Visualization and Clustering of Passenger Flow in Taipei MRT Stations

Final Paper

A05227102 Jana Wilbert, B01106047 Yayam Su, B02902071 Po-Yao Chen

1. Introduction

Navigating through a large city for both tourists and residents often means covering long distances and trying to not get stuck in traffic holdups. Besides an own or rental bike or car, the metro system is one of the most popular public transportation services in many large cities all over the world. In many cities, the metro system consists of a wide network of metro lines with the goal to connect peripheral areas with the center as well popular central places with each other. As a city grows, the metro system continuously adapts as well, for instance, by creating new lines. A metro system also adapts to a higher demand of transportation by offering more wagons or trains in smaller temporal intervals in rush-hour times, for example.

To identify different crowd flow patterns in metro systems to plan more precisely and adapt to changes in the city on the one hand but also to react to dangers arising by a high number of people on the other hand, it can be useful to visualize and classify crowd flows. Also for passengers this can be a helpful tool to get an idea of their own city's transportation system and when and where to avoid the peak-times of high crowded trains.

In Taipei with its around 2.7 million inhabitants[1] and an accelerated population and economic growth after around 1950 (Lin & Shin, 2008), the Rapid Transit System has been developed in the 1990s and started commercial service with the first line 1996 in order to decrease car traffic and improve the air quality in Taipei (Lin & Shin, 2008; Wang, 2016). Meanwhile a system length of around 131km and more than 0.7 billions passengers in 2015 can be pointed which represents Taipei among the 20-25th top cities regarding ridership and system length worldwide (Wang, 2016). Therefore, it seems likely to apply visualizations and classifications which have been

[1] Information for 2016, according to https://en.wikipedia.org/wiki/Taipei#cite_note-3, retrieved on 02 January 2017.

done for metro systems of other cities to the Taipei metro system as well. Hence, after having introduced different types of visualization cases of passenger flows of various subway system in the Midterm Report, this report focuses, more specifically, on the passenger flow of Taipei metro and its possible visualization and station classification.

In the following, after a general review of aspects of crowd flow in metro systems and classification of subway stations in general (2.), we will, therefore, decribe methods (3.), data observations (4.) and visualization results (5.) of our own visualization project of Taipei metro passenger flow. The design as well as future work will then be discussed (6.).

2. Literature Review

Regarding Taipei during rush hour, a daily average of 30 million passengers transit - a number which indicates the necessity to control crowd flow, timing and location of persons and trains. Therefore, one issue of transportation systems is to flow past each other while avoiding collision, delays or crowding (Wen, 2013). Moreover, besides representing urban efficiency, popular traffic transportation and a converging point for cultural exchange, dangers such as high density crowds, fire, spread of infectious diseases or terrorist attacks can be situated in metro transportation as well. Understanding the passenger's dynamic effectively could reduce potential danger (Wen & Chang, 2014).

Visualization and simulation of train and passenger flows is one technique to examine and understand influences on and unusual changes of crowd flow behavior in metro systems (Itoh et al., 2013; Wen, 2013). Compared to wide tables of data, location, direction and timing of flows can be presented in a vivid and demonstrative manner for experienced as well as uninformed viewers. Questions like `Where is the largest passenger flow?´, `When is the largest passenger flow?´ and `How to avoid the crowded flow?´ can be answered in an illustrative way. A common type of visualization is, thereby, flow maps which are a type of thematic map to show

movements of objects or people between different locations (Briney, 2014; Phan, Xiao, Yeh, Hanrahan & Winograd, 2005).

So visualization in general and for crowd flow, more specifically, flow maps can help to illustrate movements of goods between areas, migration patterns as well as traffic volume and stream flow. It allows cartographers, analysts and map users to easily access information about different magnitudes of various objects across space (Briney, 2014; Phan et al., 2005). Regarding crowd flow in traffics, for example, it allows commuters and city planners to see traffic patterns and act accordingly.

Typically, lines are used to show movement of people and objects between locations in flow maps. The width of lines can be varied to illustrate the magnitude or quantity of flow with a wider line usually indicating a higher traffic or flow (Simantel, 2012).

In flow maps both qualitative and quantitative data can be used. Whereas in quantitative flow mapping (e.g. about magnitudes) mostly line symbols with variable width depending on the data values are used to indicate changes in magnitude between areas, symbols of uniform width and arrows are typically used to indicate movement in qualitative flow mapping (e.g. about connections; Briney, 2014; McGraw-Hill, n.d.).

In addition, depending on the nature of data, different types of flow maps can be applied. Firstly, data in flow maps can be still, like in traditional printed maps, or it can be animated. By computers animated, interactive maps can not only visualize direction and magnitude of movement but also speed, for instance (Briney, 2014). Secondly, network, radial or distributive maps can be distinguished. Network flow maps show interconnectivity between places (Buckley, 2013) and the quality of flow on an existing network (Simantel, 2012). Therefore, they often present transportation or communication networks (Briney, 2014; see, for example, Case 1 introduced in the midterm report: ‘A day in Taipei metro’ (Appendix A)). Contrary, radial flow maps illustrate relationships between one source and many destinations (Simantel, 2012). Separate lines radiating out from the origin (or destination if data is vice versa which means many starting points and one destination) create a spoke-like pattern

(Buckley, 2013). Disruptive flow maps show the distribution of goods or other flows which start at one or few origins and end at multiple destinations (Buckley, 2013) as in radial flow maps with the difference that they often start with a single line at the origin which then divides into many smaller lines once they (almost) reached the destinations (Briney, 2014). Nevertheless, not only flow maps are used to indicate crowd flow. As it was demonstrated in the midterm report, heat maps or line charts could be useful to illustrate aspect of crowd flow as well (e.g. Itoh et al., 2013). Informations like the passenger flow change during one day as well as peak-times of passenger flow for one specific station, for example, can be displayed informativly with it (see for example Case 4 introduced in the midterm report: One day metro data in Singapore (Appendix B)), Case 5 introduced in the midterm report: Finchley central station weekday's passenger quantity (Appendix C)) and Case 6 introduced in the midterm report: Tokyo Metro Ginza Line (Appendix D)).

Based on the evaluation of these cases in the midterm report as well as the works by Itoh and colleagues (2013), for instance, it seems that different types of visualziation might be useful depending on which issues to solve and which data to visualize. Geographical flow maps could be suggested to show visualizations regarding the problems `Where is the largest number of passengers?´ as well as `Where do passengers go?´. Using line charts could the focus could be on `When is the largest quantity?´ and `What is the pattern?´. In contrast, heat views could demonstrate `Which lines have largest crowd flow number?´ and `What happened in a certain period of time?´. As demonstrated by Itoh and colleagues (2013) also a combination visualizations can be used.

In addition to visualizations, subway stations can be classified based on differences in passenger flows. For Beijing subway literature on classifying metro stations and different methods are already discussed (Wei, Lin, Chu, Tian & Fei, 2016). Stations can, for instance, simply be classified based on the passenger flow volume. More complicated, stations can be classified based on the combination of passenger flow and the service function of regional environment so that, for instance, the categories 'travel', 'leisure gatherings' and 'daily commuting' were developed in which stations were then graded into different level. Based on static (geographical

location and surrounding facilities) as well as dynamic informations (passenger flow) stations and its functions can be distinguished and classified (Wei et al., 2016).

In our project, we engaged in visualizing the metro passenger flow of Taipei, on one hand, by using different types of visualizations (network maps and line charts) to illustrate the flow as informative and coherent as possible. On the other hand, based on the data and first visualizations, we tried to classify the metro stations into clusters depending on the characteristics of the passenger flow which we then used for in our visualizations again. In the following, we will describe the process of data collection and classifying as well as the visualization results.

3. Methods

3.1. Data Source and Data Preprocessing

Taipei MRT stations data including the passenger flow checking in and checking out in each station is shown in figure 1. The data is referenced from Taipei Open Data Platform (臺北大眾捷運股份有限公司 劉亦昌 2016). Considering the holiday days to be less than in other months, we choose the whole 2016 April one month data as our data sample for reference and analysis. The data includes passenger flows checking in and out in 121 stations in Taipei for every day and every hour in April 2016.

日期	時段	松山機場	中山國中	南京復興	……	日期	松山機場	中山國中	南京復興	……
2016/4/1	5	11	42	65	35	2016/4/1	7,265	17,662	38,162	54,106
2016/4/1	6	70	349	346	409	2016/4/2	5,213	10,992	20,439	50,926
2016/4/1	7	301	1,201	1,045	1,007	2016/4/3	4,609	9,088	16,880	45,574
2016/4/1	8	393	1,766	1,600	1,430	2016/4/4	5,020	8,170	14,966	42,329
2016/4/1	9	291	823	970	1,131	2016/4/5	5,809	8,788	16,669	43,172

Figure 1. Example of the data.

For preprocessing this data, we first converted the data from “BIG5 encoding” to “UTF8 encoding” in order to unify the character encoding in different platform (tableau, web, plotly, etc.). All numbers in the data files were “string” type, so it was also necessary to convert these into “integer” type. After that, we have got four cleaned data sheet in total (‘Flow Checking in/out of each day’ and ‘Flow Checking in/out of each hour’). Cleaned data sheet would be save as date_in/out.pickle file

and hour_in/out.pickle file. “Pickled” file is the data object I use in python for different programs to share these data sheet.

In order to cluster the MRT stations into different groups, we processed the data sheet to find features of each stations. The “feature” is regarded as the property of MRT Stations flow, which is obviously different from one group stations to another one. We calculated features such as “Passenger Flow checking in/out the station in the morning/evening”, “Passenger Flow checking in/out the station in the weekday/weekend”, “Passenger Flow checking in/out the station in the high peak/ low peak”, “Passenger Total Flow”, “Passenger Average Flow” for example. Some of those are also described in previous works on Beijing subway like “Classification of subway stations in Beijing based on passenger flow characteristics” (Yin, Meng & Zhang, 2016).

Some features cannot get or observed from the source data directly. So for example the ratio of morning check-in flow and check-out flow. We used the following formula to calculate the ratio:

$$Ratio(X_1, X_2) = \frac{X_1 - X_2}{X_1 + X_2}$$

(X1 represented the check-in flow, X2 represented the check-out flow)

The range of this ratio formula is from -1 to 1. -1 means the check-in flow is none and 1 means the check-out flow is none. When the value is zero, it represents the situation in which the check-in flow and the check-out flow are equal. Due to the different size of quantity flow in different station, dividing with (X1+X2) can normalize the quantity among the stations. We also used this formula to calculate the ratio between weekday and weekends flow.

3.2. Algorithm

Based on different types of features, we could cluster the Taipei metro stations into groups based on clustering algorithm. The quantity of passenger flow in each hour or date is time series data. To select the features, some algorithms like Gaussian process model (高斯过程模型) and ARMA (Autoregressive and Moving Average Model) are mentioned by Yin, Meng and Zhang (2016). These models focus more on

the feature of the increase or decrease in the time series data. To avoid this complex processing, we used a K-means algorithm, which is famous and widely used for various clustering problems. In K-means algorithm, the score is fast to compute and it is higher when clusters are dense and well separated (Scikit-learn developers, 2016).

K-means algorithm divides the data into a set of N samples X into K disjoint clusters C, each described by the mean μ_i of the samples in the cluster. Usually, these mean values are called “centroids”. The algorithm aims to choose centroids that minimise the inertia (the sum of squared criterion):

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_j - \mu_i||^2)$$

The algorithm chooses the centroid in the groups again and again, which is usually called “iteration” until the clustering of data converges if enough time is given. In our case, the data is the quantity flow of Taipei MRT stations. Xj refers to j-th station, μ_i refers to the centroid station in i-th group. The formula can be described as ‘finding groups as well as minimizing the feature destinations between all the MRT stations and the centroid station in each group’. Comparatively, this algorithm is quite simple and easy to understand.

3.3. Visualization Tools

We used the programming language ‘python’ to process the data. During the data processing, we chose “matplotlib” as the best visualization tool to show the data which is a python 2-D and 3-D visualization library.

If we want to compare different features or stations with each other again and again, programming language can be a good choice. Once the coding part is completed, it can be used many times. However, in comparison to that, to be able to come up with an idea and to just give it a try, Tableau, for example, can be a better option. With no need to know programming language, the results still look very pretty and can be easily modified. That is why we think, once the data is processed, often Tableau is a better option than programming to show the results in data visualization.

The disadvantage of Tableau is a lack of 3-D tools and that it is difficult to share the result with friends or colleagues. Therefore, eventually, we used the web to demonstrate the result of metro stations stations clustering and integrate different graphics. Web programming is usually the last step of visualization. It can be seen as the closest part between audience and data where interactions are possible.

For web visualization tools, Google Maps API is easy to show the MRT map. It's interesting that there is a Transit layer in the Google API which can emphasize the color and line stroke on the MRT lines. Using javascript to draw different sizes of circles on the map can show the flow size of passengers.

Plotly is another web tool to make kinds of charts and dashboard which allowes to create 3-D plots. Apart from a small Plotly's advertisement on the upper right-hand corner on Plotly‘s display result, Plotly is beautiful and powerful. Also, compared to Google API and Baidu Echarts, Plotly is easier to program on the web.

4. Data Observation

4.1. Extreme Value

During the data processing, we can find that the flow of Taipei Main Station is far more than other stations. In Figure 2, each circle represents a station's passenger check in/out total quantity in April. The radius of each circle is proportional to the passenger flow. The red circle in the middle corresponds to the Taipei Main Station. The red circle covers almost the whole map and overlaps a lot with others which makes the graph a mess and hard to recognize. Figure 3 shows that tree map of each station's flow. Bigger squares in this figures go along with more passengers for each station. Obviously, Taipei Main station claimes the biggest area in this tree map.

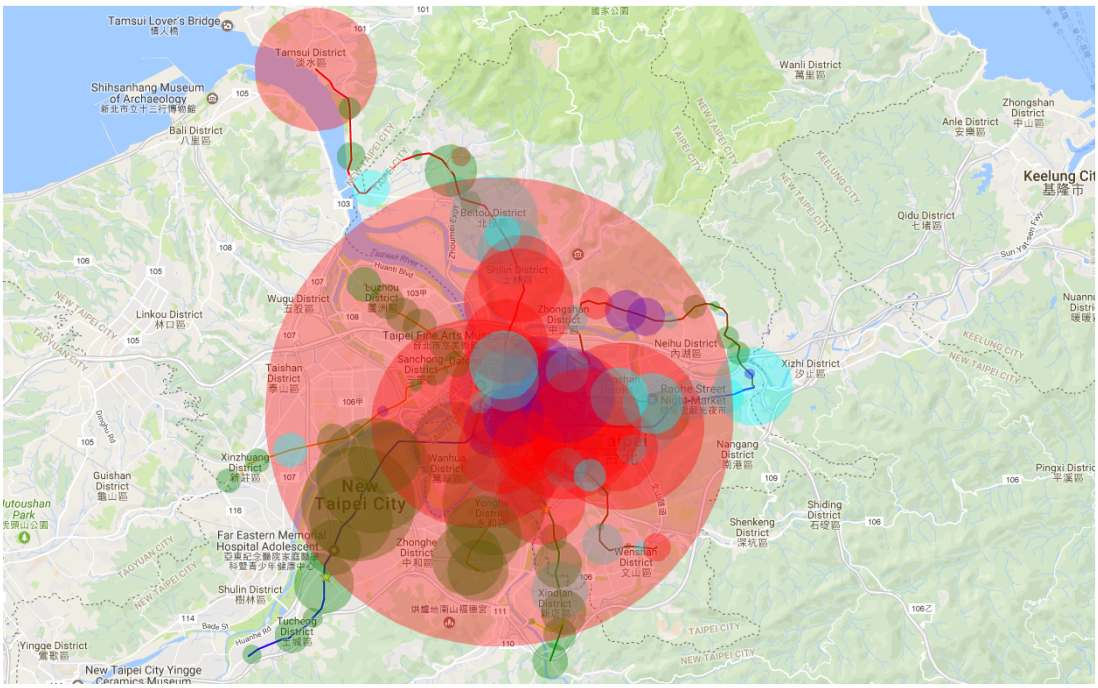


Figure 2. Taipei Geographical Map (before adjusted).

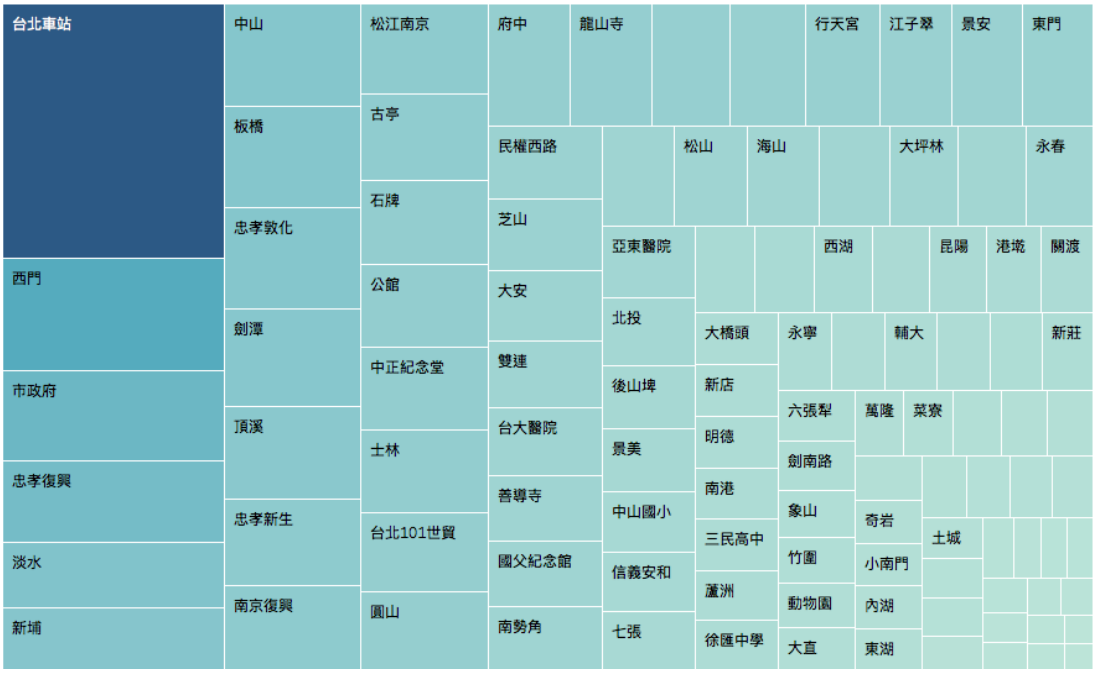


Figure 3. Tree map.

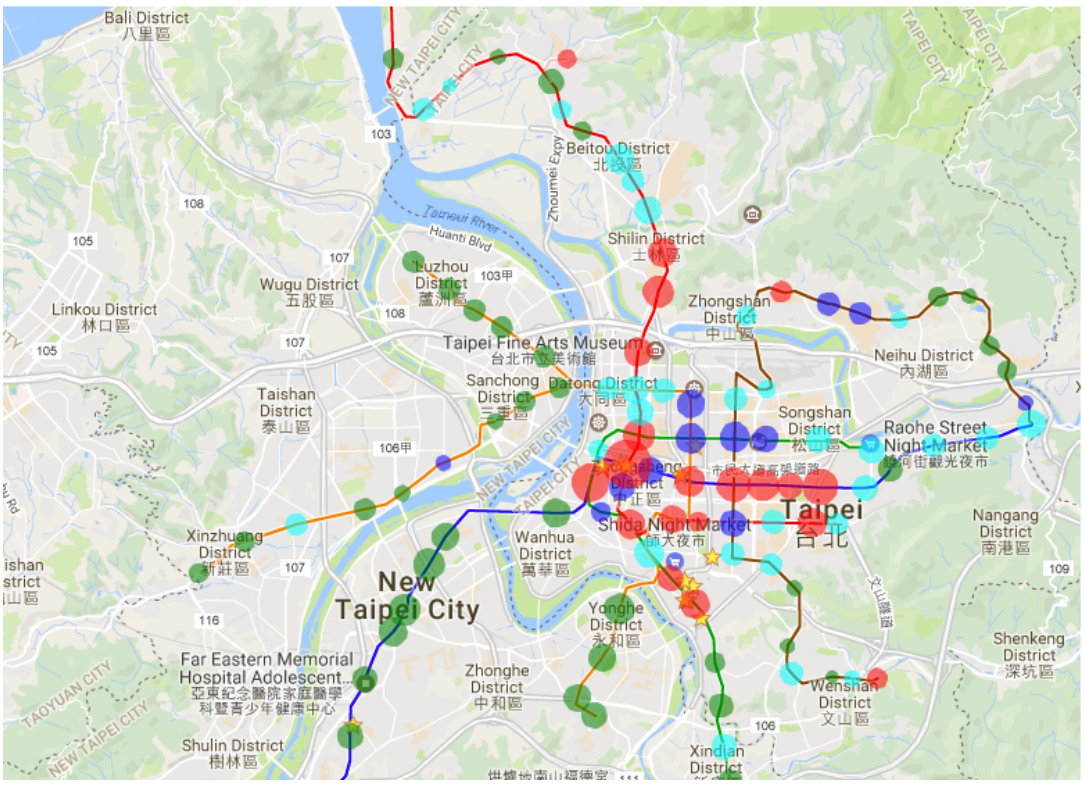


Figure 4. Taipei Geographical Map (After Adjusted).

To solve the visualization problem in Figure 2, our idea was to adapt the standard deviation of the station flow and make it smaller for all stations. So we used the formula: $X' = 5 * X^{1/3}$ to transform the data. (X is the original flow of each station, X' represents the converted data.) Then we plotted again the map which is now shown in Figure 4 and became clearer.

4.2. Correlation

During the data processing and feature selection procedures, we plotted some of the features onto Tableau to find out whether the feature has correlation to another. Usually, a good feature might make the data scatter onto different places on the graph.

In Figure 5, we can see that there's a high correlation between envening check-in-check-out ratio (evening_ior) and morning check-in-check-out ratio(morning_ior). It means that people usually get out of a station and get back into the same station. In Figure 5, we can easily reconized that the left-hand top side of the stations are almost located in office and commercial area in Taipei. While the

right-hand bottom side of the stations are almost located in residential area. For these reasons, we judged these features as useful and interesting to put into the clustering algorithm.

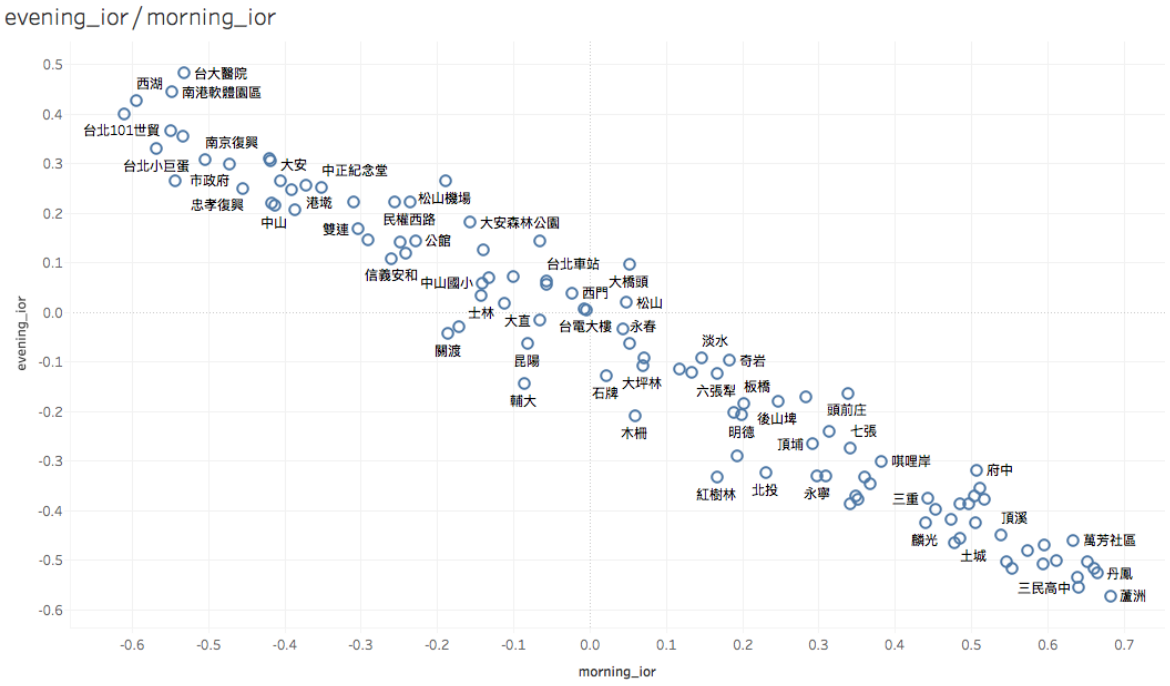


Figure 5. Scatter plot between evening-check-in-out ratio and morning-check-in-out ratio.

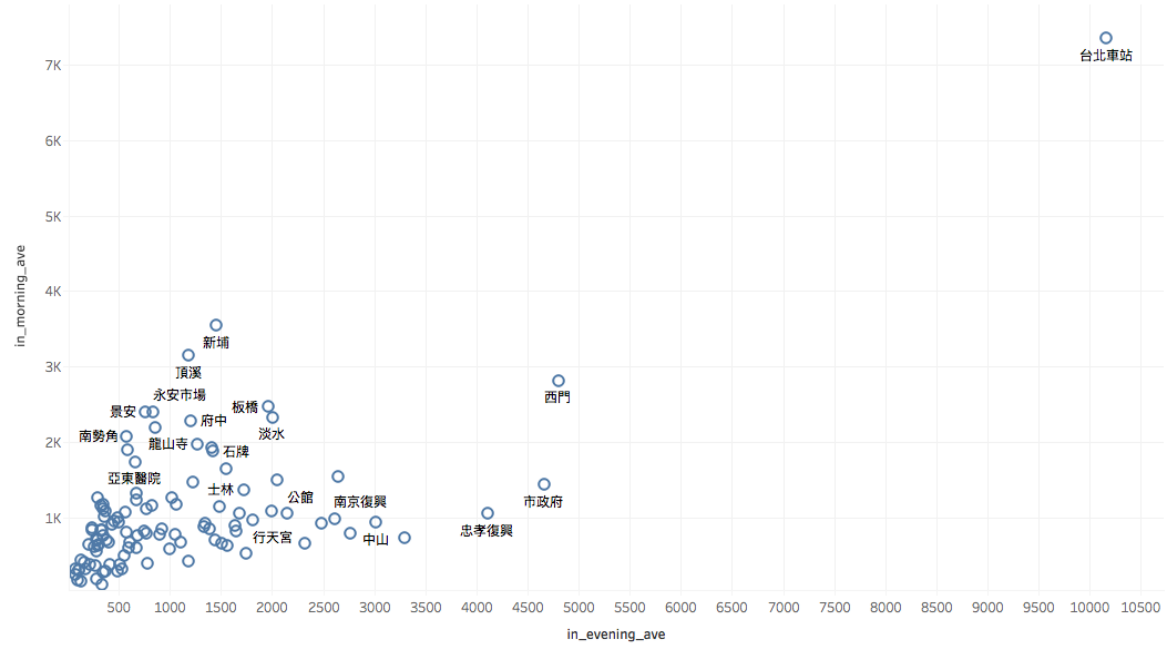


Figure 6. Scatter plot between check-in average flow and check-in evening flow.

Figure 6 is the scatter plot between evening check-in average flow and morning check-in average flow. Most of the stations gather together on the bottom left-hand side. Only Taipei Main Staion has a large flow and stands alone at the right-top corner. Having (almost) all data points in one corner, these features were not useful to cluster the data.

4.3. Adjusting Parameters

To determine the number of groups, we used ‘matplotlib’ to draw a 3-D plot which is shown in Figure 7. We can find out that the stations in right-top corner are usually in office and commercial areas. Stations at the bottom are usually relax and recreational areas, while the left-hand side are residential areas. The middle part, we referred to them as mixed area. So we tentatively decided for four groups on the map.

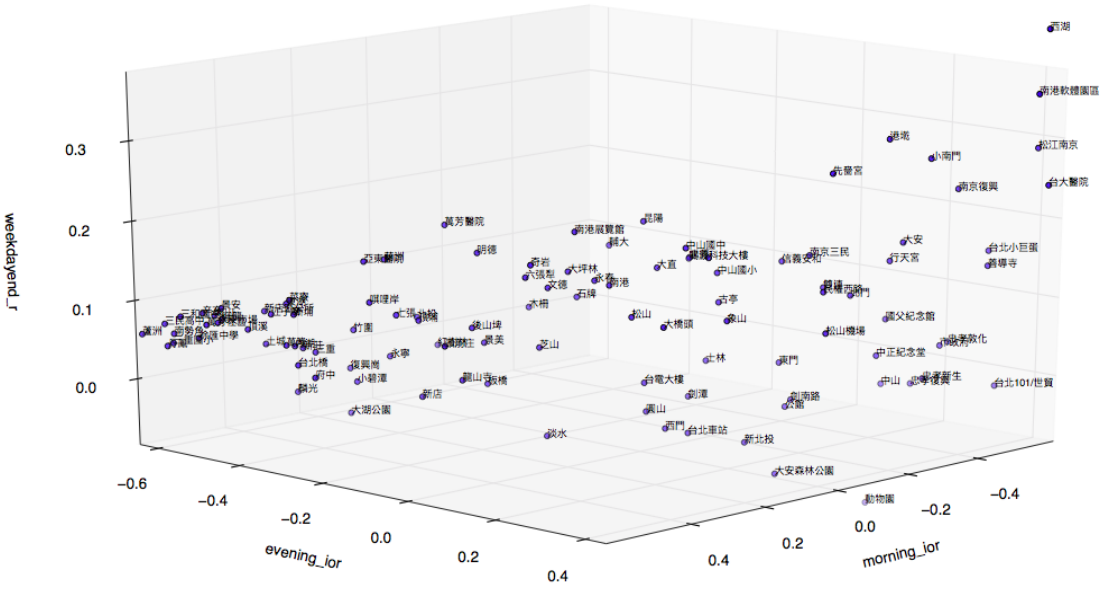


Figure 7. 3-D scatter plot among evening-check-in-out ratio, morning-check-in-out ratio, weekday-weekend ratio.

5. Visualization Results

Our visualization prototype result for the Taipei metro system and its classification is demonstrated on this website: https://goo.gl/H71dxp. In the following, the three different graphics will be decribed and explained.

5.1. 3D-Scatter Plot Clustering

We used “Plotly” to visualize the similarity or differences among the stations. In a 3-D scatter plot (Figure 8), all metro stations are represented as dots along their ratio values of the categories ‘weekday-weekend’, ‘morning-check-in check-out’ and ‘evening-check-in check-out’. While offering an interactive plot based on a web tool, in figures 8 the dimensions are presented as a screen shots example. If a user move its mouse on the right hand corner, a tool bar appeares (Figure 9), he/she can zoom in/out, rotate the graph, even save the graph as a file. If a user clicks on the label such as “混合區(Mixed Area)”, the label changes from color into grey and the plot hides the according dots on the graph.

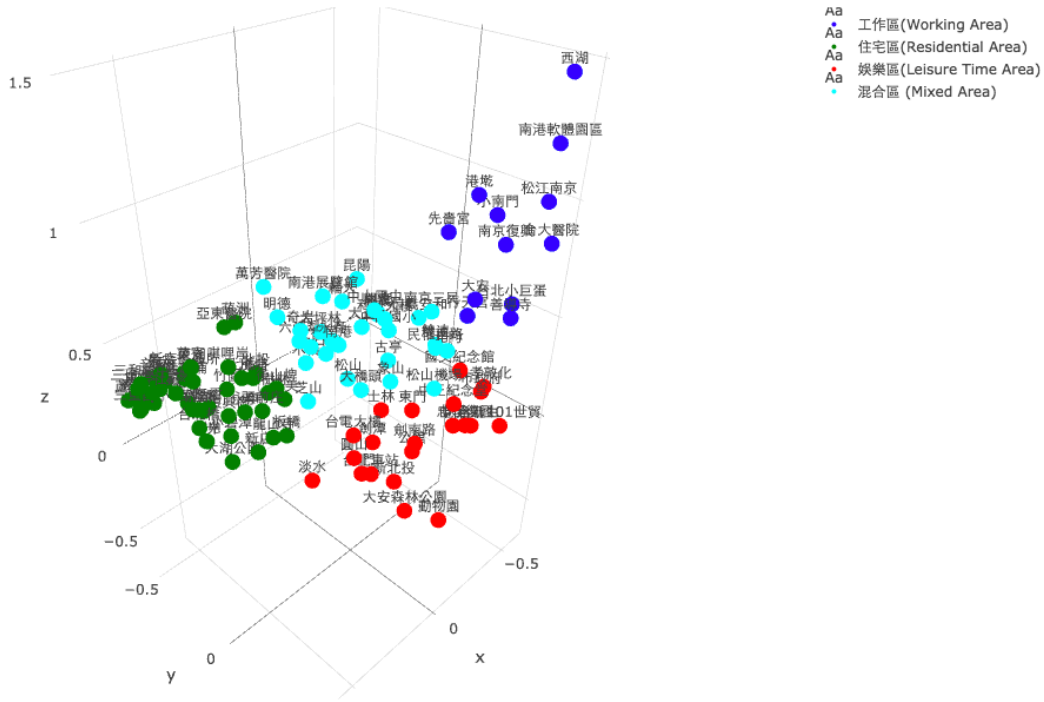


Figure 8. 3D-Scatter plot with group color among evening-check-in-out ratio, morning-check-in-out ratio, weekday-weekend ratio by Plotly.

Figure 9. A tool in 3D-Scatter plot.

We used different colors to demonstrate each group of the station. On the 3-D plot, firstly, we can discover that stations like Xihu(西湖站), NanGang Soffware Park(南港軟體園區) are to a high degree in the work area. The behavior of people going to work (check-out the station) in the morning and get back home (check-in the station) is very obvious. Secondly, the red stations in leisuring time area like Taipei Zoo(木柵動物園站), Daan Forest Park(大安森林公園) have high passenger flow at weekends and raletively low flow on weekdays. Thirdly, the green stations like Luzhou(盧州), Wanfang Community(方社區) are usually located in residential area. People check-in the station in the morning and check-out the station in the evening. For the leftest stations in the 3-D graph, the residential behavior is, therefore, more significant than others. Finally, we can find a lot of stations in cyan, mixture area plays an import role in Taipei city. Wanfang Hospital(萬芳醫院) is a center downtown in Muzha District, people going there might work there, live there or just change transportation between bus and MRT.

The advantage of 3-D scatter plot is that users can easily experience the similarity or difference between the stations. By comparing the distances between stations in one cluster we can tell a station belongs ‘likely’, ‘very’ or ‘extremely’ to ‘work area’ or ‘residential area’, for instance. For example, although Danshui Station (淡水) is a famous scenic spot and people like to go there at weekends, there is also a certain number of people who live there and go to Taipei downtown for work. So the location of Danshui(淡水) on the 3D-scatter plot is quite close to the green dots.

5.2. Geographical Network Flow Map

Figure 10 shows the map of Taipei City with MRT lines. Each station has a circle. The color represents the group that the station belongs to. The size of circle depends on the total flow of the corresponding station in April 2016. As already mentioned in Part 4, we have adjusted the size of radius among the stations to avoid the extreme value and overlap of Taipei Main Station.

Based on this map, we can easily identify the geographical location of each station. Users can also have a rough vision on the distribution of MRT stations and catch the distribution of four groups and areas on the map.

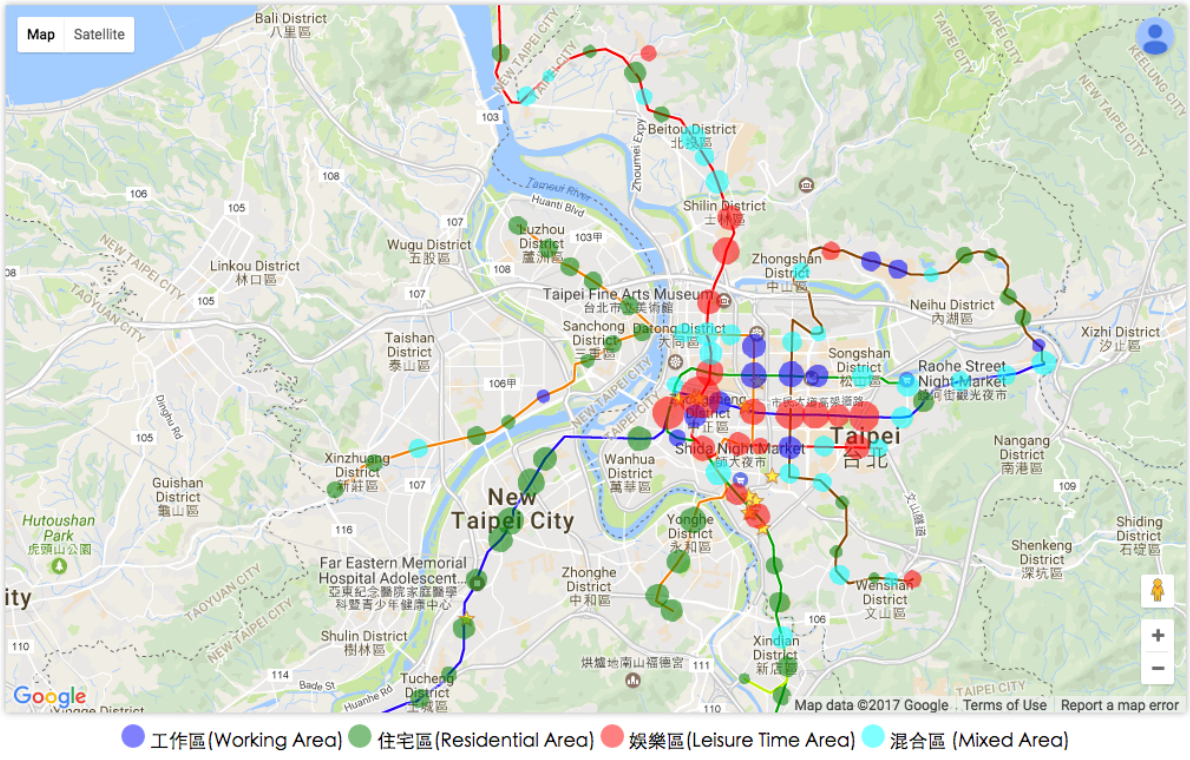


Figure 10.Geographical Network Flow Map.

Red dots represent the leisure time area cluster. Most of all the red dots are in the center of Taipei city, it seems that people in Taipei usually go to taipei center for leisure time. Taipei Zoo(木柵動物園), Danshui(淡水) and New Beitou(新北投) are few places outside the center that people like to go for fun by MRT system at weekends. Compared to Taipei City, stations in New Taipei city are usually in green which represents the residential area. Due to the assumption that the house price in New Taipei City might be quite low, it could play a role as residential area in “Big Taipei”.

5.3. Examples of Line Charts

If we click the circle of station on the geometrical map, the hour and day flow with two line charts below the map will be presented. The left one is the line chart with check-in/check-out flow in hour-series data and the right one is the flow in weekday-series data. All the values of the quantity of flow are calculated as averages for the month April 2016.

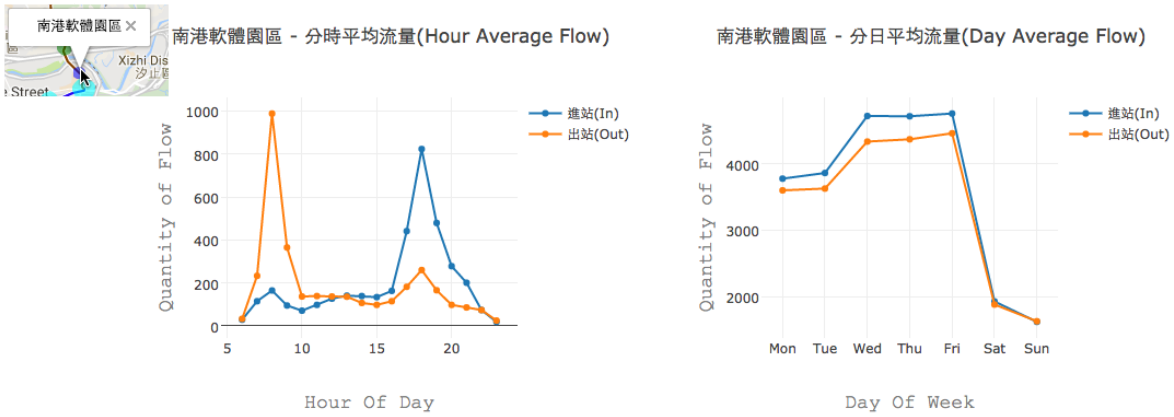


Figure 11. Hour/DayOfWeek Flow chart in Nangang Softwar Park(南港軟體園區).

In Figure 11, it shows the flow in Nangang Software Park. On the left hand side graph we can find that there's a high peak at 8 a.m., people check-out from the station and get to work at before 10 a.m. Then, people check-in to the station at about 6 p.m. On the right hand graph we can observe that the flow at weekends is very low, while the flow on weekdays is nearly three times the flow on the weekends.

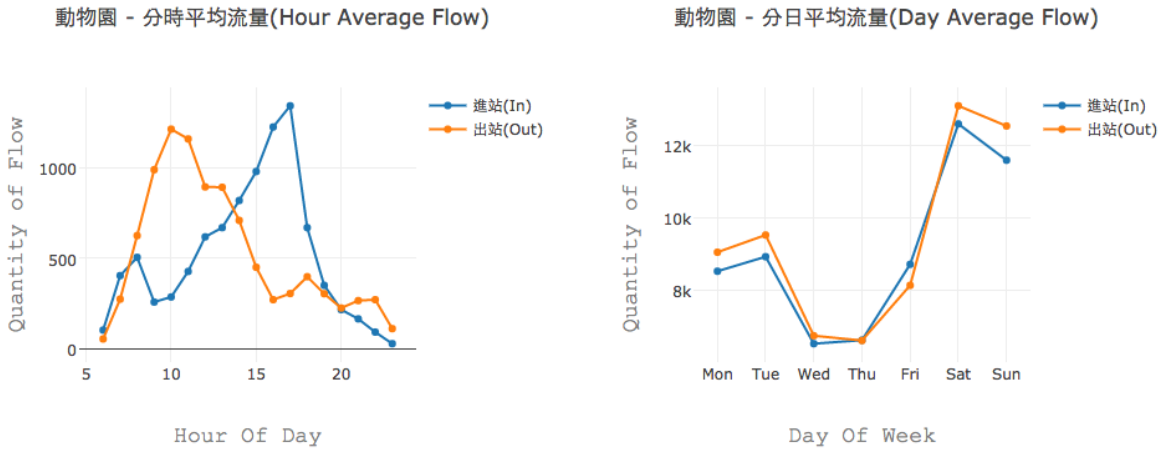


Figure 12. Hour/DayOfWeek Flow chart in Taipei Zoo(木柵動物園).

Figure 12 shows the flow in Taipei Zoo station. We chose it as the example in red leisuring time area. It's noticable that the flow at weekends is more than during the week. From the hour average flow line chart, we can find that people might arrive at the zoo around 10 a.m. and head back home at about 5 p.m..

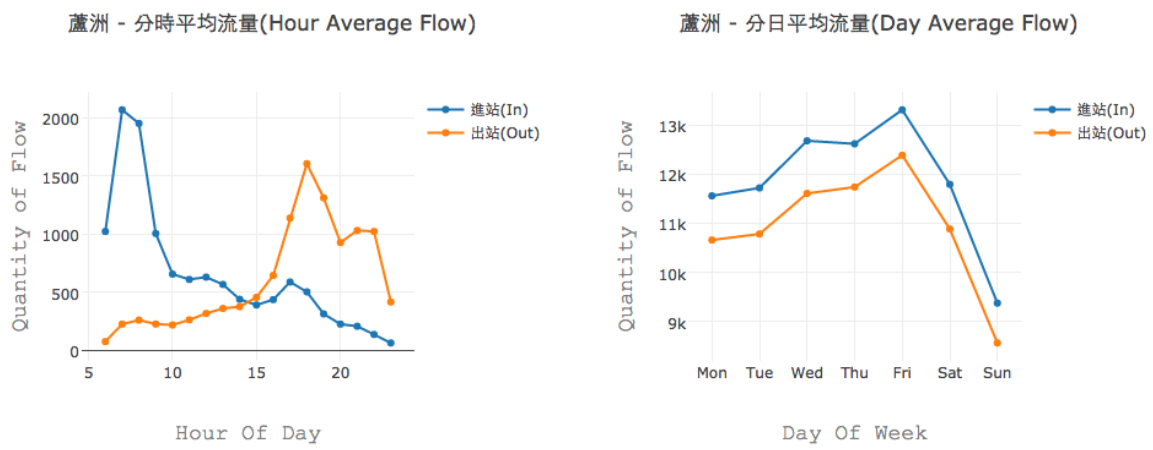


Figure 13. Hour/DayOfWeek Flow chart in Luzhou(蘆洲).

Figure 13 shows Luzhou Station's (蘆洲) passengers flow in line chart. Obviously, it is a residential station. A very high peak on the hourly line chart for the check-in blue line at 7 p.m. demonstrates that people leave for work around 7 a.m. and 8 a.m. Because the station is a bit far from Taipei center, the morning high peak is nearly one hour earlier than that in other stations (like Nangang Software Park Station).

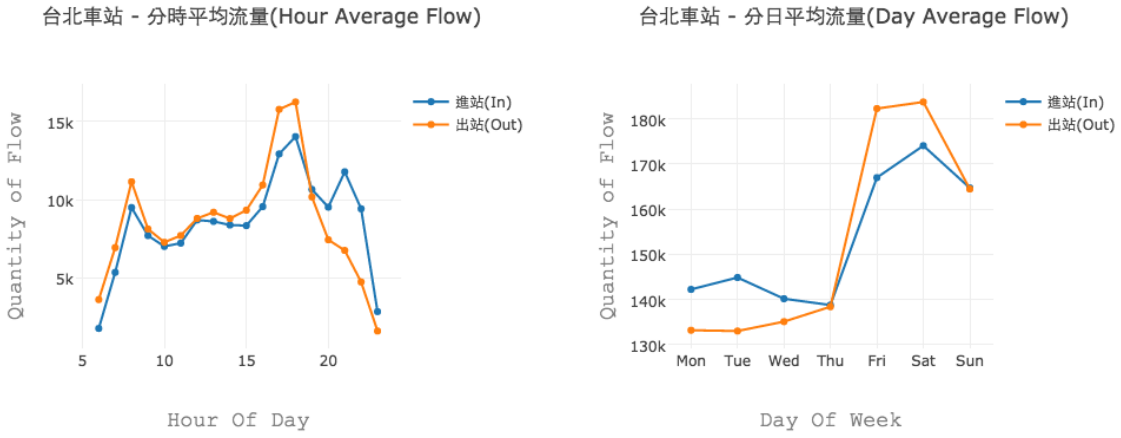


Figure 14. Hour/DayOfWeek Flow chart in Taipei Main Station(台北車站).

Lastly, we see the passenger flow in Taipei Metro station in Figure 14. The flow on Friday and weekends is higher than Monday to Thursday. Many people go to Taipei Main Station and then travel to other cities. Additionally, a lot of young people go there for shopping and for cram school. So there are two high peaks on the hourly average flow. One is on 8 a.m and another at about 5 p.m. and 6 p.m. It is interesting that we can find another small peak at the hourly check-in blue line at 9 p.m. which

shows people would get back home at that time. Thus, Taipei Main station is a lively place for people to relax and have fun after school and work.

6. Discussion of the Design and Future Work

All in all, in the web version of our project, a network flow map, line charts and a 3-D scatter plot is used to illustrate different characteristics of the data. As described above, we only used data of one month in order to create a prototype. For the discussion we will compare our work with the Principles of Analytic Graphics by Peng (2014) as learned in our course in the following.

The first principle is to show comparisons. As we only did a prototype for possible further visualizations, we did not have too much time and chance to offer comparisons. Moreover, since this data visualization process was rather explorative instead of guided by certain hypotheses, we do not have specified hypotheses to compare. This being said, with our visualized data in a 3-D scatter plot to show the position of each station along the features and the color coded groups of clusters, we offer the possibility to compare the positions of the different stations along certain features and also to observe how the stations are clustered. It can be directly compared which position of the station resulted in which cluster membership and whether another cluster membership would have been plausible as well. Furthermore, the visualization of the data in the network flow map offers the opportunity to compare the amounts of passenger flow at different stations by comparing the different sizes of dots and also the line charts. Through that comparisons, hypotheses can be drawn about which station is used more frequently and when.

The second principle is to show causality, mechanism, explanation, systematic structure. By offering the possibility to look at the passenger flows for each station for in-checking and out-checking hourly as well as for every day of the week, we allow not only to compare different aspects but also to understand the underlying conclusions we made regarding mechanisms or causality in order to choose the types of clusters. For example, for the metro station Nangang Software Park (南港軟體園區), whose color coded in green to show that people usually go there to work,

the 3-D scatter plot can show the general group cluster and the line charts can explain the membership in this category in more detail: There is a high peak of out-checking people in the morning around 7-9am and a high peak of in-checking people in the evening around 5-7pm. Furthermore, regarding the informations on the second line chart, it is observable that there is a considerable decrease of passenger flow on the two weekend days Saturday and Sunday for this station.

The third principle is to show multivariate data. In our project, we showed the position of each station along three features in the 3-D scatter plot while also indicating the color coded clusters. Furthermore, in the combination of network flow map and line charts, we also presented the geographical position, the station's cluster type and the amount of the passenger flow by size of the dot as well by different lines representing the hourly/daily data as well as the in- and out-checkings. Designing these graphics in an interactive way allows to visualize more then two different variables at the same time in a non-confusing, clear manner. Showing 3-D plots in 2-D on paper often has the difficulty that some parts of the graph hide other parts of the graph. Having a 3-D plot which is interactive in this way that you can turn the whole coordinate system and zoom in and out, the visualization can be viewed from every angle. Combining the network flow map and line charts in an interactive way to present multivariate variables allows to gain an overview about the position, cluster and rough amount of metro flow first and then decide where you want to have more detailed information on without having too much information in one graph.

The forth principle is the integration of evidence. As we have different types of graphics and elements (map, dots, lines, color, words that represent the name of the stations if you click on the station in the map, numbers of you pass the lines with the mouse on the line charts, …) to show different aspects of the data that work hand in hand in an interactive way with the user, various modes of data presentation are used. We tried to make sense out of the complete data we got in the most compresensive manner without overcrowd any visualization.

The fifth principle is to describe and document the evidence with appropriate labels, scales, sources, etc.. Looking at the network flow map as the graphic on the top of the website first, the color coded types of clusters are labelled in the legend under the map. Also the station name appears for every station when you click on the station. Ideally, another legend should provide informations about the size of the dot and the rough amount of passenger flow. Since this prototype map was made with google map, it was too difficult to integrate this type of legend. Since you can zoom in and out and the size of the dots changes, the size of the dot in the legend have to adjust as well. Nevertheless, by clicking at one specific station, you can see numbers for the amount of passenger flow in the line charts at least.

In the line charts the axes are all labeled with words and numbers and the titles explain what is shown. Furthermore, the legends explain the different colored lines. Often, it is better to have a the same intercepts/consistent axes in the line charts if there are multiple and they have a similiar axes (Quantity of Flow in this case) so that users can compare the multiples. However, since the numbers in our case are really different (on the left side the average sum of passengers in one hour is displayed, on the right side the the average sum of passengers per day), we went for different intercepts in the line charts.

The 3-D scatter plot has labels for the name of the station as well as for the color of the cluster. Furthermore, small pictogramms for actions such as zoom and rotation appear in the right upper corner if the mouse is in this area and short explanations appear if the mouse is held over the small pictogramms. The labels for each dimension on the axes are missing since, also in this case, it was too difficult to include the labels in this graphic which is supposed to be a prototype only. This would be needed to improve before publish this plot as a ‘real’ visualization.

The colors representing the clusters are the same in all graphics in order to present a consistence and homogenous appearance and to make it easier to follow. We believe, that all together, all three graphics show a complete and comprehensible story in which every graphic is able to contribute different informations which integrate well.

Which leads to the sixth and last principle that is that ‘content is king’ and analytical presentations ultimately stand or fall depending on the quality, relevance, and integrity of their content. Since visualisations of crowd flow in metro systems can help to plan better as well as to react to dangers arising by a high number of people, the visualization of metro systems, especially in big cities with rising population, is a relevant topic. Furthermore, classifying stations to better understand what kind of people use these stations and when can help as well regarding this issue. It can help to predict metro flow depending on time and date (e.g. holiday or weekday). Observing what might be the purpose of different areas why people go there and when also helps to get a broader understanding of the whole city and its compostion.

Another principle we have discussed in course was ‘Making Sense at the Speed of Thought’. By using different types of graphics the given information is not too much at once but still complete for our project. Moreover, users can zoom in after getting an overview and also choose a station to get more detailed information in the line charts. In the 3-D plot, different perspectives are possible.

Since this project only represents a prototype, there are parts which can be improved in the future. First, as a prototype work we only used data of one month. A broader and more complete overview can be achieved by integrating data of a longer time period. Change over time such as change over the seasons or also over the years could also be interesting to visualize in order to predict metro flow in the future.

Using an interaction of map and line charts, the visualization could be improved insofar that the line chart could appear right next to the station in the network flow map. Furthermore, future work could program a visualization which allows users to tick a box for every station that they want to see the line chart for so that it is possible to easily compare the line charts for different station’s passenger flows over time.

Offering different dimension and not only variables in a scatter plot which can be chosen to be shown by ticking a box each, would allow even more interaction in the future. Then correlations between different dimension can be displayed to let the users explore different patterns of clusters themselves.

Furthermore, the clustering of the metro stations resulted in four stations (work area, leisure time area, residential area, mixed area) with the mixed area as a

classification form that includes areas which could be residential but also working areas, for example. In future and more professional work classification should be more precise and specific so that there is ‘left-over’ category or, otherwise, should engage in deeper research what this category could represent.

All in all, with these visualizations and classifications an overview over the regularities and predictions regarding metro passenger flow in Taipei and its structure can be gained. This could be used in the future, for example, by adjusting the metro train length or the time intervals of the metro for peak- and off-peak-times, days or certain lines. Moreover, also different types of advertisements could be used to address different types of people based on the classification of metro stations (e.g. do you want to address people on their daily way to work or during their free time?).

7. References

Briney, A. (2014). Overview of flow mapping. Retrieved from: https://www.gislounge.com/overview-flow-mapping (06 November 2016)

Buckley, A. (2013). Mapping flow data (Presentation). Retrieved from: https://blogs.esri.com/esri/arcgis/2013/04/23/aag-flow-mapping-presentation-available-for-download/ (05 November 2016)

Itoh, M., Yokoyama, D., Toyoda, M., Tomita, Y., Kawamura, S. & Kitsuregawa, M. (2013). Visualization of passenger flows on metro. Retrieved from: http://www.tkl.iis.u-tokyo.ac.jp/top/modules/newdb/extract/1303/data/VAST2013_1.pdf (03 November 2016).

Lin, J. & Shin, T. (2008). Does transit-oriented development affect metro ridership? Evidence from Taipei, Taiwan. *Transportation Research Record: Journal of the Transportation Research Board, 2063,* 149-158.

McGraw-Hill (n.d.). Mapping Exercise – Thematic Mapper: The Flow Map. Retrieved from: http://highered.mcgraw-hill.com/sites/dl/free/0072943823/599982/Flow_Mapping.pdf (06 November 2016).

Peng, R. D. (2014). Principles of Analytic Graphics. Johns Hopkins Bloomberg School of Public Health. Retrieved from: https://github.com/jtleek/modules/blob/master/04_ExploratoryAnalysis/Principles/index.md (02 January 2017)

Phan, D., Xiao, L., Yeh, R., Hanrahan, P. & Winograd, T. (2005). Flow map layout. Retrieved from: http://graphics.stanford.edu/papers/flow_map_layout/flow_map_layout.pdf (03 November 2016)

Scikit-learn developers (2016). 2.3. Clustering (online learn platform). Retrieved from: http://scikit-learn.org/stable/modules/clustering.html#k-means (1 Jan 2017)

Simantel, B. (2012). Generating distributive flow maps with ArcGIS. Retrieved from: https://blogs.esri.com/esri/apl/2012/09/12/generating-distributive-flow-maps-with - arcgis (06 November 2016)

Wang, J. (2016).台北捷運路網與階段性通車規劃.Taipei Metro network and staging operation planning (Presentation). Retrieved from: http://www.metrotaipei20.com/index.php/en/program-at-a-glance/ (31 December 2016).

Wei, Y., Lin, S., Chu, R., Tian, Q. & Fei, W. (2016). A method of grading subway stations. *Procedia Engineering, 137,* 806-810.

Wen, K. (2013). A dynamic simulation of crowd flow in Taipei railway and MRT station by multi-agent simulation system. *Urban Planning and Design Research, 1*(4), 59-68.

Wen, K. & Chang, S. (2014). An environmental study of crowd flow transformation at Taipei MRT station. *Procedia Environmental Sciences, 22,* 43-60.

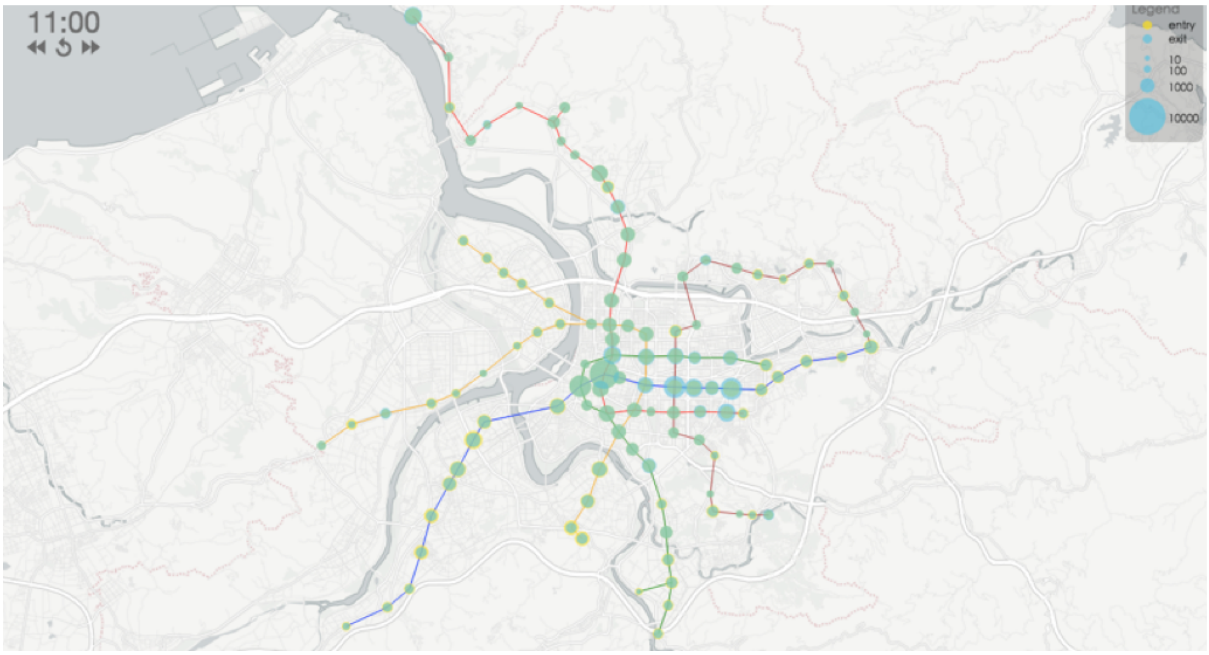
Yin, Q., Meng, B. & Zhang, L. (2016). Classification of subway stations in Beijing based on passenger flow characteristics. *Progress in Geography, 35*(1), 126-143.

Data source:

臺北大眾捷運股份有限公司 劉亦昌 (2016) 臺北捷運各站進出量統計. Retrieved from: http://data.taipei/opendata/datalist/datasetMeta?oid=1d71c478-205f-42c5-8386-35f86d74fdd1 (27 Dec 2016)

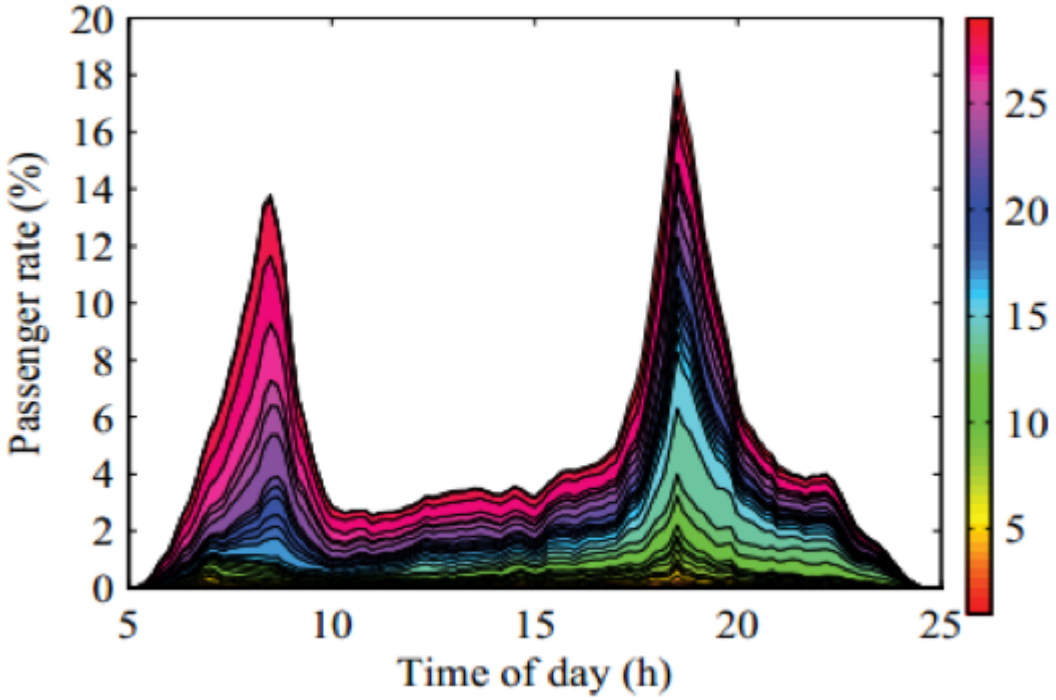
8. Appendix: List of Information Visualization Cases

Appendix A: A day in Taipei metro



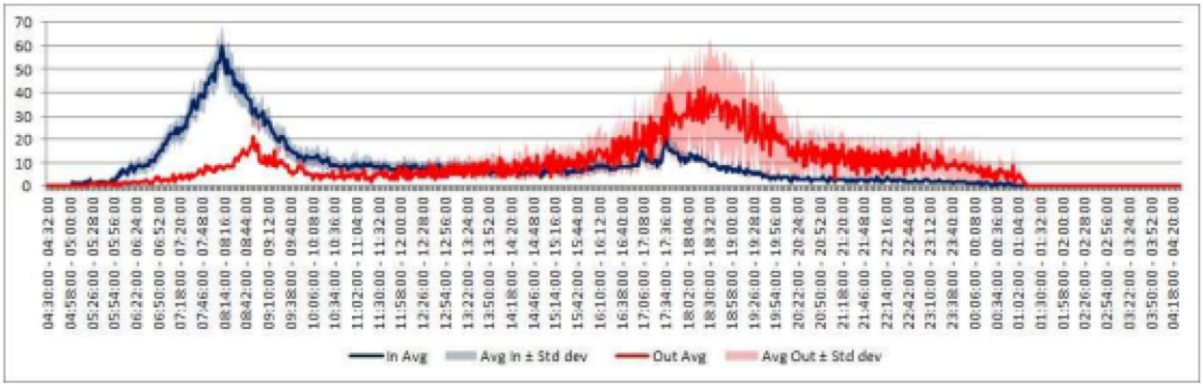
Claire Tsao. (2015). A Day in Taipei Metro. Retrieved from: http://missmoss.github.io/taipei-mrt-viz/ (07 November 2016)

Appendix B: One day metro data in Singapore



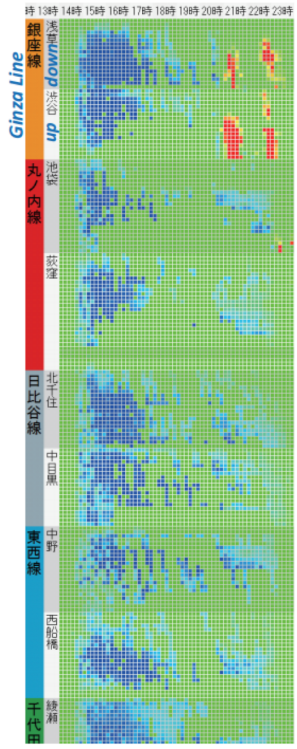
Sun, L., Lee, D.-H., Erath, A. Huang, X. (2012) Using Smart Card Data to Extract Passenger's Spatio-temporal Density and Train's Trajectory of MRT System. Retrieved from: https://pdfs.semanticscholar.org/b692/1f2bdb54e6a476e0ee0730bb957a6bd249a3.pdf (03 November 2016)

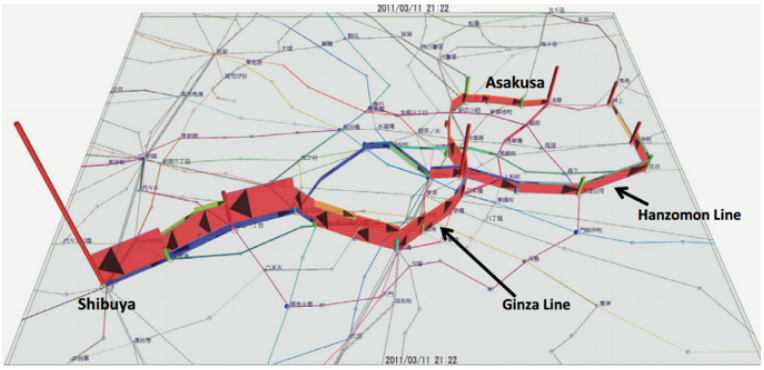
Appendix C: Finchley central station weekday's passenger quantity



Ceapa, I, Smith, C. & Capra, L. (2012). Avoiding the Crowds: Understanding Tube Station Congestion Patterns from Trip Data. Retrieved from: http://www0.cs.ucl.ac.uk/staff/l.capra/publications/urbcomp12.pdf (02 November 2016)

Appendix D: Tokyo Metro Ginza Line





Itoh, M., Yokoyama, D., Toyoda, M., Tomita, Y., Kawamura, S. & Kitsuregawa, M. (2013). Visualization of passenger flows on metro. Retrieved from: http://www.tkl.iis.u-tokyo.ac.jp/top/modules/newdb/extract/1303/data/VAST2013_1.pdf (03 November 2016).