Discovery of Tourists' Movement Patterns in Venice from Public Transport Data

ABSTRACT

The data collected by public transport tickets has become a valuable source of information for transportation analysis. There are numerous works that analyze them in case studies for subway, bus or train networks, but there are few studies referring to public transport in aquatic environments. In this paper, ticket validation data is used to analyze the movements of tourists in the centre of Venice where waterbuses are the principal public transport. The objective is to analyse the behavior of tourists and detect some relevant patterns. In order to attain this goal, first we build several complex networks which represent the flow of tourists between clusters of stops during different time periods of a day. This allows us to discover some common behaviours of tourists. Once movement flows have been identified, we consider the sequences of validations for each user and we construct a set of trajectories. By applying a hierarchical clustering algorithm, we detect the movement patterns of tourists, identifying which places they visit and in which order. For each cluster we define a representative, that illustrates visually the main routes followed by the tourists. This can represent a valuable information for the decision-maker of the local administration and public transport.

KEYWORDS

Spatio-Temporal Analyses, Hierarchical Clustering, Complex Networks, Public Transport, Trajectories

ACM Reference Format:

. 2022. Discovery of Tourists' Movement Patterns in Venice from Public Transport Data. In *Proceedings of ACM SAC Conference (SAC'22)*. ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/xx.xxx/xxx_x

1 INTRODUCTION

Venice is located between the Adriatic Sea and the Po' Valley and built on an archipelago of 117 islands in a shallow lagoon served by 177 canals with land areas connected by 409 bridges. The city is divided into the historical centre, the mainland urban settlement of Mestre (270,000 inhabitants) and the industrial district of Porto Marghera. The historical center is the largest urban car-free area of Europe, with passengers, cars and heavy vehicles moved by boats and ferries on the larger canals. The main passenger modes are motorized waterbuses (vaporetti) and private taxis, which cover regular routes along the major canals and between the city's islands, transport boats (moto-topo) and private boats. ACTV is the main public transport company that counts 151 waterbuses. It carries 145 million passengers a year on the Navigation network. It has more than 120 floating stations (jetties) and 27 well-connected lines. The bus network consists of 95 routes and the fleet is composed by 568 buses and 20 trams. The number of passengers is around 70 million a year. The company provided us a dataset containing the ticket validations in the period from September 11 2018 to November 12 2018.

In this paper, we focus on tourists with the aim of investigating their movements along the day, detecting their common itineraries and understanding how they organize their visits along the day. There are significant differences that may affect the travel behaviour of tourists compared to other cities, such as the distribution of stops, which are concentrated along the border of the islands or along the larger canals, instead of evenly spread around the city. The difficulty of travelling between areas not connected by land and the limited availability of alternatives to public transport are also considered special restrictive factors of this case of study.

The available dataset consists of the ticket validations of the users of the public transport and in particular we consider the *time limited tickets* which are usually bought by tourists since they allow them to move around for one day, two days, three days or seven days. We perform a two-step analysis. First, we provide a general descriptive analysis of the overall movements of tourists around the city centre and in the main islands. To accomplish this task we group the stops into clusters by using k-Means and Gaussian Mixtures. These clusters become the nodes of complex networks and the edges are defined by considering the tourist movement between the clusters. Moreover, a weight is associated with these edges denoting the number of people crossing them. We create four complex networks, each corresponding to a time period of a day, in order to model what happens in the morning, in the afternoon, in the evening and at night. This representation allows us to determine the main departure and destination points of the tourist trips. After this analysis, we focus on the one day tickets, and for each user we construct the sequences of the stops s/he traverses. By using a hierarchical clustering algorithm we detect the most frequent routes followed by tourists. The visual patterns we obtain give an insightful model of the movements along the different places of the area of interest.

This paper is structured as follows. Section 2 describes the related work concerning the use of public transport data to analyse the movement of people. Section 3 illustrates the dataset provided by the public transport company ACTV. Section 4 presents the general descriptive analysis based on the creation of different complex networks. Section 5 illustrates the detection of the most frequent routes followed by tourists. Finally, Section 6 draws some concluding remarks and illustrates possible future developments.

SAC'22, April 25 –April 29, 2022, Brno, Czech Republic

ACM ISBN 978-1-4503-8713-2/22/04…$15.00
https://doi.org/xx.xxx/xxx_x

2 RELATED WORK

The development of new technologies has made the implementation of smart card systems for public transport in many cities around the world possible. Among other things, the use of these cards in conjunction with the validation or recharge systems allows for a large volume of data to be collected. For instance, the system can register when a user makes use of the public transport by means of a process of validation [9]. The data, collected when a user validates the smart card, includes the validation time, the stop, or the place where it has been validated and the type of used card. This allows one to establish the spatial and temporal position of an anonymous user. One of the disadvantages of these systems, nevertheless, is that they usually do not register when the user leaves the public transport network.

By using these data, several types of studies can be carried out, being one of the most common ones the modelling of complex networks for the analysis of movements of the population. Another use of this data is the prediction of events such as traffic congestion or movement patterns. Another use can be the analysis of the movement patterns of users and how they vary over time or according to external factors [2]. A further explanation of different smart card data mining problems on public transport can be found in [5].

Focusing on the problem of identifying behavior patterns, several problems can be found in the literature. The first would be to identify the type of user, for example distinguishing between a resident user or a tourist like in [15], where the behavior of tourists in Singapore with machine learning techniques is studied. They identify tourists by the type of ticket they buy and calculate the attractiveness of the different stops for the tourists. Lastly, applying classification techniques, they extract the behavior pattern of the tourists. A related study is carried out in [10] where the different profiles of users of public transport in London are identified through clustering techniques and a spatial and temporal analysis is made.

Once the different users have been identified, the following challenge is to extract the behavior patterns of the different groups of users. In [6] both individual and collective mobility patterns in Shenzhen are analyzed, by means of a space-time study. In [16] authors analyze the cases of Singapore, Beijing and London metro to study the variability in the regularity of temporal patterns.

There are cities that, due to their orography, have rivers or maritime public transport, but the user pattern behavior in this case is not widely studied in the literature, the main contribution is the analysis of the patterns of ferry passengers in Brisbane [12].

3 AVAILABLE DATASETS

ACTV provided a dataset containing the validations of different kinds of tickets. A single validation consists of the following pieces of information: (1) Date and hour of the validation; (2) Serial number of the user; (3) Code_Profile; (4) Code of the stop in which the ticket has been validated and (5) Name of the stop. The Serial number of the user does not correspond to the ID_Client stored. This is due to privacy reason. ACTV replaced the ID_Client with a random number keeping the same number for the same user if s/he validated more than once in the period of interest. The geographical coordinates of the public transport stops are also provided by ACTV.

The dataset consists of 4,876,778 records from September 11, 2018 to November 12, 2018. On the dataset described, two cleaning operations have been performed. First, the records of ticket validations whose location was not registered are discarded (5.40% of the total). Then, duplicated records, i.e., validations whose serial number, date and time appear more than once in the dataset, are also eliminated (4.15% of the data remaining after the previous operation). After cleaning erroneous or incomplete records, the total number of validation records is 4,422,029 and the number of distinct users is 551,962. From the remaining data, we selected the ticket type having a duration between one day and seven days. These are the most common tickets bought by tourists, since the residents of Venice usually buy single tickets or monthly or annual tickets, which are not included in the dataset.

The tickets are distributed in the following way according to the different durations included in Table 1.

Ticket Type	Validations	Trips	Users	Val./Trip
24 hours	1,318,989	503,350	287,356	2.62
48 hours	722,195	284,937	98,869	2.53
72 hours	913,062	348,776	96,843	2.62
7 days	86,708	37,579	5,807	2.31

Table 1: Statistics according to the ticket type

In Table 1, the number of trips differs from the number of users because we make divisions by day (from 00:00 to 23:59). Therefore, 75.17% of the users with 24-hour tickets are divided into 2 routes. Since 24 hour tickets show the highest concentration of total validations and the highest ratio of validations per trip, the study is focused on this category of ticket duration. Finally, data from the Venice city center area and the most touristic islands have been selected, leaving aside nearby places reachable by public transport such as Mestre or Padua.

Due to their high use ratio, these tickets provide a sufficient amount of information about the user's movements to complete valuable analyses. With each ticket validation, the exact stop and the time where the user has stamped the ticket are clearly determined. Although, the exact destination of the user is unknown, it can be assumed that it will be close to the next validation of that same ticket. In Venice, many of the islands are not accessible by walking, therefore this can be considered an accurate approximation. Therefore, the daily route of a user can be deduced with the exception of the final destination. As an estimation, the first stop used by the same user on the next day, will be used as the final destination of each day, if this information is available. This hypothesis has been studied in the commuters of the New York subway in [1], and has been proven to be true in 90% of the cases. More recent works on the problem of estimating the final destination of the traveler can be found in [4, 14].

Figure 1 shows the distribution of ticket validations along each hour of the day and each day of the week. As expected, there is more tourist traffic on weekends and the peak hours are located around 10:00 and 15:00.

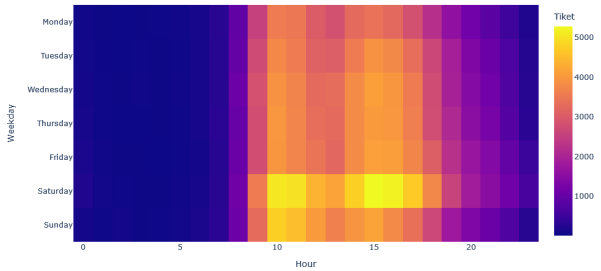

Figure 1: Heatmap: Mean number of tickets by day of the week and hours

4 DESCRIPTIVE ANALYSIS

4.1 Clustering

In the analysis carried out, the first step was to divide the city of Venice into areas of interest for public transportation. Clustering techniques are commonly used to group data into natural groups in order to be able to analyze and describe them as a whole by extracting the characteristics common to them. The objective of the analysis is to detect those areas of the city with the greatest flow of passengers, which was achieved by grouping the different public transport stops into meaningful clusters of stops, represented by the geometric center of the group. A problem to cope with is due to the fact that the geographical component of the validation is composed of fixed static stops. This makes it difficult to use density algorithms, because the distance between points corresponds to the distance between stops and the stops tend to have a homogeneous distribution with very similar distances between their coordinates. For this reason it is almost impossible to find a correct cluster distance in algorithms such as HDBSCAN [7] because they tend to either cluster all the data into very few groups or to generate almost one group per stop.

To solve this problem in the distribution of the stops, a hierarchical clustering with two levels has been carried out. In the first level, K-Means [3] has been used with a total of 12 clusters that were trained with the original already filtered set of data. The number of clusters was selected by analyzing the inertia plot [13] and we observed that with a higher number of clusters there was no significant improvement. The second level of clustering was performed only on the data belonging to the Venice and Lido clusters in the first step, since they were proportionally much denser than the others and a more fine-grained division was needed in those areas. In this case, the Gaussian Mixture algorithm [11] with 19 clusters was used, as it was able to better divide the most used stops from others with little traffic, such as the "Lido port" and the "Lido bus". The final set of clusters includes the 19 clusters from the second clustering level plus 4 additional clusters that were kept from the first level, since there was no need to divide these initially identified areas.

4.2 Building Complex Network

In the effort to understand general tourists movements around the city, the next step was to obtain complex networks describing the origin and destination of their trips around the day. Complex networks serve to abstract natural, often everyday processes, and create a network that allows us to analyze how they relate between them [8]. In this case, a directed and weighted network graph has been modelled in order to visualize and analyze an overview of tourists movements in Venice. These networks are characterized by the fact that the relationships between nodes are oriented and, in addition, a weight value is associated with the relationship. We decided to build four different networks in order to see the daily evolution of movements in the city. The time slots used are 8:00 to 12:59, 13:00 to 17:59, 18:00 to 21:59 and 22:00 to 23:59. These slots are determined according to the activity hours detected in Figure 1, using the valley hours as separators. To create these four networks, first an empty network is generated, to which the nodes are added, being these nodes the clusters previously obtained. Then, an iteration is done, processing every route and substituting any bus/vaporetti stop within a given cluster by its cluster center. The result is a set of routes that indicate movement along a series of clusters (or areas of the city) instead of a series of bus/vaporetti stops, which is clearer to see in a map. Finally, another iteration is completed along all clusters of every registered route adding the corresponding edge to the graph with a weight of one if the connection appears for the first time, or incrementing its weight if it already existed. As a result, we obtain as edges of our graphs the count of all the times that a pair of clusters appear consecutively in the routes that compose the whole dataset.

The results of this calculations can be seen in Figures 2, 4, 6, 8, that represent the different number of incoming and outgoing tourists for each cluster in the corresponding time slot and in Figures 3, 5, 7, 9 that illustrate the different movements on the synthesised network in the same time slots. The values in the bar charts represent the percentage of validations in the specific time slot, disaggregated by clusters and distinguishing if the validation is from an incoming or outgoing trip. In the graph visualizations, the same value is represented by the thickness of the graph edges. The area names in the bar charts are ordered according to the id of each node in the network graph figures.

4.3 Description of Results

Analyzing the inputs and outputs of the network nodes, several common behaviors can be identified. It can be clearly observed that in the morning (Figures 2 and 3) the main departure nodes are *Piazzale Roma* and *Ferrovia* which correspond to the bus and train station respectively. It is reasonable to think tourists start their routes there. *Fondamente Nove*, *San Zaccaria* and *Rialto* are the main points of arrival and departure of passengers at this time slot, being intermediate points for tourists' journeys to further destinations. Finally, in this time slot, two of the main destinations for tourists are the islands of *Murano* and *Burano*.

In the afternoon (Figures 4 and 5), the most obvious observation is that most of the trips are between the two islands *Murano* and *Burano*. And while the arrival and departure points remain the same, with small variations, a significant lower intensity of passenger flows is registered. Furthermore, *Piazzale Roma* and *Ferrovia* become in turn arrival points for visiting tourists: comparing Figs. 2 and 4, while in the morning both are the stops with the highest

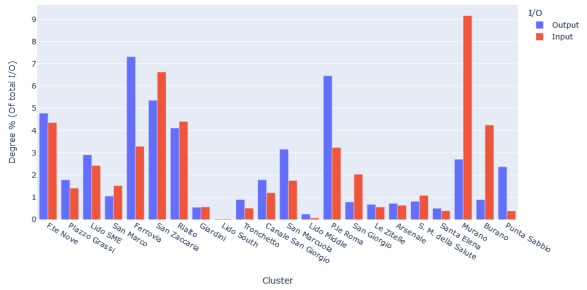

Figure 2: I/O degree values in the morning slot (8:00 to 12:59)

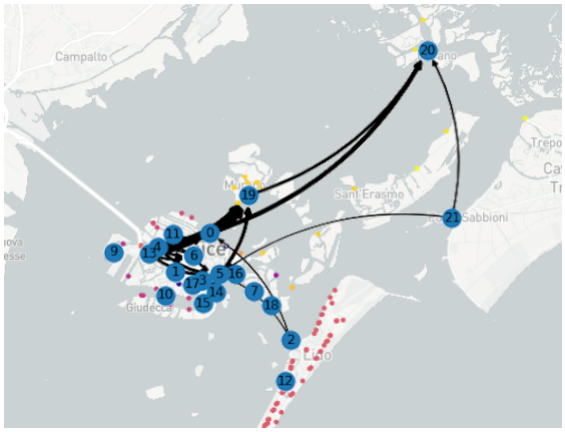

Figure 3: Morning Complex Network (8:00 to 12:59)

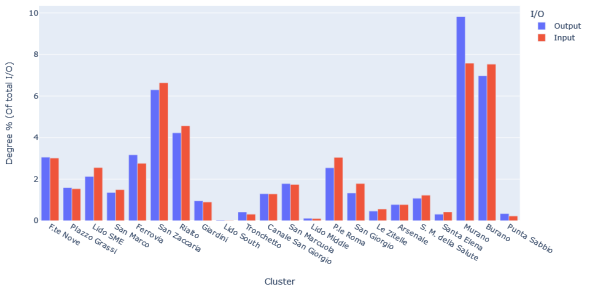

Figure 4: I/O degree values in the afternoon (13:00 to 17:59)

output degree, in the afternoon their input degree becomes higher or similar to the output one.

In the last two slots of evening and night (Figures 6 and 7 and Figures 8 and 9), traffic between *Murano* and *Burano* almost completely disappears and the flow at *Fondamente Nove* is significantly reduced. Besides, *San Zaccaria* and *Rialto* become mainly departure points and this flow of passengers is directed back to the places of origin, like *Piazzale Roma*.

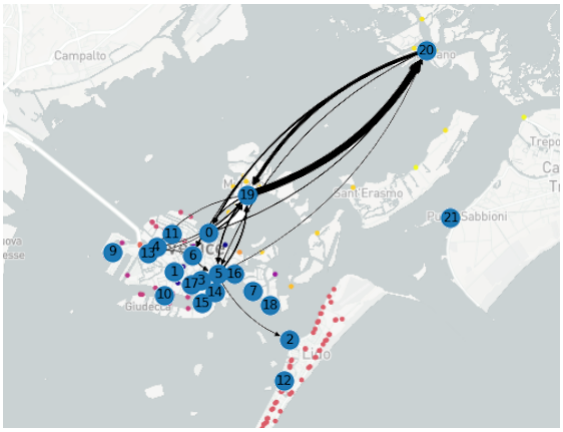

Figure 5: Afternoon Complex Network (13:00 to 17:59)

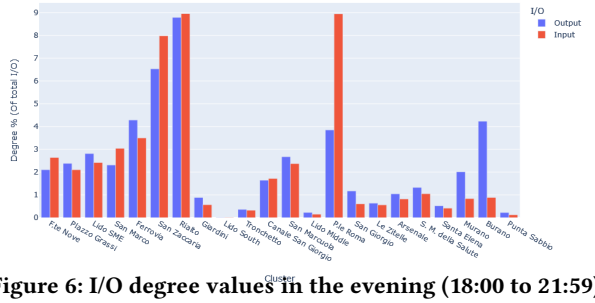

Figure 6: I/O degree values in the evening (18:00 to 21:59)

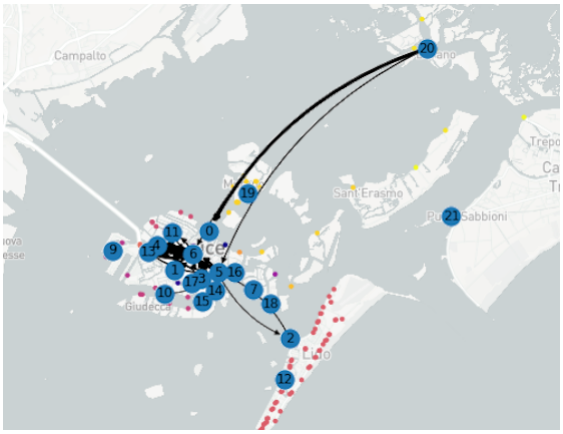

Figure 7: Evening Complex Network (18:00 to 21:59)

5 PATTERN ANALYSES

The second part of the proposed analysis aims at better understanding the movement patterns of the tourists in Venice. This includes not only discerning which areas are more visited in general, but also the order in which these places are traveled.

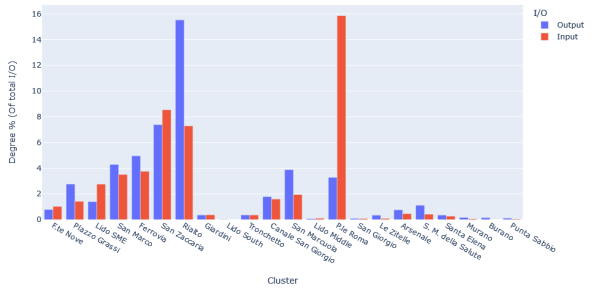

Figure 8: I/O degree values in the night slot (22:00 to 23:59)

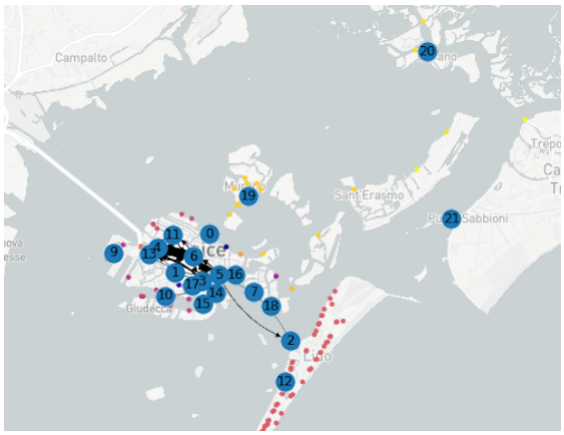

Figure 9: Night Complex Network (22:00 to 23:59)

5.1 Building trajectories

In the GTFS data, the number of stops both for buses and waterbuses is large and this makes it difficult to find meaningful behavior patterns of users. Hence, we reduced this number by merging stops together. We used different criteria to aggregate stops. Since our focus is on navigation service and the number of bus validations is small with respect to waterbus ones, for the bus transport, we built five big areas: Lido, Airport, Piazzale Roma, Mestre (Train Station), Mestre (Urban Area). These areas correspond to different functional areas in the mainland and Lido is an island where there is a bus network.

The stops of the navigation service are 144 but a lot of them correspond to the same landing stage, so we aggregated them into a unique stop like in the following example: S. Zaccaria (Pieta') "A", S. Zaccaria (M.V.E.), "B", S. Zaccaria (Jolanda) "C", S. Zaccaria (Jolanda) "D", S. Zaccaria (Danieli) "E", S. Zaccaria (Danieli) "F" are mapped into the stop named "S. Zaccaria".

The number of navigation stops is reduced to 69. After this aggregation process, from 2500 stops belonging to the bus and the navigation networks, we obtained a set containing only 74 stops.

In order to build trajectories of users, we group together the validations associated with the same serial number, and we obtain a sequence of couples of the form: *(stop, time)*, where *stop* is the identifier of the stop in which the user gets on the waterbus/bus and *time* is the moment when this action takes place. Notice that we replaced the original stops with the ones we created according to the aggregation process we have just described. The sequence of couples is ordered with respect to the time component. Some cleaning operations are done to avoid useless validations.

Stop name	# of users	Percentage	Cumulative rate
ferrovia	5,974	18.52	18.52
p.le roma	4,136	12.82	31.34
s. zaccaria	4,050	12.56	43.9
murano	2,212	6.86	50.76
rialto	2,020	6.26	57.02
TERRA	1,972	6.11	63.13
f.te nove	1,599	4.96	68.09
burano	1,445	4.48	72.57
s. marco	1,148	3.56	76.13
lido	1,072	3.32	79.45

Table 2: Frequent stops for users having a single validation

5.2 Extracting movement patterns

The hierarchical clustering algorithm was used, this time, to determine groups of tourists with the same behavior. We restrict our analyses to one day tickets and we consider only the spatial component of the sequence of validations. Hence our dataset consists of elements of this kind $s_1, \ldots, s_n$ where s_i is the identifier of a stop and the sequence represents the trip of a tourist in 24 hours. The lengths of these sequences range from one single stop to 40 stops even if only few sequences have more than 10 stops. Such a variability makes it difficult to apply the clustering algorithms since the similarity measures do not work properly. This is why we decide to split the dataset according to these lengths and we consider the users with only one stop, with two stops, with three stops and with four or five stops. This subset of trips covers the 82.9% of the trajectories of this type of ticket, which is a significant percentage. Indeed longer sequences significantly increase the complexity of the analysis but they provide very little information on the total use of the routes.

The users that validated only once are 32, 251 and over 70% is concentrated in 10 stops. Table 2 shows the most frequent stops obtained by directly grouping by the stop id. This table gives us insightful information: first of all, *Ferrovia* and *Piazzale Roma* are confirmed to be the main entrance points to Venice as observed in Section 4.3 and the islands *Murano*, *Burano* and *Lido* are points of interest which are largely visited. Moreover, the presence of *TERRA* suggests that a lot of tourists start their trip in the mainland where probably they have their accommodation since it is cheaper than to stay in Venice. *Rialto* is located in the centre of Venice where there are a lot of shops, restaurants and from there it is easy to reach museums and other touristic attractions. *San Zaccaria* is a highly used stop because it is close to San Marco and to the area of Arsenale and Giardini where several events take place, such as Biennale Arte, Architettura. Moreover, it is a connection stop towards the surrounding islands. A similar function is played by *F.te Nove* which is used to go towards Murano and Burano.

We repeated this analysis for the users having exactly two stops and three stops, without using any clustering algorithm but simply

grouping by couple and terns of stops. It is worth noting that users with 2 stops are 65, 510 and the number of distinct routes is 1, 270. However, over the 50% of the movement of these users can be described by only 15 routes as illustrated in Table 3. It is worth noticing that these routes connect the stops we described in Table 2. The only exception is *Sabbioni* which is not present among the top 10 stops for users with a single validation. This stop is located in Punta Sabbioni, a very popular destination for beach holidays. Tourists stay there and they do excursions towards Venice and also to Burano and Murano as detected in Table 4.

Route	# of users	Percentage	% Cumulative rate
murano - burano	4,071	6.21	6.21
sabbioni - s. zaccaria	3,626	5.54	11.75
ferrovia - rialto	2,702	4.12	15.87
ferrovia - s. zaccaria	2,430	3.71	19.58
ferrovia - TERRA	2,280	3.48	23.06
s. zaccaria - burano	2,261	3.45	26.51
f.te nove - burano	2,245	3.43	29.94
p.le roma - rialto	2,008	3.07	33.01
ferrovia - murano	1,999	3.05	36.06
s. zaccaria - murano	1,753	2.68	38.74
p.le roma - s. zaccaria	1722	2.63	41.37
f.te nove - murano	1,430	2.18	43.55
p.le roma - TERRA	1,414	2.16	45.71
ferrovia - s. marco	1,218	1.86	47.57
p.le roma - murano	1,144	1.75	49.32

Table 3: The most frequent routes for users with two stops

The users validating three times are 76, 410 and the different routes are 7, 207. The 15 most frequent routes cover the 25% of users while the 115 most frequent routes describe 50% of the users. In Table 4 the 15 most frequent routes are illustrated.

Route	# of users	Percentage	% Cumulat. rate
(5032, 5063, 5068)	ferrovia - murano - burano	38.36	5.02
f.te nove - murano - burano	3,591	4.7	9.72
s. zaccaria - murano - burano	2,738	3.58	13.3
p.le roma - murano - burano	1,831	2.4	15.7
f.te nove - burano - murano	1,703	2.23	17.93
sabbioni - rialto - s. zaccaria	869	1.14	19.07
s. zaccaria - burano - murano	784	1.03	20.1
sabbioni - burano - murano	712	0.93	21.03
ferrovia - s. zaccaria - murano	680	0.89	21.92
murano - burano - murano	602	0.79	22.71
murano - burano - rialto	538	0.7	23.41
sabbioni - burano - s. zaccaria	527	0.69	24.1
ferrovia - murano - s. zaccaria	504	0.66	24.76
murano - burano - s. zaccaria	494	0.65	25.41
f.te nove - burano - rialto	461	0.6	26.01

Table 4: The most frequent routes for users with three stops

The resulting routes are characterized by a visit to the islands of Murano and/or Burano. Instead, the route *sabbioni - rialto - s. zaccaria* models a tour inside the city centre of Venice.

In order to deal with longer sequences we employ a hierarchical clustering algorithm for sequences and we consider only users with 4 or 5 validations. The number of these users is 105, 000 and we tested the algorithm with different parameters and distance measurements. As distance between clusters, we chose the maximum distance as it returns the best result whereas the minimum or the average distance tended to merge the clusters too much.

In order to evaluate the clustering result we define the concept of a meaningful cluster on the basis of three parameters:

- minimum number of users inside a cluster;
- minimum number of pairs of stops validated consecutively;
- minimum percentage of users of the cluster who have been in the pairs of stops of the previous point.

Distance	250 clusters		500 clusters		750 clusters		1000 clusters	
Edit	11	9,668	26	16,471	37	21,226	42	21,998
Jaccard	13	9,460	18	11,410	22	12,479	23	12,681
$LCSS_{bil}$	16	25,665	19	32,873	54	45,680	70	49,589
LCSS	19	16,529	25	19,882	42	21,479	52	25,425

Distance	1250 clusters		1500 clusters		1750 clusters		2000 clusters	
Edit	46	23,242	48	23,314	50	23,652	51	23,251
Jaccard	27	14,482	29	16,102	31	16,599	33	17,586
$LCSS_{bil}$	77	51,101	75	50,804	77	49,392	75	48,918
LCSS	62	27,091	63	27,202	63	27,115	63	27,091

Table 5: For each number of clusters, the number of meaningful detected clusters and the number of covered data

The previous definition guarantees that a significant cluster is relevant as it contains a minimum number of users, therefore not outliers, and that within the cluster there is a common and shared path between users.

Different distance/similarity measures were used to maximize the number of users associated with a meaningful cluster. We experimented with the similarity of Jaccard, the Edit distance, the Longest Common Subsequence (LCSS) and a combination of the LCSS with the longest common substring defined as the product of these two values. We call the last measure *LCSS balance*, $LCSS_{bil}$ for short, and its aim is to give preference to *continuous* subsequences instead of subsequences with holes. The result of the hierarchical cluster is described in Table 5 and in Figure 10.

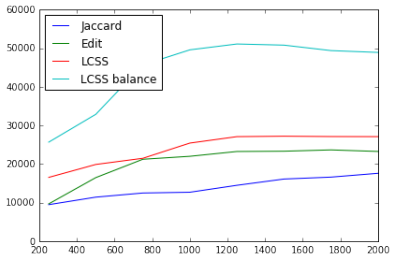

Figure 10: Plot of the covered data by the meaningful clusters. The x axis represents the number of required clusters

The best result is obtained by using $LCSS_{bil}$ with 1, 250 clusters and we succeed in covering almost the 50% of the dataset with 77 meaningful clusters.

In order to find a representative for the clusters, first we considered the user which was less distant from the others on average. Unfortunately in this way the route could include not only the most frequent stops but also some less frequent stops, as shown in Figure 11 by using a heat map. The user crossed the stops in the following order: (1) Ferrovia, (2) Fondamenta Nove, (3) Murano, (4) San Zaccaria This phenomenon is due to the fact that different users coming from different areas merge and they travel together in the central part of their trip and then they split. An example of this behaviour is well illustrated by the Sankey diagram in Figure 12. Hence, we selected only the stops of the representative validated

Figure 11: Issues with the representative of a cluster

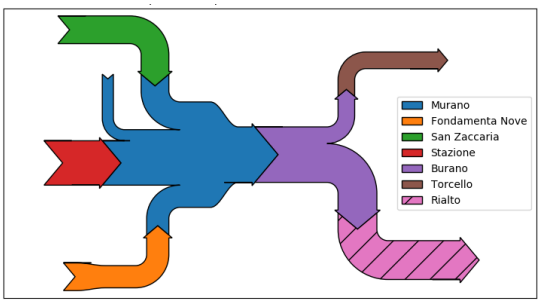

Figure 12: Users of a cluster: different origins, same central path, different destinations.

Figure 13: Resulting representative of a cluster

more than a certain threshold. For the example in Figure 11 the resulting representative route becomes the one in Figure 13, i.e., Fondamenta Nove-Murano-San Zaccaria.

Table 6 reports the top 15 meaningful clusters, with the number of users and its representative.

Let us analyse two of these meaningful clusters, i.e., Ferrovia-Murano-Burano and Murano-Burano-Terra.

The cluster Ferrovia-Murano-Burano covers 2, 623 users and it describes one of the most popular tourist routes. In fact, the tourist arrives at the train station, takes the waterbus and makes the island tour, first visiting Murano and then Burano. Figure 15 shows the different stops while the Sankey diagram in Figure 14 allows us to highlight the flows of the users. The frequency of each route

Id	# of users	representative
124	6,493	('f.te nove', 'murano', 'burano')
597	3,479	('s. zaccaria', 'murano', 'burano')
328	2,980	('p.le roma', 'murano', 'burano')
1214	2,623	('ferrovia', 'murano', 'burano')
696	1,977	('murano', 'burano', 'murano')
270	1,901	('murano', 'burano', 'f.te nove')
138	1,736	('f.te nove', 'burano', 'murano')
400	1,655	('ferrovia', 'murano', 'burano', 's. zaccaria')
1203	1,421	('murano', 'burano', 'rialto')
486	1,346	('sabbioni', 'burano', 'murano', 's. zaccaria')
607	1,323	('ferrovia', 'murano', 'burano', 'murano')
647	912	('p.le roma', 's. zaccaria', 'murano')
112	868	('ferrovia', 's. zaccaria', 'murano')
148	794	('s. zaccaria', 'murano', 's. zaccaria')
648	737	('murano', 'burano', 'TERRA')

Table 6: List of the top 15 meaningful clusters

Route		Percentage
Ferrovia	Murano	97%
Murano	Burano	97%
Burano	Rialto	36%
Burano	Ferrovia	12%
Burano	Torcello	1%
Burano	Lido	0.8%
Burano	S. Marco	0.7%

Table 7: Cluster Ferrovia-Murano-Burano: Route frequencies

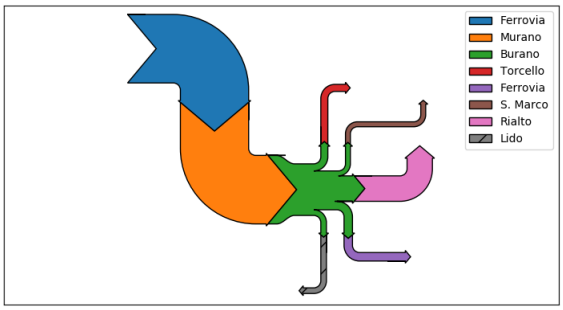

Figure 14: User flows in the cluster Ferrovia-Murano-Burano

is in Table 7. This behaviour matches very well the main routes identified in Fig. 3, where the majority of user trips flow from the stations (nodes 11 and 13) to both islands (nodes 10 and 20).

The cluster Murano-Burano-Terra describes 737 users. As for the previous cluster, Figure 17 illustrates the main and secondary movements among the stops, the Sankey diagram in Figure 16 points out the flows of the users and Table 8 the frequencies along the different routes. This cluster is interesting since it describes users starting their visit from two different stops (i.e., P.le Roma and Ferrovia) and they end their trip into Terra, that is mainland and this means that the tourists' accommodation is not inside Venice but in the mainland. Again, this is consistent with the initial results, in which the highest passenger flow at the evening is from Burano to F.ta Nove (on Fig. 7, node 20 to node 0) and continuing to the railway station (Ferrovia) and bus stop hub (P.le Roma) (in Figs. 7 and 9 nodes 11 and 13, respectively) to leave for their accommodations.

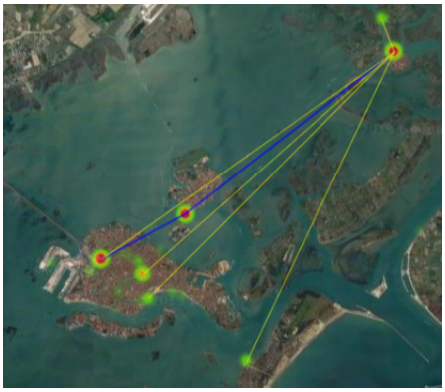

Figure 15: Trajectories in the cluster Ferrovia-Murano-Burano (main routes in blue, secondary ones in yellow)

Route		Percentage
Terra	P.Le Roma (vaporetto)	1%
Terra	Piazzale Roma (bus)	0.6%
P.Le Roma	Murano	46%
Ferrovia	Murano	45%
Murano	Burano	99%
Burano	Terra	98%

Table 8: Cluster Murano-Burano-Terra: Route frequencies

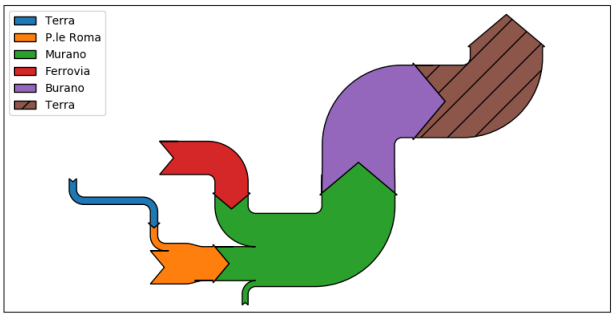

Figure 16: User flows in the cluster Murano-Burano-Terra

Figure 17: Trajectories in the cluster Murano-Burano-Terra (main routes in blue, secondary ones in yellow)

6 CONCLUSIONS AND FUTURE WORK

The city of Venice offers a very peculiar context, since a large portion of the trips are made by waterbuses (vaporetti), making this case study quite different from other similar studies in other cities in the world. We presented an analysis of the behaviour of tourists in the city of Venice by exploiting the ticket validations on the public transport network.

We attained two main goals: the first is a general description of the movements of tourists in the different areas of the city and nearby islands at various time periods in a day. The second is the discovery of patterns highlighting the most frequent routes followed by tourists. Both analyses are complementary and the results obtained are consistent, corroborating the soundness of the conclusions. These are potentially valuable insights for the city council, transport companies and museums and cultural managers.

As future work, we would like to investigate how and whether the patterns vary during the day. Moreover, it would be interesting to analyse the behaviour of the tourists in a period larger than one day, i.e., to detect the plan of the visits in different days. Additionally, we intend to obtain new datasets recording validations in different periods of the year in order to compare tourists' preferences on diverse weather conditions and in presence of specific events like the International Film Festival. Finally, it would be challenging to analyse also the validations of the monthly and yearly tickets in order to detect the patterns of residents and to try to classify people by distinguishing among workers, students and tourists.

REFERENCES

[1] J. Barry, R. Newhouser, A. Rahbee, and S. Sayeda. 2002. Origin and destination estimation in New York City with automated fare system data. *Transportation Research Record* 1817, 1 (2002), 183–187.
[2] A.-S. Briand, E. Côme, M. Trépanier, and L. Oukhellou. 2017. Analyzing year-to-year changes in public transport passenger behaviour using smart card data. *Transportation Research Part C: Emerging Technologies* 79 (2017), 274–289.
[3] Emre Celebi. 2015. *Partitional Clustering Algorithms* (1st ed. 2015. ed.). Springer, Cham. 79–98 pages.
[4] L. He and M. Trépanier. 2015. Estimating the destination of unlinked trips in transit smart card fare data. *Transportation Research Record* 2535, 1 (2015), 97–104.
[5] Tian Li, D. Sun, P. Jing, and K. Yang. 2018. Smart card data mining of public transport destination: A literature review. *Information* 9, 1 (2018), 18.
[6] Liang Liu, Anyang Hou, Assaf Biderman, Carlo Ratti, and Jun Chen. 2009. Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen. In *12th IEEE Conference on Intelligent Transportation Systems*. IEEE, 1–6.
[7] L. McInnes, J. Healy, and S. Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software* 2, 11 (2017), 205.
[8] González M. C. Menezes R., Evsukoff A. 2013. *Complex Networks*. Springer.
[9] K. Mohamed, E. Côme, L. Oukhellou, and M. Verleysen. 2016. Clustering smart card data for urban mobility analysis. *IEEE Transactions on intelligent transportation systems* 18, 3 (2016), 712–728.
[10] M. A. Ortega-Tong. 2013. *Classification of London's public transport users using smart card data*. Ph.D. Dissertation. Massachusetts Institute of Technology.
[11] M. Ouyang, W.J. Welsh, and P. Georgopoulos. 2004. Gaussian mixture clustering and imputation of microarray data. *Bioinformatics* 20, 6 (01 2004), 917–923.
[12] A. Soltani, M. Tanko, M. I Burke, and R. Farid. 2015. Travel patterns of urban linear ferry passengers: analysis of smart card fare data for Brisbane, Queensland, Australia. *Transportation Research Record* 2535, 1 (2015), 79–87.
[13] M A Syakur, B K Khotimah, E M S Rochman, and B D Satoto. 2018. Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. 336 (apr 2018), 012017.
[14] M. Trépanier, N. Tranchant, and R. Chapleau. 2007. Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems* 11, 1 (2007), 1–14.
[15] M. Xue, H. Wu, W. Chen, W. S. Ng, and G. H. Goh. 2014. Identifying tourists from public transport commuters. In *Proc. of the 20th ACM SIGKDD intern. conference on Knowledge discovery and data mining*. 1779–1788.
[16] C. Zhong, M. Batty, E. Manley, J. Wang, Z. Wang, F. Chen, and G. Schmitt. 2016. Variability in regularity: Mining temporal mobility patterns in London, Singapore and Beijing using smart-card data. *PloS one* 11, 2 (2016).