SCHOOL OF ENGINEERING AND DESIGN

TECHNISCHE UNIVERSITÄT MÜNCHEN

Semester's Thesis in Mechtronics and robotics

A resilience based approach for taxi travel demand prediction under disruptive events

Author:	Shengkai Huang
Supervisor:	Prof. Dr. Constantinos Antoniou
	M. Sc. Qinglong Lu
Submission Date:	09.02.2023

I confirm that this semester's thesis in mechtronics and robotics is my own work and I have documented all sources and material used.

Munich, 09.02.2023

Shengkai Huang

Acknowledgments

First of all, I would like to acknowledge my supervisor, Prof. Dr. Constantinos, for providing me with this research opportunity in the field of Transportation Systems Engineering.

I would particularly like to thank my supervisor, M. Sc. Qinglong Lu, for his guidance through each stage of the process. His expertise was invaluable in formulating the research questions and methodology.

In addition, I would also like to thank the university, the Technical University of Munich, for providing me with valuable learning opportunities.

Finally, I would like to thank my parents for providing financial support for my studies in Germany.

Abstract

The outbreak of the COVID-19 pandemic has brought huge impacts and changes to human mobility, which directly impacted the urban transportation system. This study explored the recovery patterns of urban taxi system in the post-pandemic era on the basis of data for taxi trips, which includes taxi origin and destination information. We used the clustering method to merge OD data with similar trends and build a resilience models for each cluster separately based on the time nodes of public policy release. The resilience model quantifies recovery patterns across community areas with similar loss rates and recovery rates within the same cluster. By determining the cluster class to which the unknown test data belongs, the corresponding resilience model can be used to predict its recovery pattern. Combining natural and man-made destructive events, the study explores the impact of both on urban transportation systems and tries to reveal some important pointers for policymakers that could potentially aid in developing urban transportation policies during the future pandemic.

Contents

1. Introduction

1.1. Motivation and Background

In recent years, there has been a growing interest in the resilience of systems during and after disruptive events such as hurricanes, earthquakes and pandemics. Whether these events are predictable or unpredictable, natural or man-made, they all have far-reaching effects on the performance and availability of transportation systems. The American legal theorist and economist Richard Allen Posner divides negative disruptive events into four categories [1]:

1 Natural catastrophes (epidemics, volcano eruptions, meteorite impact, etc.)

2 Scientific accidents or laboratory accidents (e.g. release of bacteria)

3 Unintended man-made catastrophes (climate change, nuclear accidents, social upheavals, economic crises, corruption, political structures, food shortages, "alien species", etc.) and

4 Intentional, man-made catastrophes (cyber wars, terrorist attacks, etc.)

The most common disruptive events to transport systems are natural disasters, and pandemics are no exception. However, unlike disasters such as hurricanes and earthquakes, the effects of pandemics are relatively long-lasting, often lasting months or even years, as in the case of COVID-19 pandemic. It had a major impact on taxi travel. Many people were reluctant to use taxis or other forms of public transportation because of the health risks associated with the virus. This directly led to a sharp drop in demand for taxi services. Additionally, the public measures taken to contain or mitigate the spread of the virus, such as stay-at-home orders and the closure of non-essential businesses, restricted people's mobility, further exacerbated the decline in demand for taxi services and impacted transportation systems.

Typical research on the resilience of transportation systems is based on disasters such as hurricanes and earthquakes, which are characterised by high intensity and destructive but short duration, typically 1-2 weeks. And public policy responses to such short term disasters often have no negative impact on urban transportation systems. However, public health policies in response to the pandemic restrict people's mobility, reduce travel demand, and the release time is unknown. So it's often more complex than a single disruptive event. This study explores the impact of longer lasting pandemics and public health policies on the resilience of taxi systems by combining natural and man-made disruptive events.

1.2. Concept and Framework of Resilience

The concept of resilience originated in engineering and construction and is defined as the degree to which a system may react adversely during a disruptive event. It is the ability of systems to absorb and recover from the effects of disruptive events without fundamental changes in function or structure [2]. Bruneau et al. have developed a conceptual framework to identify and quantify the extent of resilience [3]. This framework can be illustrated by the following triangle.

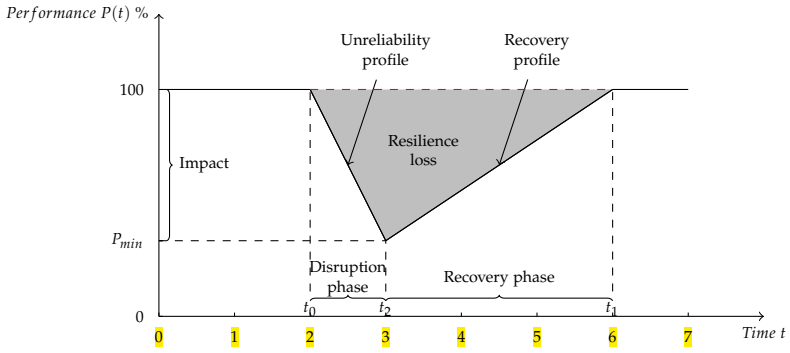


Figure 1.1.: Resilience triangle

In order to quantify resilience, it is necessary to examine the variation of performance along the time of disturbances. This approach is based on the notion that a measure of performance P(t) varies with time. Specifically, performance can range from 0% to 100%, where 100% means no degradation in of service and 0% means no service is available. Normally, system performance will not drop to 0% after disruptive events, but will reach a minimum point P_{min}, and then steadily increase until it reaches a new steady state. The area of the shaded triangle is the resilience loss. Mathematically, this loss is represented as follows [3]:

$$R = \int_{t_0}^{t_1} [100 - P(t)]dt \tag{1.1}$$

In the field of taxi transport, the number of taxi trips is a good performance indicator, more precisely the origin-destination (OD) demand, which is discussed in more detail in section 3.2. It can reflect well the travel demand of people and the activity of taxis. For a stable urban taxi system, the OD demand from area A to area B should fluctuate within a reasonable range. When a disaster occurs, the number of trips will quickly drop to an extremely low value and then recover.

1.3. Resilience Quantification

Unreliability and recovery are two areas that resilience studies focus on. Typically, unreliability and recovery processes all follow similar patterns. That is, the rates of evacuation or recovery

follow an S-shape [4]. Such process can be modelled using a logistic function, which will be discussed in more detail in Section 2.5. This is done to build a more accurate resilience model, as shown in the figure 1.2.

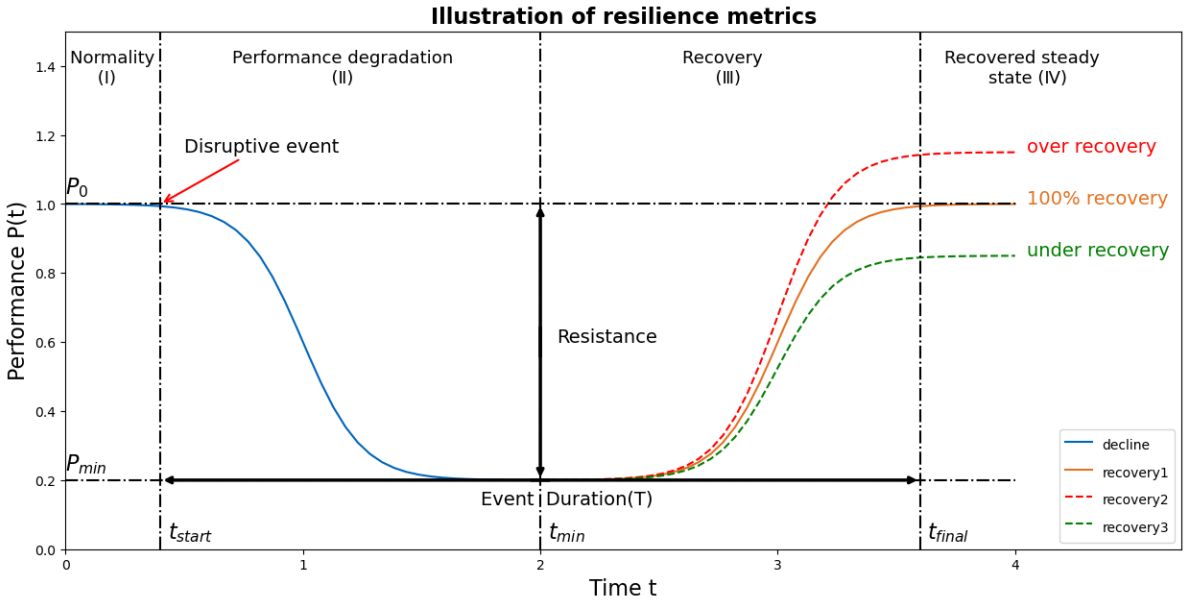


Figure 1.2.: Illustration of resilience metrics

As the diagram above shows, a complete resilient cycle should include the following four states:

1 Normality (I): Baseline, or the original state P_0 in which the system was operating normally prior to the occurrence of the disturbing event.

2 Performance degradation (II): Unreliable state, when the system degrades to P_{min} at time t_{min} after a disturbing event.

3 Recovery (III): Recovery state, when the system improves its performance functions as a result of restorative efforts.

4 Recovered steady state (IV): System performance reaches a new steady state after successfully completing the recovery state.

Suppose t_{start} is the time when the pandemic starts and t_{final} the time when OD demand has recovered to a new state of equilibrium. This new steady state may also be above or below pre-pandemic baseline demand. t_{min} is a time between t_{start} and t_{final} when the system starts to recover.

A resilient system includes high resistance, recovery percentage and event duration. Their concepts and formulae are shown in the table 1.1.

Table 1.1.: Resilience Metrices Description

Metrics Description	Formula
Resistance(Re): is the level of deviation of the performance from the normal state. It indicates the degree of susceptibility to disturbance.	$Re = \frac{P_{min} - P_0}{P_0}$
Event Duration(T): is the time it takes from the occurrence time of the disruptive event to the time that the system reaches a new steady state	$T = t_{final} - t_{strat}$
Recovery Percentage(RP): is the percentage of improvement or reduction of performance after a disruptive event. Recovery percentage indicates the impact level of an event on the system performance.	$RP = \frac{P_{final} - P_{strat}}{P_{start}}$

2. Methodology

2.1. Experiment Flowchart

Based on the OD pair concept of above, the following research process is designed. Firstly, the number of trips will be aggregated based on their OD positions and a comprehensive visual statistical analysis of the data collected during different Covid periods will be carried out. The purpose of this study is to investigate whether there is a commonality in travel between community nodes, i.e. to use travel data between any community to build a resilience model, and use this model to predict the change in taix trips in other OD pairs after the impact of COVID-19. In order to obtain a more accurate resilience model, it is necessary that the OD pairs contain sufficient data samples. Therefore, the OD pairs with too few trips were discarded in this study.

The OD pairs obtained after screening are divided into a training data set and a test data set. For the training data, further clustering is required and those datasets with similar trends are classified into the same cluster. The resilience model is then modelled separately for each cluster using logistic regression. Finally, the test data set is classified into known clusters and the resilience model of the corresponding cluster is used to make predictions. The methods used in clustering and logistic regression are explained in more detail in the following chapters.

The entire analysis process is based on the following flowchart 2.1:

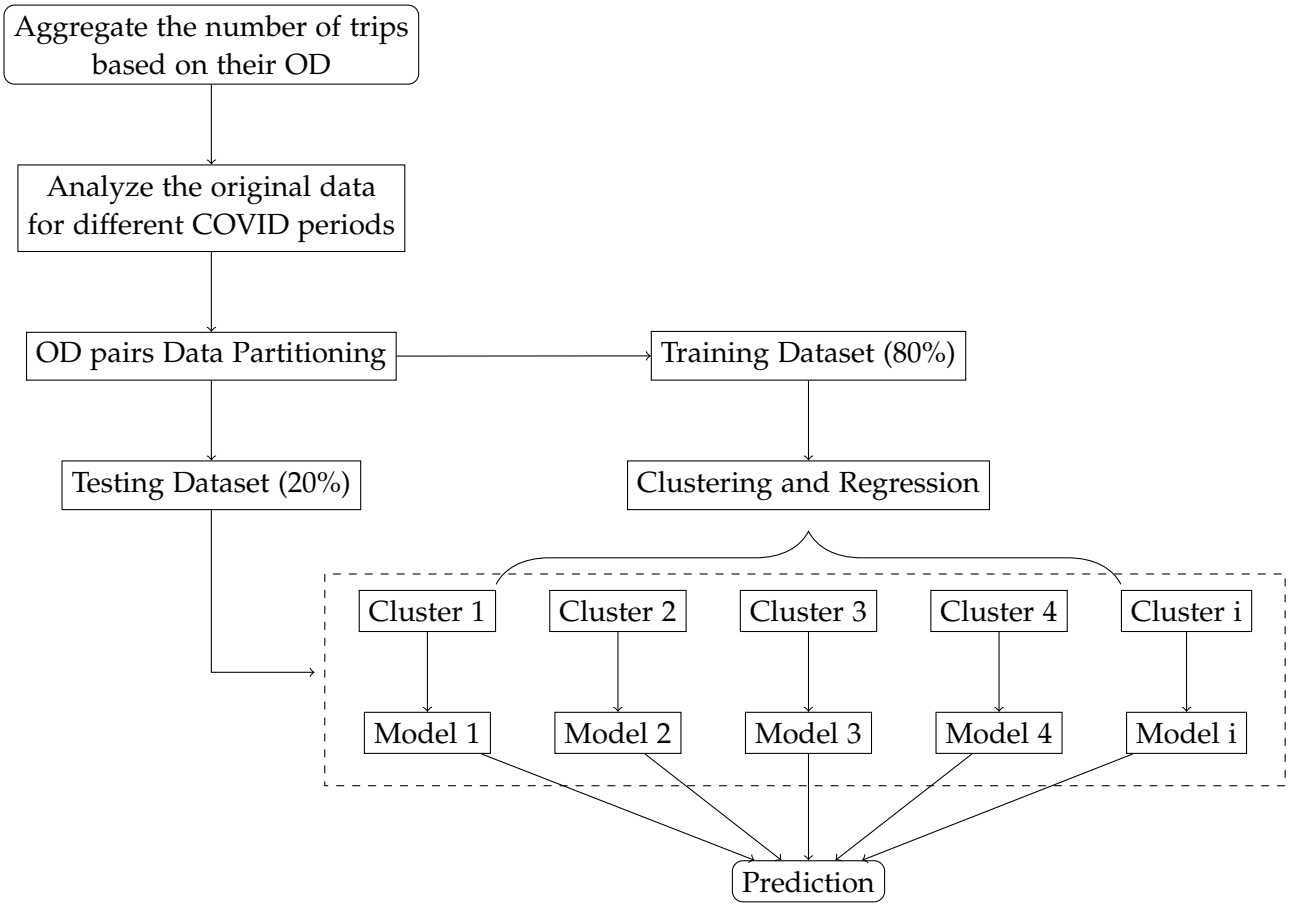


Figure 2.1.: Experiment flow chart.

2.2. Degree Centrality

To reflect the importance of community nodes, the concept of degree centrality is introduced, which is measured by the total number of direct links with the other nodes [5]. By considering the number of taxi trips between two nodes in a network, a more meaningful network metric is obtained. We define the weighted degree centrality $(C_D(i))$ of a node i as follows:

$$C_D(i) = \sum_{j=1}^{N} A_{ij} w_{ij}, \tag{2.1}$$

where w_{ij} is the weight of a link, estimated as the number of taxi trips between two nodes i and j, and A_{ij} is the adjacency matrix (Figure: 2.2) of the network, whose elements take the value of 1 if two nodes i and j are connected, and 0 otherwise.

	A	B	C	D
A	0	1	1	0
B	1	1	1	0
C	1	1	0	1
D	0	0	1	0

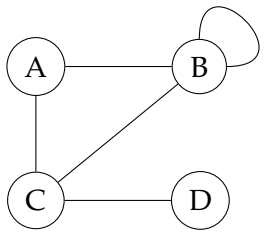


Figure 2.2.: Adjacency Matrix.

Typically, the community area with a high degree of centrality corresponds to areas with city centers, train stations, airports, and other commercial areas. This allows we to identify the community nodes with high degree centrality in Chicago. The changes in the degree centrality of these community nodes before and after Covid-19 will help to understand the changes in the urban traffic structure more intuitively.

2.3. Dynamic Time Warping, DTW

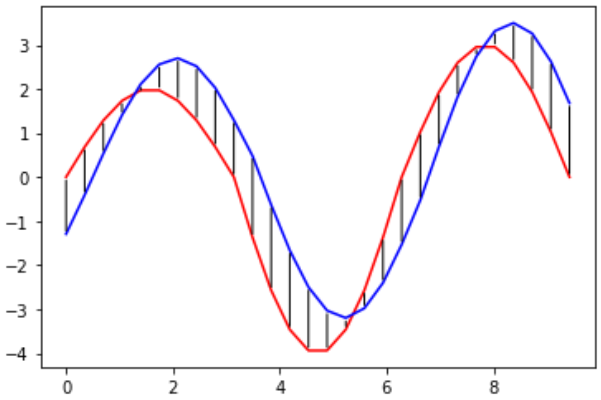
In order to complete the clustering of long time series based on the spatial similarity, this study adopts the method, the Dynamic Time Warping algorithm (DTW). It is a well-known algorithm for measuring the similarity between two time series, especially for time series of different lengths and rhythms. This algorithm was proposed around the 1970s and was first used to solve with the problem of automatic speech recognition and classification [6], to cope with different speaking speeds. Later it has been widely used in many fields: isolated word recognition [7], handwriting and online signature matching [8], sign language and gestures recognition [9].

Consider two time series **Q** and **C** where,

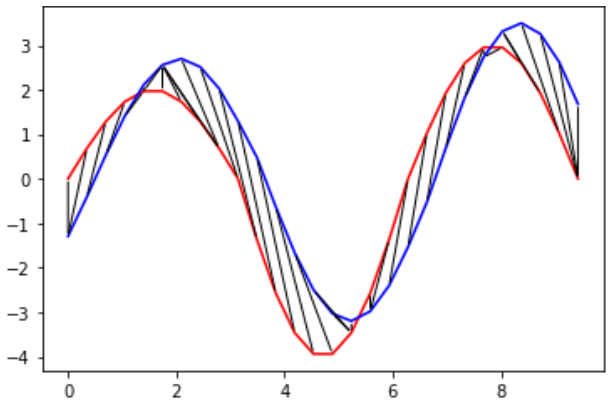
$$\mathbf{Q}: \ q_1, q_2, q_3 \cdots q_n,$$

$$\mathbf{C}: \ c_1, c_2, c_3, \cdots c_m$$

The simplest way to compare their similarity is to calculate the Euclidean distance from point to point and add them up. The smaller the sum of the distances, the more similar the two sequences are. The figure 2.3a illustrates this concept.



(a) Euclidean distance.



(b) DTW distance.

Figure 2.3.: Two time series similarity measurement.

Although the Euclidean distance is a very simple and intuitive method, if the sequences are out of phase, the result obtained by calculating the Euclidean distance will be much greater than the actual minimum distance. For example, both sequences have very similar shapes overall, but the shapes are not aligned on the x-axis. So before comparing their similarities, one of the sequences needs to be shifted down the time axis to achieve better alignment. This method of "corresponding a point at one time in one sequence to points at several successive times in another sequence" is called time warping. DTW is an effective way to achieve this warping distortion. It computes the similarity between two time series by lengthening and shortening the time series. The figure 2.3 shows this different adjustment.

To calculate the distance between the two, first draw a two-dimensional array $[n \times m]$, where each point $w(i, j)$ in the array represents the distance between $\boldsymbol{Q_i}$ and $\boldsymbol{C_j}$:

$$w(i, j) = \sqrt{(q_i - c_j)^2} = |q_i - c_j| \tag{2.2}$$

Theoretically, it is possible to exhaustively enumerate all possible warping forms of the two

sequences, and calculate the distance between them one by one, and the one with the smallest distance is the required warping. However, the amount of computation is too large, so dynamic programming is used to perform the computation efficiently.

Steps of dynamic programming algorithm:

1 Compute the distance matrix between individual points in two sequences.

2 Find a path from the bottom left corner to the top right corner of the matrix such that the sum of the elements on the path is the smallest.

Assuming the warp matrix is M, the length of the shortest path from the lower left corner of the matrix $\boldsymbol{M}(1,1)$ to any point $\boldsymbol{M}(i,j)$ is $\boldsymbol{L}_{min}(i,j)$.

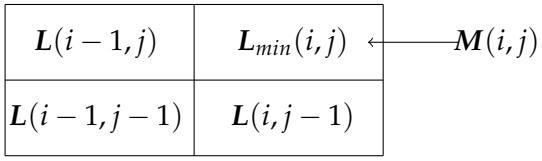


Figure 2.4.: Path matrix.

Using the recursive algorithm to find the shortest path length:
Starting conditions:

$$\boldsymbol{L}_{min}(1,1) = \boldsymbol{M}(1,1)$$

Recurrence relation:

$$\boldsymbol{L}_{min}(i,j) = \boldsymbol{min}\{\boldsymbol{L}_{min}(i,j-1), \boldsymbol{L}_{min}(i-1,j), \boldsymbol{L}_{min}(i-1,j-1)\} + \boldsymbol{M}(i,j) \tag{2.3}$$

This algorithm is based on finding a path through several grid points in this grid, and the grid points passed by the path are the aligned points for calculating the two sequences. The smaller the distance of the final regularised path, the greater the similarity between the two original time series.

The example in the figure below simply shows how dynamic programming can be used to find the best path.

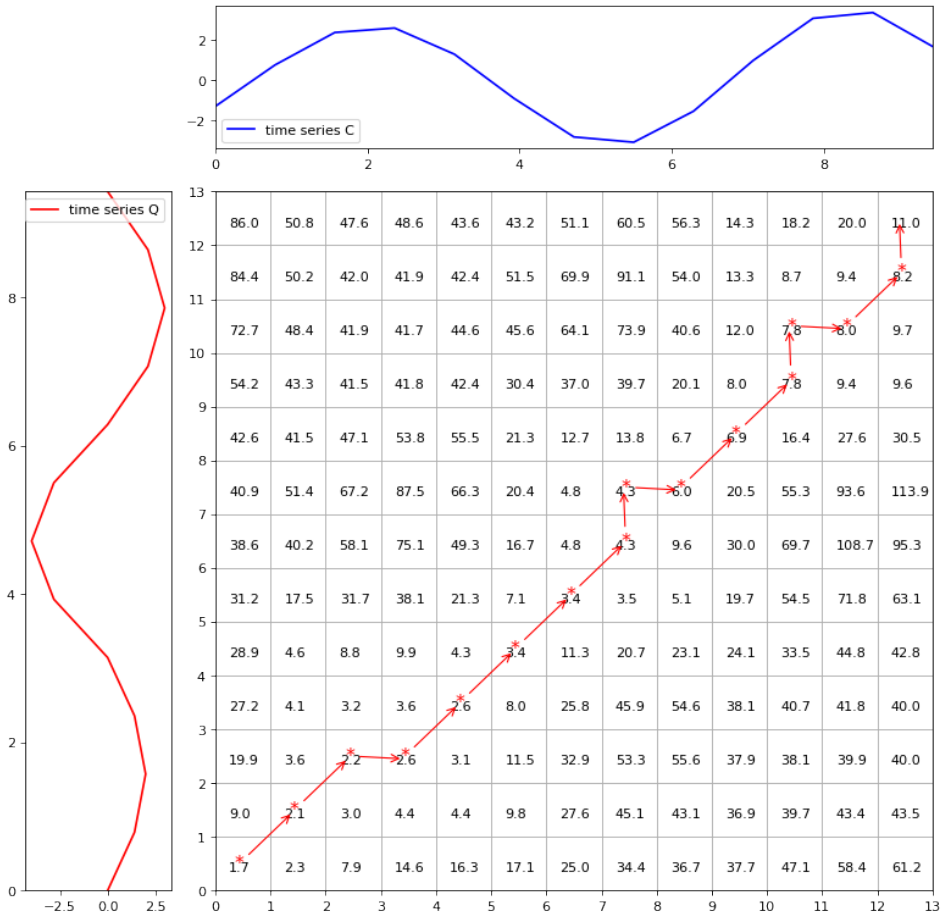


Figure 2.5.: Warping matrix construct and the optimal warping path search.

The red line represents the optimal path between the time series **Q** and **C**, e.g.

$$(1,1),(2,2),(3,3),(4,3)\cdots(12,11),(13,12),(13,13)$$

and its length is 11. Although DTW is a good method for measuring time series similarity with high accuracy and robustness, the space and time complexity of its computation is related to the length of the two sequences, which is $\boldsymbol{O}(mn)$. Obviously, it requires more time and memory than the Euclidean distance, especially when comparing long time series.

2.4. LB_Keogh Distance

To speed up the DTW calculation process, an approximate method is used to compute the Lower Bounding (**LB**). It can exclude most of the sequences that cannot be the optimal match, and the remaining sequences are compared one by one using DTW. There are several different functions for lower bounds on different time series distance measures, LB Keogh lower bound is used in this study. It was introduced in 2002 as the first non-trivial lower bound for

DTW [10] and it is still the fastest known technique for indexing DTW.

The distance between time series **Q** and **C** is defined as [10]:

$$LB_{Keogh}(Q,C) = \sqrt{\sum_{i=1}^{n} \begin{cases} (c_i - U_i)^2 & if\ c_i > U_i \\ (c_i - L_i)^2 & if\ c_i < L_i \\ 0 & otherwise \end{cases}} \tag{2.4}$$

where $\boldsymbol{U}_i$ and $\boldsymbol{L}_i$ are upper and lower bounds, i.e. an envelope for time series **Q**, which are defined as:

$$\boldsymbol{U}_i = max(x_{i-r}, \cdots, x_{i+r}),$$
$$\boldsymbol{L}_i = min(x_{i-r}, \cdots, x_{i+r})$$

for all $i \in \{1, \cdots, n\}$ and **r** is the radius of the envelope.

The following illustration is a visual intuition of LB_Keogh, a protective envelope is built around the red time series **Q**. The squared sum of the distances from each part of the blue candidate sequence **C** that does not fall within the bounding envelope, to the nearest orthogonal edge of the bounding envelope is returned as the lower bound of the DTW.

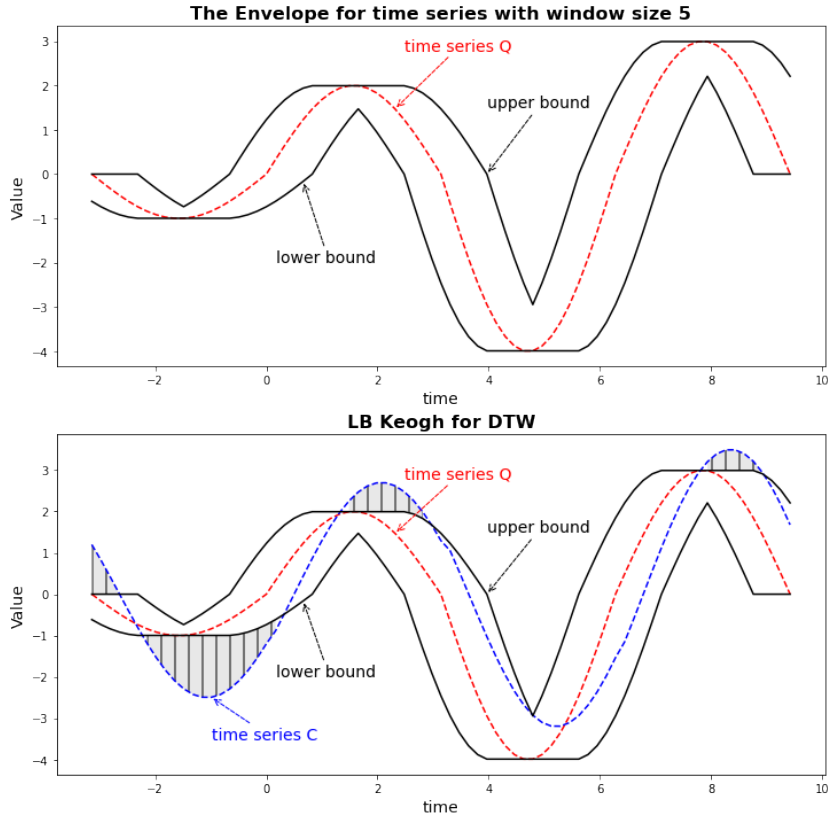


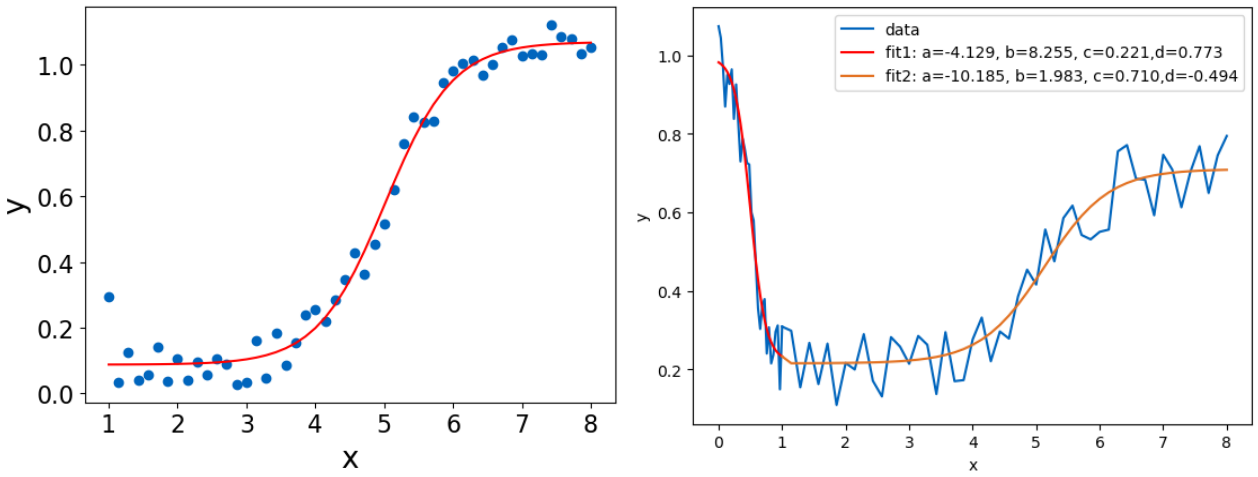
Figure 2.6.: Visualization of the LB Keogh lower bounding function (Q,C).

2.5. Logistic Function based Regression Model

The resilience model can be seen as a combination of two S-shaped curves [11], slightly reminiscent of the sigmoid function, and more generally, the logistic function, which was introduced by Belgian mathematician Pierre François Verhulst in 1838 for the purpose of analyzing population growth in Belgium [12]. Its most general equation is:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} \tag{2.5}$$

where L is the supremun of the values of function and k is the logistic growth rate of the curve.



(a) Logistic function.

(b) Fitting a random resilience model using a logistic function.

Figure 2.7.: Regression based on logistic function.

The figure above shows the most common logistic function graph. The initial phase of growth is approximately exponential. Then the growth rate levels off. Finally, when saturation is reached, the rate slows down and stops growing. Typically, the range of the logistic function should be between 0 and 1, representing the range from the worst service status to full recovery of the service. In practice, the trough of the recovery curve may be greater than zero (partial disruption of the system) and the peak at the end of the analysis period may be less than 1 (system not fully recovered during the COVID-19 epidemic period). To obtain a wider range of function values to fit more models, the function can be more generalised as:

$$f(x) = \frac{d}{1 + e^{a+bx}} + c \tag{2.6}$$

The range of the function is $(c, c + d)$, theoretically, it is possible to fit any range of S-shaped curves. This function is then used to model and predict taxi recovery activities.

3. Application with Data

3.1. Data and OD Networks

The city of Chicago is divided into 77 community areas [13]. City Hall adopted these as the official community boundaries and uses them for much of the official city government statistics, including taxi trip data. The data includes information such as pick-up location, drop-off location and travel time. The mobility of Chicago taxis between pick-ups and drop-offs in community areas can be represented as a dense network of origin-destination (OD) networks. A large dataset of Chicago taxi trips between 1 January 2019 and 31 October 2022 is employed in this study [14]. The taxi-trip management system in Chicago collects the geospatial information of the community areas and the date and time of the trips. The geospatial information on origin-destination (OD) pairs provides a platform for studying human mobility patterns. Any change in human mobility patterns is reflected in the OD locations of taxis. Therefore, we construct a weighted and undirected network with different community areas in the city of Chicago as the nodes of the network and the total number of trips between OD locations as the edge of the network. The pairs of nodes representing OD locations are connected if there is at least one taxi trip between two nodes. Edge weights can be aggregated over any time period such as daily or weekly taxi trips. Theoretically, there are $77 * 77$ different OD pairs that can be used as research objects.

3.2. The Original Data for Different COVID Periods

In the state of Illinois, one of the first cases was reported in Chicago when a person travelled to the city from Wuhan, China on 24 January 2020 [15]. In the early stages of the pandemic, there was not much impact on travel. Before 6 March, the number of journeys fluctuated around 45000, as the situation had not yet become pandemic. With the increase in reports of new cases and the introduction of a number of public health measures to deal with the pandemic, people began to commute less frequently. This change was reflected in the taxi travel data as a rapid decline in the daily number of trips between communities. The following two graphs (Figure 3.1 and 3.2) illustrate the number of trips per day between 01 February 2020 and 31 November 2021.

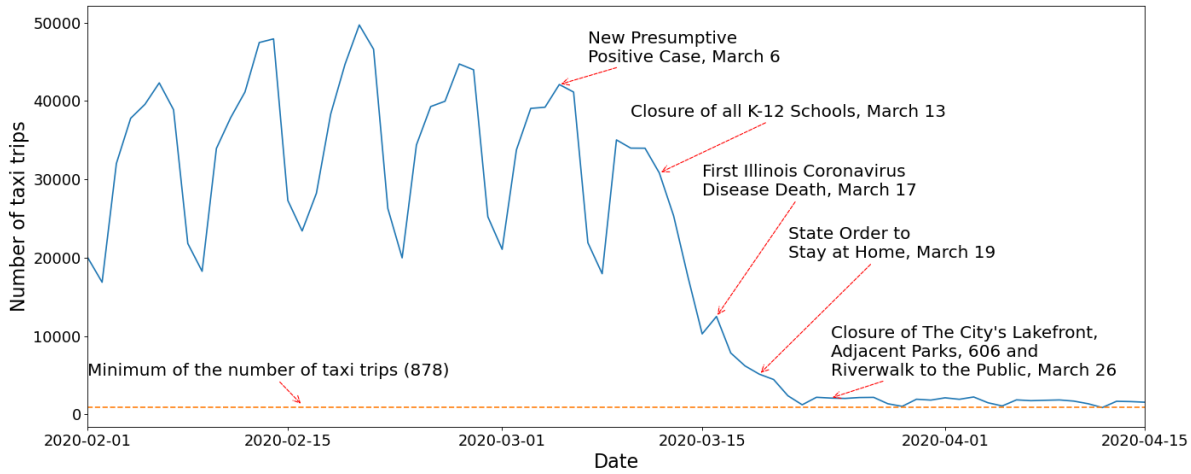


Figure 3.1.: Impact on the number of taxi trips in the initial phase of the COVID-19.

Over the next two weeks, starting on 9 March 2020, the daily number of taxi trips steadily declined as the number of COVID-19 infected cases began to rise. The steady decline in the average number of trips coincides with the closure of schools (13 March), the first Illinois coronavirus death (17 March), the state order to stay home (19 March), and the closure of the city's lakefront, adjacent parks, 606, and Riverwalk to the public (26 March) [16]. From 26 March 2020, the number of trips is close to the minimum of 878 and continues to fluctuate around a value of 1000. See the appendix A.2 for the City of Chicago Public Health Policy.

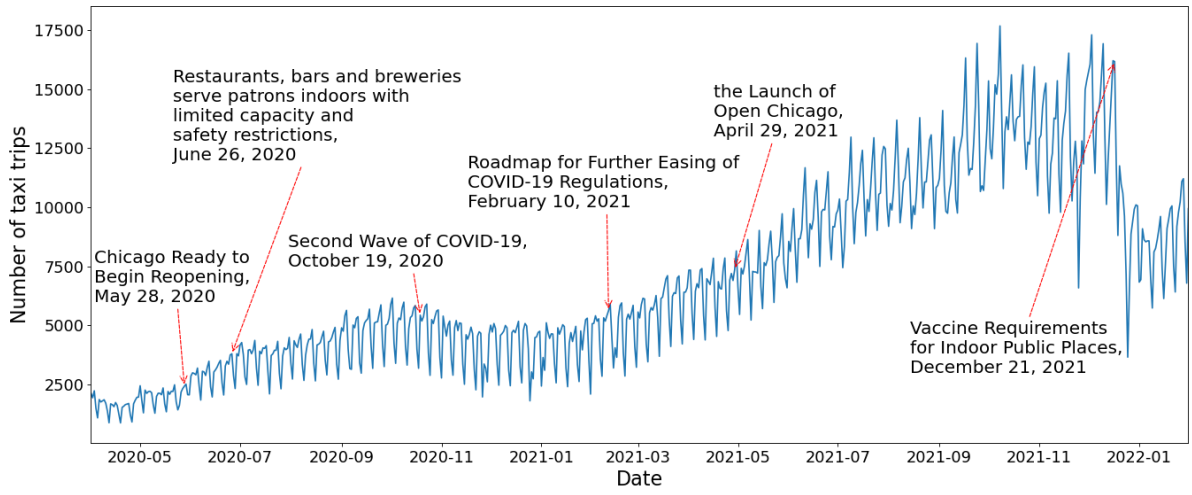


Figure 3.2.: Events during the recovery process.

From May 2020, the number of daily taxi trips began to slowly recover. When the Mayor announced that Chicago would open in June and that restaurants and bars would open with restrictions, the rate of recovery increased significantly. Although the second wave of the pandemic arrived again in October 2020, this time it did not have as big an impact as the first

wave, and the trip curve continued to grow after a slight dip. However, until 21 December, when the City of Chicago announced vaccination requirements for indoor public places, the number of taxi trips was again affected, as the mobility of those who were not vaccinated was restricted due to the constraints of the executive order. To date, the taxi industry in Chicago has not returned to pre-pandemic levels.

As can be seen from the two graphs above, the data shows a clear periodicity. It is easy to understand that people's travel needs are quite different on weekdays and weekends. In order to eliminate this volatility, the following discussion aggregates the data on a calendar week basis.

3.3. Aggregate Travel Data on a Weekly Basis

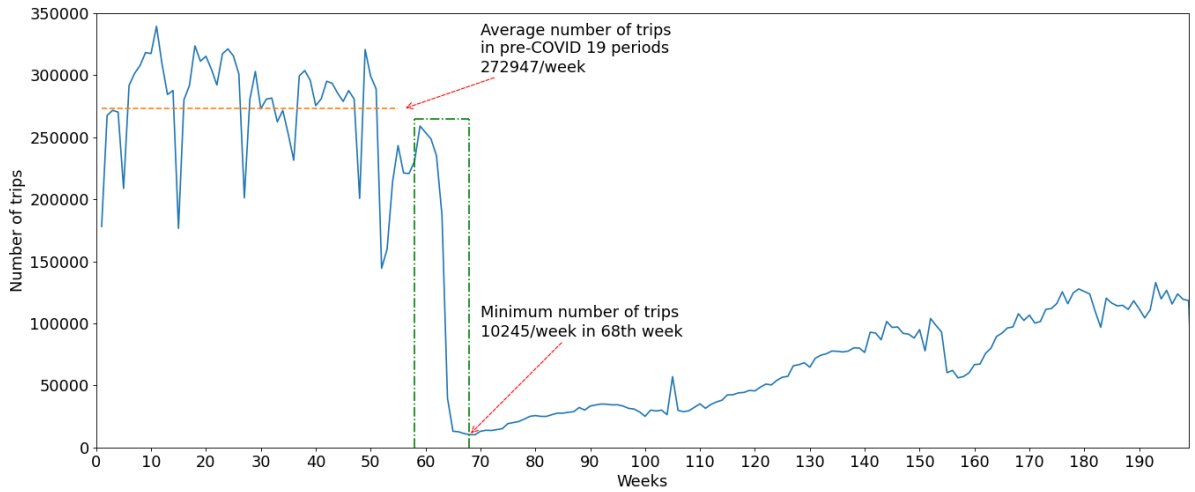


Figure 3.3.: Number of taxi trips per week.

In the first 50 weeks, the number of trips fluctuates around the average of 272947 times/week, as the situation with the spread of the virus has not yet become a pandemic. Although there is a sharp drop in the number of trips at certain times, e.g. 15, 27 and 48 weeks, this can be quickly recovered in a short period of time. This may be due to a public holiday or extreme weather conditions. In stark contrast to this was the impact between weeks 59 and 65. In less than two months, the total number of weekly taxi trips fell rapidly from 250,000/week to 10,000/week. This number reached a low of 10245/week in week 68. Detailed travel data are given in Table 3.1.

Table 3.1.: The number of trips affected by the pandemic

week	demand	week	demand
58	259050	64	13074
59	253793	65	12650
60	248624	66	11242
61	235127	67	10343
62	187022	68	10245
63	39759		

The resistance value can be easily calculated based on the usual average weekly travel times as follows:

$$Re = \frac{P_{min} - P_0}{P_0} = \frac{10245 - 272947}{272947} = -96.25\%$$

This means that taxi trips across the city were less than 4% of normal levels. This number intuitively show the impact of the pandemic on the taxi system.

3.4. The Weighted Degree Centrality of Nodes

Figure 3.4 shows the strength of nodes (community areas) for the selected weeks, which provides a reasonable representation of the level of interaction between origins and destinations in Chicago. The colour spectrum of the strength of community areas could be interpreted as the range of activities in different locations in Chicago. To make the colour difference in the figure less significant, the logarithm is taken for all values. We provide the visualisation for six weeks, from week 60 to week 65.

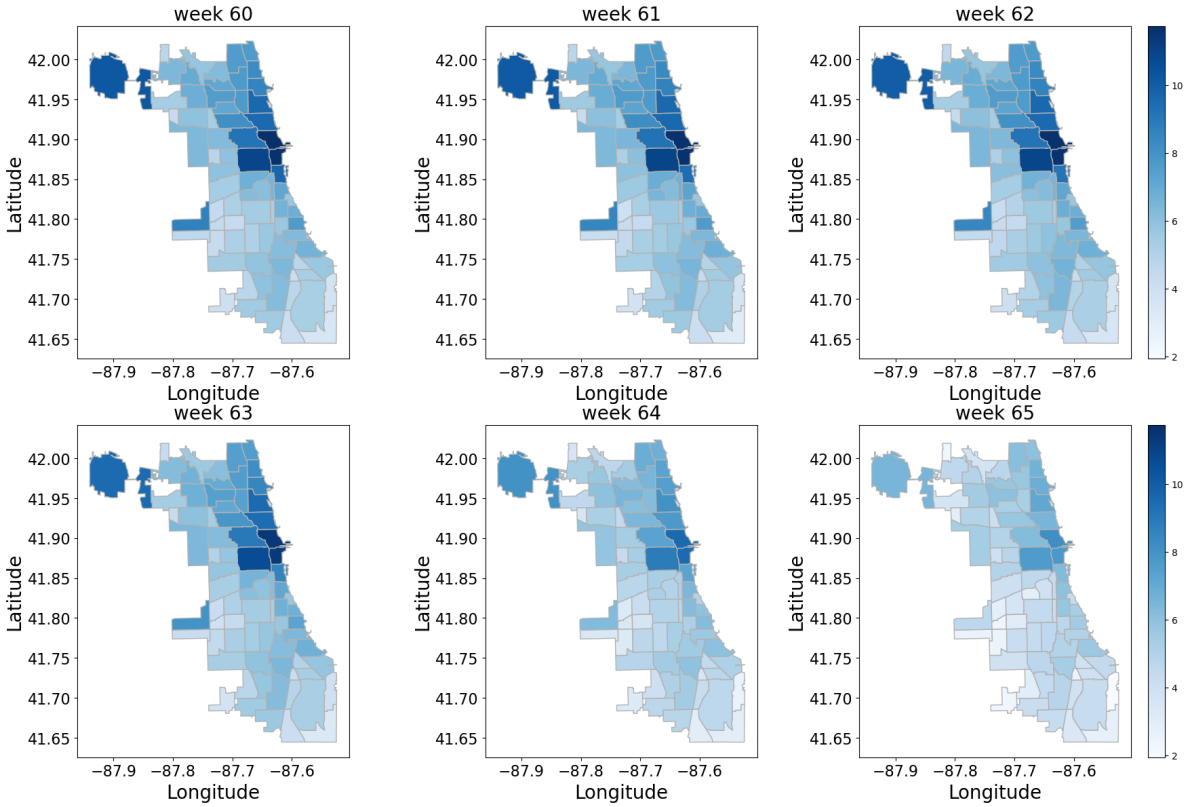


Figure 3.4.: The degree centrality of Chicago's neighbourhoods according to COVID-19 per week.

In week 60, there are higher values of node strength in the Near North Side (community node 8), Loop (community node 32), Near West Side (community node 28), O'Hare (community node 76), Near South Side(community node 33) and Lake View(community node 6). (The map of Community Areas in Chicago can be seen in the Appendix A.1). Since the Loop and the Near North Side are the commercial centre of Chicago, containing several retail establishments, restaurants and commercial workplaces, higher values of node strength in and around the Loop are justified. And the Near West Side has so many high schools, major hospitals and a university that it also has a high degree of centrality. A similar argument can be made for Chicago O'Hare, which is mainly occupied by O'Hare International Airport. The table and graph below show the data for 6 community areas with the highest degree centrality after the outbreak of the COVID-19 pandemic.

Table 3.2.: The strength of six communities with the highest degree of centrality.

Index	Name	60	61	62	63	64	65
8	Near North Side	11.73	11.67	11.63	11.4	9.58	8.17
32	Loop	11.65	11.64	11.59	11.34	9.53	7.79

Continued on next page

Table 3.2 – continued from previous page

Index	Name	60	61	62	63	64	65
28	Near West Side	10.95	10.97	10.92	10.67	8.90	7.67
76	O'Hare	10.22	10.15	10.03	9.51	8.05	6.51
33	Near South Side	9.71	9.44	9.16	8.63	7.28	6.28
6	Lake View	9.61	9.62	9.61	9.50	8.06	7.13

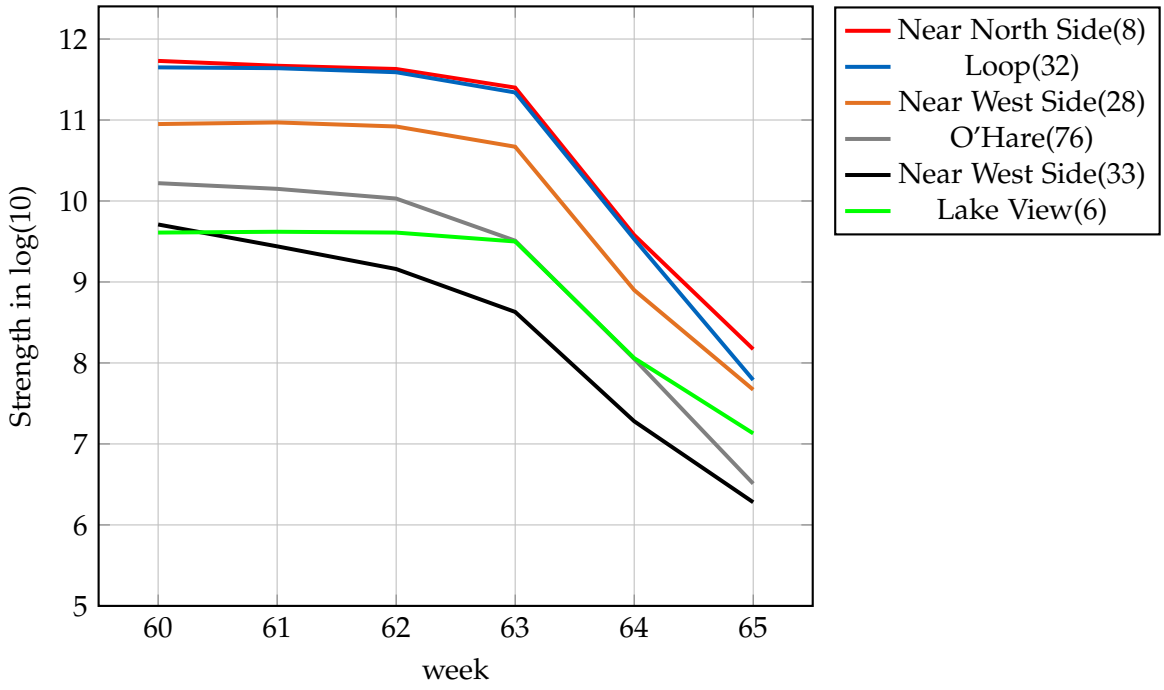


Figure 3.5.: The six communities with the highest degree of centrality.

The figure 3.5 shows that the strength of node 33 starts to decline from week 60. This includes some of Chicago's most famous structures: The stadium, Chicago's main convention centre, the museum campus, etc. These areas for leisure and entertainment are the first places where people consider reducing travel when faced with infectious disease. At the same time, the node strength of Chicago O'Hare starts to decrease, indicating less movement of domestic and international travellers. It's easy to understand, since the first cases usually come from international flights. Travelling in a relatively closed cabin for a long time greatly increases the likelihood of infection, so people will try to minimise air travel. The colour of the zones adjacent to the node Loop shifts towards the light blue spectrum from week 63, the corresponding date being 9 March to 15 March. The time-lagged colour transition of node Loop shows its traffic importance. In contrast, activity in the southern and southwestern communities of Chicago remains relatively stable around the light blue spectrum throughout the pandemic. This can be explained by the sparse travel patterns within these community areas. We can conclude that regions with high economic activity in the pre-COVID-19

period (such as airports, city centres and business districts) experienced a sharp decline in travel-related demand at the onset of the pandemic. On the other hand, regions with low travel-related demand were not significantly affected by the spread of the disease. The above observation suggests that COVID-19 has drastically affected the propensity of individuals to travel for work and leisure.

3.5. Clustering

As explained in Section 2.1, we filter out the OD pairs whose total number of trips between 01 January 2019 and 31 October 2022 is less than 10000, and finally obtain 182 pairs, of which 145 pairs are used as the training set and 37 pairs as the test set. The test data set was clustered by combining DTW and LB_Keogh methods. In the end, a total of 5 clusters were obtained and their centroids are shown in the figure below. All data are shown with normalisation.

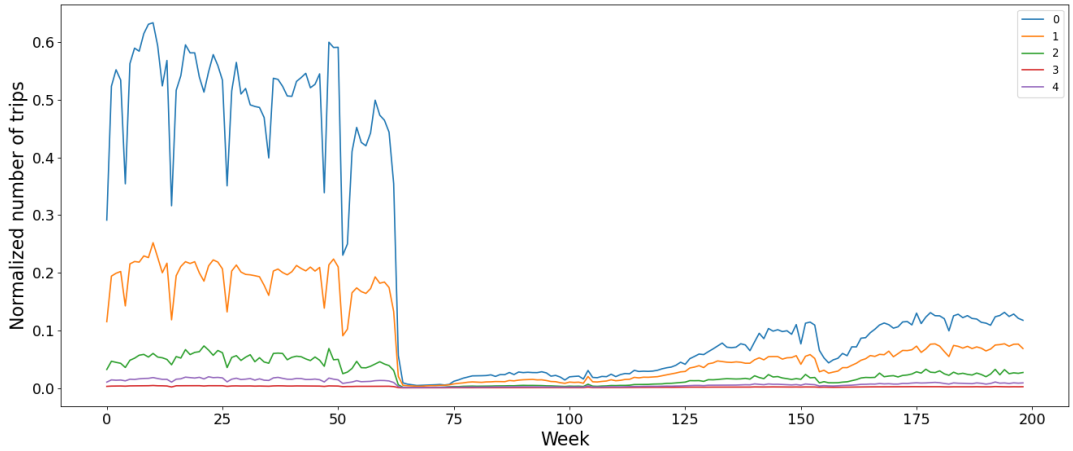
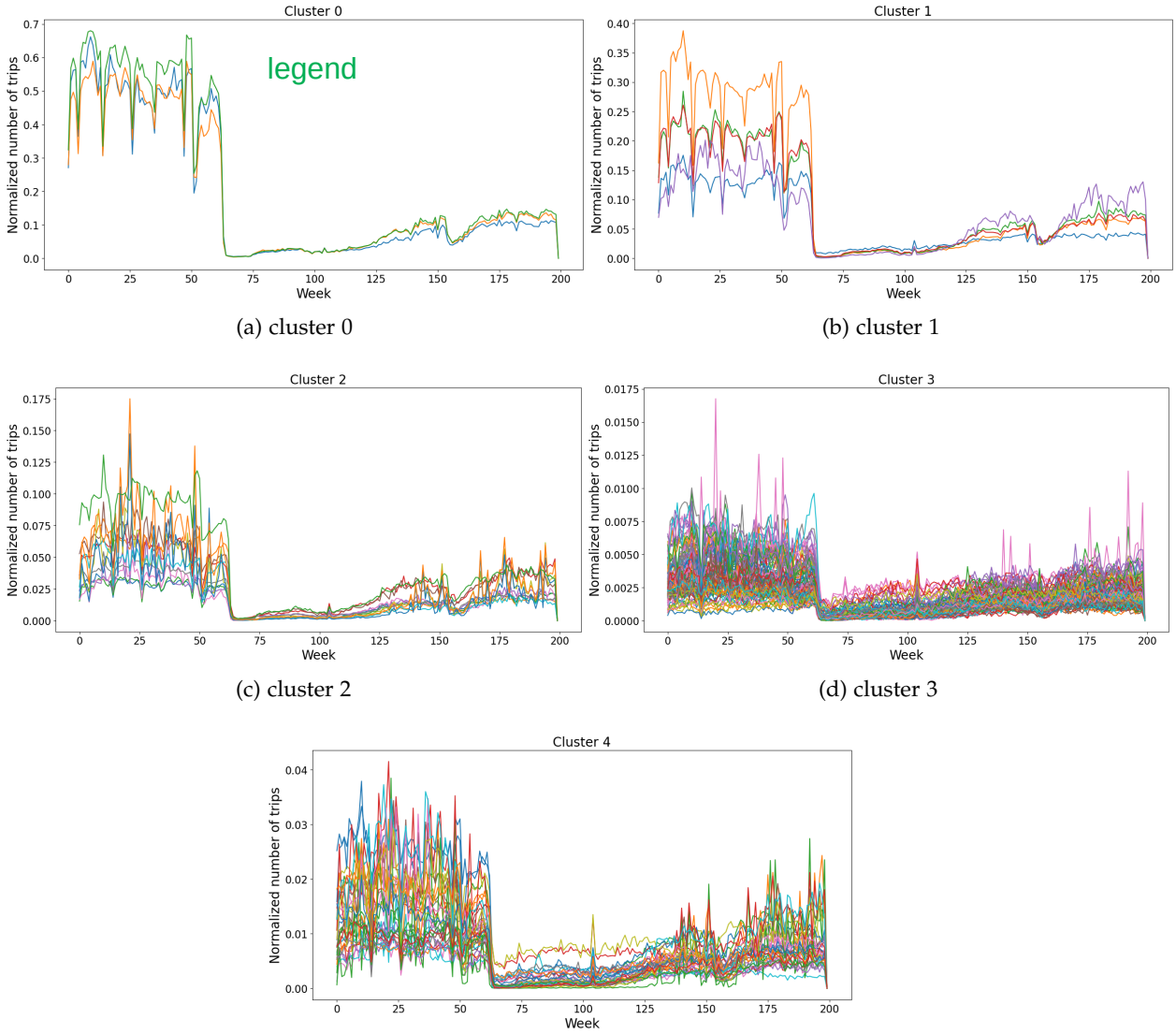


Figure 3.6.: Clustering results in representation of centroids.



(a) cluster 0

(b) cluster 1

(c) cluster 2

(d) cluster 3

(e) cluster 4

Figure 3.7.: Prediction results for 5 clusters.

Obviously, the larger the data volume of the centroid, the fewer samples it contains. Classifying the data in the test set on the basis of these centroids gives the following results:

Table 3.3.: Testing dataset classification

cluster	Index of testing dataset
0	9
1	22
2	2,11,18,26,29,31,33
3	0,1,4,5,6,7,8,13,15,16,19,20,23,24,25,27,30,32,34,35
4	3,10,12,14,17,21,28,36

3.6. Regression

The impact of traditional natural disasters such as earthquakes or extreme weather on the transport system is often severe and short-lived. Typically, the system goes through a resilience cycle to return to a new steady state. However, under the dual influence of the pandemic and public policy, the subject of this research has become complex and changing. During the recovery process, the system is often impacted again by changes in the external environment, such as increased viral infectivity or new executive orders. A single resilience cycle model is not sufficient to fit the overall trend of the curve, so the entire curve must be broken down into multiple resilience cycle processes, as shown in the figure 3.8 below.

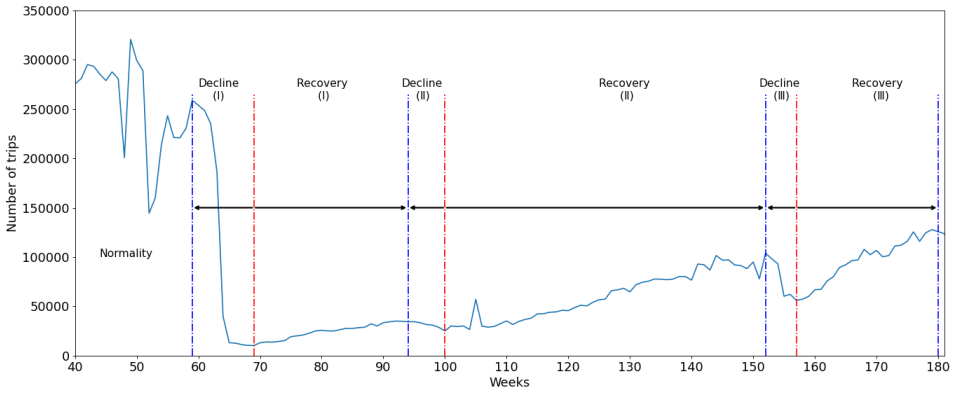


Figure 3.8.: Division of time intervals.

Section 3.1 looks in more detail at key events during the pandemic. Combining the trend of the curve and these events, the curve can be roughly divided into three resilience periods, with the blue dotted line as the dividing line. The dashed red line marks the trough of the cycle. The landmark events at the beginning of these three cycles are:

1 The new positive case in Chicago,

2 The second wave of the pandemic,

3 Vaccine requirements for indoor public places.

Each resilience cycle includes two processes of decline and recovery, so the curve during the pandemic is divided into 6 time periods, i.e. [59,69], [70,92], [93,103], [105,150], [151,156], [157,180]. For each small time interval, the logistic function is used to fit the parameters and they are finally connected to obtain a finished model. The centroid curve of cluster 0 is divided into 6 segments according to the above time intervals. The fitting results based on the logistic function are shown in the following figure 3.9.

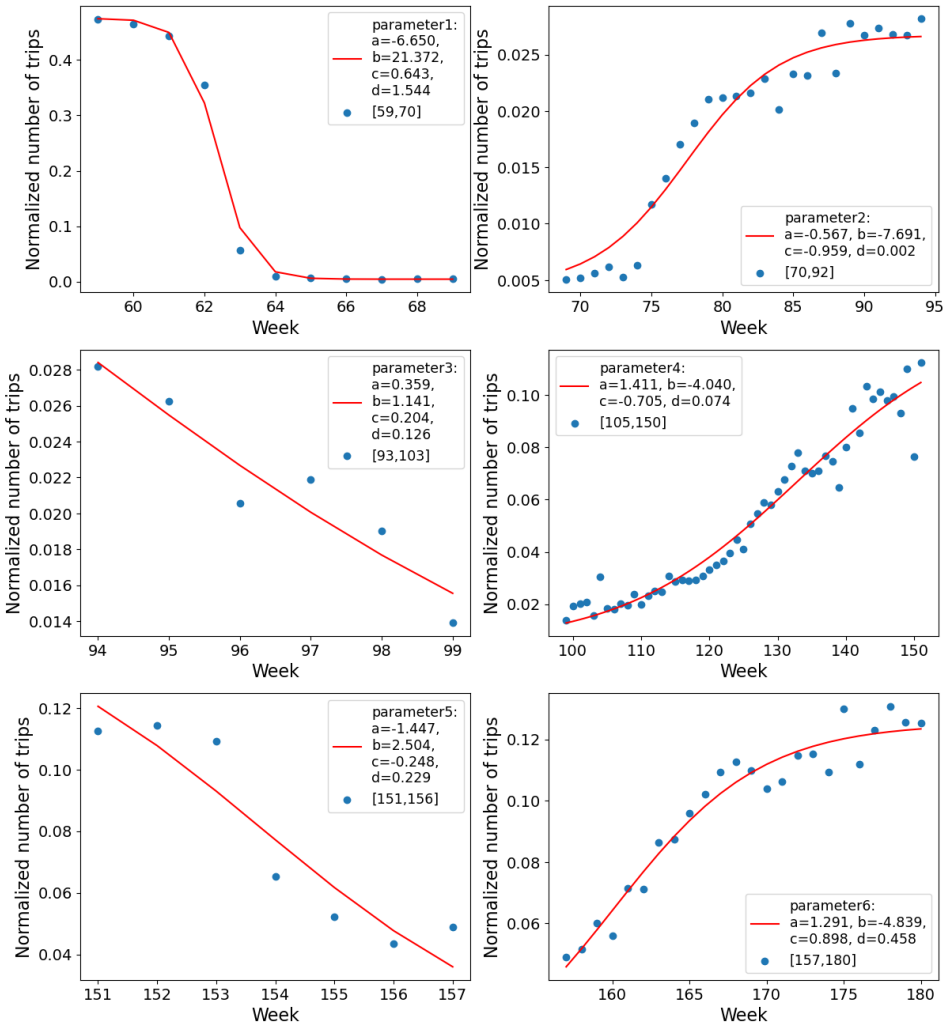


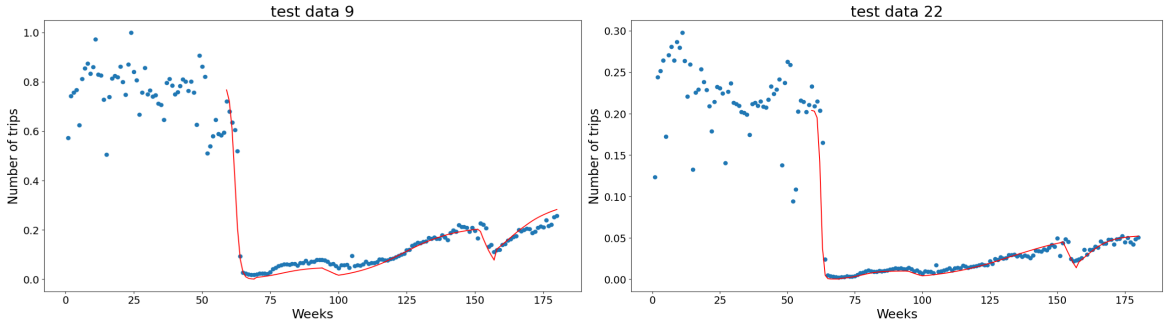
Figure 3.9.: Logistic regression for Cluster 0.

Similar results can be obtained by segmentally fitting the central curves of the remaining four clusters, which can be found in the Appendix A.3. These parameters can be used to build 5 benchmark models. These are then used to make predictions on the test data set in Table 3.3.

4. Results and discussion

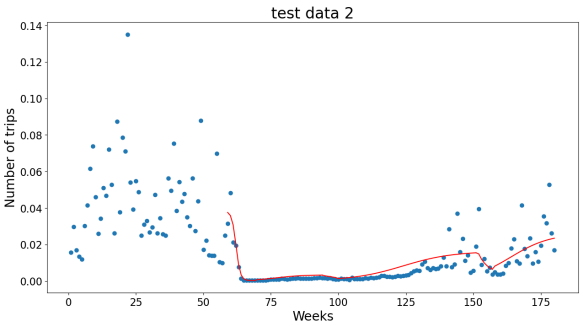
4.1. Prediction Results According to Different Clusters

The blue scatter points are the real test data and the red curve is the predicted curve based on the baseline model for the cluster to which the test data belongs. The full results can be found in the Appendix A.4, and only some of the results are shown in the figure below 4.1 and 4.2. As can be seen from the resulting figure, the prediction curve of cluster 0, 1, 2 can be a very accurate representation of the real data. However, there is a significant deviation between the prediction curve of cluster 3 and the real data.

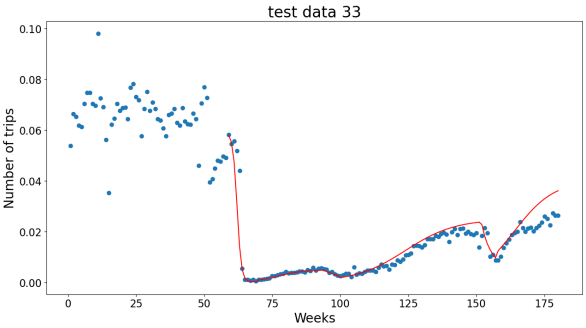


(a) cluster0

(b) cluster1

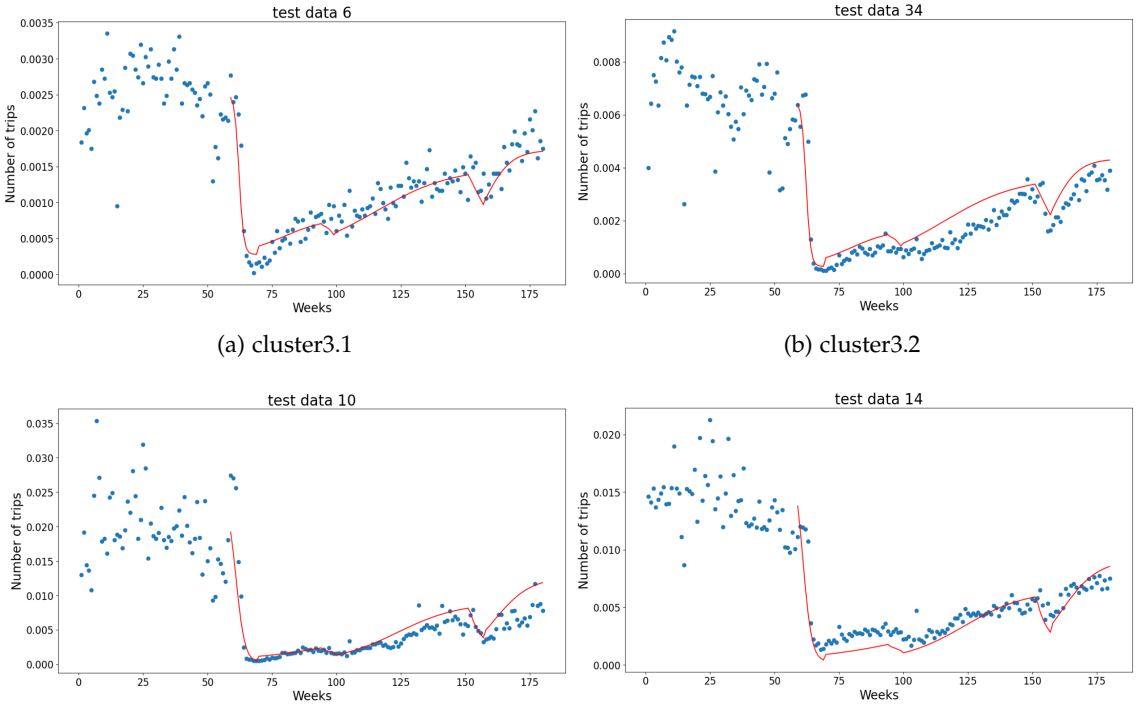


(c) cluster2.1



(d) cluster2.2

Figure 4.1.: Partial prediction results (1)



(a) cluster3.1

(b) cluster3.2

(c) cluster4.1

(d) cluster4.2

Figure 4.2.: Partial prediction results (2)

4.2. Evaluation of Prediction Results

In order to quantify the difference between each curve and the true value and to evaluate the quality of the prediction results, regression evaluation indicators are introduced: Mean Squared Error (MSE) and Coefficient of Determination (R^2). The MSE is the average of the sum of the squares of the corresponding point errors of the predicted data and the original data. Smaller values indicate smaller differences between the prediction curve and the actual data. R^2 is a measure of the goodness of fit of a model, usually ranging from 0 to 1. The closer R^2 is to 1, the better the model fits. If the fit is poor, the value will be closer to 0. However, in regression without intercept, the usual expressions of R^2 may give unexpected results, i.e. the R^2 obtained is negative or greater than 1 [17]. A negative R^2 indicates that the model's prediction is worse than the mean value. The logistic function fitting used in the study is a nonlinear function without an intercept, so it is possible for R^2 to have negative values.

In addition, two indicators are used to describe the characteristics of different datasets: data volume and recovery rate. The normal data volume is the average of the OD to the pre-pandemic data volume, and all data are normalised. This shows the volume of taxi journeys for this pair of ODs before the pandemic. The recovery rate is the improvement or reduction in the normal data volume after a disruptive event.

Table 4.1.: Evaluation of prediction Results.

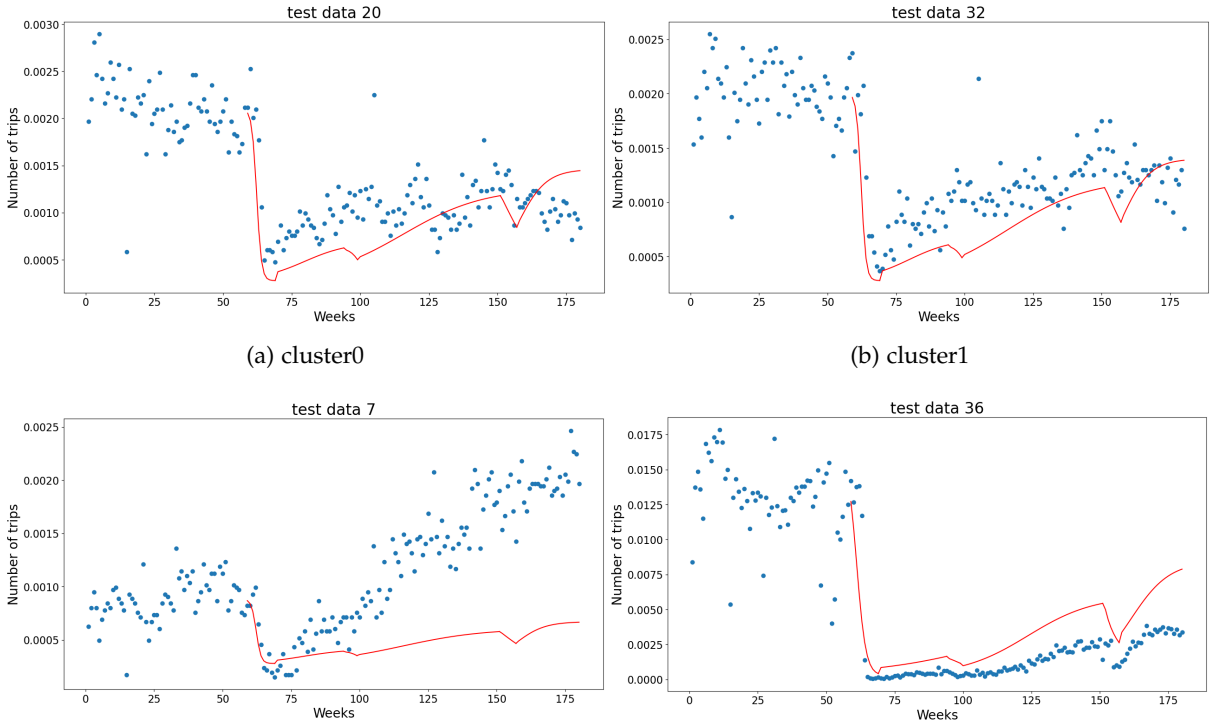
Index	Normal data volume	Recovery rate[%]	MSE$[1e^{-5}]$	R^2
Cluster0	0.241			
9	0.755	29.8	2.491	0.980
Cluster1	0.330			
22	0.218	22.9	93.016	0.924
Cluster2	0.479			
2	0.041	60.9	4.566	0.586
11	0.062	66.2	4.876	0.789
18	0.093	73.8	9.815	0.822
26	0.063	75.1	7.464	0.750
29	0.048	22.9	4.612	0.340
31	0.059	36.4	3.088	0.544
33	0.064	35.7	1.218	0.892
Cluster3	0.607			
0	0.002	44.7	0.009	0.415
1	0.006	88.3	0.050	0.814
4	0.003	28.2	0.040	-0.892
5	0.003	80.8	0.014	0.768
6	0.003	76.3	0.004	0.851
7	0.001	264.9	0.086	-1.213
8	0.002	38.4	0.019	0.145
13	0.004	40.5	0.031	0.077
15	0.005	39.0	0.069	0.237
16	0.007	77.0	0.071	0.831
19	0.003	31.9	0.022	-0.036
20	0.002	41.4	0.015	-0.451
23	0.002	41.4	0.007	0.579
24	0.006	83.0	0.079	0.674
25	0.003	47.8	0.019	0.614
27	0.005	72.6	0.048	0.634
30	0.002	52.8	0.006	0.698
32	0.002	51.8	0.014	-0.402
34	0.007	51.8	0.040	0.770
35	0.006	75.2	0.035	0.825
Cluster4	0.532			
3	0.016	53.0	0.299	0.741
10	0.019	49.3	0.421	0.721
12	0.014	71.6	0.247	0.741

Continued on next page

Table 4.1 – continued from previous page

Index	Normal data volume	Recovery rate[%]	MSE$[1e^{-5}]$	R^2
14	0.014	45.6	0.094	0.792
17	0.026	62.4	0.682	0.838
21	0.016	45.9	0.244	0.715
28	0.008	60.7	0.147	0.252
36	0.013	26.5	0.572	-0.024

Considering that there is a correlation between MSE and data volume, and that the data volume of different OD pairs is quite different, the results of MSE can only be used as a reference, and the main focus is on R^2. From the results of R^2 in the table, it can be seen that not all prediction curves agree well with the real data, and there are even cases where R^2 is negative, e.g. 4, 7, 19, 20, 32, 36. For these test sets the prediction curve shows large errors or even distortions.



(a) cluster0

(b) cluster1

(c) cluster2.1

(d) cluster2.2

Figure 4.3.: Prediction errors

The causes of the above prediction errors can be broadly grouped into the following categories:

1. The time interval is divided on the basis of specific events, which can be understood as shape-based parameter fitting. If the real data does not show an obvious decline when faced

with new external shocks at the corresponding time, i.e. the shape of the real data curve is very different from the benchmark model, this will directly lead to the invalidation of the prediction results.

2. The size of the data sample is also an important factor in determining whether the prediction is accurate. It is intuitive that most of the inaccurate predictions are concentrated in cluster3. This is because samples with a small amount of data have relatively large fluctuations, such as datasets 19, 20, 32, which can be understood as a resilience model with relatively large noise, while the model we obtained through data fitting is relatively smooth. The result is that the real data is scattered above and below the predicted curve. In addition, the small amount of data makes the recovery rate anomalous, and the recovery rate tends to spike, as in data 7 4.3c.

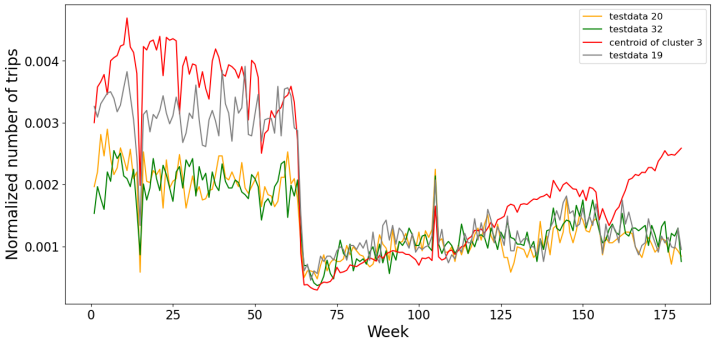


Figure 4.4.: Comparison of test data 19, 20, 32 with centroid of Cluster 3.

3. Since the benchmark model is based on the cluster centroid curve, if the recovery rate of the real data is very different from the cluster centroid curve, this error will increase over time. For example, the recovery rate of dataset 7 is as high as 264.9%, and the recovery rates of datasets 4 and 36 are only over 20%, which is far from the average recovery rate of their clusters, resulting in the overall prediction curve being higher or lower than the real data.

In order to reduce the above errors and to improve the prediction accuracy, the following improvements are proposed:

1 10,000 trips in three and a half years may be too few, and this threshold can be increased to remove more OD pairs with a small sample size.

2 After primary clustering, secondary clustering is performed on the cluster with the smallest amount of data, i.e. cluster 3, screening out data sets with similar recovery rates and obvious resilience characteristics.

3 Use the idea of stepwise regression forecasting, i.e. introduce fixed period data at each step, e.g. every four weeks, calculate the error between the forecast curve and the real data, and use it to compensate the forecast curve to reduce the recovery rate caused by the continuous change in time error.

5. Conclusion and future works

5.1. Conclusion

This study attempts to build an OD network to observe changes in taxi travel patterns in the city of Chicago, and to use travel data between some community nodes to predict taxi trips in other OD pairs using a resilience model. This exploratory analysis shows how the COVID-19 pandemic has profoundly affected taxi travel patterns in Chicago.

Here is a brief summary of the COVID-19 developments in Chicago City. Mobility decreases rapidly in the early days of the pandemic in March 2020. By the end of April 2020, the total number of taxi trips reaches its lowest point and remains at roughly the same level. From June 2020, we also see a recovery in transport activity in Chicago. While mobility has increased, it does not reach pre-pandemic levels. This suggests that people may have reduced non-essential leisure and entertainment travel and adapted to working from home. Only those whose work structures do not allow for the establishment of home offices could contribute to the current level of mobility. In October 2020, when the second wave of the pandemic hit, the number of weekly trips fell again. However, the duration of this decline was very short, the negative impact on mobility was limited and transport activity soon began to recover. Compared to the pandemic, the vaccination requirements for indoor public places in December 2021 had a greater impact on taxi trips and people's mobility, with the number of taxi trips falling by 46% in a short period of time.

Combining the overall trend in travel numbers and the policy news provided by the CDPH COVID-19 Press Room, the curve since the pandemic can be roughly divided into three segments, and each segment is approximated as a resilience model. Using these three time nodes above and the training set data, three continuous resilience models were built for each cluster. Based on these benchmark models and combined with the average volume of data before the epidemic, the change in the number of taxi trips after the pandemic was predicted. From the results in Chapter 4, most of the prediction curves are more in line with the trend of the real data, especially for OD pairs with a large amount of data, the prediction results will be more accurate. Therefore, we believe that for the same city, changes in the number of taxi trips between different regions will be similar. Changes in other OD pairs can be predicted using data from an OD pair belonging to the same cluster. However, the drawback is that this method will fail for OD pairs with a small sample size. The reason for this is that those OD pairs with a small number of trips have large data fluctuations, which weakens the resilience properties and increases the randomness of the recovery rate. The recovery rate can be in the range of 20% - 250%, which makes prediction difficult. Therefore, the prediction method

discussed in this study is more suitable for predicting OD pair with a large number of data samples. On the other hand, in addition to the epidemic itself, public health policies have had a major impact on the taxi system. Policy makers should take this into account and compensate the public transport system accordingly.

A limitation of this study is that the dataset only contains taxi data for two years after the outbreak. Clearly, as of October 2022, the transportation situation has not yet fully recovered to pre-outbreak levels. Notably, President Joe Biden announced to the US Congress on 11 May that he would end the country's public health emergency under COVID-19 [18]. This landmark event could have a positive impact, increasing people's willingness to travel and encouraging further recovery of public transport. There is reason to believe that the introduction of more data will help to build a more complete model of resilience and improve the accuracy of forecasts. In addition, another limitation of this method is the poor prediction results for these weakly resilient OD pairs. Simply using the size of the data set is not sufficient to identify those OD pairs with weak resilience properties and it may be necessary to introduce quadratic clustering of the data set with small sample sizes to identify them.

In summary, the results and conclusions of this paper have practical implications for the network management of taxi systems during major public health events. The characteristics of transport resilience can help policy makers to clearly understand the process of resilience between different community nodes. The proposed framework and prediction results would help transport management to understand the trends of taxi systems before and during disasters, and build efficient emergency response procedures.

5.2. Future Research Works

In future research, we can try to generalise this prediction method from different regions of a city to different cities in a country. A network could be created with the city's airport or central railway station as the hub, and other inter-regional transport systems such as planes and trains could be predicted and analysed. However, given that public policy will have a major impact on the transport structure and people's willingness to travel, it is worth investigating whether it can be accurately predicted when the policy of the whole system is inconsistent. Obviously, central, state and local governments will make decisions in response to disasters. Sometimes these decisions are consistent, and sometimes they are inconsistent or even contradictory. This can make forecasting very difficult. On the other hand, due to the different political systems in different countries, the decision-making between central and local governments will also be different. For example, in countries with a federal system, the state government retains some autonomous power to manage internal affairs and its decision-making has a certain degree of independence. If the resilience model can predict well the recovery of the future transport system after a shock, it will help policymakers to anticipate losses after a disaster and formulate more rational public policies that help the system recover faster.

A. Appendix

A.1. Community Areas in Chicago

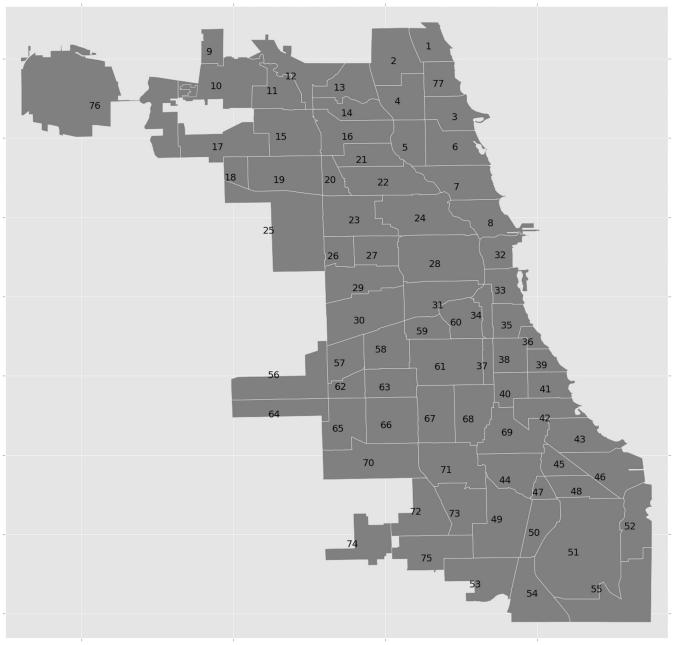


Figure A.1.: 77 Community Areas in the City of Chicago [13].

Table A.1.: Name of the Community Areas of Chicago.

Num	Community Area
1	Rogers Park
2	West Ridge
3	Uptown
4	Lincoln Square
5	North Center
6	Lake View
7	Lincoln Park
8	Near North Side

Continued on next page

Table A.1 – continued from previous page

Num	Community Area
9	Edison Park
10	Norwood Park
11	Jefferson Park
12	Forest Glen
13	North Park
14	Albany Park
15	Portage Park
16	Irving Park
17	Dunning
18	Montclare
19	Belmont Cragin
20	Hermosa
21	Avondale
22	Logan Square
23	Humboldt Park
24	West Town
25	Austin
26	West Garfield Park
27	East Garfield Park
28	Near West Side
29	North Lawndale
30	South Lawndale
31	Lower West Side
32	Loop
33	Near South Side
34	Armour Square
35	Douglas
36	Oakland
37	Fuller Park
38	Grand Boulevard
39	Kenwood
40	Washington Park
41	Hyde Park
42	Woodlawn
43	South Shore
44	Chatham
45	Avalon Park
46	South Chicago
47	Burnside

Continued on next page

Table A.1 – continued from previous page

Num	Community Area
48	Calumet Heights
49	Roseland
50	Pullman
51	South Deering
52	East Side
53	West Pullman
54	Riverdale
55	Hegewisch
56	Garfield Ridge
57	Archer Heights
58	Brighton Park
59	McKinley Park
60	Bridgeport
61	New City
62	West Elsdon
63	Gage Park
64	Clearing
65	West Lawn
66	Chicago Lawn
67	West Englewood
68	Englewood
69	Greater Grand Crossing
70	Ashburn
71	Auburn Gresham
72	Beverly
73	Washington Heights
74	Mount Greenwood
75	Morgan Park
76	O'Hare
77	Edgewater

A.2. Public Health Policy

Public Policies and news during the COVID-19 pandemic in Chicago:

- January 24, 2020: City of Chicago Announces First Local Patient with Travel-Related Case of 2019 Novel Coronavirus
- March 6, 2020: Public Health and Chicago Public Schools Officials Announce New Presumptive Positive Case of Coronavirus Disease 2019

- March 13, 2020: City of Chicago Prepares for Closure of all K-12 Schools, as Mandated by the State of Illinois
- March 17, 2020: Public Health Officials Announce First Illinois Coronavirus Disease Death
- March 19, 2020: City of Chicago Orders Sick Residents to Remain Home to Prevent Further Spread of COVID-19
- March 20, 2020: Mayor Lightfoot Joins Governor Pritzker To Announce State Order to Stay at Home to Prevent Further Spread of COVID-19
- March 26, 2020: Mayor Lightfoot Orders the Immediate Closure of The City's Lakefront, Adjacent Parks, 606 and Riverwalk to the Public
- March 28, 2020: Public Health Officials Announce the First Death of an Infant with Coronavirus Disease
- May 28, 2020: Mayor Lightfoot and CDPH Announce Chicago Ready to Begin Reopening Cautiously on Wednesday, June 3, 2020
- June 26, 2020: Restaurants, bars and breweries serve patrons indoors with limited capacity and safety restrictions
- October 19, 2020: Mayor Lightfoot and CDPH Commissioner Dr. Arwady Sound the Alarm on Second Wave of COVID-19
- February 10, 2021: Mayor Lightfoot Announces Roadmap for Further Easing of COVID-19 Regulations
- April 29, 2021: Mayor Lightfoot Announces the Launch of "Open Chicago"
- December 7, 2021: Illinois and Chicago Departments of Public Health Confirm State's First Case of the Omicron COVID-19 Variant
- December 21, 2021: City of Chicago Announces Vaccine Requirements for Restaurants, Bars, Gyms, and Other Indoor Public Places

A.3. Parameter Results for Linear Regression

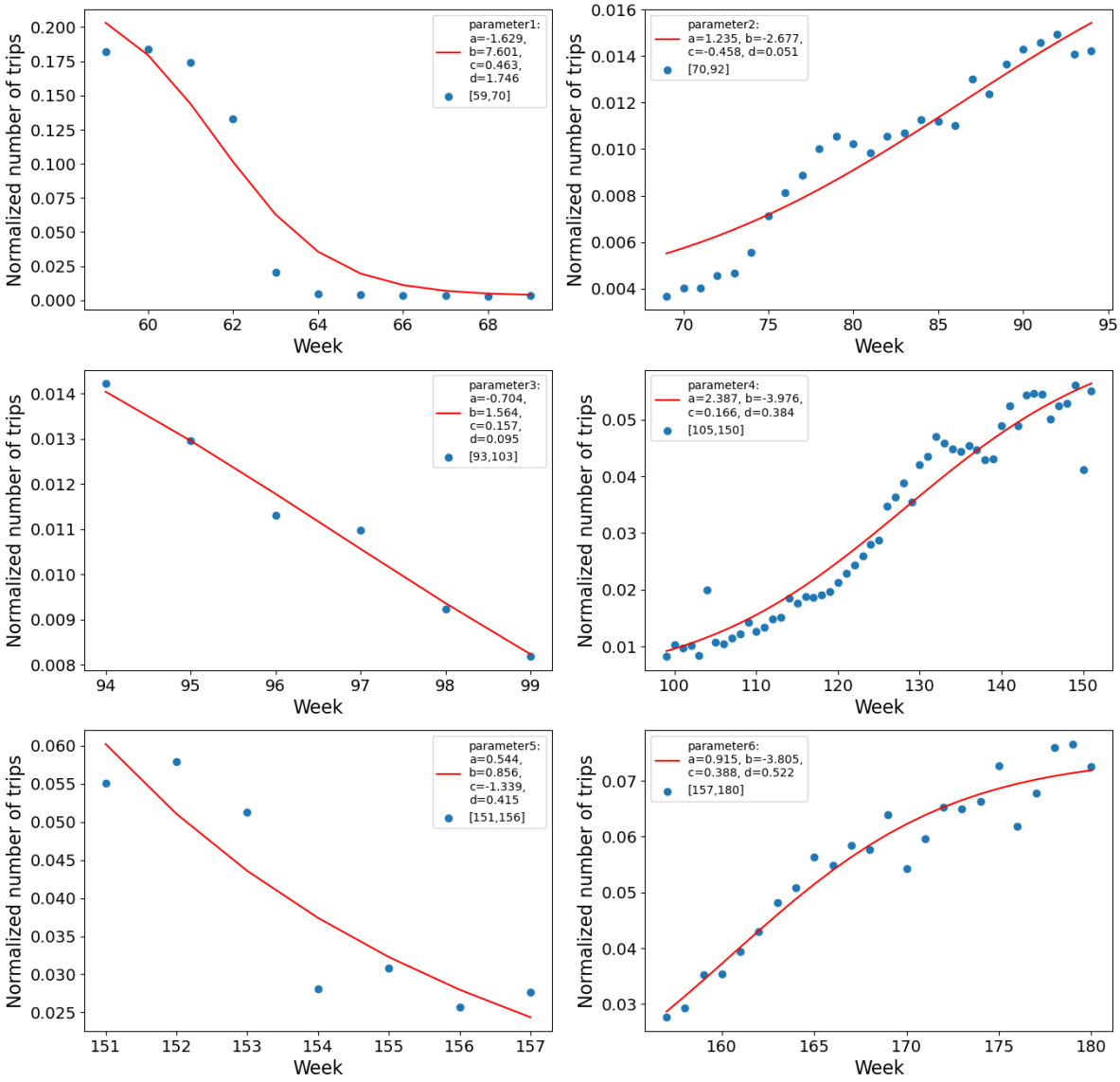


Figure A.2.: Parameter results for cluster 1

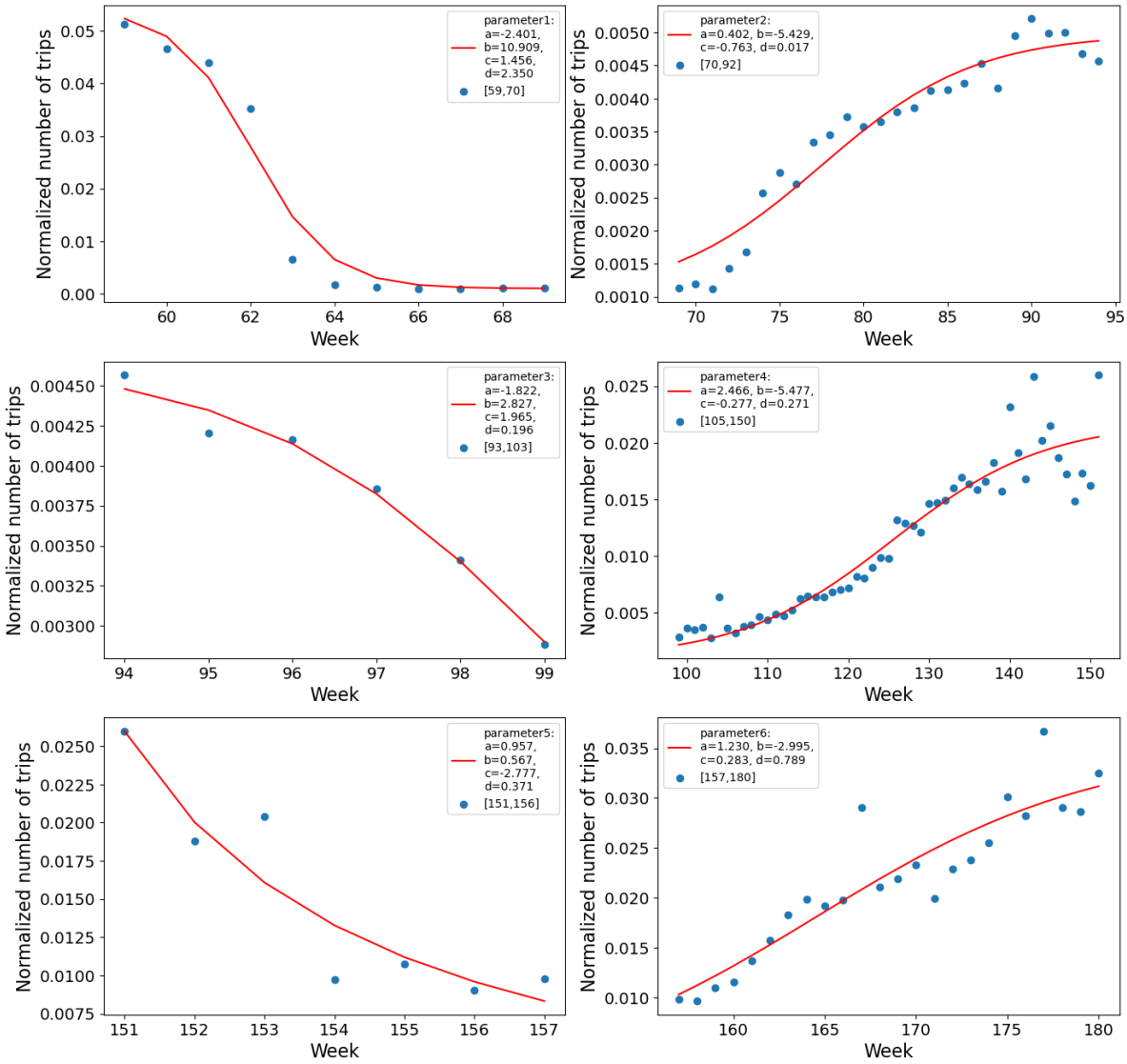


Figure A.3.: Parameter results for cluster 2

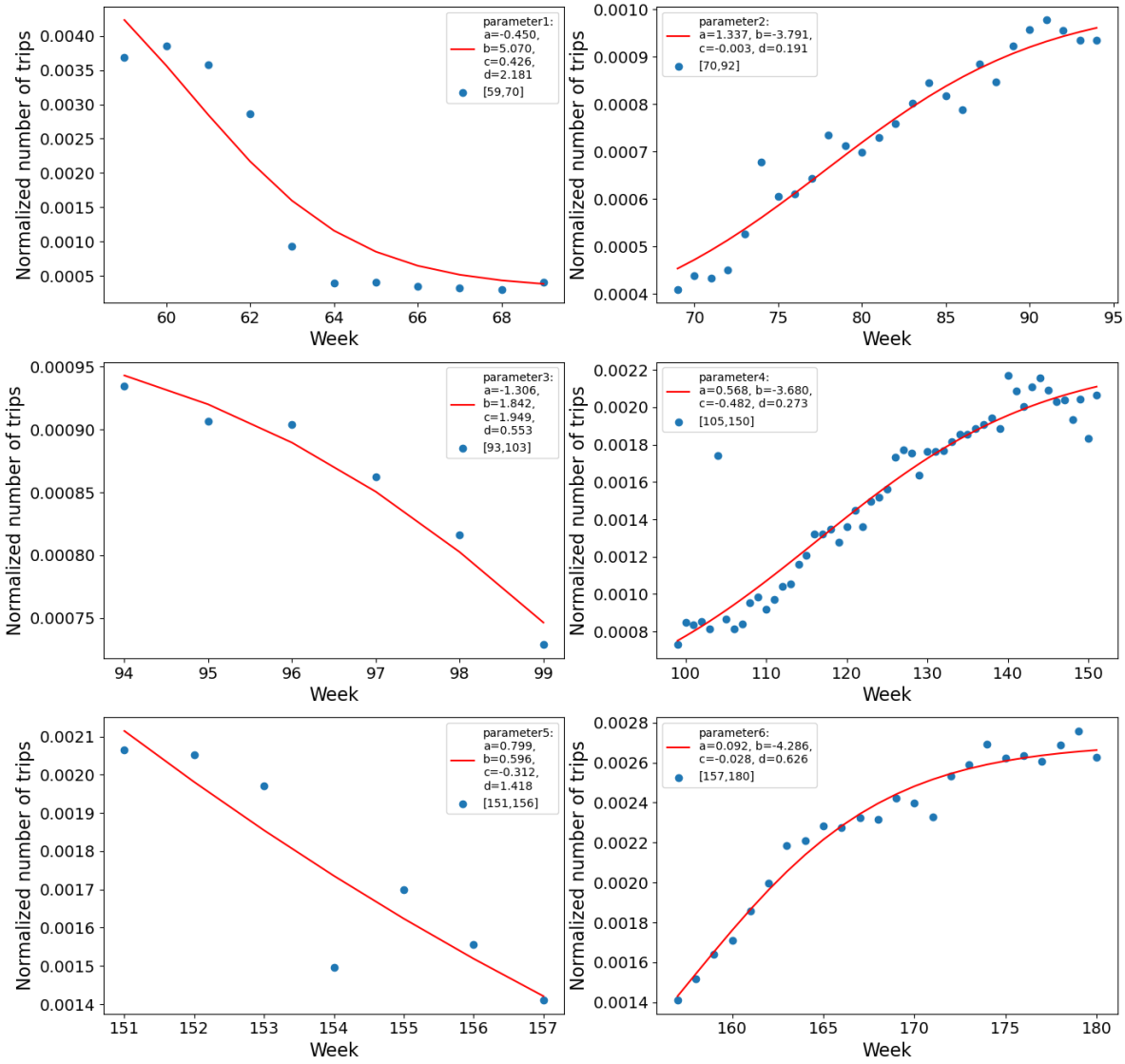


Figure A.4.: Parameter results for cluster 3

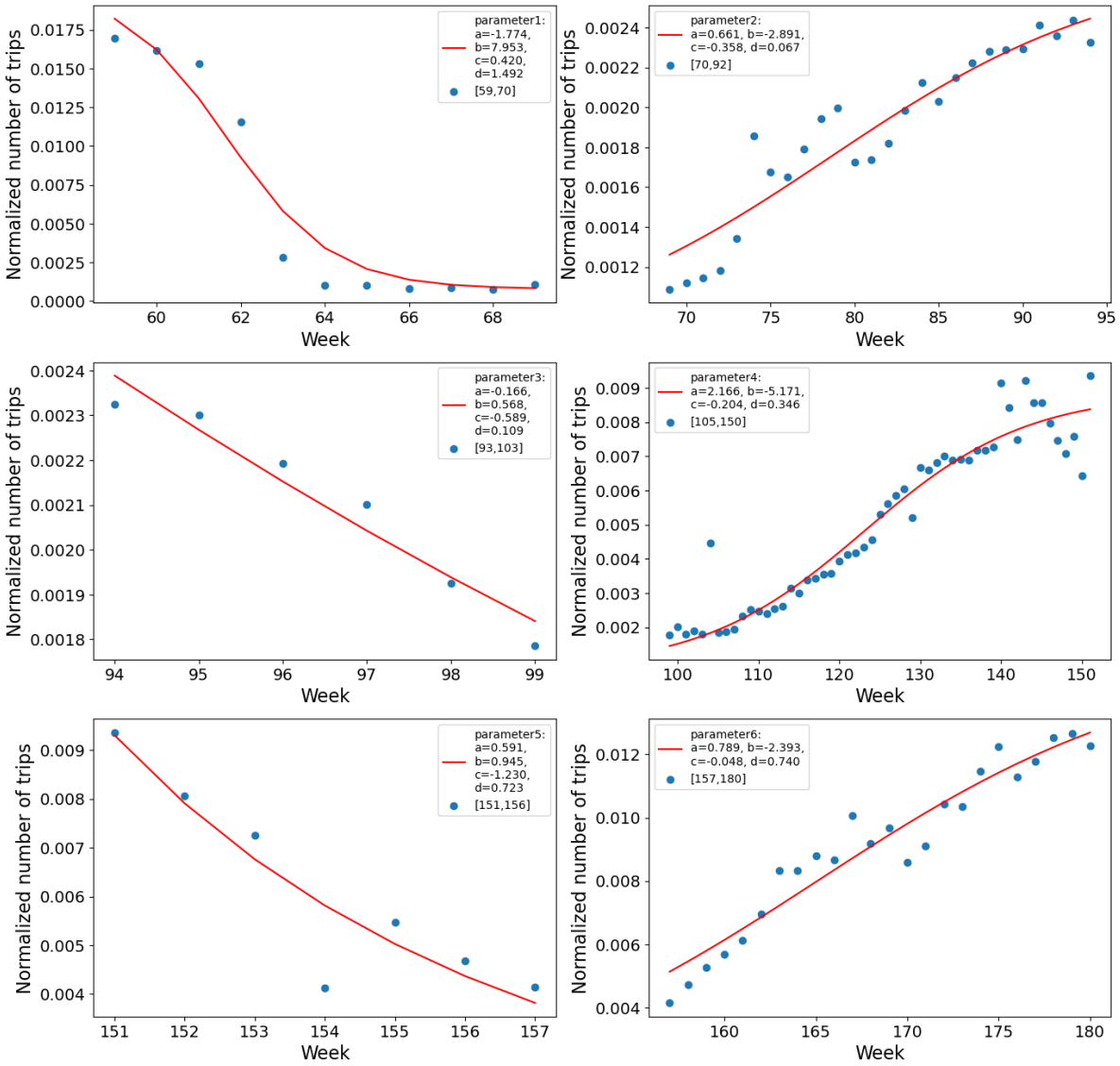


Figure A.5.: Parameter results for cluster 4

A.4. Complete Prediction Results for Cluster 2

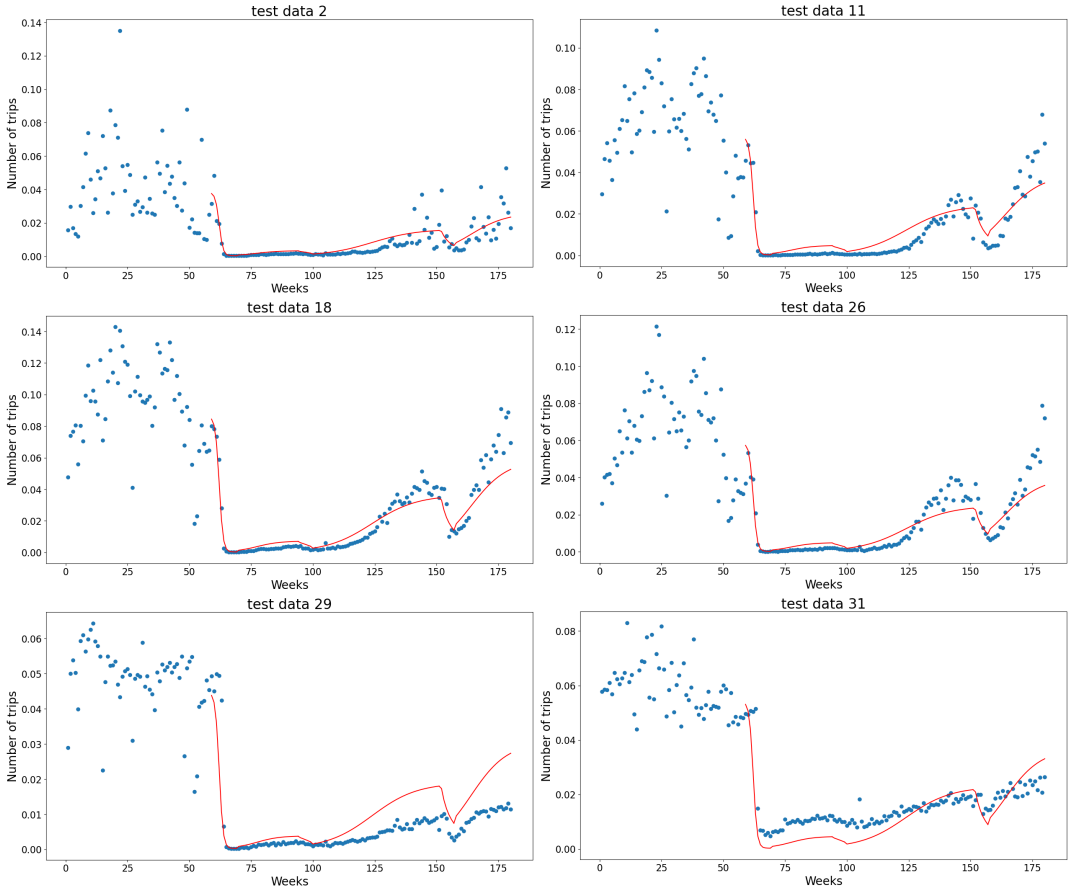


Figure A.6.: Prediction results for cluster 2 (1-6)

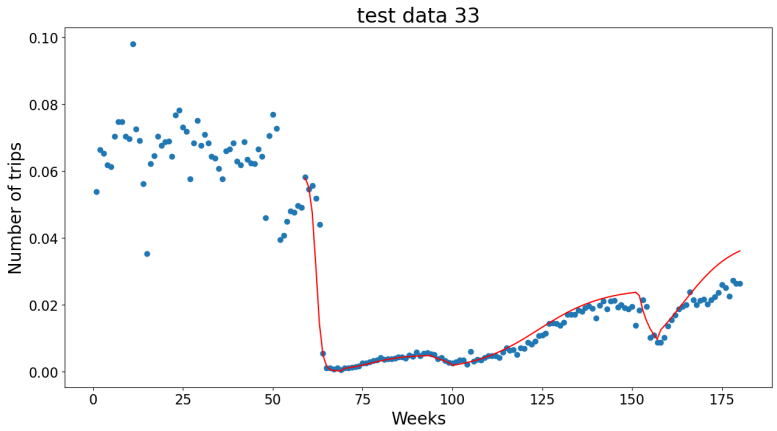


Figure A.7.: Prediction results for cluster 2 (7)

A.5. Complete Prediction Results for Cluster 3

(a) cluster3 01

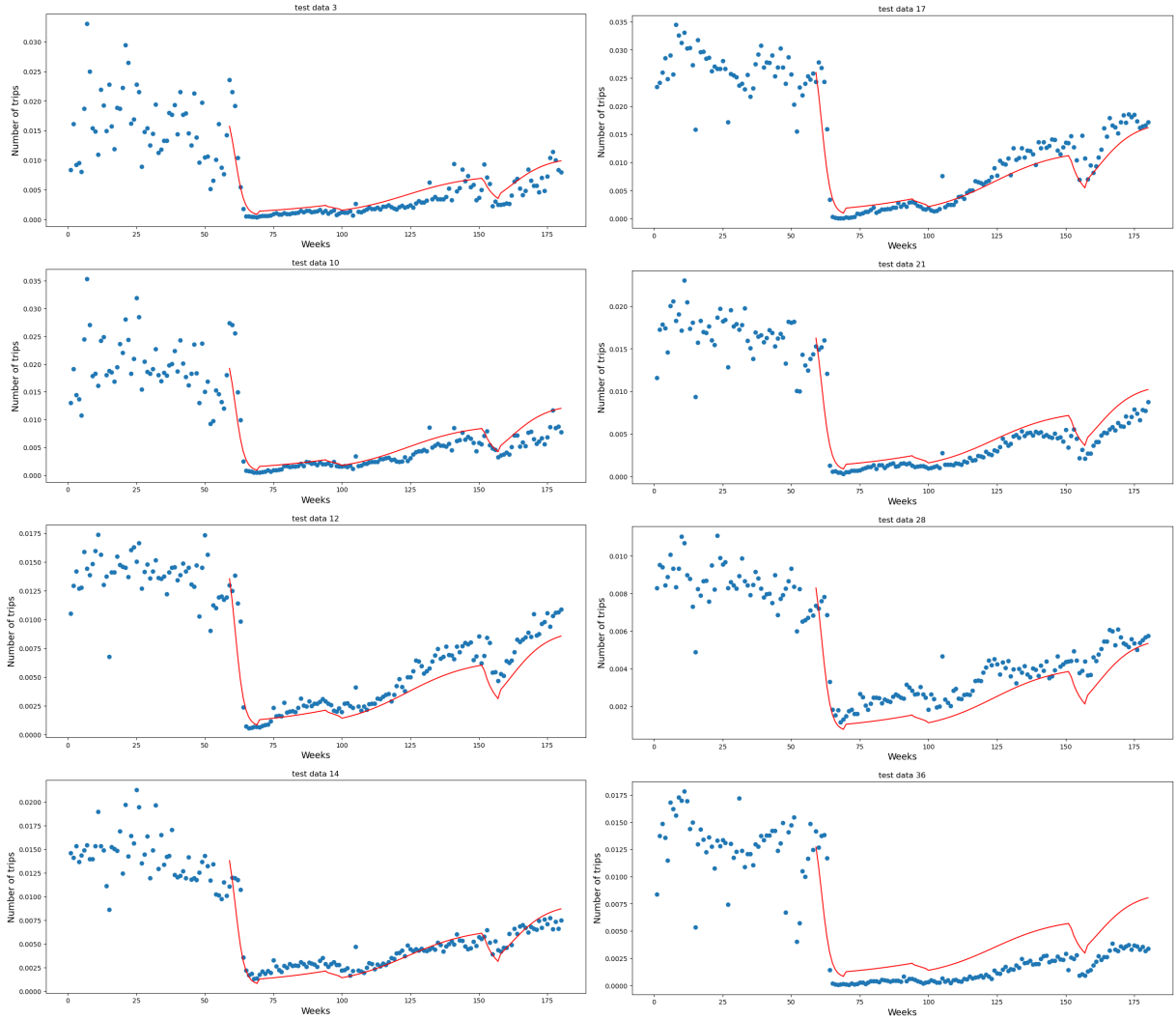
(b) cluster3 02

Figure A.8.: Prediction results for cluster 3.1.

(a) cluster3 03 (b) cluster3 04

Figure A.9.: Prediction results for cluster 3.2.

A.6. Complete Prediction Results for Cluster 4



(a) cluster4 01 (b) cluster4 02

Figure A.10.: Prediction results for cluster 4.

List of Figures

List of Tables

References

[1] R. A. Posner. *Catastrophe: risk and response*. Oxford University Press, 2004.

[2] V. Proag. "The concept of vulnerability and resilience". In: *Procedia Economics and Finance* 18 (2014), pp. 369–376.

[3] M. Bruneau, S. E. Chang, R. T. Eguchi, G. C. Lee, T. D. O'Rourke, A. M. Reinhorn, M. Shinozuka, K. Tierney, W. A. Wallace, and D. Von Winterfeldt. "A framework to quantitatively assess and enhance the seismic resilience of communities". In: *Earthquake spectra* 19.4 (2003), pp. 733–752.

[4] J. Li, K. Ozbay, B. Bartin, S. Iyer, and J. A. Carnegie. "Empirical evacuation response curve during hurricane irene in cape may county, new jersey". In: *Transportation research record* 2376.1 (2013), pp. 1–10.

[5] J. M. Bolland. "Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks". In: *Social networks* 10.3 (1988), pp. 233–253.

[6] H. Sakoe and S. Chiba. "Dynamic programming algorithm optimization for spoken word recognition". In: *IEEE transactions on acoustics, speech, and signal processing* 26.1 (1978), pp. 43–49.

[7] C. Myers, L. Rabiner, and A. Rosenberg. "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.6 (1980), pp. 623–635.

[8] A. Kuzmanic and V. Zanchi. "Hand shape classification using DTW and LCSS as similarity measures for vision-based gesture recognition system". In: *EUROCON 2007-The International Conference on" Computer as a Tool"*. IEEE. 2007, pp. 264–269.

[9] A. Corradini. "Dynamic time warping for off-line recognition of a small gesture vocabulary". In: *Proceedings IEEE ICCV workshop on recognition, analysis, and tracking of faces and gestures in real-time systems*. IEEE. 2001, pp. 82–89.

[10] E. Keogh and C. A. Ratanamahatana. "Exact indexing of dynamic time warping". In: *Knowledge and information systems* 7.3 (2005), pp. 358–386.

[11] Y. Zhu, K. Ozbay, K. Xie, and H. Yang. "Using big data to study resilience of taxi and subway trips for hurricanes Sandy and Irene". In: *Transportation research record* 2599.1 (2016), pp. 70–80.

[12] J. S. Cramer. "The origins of logistic regression". In: (2002).

[13] *Current community area boundaries in Chicago. Chicago Data Portal:* `https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6`. (cited 15 Jan 2023).

[14] *City of Chicago Taxi Trips. Chicago Data Portal:* https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew. (cited 15 Nov 2022).

[15] *City of Chicago announces first local patient with travel relate. An official website of the City of Chicago:* https://www.chicago.gov/city/en/depts/cdph/provdrs/health_protection_and_response/news/2020/january/city-of-chicago-announces-first-local-patient-with-travel-relate.html. (cited 15 Jan 2023).

[16] *Public Health Policy of the City of Chicago. CDPH COVID-19 Press Room:* https://www.chicago.gov/city/en/sites/covid-19/home/press-releases.html. (cited 15 Jan 2023).

[17] A. P. Barten. "The coefficient of determination for regression without a constant term". In: *The Practice of Econometrics: Studies on Demand, Forecasting, Money and Income* (1987), pp. 181–189.

[18] W. N. LaFraniere S. *US plans to end public health emergency for covid in May. New York Times 2023 Jan 30.* https://www.nytimes.com/2023/01/30/us/politics/biden-covid-public-health-emergency.html. (cited 31 Jan 2023).