Identifying Hubs and Trends in Origin-Destination Data

Daniel Devine, Adam Gaventa, Daniel Hammocks, Tim Laibacher, and Mansoor Yousuf

Abstract—In this paper we review current literature on origin-destination (OD) data and how they address two main groups of analytical questions that are asked of OD data. We categorise these as 'identifying hubs' and 'identifying trends'. We discuss the visual and analytical methods used to answer these questions, their relative advantages and disadvantages, and common pitfalls of OD data.

1 Introduction

Origin-Destination (OD) analysis is a pivotal part of visual analytics, with applications spanning numerous domains, such as transportation and journey data, migration and movement of goods. OD data focuses on the movement of an entity, from one point in space to another, without considering the path taken. OD data is similar to 'trajectory' data, but for OD data the trajectory of the flow is either unimportant, such as for migration or train journeys, or does not exist, such as in mobile phone records. These spatial OD flows are often considered temporally; that is, the same spatial paths are analysed at different moments in time. Over the past century our means of transportation has developed rapidly, resulting in a spike of circular migration; often for the purpose of commuting to and from a location of employment due to urbanisation. In conjunction with the advancement of GPS devices and computerised ticketing services, petabytes (if not exabytes or greater) of movement data are being continuously collated, using a variety of methods such as GPS data from taxis, smartcard data and geolocation phone data. The aforementioned points combined with the availability of open-source transport data has resulted in a flourishing of pioneering methods and techniques for OD data.

We have chosen to focus on evaluating papers relating to the transport and migration domains. By comparing recent literature in those domains, we have identified two main categories of analytical questions:

- Identifying hubs, where a large proportion of flows either originate from or terminate.
- Identifying trends, whether in direction and length of flow, spatial distribution of flows, or temporal trends.

The paper is organised as follows: Section I discusses the methods used to identify hubs. Section II discusses the methods used to identify various types of trends. The final section discusses some important issues to be aware of when working with OD data.

2 Identifying Hubs

We found that one of the most common analytical questions that arise out of OD datasets is the identification of major hubs and areas of interest. These are locations that have a large amount of in- or outflows and areas with a high concentration of individual locations that exhibit high amount of in- and out-flows. The motivation for this question might be to identify the best pick-up points for tai drivers - where are the major areas for collectios or drop-offs. In this section we review four main techniques that can be used to answer those questions: flow diagrams, color schemes, heatmaps and clustering.

2.1 Flow Diagrams

A novel approach to address the question of identifying major hubs is the flow diagram approach proposed by Adrienko et al. [1]. Based on an informal survey amongst professionals familiar with OD-maps, they found that flow diagrams, as displayed in Fig. 1, are well suited to identify major hubs. Flow diagrams separate flows into discrete classes according to their length and direction, while the size of each'interval spoke' indicates the magnitude of flows in a certain direction for a certain length class. Round trips are indicated by the size of the radius of a grey ring around the centre. In the upper part of Fig. 1 the Waterloo and Kings Cross docking stations can be easily identified as the main outflow hubs by looking at the overall magnitude of the spokes. In contrast to traditional direct flow maps, flow diagrams therefore ease the identification process of hubs by clever aggregation which reduces the amount of occlusion and clutter. That being said, in cases where several major hubs are located closely to each other, intersections of their flow diagrams can occur. In this case some form of partition clustering might be necessary. We will further explore flow diagrams in Sect. 3.1, where we evaluate their usefulness in the identification of spatial flow trends in terms of direction and distance ranges.

2.2 Colour

In a study by Guo and Zhu [6] to examine the flow of migration in the USA the author demonstrates how colour can illustrate flows of net migration and eliminate clutter from the visual representation of the data map. In Fig. 2 blue colours are used to represent places with net out-migration while red areas are the 'hot' destinations with more inward migration. The varying shades between blue and red are also a useful analytical tool to illustrate the migration patterns clearly. In the research of Fanhas and Saptawati [3] frequent OD movements of taxis were visualised using GPS data. To visualise this data the author used 2-D graphical tools. A unique colour was assigned to each specific neighbourhood in the city of Bandung, Indonesia. Following the analysis of the movement data the unique identifying colour of the origin area is mapped onto the destination neighbourhood. This provides a clear visual representation of the correlation of flows between origin and destination. In the real world it provides the taxi driver with information of the likely destination of a passenger depending where they have been picked up.

The methods used by Guo and Zhu and Fanhas and Saptawati show us that there are different ways to identify origin and destination hubs. To successfully analyse flows of migration Guo and Zhu associated the colours with the magnitude of the flow, which is useful in highlighting flows on a national scale. However, the analysis of taxi trip data in Indonesia looks at a smaller dataset with movements on a more local level. In their example the motivation was primarily to link movements between pick-up and drop-off points. A more specific correlation between origin and destination movements is achieved here which has the result of identifying major destination hubs. The key point here is that

- *Daniel Devine is a postgraduate student at City, University of London. E-mail: daniel.devine@city.ac.uk.*
- *Adam Gaventa is a postgraduate student at City, University of London. E-mail: adam.gaventa@city.ac.uk.*
- *Daniel Hammocks is a postgraduate student at City, University of London. E-mail: daniel.hammocks@city.ac.uk.*
- *Tim Laibacher is a postgraduate student at City, University of London. E-mail: tim.laibacher@city.ac.uk.*
- *Mansoor Yousuf is a postgraduate student at City, University of London. E-mail: masoor.yousuf@city.ac.uk.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

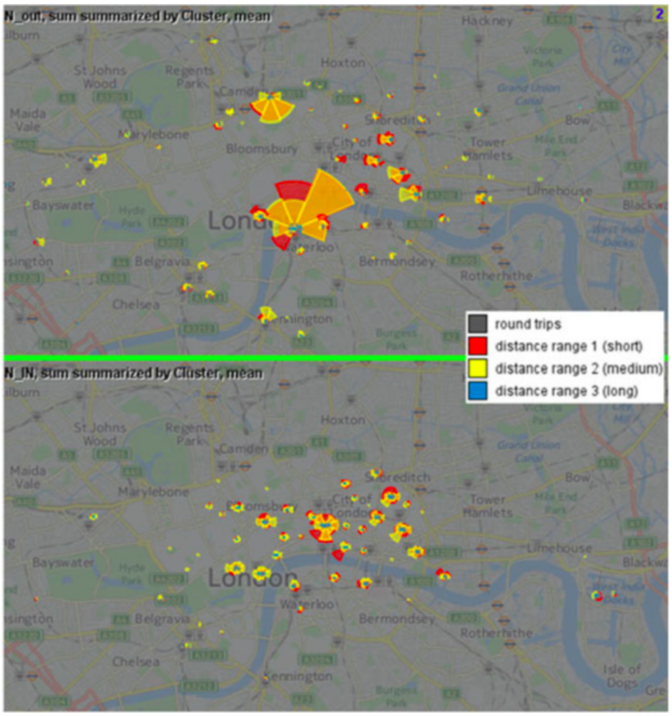


Fig. 1. OD-Flow Diagram, displaying in and out flows for a selected time cluster (summer weekday mornings), from [1].

the precise correlation enables the taxi driver to gain a robust estimate for the earning potential of a fare.
The strength of using colour to visualise OD movements is that the readability of the visualisation is greatly improved. Furthermore, there are a multitude of colours and shades that can be assigned to particular nodes and edges. However, colour choice can be subjective. In the migration example the areas of increased net inflow were assigned a red colour. They are 'hot' areas. An analyst must be aware that this may not be the case in other visualisation examples: in the taxi movement case the colour red does not necessarily mean that this area is a major hub for origin and destination movements.

2.3 Heatmap

Another example of a method to identify hubs in OD data is the heatmap solution presented by Kim et al. [9]. To analyse the origin-destination data of flows by using a heatmap we are required to visualise the data in two separate time intervals as shown in Fig. 3 The flow between origin and destination is visualised by interpreting the difference in heat between the time periods. This solution provides a broader representation of the OD movement data. The author outlines that this method of analysing movements can be used to highlight origins and destinations of disease outbreaks. However, in this case it is necessary to discard all trajectory information of the origin-destination dataset for disease outbreak flows. An advantage of the heat map is that an analyst can look at multiple temporal trends. However, a heatmap does not allow the isolation of specific detailed movements.

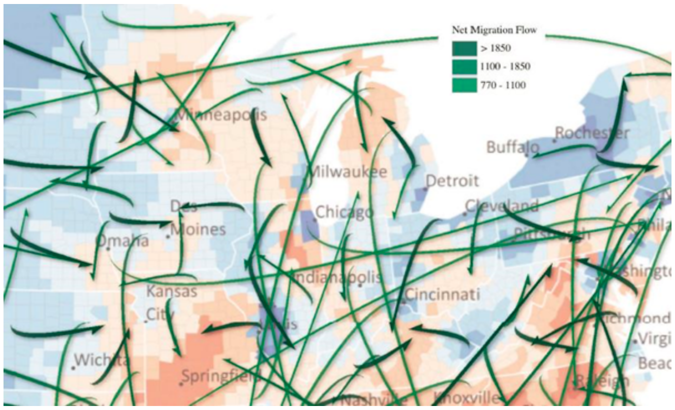


Fig. 2. Visualisation illustrating variation in origin and destination densities, from [6].

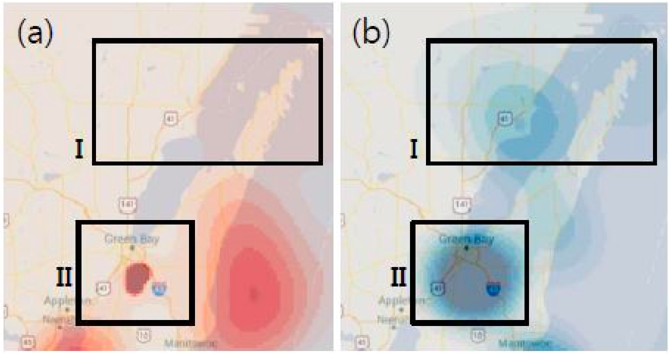


Fig. 3. Heat map with origin and destination flows, from [9].

2.4 Clustering

2.4.1 Hierarchical

With real-world OD data, there will usually be too much data for a visualisation to be meaningful to an analyst. Taking the example of Shenzhen taxi trip data in Zhu and Guo [17], an analyst would want to identify similar flows and nodes, for a variety of uses. For instance, can one identify where major origin points are in the city, such that more taxis should congregate there? Are there certain times where certain locations are more busy than others? Are there some trips which are of a similar profile to others (e.g. distance, number of passengers). When the port is particularly busy, where have people come from? Transport planners may use this information for traffic management or infrastructure planning. Taxi drivers and companies may use it for pricing or directing taxis. Passengers could use it to identify locations where they are most likely to find a driver. In all these use cases, and others, the challenge is both to find these patterns and convey them, from amongst an abundance of other information. One attempt to create a method to answer these questions is Zhu and Guo's [17] hierarchical (agglomerative) clustering method. They define the k-nearest neighbour of a flow F as the set of flows where origins are k-nearest neighbours of the origin of F, and the destinations are k-nearest neighbours of the destination of flow F. Using these 'contiguous flow pairs', they define a 'shared nearest neighbour distance' between two flows p and q as the product of intersection sets of origins and destination respectively, normalised by k. The effect is that if the origins and destinations of flows p and q are identical the distance is 0, whereas if there are no shared origins and destinations the distance is 1. Fig. 4 demonstrates how the shared nearest neighbour distance works.

The clustering algorithm based on these definitions works by first sorting contiguous flow pairs in ascending distance, and defining each flow as a unique cluster. Next the algorithm calculates the distance between the two clusters that a pair of contiguous flows belongs to, merging them if the distance is less than 1. When calculating distance between clusters, calculating the mean distance of all flow pairs between clusters would be computationally prohibitive for large data-sets. Zhu and Guo instead propose calculating the distance between median flows of the clusters, defined as the flow between the closest origin point to the origin centroid and the closest destination point to the destination centroid found in the original point set. Flow clusters are mapped using the centroids of origins and destinations, and total flow volume. The overall algorithm complexity is $O(kmnlog(mn))$, where m

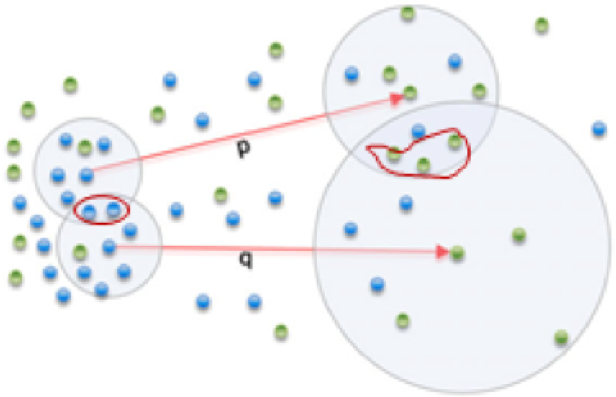


Fig. 4. Shared nearest neighbour distance, with intersection sets for origins and destinations highlighted, from [17].

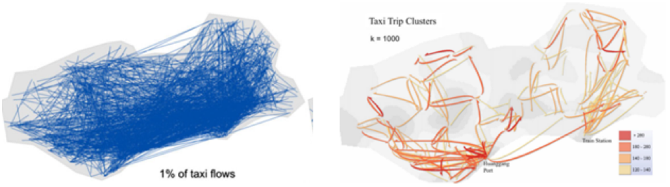


Fig. 5. ’Before’ and ’After’ application of hierarchical clustering, from [17].

is the average number of flow neighbours, n the number of flow clusters, and *k* the nearest-neighbour parameter. The flow map produced with this method enables a generalisation of all flows while preserving flow characteristics and patterns. In Zhu and Guo’s example of Shenzhen taxi trips, major locations (such as train stations) are identifiable, and major flows are visible, along with density of flows coded by colour. Fig. 5 shows the results of applying the method to the Shenzhen data-set. The visualization can be adjusted to show clusters of various sizes if more detail is required. The method can be configured with a different choice of distance metric, and the initial choice of *k*.

Zhu and Guo’s method succeeds in identifying significant hubs and flows, and can be used successfully with large data-sets. It validate itself by successfully highlighting known major hubs. It differentiates itself from other methods by defining clusters according to the data itself rather than from an arbitrary split (such as with pre-defined administrative areas or some forms of partition-based clustering), and is both computationally-efficient and memory-efficient, unlike density-based clustering. However, multi-variate analysis, such as an analysis of temporal patterns within the flows, can only be done by comparing multiple flow maps. One area of exploration would be of a suitable spatial-temporal distance metric. Alternatively, one could build an ’agglomerated flow map time cube’, although space-time cubes have their own readability issues.

2.4.2 Density-based and Partition-based Clustering

Two major types of clustering are partition-based algorithms and density-based algorithms. Both types of clustering algorithm handle regions of high- and low-density differently, as displayed in Fig. 6. Analysing data with a density-based algorithm would be more appropriate for identifying hubs than a partition-based algorithm but there are often too many dense data points to achieve this. Kumar et al. [10], suggest an innovative method of pre-clustering, sampling and then clustering the data once more to overcome computational limits and to aid with the initialisation of density-based clustering algorithms. By using the clusiVAT sampling method, an improvement of the Visual Assessment of Cluster Tendency (VAT) algorithm, it partitions the data quickly, without the need for initialisation, and suggests a number of clusters by using comparably less computational memory than an initial Density Based Spatial Clustering of Applications with Noise (DBSCAN) clustering. It then takes a random sample from each cluster, reducing the number of data points but ensuring the geometry of the dataset is retained. This sample set can then be used with a density-based clustering algorithm, in this particular case DBSCAN. All points not contained within the sample are then ‘assigned to the cluster in which their nearest sampled trip belongs’ [10].

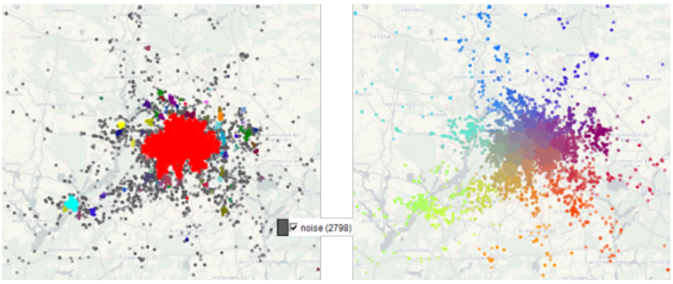


Fig. 6. Partition-based (left) and Density-based (right) clustering of a varied density dataset, from [4].

Although the analytical questions of the paper were to provide an insight into Singapore’s mobility patterns, urban hotspots, general crowd movement and to help with identifying the best places for taxis to pick up their next customer, these ‘hub-identification’ methods can be generalised. The techniques and methods stated could be used to cluster other data within the transport domain such as smart-card OD data. Outside transport, the principle of partitioning and sampling dense data prior to density-based clustering can be used with other OD datasets such as circular migration within London’s metropolitan boroughs.

3 Identifying Trends

A major group of analytical questions concerns the identification of spatial and temporal trends. In this section we will explore methods to identify trends: in direction and length of flows, in spatial distribution of flows, in temporal patterns, to to answer specific analytical questions related to our two main domains, transport and migration. In the transport domain these trends would be of interest to multiple parties. For instance, transport and infrastructure planners would be interested in finding areas which are under-served by the existing network. Network users would be interested in finding the least busy route, or the best housing locations for ease of commuting. In this section we discuss the various methods used to answer these types of questions, and their relative advantages and disadvantages.

3.1 Direction and Length

In the transport/commuting domain it is often of interest to planners to analyse the direction and length of trips. One could, for example think of an urban development planner investigating the major direction of cycling hire trips undertaken from various hubs to evaluate the demand for additional cycling lanes.

In Sect. 2.1 we introduced flow diagrams as a useful technique to identify major hubs in spatial situations and hinted at their ability to aid the discovery of major trends in terms of direction and length of trips. This feature is best illustrated by looking at an example application to the London Bike Hire dataset. In this case three length intervals, short (0.2km, red colour code), medium (2km, yellow colour code) and long (5 km,blue colour code) and eight directions classes equivalent to the eight principal compass directions have been proposed by Adrienko et al. [1]. The upper part of Fig. 1 illustrates the out-flows, the lower part shows the in-flows for bicycle trips for a chosen time cluster for all nodes. The analyst can identify the main trends in terms of direction and length by looking at the magnitude of each location’s spokes, their colour and their direction. In the cycling hire example, there is a clear trend of outflows towards the City of London from major hubs such as Waterloo and Kings Cross, with the majority of them being of medium length. We find that the true benefit of flow diagrams therefore lies in their ability to address the three discussed analytical questions, identification of major hubs, trends in directions and lengths of flows, simultaneously. As this technique has been introduced only recently, it has so far only been applied on journey-related OD-data but we could imagine that it will also find applications in other popular domains such as migration. It should also be noted that this technique can be amended with interactive elements, such as the display of flows links between certain OD pairs on demand.

One of the drawbacks of this approach is its sensitivity to the size of the aggregation classes that have been defined. The classes have to be selected by the analyst, either according to his domain knowledge or based on the statistical distribution [1]. Furthermore this kind of discretization can lead to classifying inherently similar flows into different classes or inherently different flows into the same class, depending on the definition of the breaks.

3.2 Spatial Distribution

One of the main analytical questions in the migration domain is to identify the spatial distribution of migration flows and areas of attraction and how to identify similarities or differences in regards to a selected attribute. In this section we present three techniques for answering these questions: Map^2, OD Maps and bi-partite maps.

3.2.1 $Maps^2$

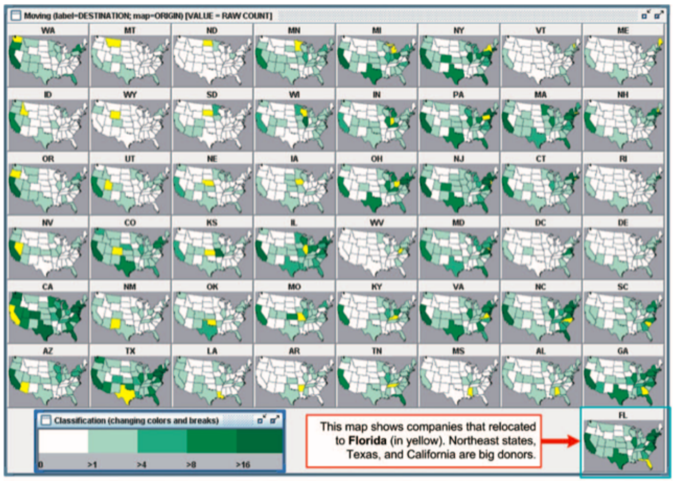
Depending on the time horizon and the size of the dataset, migration flow maps within or between countries tend to become cluttered very rapidly. To avoid the issue of occlusion and intersecting flows recent research suggests to use "map of maps" approaches which completely refrain from displaying the links between locations. Two often cited and used techniques are Map^2 by Guo et al. [5] and OD-Maps by Wood et al. [16].

A Map^2 is made out of multiple smaller nested maps of the same size, ordered in a matrix according to their geographical location. Each nested map highlights the destination in a distinct colour, and shades the origins according to the magnitude of the flows to the destinations. Guo et al. apply this to a dataset which contains the relocation movements of companies within the United States as displayed in Fig. 7. The overall map provides a comprehensive overview of the relocation activities. It is for example straightforward to identify regions with minimal inflows, like Montana and North Dakota in the Great Plain region, and high inflows, like in New York, Massachusetts and Pennsylvania.

Depending on the kind of OD-data, this method might require some application of computational methods, such as spatial clustering to avoid a cluttered matrix with too many nested maps which would be difficult to read and interpret. A common form of aggregation applied in Map^2 is the use of pre-existing regions. This aggregation requirement is at the same time one of the limitations of this technique, as local patterns within states or patterns for group of states might be missed [5]. Another disadvantage of Map^2 is that small areas are hard to identify without zooming in. In the case of company relocation, it is for example very difficult to infer the number of companies relocating from Washington DC, the smallest state in terms of geographical area, to any other destination state.

3.2.2 OD Maps

Wood et al.'s [16] OD Maps takes a similar "map of maps" approach as seen in Map^2. Instead of using pre-defined geographical areas as the base of the inner maps and then ordering them according to their geographic location in a matrix, the OD Map divides the geographical space into a 2-dimensional grid that has the same number of grid cells for both the nested maps and the base map. Another way of interpreting OD Maps is to see them as a special kind of OD-Matrix, whereby each nested map is a geographically reordered representation of a row in an OD-Matrix [14]. An example of an OD Map is shown in Fig. 8, which uses county-to-county migration data in the United States between 1995 and 2000 and utilises a Brewer colour scheme [16] to illustrate the absolute number of movements. The 'home' county is located in the base map according to it's geographical location and contains a nested map with the same number and ordering of cells as the base map. As with Map^2, this kind of visualization makes it relatively easy to grasp the spatial relationships between the defined areas and to identify the spatial distribution. For example the dark red cells, indicating high OD density, are mostly distributed around the home counties, indicating that people are more likely to migrate to states that are close to their home state [16].



Moving (label=DESTINATION; map=ORIGIN) [VALUE = RAW COUNT]

This map shows companies that relocated to **Florida** (in yellow). Northeast states, Texas, and California are big donors.

Fig. 7. Map^2 showing the movement of companies in the US. Origins are highlighted in yellow, the magnitude of flows from the origin states in shades of green, from [5].

OD Maps have also been used to analyse the county-to-county migration within Ireland where they were found to be an effective method for the identification of spatial distributions and patterns, for example, how popular certain destination counties are for origins and the degree of spatial autocorrelation of migration flows [14]. Additionally OD Maps can help identify similarities or differences in regards to selected attributes. In this case maps are plotted next to each other, each with a different attribute value. In the Ireland migration example Slingsby et al. [14] plotted two OD Maps next to each other, one with migration flows with the gender attribute male, and one with the gender attribute female and by doing so revealed interesting differences in the distribution of migration flows between women and men. This works especially well for binary attributes, as the number of maps that have to be plotted next to each other is limited to two. One drawback of OD Maps is the importance of the appropriate grid-size which has to be defined by the user by dynamically adjusting the grid parameters. Depending on the grid size, the resulting aggregation can also lead to aliasing effects - a form of the Modifiable Area Unit Problem (MAUP) [16], which is explained in more detail in Sect. 4.2. Additionally OD Maps reach their limit when geographical areas of interest have very elongated aspect ratios, such as Chile, which results in even further elongated nested cells that are difficult to read and interpret [16].

3.2.3 Graph Theory and Bi-Partite Graphs

A significant area of research is the viability of using new 'big data' sources in place of more traditional data-gathering techniques. One such area is the use of mobile phone records to estimate travel habits. Call details records (CDRs) contain time-stamped co-ordinates of phone users. Origin-destination data can be generated from this data, and used to help answer analytical questions related to travel. For example, can OD data generated from CDRs be used to identify 'geographic accessibility' i.e. to see whether trips between certain areas are limited to a particular socio-economic group. Another analytical question to identify spatial trends asks: are there some areas which have a wider spatial distribution of flows, and is the trend linked to socio-economics? To answer this question, Toole [15] proposes splitting a single OD flow into individual flows for each road/transport connection between the relevant origin and destination. This is still considered OD data, as

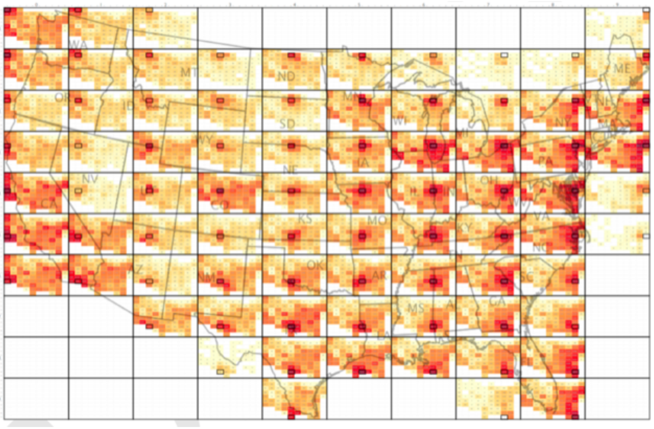


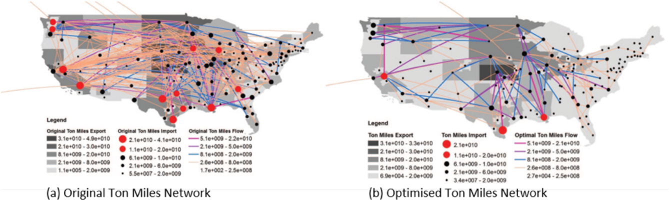
Fig. 8. OD-Map US County to County Migration, showing destination densities, from [16].

the trajectory of the routes remains unimportant. A bi-partite graph is then constructed between road segments and driver 'sources' i.e. locations within the city. We calculate the degree of each node and an edge 'betweenness' metric, measuring the importance of a road by the 'number of shortest paths between any two locations which pass through the edge.' These two analytical metrics can be used to evaluate the relative importance of a city's road network, and visualisations can be constructed to identify patterns based on these metrics. To analyse patterns of road usage based on socio-economic groups, one can colour roads on a map based on their betweenness/degree classification, and can code socio-economic attributes into the visualisation. One would be able to identify whether certain areas produce more traffic (trip volume) than others, which routes form the predominant portion of this traffic, and also where there are differences between users of various types. Of course, one should be aware of potential biases in methods such as these. As the source of the analyses is CDRs, an awareness of their limitations is prudent. Firstly, CDRs can only approximate a user's location, as network connections can be intermittent. Furthermore, by their nature, CDRs are skewed towards heavy mobile phone users. An analysis of socio-economic or age-related road usage would have to account for this, potentially by an appropriate weighting factor, taking care that the choice of weighting factor does not introduce additional bias. Despite this, travel estimates produced by CDRs are regularly validated in comparison to traditional travel surveys [2], and have also been used to infer trip purpose, by categorising origins and destinations as 'home', or 'work' based on CDR time-stamps. The measure of road importance can be used to answer other analytical questions. For example, a transport planner could decide where to place an alternative transport route by identifying the sources of traffic along a certain route, and considering the relative importance of this route to others. Companies may wish to look at the data to decide whether opening a new retail outlet would be supplied with enough traffic to be viable.

Toole [15] accepts that, like most aggregation methods, the method is subject to the Modifiable Area Unit Problem (discussed Sect. 4.2). He sees potential in extending the method with other data sources such as bike-sharing, public transport, or even water transportation networks.

While Toole used bi-partite graphs in the transport domain to analyse road networks Robinson et al. [12], have used properties of bi-partite graphs to optimise food flows in the United States. The main aim of this study was to optimise the US food network in terms of sustainability and economic efficiency whilst retaining resilience to uncontrollable effects to parts of the supply chain.

Using data provided by the commodity flow survey (CFS) and published by the US Census Bureau, Robinson et al. [12] selected five food categories and translated all related shipping records into a bi-partite graph, illustrated in Fig. 9, with two sets of nodes (origin and destination) with two sets of in- and out-flows corresponding to the weight of shipment to and from the node and the mileage for which the shipment has travelled.



(a) Original Ton Miles Network

(b) Optimised Ton Miles Network

Fig. 9. Comparison of the original and optimised networks in terms of food mile flows, from [12].

The goal was then to maintain each nodes import and exports on a local level whilst optimising the flows, in this case the food mileage and weight, on a global level. The optimisation was dealt with using linear programming and solved by using off-the-shelf solvers [12]. It was identified that an optimised network was subject to potential node failure and could be less able to adapt to future demands. It was therefore suggested that a partially optimised network would be a more appropriate model.

A variety of visual displays have been used to communicate the data. The colour luminance of the CFS areas indicates the export (in terms of weight) of each region, the size of each node indicates the weight of imports for each region and the amount of food product (in weight) from an origin to a destination is indicated by coloured flow lines.

These methods can be generalised in order to reduce or redirect flows in other contexts. For example, we could consider supply chains of different types of products or pieces from various factories, required to manufacture a product. In terms of other domains; transport could use network optimisation but this would most likely involve analysing sets of OD pairs (i.e. a trajectory).

3.3 Temporal Patterns

Besides questions addressing the spatial dimension of a dataset, temporal analytical questions often arise. Common analytical questions that were identified in the literature are: Are there periodic temporal patterns? Are there temporal outliers? Are there temporal trends? Are there differences or similarities between two time intervals?

3.3.1 Time Arranger

Andrienko et al. [1] propose an interactive and visually supported clustering method to address such questions. This approach combines visualisation methods in the form of a 'Time Arranger' and colour planes, with partition based clustering such as k-means [1]. The Time Arranger represents chronologically ordered intervals in the form of blocks that are arranged in rows, where the row length describes a relevant time cycle. The time intervals and time cycle have to be selected based on either the domain knowledge and the questions that the analyst is trying to answer, or based on a histogram analysis of the distribution. In the London Cycling Hire dataset case, mentioned in Sect. 3.1, each row represents one week and each day is divided into four time-interval blocks as depicted in Fig. 10, resulting in a total of 28 blocks per row.

To assign colours to the identified clusters the cluster centres are projected onto a two-dimensional continuous colour plane. Patterns in the Time Arranger can then be detected in three ways:

1. Periodic patterns are revealed if blocks are vertically aligned in the same colour.
2. Trends are revealed by gradually shifting colour between blocks.
3. Outliers are revealed by high blocks with a contrasting colour in comparison to neighbouring blocks.

Adrienko et al. [1] propose the following interactive method to detect temporal patterns, trends and outliers. During each of the steps described below, the analyst has to check for periodic patterns, trends and outliers and stop the process if no additional patterns are detected.

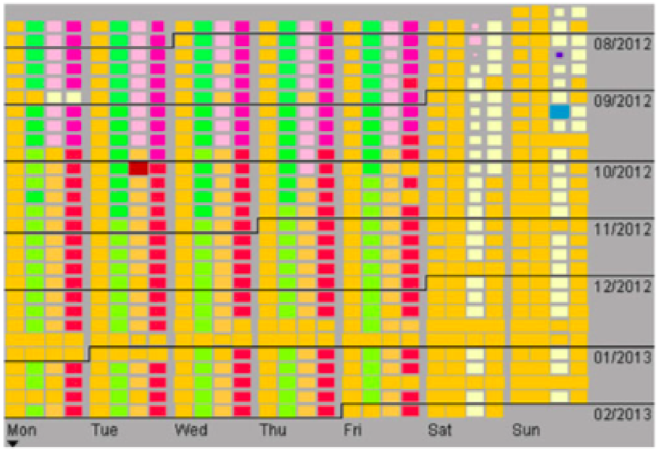


Fig. 10. Time Arranger for the London Bicycle Hire Data set, configured with a weekly time period and four time interval blocks per day, from [1].

1. Select a partition-based clustering method, such as k-means, and start with a small initial number of *k* clusters.

2. Evaluate the clusters based on their distances to the centre, which can be depicted as the block size in a Time Arranger View or on a colour plane.

3. If the distances between the cluster centres are large , increase *k*. If the distances between the cluster centres are small, decrease *k* to the previous value.

4. If necessary apply progressive clustering of a subset of clusters that need adjustment.

We found that the Time Arranger can be a useful tool to highlight temporal patterns which can then be further analysed in combination with other visualisation techniques. This relies however on the analyst's judgement of the quality of the separated clusters.

3.3.2 Multiple Maps

In cases where the focus is on the differences or similarities in patterns between two selected time intervals, a multiple map approach is often taken whereby two maps are plotted next to each other. As mentioned in Sect. 3.2.2, OD Maps can be used to compare differences in regards to selected attributes and can therefore also be utilized to compare different time periods. For example Slingsby et al. [14] found that in Ireland migration to Antrim from close by counties, except Down, was higher in 1911 compared to 1851. This can work well for simple visual analysis of differences, however if the number of relevant time periods is high, a large number of maps must be displayed, reducing the readability of the maps. If multiple time periods are of interest, the 'time-arranger' technique is preferable for answering temporal analytical questions.

3.4 Density of Flows

In a study by Scheepens et al. [13] into interactive density maps for moving objects the author asks the questions - 'How do we find the right filters and how do we display the density flow maps together as one?' Guo and Zhu [6] tackle the problem of how to analyse these different fields together in one single visualisation.

A feature of the density-normalising technique used by Guo and Zhu is that duplicate flows are not represented by the visualised maps. To do this only the 'smoothed flows' between non-overlapping geographic areas were calculated. In addition, flows which are too close to each other are not considered. Using parameters to set the minimum movement within a defined space and the desired number of total flows is effective to cluster the OD movements. For example, the top 100 flows over a minimum distance of 500km can be selected for the national level.

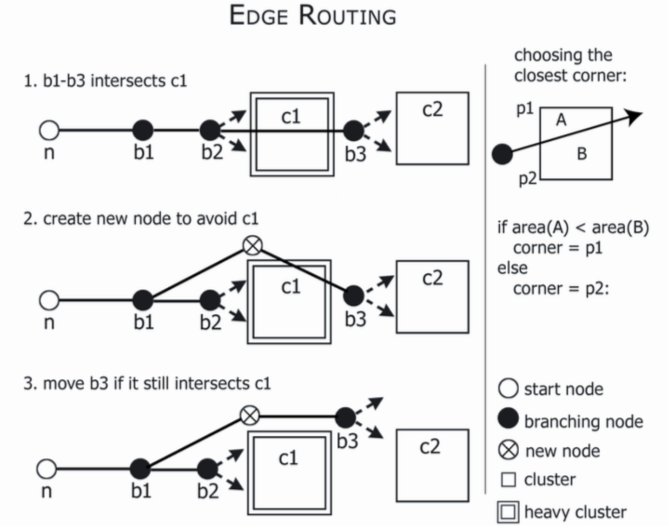


Fig. 11. Diagram showing implementation method of the edge routing technique from [11].

Another approach is representing movements of flows by adjusting the size of the flow arrows based on the magnitude of the movement or migration pattern. The method used is an effective tool in allowing us the ability to locate heavy flow patterns efficiently. However, in aggregating heavy flows over larger distances an analyst may miss out on heavy patterns at a local level.

3.5 Edge Routing

To answer the question of how we can analyse high density patterns in flow data Guo and Zhu [6] propose 'edge routing' or 'edge bundling'. These methods are acknowledged as being effective methods in improving the visual representation of flow data. A study by Phan et al. [11] details how an edge routing computation is implemented. In this paper the author illustrates this principle by explaining that 'the edges are routed around the bounding boxes within the same hierarchical cluster'. The main principle of edge routing is illustrated by Fig. 11.

In Fig. 11 the author explains that the edge between the branching points $b1$ and $b3$ intersects a heavily clustered point, $c1$. To avoid cluttering the visual display a new 'boundary box' is firstly created around $c1$. An additional node is subsequently added to avoid this boundary box, the position of which is decided by finding the closest corner of the '$c1$ boundary box' to the line between $b1$ and $b3$. The edge is then routed to the new node and carries on to join up with $b3$. If the line between the new node and $b3$ still intersects the $c1$ boundary box, the position of $b3$ is again updated.

3.6 Edge Bundling

In the research carried out by Guo and Zhu [6] the method of 'edge bundling' is often used in tandem with 'edge routing'. These two methods are considered to deliver something similar in terms of an 'end product' when it comes to analysing high density patterns within flow maps. A visualisation by Holten [7] gives us an appreciation of how edge bundling and edge routing methods differ. The author describes the process of edge bundling, which begins by defining a path an edge will take from one node to another. The adjacent edges are bent towards a pre-defined path, by spline curves in this case. Holten uses edge bundling with the aim to link and organise a software system. The author compares edge bundling to 'merging electrical wires or network cables together to make an otherwise tangled web of wires and cables more manageable'.

Both the edge routing and edge bundling techniques result in a clear display of high density movements, but Guo and Zhu emphasise that there are limitations to the methods and states that, 'whilst visual clarity is greatly improved the limitation of these approaches is that it is difficult to perceive the actual connection between specific pairs of origins and destinations'. A way of overcoming this issue is suggested by the author which involves assigning each edge with a unique colour to make them identifiable.

4 OD Issues

We have discussed the major types of analytical questions asked of OD data, the methods used to answer these questions, and their relative advantages and disadvantages. It would be remiss to discuss OD data and OD visualisations without a brief discussion of some significant issues with analysing large sets of OD data, namely the 'clutter-occlusion problem' and the 'MAUP problem'. We also briefly discuss the issues involved in collecting and generating OD data.

4.1 Clutter

Although a problem for many types of data visualisations, OD direct flow maps are particularly susceptible to clutter, as the flow data is often overlaid on a geographic map, and in the case of a large number of flows the map fast becomes cluttered. A variety of methods exist to solve the problem, broadly categorised as 'avoid clutter' and 'mitigate clutter'. Clustering-aggregating methods are a form of avoiding clutter, grouping similar flows/origins/destinations and reducing the number of flows visualised. Mitigation methods include edge re-routing in order to minimise edge intersections, and other design principles aimed at improving readability of flow maps, as shown by Jenny et al. [8], who demonstrates that (for example) curved flow lines are preferable to straight flow lines, arrows are preferable to a tapered width flow, that quantities are best represented with line width or colour brightness, and that if flows must intersect, the greater the angle of intersection the better.

4.2 MAUP

The research of Guo and Zhu [6] looks at issues when representing different aggregations of data. In this research the author looks at migration patterns which reflect areas of various sizes in terms of population. From the research conducted we acknowledge that the common approach would be to aggregate these locations into subsets. However we are confronted in this case with an issue known as the modifiable areal unit problem (MAUP), where results are highly dependent on the choice of aggregation unit. To address the issue of the MAUP and density-normalising of aggregations of data, Guo and Zhu implement a calculation to their algorithm to consider flow density. A 'smoothed flow value' from the origin to destination is calculated by normalising the magnitude of the values. In this specific case the density was normalised by considering the size of the populations within the geographic units in the data. However, there are limitations to such a simple technique. The root cause of MAUP is excessive aggregation which as explained 'results in a severe loss of spatial resolution'. It is also noted that there is a risk that 'over-aggregating' results in major patterns being overlooked, for example, large flow patterns at a more local scale.

5 Conclusion

We have looked at several techniques used to answer analytical questions in origin-destination data. Broadly, these are either methods to identify hubs, or to identify trends such as temporal or spatial. We have discussed the relative merits of these methods, in terms of ease of use, interpretability, computational efficiency, how well the analytical question is answered, and how well each method copes with clutter and the modifiable areal unit problem. We have also highlighted the importance of combining computational methods with visual techniques when analysing movement data.

There is scope for further research in several key areas. In many cases a method is proposed, and the author requests for further research in avoiding the modifiable area unit problem. A method's efficiency in coping with large data-sets can always be improved, whether by algorithmic improvements or sensible approximations. More consideration might also be given to user interaction, and the desired end-user. For example, the results shown to a taxi driver may depend on an analyst's choice of parameter. If an interactive tool were available, a taxi driver or domain expert may choose differently. Similarly in migration research we saw that patterns were not visually represented unless they traversed a distinct boundary - so whilst this type of analysis may be beneficial to a government official it may not have the same value to a real estate company.

References

[1] G. Andrienko, N. Andrienko, G. Fuchs, and J. Wood. Revealing Patterns and Trends of Mass Mobility Through Spatial and Temporal Abstraction of Origin-Destination Movement Data. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2120–2136, 2017. doi: 10.1109/TVCG.2016.2616404

[2] M. G. Demissie, F. Antunes, C. Bento, S. Phithakkitnukoon, and T. Sukhvibul. Inferring origin-destination flows using mobile phone data: A case study of Senegal. In *2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pp. 1–6, June 2016. doi: 10.1109/ECTICon.2016.7561328

[3] R. S. Fanhas and G. A. P. Saptawati. Discovering frequent origin-destination flow from taxi GPS data. In *2016 International Conference on Data and Software Engineering (ICoDSE)*, pp. 1–6, Oct. 2016. doi: 10.1109/ICODSE.2016.7936153

[4] G. Andrienko and N. Andrienko. Use of Density Based Clustering, INM433 Lecture, Nov. 2017.

[5] D. Guo, J. Chen, A. M. MacEachren, and K. Liao. A Visualization System for Space-Time and Multivariate Patterns (VIS-STAMP). *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1461–1474, Nov. 2006. doi: 10.1109/TVCG.2006.84

[6] D. Guo and X. Zhu. Origin-Destination Flow Data Smoothing and Mapping. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2043–2052, Dec. 2014. doi: 10.1109/TVCG.2014.2346271

[7] D. Holten. Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):741–748, Sept. 2006. doi: 10.1109/TVCG.2006.147

[8] B. Jenny, D. M. Stephen, I. Muehlenhaus, B. E. Marston, R. Sharma, E. Zhang, and H. Jenny. Design principles for origin-destination flow maps. *Cartography and Geographic Information Science*, 45(1):62–75, Jan. 2018. doi: 10.1080/15230406.2016.1262280

[9] S. Kim, S. Jeong, I. Woo, Y. Jang, R. Maciejewski, and D. Ebert. Data Flow Analysis and Visualization for Spatiotemporal Statistical Data without Trajectory Information. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2017. doi: 10.1109/TVCG.2017.2666146

[10] D. Kumar, H. Wu, Y. Lu, S. Krishnaswamy, and M. Palaniswami. Understanding Urban Mobility via Taxi Trip Clustering. In *2016 17th IEEE International Conference on Mobile Data Management (MDM)*, vol. 1, pp. 318–324, June 2016. doi: 10.1109/MDM.2016.54

[11] D. Phan, L. Xiao, R. Yeh, and P. Hanrahan. Flow map layout. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pp. 219–224, Oct. 2005. doi: 10.1109/INFVIS.2005.1532150

[12] C. Robinson, A. Shirazi, M. Liu, and B. Dilkina. Network optimization of food flows in the U.S. In *2016 IEEE International Conference on Big Data (Big Data)*, pp. 2190–2198, Dec. 2016. doi: 10.1109/BigData.2016.7840849

[13] R. Scheepens, N. Willems, H. v. d. Wetering, and J. v. Wijk. Interactive Density Maps for Moving Objects. *IEEE Computer Graphics and Applications*, 32(1):56–66, Jan. 2012. doi: 10.1109/MCG.2011.88

[14] A. Slingsby, M. Kelly, J. Dykes, and J. Wood. OD Maps for Studying Historical Internal Migration in Ireland. Seattle, Washington, US, 2012.

[15] J. L. Toole, S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. Gonzlez. The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, 58(Part B):162–177, Sept. 2015. doi: 10.1016/j.trc.2015.04.022

[16] J. Wood, J. Dykes, and A. Slingsby. Visualisation of Origins, Destinations and Flows with OD Maps. *Cartographic Journal, The*, 47(2):117–129, 2010. doi: 10.1179/000870410X12658023467367

[17] X. Zhu and D. Guo. Mapping Large Spatial Flow Data with Hierarchical Clustering. *Transactions in GIS*, 18(3):421–435, June 2014. doi: 10.1111/tgis.12100