[127, 156, 126, 154, 24]	[82, 80, 67, 76, 72]
Top 5 nodes based on betweenness centrality	[140, 1, 166, 188, 138]	[1, 109, 407, 6, 10]	[402, 161, 206, 127, 242]	[158, 261, 104, 398, 131]	[126, 156, 127, 132, 121]	[80, 71, 67, 72, 68]

We constructed the table above by regrouping all the metrics we've discussed before, and also considering the top five most and the top five less active, reachable or visited nodes according to those metrics.
From that table is clear the difference between the mobility networks of the various social classes. It is interesting to see that the networks of the classes 2 and 3 are the most densely connected and the most active as shown by the higher number of edges, values of the average degree, and the values of their average strength. We were expecting these results since the classes 2 and 3 correspond to the middle-income population, hence the majority of the population.
Another interesting result is the difference among the values of the average path lengths (in both the weighted and unweighted versions). These values are smaller for classes 2 and 3 in comparison to the other mobility networks of the other classes.

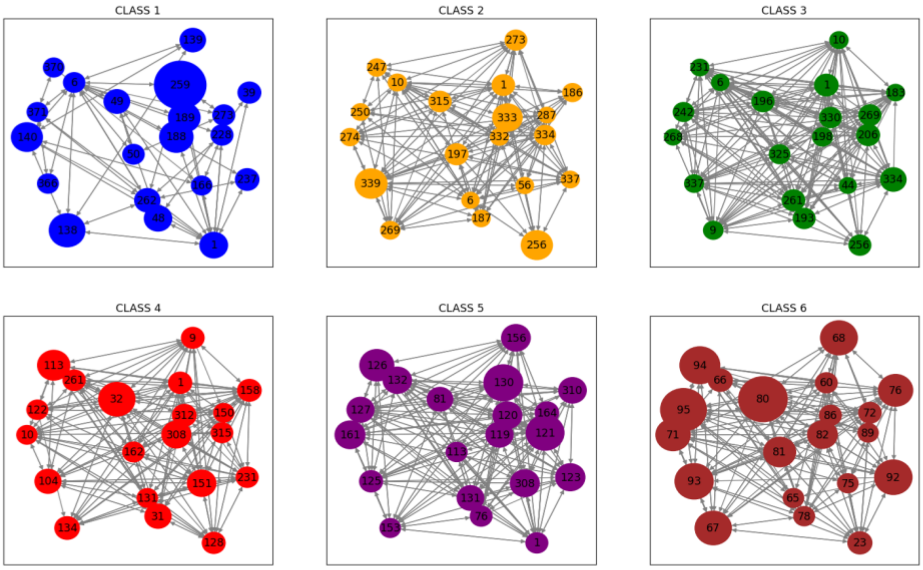
However the most relevant results are the differences in values between the poorest social class (class 1) and the richest social class (class 6). In particular, class 1 shows a larger value for the average path length, higher strength and graph clustering coefficients. These differences are significant because they allow us to understand that trips of poor peoples in the city of Medellin tend to be dispersed across the city whereas the trips of rich peoples tend to be more localized, redundant and near to the starting point.

These results could give some insights about the behavior of people belonging to different social classes, and even about the infrastructures and the road connections of different areas of the city.
For instance, the mobility network of the poorest social class shows a relatively low redundancy, this could be due to poor wiring in these districts, perhaps due to less road building works in these areas, the higher average path length instead could be explained by the fact that poor districts could potentially have fewer commercial activities or public infrastructures such as hospitals and schools hence people belonging to this category needs to travel longer distances to accommodate their needs. Also this lack of commercial activity could lead workers in this social class to travel longer distances to reach their working place.

On the other side, richer people may live in more industrialized districts that have all these infrastructures, therefore the spatial pattern of movements become more localized since they don't need to move much to reach these infrastructures or services. Moreover, there could be a prevalence of office jobs in this social class, that sometimes could be done remotely in smart working, and they could be less prone to move far away to reach their working place.

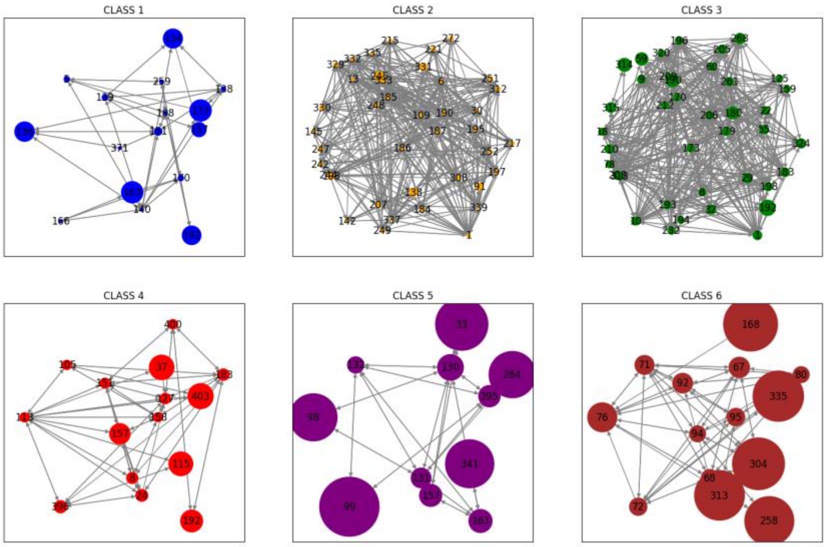
Middle-wealth state social classes instead show multimodal mobility as they cover most of the areas of the city and display dense mobility networks as confirmed by the higher average degree, average strength and number of edges.

TOP 20 STRENGTH NODES PER CLASS



The above Figure shows the 20 nodes with higher strength for each class. The size of each node is proportional to the actual value of such a metric. It is possible to see that classes 1 and 6 have bigger nodes compared to e.g. class 3 and this reflects the fact that the poorest and richest classes tend to go respectively in the same zones in contrast to the most populated 3, where people are spread along the city. Furthermore, node 1 is present in almost all the figures and then, it is reasonable to assume it as a central zone.

TOP 5 CLUSTERING COEFFICIENT NODES (WITH NEIGHBORS) PER CLASS

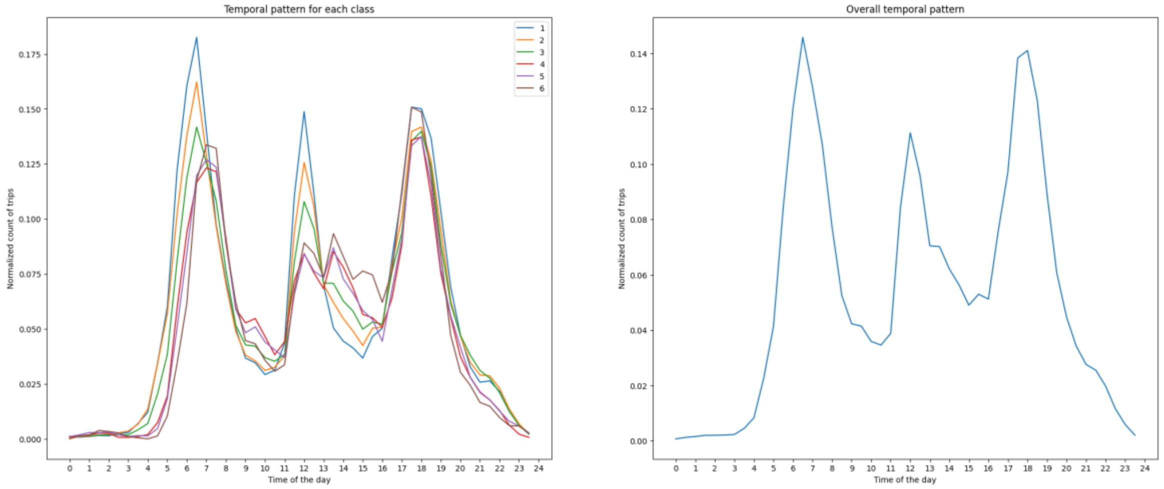


The above Figure shows the 5 nodes with higher clustering coefficient (and the respective neighbors) for each class. The size of each node is proportional to the actual value of such a metric. From here, it is evident how classes 2 and 3 are highly distributed across the city. Indeed, the size of their nodes is very small and comprises a high number of neighbors. The richest classes 5 and 6 instead, have bigger nodes and this translates into being highly dense in particular zones.

Temporal analysis results

The two figures below show the results obtained from the analysis of the distribution of the trips during the time of the day. The y-axis tells the fraction of active trips with respect to the total number of trips performed in a day.

In the left panel we show the different distribution for each economic class, while in the right panel we reported the distribution for the whole city, without distinction between classes.



The first thing to observe is the presence, for all classes, of three time windows with a spike of activity, corresponding to the main period of traffic activity in a day: the morning (from 6 to 8), the midday (from 12 to 14) and the evening (from 17 to 20). These peaks are easily explainable considering the normal working or studying routine of most people.

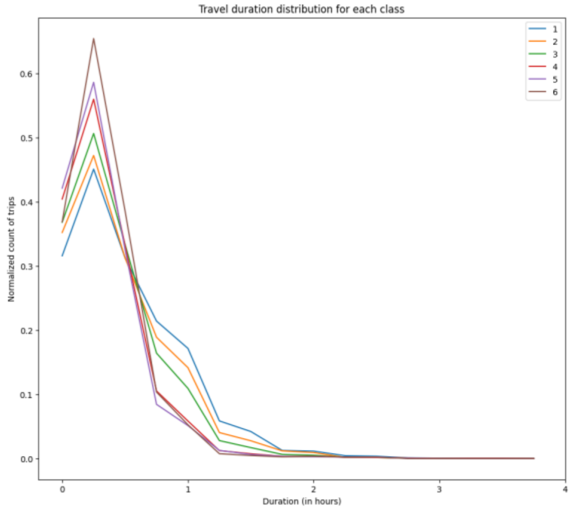
However it is interesting to observe that there are some significant differences that happen following one-by-one the wealth ordering going from class 1, the less wealthy, to 6 the most wealthy.

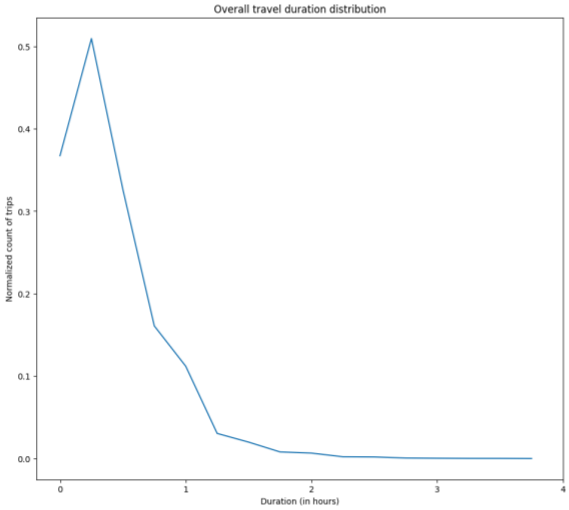
In particular, going from one class to the next in wealth order, we can see an almost constant shift to the right in the morning peak of movements. This means that the richer classes tend to start moving later in the morning, with a shift of more than one hour between class 1 and 6. Such behavior can be explained by saying that the wealthier classes start working later, or simply by considering the fact that wealthier people are more likely to live near the center and the most active part of the city, so they can get to work quicker than someone who lives in the suburb (see also “length of trip analysis” below).

Considering the peak during the mid part of the day, there is a big change in the shape of the curve: the activity for the wealthier groups is spread across the whole afternoon, probably reflecting the possibility of moving around more for other reasons besides work and study.

The intense activity in the evening instead is very consistent for all classes, but still presents a small shift, this time in inverse order with respect to the one observed in the morning, implying that the poorest groups get off work later or it takes them more time to get home.

Length of trip analysis results





```
Class 1 - mean travel duration across all trips: 0.567934638694639

Class 2 - mean travel duration across all trips: 0.5208814961575965

Class 3 - mean travel duration across all trips: 0.48580818459141645

Class 4 - mean travel duration across all trips: 0.43717015114288327

Class 5 - mean travel duration across all trips: 0.40400534943411653

Class 6 - mean travel duration across all trips: 0.41708781030012276
```

The figures above show the travel duration distribution for each class (left panel) and for the overall population (right panel). The y-axis tells the fraction of trips (within a class for the left panel) that have the duration specified by the x-axis. Below the left panel we also reported the average duration of a trip, in hours, for each class.
We can see that the trips are mostly less than 1 hour long and the distributions present a peak at 0.5 hour. This data can be useful to give an idea about the dimension of the city.
Regarding the differences between socio-economic groups, it is interesting to observe that there is a clear tendency of the distribution to get more right-skewed with the decreasing of wealth, from 6 to 1, meaning that the trips tend to get longer. This is reflected also in the mean durations: the poorer a class is, the higher the mean duration of the trips, with the exception of class 5 and 6 that, however, present similar values.
These observations were expected since the poorer classes live more likely outside the city center and are also more likely to use slower public transportation to move across the city.

Contribution of each member

We tried to work together as much as possible, so each member has given the same contribution to the project. Also the code was written while working in group, so most of the parts are the result of the contribution of more members.

References:

[1] https://datadryad.org/stash/dataset/doi:10.5061/dryad.hj1t4
[2] https://networkx.org/documentation/stable/reference/index.html