Generating Insight from Data

Tailoring Analytic Algorithms and Visualization to Address User Requirements

© 2016 IEEE.
Table of Contents

1 Abstract........................................................................................................................................3
2 Classification of stations by type of usage..................................................................................3
  2.1 Proximity Measure................................................................................................................3
  2.2 Clustering Procedure..............................................................................................................4
  2.3 Clusters....................................................................................................................................4
  2.4 Visualization..........................................................................................................................5
3 Model-based traffic-flow analysis tool .........................................................................................6
4 Acknowledgment ..........................................................................................................................7
5 References......................................................................................................................................7
1 Abstract

There are many different tools available for web analytics for business intelligence and empowerment. To be useful for a user community, data analytics requires ascertaining the users' needs to drive a combination of appropriate analytical algorithms and effective visualization. Should any of these three be missing or tackled without regard for the others, data analysis will be carried out without enabling the users to move from data to action.

Using the example of the Transport for London (TfL) open data set on tube journeys we provide two examples of the combination of algorithms, visualization and user requirements, one of which is described in detail here, while the other is described at a summary level.

Keywords— Cluster Analysis, Data Analytics, Transport for London (TfL), Data Visualization

2 Classification of stations by type of usage

It is easy to determine total and peak traffic flows through the various stations on the tube network – and this allows determination of busy and quiet stations. However, it is clear that there should be another intuitively sensible classification of stations, that is by type of usage, e.g. commuter source station in the suburbs or commuter destination stations near major sources of employment. Classification of stations into these groups will be useful for a range of users interested in the type of usage of a station, e.g. for setting appropriate staffing levels for different user groups, or tailoring advertising to travelers.

Unsupervised data-driven classification [1] can be carried out by the use of various clustering algorithms. There are a wide range of different clustering algorithms available including both agglomerative and divisive hierarchical clustering techniques and optimization-based techniques such as k-means. Several different clustering procedures and proximity measures were evaluated, and while all of them have some mathematical validity, the utility of them for tackling these user requirements varied widely. We expect a successful technique to produce a large number of results which confirm prior expectations, but also to provide some unexpected insights.

To address this specific problem we selected an agglomerative hierarchical clustering approach using a proximity measure based on relative passenger flows as delivering suitable performance.

2.1 Proximity Measure

A proximity measure between two stations was developed. Great care was taken to ensure the following two properties:

- The measure was insensitive to certain differences (total traffic volume, timing of rush-hour peak flows) which were considered easy to investigate and unlikely to lead to the development of new insights;
- The measure was otherwise sensitive to differences in both arrival and departure rates.

First, a station profile was developed for each station by the following process:

- Count the number of arrivals at the station in each 30 minute window, averaged over all weekdays, to form an arrival profile vector a. Similarly generate the departure profile, d.
- Scale these profiles to have a maximum value of 1 – this generates the relative arrival and departure profiles ar and dr.
- Concatenate ar and dr to produce s, the relative station profile.

The use of the scaling was necessitated to remove the dominant effect of station traffic volume, which otherwise dominated the clustering. Due to the variability in peak flows in the morning rush hour as compared to the evening rush hour we considered it sensible to normalize the arrival and departure profiles separately. The choice to normalize via the maximum value, rather than normalizing to unit
power, was made as it was found to emphasize the difference between stations which had similar traffic except for sections of time when one station has no traffic.

The proximity measure between any two stations was then developed, with the desire that it be insensitive to differences between stations which can be expressed as small shifts earlier for departures and later for arrivals. This was achieved by using:

- The Euclidian distance metric as the base proximity measure between relative station profiles.
- The overall measure of proximity was given by taking the minimum over a set of Euclidian distance metrics between relative station profiles with small time shifts, allied in opposite directions to the arrival profile vector \( a \) and the departure profile vector, \( d \).
- We allowed time shifts of \( \pm 1 \) bins, although the analysis was unchanged for larger values than 1.
- Note that this proximity measure is not a distance metric as it does not satisfy the metric inequality.

### 2.2 Clustering Procedure

An agglomerative hierarchical clustering algorithm, [1] was used. Other option considered were k-means type approaches and divisive techniques. Although attractive for investigatory data analysis, k-means (and other similar optimization techniques) were unable to use the non-Euclidian proximity measure we had developed. Divisive clustering techniques would be an interesting area for further study.

The agglomerative technique used group average linkage, [3] to merge clusters. This was preferred over some of the other options:

- Minimal distance led to the formation of a single super-cluster which absorbed one station at a time – this was non-informative for the required use case although was relatively informative about outlier stations.
- Maximal cluster distance failed to link stations with similar behaviour if the peaks in this behaviour were time-shifted relative to each other so the peaks didn’t interact.

These observations emphasize the advantages of using a space-conserving linkage method over space-contracting or space-enhancing linkage methods, [2].

### 2.3 Clusters

Using the method described we created a full clustering dendrogram which enables the stations to be split into any number of clusters. Splitting at 6 clusters provided the greatest insight. We labelled these 6 clusters by observing their behaviours, as follows:

- Commuter Source: 168 stations in this cluster, characterized by a morning departure peak and a flatter evening arrival peak. Usually located in the suburbs of London. Example: Barnet.
- Commuter Destination: 44 stations in this cluster, characterized by a sharp morning arrival peak and an evening departure peak. Usually located in the center of London. Examples: Bank, Canary Wharf.
- Transit: 44 stations in this cluster, characterized by peaks as for a commuter destination, but also a moderate level of both arrival and departure throughout the day. Investigations showed that this cluster includes all of the main-line rail interchanges and many major tube/bus interchanges Example: Kings Cross.
- Social: 3 stations in this cluster, characterized by peaks as for a commuter destination, but with extra arrivals in the early evening and departures in the late evening. Example: Covent Garden
- 2 x Heathrow: Two separate 1-station clusters, characterized by highly variable traffic flows all day with no commuter-based peaks.
This clustering allows a clear split by use-type, and confirms some prior expectations, e.g. that the major rail interchange stations all lie in a single cluster along with several major bus interchanges. The two Heathrow tube stations were identified as outliers; and are clearly separate from each other as well. This distinction is due to the flight patterns, which drive the tube traffic, being significantly different at the two sets of terminals serviced by the tube stations.

2.4 Visualization

We produced a geographic web-based visualization (using D3.js) to both overlay the cluster of each station and show the current usage of it varying across a day. Fig. 1 shows an example screenshot of this visualization. The basic geographic framework on which it is based is the Google JavaScript API [5], which easily allows a set of familiar map-based operations such as pan and zoom, with rapid re-scaling of the map.

There are a number of ways that a user can interact with this visualization – in particular the time shown runs through the day, but a user can pause, adjust and re-start this through a simple and intuitive set of mouse inputs.

Extra information on each station was available on selection (e.g. proportion of current usage due to arrivals vs. departures, or due to users with various ticket types), but some of the most interesting insights came from the geographical nature of the visualization.

Much of the information provided by the visualization is of a confirmatory nature – in that it confirms the prior expectations of anyone with a basic knowledge of London and the tube system. However there are some anomalies which are surprising and offer new insight into the London tube system, e.g. Uxbridge, at the far western end of one line, is a commuter destination station, even though its location suggests it would be a commuter source station.

A user can interact with the visualization by selecting a station for closer observation. Fig. 2 shows a detail of the visualization, for several exemplar stations; this detail displays the station profile in terms of the arrival and departure rate combined into a 2-D parameterized plot. The station indicator moves with the time set in the overall visualization of Figure 1 In (a) the clear structure of a commuter destination
station is obvious, with a large morning arrival peak and a large evening departure peak, (b) shows a very different clear structure with arrivals and departures generally matched, and varying through the day, while (c) shows the mainly unstructured behavior of one of the two Heathrow stations.

The same analysis and visualization tools can be used on other parts of the same data set – e.g. weekend travel data, travel data limited to season ticket holders etc. It would also be possible to combine all of these, by generating and then concatenating station profiles for each group (split by user type, or time of week or by any other way the data allows) before carrying out the clustering and visualization. The selection of a suitable group or set of groups on which to carry out this analysis must be chosen to align with the interests of the users.

![Station Visualization](image)

**Fig. 2.** Close-up of station profile visualization for (a) Canary Wharf, (b) Victoria Station and (c) Heathrow Terminals 1, 2 & 3 at 17:10pm

### 3 Model-based traffic-flow analysis tool

The TfL data set describes over 4 million journeys in terms of their start and end point. It is possible to develop a model describing the likely path and timing of each journey across the tube network, and hence to obtain a time-varying model of the traffic in each section of the tube network. By combining this model with an interactive visualization tool a user can investigate the traffic flow across the network; furthermore they can modify the network, e.g. by closing a tube station, and observing the effects this would have upon traffic flows.

Similarly to the station visualization, this visualization is based upon the Google JavaScript API [5], with the tube lines overlaid. Thickness of the lines was used as an indicator of the amount of traffic on each section of line – other options for line thickness, which we have not fully implemented, which may be of specific interest to some users are:

- Quantity of traffic referenced to the usual quantity of traffic – this visually emphasizes the effect of any changes on the network;
- Quantity of traffic referenced to a measure of capacity (based upon rate and size of trains on the line) – this visually emphasizes likely congestion spots on the network.

The visualization allows a user to address a series of what-if questions in a rapid manner, where the complexity of the underlying model is hidden from them. A tool like this will be useful to people tasked with emergency planning or other similar tasks.

Fig. 3 shows a detail from this visualization demonstrating the predicted effects of closing Leicester Square station. The consequential effects include significantly higher traffic loads upon both the Victoria and Central lines, neither of which directly passes through Leicester Square, but both provide suitable bypass routes.
A further improvement would fully integrate the carrying capacity of the network into the visualization, highlighting where changes have driven required use close to or over the capacity.

Fig. 3. Close-up detail of traffic flow visualization tool for normal for (a) normal rush-hour traffic and (b) traffic compensating for a closure at Leicester Square: Note increase in flow on Victoria line (light blue) and Central Line (red) as they take the bulk of the diverted traffic.

4 Acknowledgment

This work was powered by TfL Open Data [4]

5 References