Politecnico di Torino

ICT for Smart Societies

Lab Report for:
ICT for Transport Systems

Professor:
Prof. Mellia Marco

Authors:
Matteo Nisi 238188
Navid Yamini 235935
Stefano Calleris 243892

February 2018

Step 1 – Preliminary data analysis

In the first step we were asked to make some preliminary analysis on the MongoDB database that was provided to us containing data collected from carsharing system divided in four different collections: ActiveBookings, ActiveParkings, PermanentBookings and PermanentParkings.
Please note that these documents are updated in real time and the results that we are going to show refer to January 28th, 2018.

- **How many documents are present in each collection?**

ActiveBookings	4583
ActiveParkings	8929
PermanentBookings	27944897
PermanentParkings	28072946

- **Why the number of documents in PermanentParkings and PermanentBooking is similar?**
  The number of documents in PermanentParkings and PermanentBooking are similar because the number of cars is constant, and we expect that whoever has taken a car also has parked it. Nevertheless, numbers are not the same because of some missing values and anomalies in the system.

- **For which cities the system is collecting data?**
  By taking query from the ActiveBooking collection we will see that the system is collecting data for 24 cities:

  ["Amsterdam", "Austin", "Berlin", "Calgary", "Columbus", "Denver", "Firenze", "Frankfurt", "Hamburg", "Madrid", "Milano", "Montreal", "München", "New York City", "Portland", "Rhineland", "Roma", "Seattle", "Stuttgart", "Torino", "Toronto", "Vancouver", "Washington DC", "Wien"].
  But if we take the query on the PermanetBooking we will find two more cities there: Stuttgart and Twin cities.

- **When the collection started? What about the time zone of the timestamps?**
  By running the query on the PermanentParkings we will understand that the collection started at 1481650658 in Unix time which is equal to December 13th ,2016 at 16:37:38 GMT.
  For the time zone we can say that we have two different times in the system:
  Init_time: it is a server time in UNIX which system uses to save the data on server.
  Init_date: it refers to the local time zone of the city that the data is come from. It is different for each city.

In the following part we need to consider only cities that were assigned to our group, which are: Turin, Madrid and New York city.

- **How many cars are available in each city?**
  To solve this task, we decided to sum cars in Active Bookings and Active Parkings. To have only the unique car and not a possible repetition due to possible anomalies in the DB, we projected the plates and took only unique values.

Madrid	436
New York City	550
Turin	412

- **How many bookings have been recorded on the October 1st, 2017 in each city?**

Madrid	6217
New York City	3554
Turin	3163

- **How many bookings have also the alternative transportation means recorded in each city**?

  We ran the query for all cities that were assigned to us. As result we found that only Turin city has this option and the result is 295057 objects have walking duration and 261214 have public transportation system as the alternative.

Step 2 – Analysis of the data

Task 1

In the first task, we consider the three assigned cities and we evaluate the bookings and parking for every hour of a defined period of time (September 2017). Please note that we filtered the outliers to reach to these results. A better explanation of the filter will be given in task 3.

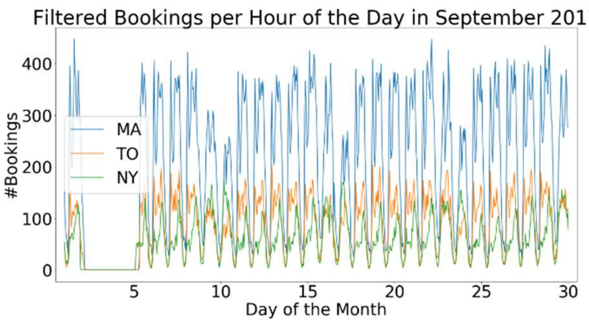


Figure 1 – Filtered Bookings per Hour of the Day

In Fig.1 it can be seen how the system was not working for three days. Due to the fact that we have the same period of time for all the three cities and presuming that at least for America and Europe Car2Go runs on different servers, we think that it is a possible problem in the acquisition data in the database.

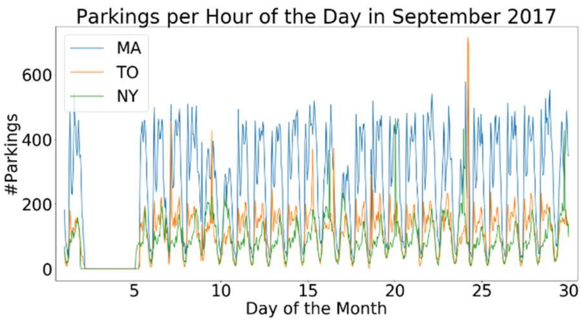


Figure 2 – Filtered car park per Hour of the Day in September 2017

Both in figure 1 and 2 it can be noticed how the green shape is shifted by some hours with respect to the two others.
That is because every computation is done on 720 samples (24 samples per 30 days) starting from midnight of the city local time so the shift seen represents the different time zone of New York City.

Since the three cities seem to have a common regular behavior (with different amplitude), we focused on the shape of the analysis, aggregating rentals per hour of the day (0-24).

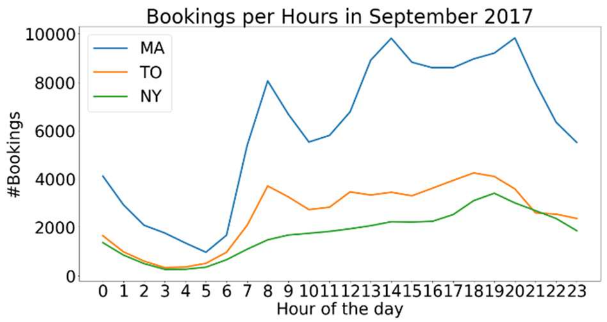


Figure 3

In Figure 3 then, it is shown how the usage has common shape for all the cities with peaks in commuting hours. Taking in account Torino, of which we know better the time schedule of a typical working day, it can be seen how the peaks are at 8 (bookings from 8:00am to 8:59am) and at 18 (bookings from 6:00pm to 6:59pm). Figure 4 instead, shows the percentage of booked cars during the day. It can be seen how during the night the usage of the cars is smaller and how in general New York city presents a smaller share of used cars during the day.

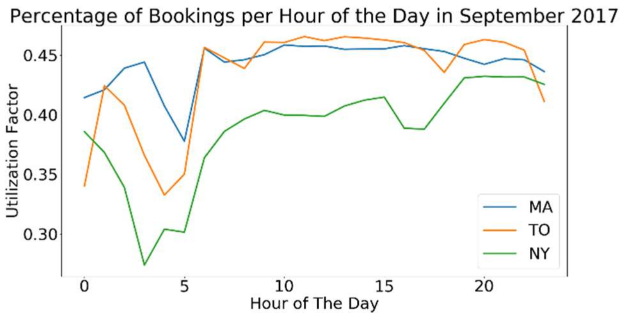


Figure 4 – Percentage of booked cars during the day

Task 2

In task two, the focus was to produce a CDF of bookings and parking for every assigned city.
From the CDF in Fig. 5 it can be seen how bookings are the shortest in NYC and the longest in Torino.

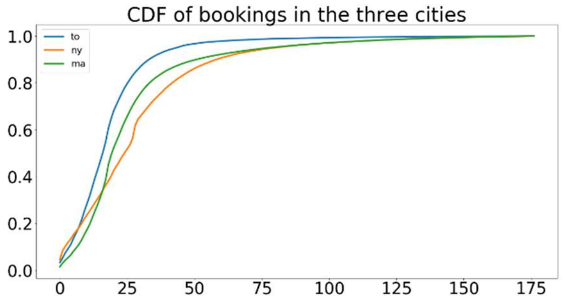


Figure 5 – CDF of Bookings duration in September 2017

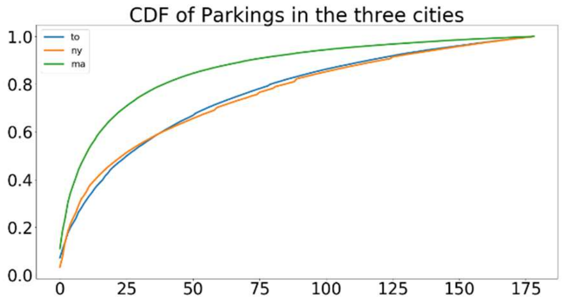


Figure 6 – CDF of car parks duration in September 2017

The CDF on the car parks confirm us what the data in Fig. 1 and 2 were forecasting. In Madrid the usage of the car sharing services seems to be more dynamic. Also, in Fig.3 could be seen how even if the total number of cars in Madrid is slightly bigger than the one in Turin and smaller than NYC, cars are way more rented during the day.
A confirmation to this is given by Fig. 7 in which the percentage of used vehicles is shown. It can be noticed how Torino and New York differently from Madrid present slots with 10% of usage.

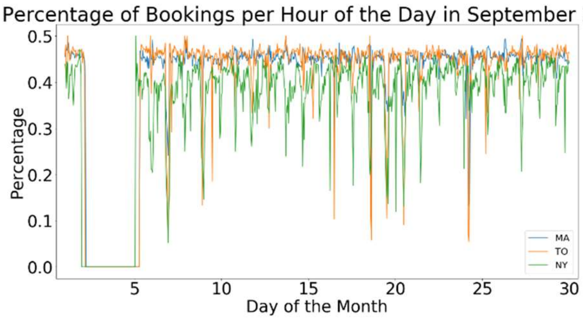


Figure 7 – Utilization factor in the three cities

A further analysis has been done by grouping data weekly (Fig.8) and daily (Fig.9) to see if the shape of the CDF is affected.

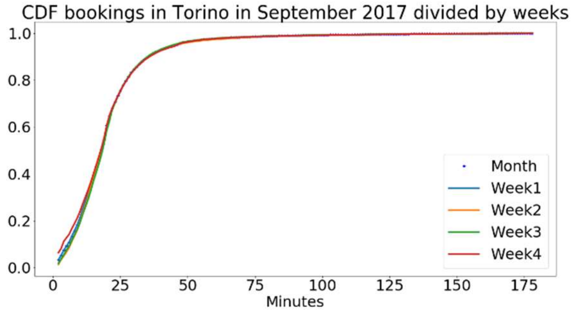


Figure 8 – CDF comparison on 4 weeks (SU-MO)

As it can be easily seen from Fig. 8 the shape of the CDF does not change considering 4 complete weeks of the month.

An interesting output instead is given by the daily division (Figure 9). As it can be seen, the shape of the CDF is changing for Sunday and Monday. If the difference in the shape of Sunday could be expected, the difference in Monday is interesting.

As we know on Monday most of the commercial activity are closed and this can influence the result.
The shortness of the bookings duration in that day could be led by the absence of traffic jam for the weekly closure.

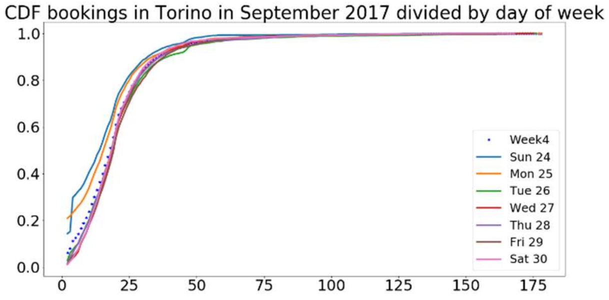


Figure 9 – CDF on different days of a week

Task 3

Since the MongoDB that we were working on it, has some missing values and anomalies, it is required to drive a filtering criterion to filter possible outliers. For booking collection we defined two filters:

- **Not moved**: because of some error in the system and the possibility of the cancelation from the user, we need to make sure that the car has been moved from the original place. To obtain this, we need to compare the origin and the destination of the car by checking the related latitude and longitude. We filtered bookings periods that have the same origin and destination.

- **Too Short/ Too long**: sometimes users change their minds and delete the bookings made few minutes ago. It this case, we decided to ignore the booking periods shorter than two minutes. In the second case, due to errors in the system or car not in the system for maintenance, we can find some bookings with a long period.
To ignore these cases, we filtered out bookings lasting more than 3 hours. We tried to select the good range to make sure that we are covering all possibilities.

For parking collection, we only filtered out those ones that last for a very short amount of time (2 minutes). We did not act a threshold for the maximum parking period because it is possible for cars to be parked and not used for many days.

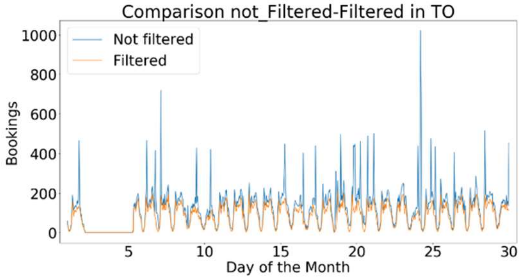


Figure-10

Task 4

In this section, after filtering the data, we calculate the required statistics which are: average, median, standard deviation, and 75th percentiles at a given time (September 2017) for the three assigned cities.

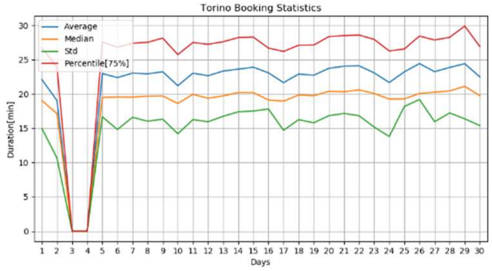


Figure 11 – Booking statistics For Turin

As it can be seen in figures from 11 to 16 we have a significant fall and rise from the 2nd to the 5th of September because of the abovementioned missing values in the collections.

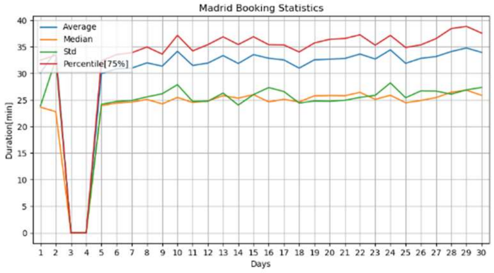


Figure 12 – Booking statistics For Madrid

After those days in the booking collection we have smooth and predictable behaviors during the weekdays and weekends (9th-10th,16th-17th,23rd-24th) but lines graph related to parking duration fluctuate a lot.
By analyzing the graphs, we can find the relation between bookings and parking duration.

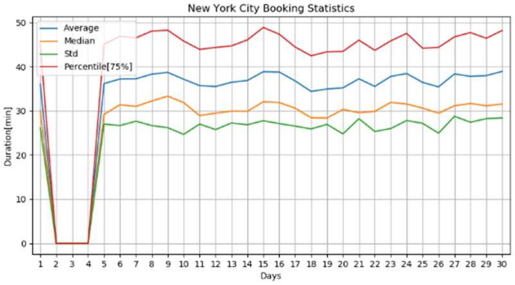


Figure 13 – Booking statistics For New York

Most of the times, as expected, when the duration of rental decreases parking duration increases.
Another outcome is that users have different behaviors in each city. For instance, people in Turin use the carsharing less during Sundays while the usage is mostly the same for the other days; in Madrid carsharing is less used in the last two days of the week; in the end, for New York City parking duration increases after the weekends and this could be translated as a lack of usage of carsharing during the working days.

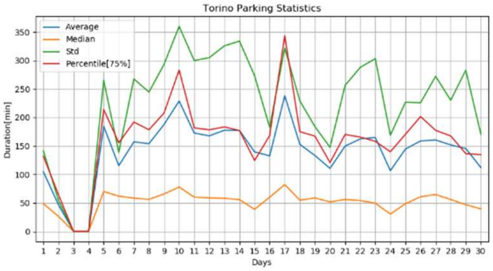


Figure 14 – Parking statistics For Turin

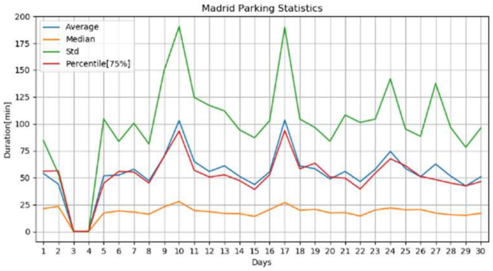


Figure 15 – Parking statistics For Madrid

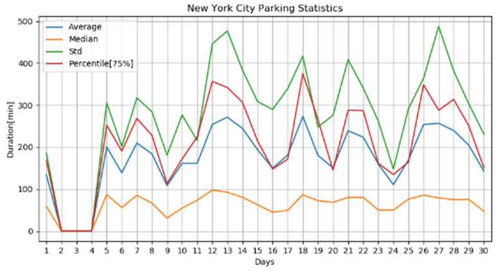


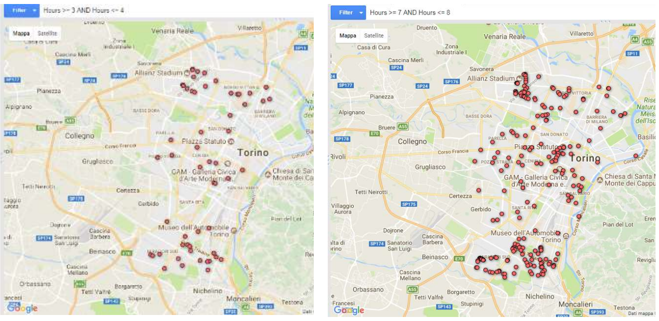
Figure 16 – Parking statistics For New York

Task 5

Point A

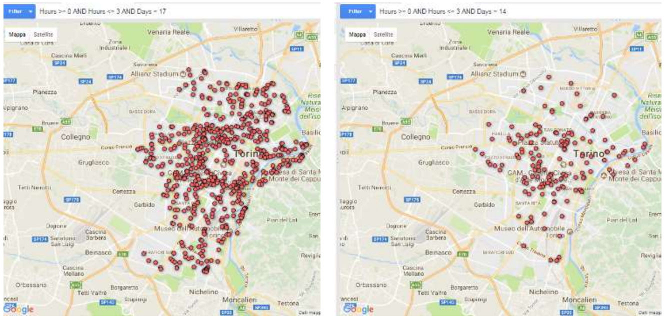
For this task we have chosen Torino as a case study.
Using fusion tables and data extracted from Mongo DB about parking in September 2017, we have tried to make some considerations and comparisons.
It is worth noticing that the data shown are referred to the act of parking in a set period and not to the overall parked car in the period. We decided to make some comparison based on:

- Total parking in different time of the day [Fig.17]
- Parking in a night slot in weekends and working day[Fig.18]
- Parking in peak hours during weekend and working day [Fig.19]



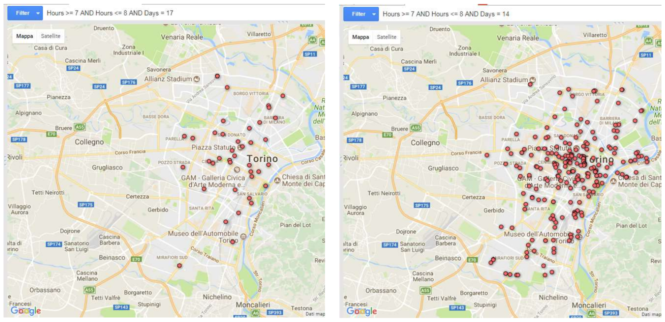
From 3am to 4am (GMT) = 1337 From 7am to 8am(GMT) = 6075

Figure 17 – Parked car in different time slots


Filter
Hours >= 0 AND Hours <= 3 AND Days = 17
Filter
Hours >= 0 AND Hours <= 3 AND Days = 14

Sunday 17th September =598 Thursday 14th September=155

Figure 18 – Car parked on two different days from 0am to 3am (GMT)


Filter
Hours >= 7 AND Hours <= 8 AND Days = 17
Filter
Hours >= 7 AND Hours <= 8 AND Days = 14

Sunday 17th Septemer =52 Thursday 14th September=232

Figure 19 – Car parked on two different days from 7am to 8am(GMT)

Point B

In task 5.b we were asked to divide the area of Turin in grids of 500x500 meters and to visualize the density of cars in each area.

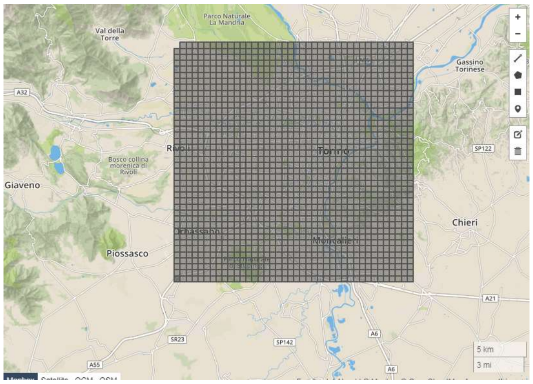


Figure 21 – City division in squares of 0.25 km^2

We decided to take a square larger than Turin in order to take into account also possible outgoing traffic, and then by filtering out the zones with 0 traffic ingoing and outgoing we realized with the support of Fusion Tables the figure 22. It is noticeable that one point is separated from others, that point refers to the Le Gru shopping center.

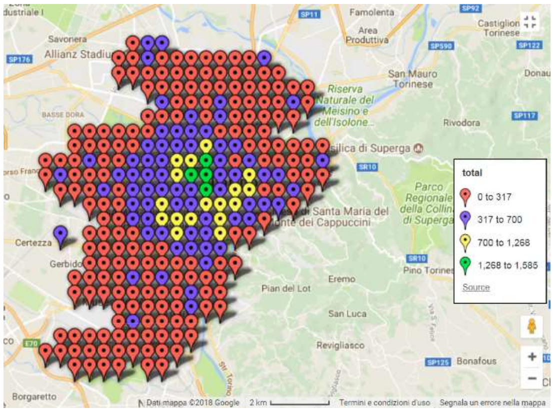


Figure 22 – Density of cars in each computed zone

In Table 1 then, it is shown the difference between two different days in car park density in different hours.

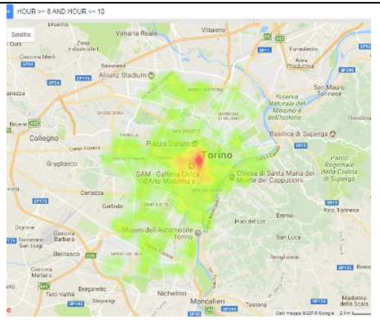
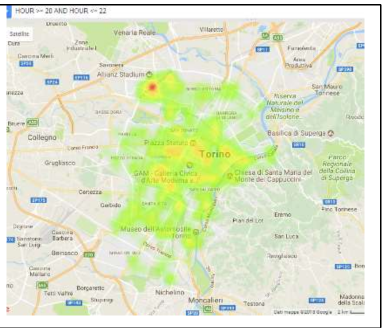
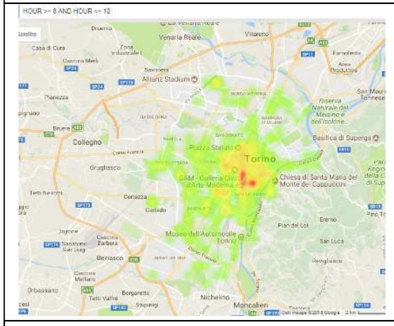
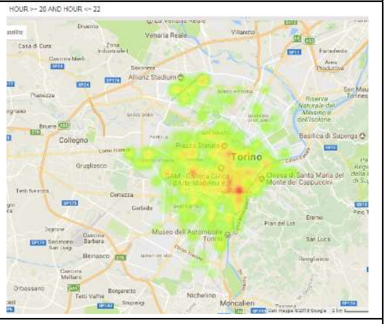
	
20th September from 8am to 10am	20th September from 8pm to 10pm
	
24th September from 8am to 10am	24th September from 8pm to 10pm

Table 1 – Comparison density 20th September (Working Day) and 24th of September (Sunday) in two different time slots Point B

It is interesting noticing how the red area during Sunday are close to the center and to the main Station of Porta Nuova while the area covered in the 20th of September is more spread.

Another factor interesting to notice is the presence of an "out of the ordinary flow" during the night of the 20th of September at the Juventus Stadium.

That night a football match of the major league took place, and this confirms the importance of taking into account events and attractions in the transport supply planning.

Point C

In this task we created an origin destination matrix by using the square zones created in Point B.

By running the query on the permanent booking collection, we retrieved all the coordinates related to starting point of a trip and ending point. Using a Python script, we assigned these points to related square zones and we insert the total number of the points in each zone to origin destination matrix. For visualizing the result, we used MATLAB to create surface plot. (figure 23)

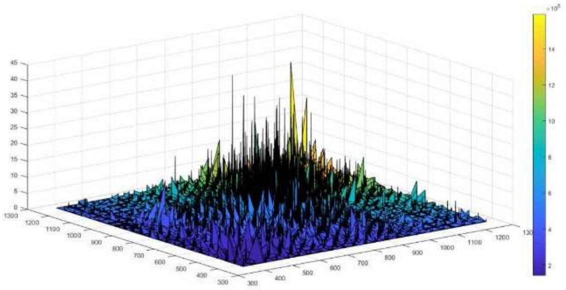


Figure 23 – Surface Plot of Origin Destination Matrix

As shown in the plot most of the peaks are located in the corner. Those zones refer to the coordinates of Turin's center.

For further analysis, we decided to create a heat map both for in going and outgoing flows entering/exiting in/from the zones to see which parts of the city are more involved in the trips. The results shown in figure 24 and 25, confirm the data of Fig. 23, as most of trips ended in the center (Fig. 24) as well as most of them started from there (Fig. 25).

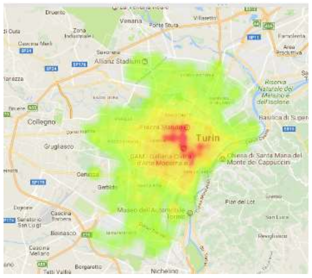


Figure 24 – in going flows to zones

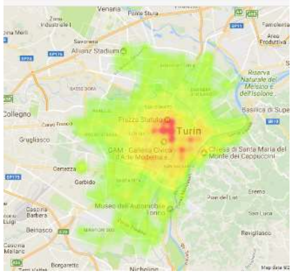


Figure 25 – outgoing flows from zones

Task 6

In order to correlate the probability of a rental with the availability of other transport means, we applied the same filters explained in task 3.

Moreover, we only considered those tuples for which an evaluation of the same path by using PT has been made and is available on the DB.

Once extracted the public transport trip duration, we divided it into time bins, [0,5) min, [5,10) min, [10,15) min and so on and then we computed the number of rentals for each bin.

The result of this analysis is shown in Figure 26.

where the histogram diagram shows that people most of the times use carsharing for duration between 15 and 25 minutes, while the maximum is between 15 to 20 minutes.

This remark is compatible with results that we have in task four related to average rental duration in Turin.

In Fig. 26, we can also see that the average rental duration is between 20 to 25 as already shown in the CDF of Fig. 5.

Furthermore, we understand that for shorter and longer travel duration number of bookings decreases and people prefer to use public transport system instead of carsharing.

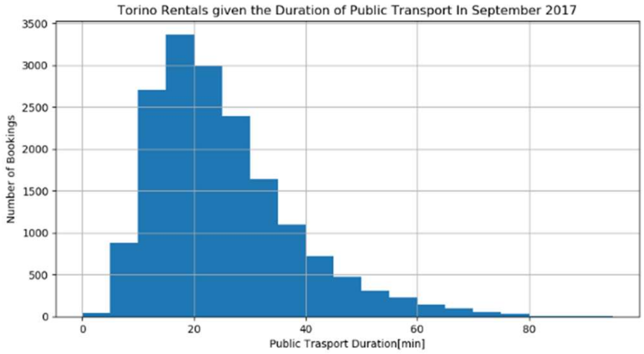


Figure 26 – Probability of a rental with the availability of PT in Turin.

Laboratory 2

The objective of this laboratory was to find out which are the most likely type of users of carsharing system. For answering this purpose, we need to compare Origin- Destination matrixes of carsharing users with a given dataset to obtain the most similar OD matrix between them.

Datasets

IMQ dataset contains movement of people considering all modes of transport. The data has been collected through phone interviews in 2013 in Piedmont region, but we only focused on a set of 23 zones related to Turin city. It includes trips from Monday to Friday, people profile like gender, age and trip motivation like Work, study, shopping etc. Tables 1 to 3 show categories related to people profile and trip motivations.

Gender	
1	Male
2	Female

Table 1- Gender

Age Range	
1	From 11 to 19
2	From 20 to 49
3	From 50 to 64
4	65+

Table 2- Age Range

Purpose	
1	Go to work
2	Working reason
3	Study
4	Shopping
5	Bring someone
6	Cures or medical visit
7	Sport or leisure
8	Going back home
9	Visiting relatives or friends
10	Other
11	Going back home on the day of interview

Table 3- Trip Motivations

For carsharing users' dataset, we have access to two months rentals data for Car2go and Enjoy with origin and destination, indexed for running geospatial queries, which are stored on MongoDB.

Methodology

To create different OD matrixes from the carsharing datasets and finding the best matches with smallest distance respect to IMQ data, we decided to define different categories. First, three different OD matrixes related to carsharing have been created: one matrix related to the Car2Go service, one to Enjoy and one having both as total. In the second step we divided the datasets in four subsets: working days, weekends, morning and afternoon hours, where morning goes from 5am to 10am while afternoon from 4pm to 8pm. For comparing morning and afternoon time, we only extracted data related to the working days since, the IMQ dataset does not have data related to the weekends. At the end, we have 15 different OD matrixes coming from carsharing datasets.

Distance Calculation

To extract the set of gender, age and scope that better matches the carsharing set, we decided to run all the possible combinations of the three set (excluding the empty subset for each of them) having in the end 92'115 possible subset.

Thanks to a Python script using Pandas, we read the OD matrix related to the transport survey, we normalize the data in OD matrixes by dividing all cells by the sum of all trips, then we computed the difference by using equation 1 with three different OD matrixes related to carsharing: one matrix related to the Car2Go service, one to Enjoy and one considering both. Results of this analysis are shown in table 4.

$$d(A,B) = \sum_{i=0}^{n} \sum_{j=0}^{n} |a_{ij} - b_{ij}|$$

Equation 1- Distance Equation

			Gender		Age Range				Purpose											
No.	Dataset	Time	1	2	1	2	3	4	1	2	3	4	5	6	7	8	9	10	11	Distance
1	Enjoy	All Days	✓	✓	✓	✓	✓		✓	✓	✓			✓		✓			✓	0.652438359252
2		Working Days	✓	✓	✓	✓			✓	✓	✓					✓				0.650100604734
3		weekends	✓	✓	✓	✓	✓		✓	✓	✓			✓		✓				0.675765999961
4		Morning	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓				✓		0.709848528764
5		Afternoon	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓	✓	✓			0.737577927716
6	Car2go	All Days	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓	✓			✓	0.641421596945
7		Working Days	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓	✓			✓	0.648928599071
8		weekends	✓	✓	✓	✓			✓	✓	✓				✓	✓			✓	0.637642633493
9		Morning	✓	✓		✓	✓	✓	✓	✓	✓						✓	✓		0.752774468214
10		Afternoon	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓	✓	✓			0.693640036632
11	Total	All Days	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓	✓			✓	0.626681362256
12		Working Days	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓	✓				0.626573137011
13		weekends	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓	✓			✓	0.645441468321
14		Morning	✓	✓		✓	✓	✓	✓	✓	✓						✓	✓		0.701732476943
15		Afternoon	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓			0.685842819342

Table 4- Best combinations of IMQ dataset that have smallest distances with carsharing datasets

Results

In table 4, is reported the best match for any subset of days and carsharing services abovementioned. The smallest distance from the IMQ OD matrix in every dataset is presented with red color. As it is noticeable in this table, both female and male category take part in all combinations and the class of age range more than 65 participate less in the matrixes. The main purposes that have most similarity between IMQ dataset and carsharing systems are: going to work, working reasons, studying, cures or medical visit and going back home on the other hand shopping and bring someone are not take place in the best combinations. Another interesting result that we can notice is by comparing morning and afternoon times it seems from the analysis that people use carsharing in the afternoon for cures or medical visit, sport or leisure, going back home and visiting relatives or friends.

The plot in Figure 1 shows the distance of the optimal match (computed without any filter on days or carsharing service) from the IMQ OD matrix.

It can be noticed how the main distance from the IMQ matrix is computed on the main Diagonal. From our point of view, the difference is due to the statement made by the interviewed people during the survey. If we look at the IMQ matrix, we can see how the main movement are intrazonal. This behavior is not replicated by carsharing users and this leads to the distance we can appreciate in the following plot.

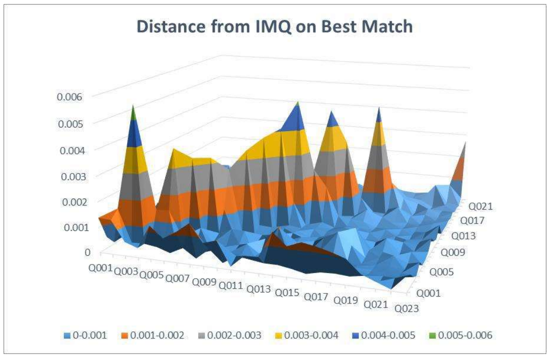


Figure 1- Best Match distance from IMQ matrix

We tried then to filter out only the working days to appreciate the same plot shown in Figure 2. This plot shows how reducing the data set, the optimum is going to be farther than the previous solution with respect to the IMQ matrix.

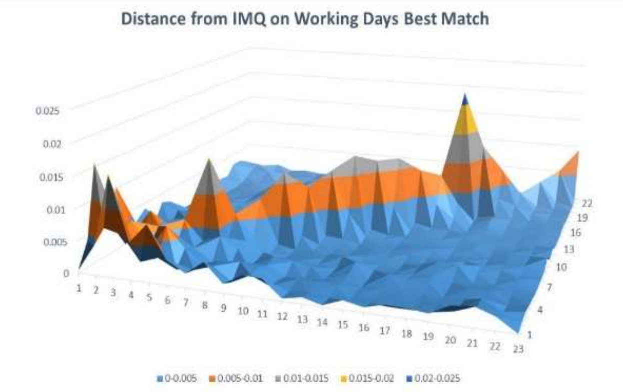


Figure 2- Distance from IMQ of Working Days best match

Further Analysis

In the following charts, it is important to remind that, what is showed in the graph is the distance from the survey sample hence, the smaller the column, the closer is that category to the car sharing customer type.

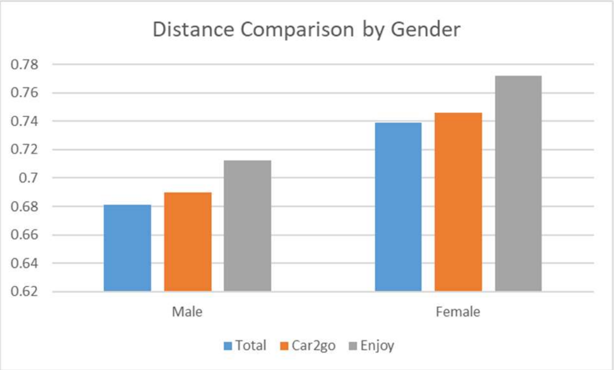


Figure 3

In Figure 3 we are comparing distances between carsharing datasets and IMQ data with respect to gender. The first thing that can be noticed is the male users have the best match comparing with female users. and in general, car2go data set fits better consider with Enjoy.

For more analysis we decided to calculate the distance with respect to the age categories to understand which category in IMQ matches better with carsharing data. As we saw before in table 4 in this part we ignore people older than 65. As it shown in Figure 4, the best match is second category, so we can conclude that people from age 20 to 49 uses the carsharing system more than the other groups.

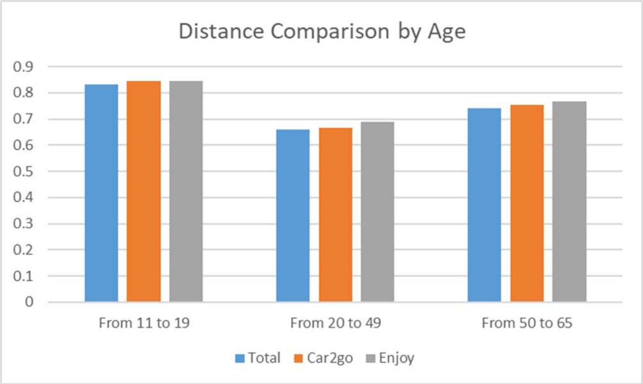


Figure 4

In addition, since it is not clear in table 4 which purpose is more suitable with carsharing, for this category we decided to compare them one by one (please refer to table 3). Please note that in figure 5 we compare the more relevant purpose of table4.

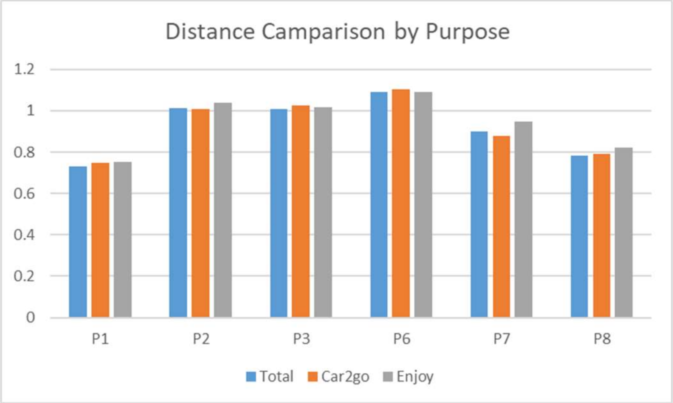


Figure 5

It is easily noticeable that category P1 which is related to work purpose has the smallest distance.

Conclusion

In this laboratory we compared origin destination matrixes coming from two different datasets: one is the IMQ dataset that data is gathered by interviews and the other one is carsharing system which include car2go and

Enjoy. By taking differences between these matrixes we obtained interesting results

- Comparing Male and Female, trips taken by men has smallest distances.
- Comparing the age ranges, trips taken by people from 20 to 49 has the best fit.
- The main purpose for taking the trips are: going to work, working reasons, studying, cures or medical visit and going back home but in general trips taken for going to the work has the most similarity.
- Comparing morning and afternoon time it is noticeable that in the morning time the main purposes for taking trips are: going to work, working reason and study while on the other hand in the afternoon we can see four more purposes which are: cures or medical visit, sport or leisure, going back home and visiting relatives or friends.

At the end we should mention that IMQ dataset refers to 2013 and there is 4 years gap, maybe by having newer data it is possible to find better results due to the increased popularity of carsharing systems four years difference can affect the age ranges categories.