THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Time-dependent Edge Cost Estimation

Chapter 3 describes the general procedures for constructing a landmark graph. To find a time-dependent shortest route on a landmark graph, the time-dependent edge cost must be calculated. In this project, the time-dependent edge cost specifically refers to travel time, but in theory, it can refer to any quantity that can be described as a time-dependent edge cost function of the form $w : E, t \rightarrow \mathbb{R}$. In practice, it sometimes also refers to fuel consumptions or taxi fares. This chapter introduces a machine learning-based approach to estimate the travel time of each *significant edge* in a landmark graph at a particular moment in time.

Definition 9 (*Significant Edge*)**.** A significant edge in a landmark graph $G = (V, E)$ is an edge $e \in E$ that appears at least m times, where m is a parameter specified in advance.

The purpose of defining significant edges is to eliminate those edges that are seldom traversed by taxi drivers, as estimating the travel time of those edges will not

be very accurate. The parameter m also represents a level of *confidence*[1], that is, to what extent it is true that this edge *really* exists in the real world.

Everyday experiences show that the travel time of a particular road usually has different time-varying patterns in weekdays as compared to that in weekends or public holidays. For instance, it is likely that, on weekdays, the travel time of a particular road has one *peak* at 8 a.m. when people travel to work and the other peak at 6 p.m. when people return home after work. But when it is weekends or public holidays, the travel time of that road may have a peak at 10 a.m. when people go for holiday activities with families and the other peak at only about 8 p.m. when the whole day's celebrations are over.

Based on this intuition, two separate landmark graphs were built in this project, with one for weekdays and the other for weekends or public holidays. Moreover, as mentioned in Section 2.3.3, two data sets, bjtaxigps_30m and bjtaxigps_50m, remained after outlier removal based on different thresholds set for removing outliers. Therefore, in total, *four* landmark graphs were built in this project and they are summarised in Table 4.1, although their names are self-explanatory.

Landmark Graph	**Data Source**
wrkd_ldmkgraph_30m	weekday trajectories in bjtaxigps_30m
holi_ldmkgraph_30m	holiday trajectories in bjtaxigps_30m
wrkd_ldmkgraph_50m	weekday trajectories in bjtaxigps_50m
holi_ldmkgraph_50m	holiday trajectories in bjtaxigps_50m

Table 4.1: An summary of landmark graphs

[1]Sometimes it is also referred to as *support*

4.1 Travel Time Distribution

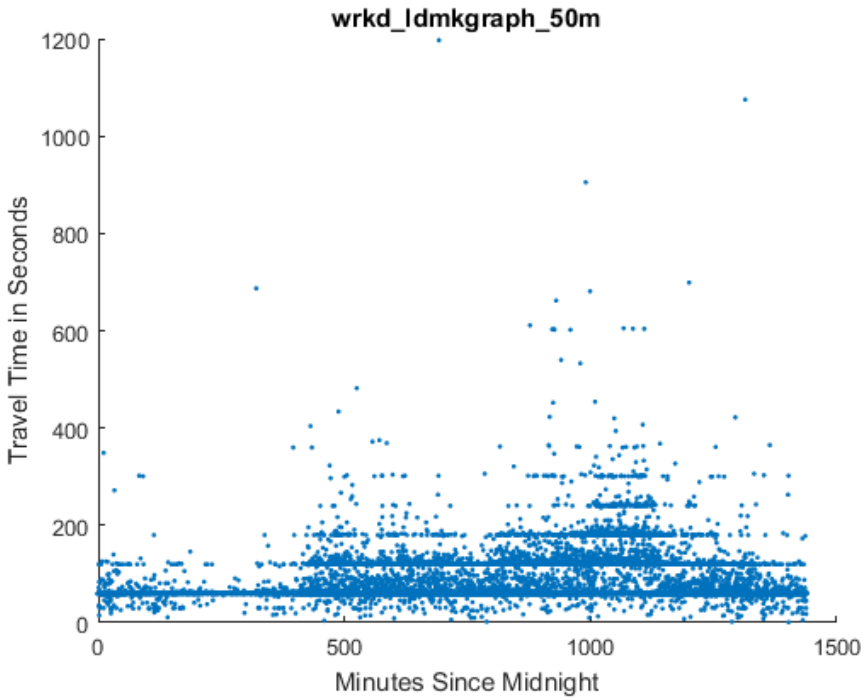

Figure 4-1: An example of travel time patterns

Figure 4-1 illustrates a scatter plot of the travel time of a particular landmark graph edge during the course of a weekday. It can be observed that the travel time does not seem to be a single-valued function with respect to time of the day, as one may expect; rather, the scatter points tend to gather around some values and form some *clusters*. For instance, when it is 500 minutes since midnight, namely 8:20 a.m., the travel time seems to have three main clusters which are represented by three horizontal lines formed by the scatter points. When it is 1,000 minutes since midnight, namely 4:40 p.m., there are about five such lines. This pattern is attributable to three possible reasons:

1. Drivers may actually choose different routes to travel between the two land-

marks, which cannot be captured by the landmark graph since it only knows a driver has traversed between the two landmarks but not the exact route. Different routes have different traffic conditions and speed limitations, therefore the travel time varies;

2. Drivers have different driving skills, preferences and behaviours. Some drivers just drive faster than others, even if the road conditions are similar and;

3. The GPS devices on taxis reported locations *periodically*, therefore, durations like 60 secondsr or120 seconds are very commonly seen in the landmark graphs. Even if the *actual* travel time is 53 seconds, it is still recorded as 60 seconds. This corresponds to the low-sampling-rate problem mentioned in Section 1.2.

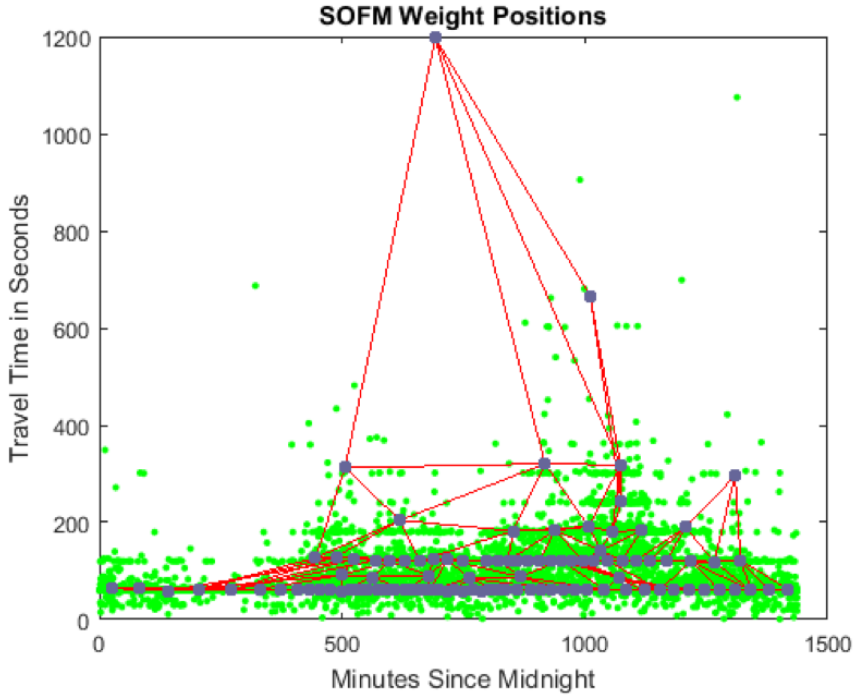

Figure 4-2: Final positions of the neurons

Therefore, it is not possible to fit the scatter points with a single-valued function. Rather, the clustering technique should first be employed to identify the travel time clusters. Like in Section 2.3.2, a **self-organising feature map** is used to cluster the scatter points. Figure 4-2 shows the final positions of the neurons at the end of the training.

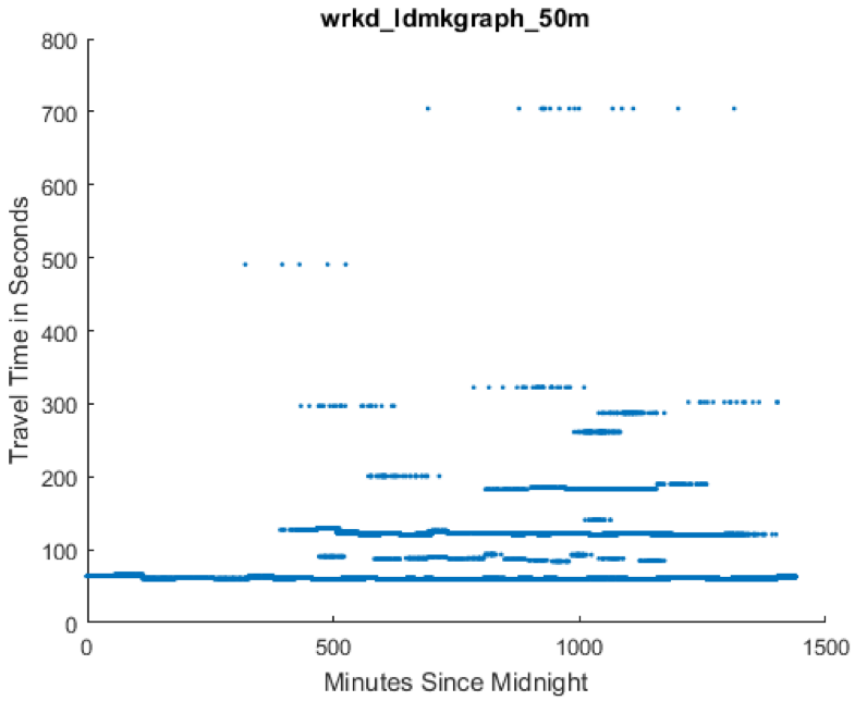

Figure 4-3: An example of representing data points with centroids

One merit of SOFM clustering is *feature extraction*. In this case, after the training is completed, the SOFM has learned some features about the travel time at different time of the day and expressed its understanding by moving its neurons to the centroids of the clusters. Now, **the data points can be represented by their respective centroids**. Figure 4-3 shows the effect of replacing each data point's value with their centroids'. The clusters are clearly shown by the horizontal lines formed by those centroids.

Representing data points with cluster centroids or features learned provides the benefit of *generalisation*. The data set used in this project, albeit large in size, is nevertheless a small *sample* of the *population* of taxi trajectories over time. In other words, these trajectory records are only the *observed* ones but there are potentially infinite number of trajectories that cannot be all observed. The only information known about them is that **they must fit into one of the clusters** after undergoing the same procedures described in the previous chapters. The use of centroids instead of real data points also takes into consideration the unobserved trajectories and reflects the real underlying patterns. In fact, generalisation is an advantage that all neural network-based methods share.

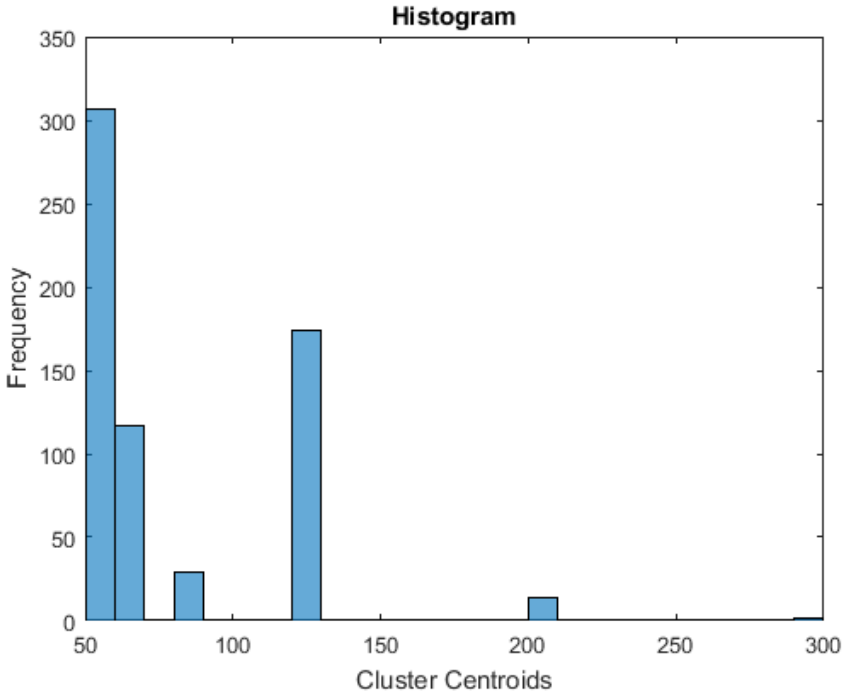

Figure 4-4: An example of distribution of clusters

Now that the clusters are identified, Figure 4-4 shows a histogram of clusters within the 10:00 a.m. to 10:30 a.m. interval. It can be observed that within a

particular time interval, there are many clusters of travel time and each cluster has different sizes. To describe the travel time pattern within a particular time interval, the histogram is converted into a cumulative probability distribution of clusters which is then fitted with Weibull Distribution described in Theorem 4.

Theorem 4 (*Weibull Distribution*)**.** A Weibull Distribution is described by two strictly positive parameters, the scale parameter α and the shape parameter β, and has a **probability distribution function** of [8]

$$f(x|\alpha,\beta) = \frac{\beta}{\alpha} \cdot (\frac{x}{\alpha})^{\beta-1} \cdot e^{-(x/\alpha)^{\beta}}. \tag{4.1}$$

Correspondingly, its **cumulative distribution function** is given by [7]

$$F(x|\alpha,\beta) = \int_0^x f(t|\alpha,\beta)dt = 1 - e^{-(x/\alpha)^{\beta}}. \tag{4.2}$$

Figure 4-5 shows an example of fitting a cumulative probability distribution with Weibull Distribution. The red line represents the cumulative probability distribution and the blue line indicates the best Weibull Distribution fit estimated by maximum likelihood.

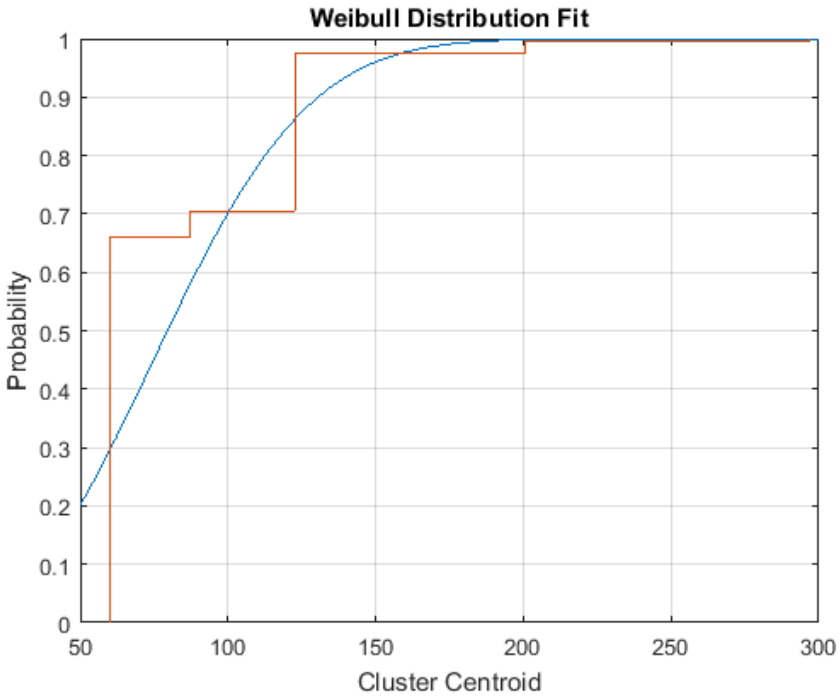

Figure 4-5: Weibull Distribution Fit