POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Analysis of Milan public transport surface network: offer performances in an intermodal perspective

Tesi di Laurea Magistrale in
Mobility Engineering - Ingegneria della Mobilità

Andrea Parravicini, 969331

Advisor:
Prof. Simone Vantini

Co-advisors:
Arianna Burzacchi

Academic year:
2022-2023

Abstract: Public transport offer in the city of Milan consists of two main networks: the metro network and the surface network, both managed by *Azienda Trasporti Milanesi S.p.A.* (ATM). Since the lack of studies regarding the latter, this work took it as the topic of the study, in particular considering just data about the offer. The methodologies adopted are the ones of Network Analysis, in order to analyze urban mobility. Results obtained could be then exploited by ATM itself, which eventually could take potential measures to improve the service offered to users.
Starting from GTFS input data of a standard week, it has been possible to build the network first, considering all its aspects: nodes, edges and additional information, such as lines and paths that transit through them. The nodes of the network built have then been subjected to a clustering process: this has been done by simply grouping together stops close to each other, to make the network more similar to how it is perceived by users. A weight has then been assigned to each edge of the network, based on the number of seats available along it during a predetermined time period. Some representations of this weighted network have been done, allowing to better comprehend how the offer changes during the week considered and how it spreads geographically. These weights have then been used for the definition of some indicators to assess the performances of the service.
As last considerations, since the importance of intermodal transportation and because ATM manages them all, the metro network stations have been considered in order to relate surface performances to their location: in particular, surface stops have been classified based on the distance to their closest metro station, and then a final comparison of performances has been done, to see whether the overall network is well integrated or not.

Key-words: Public transport network, network analysis, mobility offer, service performance, intermodality

1. Introduction and objectives

Public transport service available in the city of Milan consists of two main networks: the metro network and the surface network. For what concerns metro network systems in general, since the availability of data, many studies have already been developed (see [3, 6, 9]), and cover the main aspects and issues of both offer (i.e.

the actual network and the infrastructure present) and travel demand (i.e. flows of passengers and waiting passengers at the stations).
For what concerns instead surface network services, the situation is a little bit different: studies and analyses are not as well developed as the ones regarding the metro due to a bit of lack of information. For instance in Milan, the study area considered, it is not possible, at the moment, to have complete access to data about passengers' flow among the lines, since most of the vehicles of the city do not have APC (Automated People Counter) systems installed on-board. This is though a transitioning phase, since in the future all vehicles in the city will have such systems, and so complete analyses will be possible. Despite the lack regarding passengers' flow data, it is still possible to consider, as the object of the study, data about the offer and analyse the surface services network.
But why are studies about urban transportation networks so important? Possible analyses could lead in fact to the identification of potential criticalities that affect urban mobility; from criticalities then some measures to improve the service can be taken. The first one who can benefit from this kind of analysis is ATM (*Azienda Trasporti Milanesi S.p.A.*), who manages the public transport service in Milan; then, results of any potential measure directly affect users, who are offered with a better transportation service and a more pleasant travel experience.
Furthermore, since ATM also manages the metro system, it is interesting to see whether there is a good integration between surface services and underground ones and try to understand how surface performances change depending on the distribution of underground stations. Network analysis methodologies will allow to achieve the objective of the study: construction and analysis of Milan public transport surface network to evaluate offer performances in an intermodal perspective.
The concept of intermodal transportation represents in fact a possible solution for a more efficient and integrated travel experience, and it constitutes a great step towards a more sustainable future, a better urban mobility and a reduction of road traffic. The perspective of intermodal transportation should be increasingly considered in order to bring some improvements in mobility performances, to offer better services and a more flexible travel experience to users. The work done and the results obtained could be then considered as a starting point for further and future analysis, including those whose main objective will be that of the analysis of the passengers' flow inside the network (travel demand).
For all the analyses, *Python* and *RStudio* methodologies contained in [2, 4, 5, 8, 10, 13] were followed.

2. Input data and network construction

All the analyses and the reasoning of the project have been possible thanks to the availability of very detailed data. These data, provided by ATM, consist of a GTFS file format, which is a very popular file format when considering public transport data. GTFS file is composed by a series of text files, where each of them regards a specific aspect of the information of the public transport, such as stops, lines, routes and date and time service information. The data available are referred to a standard week of the winter service, starting from October 17th, 2022 (Monday) to October 23rd, 2022 (Sunday). These periods of interest allowed to have a vision of the offer both for time slots and for workdays and week-end days, and so to also understand the differences among the service and how it changes through a standard week. To sum up, the data contained in the GTFS file allowed to have all the useful information to find the location of the public transport stops (4852 stops), the description of the lines (159 lines), the list of paths of each line (a path is a specific sequence of stops) and the stop sequence for each path with the arrival and departure time at each stop for each day of the week considered. Alongside GTFS data, ATM provided also information about the service vehicles, specifying for each public transport line, the type of vehicle and its capacity.
With these initial data it has been possible to build the basic network first, composed by nodes and edges, and then to fill it with the offer data.

2.1. Nodes and edges

The network construction started with the identification of the nodes, that represent the actual public transport stops: the GTFS file already contained a specific dataset filled up with all the necessary information, such as the stop ID, its name and location and its coordinates. Coordinates values have been also converted from longitude and latitude into UTM coordinates, that lead to a better and more precise visualization.
The next step is about the identification of the edges. Edges represent the actual connections between stops: two stops are connected by a directed edge whenever there is at least a surface line path from the first stop to the second. Disposing of the sequence of stops for each path of the network, it was possible to create a dataset containing all the information regarding the edges, such as starting and ending stops and the lines and paths that transit through it. The first dataset allowed to build a multi-edge network: the number of edges between

each couple of stops will be equal to the number of paths that transit between those stops, that corresponds to every travel possibility available to users. There will be one edge per path, and all the edges will be directed. From the multi-edge dataset, by simply collapsing the edges between the same couples of stops, but keeping the information about paths, it has been possible to obtain a dataset for the construction of a single-edge network, always directed, which is the one needed for all the further analyses.
Nodes and edges allowed to build the first network, and to perform some initial qualitative analysis of its plot. In Appendix A, figure 13 shows the Multi-edge Network, while figure 1 below shows the Single-edge one.

Figure 1: *Single-edge network*

This representation shows the geographic extension of the network and, furthermore, it allows to visualize the distribution of the stops among the city considered, Milan. In addition, it can give an initial idea of whether the input data are correct, in terms of stops coordinates; since the network is the one of surface public transport of the city of Milan, a first and basic check could be done by simply visualizing the city on Google Maps, for example, and see if the network obtained tends to follow the real road network. In Appendix B, figure 14 shows a detailed plot to better see the differences between the multi-edge network and the single-edge one.

2.2. Nodes aggregation

A deeper investigation of the network lead to notice that there are some situations in which two or more stops, belonging to different lines, are very close to each other, and their respective distance is just a few meters. In all these cases, from the user point of view, those stops are perceived as just one unique stop. Based on this, it has been established to manipulate the network by aggregating those stops.
The objective of the aggregation is to bring together stops that are "close to each other" in order to make the entire network more similar to how it is perceived by users, who are the ones that actually benefit it. A collateral effect is, by the way, the simplification of the network in terms of reduced number of nodes (stops), which on the one hand it allows to make the analyses less time consuming, but on the other it may not represent the original network anymore. The choice of the most appropriate methodology and of the optimal distance value is then very important.
The methodology that better fits this purposes, in this case, is *Hierarchical Clustering* (see [7]). This method in fact creates a sequence of nested partitions of the network with a bottom-up approach. This means that, starting from the most similar nodes (stops closer to each other in terms of physical distance), it gradually aggregates all the nodes of the network until all of them are aggregated; then it's just a matter of "when" to interrupt the procedure and obtain a good, clustered network. Another important aspect of this methodology is that it does not require in advance to assume the number of clusters. Since the size of the clustered network is still not known, this is very convenient. Lastly, the pairwise distance between nodes can be carried out with different methods. The one chosen for this case is "Complete-linkage clustering" (or farthest neighbour

clustering): the value of the distance chosen represents the maximum distance between two stops belonging to the same cluster, and hypothetically, it is meant as the maximum walking distance that a generic user has to do to reach one specific stop inside the cluster.

2.2.1 Identification of clustering parameters

Figure 2 shows the curve of possible outcomes of the clustering procedure, in terms of number of clusters and clustering distance.

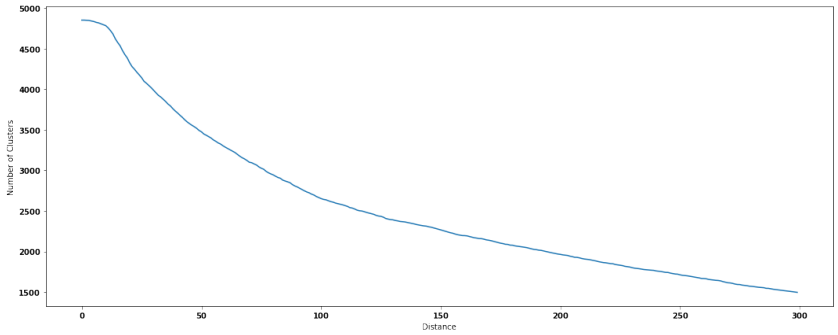


Figure 2: *Size of the network (Y axes) with respect to clustering distance (X axes)*

With a distance equal to 0 meters, the size of the clustered network will be the same as the original one. By increasing the distance, the number of clusters, and so the size of the network, will decrease, eventually coming to be equal to 1 when the distance is big enough to cover all the network. Note that this plot focuses though on just the upper part of the curve, which is defined for lower values of distances.
With the curve of possible results defined, it is now a matter of finding the optimal values for the network considered, precisely in terms of number of clusters and distance, taking some characteristics of the network itself as a starting point. These values will then lead to better results coming from Hierarchical Clustering algorithm. The number of lines and the number of stops per each path of the lines were considered as initial characteristics of the network. These would lead to the definition of the number of clusters, or a range of possible values of them.
A range of possible theoretical values has been set, defined by two cases: *Average Maximum Size* and *Average Mean Size*. The *Average Maximum Size* has been computed by considering, for each line and for each direction, the path with the maximum number of stops; then, by simply computing the average of the number of stops and multiplying the result for the number of lines considered, it is possible to obtain the *Average Maximum Size*, which is equal to 3597 (this reasoning has been done by assuming that each line had the maximum number of stops). The *Average Mean Size* instead has been computed by considering, for each line and for each direction, all the existing paths; then, by simply computing the average of the number of stops of the paths and multiplying the result for the number of lines considered, it is possible to obtain the *Average Mean Size*, which is equal to 2955.
These two theoretical values about the size of the clustered network are represented in figure 3 with respect to the clustering curve (yellow and green horizontal lines), ad define a range of possible outcomes. Two respective distance values correspond to those size values, which defines two possible scenarios.

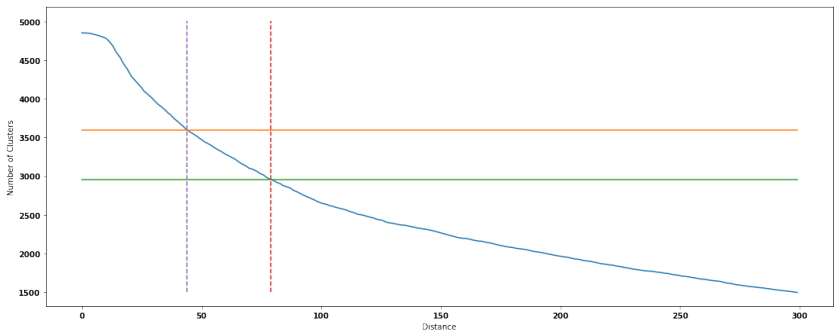


Figure 3: *Scenarios identification on Distance – Size curve*

***Precautionary Scenario*:** it corresponds of having a number of clusters equal to the *Average Maximum Size* (3957), and a corresponding distance value equal to 44 meters. This means that stops inside the same

cluster will be far from each other 44 meters at most. In this scenario the network that would be obtained will be as similar as possible to the original one, and the clustering distance will be lower.

***Worst Case Scenario*:** this scenario instead corresponds of having a number of clusters equal to the *Average Mean Size* (2955), and a corresponding value of distance equal to 79 meters. With respect to the precautionary scenario, in this case the network would be more distorted and as dissimilar as possible to the original network (of course with respect to the initial considerations), but it would be more simplified.

The choice of the scenario, and so of the clustering distance to adopt, went on the *Worst Case Scenario*, in order to have a simpler network and that was easier to manage. This, though, does not prevent to consider the other scenario, or even other initial considerations, for future analyses. It is now necessary to evaluate the clustering distance, equal to 79 meters, and see whether it can be acceptable or if the whole process needs some more reasonings. The distance assessment involves two steps:

- Evaluation from the point of view of users
- Evaluation based on the infrastructure of the network.

For what concerns the first one, it is not scientifically proven, but it is reasonable to say that 79 meters is an acceptable value of the maximum walking distance from one stop to another inside the same cluster (most people would walk that distance without problems). Regarding instead the second step of the evaluation, it is necessary to see what 79 meters as clustering distance would involve inside the network, especially for what concerns stops belonging to the same paths. Some checks were done done, with particular attention to all the cases in which the distance between two consecutive stops of the same path is less than 79 meters. In these cases, in fact, there is a high chance that those consecutive stops end up in the same cluster.

To verify if this situation affects the network, it is possible to just compute the length of all edges and, for each path, extract the shortest ones, and then compare them to the distance value considered. If lengths are less than 79 meters, then it is possible to check them directly on the map and understand the reasoning. This kind of situations can happen, for example, near the terminal of the lines, or if multiple stops are located around the same square. There are just 18 cases where the length of the edges was less than 79 meters, and with respect to the totality of the network they represented an extremely poor percentage of cases. Furthermore, since 79 meters have been considered acceptable from the users' point of view, the clustering of the network can be done. The result will be a new network with 2955 nodes (clusters), in which the maximum distance between stops belonging to the same cluster is 79 meters. Please note that the clustering procedure followed for the case considered is not the only one possible, but it seemed that it fitted the purposes of the work .

2.2.2 Nodes clustering evaluation

This new network is the one considered for the future steps of the work, and for all the further analyses. It is fair, now, to question about how the new network was compared to the original one, and if it was distorted too much.

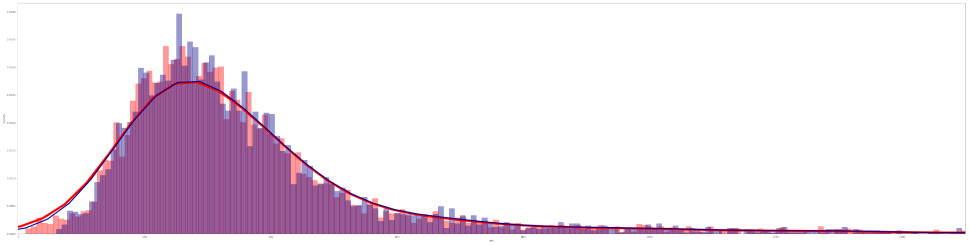


Figure 4: *Edges length distribution*

First, lengths of the edges, and so the distances between the stops have been checked. Figure 4 shows the distributions of length of edges of the Original Network (red) and of the Clustered Network (blue) overlapped, with their respective density curve. It is possible to see that these distributions are basically the same, and that there are not radical changes. This means that the distances between stops have been preserved, and it can be considered as a good indicator of the clustering procedure. Then, in addition to the distribution of the lengths of the edges, it is also possible to compute some indicators of the Macro-scale analysis of the networks, to further evaluate the result and compare the two networks (table 1).

All values show that the clustering procedure has been carried out in a good way and, furthermore, the diameter and the mean distance confirm what it has been possible to deduce from the graphs of the distribution of the lengths of the edges, so that the geographical extension of the network basically does not differ from the original one. The number of components decreases from 3 to 2. The fact that it is not equal to 1 does not represent an issue for the analyses though, since the second component is related to an independent line (*line 171, Cimitero Maggiore*) that will not be considered for further analyses.

	Original Network	Clustered Network
Size	4852	2956
Density	0.000504	0.00127
Number of Components	3	2
Diameter [m]	43481	40862
Mean Distance [m]	12358	11312
Clustering Coefficient	0.03075	0.10324

Table 1: *Macro-scale indicators*

In Appendix B, figure 15 shows a detailed comparison between the Original Network and the Clustered one, while figure 16 shows the entire clustered network. In the Clustered Network, the "new nodes" obtained will be called, from now on, *"Macro-stops"*.

3. Offer computation and indicators definition

After the first phase the network of Macro-stops, edges and paths per edges has been built. In this second phase the available information of the actual offer was used to enrich the network and analyse it in terms of passenger offer. Offer available means, at first, the number of trips that each line (and more specifically each path of a line) performs in a certain time slot of the day; then, going more into detail, the offer will be referred to the number of seats available, so that it could be more easily related, in future analyses, to the number of effective passengers.
The input data provided by ATM, as already said in chapter 2, consist of a GTFS file, with information that spread from October 17th, 2022 (Monday) to October 23rd, 2022 (Sunday), that is a standard week without any event that could compromise the service regularity. These data allowed to perform a pretty detailed analysis of the offer since, depending on the time slots defined, it is possible to obtain more specific or more generic results. For this work, hourly time slots were considered for each day. Furthermore, in order to align the results with the standards adopted by ATM, time slots of each day were shifted by three hours: this means that each day "starts" at 03:00 a.m. and "ends" at 03:00 a.m. of the next day. Time slots after midnight will be changed into 24:00 – 25:00, 25:00 – 26:00 and 26:00 – 27:00.
Once the time slots have been defined, it is possible to compute the number of trips that each path performs during each time slot. Starting from the dataset containing the arrival and departure time from each stop of the paths, for each trip performed, by considering the departure time of the trip of the path at the first stop it was possible to count the total number of trips during each time slot. More precisely, what has been computed is the number of trips that each path begins in the time slot. This was done for all the paths. The output of this procedure is a set of datasets, one for each day of the week, with the list of all the paths existing in the network, and for each of them, for each time slot, the number of trips that they perform (begin).
Alongside these datasets just created, ATM also provided data about service vehicles, specifying for each of them their respective service lines and their maximum capacity. By merging these two datasets it is possible to compute the final offer, in order to assign the seat availability as weights to edges. When building up the dataset about the edges, the information of the paths that transit through each of them has been maintained, and so it has been possible to assign to each of them the number of seats available, based on just the paths that transit through it. This has been computed according to the following formula:

$$Edge\ Offer_{time-slot} = \sum_{paths} Seats_{path} * Number\ of\ trips_{path,time-slot} \tag{1}$$

Formula (1) refers to a single edge; it takes just the paths that transit through it and their respective number of trips (per each time slot); then, by just multiplying the number of trips for the number of seats available for the path, and then summing up the values for all the paths considered, it is possible to obtain the final offer of the edge (for each time slot), which is expressed in number of seats available per hour.
By computing the average value of the offer per each time slot and per each day, it is possible to show its distribution in time (figure 5). The trend shown here is as it could have been expected, since it follows the one typical of public transport mobility [12]. Three different trends can be identified: one for all the working days (from Monday to Friday), one for Saturday and the last one for Sunday. Working days present the highest trend of the three, and two separate peaks can be identified: the first one coincides with the morning peak hour (from 06:00 to 09:00 a.m.), and the second one, instead, coincides with the evening peak hour (from 16:00 to 20:00). The first peak is higher than the second one, but also shorter in terms of time duration. Saturday

trend approximately follows the one of working days, even if the peaks are less pronounced. More in general though the overall trend is lower. Sunday trend is the lowest one, with a pretty constant offer and without any particular peaks along the course of the day. Lastly, it is interesting to notice that the offer during night hours is basically the same regardless of the day of the week.

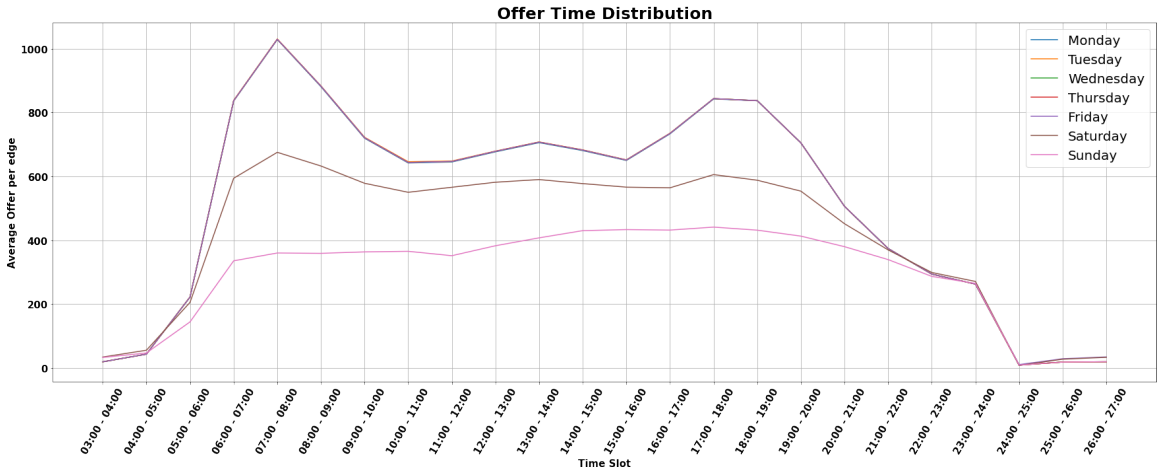


Figure 5: *Offer time distribution*

The offer trend and the conclusions that can be drawn from it confirm what it could have been expected from the offer of a public transport service in a city like Milan, during a standard week of the year. Figure 5, though, allows only to make quantitative considerations about the offer in aggregated terms only. In fact, it doesn't give any information about how the offer is geographically distributed inside the network. For a more detailed analysis, some representations of the network during different time-slots and different day were made. Since the quantitative differences between working days are basically null, future graphs and visualizations of the network will be referred to just the type of day: working day, Saturday, and Sunday. Regarding working days, Thursday is the one that shows slightly lower values about the offer, and so it will be the one that will represent and summarize the behaviour of all working days.

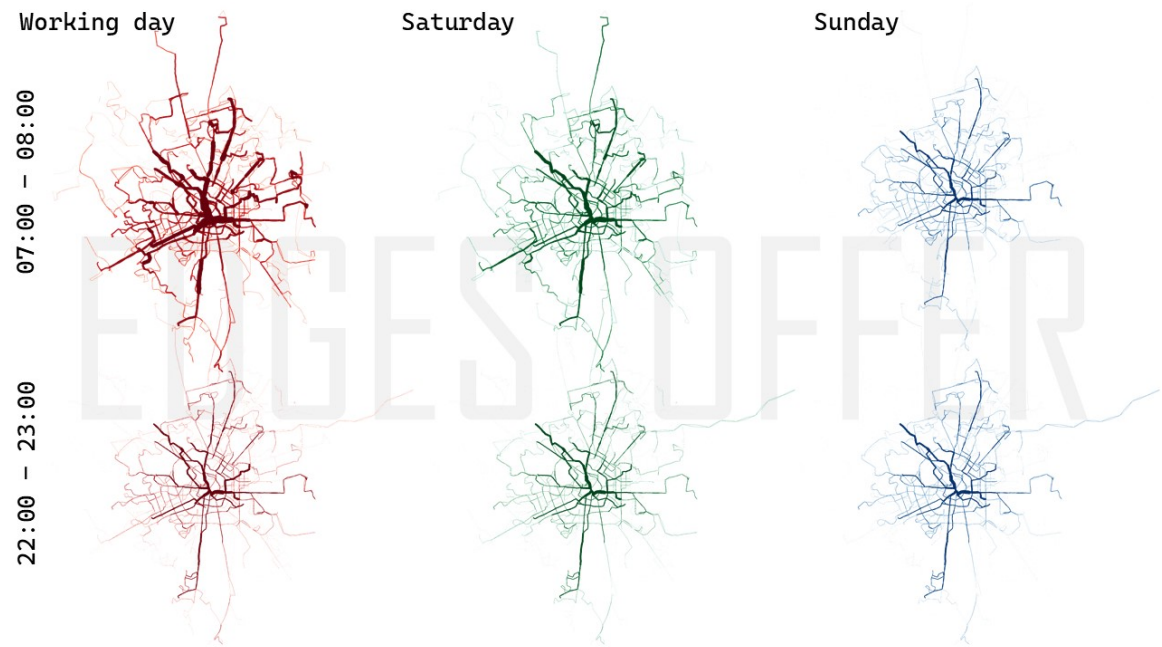


Figure 6: *Offer distribution in the network. Each plot represents the situation of the network for a specific day and for a specific time slot*

Figure 6 shows not only the differences in quantitative terms of the offer, but also its spatial distribution inside the network. For each type of day two maps are shown: one representing the situation of the network during the morning peak hour (from 07:00 to 08:00); the other one instead shows the network at late evening (from 22:00 to 23:00).

Immediately it is possible to see how the offer is being reduced at late evening, both in quantitative terms and spatial terms. Another thing that is possible to notice is that, by looking at the maps during the morning peak hour, from working days to Saturday the major difference basically consists of a reduction of the offer available just in quantitative terms; from Saturday to Sunday instead, the reduction of the offer isn't just in quantitative terms, but also in spatial terms. There's a reduction of the areas covered by the service.
In addition to the available offer related to edges, it is also possible to relate it to the Macro-stops. In particular, a pretty useful and easy to understand indicator is *headway*. Headway is the time interval between two successive trips that transit through the same stop, and from the point of view of users it is very useful since it can give an idea of the average waiting time at the stop. Starting from the number of trips that each path performs per each time slot during the day, it has been possible to compute the number of trips passing through each stop. To adapt those values to the clustered network used for the analyses, a simple summation was done for all the trips of the stops belonging to the same Macro-stop. This was done for all the Macro-stops and for each time slot. From the number of trips to the headway is then very simple: it is necessary just to divide 60 (minutes per hour) for the number of trips per each Macro-stop, as shown in formula 2:

$$Headway_{MACRO-Stop} = \frac{60}{Trips_{MACRO-Stop}} \tag{2}$$

The result is a dataset containing all the headway values, per each Macro-stop in each time slot for each day of the week. Then, similarly to what has been done for the edge offer, by computing the average of the headway values per each time slot and per each day, it is possible to show its distribution in time.

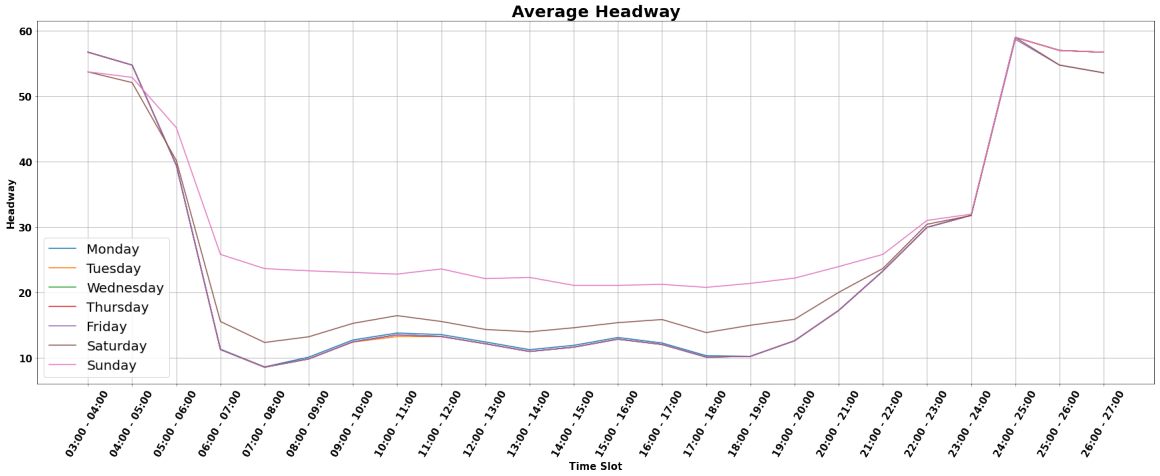


Figure 7: *Average headway time distribution*

Also in figure 7 it is possible to identify three separate trends: one for working days, one for Saturday and one for Sunday. With respect to the quantitative offer though, here lower values are better than higher ones: the lower the value, the lower will be the average headway, and so the time interval, in minutes, between a trip and its consecutive one. Lower values of average headway are referred to working days, while the higher ones are referred to Sunday, as it was possible to expect. The three different trends collapse into one trend though when looking at night time slots.
It is important to say that, for all the cases in which there were no trips during the time slot considered, the formula for the computation of the headway would have given, as result, an infinite value. Those values have been substituted with a value equal to 60 minutes (equivalent to the time length of the time slots) just for the purpose of the computation of the average value of the headway. The graph with the distribution of the headway does not give any information on how headway values are geographically distributed inside the network. In Appendix C, figure 17 shows its geographical distribution.

3.1. Indicators definition

To quantify, and then evaluate service performances, some new indicators are defined, to better summarize the overall level of service of the network. Apart from the more classical *micro-scale* indicators (see [11]), that can help to better understand some aspects related to nodes, the new indicators defined could be useful for future analysis when combined with travel demand data. For this purpose, indicators computed will be referred to Macro-stops, so that they could be more understandable and practical from users' perspective.

Two indicators have been defined. The first one represents the *Potentiality* of Macro-stops: it indicates, for each Macro-stop, the number of seats available that users can have and benefit from. Its definition is equivalent to the one of the *out strength*, that is:

$$Out\ Strength_{Macro-stop} = \sum_{Macro-stop} Outgoing\ Edge\ Offer_{Macro-stop} \tag{3}$$

Potentiality allows to understand which of the Macro-stops have the largest offer, and so the nodes in which users have more possibilities to take advantage of surface public transport service. Since *Potentiality* refers to the available outgoing offer from each Macro-stop, it could be useful to consider it alongside headway, after a slight reformulation, since both these information directly affects the level of service for users. *Potentiality* has been recomputed to be referred to one single trip of the Macro-stop: this "new" parameter is called *"out strength per trip"*, defined as:

$$Out\ Strength_{Trip} = \frac{Out\ Strength_{Macro-stop}}{Trips_{Macro-stop}} \tag{4}$$

Figure 8: *Potentiality per trip - Headway subplot. Each plot refers to a time slot. Along the X axes there is the headway; along the Y axes there is the out strength per trip. Dimension and colour of points indicate their Potentiality; position in the graph instead indicates the type of service.*

Figure 8 presents 24 smaller plots of potentiality per trip over headway, one for each time slot of the day. By start looking at the one referred to working days, it is possible to see that most of the point tend to lay on the left side of the graph, and so to have low headway values. Furthermore, the distribution of points during central hours of the day basically does not change. This allows to conclude that the overall service is pretty much constant and continuous inside the network for most of the day. Points then start to spread during late evening, and the number of points starts to decrease, meaning that active Macro-stops are fewer.
Appendix C shows these same representations but related to Saturday (figure 18) and Sunday (figure 19). Saturday situation is not that much different with respect to working days. Points tend to be a little bit more spread in all time slots, but also here the service could be considered pretty much continuous. From Saturday

to Sunday the difference is evident instead, and graphs show much more dispersed points during the whole day: the service here is less continuous than the other days of the week.
Representations of service offered like the one showed in figure 8 do not allow, though, to visualize how performances are distributed in the network, since two parameters are considered together. The second and last indicator defined for the analysis aims at grouping those two parameters combining them into one, in order to sum up all service information. This indicator is *Continuity*, defined as:

$$Continuity = \frac{Out\ strength_{Trip}}{Headway} \tag{5}$$

Continuity considers information not only about the quantity of the offer available at each Macro-stop, but also about the way in which this offer is managed and made available to users. It is expressed in "number of seats per minute", and it allows to make plots with geographical distribution of performances (figure 9).

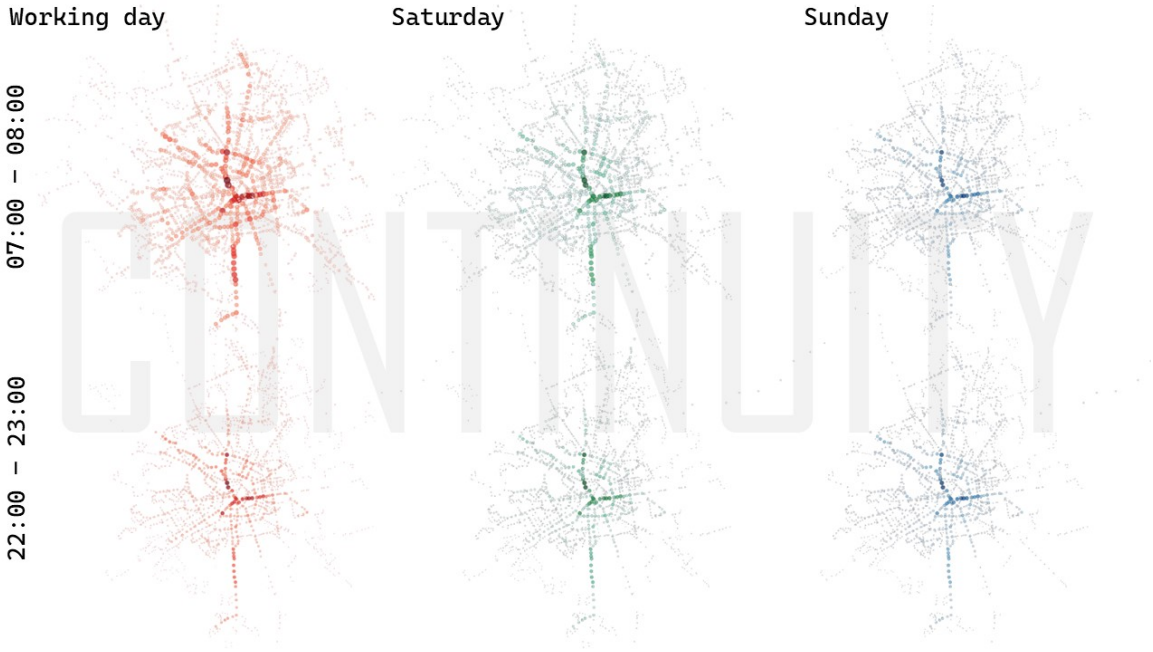


Figure 9: *Continuity distribution among Macro-stops. Each plot represents the situation of the network for a specific day and for a specific time slot*

Pattern of best Macro-stops repeats itself for different days, even if *Continuity* decreases between working days and Saturday, and even between Saturday and Sunday. This reduction happens also between morning peak hour and late evening. Another interesting thing to notice is that performance differences between high *Continuity* and lower *Continuity* nodes are less pronounced in working days, in particular during morning peak hour. This leads to a more homogeneous service inside the network.
Last consideration about *Continuity* regards its formulation: it took as input the "out strength per trip" and the headway, considering both terms equally, even if in reality it may not be like this. For future studies, since both those parameters determine the service offered and affect users, it could be useful to understand their preferences (whether more offer available or lower headway) and, as a result, re-arrange the formula with the introduction of a weight factor. This would lead to better and more accurate results, that are aligned with users' preferences. A possible way to obtain users' preferences is by submitting them surveys, maybe differentiated with respect to different aspects, such as, for example, category of user, purpose of the trip, usual time slots, usual stop, ecc...

4. Relationship with metro network

The analyses and the indicators computed until this point allowed to describe and understand the behaviour of the offer available in the surface network, especially considering practical aspects from users' point of view. It might be a little bit reductive though to simply consider the results obtained within the surface services network, and not to see how these results relate with the other crucial public transport network in the city always managed by ATM: the metro network. Metro network in fact claims a very powerful and reliable offer,

characterized by high capacity, high regularity and continuity, and the fact that it is not affected by road traffic, which leads to faster transfers.
Considering how surface network performances relate with the metro network could lead to a better comprehension of the service offered, especially in a perspective of intermodal transportation. Furthermore, considering that users can benefit from metro network and surface network systems with one unique ticket, this integration makes even more sense. Given initial data provided for the surface network system and the results obtained up to now, the only aspect that is reasonable to consider about metro network is the location of its stations (Appendix D, figure 20).
All this brief introduction leads to the definition of the purpose of this last analysis: see how surface performances change (considering *Continuity*) with respect to the distance from metro stations and identify particular situations that might be critical. Two opposite scenarios can exist: the first one where performances are better for Macro-stops closer to metro stations; the other one where, instead, performances are better for Macro-stops further from metro stations.

What if performances are better for Macro-stops near metro stations?
In an intermodal transport perspective, which involves one or more changes of mean of transport, it represents an optimal solution. The interchange between metro and surface systems is well guaranteed, and the whole transportation experience is integrated.

What if performances are better for Macro-stops far from metro stations?
This alternative possibility would mean that there is no such good integration between metro and surface network; the two systems are practically complementary, and each one of them tries to compensate the lack of performances of the other. This solution, maybe, would lead to a slightly more homogeneous level of service, but from an intermodal transport perspective it is not an ideal solution: it would make the interchange critical.

Obviously a third situation can exist too: performances for closer and further Macro-stops are almost the same.

4.1. Metodology and results

Let's now define when a Macro-stop can be considered near or far from a metro station. First of all, the physical distance (euclidean distance) between metro stations and Macro-stops has been computed, considering then for each Macro-stops its nearest metro and its distance to it. Then, also performances values have been considered, to try to relate some possible common trends to the distance. A simple scatter plot with distance (X axes) and Continuity (Y axes) can show some potential relationship between the two parameters, that could lead to the possible identification of threshold of distance values that separate similar trend zones; these thresholds could then allow to identify some Macro-stops categories. This first possible categorization needs though to be statistically validated then, to get confirmation of the fact that the trends identified are actually valid and truthful. ANOVA test (see [1]) has been done for the statistical validation: it compares mean values of multiple groups and it evaluate differences in mean of at least one group by analyzing their variance. If the test confirms the validity of the categories, classification process can finish and some evaluations can be done.
This methodology just described was applied to three reasonable mobility scenarios (Working day, 07:00 – 08:00; Saturday, 22:00 – 23:00; Sunday, 12:00 – 13:00) that describe the network in different situations.

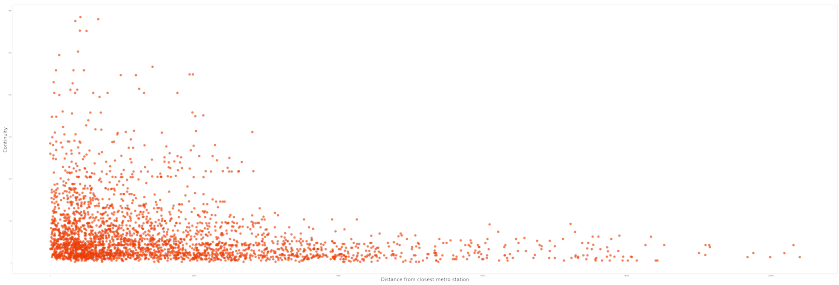


Figure 10: *Distance (X axes) - Performances (Y axes), Working days*

Figure 10 shows the scatter plot of working days. The ones of Saturday and Sunday present similar trends to the one showed here, and this allowed to define the same distance values for the categorization of Macro-stops. The proposed categories identified, based on the distance from the metro, are:

- Distance $<$ 79 meters: ***Neighbourhood***
- 79 meters $<$ distance $<$ 750 meters: ***Walkable Stops***
- 750 meters $<$ distance $<$ 3000 meters: ***Challenging Stops***
- Distance $>$ 3000 meters: ***Off-limits Stops***

ANOVA test was then applied, and for all scenarios it gave a valid and statistical evidence of the goodness of the categories (pvalue almost 0), confirming then the groups proposed.
The categorization allows now to see if there is a good integration between the surface network and metro network. For the three scenarios considered, average performance values of Neighbourhoods are better than the other categories, and average performances decrease by moving away from metro stations. This indicates that surface service is well integrated in the totality of transportation system, and that it's not meant to be like an alternative to underground service. This is the optimal solution from an intermodal transportation perspective, even if Saturday and Sunday scenarios show minor differences between categories. Working days during morning peak hours are better instead, showing an average increase of performances between Walkable Stops and Neighbourhoods equal to almost 29%. The categorization allows also to see how categories change between different scenarios. Here the differences are even more pronounced: from working day scenario to Saturday, Neighbourhoods' performances decrease of almost 65%, while from Saturday to Sunday they increase of almost 19%.
Considerations done up to now refer to the network in a general way, especially considering average values for the categories. A more detailed analysis of the network in the most stressed situation, which is the one during working days in the morning peak hour, was then conducted.

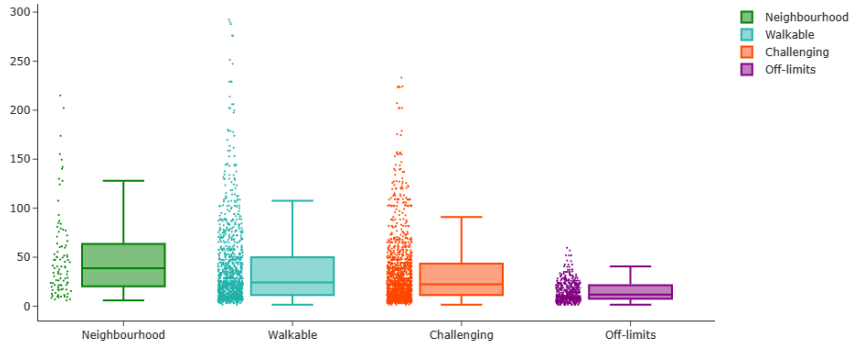


Figure 11: *Boxplot, Macro-stops categories and related performance values*

From figure 11, that is a box-plot representing performances of categories, one main thing can be noticed: performances range of Walkable Stops and Challenging Stops does not differ too much from the one of Neighbourhood. This is a very interesting aspect, since it means that, despite their average performance value is lower, some high performing Macro-stops still occur.
It is possible, then, to explore this latest aspect, to see which of the Macro-stops not belonging to any Neighbourhood are instead more performing, focusing, in particular, to those situations in which Macro-stops seems to follow specific paths or directions (directrixes). This configuration can lead to intend those paths as a really valid alternative to metro system.
From figure 12, a pretty clear directrix, that seems to be designed as an extension of metro *line M2*, is the one from one of its terminals (*Abbiategrasso*) to approximately Rozzano municipality (in the south area of Milan). Two main surface lines pass through this directrix: *line 3* and *line 15*, both served by high capacity vehicles with high frequency, that results in high performances. Because of this, *Abbiategrasso* can be considered as an interchange station; it's necessary to evaluate the performances of its neighbours, to see whether interchangeability is ensured:

- if they are low, the neighbourhood is a bottleneck for passengers' flow, and some measures should be taken to avoid it
- if performances are high instead, or at least aligned to the ones of the Macro-stops composing the directrix itself, the neighbourhood is good and intermodal travel experience is guaranteed to users

By checking its performances, *Abbiategrasso*'s neighbourhood shows good characteristics, aligned to the ones of the directrix, and so an interchange in that station is viable. Obviously, since it is the extension of a metro line from a terminal to peripheral area, it is reasonable to have lower performances with respect to metro's ones, but high enough to guarantee a good homogeneity of the service.
Another directrix is the one that runs from *Maciachini* metro station (*line M3*) to Niguarda district, in the northern part of the city. Surface *line 4* is the main one that serves this directrix, together with some other bus lines. Similarly to previous case, *Maciachini* can be meant as an interchange station, and its neighbourhood needs to be evaluated. This case though shows performance values lower than average, even if it's comparable

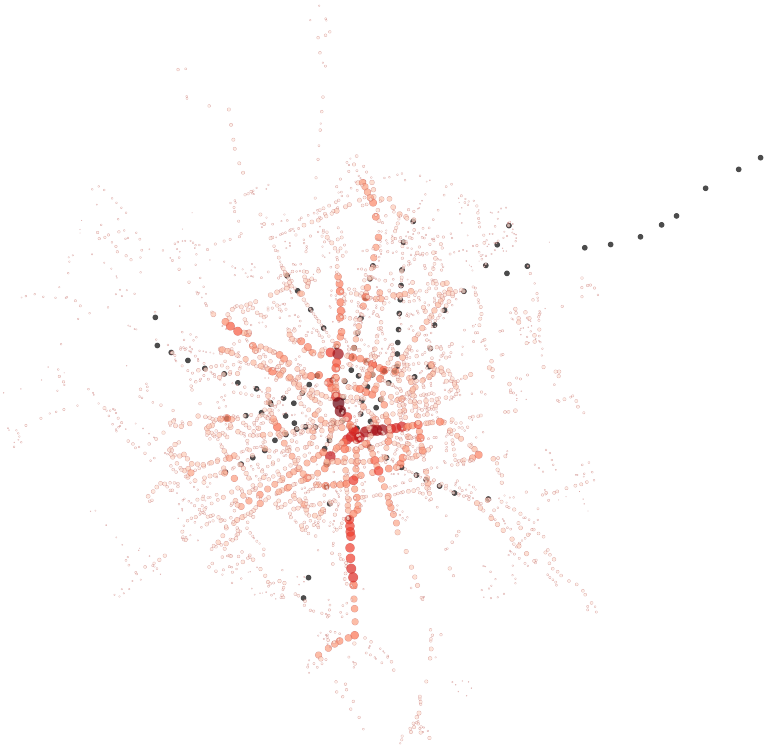


Figure 12: *Continuity values distribution in the network with respect to the metro stations*

to the one of Macro-stops belonging to the directrix. This neighbourhood, so, may be critical (just from the available offer perspective), and further analysis considering passengers' flow should be conducted.
Last big directrix that pops out is the one that goes from the city centre to the eastern part of the city, and it almost follows the path of the future metro *line M4*, which still needs to become fully operative. Surface services now show very high values of performances, but it's necessary to see how the situation evolves once the new metro line service begins, in order to understand how passengers' flow reallocates in the network and, eventually, evaluate possible measures to take to adapt service lines.
These were the three most interesting cases concerning the study of high performance Macro-stops not belonging to any neighbourhood, and that were lined up on some directrixes. Of course further considerations can be done, considering other parts of the city or other minor cases.

5. Conclusions

In this work the public transport surface network of Milan has been analysed. Starting from GTFS data, collected directly from service vehicles, it has been possible to make various analyses and to evaluate service performances for what concerns the mobility offer, also relating then the results obtained to the underground network.
Results obtained showed that Milan is a pretty connected city, with an integrated transportation network that can guarantee to users an intermodal experience of travel. As it could have been expected, working days present the best service, in particular during morning peak hours; Saturday and Sunday, instead, show some performances reduction, always allowing users, though, to enjoy a pleasant travel experience (although with a few more difficulties). Furthermore, the analyses allowed to identify potential service criticalities, which could be then assessed with the integration of additional data.
Just data about the offer available have been considered, and a very useful and interesting development of the project could be that of considering also passengers' flow data, so to better justify the results obtained and fully understand whether service potentialities are justified or not.
Additional data to be related with service performances that could be considered in the future are the ones concerning various services in the city: schools, universities, supermarkets, hospitals, work offices and green areas. These elements could allow to widely comprehend the performances of the public transport service.
Lastly, since all the analyses started from a GTFS file, methodologies defined and applied in this work are not

specific for the network considered, and so they could be easily adapted to also other transportation networks.

References

[1] Chao De-Yu. Anova test, with python.

[2] Aric Hagberg and Drew Conway. Networkx: Network analysis with python. *URL: https://networkx. github. io*, 2020.

[3] E Kluzer. The milan metro network from the first years of operation up to the present. *Vie e Trasporti*, 53(520-521), 1984.

[4] Eric D Kolaczyk and Gábor Csárdi. *Statistical analysis of network data with R*, volume 65. Springer, 2014.

[5] Douglas A Luke. *A user's guide to network analysis in R*, volume 72. Springer, 2015.

[6] Carlo Mannino and Alessandro Mascis. Optimal real-time traffic control in metro stations. *Operations Research*, 57(4):1026–1039, 2009.

[7] Frank Nielsen and Frank Nielsen. Hierarchical clustering. *Introduction to HPC with MPI for Data Science*, pages 195–211, 2016.

[8] Lothar Richter. Jure leskovec, anand rajaraman, and jeffrey d. ullman. mining of massive datasets. cambridge, cambridge university press., 2018.

[9] Agostino Torti, Marta Galvani, Valeria Urbano, Marika Arena, Giovanni Azzone, Piercesare Secchi, and Simone Vantini. Analysing transportation system reliability: the case study of the metro system of milan.

[10] Mohammed J Zaki and Wagner Meira. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

[11] Junlong Zhang and Yu Luo. Degree centrality, betweenness centrality, and closeness centrality in social network. In *2017 2nd international conference on modelling, simulation and applied mathematics (MSAM2017)*, pages 300–303. Atlantis press, 2017.

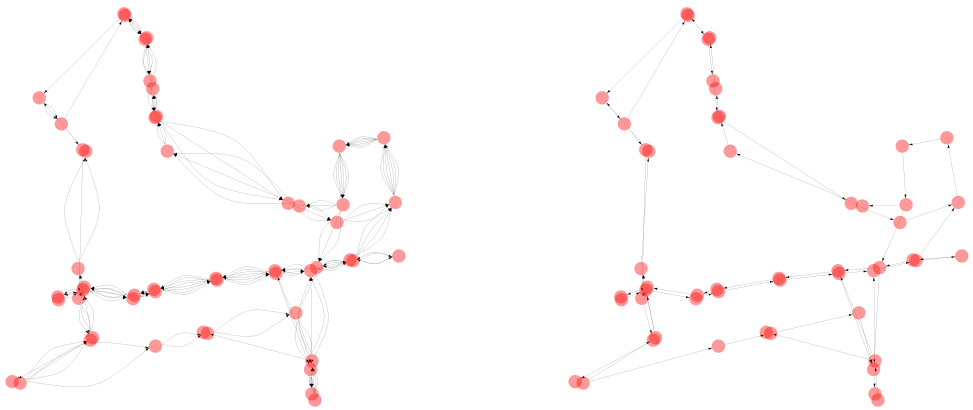
[12] Cheng Zhong, Peiling Wu, Qi Zhang, and Zhenliang Ma. Online prediction of network-level public transport demand based on principle component analysis. *Communications in Transportation Research*, 3:100093, 2023.

[13] Dmitry Zinoviev. Complex network analysis in python: Recognize-construct-visualize-analyze-interpret. *Complex Network Analysis in Python*, pages 1–200, 2018.

A. Appendix A

Figure 13: *Multi-edge network*

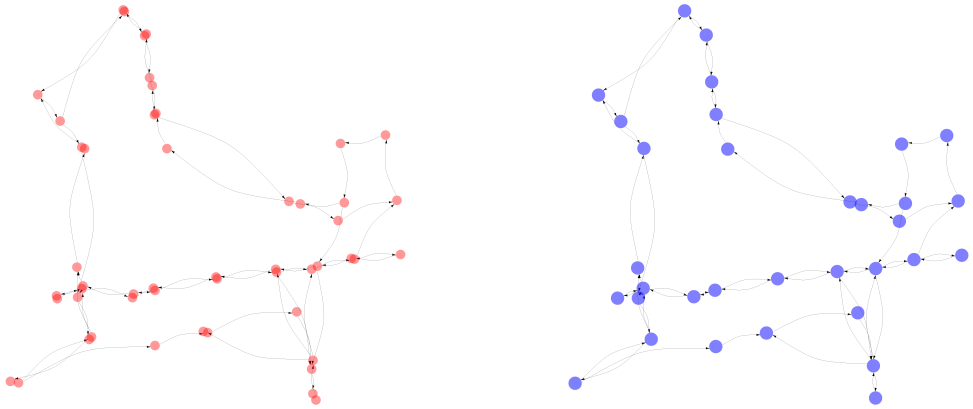
B. Appendix B



(a) Multi-edge network, detail

(b) Single-edge network, detail

Figure 14: *Multi-edge and Single-edge networks, detail comparison*



(a) Original network, detail

(b) Clustered network, detail

Figure 15: *Original and Clustered networks, detail comparison*

Figure 16: *Clustered Network*

C. Appendix C

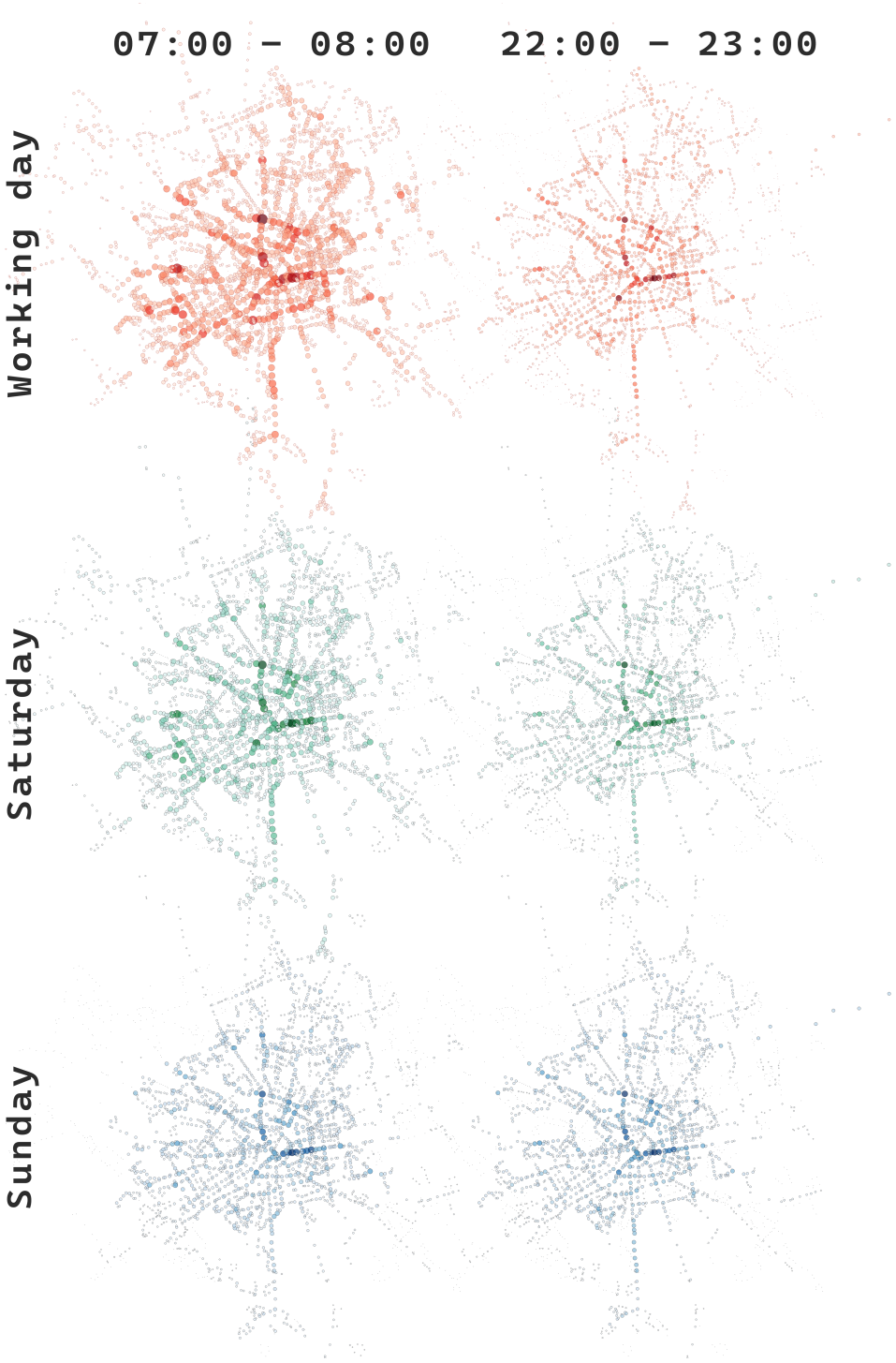


Figure 17: *Average headway values distribution among Macro-stops*

D. Appendix D

Figure 18: *Saturday: Potentiality per trip - Headway subplot. Each plot refers to a time slot. Along the X axes there is the headway; along the Y axes there is the out strength per trip. Dimension and colour of points indicate their Potentiality; position in the graph instead indicates the type of service.*

Figure 19: *Sunday: Potentiality per trip - Headway subplot. Each plot refers to a time slot. Along the X axes there is the headway; along the Y axes there is the out strength per trip. Dimension and colour of points indicate their Potentiality; position in the graph instead indicates the type of service.*

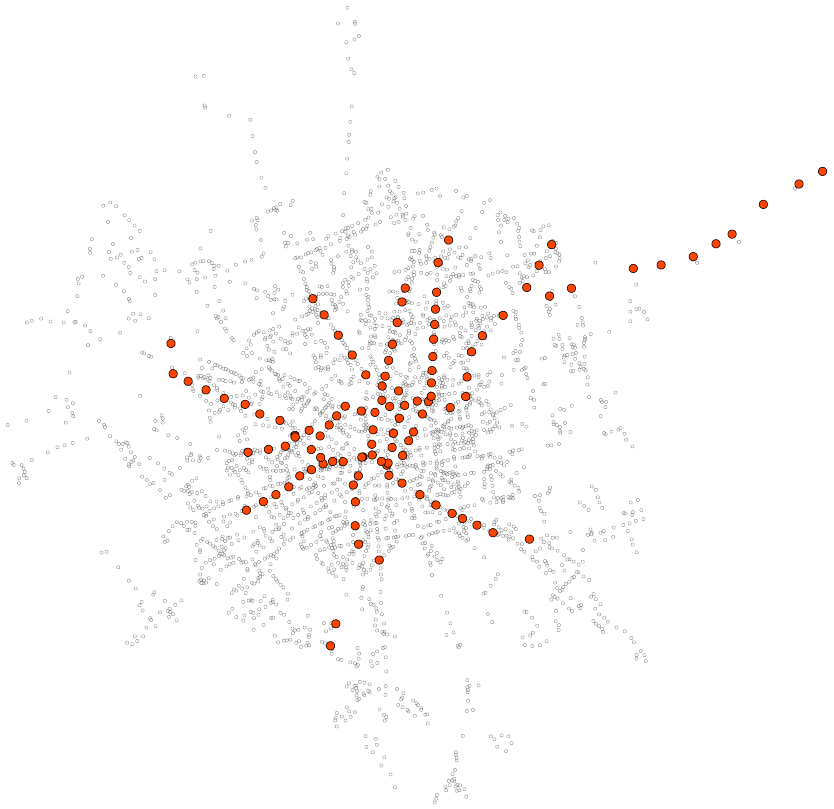


Figure 20: *Metro stations distribution*

Abstract in lingua italiana

L'offerta di trasporto pubblico a Milano consiste in due reti di trasporto principali: la rete metropolitana e la rete di superficie, entrambe gestite da *Azienda Trasporti Milanesi S.p.A.* (ATM). Vista la scarsità degli studi riguardo quest'ultima, questo lavoro la considera come oggetto di studio, in particolare considerando solamente i dati relativi all'offerta. Le metodologie adottate sono quelle dell'analisi delle reti (Network Analysis), al fine di analizzare la mobilità urbana. I risultati potranno poi essere sfruttati dalla stessa ATM, che eventualmente potrà prendere dei provvedimenti per migliorare il servizio offerto agli utenti.
Partendo dai dati in input, che consistevano in un file GTFS relativo ad una settimana standard, è stato possibile costruire la rete iniziale, costituita da nodi, edge e informazioni aggiuntive come linee e percorsi passanti da ciascun edge. I nodi della rete costruita sono stati poi sottoposti ad un processo di clustering: questo è stato fatto semplicemente raggruppando tutte le fermate vicine tra di loro per rendere la rete più simile a come viene percepita dagli utenti. Ad ogni edge della rete è stato poi assegnato un peso sulla base del numero di posti disponibili che vi transitavano in un determinato periodo di tempo. Sono state poi fatte delle rappresentazioni della rete pesata per poter comprendere meglio come l'offerta si distribuisse geograficamente e come variasse durante la settimana. Questi pesi sono poi stati usati per definire degli indicatori per valutare le performance del servizio offerto.
Come ultime considerazioni, vista l'importanza del trasporto intermodale e siccome è sempre ATM che gestisce il servizio, sono state considerate le stazioni della metropolitana per poter relazionare le performance di superficie rispetto alla loro posizione: in particolare, le fermate di superficie sono state classificate sulla base della distanza dalla fermata della metro a loro più vicina, e poi le performance sono state paragonate, per capire se ci fosse una buona integrazione tra i vari servizi di trasporto offerti.

Parole chiave: Rete di trasporto pubblico, analisi delle reti, offerta di mobilità, performance di servizio, trasporto intermodale

Acknowledgements

Scrivendo questi ringraziamenti mi sono reso conto di come il lavoro di tesi non rappresenti solo la fine dei due anni di magistrale, ma bensì chiude l'intero capitolo della vita legato allo studio. Non ho raggiunto questo traguardo da solo, non ce l'avrei mai fatta, ma è solamente grazie a tutte le bellissime persone che ho incontrato durante questo percorso che è stato possibile.
Ringrazio innanzitutto il Prof. Simone Vantini, che ho avuto il piacere di avere non solo come relatore di questo lavoro di tesi, ma anche come docente di un corso. È stato lui a trasmettermi l'entusiasmo e l'interesse verso la Network Analysis, e poi successivamente a coinvolgermi nella realizzazione di questo progetto di tesi. Le sue numerose riflessioni e i suoi apprezzamenti sul lavoro svolto sono state fonte continua di motivazione durante tutto il lavoro.
Ringrazio poi anche la mia correlatrice Arianna Burzacchi, fondamentale per la buona riuscita di tutto il progetto e che, grazie ai suoi preziosi consigli e alle numerose revisioni, ha dato un contributo importantissimo.
Un ringraziamento va anche a tutto il team di ATM, in particolare agli Ingegneri Andrea Mazzola, Marco Pivi e Maurizio Vazzana di ATM, per tutti i dati necessari al progetto.
Non voglio poi fare classifiche..., ma la prima delle altre persone da ringraziare è sicuramente Pietro, il numero 1 del corso di Mobility, nonché mio Presidente. Ci siamo conosciuti durante l'ultimo anno di triennale, e la nostra prima collaborazione risale a quell'ormai noto "progetto di Trucco". Durante la magistrale ho avuto il piacere di essere sempre tra i tuoi compagni di progetto (tranne per una volta, mannaggia), e letteralmente non avrei mai potuto fare questo lavoro di tesi senza il tuo contributo decisivo. Per questo ti ringrazio.
Altro ringraziamento importante va a Matteo, sempre disponibile per una consulenza e un consiglio e sempre sul pezzo su qualunque cosa, davvero qualunque cosa, anche se non era di tuo interesse o competenza. Sei stato veramente una salvezza.
Ringrazio poi tutti i compagni con cui ho avuto il piacere e la fortuna di collaborare, e che ho anche potuto frequentare al di fuori delle aule. Non posso non fare i nomi di Carlotta, per la tua ospitalità e generosità, per le serate Milanesi e San Benedettesi e per le partite di padel, AndreG, per le altre partite di padel, LucaP, Ema, l'altro LucaP e PietroMan.
Un grazie speciale anche a tutti quelli del gruppo Mobuility, che tra pranzi al Date, aperitivi, cene e partite di "Lupus in Fabula" avete reso molto più piacevole questi due anni di università.
Grazie anche a tutti gli altri miei amici. Non ci sarà più la scusa del "progetto" o della "presentazione".
Infine, ringrazio la mia famiglia, mia mamma, mio papà e mia sorella, che mi hanno supportato in tutti questi anni di studi e che, a suo tempo, mi hanno incoraggiato ad intraprendere il percorso universitario della magistrale, senza il quale non avrei conosciuto molte splendide persone.
Ora la pacchia è finita. È finita per davvero. Bisognerà cominciare a lavorare.
Tanti magari si stanno ancora chiedendo che cosa abbia effettivamente studiato in questi due anni di magistrale. Bhè, pensate a questo: se sarete su un treno in ritardo, o se sarete imbottigliati nel traffico allora NON sarà colpa mia, ma se invece quel treno sarà in perfetto orario e se il traffico della city sarà scorrevolissimo, allora sarà tutto merito del vostro Mobility Engineer preferito.

Andrea