UCL DEPARTMENT OF GEOGRAPHY

YEAR 2016-17

EXAM CANDIDATE ID:	**QRSH1**
MODULE CODE:	**GEOGG153**
MODULE NAME:	**Mining of Social and Geographic Datasets**
COURSE PAPER TITLE:	**Mobility Analysis of Gowalla Users**
WORD COUNT:	**2985**

Are you registered as dyslexic with UCL Student Disability Services (SDS) and been given labels to 'flag' your written work

~~YES~~ / NO ***(please delete as applicable)***

Mobility Analysis of Gowalla Users

Contents

1. Introduction

This report covers the analysis of mobility patterns of Gowalla users, a mobile-based geo-social network that was active between 2007 and 2012. It allowed users to 'check in' to visited locations and share photos and experiences with other users (Gowalla Incorporated, 2010).

The aim of this analysis is to visualise and characterise the mobility patterns of three users, 177, 486, and 551, through in-depth comparisons investigating both the distance travelled, and the locations visited whilst using Gowalla. *K*-means clustering is also implemented to investigate the temporal distribution of check-ins, and identify significant locations for each user at different times of day. A summary of the activity dates and 'check-ins' for each user can be found in Table 1.

The data used for this analysis has been obtained from the Stanford Large Network Dataset Collection (Leskovec & Krevl, 2014). All analysis has been carried out using Python, with the visualisations created using Folium.

Table 1: Summary of Gowalla user activity.

User	*Total Check-Ins*	*Date Active*	*Date Inactive*	*Total Days*
177	1550	19/12/2009	29/03/2010	100
486	1425	03/10/2009	20/10/2010	382
551	1950	15/01/2010	19/10/2010	277

Figure 1 shows the mobility patters of each user. Although each user's mobility pattern is clearly different, each user's spatial footprint overlaps with that of another user; users 177 and 486 have both visited San Francisco, and users 486 and 551 have both spent time in and around Dallas, Texas. The spatial and temporal check-in distributions are further analysed in section 3.

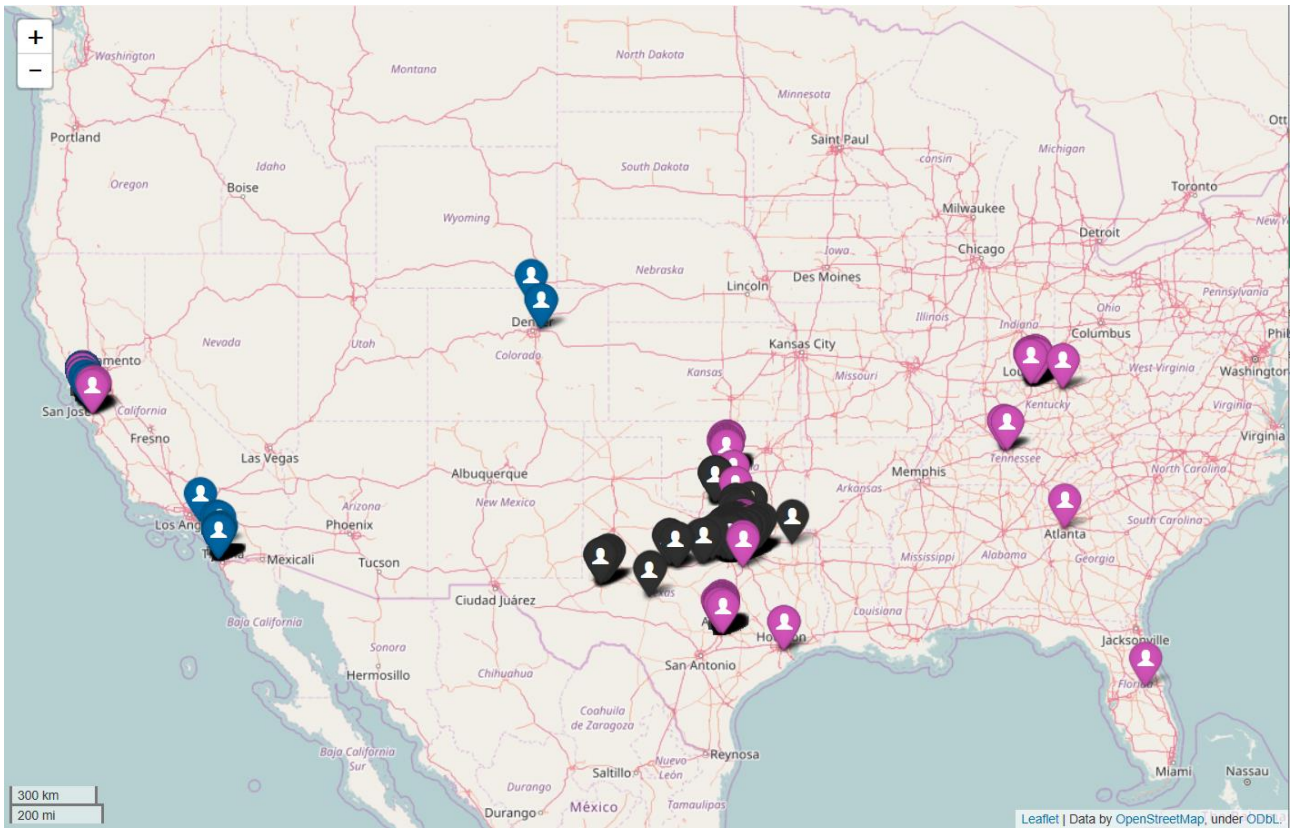


Figure 1: All check-in locations users 177 (blue), 486 (pink), and 551 (black). Whilst user 486 has checked in at many locations across the United States, user 551's check-ins are limited to the region surrounding Fort Worth, Texas. In comparison, user 177s check-ins are restricted to 3 areas; Southern California, San Francisco, and Denver.

2. User Characterisation

For first 7 days each user was active, the distances of each individual 'trip' were calculated and statistical analyses conducted. A trip is defined as the distance between two consecutive check-in locations, and therefore assumes that the user always checks in at their destination. Therefore, if a user checks in at only one location in a day, this was interpreted as zero trips having been taken. The results are detailed in Table 2.

Despite user 486 having taken only 5 trips, compared to user 177's 440, and travelling on only four of the seven days, user 486 has the highest mean daily trip distance of the three users, at 8.75 km per day, vs user 177's 0.13 km. User 486 also travelled further than user 177, covering a straight-line distance of 80.61 km, compared to 55.08 km.

This suggests that user 486 was perhaps more discerning about the places they checked into, perhaps reserving their check-ins for places of significance, rather than recording their everyday movement, as did user 177. This is reflected in the standard deviation of the mean distance travelled per day, where user 177 shows the least variation from the mean at 0.19 km, indicating that the users' mean daily trip lengths are relatively consistent. Users 486 and 551, however, show significantly greater variation with standard deviations of 8.64 km and 6.42 km respectively. This indicates a more erratic distribution of average daily trip lengths over the week. However, this may also be a function of the lower number of daily check-ins, since user 486 made zero trips on 3 of 7 days. User 177 averages 62.86 check-ins a day, whilst user 486 averages 0.71.

The maximum distance travelled by any user on any one trip was 30.94 km, a trip made by user 551 on day 3. This user also travelled the furthest in this time period, totalling 257.82 km. In contrast, user 177 has the lowest maximum trip distance of all the users, with a maximum of 1.77 km recorded on both days 1 and 6. This low maximum trip distance is reasonable given the frequency with which the user checks in compared to the other two users, since more frequent check-ins would logically lead to shorter distances travelled between locations.

Table 2: Descriptive statistics characterising the mobility patterns of users 177, 486, and 551 for the first 7 days they were active on Gowalla.

			Day							Daily Mean
User		All 7 Days	1	2	3	4	5	6	7	
177	Maximum	1.77	1.77	0.22	0.25	1.05	0.46	1.77	0.97	0.93
	Mean	0.13	0.25	0.09	0.08	0.12	0.10	0.12	0.17	0.13
	S.D.	0.19	0.37	0.05	0.05	0.18	0.09	0.20	0.21	0.16
	Trip Count	440	20	23	30	74	76	151	66	62.86
486	Maximum	25.38	0.00	25.38	0.00	14.48	2.17	25.20	0.00	9.60
	Mean	16.12	0.00	19.38	0.00	14.48	2.17	25.20	0.00	8.75
	S.D.	8.64	0.00	6.00	0.00	0.00	0.00	0.00	0.00	0.86
	Trip Count	5	0	2	0	1	1	1	0	0.71
551	Maximum	30.94	20.20	17.72	30.94	11.53	7.58	18.56	19.54	18.01
	Mean	4.37	2.93	17.72	5.63	7.37	2.36	18.25	3.81	8.30
	S.D.	6.42	4.59	0.00	10.58	3.65	2.36	0.31	5.40	3.84
	Trip Count	59	18	1	7	3	12	2	16	8.43

3. Analysis and Discussion

3.1. Spatial Distribution and Trip Analysis

The mobility patterns of all users were visualised using marker clusters, and through joining temporally-adjacent markers to infer the routes taken between locations.

Analysis of the spatial distribution and check-in frequency of user 177 reveals that over 94% of their check-ins are within San Francisco, with the San Diego and Denver areas receiving 5.7% and 0.3% of check-ins respectively (Figure 2). Figure 3 indicates that the user travels to San Diego and Denver from San Francisco, with 79% of this user's check-ins are in the vicinity of the 4th & King Street Station. This person is therefore likely to live or work in this area. Whilst active on Gowalla, user 177 went on two trips, one to Southern California and one to Denver. The user travelled to Denver via plane, checking into the airport upon either their arrival or departure. This user did not check into San Diego airport during their visit, suggesting that the user may have driven to San Diego. However, this cannot be certain since the user did not check into any destinations between the two cities.

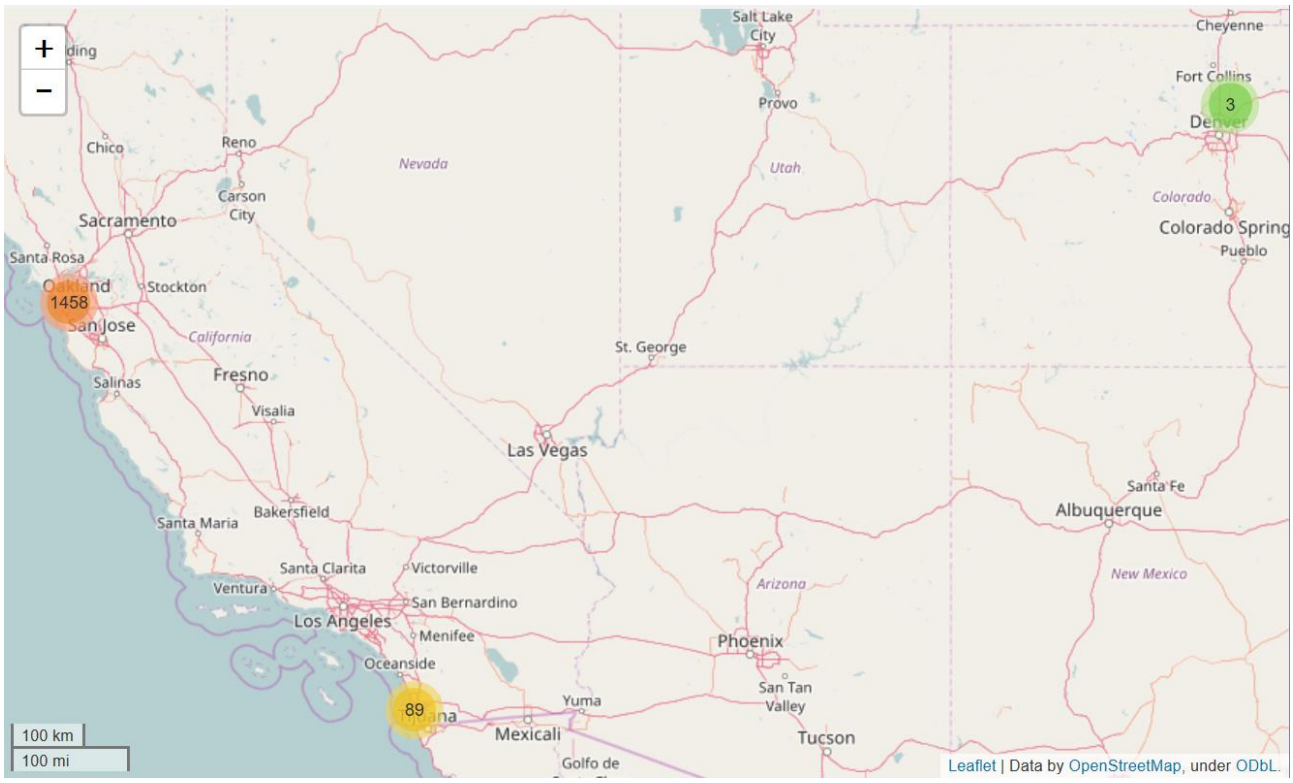


Figure 2: Check-in locations for user 177 showing spatial distribution of check-in frequency.

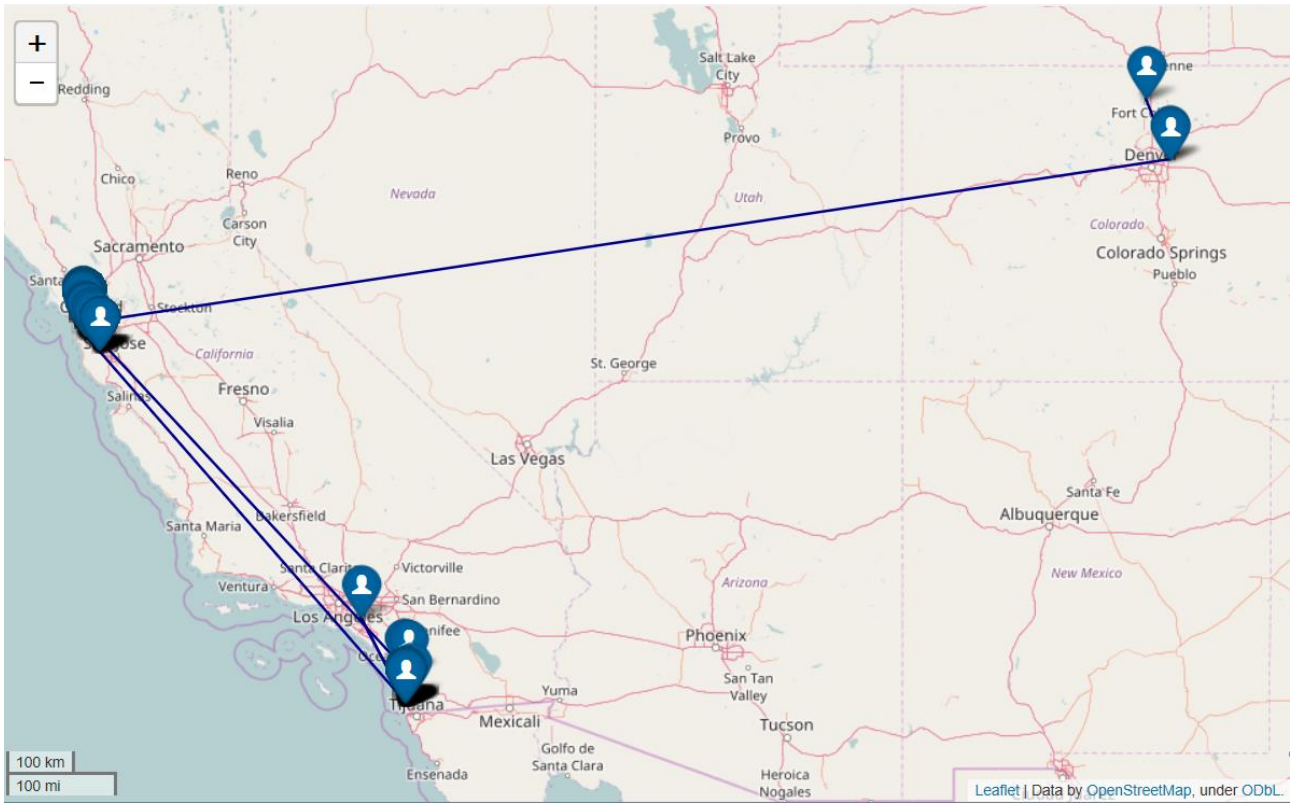


Figure 3: Check-in locations for user 177 showing inferred routes between destinations. Their Denver trip was either their first or last trip recorded using Gowalla, as indicated by the termination of their route line at the Odell Brewing Company, Fort Collins.

Analysis of 486's movements indicates similar a check-in density in both Dallas, and Austin, Texas, with 42% of their check-ins located in each area respectively. The rest of their time is split between the areas surrounding Oklahoma (7.4%), San Francisco (2.9%), and Louisville (3.4%), with single check-ins recorded in Atlanta and Orlando (Figure 4). The majority of trips originate from Dallas,

indicating that this is likely where user 486 lives, rather than Austin, despite having checked into Austin-based locations as frequently as Dallas-based locations (Figure 5). Unlike user 177, where the trips away from home were to one geographic area, user 486's trips are to multiple destinations. For example, one trip comprises check-ins in Florida, Atlanta, and Lexington, and another trip includes both San Francisco and Louisville. A third comprises Austin, Houston, and Nashville (Figure 5). For almost all country-wide trips, the user checks into the airport in each destination city, indicating that the user flies to most of his or her destinations. The combination of cities visited would be considered unusual for a holidaymaker and, combined with the frequency with which this user travels, it is highly likely that they were travelling for work. When in San Francisco, the user checked into the Moscone Convention Center multiple times, supporting the business trip theory.

Trips from Dallas to Oklahoma are the most frequent, with over 20 trips recorded. These trips are likely taken by road, since the user has often checked-in to locations situated along the main freeway (I35) between the two cities, and the user did not check into any airports in the Oklahoma vicinity.

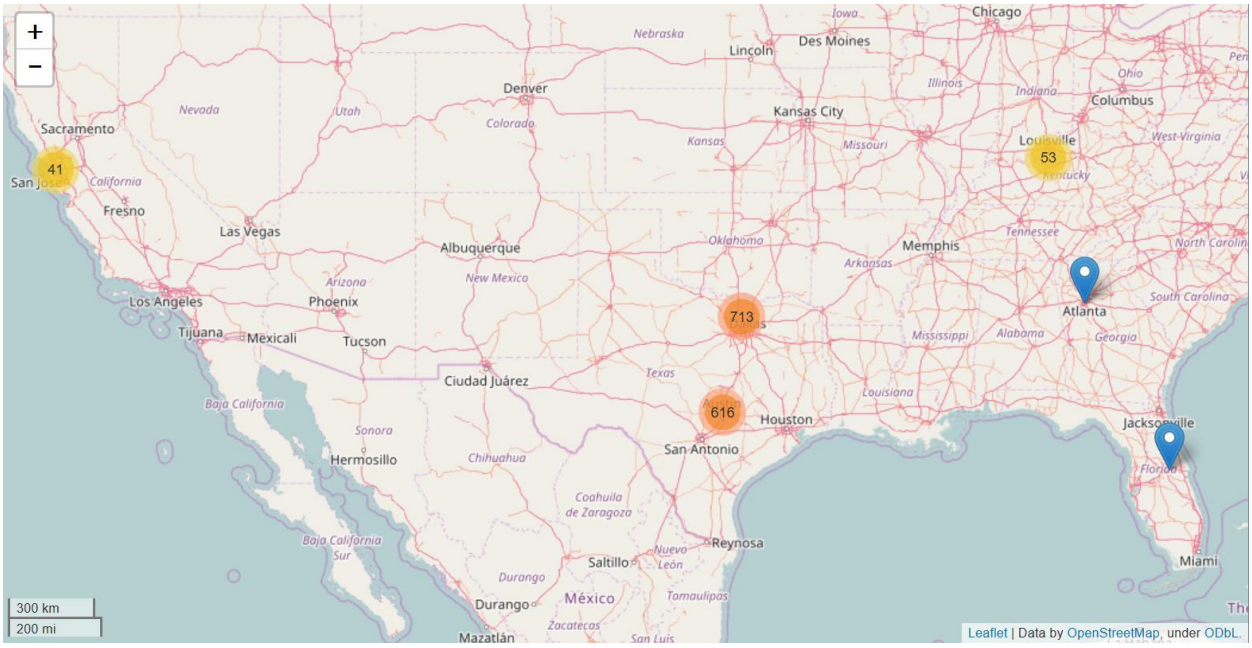


Figure 4: Check-in locations for user 486 showing spatial distribution of check-in frequency.

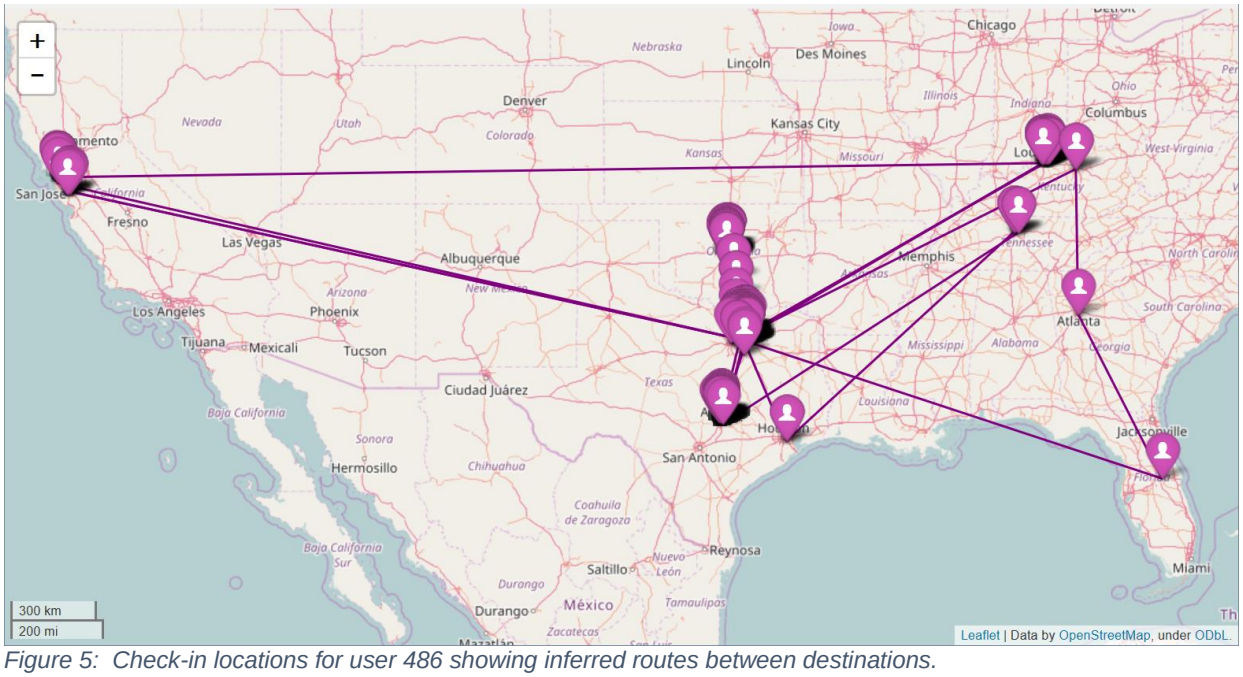


Figure 5: Check-in locations for user 486 showing inferred routes between destinations.

User 551's mobility patterns are more difficult to analyse due to the smaller spatial footprint and less obviously clustered check-ins compared to the other users. The most frequently checked-into city is Fort Worth, with 64% of check-ins. 9% of check-ins were located in Dallas, however an additional 13% of check-ins were recorded in area between both Dallas and Fort Worth (Figure 6). It is therefore likely the user lives or works in Fort Worth. The majority of trips originate from a location just north of the Fort Worth Convention Center. Outside of Fort Worth, user 551 took two separate trips: one to Odessa/Midland, and the other to San Angelo. Both of these trips were taken via road, since the user checked into locations along the main highway/freeway en route to these locations, for example Abilene, Baird, and Thurber. Other trips taken by user 551 include to Duncan, various locations surrounding Azle and Roanoke, Gainesville, and Sherman (Figure 7).

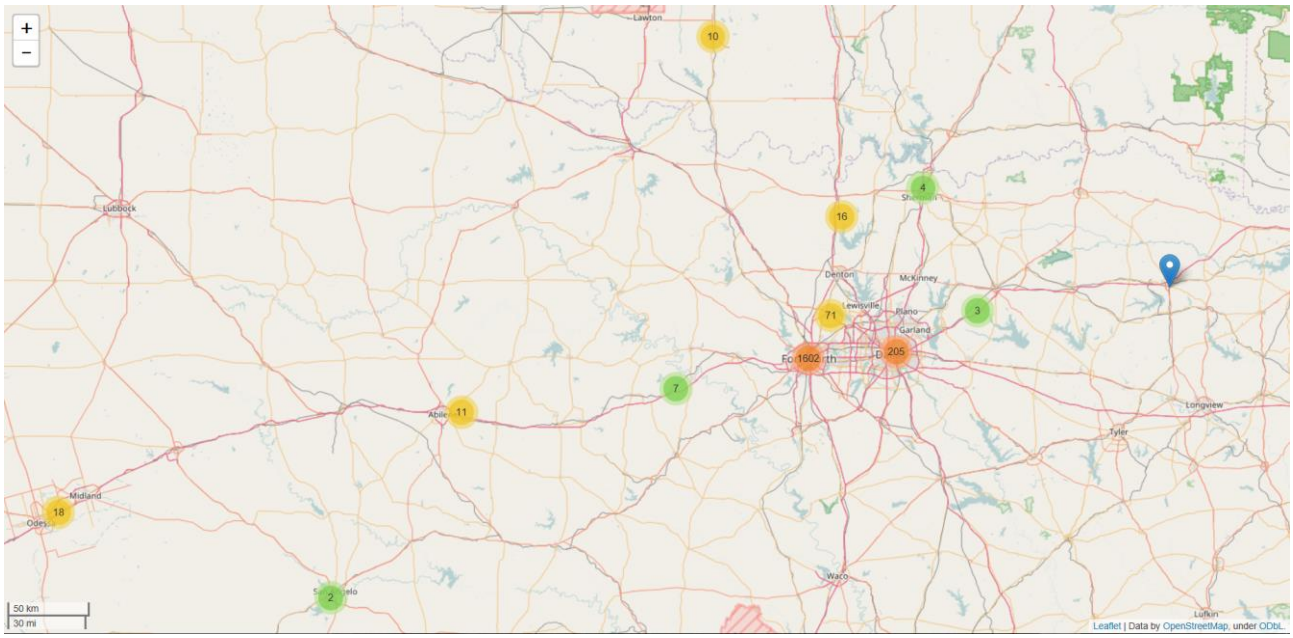


Figure 6: Check-in locations for user 551 showing spatial distribution of check-in frequency.

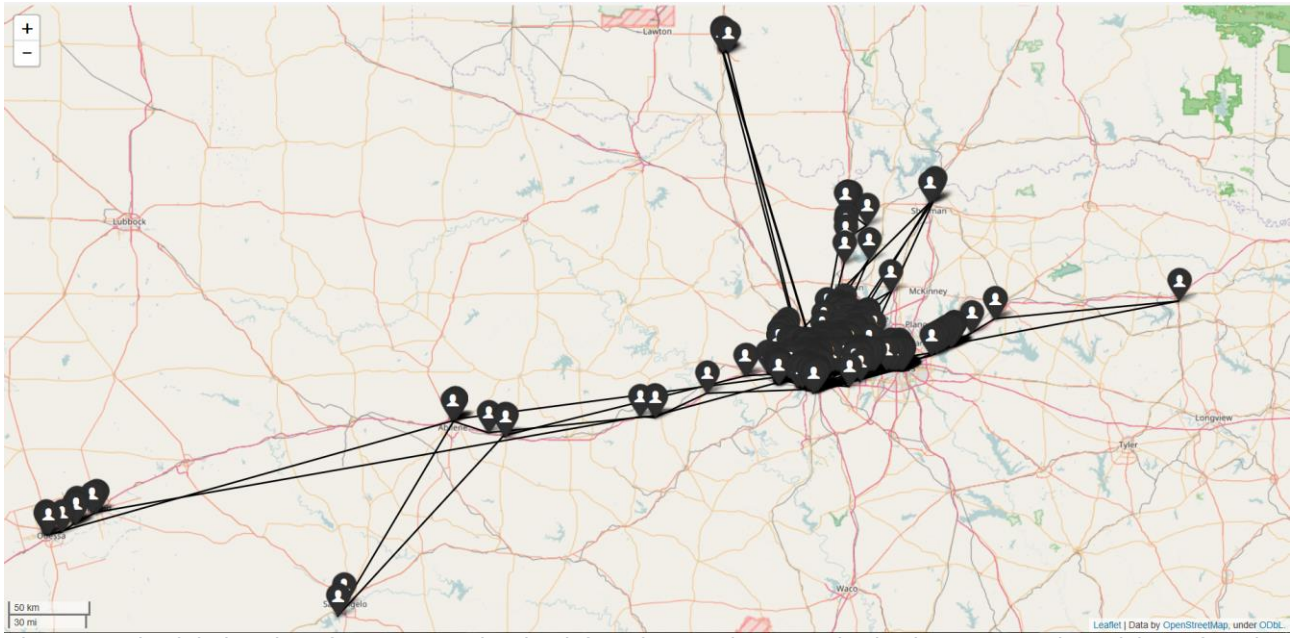


Figure 7: Check-in locations for user 551 showing inferred routes between destinations. Most trips originate from just north of the Fort Worth Convention Center

3.2. Temporal Analysis and *K*-Means Clustering

Method

K-Means clustering was used to identify significant places for each user during the day. The data was divided into three subsets of 8 hours, splitting the data at 00:00, 08:00, and 16:00, the intervals between which are referred to as 'night', 'day', and 'evening' respectively (Table 3).

Table 3: Number of check-ins per user, per time period, used in the k-means analysis.

	Night 00:00-08:00	***Day 08:00-16:00***	***Evening 16:00-00:00***
User 177	990	22	538
User 486	418	41	966
User 551	335	526	1089

K-Means clustering is an unsupervised machine learning algorithm used to divide datasets into *K* categories based on their attributes, in this case geospatial location. The algorithm randomly selects *K* points from the dataset as the centroid of each cluster, and then calculates the distance between each observation and the cluster centroids, assigning observations to their closest centroid. The cluster centroids are then re-calculated, and the process repeated until no observation changes cluster classification. This analysis was performed using different *K* values to determine the most suitable number of clusters for each user (Tan, et al., 2005).

The algorithm often produced empty clusters, a phenomenon that occurred for all users and at all time intervals. This is a common problem and results from poor initialisation of the centroids (Pakhira, 2009). The algorithm was adapted to re-iterate until no empty clusters were returned. As the algorithm produces locally, rather than globally, optimal solution, the final clusters were found to vary considerably each time the algorithm was run.

Analysis

User 177's locations were grouped into 4 clusters for each time period (Figures 8 – 11). Although the user's trips to Southern California and Denver are visually independent, these locations were grouped together. This could be due to the small number of Denver check-ins compared to other destinations; since the initial centroids are selected randomly from the entire dataset, the likelihood of selecting one of these points as a centroid would be negligible.

This user checks in least frequently during the daytime, with locations restricted to the vicinity of the San Francisco 4th & King Street Station. This location is also significant during the evening and night, with 79% of all check-ins located here, suggesting that the user lives within this region. A potential residential address is Avalon at Mission Bay, with several check-ins corresponding to this apartment block.

One cluster of evening check-in locations correspond to restaurants and retail locations in the city centre, whilst two additional clusters are present along the railway line to the south. This correspond to mixture of commercial and residential addresses. A higher *K* number could have been beneficial for identifying significant locations within the city centre, since these are all grouped in one cluster. User 177 also checks-into the city centre frequently during the night, and in similar locations to those visited during the evening. These check-ins have grouped into 3 clusters (Figure 11).

During their trip to Southern California, the user most likely stayed in either Hillcrest or the Gaslamp Quarter or San Diego, due to the large number of evening and night check-ins in these locations. The precise location cannot be pinpointed (Figures 8-10).

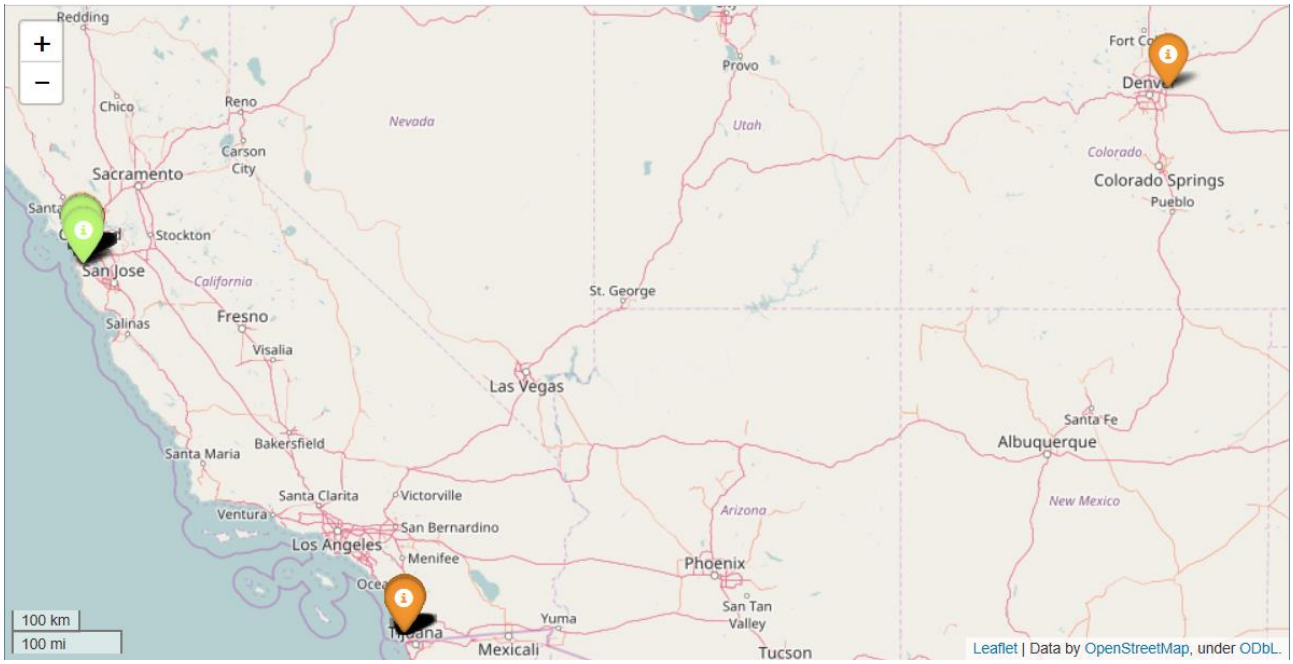


Figure 8: K-means clusters for user 177, 00:00 to 08:00. The locations are grouped into 4 clusters, 3 in San Francisco (lime green, red, peach), and one comprising both Denver and San Diego (orange).

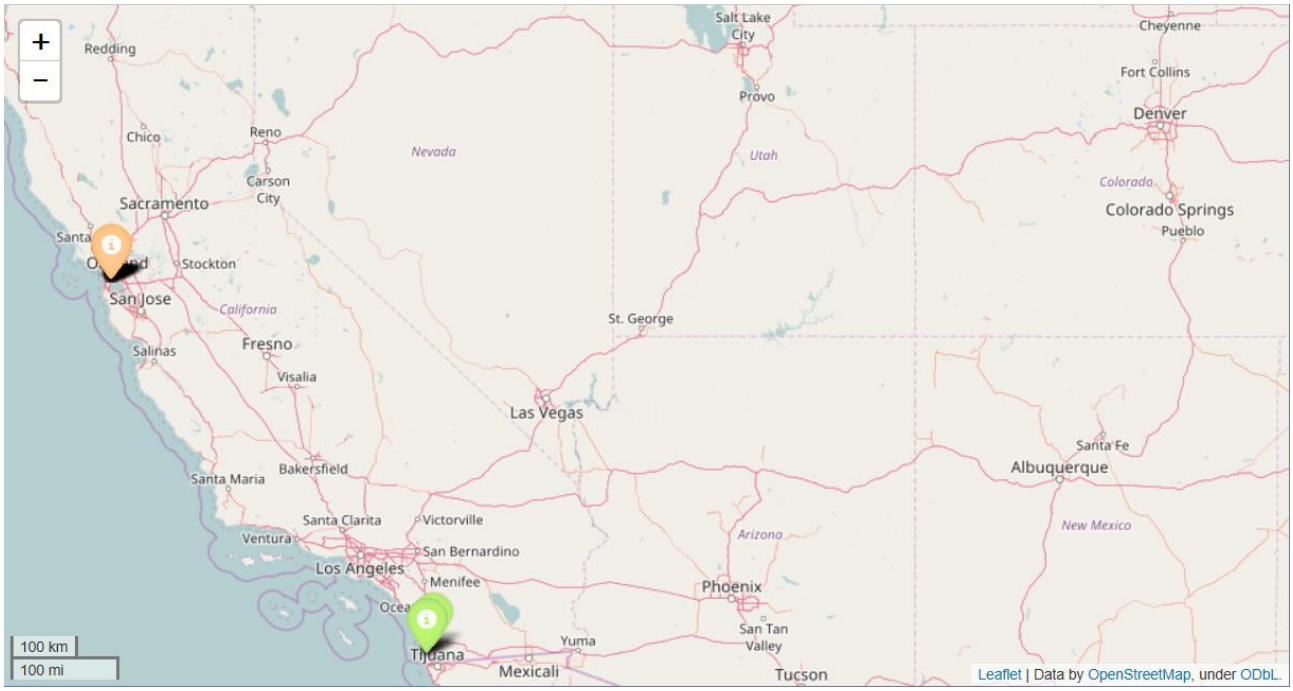


Figure 9: K-means clusters for user 177, 08:00 to 16:00. The locations are grouped into 4 clusters, 3 in San Francisco, and one in San Diego.

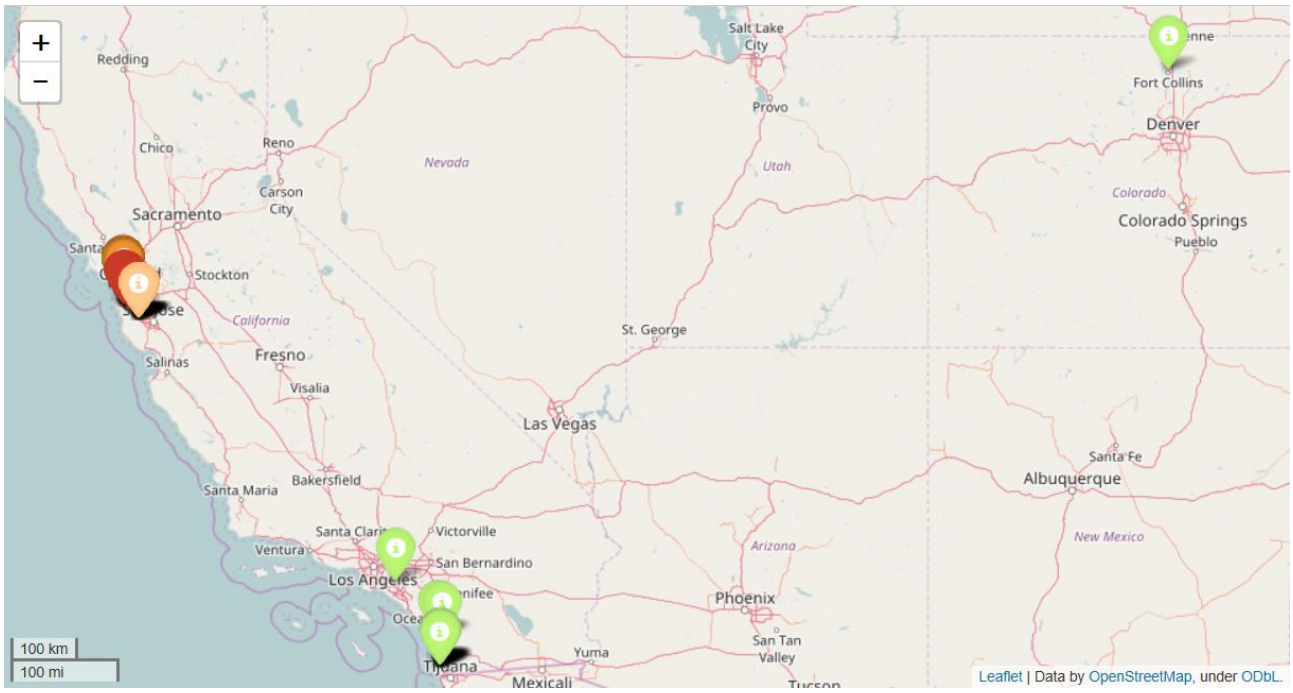


Figure 10: K-means clusters for user 177, 1 6:00 to 00:00. The locations are grouped into 4 clusters, 3 in San Francisco (red, orange, peach), and one comprising both Denver and Southern California (lime green).

User 486's locations were grouped into 6 distinct clusters at all time intervals, largely based on city (Figures 12 – 16). Re-running the algorithm resulted in the creation of similar groupings each time, and increasing the number of clusters produced less spatially-meaningful groupings.

For all three time periods, the Dallas/Fort Worth locations have been grouped into two clusters, however the clusters are not consistent between time periods. Ignoring the *k*-means clusters, the following locations stand out visually: Dallas city centre, which is visited both evening and night, Lake Carolyn and surrounds, MacArthur Park shopping centre, Southlake Town Square, and Vista Ridge, all of which are visited most frequently during the evening and night. It is highly probable that the user lives in the vicinity of Irving, given the large number of recreational locations frequented in this region during the evening and night. However, no clusters are clearly centred on residential or commercial locations so it is difficult to determine precisely where this user lives.

Previous route analysis highlighted that Austin was unlikely to be the user's home; this is supported by the lack of Austin-based check-ins during the day. If they lived in Austin, some daytime check-ins would be expected, for example at weekends. Evening check-in locations are almost exclusively limited to locations stretching from the city centre northwards, along the I35 (Figure 15), whilst night time locations are confined to the city centre (Figure 16).

Another significant location highlighted by the *k*-means clusters is Oklahoma City. The user also checks-in here throughout the day, however locations are spread out and mostly correspond to retail areas. It is unclear whether these trips are for business or recreational purposes. The remaining cities are checked-into during all time periods, indicating that this person stayed in these locations for at least one full day.

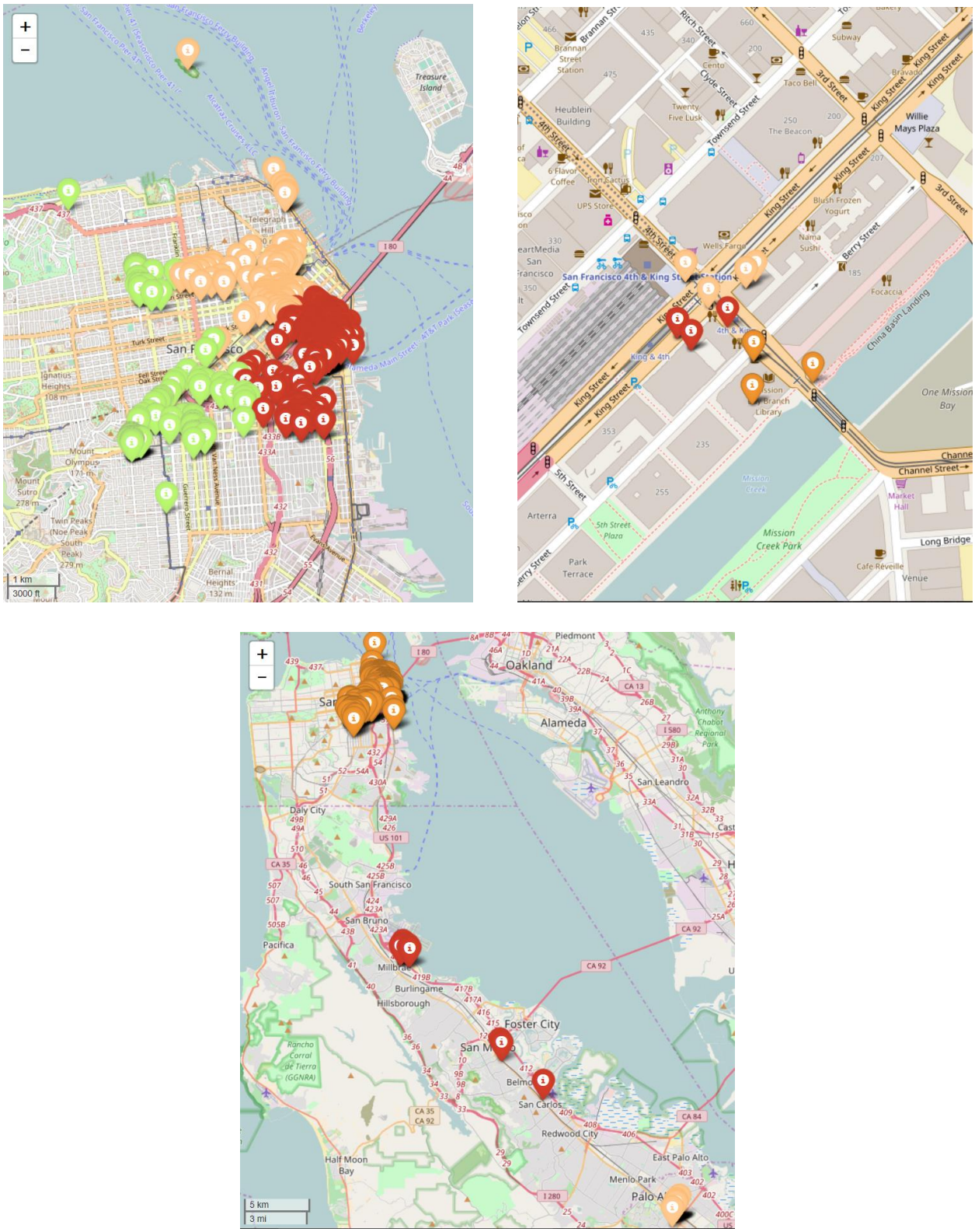


Figure 11: K-means clusters for user 177 in the San Francisco area, 00:00-08:00 (top left), 08:00 – 16:00 (top right), and 16:00 – 00:00 (bottom). ***Night-time locations:*** *The Bay area to the east of James Lick Freeway (close to the user's residence, red); north-west of James Lick Freeway, clustered around 3rd and 4th street (peach); and other, more spread-out, locations to the west (lime green). Further analysis of this third cluster shows that this user frequents the shopping area located on the intersection between Castro and 18th Street.* ***Daytime locations:*** *The user's check-ins are limited to the 4th & King Street Station (all colours).* ***Evening locations:*** *city centre (orange), Palo Alto (peach), and a collection locations along the railway line south of the city; Millbrae, San Carlos, and San Mateo (red).*

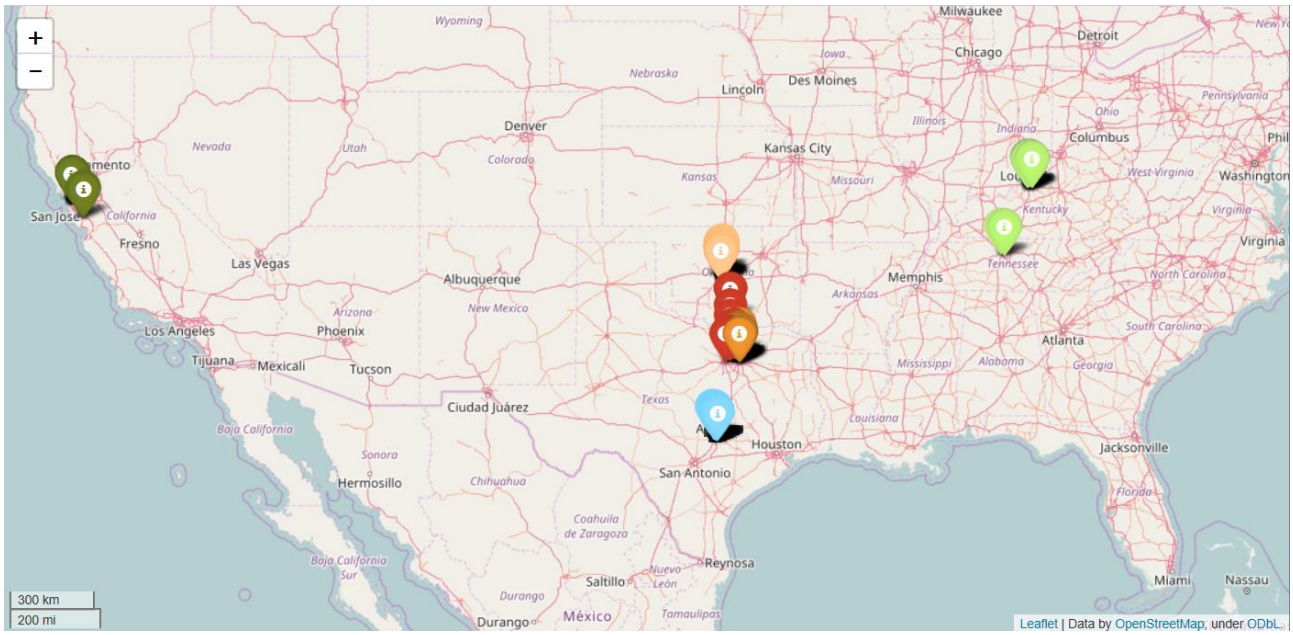


Figure 12: K-means clusters for user 486, 00:00 to 08:00. Locations are grouped into 6 clusters, based mainly on city; Dallas (orange), Fort Worth (red), San Francisco/San Jose (olive green), Louisville/Nashville (lime green), Oklahoma (peach), and Austin/Huston (blue). The user's trip to San Francisco was to the Moscone Convention Center, and so was likely a business trip.

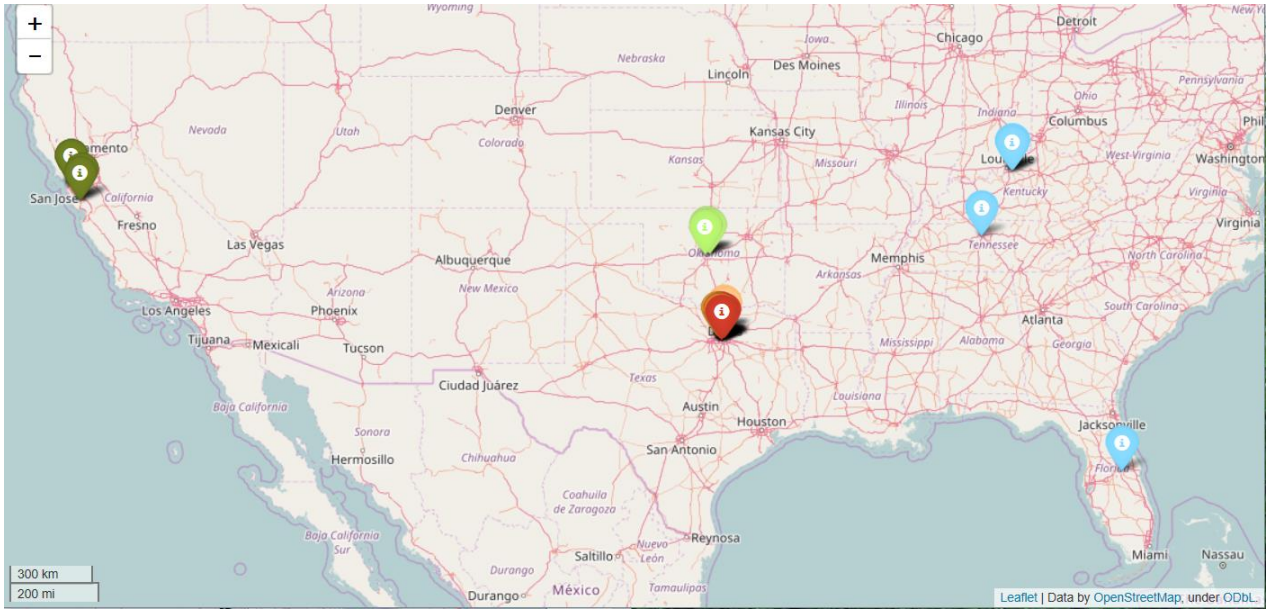


Figure 13: K-means clusters for user 486, 08:00 to 16:00. Locations are grouped into 6 clusters, based mainly on city; Dallas (red), Irving (orange), San Francisco/San Jose (olive green), Louisville/Nashville (blue), Oklahoma (lime green), and Frisco (peach).

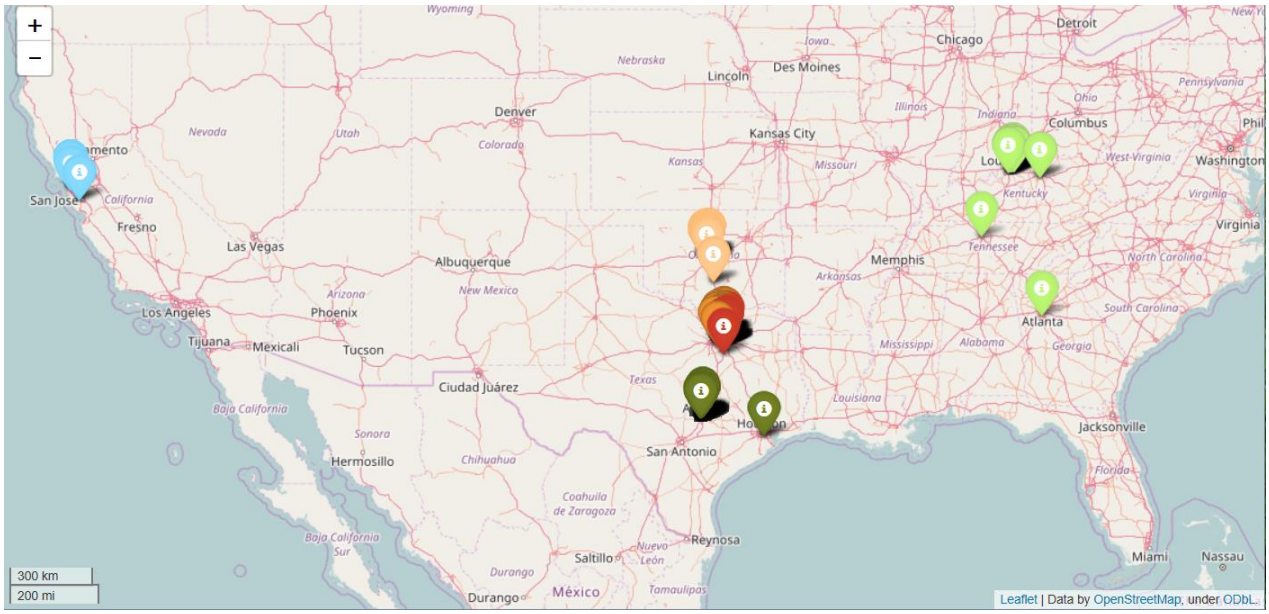
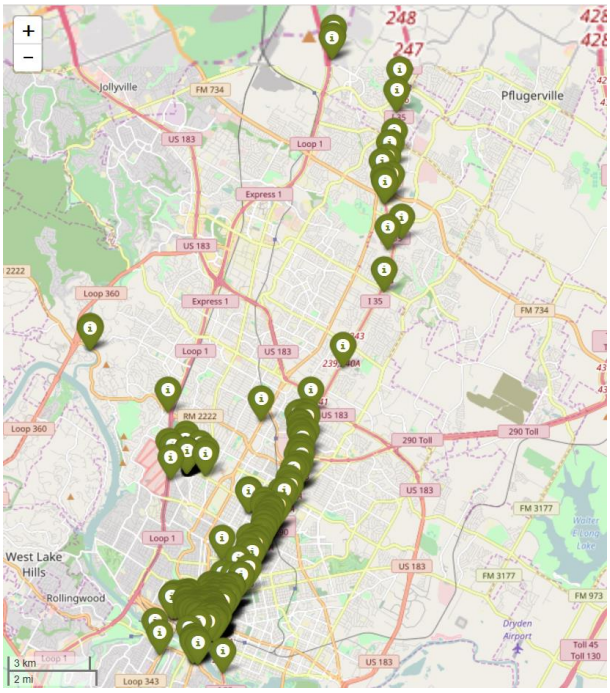


Figure 14: K-means clusters for user 486, 16:00 to 00:00. Locations are grouped into 6 clusters, based mainly on city; Dallas (red), Irving (orange), San Francisco/San Jose (blue), Louisville/Nashville (lime green), Oklahoma (peach), and Austin/Huston (blue).



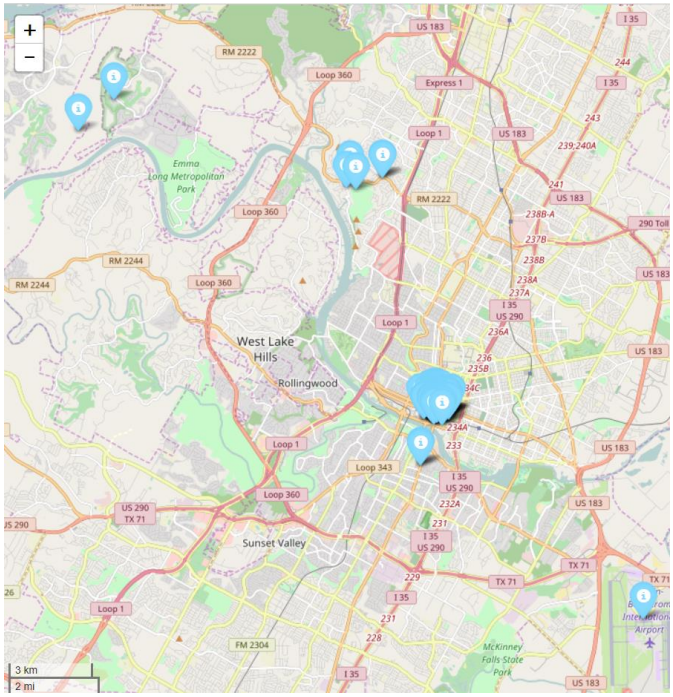


Figure 15 (left): K-means clusters for user 486, 16:00 to 00:00. The user's locations are limited to the I35 and adjacent roads, and a residential area near Camp Mabry (western point cluster). Referring back to Figure 5, the majority of these locations were visited on the same trip, where the user appears to have driven or walked along the road, checking-in to multiple locations along the way.

Figure 16 (right): K-means clusters for user 486, 00:00 to 08:00. The user's locations are limited to Austin city centre, and a cluster of locations around St. Theresa of Lisieux Catholic Church (north). The city centre locations mostly correspond to recreational locations, such as theatres and restaurants, and also a few apartment complexes.

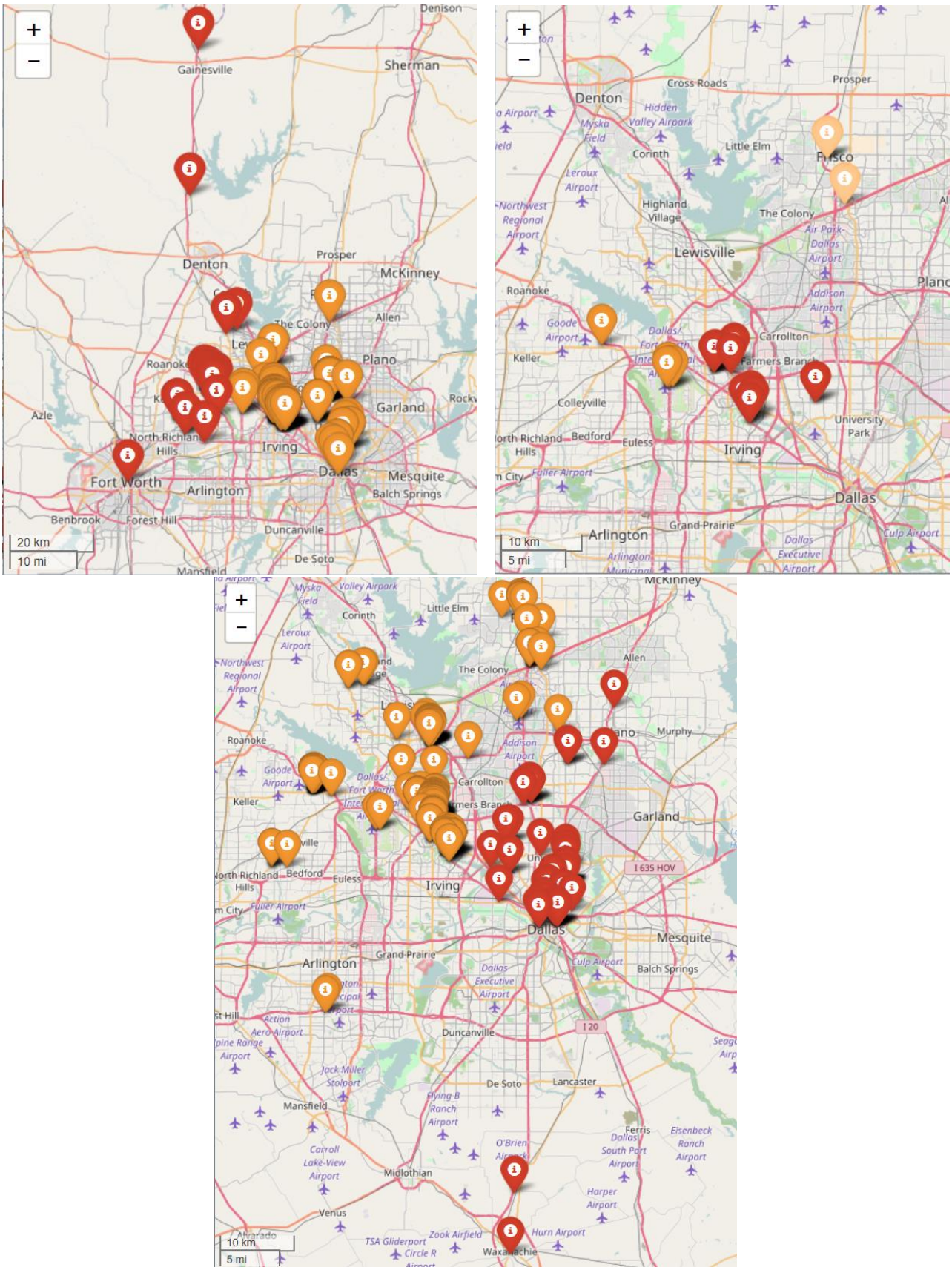


Figure 17: K-means clusters for user 486 in the Dallas/Fort Worth area, 00:00-08:00 (top left), 08:00 – 16:00 (top right), and 16:00 – 00:00 (bottom). A common cluster between all three time periods is the locations just north of Irving. This is where the user is presumed to live, since they spend the majority of their time in the vicinity of this location, visiting recreational locations such as Lake Carolyn and surround, MacArthur Park shopping centre, Southlake Town Square, and Vista Ridge. Given the large number of shopping centres in the area, it would make most sense for the user to visit the one closest to their place of residence.

User 551's locations were more difficult to classify, with dramatically different clusters found each time the algorithm was run. The evening and night-time periods were grouped into 5 clusters (Figures 18, 19, 23, and 24), whilst 8 clusters was found most suitable for the daytime (Figures 20 and 21). Many of the resulting clusters contained visually-distinguishable 'sub-clusters', however increasing the *K* values did not improve the classification. Given the high density of check-in locations for all time periods, and the similar geographic footprint across the day, it is difficult to differentiate between business and leisure trips, and to identify key locations specific to a particular time of day.

Two of the most well-defined daytime clusters are centred on University Park Village, and Sundance Square, both of which are shopping centres (Figure 21) . Further analysis highlights many other retail zones visited by this person all over the Dallas-Fort Worth region, throughout the day. Evening locations for user 551 are less obviously clustered, with a large number of check-ins distributed along major roads within the Fort Worth area (Figure 23). An area common to both daytime and evening locations is the area north of the Convention Center, although this is only identified as a separate cluster during the daytime period (Figure 20). This was highlighted previously as the origin of a high percentage of trips. There is a FedEx office located in this region, so it is possible the user is a delivery driver. This would explain the similarity in the locations visited across time periods, and the large number of daily trips and retail parks visited. It would also explain the somewhat random nature of the user's movements, since delivery drivers would be expected to deliver to anyone within their catchment area. There are FedEx offices in central Dallas, another significant location for this user.

Trips outside the Dallas-Fort Worth area correspond to hotels and eating establishments, or points of interest, such as the W.K. Gordon Center for Industrial History of Texas, Thurber. These are likely recreational trips. Odessa and San Angelo check-ins are also likely personal due to the nature of their destinations (Figure 22).

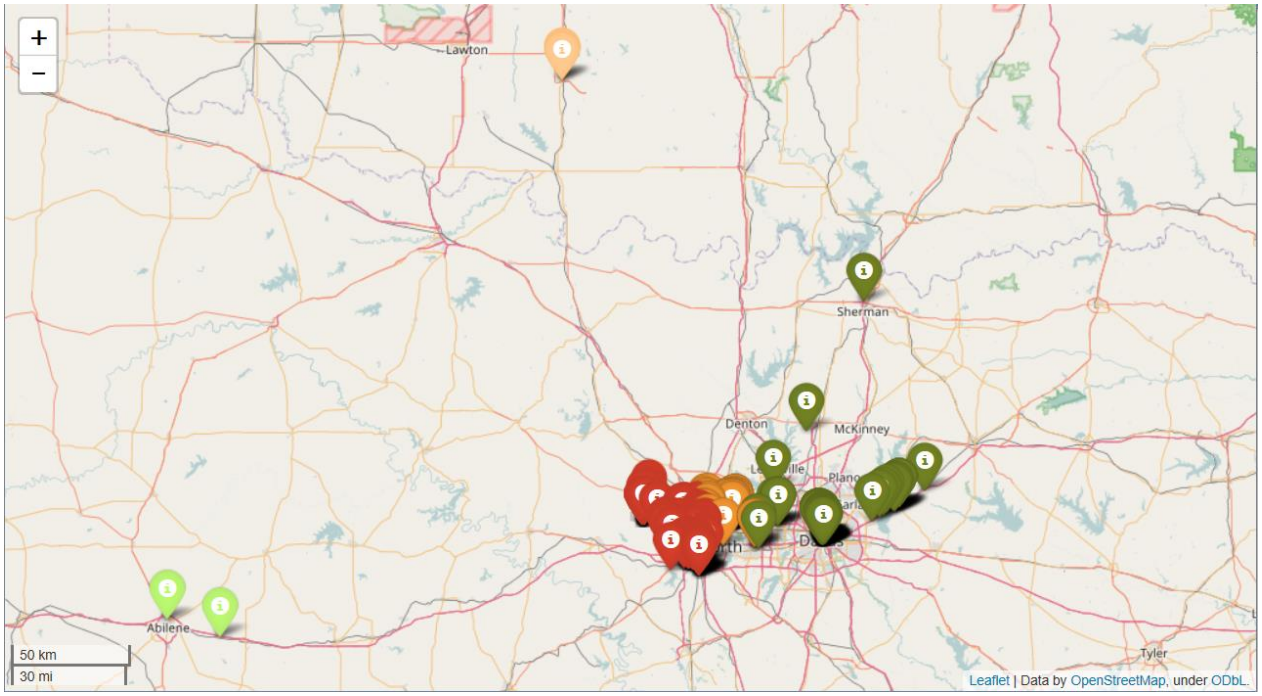


Figure 18: K-means clusters for user 551, 00:00-08:00. The 5 clusters are: Abilene (lime green), Dallas and surrounds (olive green), Fort Worth and west (red), between north-east Fort Worth and Arlington (orange) and Duncan (peach). The Abilene cluster locations are Hardin-Simmons University warehouse in Abilene, and the Happy Trails Enterprises in Clyde both of which are unusual locations to visit for non-business purposes, particularly between 00:00 and 08:00.

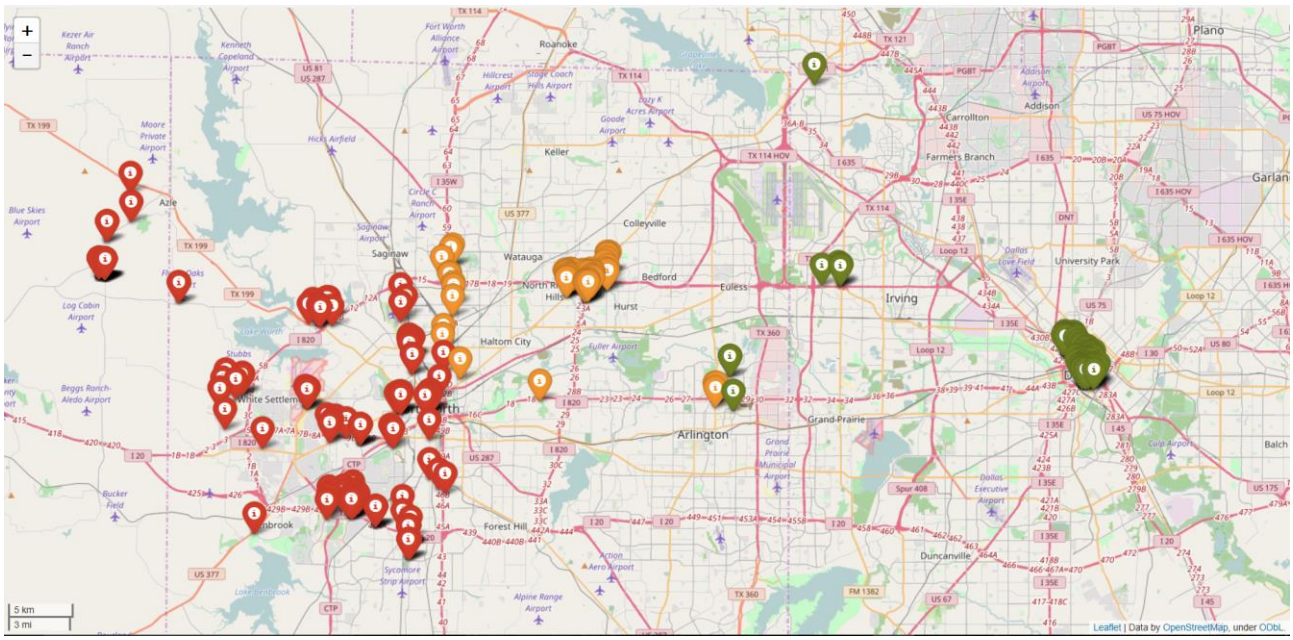


Figure 19: K-means clusters for user 551, 00:00-08:00, Dallas/Fort Worth vicinity. The dense sub-cluster of orange points surrounds the North East Mall, the northernmost red cluster is the Lake View Boulevard shopping area, and the southernmost red cluster is the Hulen Mall and Cityview Center. The majority of these locations appear to be recreational, however this is difficult to fully determine, since the user spends such a significant proportion of their time at retail locations. Although these locations are classified as night time, it is possible that this person's job requires them to work from early in the morning before the shops open.

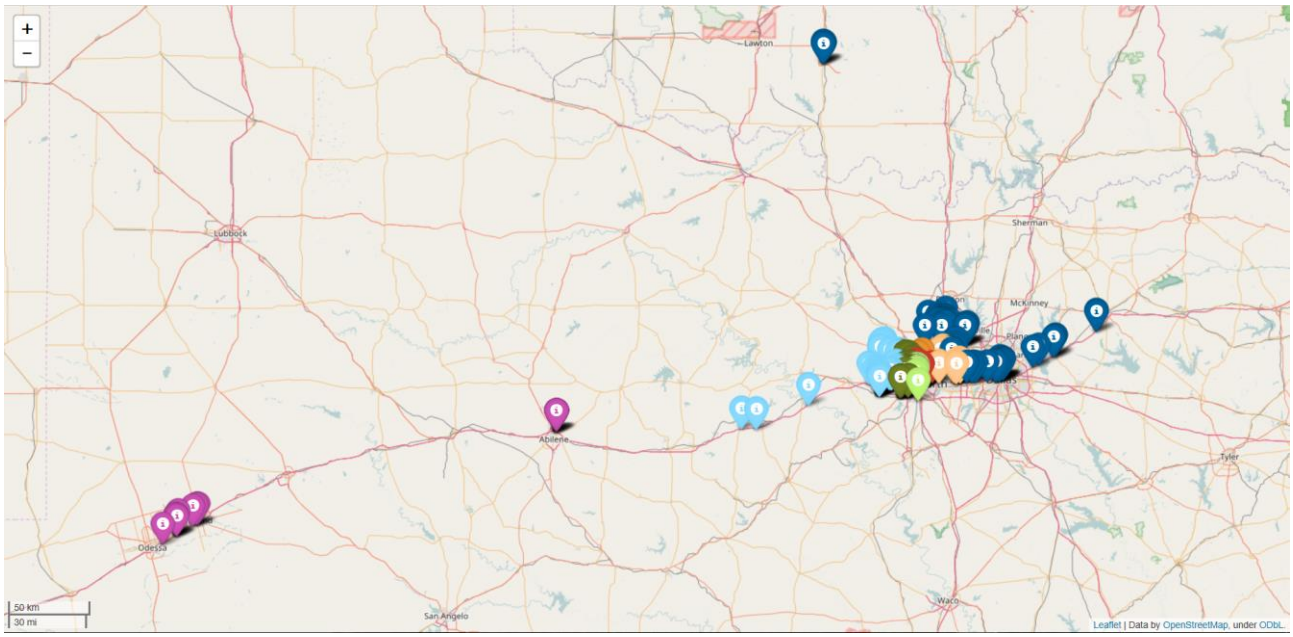


Figure 20: K-means clusters for user 551, 08:00-16:00. 8 clusters are: Abilene (pink), main roads in outskirts of Fort Worth (light blue), western Fort Worth (olive green), Central Fort Worth, north of Convention Center (red) University Park Village (lime green), inter-city region including Arlington and Hurst (peach), Dallas and Surrounds including Roanoke (dark blue), three ambiguous points in north-west Fort Worth (orange). Dark blue markers to the east of Dallas along the I30 represent Plaza Rockwall, and Rockwall Market Center, also shopping centres.

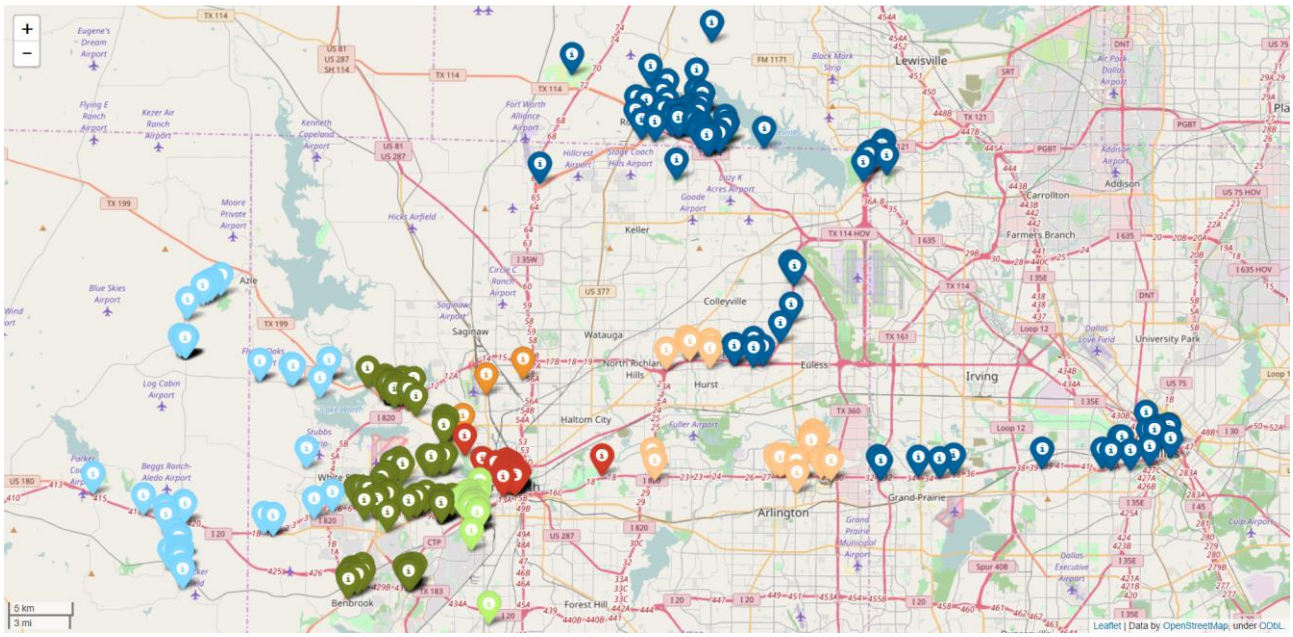


Figure 21: K-means clusters for user 551, 08:00-16:00. Further investigation reveals that high-densities of check-ins correspond to other shopping malls and retail areas, such as Cityview Center (southernmost olive green cluster), and Lake View Boulevard (northernmost olive green cluster) in the Fort Worth suburbs. This trend is repeated in Dallas, where the user checks-into retail locations along Main St.

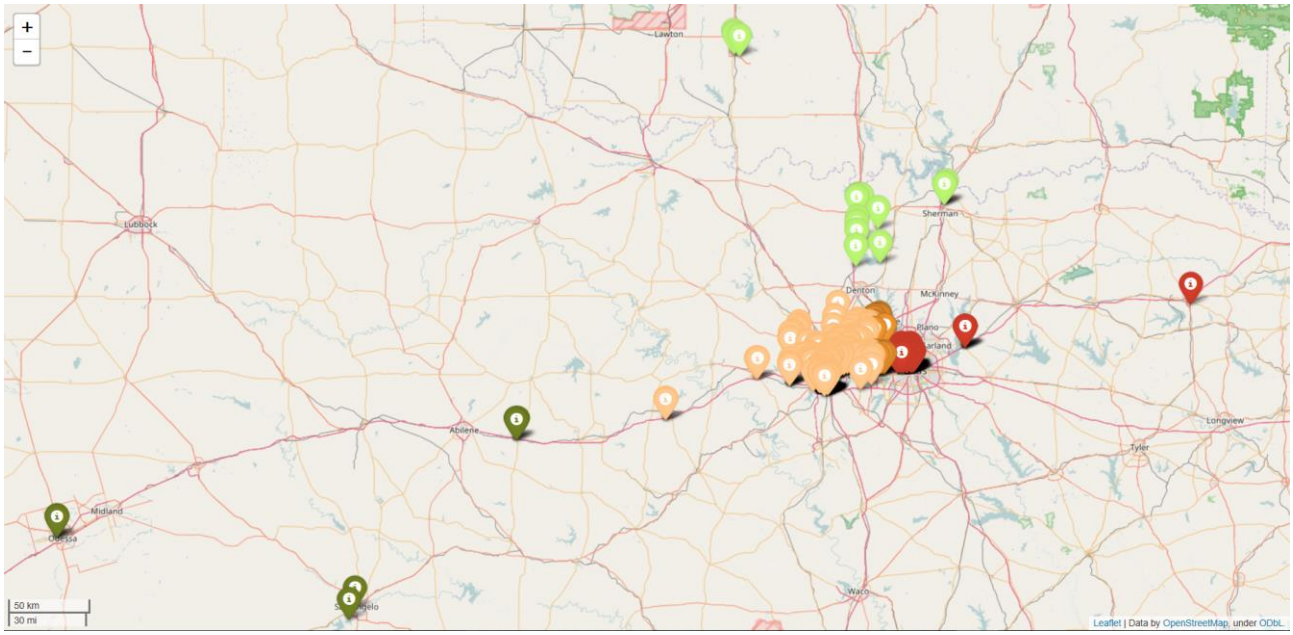


Figure 22: K-means clusters for user 551, 16:00-00:00. 5 clusters are: Fort Worth (peach): k-means struggled to divide this cluster meaningfully, although obvious sub-clusters exist within it, Abilene (olive green), inter-city region (orange), central Dallas (red), and Gainesville area (lime green). Gainesville is only visited during the evening, as are San Angelo and Odessa (easternmost olive green points). In Odessa, the user checks into RadioShack and Walmart Tire & Lube Express however, due to their presumed as a delivery driver, it is difficult to determine whether these locations were visited for work or leisure purposes. Given that these locations were visited during the evening, and are away from the user's home, it is likely that they were visited for personal reasons rather than business. Evening visits to St Paul Presbyterian Church, and the Packsaddle Bar-B-Que in San Angelo are also more likely to be recreational, further supporting the conclusion that these trips outside of Fort Worth are for recreational purposes.

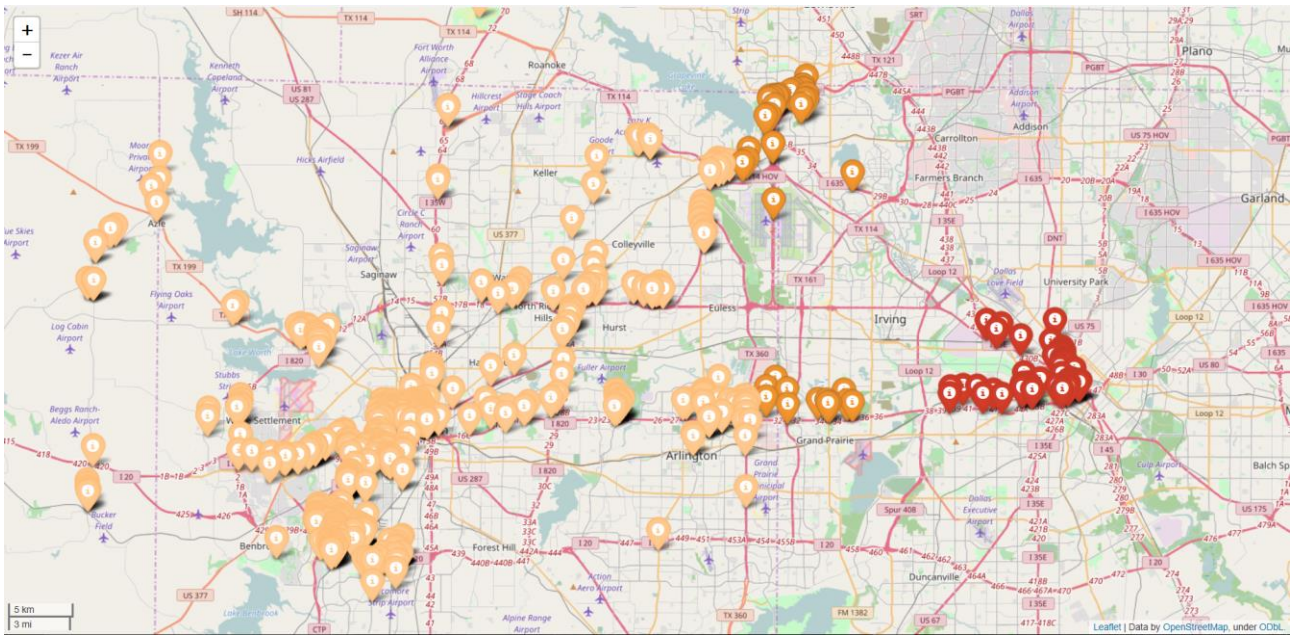


Figure 23: User 551 checks-in least frequently between 00:00 and 08:00, and so significant places are easier to determine.

3.3. Markov Mobility Modelling

Markov mobility modelling was implemented to calculate the probability of the users travelling to a location in another cluster, given the cluster in which they are currently located. Figures 24 – 26 show the resulting transition matrices for each user. Due to the algorithm set-up, it was not possible visualise the clusters using Folium, nor was it possible to enforce non-empty clusters.

```
From cluster:  0
    To cluster:  0  | Probability:  0.8823529411764706
    To cluster:  1  | Probability:  0.029411764705882353
    To cluster:  3  | Probability:  0.0
    To cluster:  4  | Probability:  0.08823529411764706
From cluster:  1
    To cluster:  0  | Probability:  0.0
    To cluster:  1  | Probability:  0.9887640449438202
    To cluster:  3  | Probability:  0.0
    To cluster:  4  | Probability:  0.011235955056179775
From cluster:  3
    To cluster:  0  | Probability:  0.0
    To cluster:  1  | Probability:  0.0
    To cluster:  3  | Probability:  1.0
    To cluster:  4  | Probability:  0.0
From cluster:  4
    To cluster:  0  | Probability:  0.0028089887640449437
    To cluster:  1  | Probability:  0.0
    To cluster:  3  | Probability:  0.0007022471910112359
    To cluster:  4  | Probability:  0.9964887640449438
```

Figure 24: Markov transition matrix for user 177 clusters. Each value is the probability of the user's next location existing in each corresponding cluster, given their current location. If the user is in cluster 3, the probability of them staying there is 1. This is potentially Denver, since the user's Gowalla check-ins end at this location. The highest probability of subsequent cluster for each cluster is itself, with no cluster showing significant variation in next cluster. This is likely due to the algorithm not enforcing non-empty clusters, and so it is likely that clusters within San Francisco city centre are not identified.

```
From cluster:  0
    To cluster:  0  | Probability:  0.8962264150943396
    To cluster:  1  | Probability:  0.0
    To cluster:  2  | Probability:  0.009433962264150943
    To cluster:  3  | Probability:  0.0
    To cluster:  5  | Probability:  0.09433962264150944
From cluster:  1
    To cluster:  0  | Probability:  0.0
    To cluster:  1  | Probability:  0.9512195121951219
    To cluster:  2  | Probability:  0.0
    To cluster:  3  | Probability:  0.0
    To cluster:  5  | Probability:  0.04878048780487805
From cluster:  2
    To cluster:  0  | Probability:  0.2
    To cluster:  1  | Probability:  0.0
    To cluster:  2  | Probability:  0.4
    To cluster:  3  | Probability:  0.0
    To cluster:  5  | Probability:  0.4
From cluster:  3
    To cluster:  0  | Probability:  0.0
    To cluster:  1  | Probability:  0.018518518518518517
    To cluster:  2  | Probability:  0.0
    To cluster:  3  | Probability:  0.9259259259259259
    To cluster:  5  | Probability:  0.05555555555555555
From cluster:  5
    To cluster:  0  | Probability:  0.008210180623973728
    To cluster:  1  | Probability:  0.0008210180623973727
    To cluster:  2  | Probability:  0.0016420361247947454
    To cluster:  3  | Probability:  0.003284072249589491
    To cluster:  5  | Probability:  0.9860426929392446
```

Figure 25: Markov transition matrix for user 486 clusters. With the exception of cluster 2, the probability of the user remaining in their current cluster is ≥ 90%, with the probability of them changing clusters negligible. If the user is in cluster 2, however, the probability of them staying in cluster 2 is only 40%, with clusters 0 and 5 20% and 40% respectively. This could potentially be their home or work location.

```
From cluster:  0
    To cluster:  0  | Probability:  0.890625
    To cluster:  1  | Probability:  0.0
    To cluster:  3  | Probability:  0.0
    To cluster:  4  | Probability:  0.0390625
    To cluster:  5  | Probability:  0.0703125
    To cluster:  6  | Probability:  0.0
From cluster:  1
    To cluster:  0  | Probability:  0.2
    To cluster:  1  | Probability:  0.2
    To cluster:  3  | Probability:  0.4
    To cluster:  4  | Probability:  0.0
    To cluster:  5  | Probability:  0.2
    To cluster:  6  | Probability:  0.0
From cluster:  3
    To cluster:  0  | Probability:  0.04
    To cluster:  1  | Probability:  0.04
    To cluster:  3  | Probability:  0.76
    To cluster:  4  | Probability:  0.0
    To cluster:  5  | Probability:  0.16
    To cluster:  6  | Probability:  0.0
From cluster:  4
    To cluster:  0  | Probability:  0.03225806451612903
    To cluster:  1  | Probability:  0.0
    To cluster:  3  | Probability:  0.005376344086021506
    To cluster:  4  | Probability:  0.8279569892473119
    To cluster:  5  | Probability:  0.10215053763440861
    To cluster:  6  | Probability:  0.03225806451612903
From cluster:  5
    To cluster:  0  | Probability:  0.004338394793926247
    To cluster:  1  | Probability:  0.0021691973969631237
    To cluster:  3  | Probability:  0.0021691973969631237
    To cluster:  4  | Probability:  0.015907447577729574
    To cluster:  5  | Probability:  0.9667389732465654
    To cluster:  6  | Probability:  0.008676789587852495
From cluster:  6
    To cluster:  0  | Probability:  0.0
    To cluster:  1  | Probability:  0.0
    To cluster:  3  | Probability:  0.0
    To cluster:  4  | Probability:  0.02252252252252252
    To cluster:  5  | Probability:  0.05855855855855856
    To cluster:  6  | Probability:  0.918918918918919
```

Figure 26: Markov transition matrix for user 551 clusters. With the exception of cluster 1, the probability of the user remaining in their current cluster ranges between 76% and 97%. Clusters 1 shows the largest variety in subsequent locations, with clusters 0, 1, and 5 having probabilities of 20%, and cluster 3 being the most probably destination at 40%. This could also indicate their work location.

4. Privacy Implications and Conclusions

Through this analysis it was possible to determine patterns in the user's movements, and identify significant locations for each user during the course of the day. For all users it was possible to determine either their home or work locations, although the precision to which this was achieved depended on the spatial footprint and degree of clustering of the check-in locations. For user 177, their most probable home location was narrowed down to a specific address, whilst user 486's home location was pinpointed to within a large suburb. User 551's work location, and potential job, were identified, whilst their home could not be located.

Although user details are not included in the dataset, this geo-located information could be used to extract street addresses from various online mapping services, such as Google Maps, to determine the exact establishments visited by users. This would likely be a more accurate method of determining the user's home and work addresses, and would also provide in-depth information on

the interests of the user. This information could easily be cross-referenced with other data, such as social media accounts, and potentially result in the identification of the Gowalla account owner. Social media information from public profiles can easily be retrieval using web crawlers, or other data mining methods. Although the precise timing of the check-ins was not scrutinised in this analysis, this information is available and could also be cross-referenced with known events, such as demonstrations, or other locational data such as mobile phone providers' locational datasets, number plate/facial recognition data from CCTV, or in-store financial transactions. These methods could also have been utilised by Gowalla when the platform was active for targeted advertising purposes.

Despite the dataset's anonymity, combining this information with other datasets could lead to the identification of individuals. Under UK law, this would be considered a breach of Article 8 of the Human Rights Act, that protects the 'right to respect for private and family life' (Human Rights Act 1998, 2017). Whether this would also be considered a breach of the Data Protection Act (DPA) is debatable. One function of the DPA is to safeguard the 'right to prevent processing likely to cause damage or distress' (Data Protection Act 1998, 2017). The Gowalla data is made available anonymously, and it would not be possible to identify individuals using this data alone. However, with advances in computing and machine learning algorithms, and in combination with other "anonymous" datasets, it would likely be possible to de-anonymise the data, and likely causing 'damage and distress' to the individuals involved. Various methods exist for further anonymising data, such as adding noise through obfuscation techniques (Brunton & Nissenbaum, 2015).

Limitations and Further Work

A key limitation of this analysis was the unsuitability of the *k*-means algorithm for identifying notable locations visited by the user. Using other clustering methods, such as nearest-neighbour or density-based clusters, could result in more meaningful clusters, as *k*-means is known to struggle when clusters are of differing sizes and densities, are non-circular in shape, and when the data contains multiple outliers (Tan, et al., 2005).

The analysis also assumes that the user records their every movement, whereas in reality users may only record their visits to more significant locations, therefore providing an inaccurate view of users' mobility patterns. It is also assumed that locations are accurate, however the accuracy depends on the accuracy of the GPS position recorded by the user's mobile phone. In locations where the GPS signal weaker due to obstructions, such as tall buildings, the locational data will be less accurate, leading to inaccurate conclusions.

Bibliography

Brunton, F. & Nissenbaum, H., 2015. *Obfuscation: A User's Guide for Privacy and Protest.* London: MIT Press.

Data Protection Act 1998, 2017. *Data Protection Act 1998.* [Online]
Available at: http://www.legislation.gov.uk/ukpga/1998/29/contents
[Accessed April 2017].

Gowalla Incorporated, 2010. *Gowalla.* [Online]
Available at: http://blog.gowalla.com/
[Accessed April 2017].

Human Rights Act 1998, 2017. *Human Rights Act 1998.* [Online]
Available at: http://www.legislation.gov.uk/ukpga/1998/42/data.pdf
[Accessed April 2017].

Leskovec, J. & Krevl, A., 2014. *SNAP Datasets: Stanford Large Network Dataset Collection.* [Online]
Available at: http://snap.stanford.edu/data
[Accessed April 2017].

Pakhira, M. K., 2009. A Modified k-means Algorithm to Avoid Empty Clusters. *International Journal of Recent Trends in Engineering,* 1(1), pp. 220-226.

Tan, P.-N., Steinbach, M. & Kumar, V., 2005. *Introduction to Data Mining.* 1st ed. Boston: Pearson.