POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

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

Laurea Magistrale in Mobility Engineering - Ingegneria della Mobilità

Author: Andrea Parravicini

Advisor: Prof. Simone Vantini

Co-advisor: Arianna Burzacchi

Academic year: 2022-2023

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, both managed by *Azienda Trasporti Milanesi S.p.A.* (ATM).
For what concerns the metro network system, many studies have already been developed, 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 onboard. This is though a transitioning phase, since in the future all vehicles 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.
Furthermore, since the metro system is always managed by ATM, 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 [1–4] were followed 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 work done and the results obtained could be then considered as a starting point for deeper and future analysis, including those whose main objective will be that of the analysis of the passengers' flow inside the network (travel demand).

2. Input data and network construction

All the analyses and the reasoning done during the course of the entire 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. Data 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 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.

Figure 1: *Single-edge network*

Figure 1 shows the geographic extension of the network and 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.

2.1. 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 methodology that better fits this purposes is Hierarchical Clustering.

2.1.1 Identification of clustering parameters

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. The number of lines and the number of stops per each path of the lines were considered as initial characteristics of the network, and by considering different aspects of them it was possible to define two possible scenarios: the worst case one, which consists of a higher clustering distance and a lower number of nodes, and that leads to a more distorted network, has been chosen for the analysis.
This scenario, and the corresponding network, presents a clustering distance of 79 meters: this means that stops that are grouped together in the same cluster will be at most 79 meters apart from each other. It is necessary to evaluate this clustering distance 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 reasonable to say that 79 meters is an acceptable value of 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, 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, 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. Just very few cases where this happens were identified, but they did not compromise the clustering process. Clustering process can be concluded. The final network is constituted of 2955 nodes, with maximum distance between stops belonging to the same cluster equal to 79 meters.

2.1.2 Nodes clustering evaluation

This new network will be the one considered for the future steps of the work, and for all the further analyses, and it must be evaluated with respect to the original one.

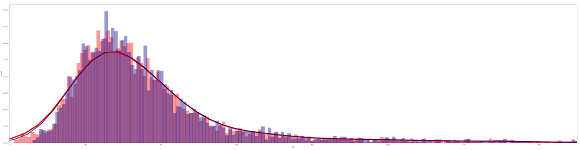

Figure 2: *Edges length distribution*

First, lengths of the edges, and so the distances between the stops have been checked. Figure (2) shows that the distributions of length of edges of the Original Network (red) and of the Clustered Network (blue) are basically the same. This means that the distances between stops have been preserved and that the geographic extension of the network did not change. Also other indicators of the macro-scale analysis have been computed, confirming the goodness of the process. From now on, the "new nodes" obtained will be called *"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. First, hourly time slots were considered for each day, and then the number of trips that each path performs during each time slot has been computed. By combining these information with the additional data about service vehicles provided by ATM, it has been possible to compute the final offer, in order to assign the seat availability as weights to edges, according to the 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.

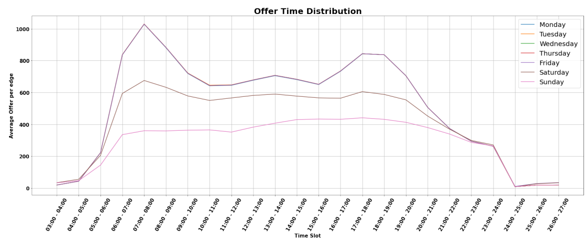

Figure 3: *Offer time distribution*

Figure 3 shows the distribution of the average value of the offer per each time slot and per each day. Three trends can be identified, one for each type of day: the highest one is referred to working day, the lowest to Sunday, and the one in the middle to Saturday. The results obtained follow typical trend of public transport mobility. 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.

In addition to the available offer related to edges, it is also possible to relate it to Macro-stops. In particular, a pretty useful and easy to understand indicator is *headway*, which 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.

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. Indicators computed are 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:

$$Out\ Strength_{Macro-stop} = \sum_{Macro-stop} Outgoing\ Edge\ Offer_{Macro-stop} \tag{2}$$

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 then recomputed to be referred to one single trip of the Macro-stop: this "new" parameter is called *"out strength per trip"*, defined as:

$$Out\ Strength_{Trip} = \frac{Out\ Strength_{Macro-stop}}{Trips_{Macro-stop}} \tag{3}$$

Out strength per trip and headway are suitable to be grouped together, in order to sum up the service performance for each Macro-stop into one single indicator. This second indicator is *Continuity*, defined as:

$$Continuity = \frac{Out\ strength_{Trip}}{Headway} \tag{4}$$

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. *Continuity* is the parameter that will be used in final analyses to summarize the performances of Macro-stops.

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.

In this last analysis the aim is to see how surface performances change (considering *Continuity*) with respect to the distance from metro stations and identify particular situations that might be critical.

If Macro-stops near metro stations have better performances than further ones, the overall service is integrated, allowing an easy and intermodal travel experience to passengers. If the opposite situation occurs, metro system and surface one are practically complementary, and each one of them tries to compensate the lack of performances of the other.

4.1. Metodology and results

To categorize Macro-stops with respect to the distance to their closest metro station, physical distance (euclidean distance) has been computed. Then, also performances values have been considered, to try to relate some possible common trends to the distance. This first draft of categorization needs then to be statistically validated, to get confirmation of the fact that the trends identified were actually valid and truthful. ANOVA test has been done for the statis-

tical validation. If the test confirms the validity of the categories, classification process can finish and some evaluations can be done.
The 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.
The performance trends identified with respect to distance are similar for the three scenarios, 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 $< 79\ m$: ***Neighbourhood***
- $79\ m <$ distance $< 750\ m$: ***Walkable Stops***
- $750\ m <$ distance $< 3000\ m$: ***Challenging Stops***
- Distance $> 3000\ m$: ***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.
These results look at the network in a general way, considering just average values for the categories. Some deeper analyses of the network in the most stressed situation, which is the one during working days in the morning peak hour, should be conducted.
From figure 4 it is possible to notice that there are some Macro-stops not belonging to any Neighbourhood that present pretty high values of performances. Some of them even seem to follow specific paths or directions (directrixes). This configuration can lead to intend those paths as a really valid alternative to metro system. Three main directrixes were investigated.
A pretty clear one with high performances, that seems to be designed as an extension of metro

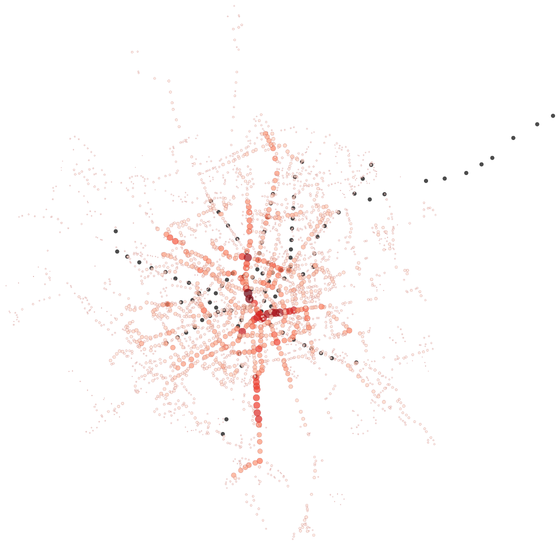

Figure 4: *Continuity values distribution in the network with respect to the metro stations*

line M2, is the one from one of its terminals (*Abbiategrasso*) to approximately Rozzano municipality (in the south area of Milan).
Abbiategrasso can be then considered as a sort of interchange station between underground and surface, and so its neighbourhood must be evaluated: by checking its performances, it shows good characteristics, aligned to the ones of the directrix itself, and so an interchange in that station is viable.
The second directrix, with lower performances, is the one that runs from *Maciachini* metro station (*line M3*) to Niguarda district, in the northern part of the city. Similarly to previous case, *Maciachini* can be meant as an interchange station, but in this case performances are lower than average, and so its neighbourhood may be critical (just from the available offer perspective).
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 con-

cerning the study of high performance Macrostops 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.
From analysis done, it results that Milan is a pretty connected city, with an integrated transportation network that can guarantee to users an intermodal experience of travel. Some criticalities concerning just offer service have been found, and to have a complete vision of how the service actually is, data about passengers' flow must be considered.
Additional data that can further help to obtain better and more precise results are the ones related to the various services in the city, such as schools, universities, supermarkets, hospitals, work offices and green areas.
Lastly, it's important to remark that all the analyses started from a GTFS file, and so methodologies defined are not specific for the network considered, but could be easily adapted to also other transportation networks.

References

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

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

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

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

6. Acknowledgements

This work has been developed thanks to Prof. Simone Vantini and Arianna Burzacchi. Eng. Andrea Mazzola, Marco Pivi and Maurizio Vazzana (ATM) provided all the necessary data. A special thanks goes also to Pietro Mariano, who helped a lot during the realization of a previous similar work.