GEO881 - Data Challenge 3 Report

1. Explore the data

1.1 Explore the network data

In this dataset, Zurich has 2084 streets segments and 1313 street intersections, New York has 6717 streets segments and 4120 street intersections. The street segments in New York are generally longer than those in Zurich (Figure 1).

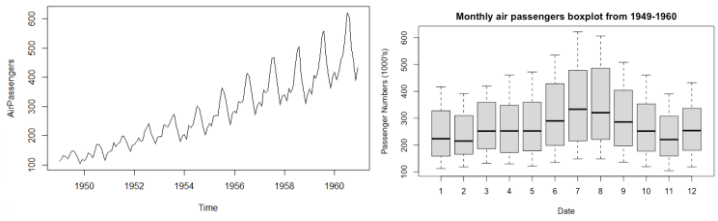
```
[1] "Number of streets in Zurich"
[1] 2084
[1] "Number of street intersections in Zurich"
[1] 1313
[1] "Summary of street lengths in Zurich"
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  0.2756   6.9441  15.3952  26.4738  36.7709 302.1729
[1] "Number of streets in New York"
[1] 6717
[1] "Number of street intersections in New York"
[1] 4120
[1] "Summary of street lengths in New York"
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  0.0249   9.6189  18.5387  46.7943  65.2160 894.5724
```

Figure 1. Basic statistics and visualization of network data (left: basic statistics, middle: Zurich network, right: New York network)

1.2 Explore the time series data

AirPassengers data seems to have an increasing trend and seasonal pattern in the time series plot, and the box plot show that the air passenger flows are larger in July and August than in other months; no obvious patterns can be observed in Zurich pedestrian counts in 2020, but there is an obvious break point (Figure 2).

```
[1] "summary of AirPassengers data"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  104.0   180.0   265.5   280.3   360.5   622.0
```




```
[1] "summary of Zurich pedestrian count in 2020"
      DATE                PED_TOTAL
 Min.   :2020-01-01   Min.   : 311
 1st Qu.:2020-04-02   1st Qu.:1450
 Median :2020-07-02   Median :2485
 Mean   :2020-07-01   Mean   :2365
 3rd Qu.:2020-10-01   3rd Qu.:3087
 Max.   :2020-12-31   Max.   :4504
```

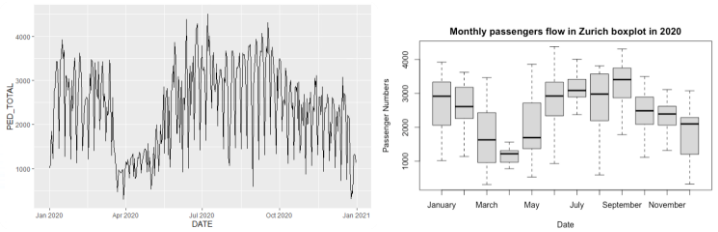


Figure 2. Basic statistics and visualization of time series data (top: AirPassengers data, bottom: Zurich pedestrian counts in 2020)

2. Network analysis

2.1. Create street graphs and dual graphs

The first step is to filter street segments because there are some street segments sharing the same start and end nodes, among which, the shortest street segments are kept. After filtering, there are 2079 street segments in Zurich and 6679 street segments in New York.

The second step is to create undirected weighted street graphs for Zurich and New York (Figure 3). The nodes of the graph are the street intersections, the edges of the graph are the street segments, the weights of the graph are the lengths of street segments.

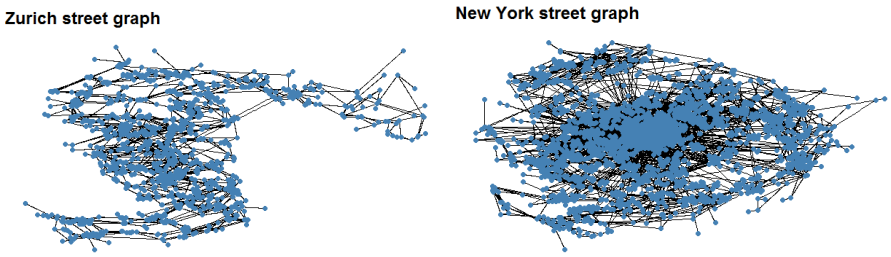


Figure 3. Street graphs

The third step is to create undirected unweighted dual graphs for Zurich and New York (Figure 4). The nodes of the graph are the street segments, the edges of the graph are the street intersections, and the dual graphs are more complicated than street graphs.

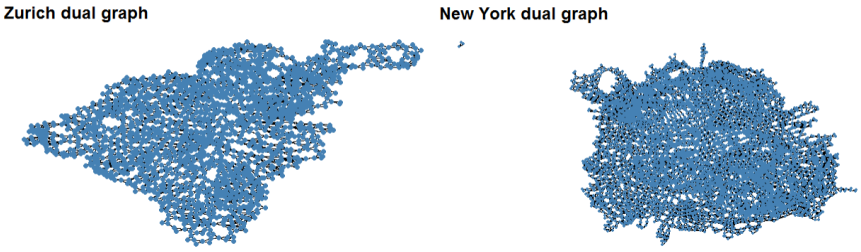


Figure 4. Dual graphs

2.2. Centrality calculation of the graphs

2.2.1. Centrality of the street graphs

Add *degree_weighted, closeness_weighted, betweenness_weighted, degree_unweighted, closeness_unweighted, betweenness_unweighted* columns to Zurich street graph and New York street graph, and visualize them.

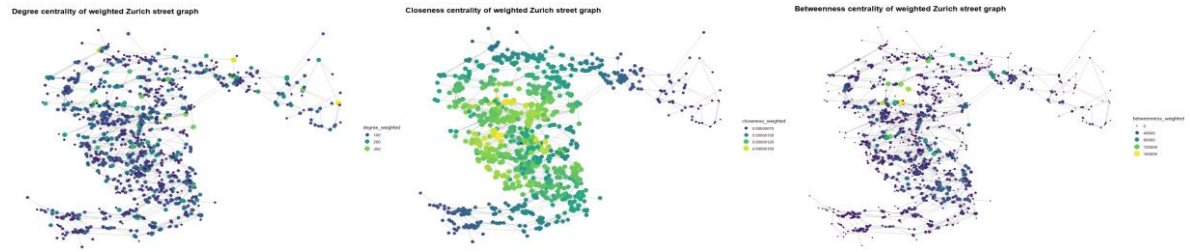


Figure 5. Centralities of the weighted Zurich street graph

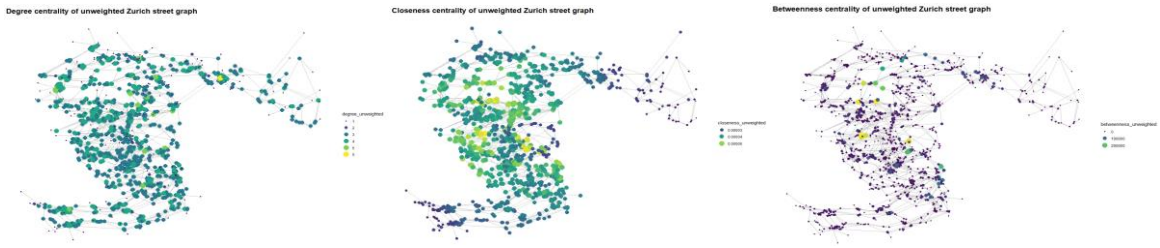


Figure 6. Centralities of the unweighted Zurich street graph

Figure 7. Centralities of the weighted New York street graph

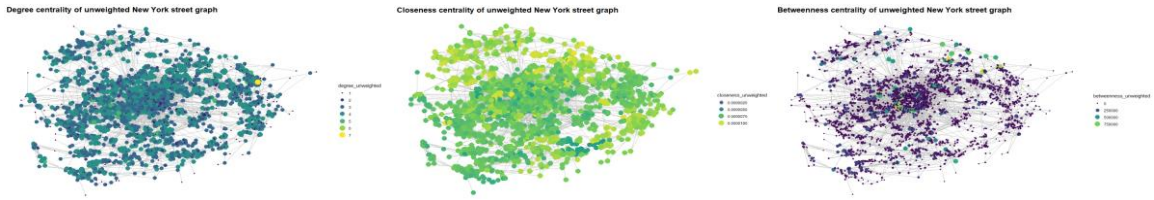


Figure 8. Centralities of the unweighted New York street graph

2.2.2. Centrality of the dual graphs

Add *degree, closeness, and betweenness* columns to Zurich dual graph and New York dual graph, and visualize them (Figure 9, Figure 10).

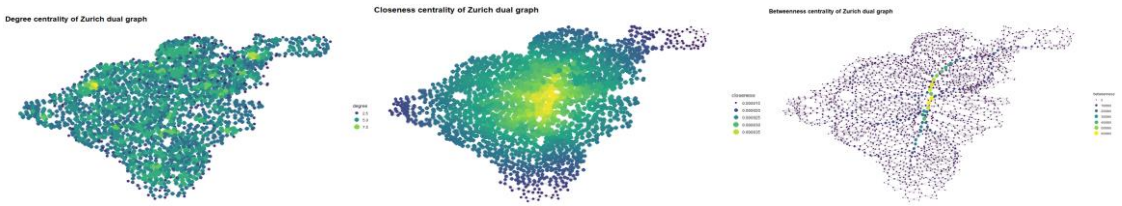


Figure 9. Centralities of the Zurich dual graph

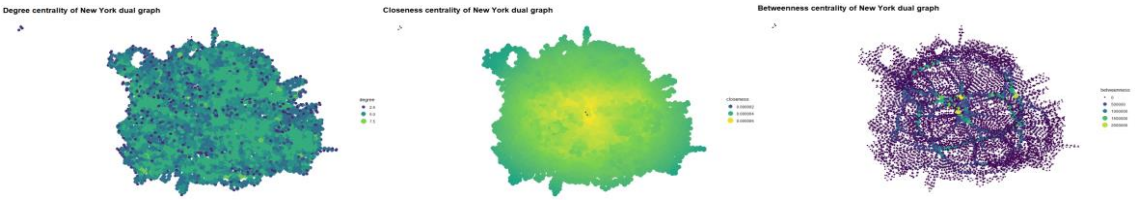


Figure 10. Centralities of the New York dual graph

2.3. Visual and numerical analytics

Zurich and New York in this report mean the study areas of these two cities, instead of the whole cities.

2.3.1. Centrality within the same graph

Zurich street graph: Most street intersections have similar degree centrality, being around 50 to 150 (Figure 5, Figure 11). As for the closeness and betweenness centralities, street segments in the center of Zurich have higher centralities. The closeness centrality ranges from 0.0000005556 to 0.0000015561 and is mainly distributed around 0.0000010 to 0.0000014 in central area. The betweenness centrality ranges from 0 to 280078, and is mainly distributed around 0 to 161606 (Figure 11, Figure 13). Based on the closeness and betweenness, most street intersections in Zurich have similar closeness to each other (street intersections with extremely high or low closeness values are rare); and they are not so important as bridges connecting other street intersections because except the center street segment, most of them have low betweenness values, which means there are not many hub street intersections in Zurich. The unweighted street graph shows similar distribution pattern of centralities, but it does not consider the length of street segments, thus the results are different from

the weighted one. For instance, there are more street segments around the Zurich center with similarly high closeness values in the weighted street graph, this is because these street segments are not long, which means that they are close to the center (Figure 6, Figure 11, Figure 13).

New York street graph: For the weighted street graph, most streets have low degree centrality and betweenness centrality, and there is no specific distribution patterns in maps for the two centralities in New York, while the street segments in the center of New York have higher closeness centrality and the centrality decays as they are away from the city center (Figure 12, Figure 14). Like Zurich, most street segments in New York have similar degree and closeness, as for the betweenness, not many of them serve as bridges between different areas. The unweighted street graph also indicates that the street segments in the city center have higher centralities than those in marginal areas, but this pattern is not as obvious as in the weighted street graph.

Zurich dual graph: The degree centrality ranges from 1 to 9, and most values distributed around 4 to 5 (Figure 9, Figure 15), which means that most streets in Zurich are linked with 4 to 5 streets. As for the closeness centrality, it has higher values in the center of Zurich. Its frequency distribution looks like a normal distribution, thus Q-Q plot and Shapiro-wilk normality test are performed, the p-value is 0. 0.00008092 (less than 0.05), which indicates that it is not normally distributed. The betweenness centrality shows that except for the street segments in the city center, most streets are not very important in connecting others.

New York dual graph: The degree centrality ranges from 1 to 9, and most values distributed around 5 to 6 (Figure 10, Figure 16), which means that most streets in New York are linked with 5 to 6 streets. The distribution patterns of centralities are similar with Zurich dual graph, closeness centrality and betweenness centrality are higher in the city center. Figure 16 shows that the frequency distribution of closeness centrality is skewed, this is because there are several isolated street segments and intersections not connected with the main network in New York, and it results in an extremely low closeness centrality.

2.3.2. Centrality between the street graph and dual graph

Degree centrality: The degree centrality is calculated based on the number of edges linked to the node and it shows the how easy the node is influenced by something flowing in the network. For the weighted street graph, the degree centrality is the sum of the weighted of the edges. In the Zurich street graph, most street intersections are connected with 3 to 4 street segments (Figure 11, Figure 13); in the Zurich dual graph, most street segments are connected with 4 to 6 street intersections (Figure 9, Figure 15). The degree centrality of New York street graph and dual graph have similar frequency distribution and distribution patterns in the map (Figure 10, Figure 12, Figure 14, Figure 16).

Closeness centrality: The closeness centrality shows how close the node is to other nodes in the street graph, and in the dual graph, it shows the closeness among edges. The street graph and dual graph of Zurich show no matter the street segments or the street intersections are closer to each other in the city center (Figure 9, Figure 11). The histograms suggest that there are almost no extremely high or low values, most street segments and street intersections have the closeness centrality distributed around the medium value, and geographically, they surrounded around the city center. In general, there is a decaying pattern in close centrality from the city center to the marginal areas (Figure 9, Figure 11, Figure 13, Figure 15). The same pattern applies to the New York street graph and dual graph.

Betweenness centrality: The betweenness centrality shows how important the node is in acting bridge between different subgroups in the street graph, and shows the importance of edges in the dual graph. Based on the histograms, the betweenness centralities in both street graphs and dual graphs of Zurich and New York are skewed distributed (Figure 13, Figure 14, Figure 15, Figure 16). The maps also validate that most street segments and intersections have low values over the whole city, and only those along the main streets in the city center can be the bridges or hubs for the city network(Figure 9, Figure 10, Figure 11, Figure 12, Figure 13, Figure 14, Figure 15, Figure 16).

2.3.3. Centrality between the same type of graphs

This section mainly compares the centralities of the weighted street graphs of two cities and explains the reasons for the differences.

Degree centrality: Based on the dual graphs, in Zurich, street segments in the city center have higher degree centralities; in New York, there are many street segments with low degree centralities in the city center (Figure 9, Figure 10). Two cities have similar degree centralities, with a mean around 5, and most street segments connect with less than 6 street segments (Figure 15, Figure 16). The reason for different degree centralities might be the purity of branch streets. The branch streets in Zurich connect with several main streets, while those in the center of New York connect with only 1 or 2 main streets, so they have low degree centralities.

Closeness centrality: Street segments along Bahnhofstrasse in Zurich, and street segments stretching along Lincoln Tunnel in New York are closer to others. Compared with New York, streets segments in Zurich have higher closeness centralities, with a mean being 0.0000011616, while the mean of New York is 0.00000011894 (Figure 11, Figure 12). The grid system of New York network makes it easy to find a way though, the vertically intersected street segments, to some extent, enlarge the distance between each other and decrease the closeness centrality. The street segments in Zurich are not as regularly-organized as those in New York, and many branch streets help to increase the closeness centrality. The border effect of the Zurich street graph is more obvious than the New York street graph, and it might due to the denser network in Zurich.

Betweenness centrality: The betweenness centralities of two cities are distributed similarly with the closeness centralities, and street segments in the center areas have higher betweenness centralities. The street segments in New York are more important in bearing network flow, with a mean being 84409, while the mean of Zurich is 16140 (Figure 11, Figure 12). The reason for the differences between two cities might be how regularly the network is planned. In New York, every street segment is indispensable in building the network grid system and shares similar length with each other. As for Zurich, a large amount of branch streets do not serve as bridges, which results in their low betweenness centralities.

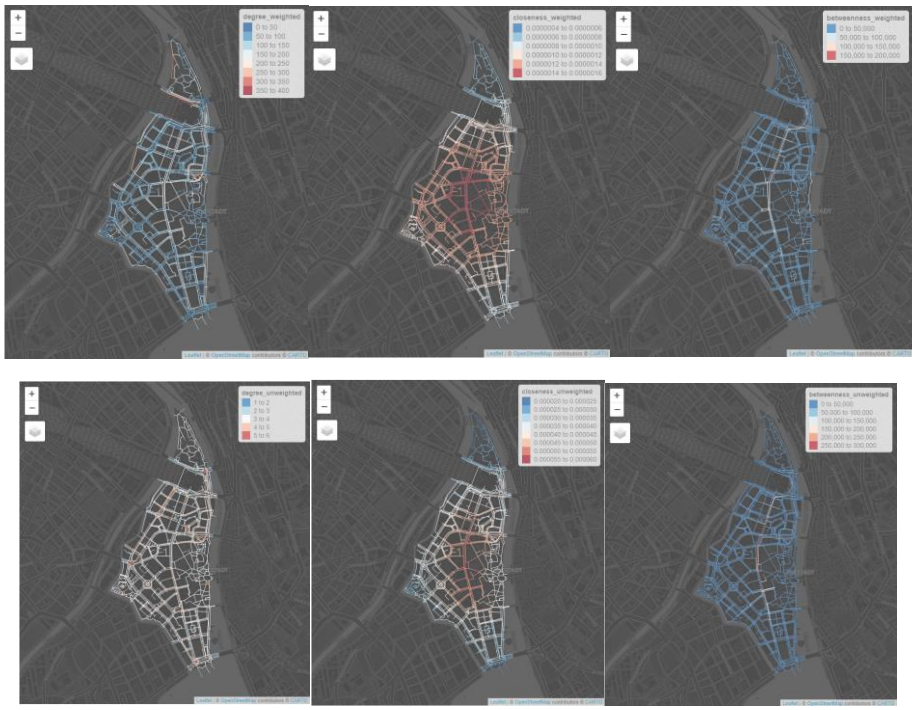


Figure 11. Centralities of the Zurich street graph in edges (top: weighted, bottom: unweighted)
Note: the centrality of each edge is calculated by the mean centrality of its start and end nodes

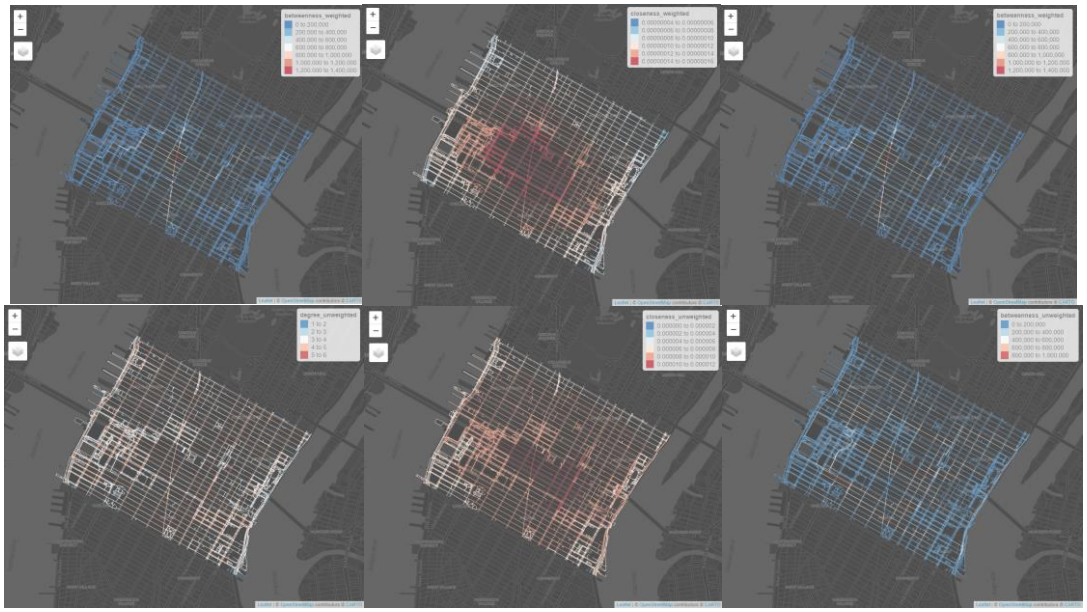


Figure 12. Centralities of the New York street graph in edges (top: weighted, bottom: unweighted)
Note: the centrality of each edge is calculated by the mean centrality of its start and end nodes

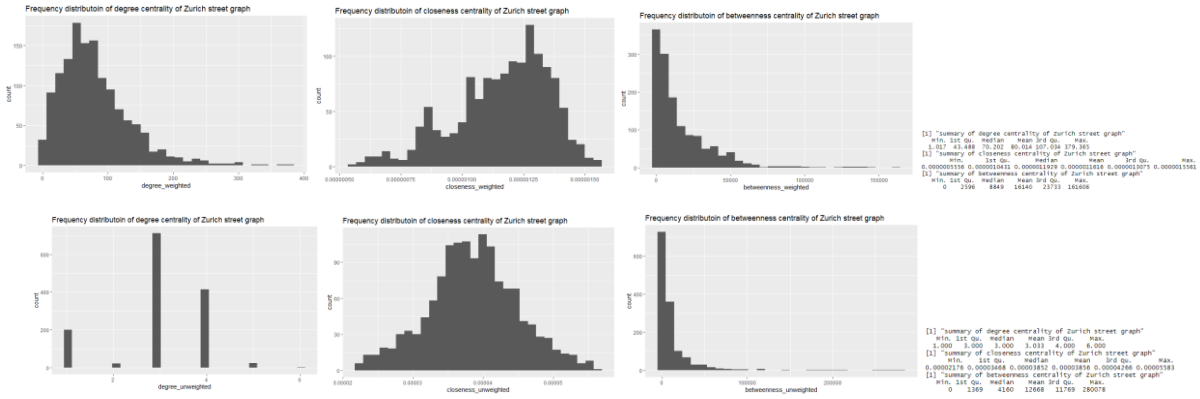


Figure 13. Frequency distributions of the centralities of Zurich street graphs (top: weighted, bottom: unweighted)

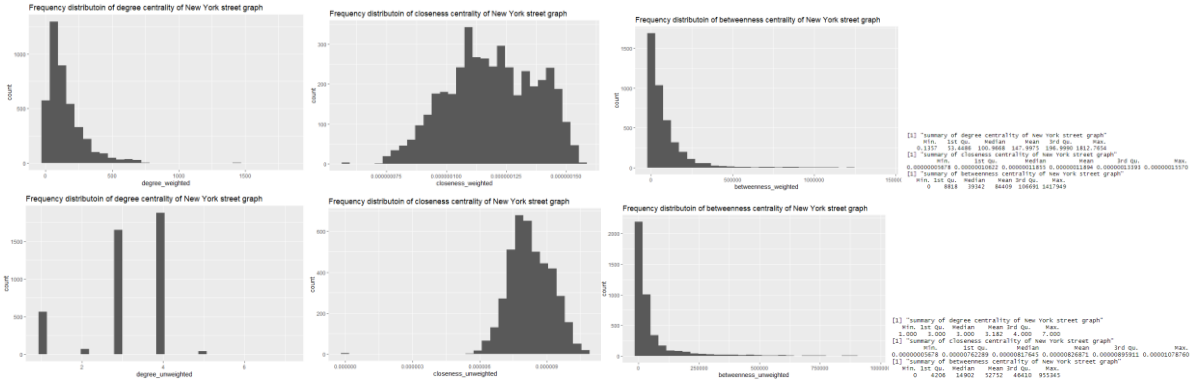


Figure 14. Frequency distributions of the centralities of New York street graphs (top: weighted, bottom: unweighted)

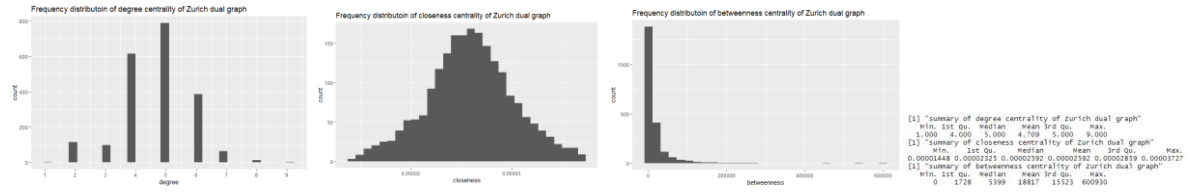


Figure 15. Frequency distributions of the centralities of Zurich dual graph

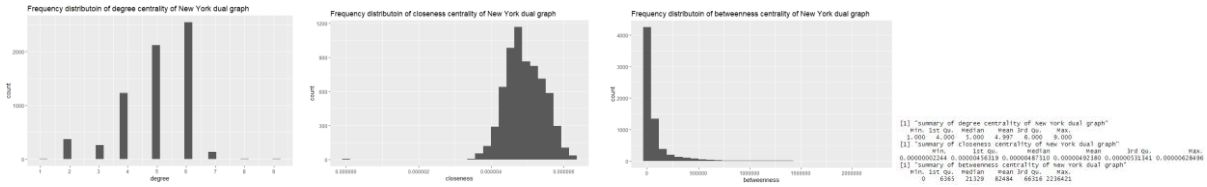


Figure 16. Frequency distributions of the centralities of New York dual graph

3. Time series analysis

3.1. Analytics of the AirPassengers data

3.1.1. Decomposition analysis

The ACF plot has a decreasing trend along the lag, which suggests that AirPassengers time series data has autocorrelation on itself and there is a trend pattern in the data. From the visualization of AirPassengers along the time, the data seems to have an upward trend, and a seasonal pattern is also observed (Figure 17). For the detection of specific patterns, further analysis is needed.

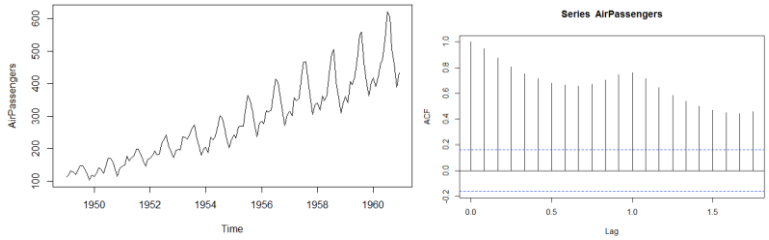


Figure 17. AirPassengers time series data

To decompose the AirPassengers data, three different window sizes (periodic, 3, 5) are applied. Figure 18 shows the decomposed AirPassengers data, and it is obvious that the data has an upward trend pattern and a seasonal pattern with an annual seasonality. But there seems to be some seasonal patterns in the remainder whose decomposing window size is periodic, and the remainder might not be merely white noises, thus the decomposition results will be investigated based on the remainder later. Figure 19 suggests that the data is based on an additive process because the additive model of three components is similar with the original AirPassengers time series data, while the multiplicative model cannot reveal the original data.

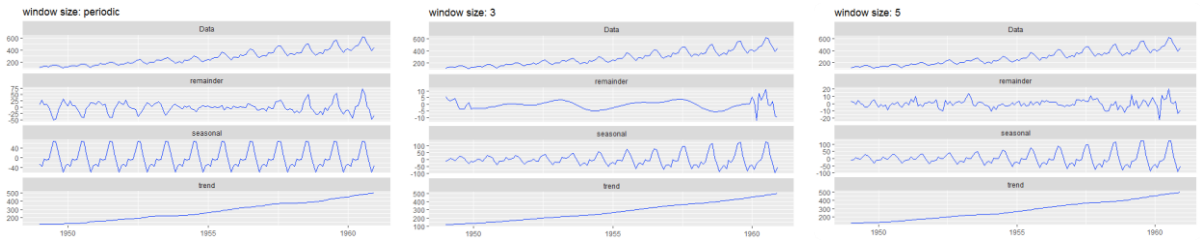


Figure 18. Decomposition of AirPassengers data with different window sizes

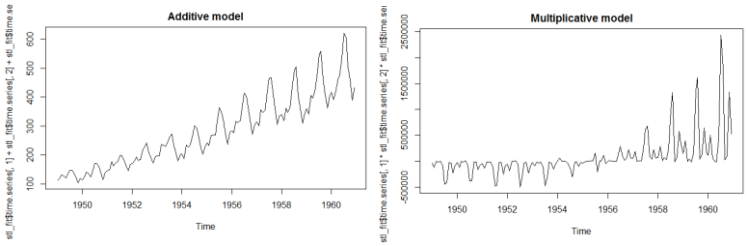


Figure 19. Composition of three components with different models

To qualify the decomposition results, ACF and tests on the distribution of remainders are performed. If the remainder is white noise, its ACF should be around 0, and it should be normally distributed. The remainder with window size of 5 seems to be white noise. Q-Q plots show that none of the remainders obey normal distribution. The Shapiro-wilk normality test also validates it because all three remainders have p-value less than 0.05 (Figure 20).

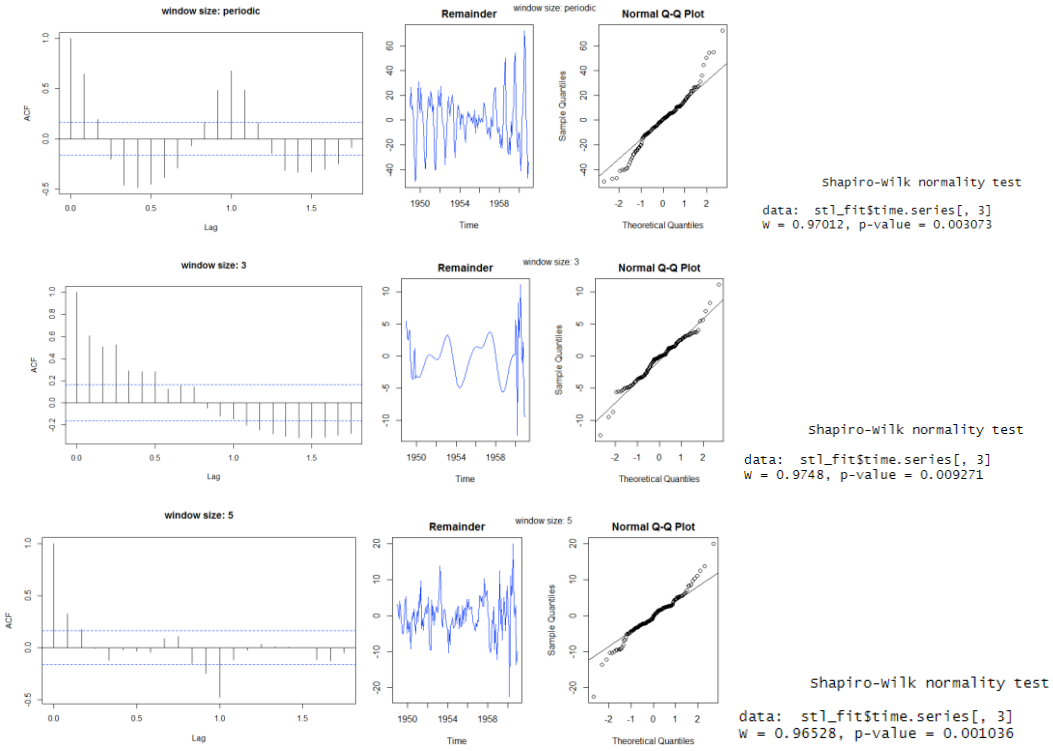


Figure 20. Analysis on the remainder

3.1.2. Autoregressive modeling

Exponential smoothing model and ARIMA model are applied to predict the AirPassengers data. The subset of AirPassengers data from 1949 to 1959 is used as the training dataset, the subset of AirPassengers data in 1960 is used as the testing dataset. The AIC and BIC of indicate that the ARIMA model (AIC: 899.9, BIC: 905.46) fits the training dataset better than exponential smoothing model (AIC: 1244.458, BIC: 1296.348). The residuals of two models seem not to be white noise, which indicates that two models cannot capture all the dynamics in the training dataset well. Combined with the root-mean-square deviation (RMSE), mean absolute percentage error (MAPE) and mean absolute scaled error (MASE), we can find that though the ARIMA model fits the training dataset better, the exponential smoothing model predicts the testing dataset more accurately (Figure 21, Figure 22).

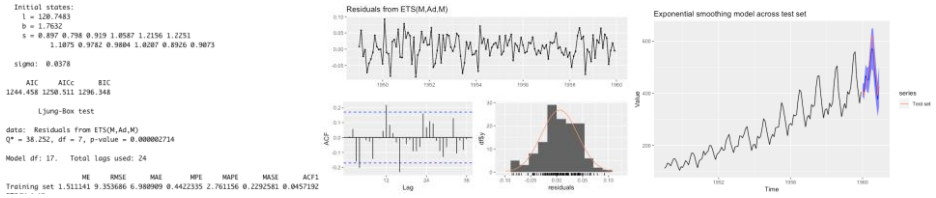


Figure 21. Exponential smoothing model

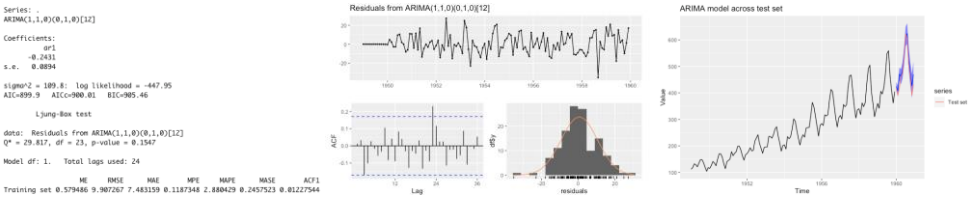


Figure 22. ARIMA model

The combinations of STL decomposition and autoregressive models are also applied. For the ARIMA model, the same pdq order and PDQ seasonal order are applied in the combination model. The results show that the autoregressive models with STL decomposition fit the training dataset better than previous models, but the residuals are still not merely white noises. As for the prediction performances on the testing data, the STL decomposed autoregressive models also have less prediction errors (Figure 23, Figure 24).

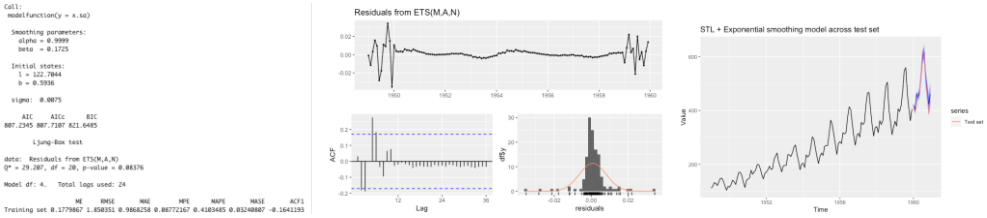


Figure 23. STL + Exponential smoothing model

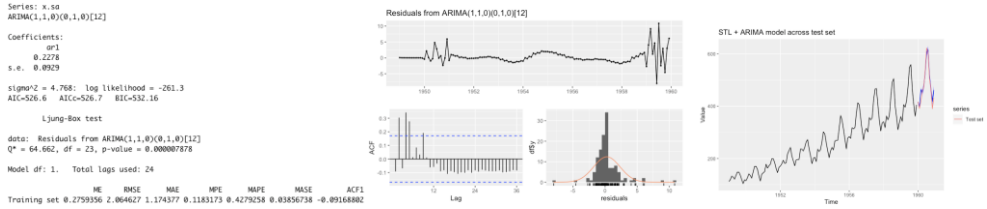


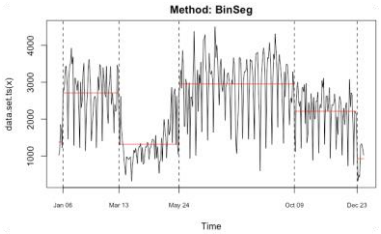
Figure 24. STL + ARIMA model

3.2. Analytics of the 2020 data

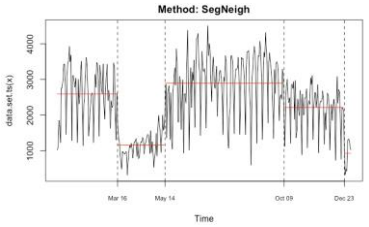
In the educational guess, there might be two break points in the Zurich pedestrian flow data in 2020, which were at the end of March and in the beginning of October, because the government imposed the lockdown and strict public health measures to restrict the pedestrian flow in these two periods.

3.2.1. Break points detection with changepoint package

The changepoint package is applied to detect the break points, and BinSeg method and SegNeigh method are used here. The BinSeg method discovers five break points; the SegNeigh method discovers similar four break points, and it does not determine there is a break point during January as the BinSeg method does (Figure 25). The break points detected are close to the educational guessed break points at the end of March and in the beginning of October, furthermore, there are two more break points detected in the late of May and December.



date_start <date>	ped_mean <dbl>
2020-01-01	1376.4000
2020-01-06	2711.0448
2020-03-13	1315.4028
2020-05-24	2948.8551
2020-10-09	2209.5200
2020-12-23	927.3333



date_start <date>	ped_mean <dbl>
2020-01-01	2591.7200
2020-03-16	1152.4576
2020-05-14	2890.5338
2020-10-09	2209.5200
2020-12-23	927.3333

Figure 25. Break points detected by the changepoint package (top: BinSeg method, bottom: SegNeigh method)

3.2.2. Break points detection with strucchange package

Before applying the strucchange package to detect break points, the CUSUM test and F test are used to find the optimal number of break points (Figure 26). BIC is also used and all of them suggest that there should be three break points. This package finds the same break points with those detected by the SegNeigh method in changepoint package, the only difference is that the break point in 23rd December is not detected here (Figure 27). In general, the results generated by two packages fit the educational guess and there are 2 to 4 more break points discovered. According to the time series plot, it is reasonable that there is a break point in May because the pedestrian flows increased greatly to the same level in January. As for the break point in December, there might be a break point, but due to the limited data at the end of December, we cannot draw a conclusion that it is absolutely a break point, and it is also not detected by the strucchange package.

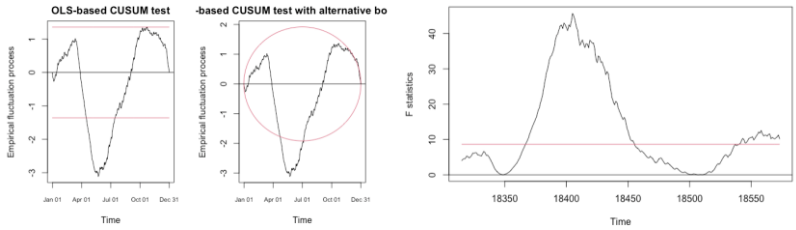
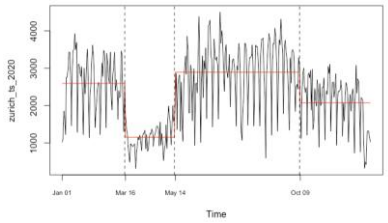


Figure 26. Statistical tests to identify potential change points (left: CUSUM test, right: F test)



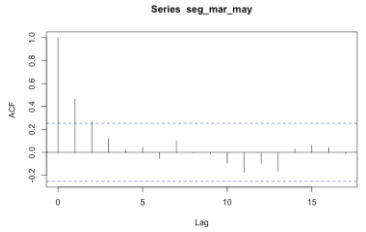
date_start <date>	ped_mean <dbl>
2020-01-01	2591.720
2020-03-16	1152.458
2020-05-14	2890.534
2020-10-09	2072.143

Figure 27. Break points detected by the strucchange package

3.2.3. Compare different segments with ACF

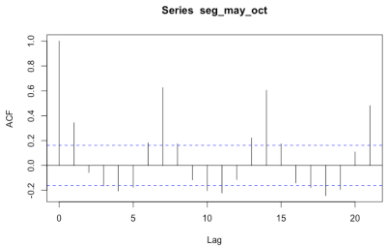
The break points detected by the strucchange package are used to select segments for comparison. The first segment is from 16th March to 14th May, and the second segment is from 15th May to 9th October. The ACF plots show that the pedestrian flow in the first segment is distributed randomly, while in the second segment, it seems to have a seasonal pattern. Based on the ACF features, the sum

of squared of first ten autocorrelation coefficients (acf10, diff1_acf10, diff2_acf10) also indicates that the pedestrian flow in the second segment is more autocorrelated with itself (Figure 28, Figure 29). This result aligns with the reality. In the middle of March 2020, the residents in Zurich faced an unusual travel situation due to the lockdown, which resulted in the low autocorrelation coefficient; from the middle of May, the weather got warmer and the government lifted the travel restrictions, so residents could come back to their normal life and commute regularly, which led to the seasonal pattern of pedestrian flow.



acf1	acf10	diff1_acf1	diff1_acf10	diff2_acf1	diff2_acf10
0.4641798	0.3304538	-0.2472719	0.2509664	-0.5921941	0.7708779

Figure 28. ACF of Zurich pedestrian data from break point 1 to break point 2



acf1	acf10	diff1_acf1	diff1_acf10	diff2_acf1	diff2_acf10
0.3432234	0.7303544	-0.1920552	0.6861183	-0.4785868	0.9223734

Figure 29. ACF of Zurich pedestrian data from break point 2 to break point 3