Australasian Transport Research Forum 2018 Proceedings

30 October – 1 November, Darwin, Australia

Publication website: http://www.atrf.info

Adapting truck GPS data for freight metrics

Richard Green[1], David Mitchell[1]

[1]Department of Infrastructure, Regional Development and Cities

Email for correspondence: Richard.Green@infrastructure.gov.au

Abstract

The Bureau of Infrastructure, Transport and Regional Economics (BITRE) has developed a process to use telematics data from private firms to provide information about the Australian road freight network and industry to assist decision making by firms, government and the public. This paper outlines the intent of the project and describes the process in a non-technical way.

1. Introduction: Making sure we can get the best use of in vehicle telematics data

Many freight providers have adopted in-vehicle telematics to assist drivers and to monitor the location and activity of their fleet. However, collectively this data is also a rich potential source of information for the industry, government and the community at large. It can provide a cheaper, more detailed and more timely alternative to traditional data gathering means, including surveys and road monitoring equipment, as well as providing data on things that we currently have little formal knowledge about.

The BITRE has obtained a sample of telematics data from a number of firms with an understanding that they may provide more in the future as part of a larger and more representative data set. This paper describes, in a largely non-technical way, what we have learned so far about how we can make the best use of this data for potential users.

The number of things we could do with the data is immense, so we first decided to identify a small number of areas where users would gain the most from the data. This would ensure we developed a process that produced what users need the most.

1.1. Activity

Firstly, users need timely information about changes in freight patterns, including changes in activity overall, between regions, and in vehicle utilisation. Currently the only source of data is an irregular survey of freight patterns conducted by the ABS. As well as being irregular, this survey only provides information at some delay, and it costs a great deal both financially to government and in compliance costs by industry.

Without aggregate activity data, road freight firms are unable to tell whether changes in demand for their business are unique to them or reflect changes that affect everyone in the industry. This lack of knowledge reduces their ability to adapt to changing circumstances and meet the needs of customers.

Timely measures of freight are also a potential tool in monitoring macroeconomic activity. Freight flows reflect current economic activity as well as businesses' expectation of future demand, particularly in retail. As such timely information on how changes in freight flows may be a useful leading indicator in economic forecasting.

Neither of these needs necessarily requires an estimate of *total* freight activity at any time, rather they both require information on trends in freight activity. We can meet this need with an index that tracks activity relative to an arbitrarily chosen base period. Nonetheless statistical techniques, especially when combined with less frequent survey data, can allow reasonable aggregate estimates.

1.2. Road information

Telematics also can provide an open, consistent and detailed source of data on the road freight network. Such data can identify traffic patterns, and congested areas and times to help route planning and network investment.

The industry does have this kind of information, but usually it is informal, subjective and uncollated, for instance stored in drivers' heads and shared orally. This limits the ability for the whole industry, and government in particular to make use of it.

Where there is formal information it relies on traditional methods of traffic monitoring. These require fixed, physical infrastructure such as road sensors, pneumatic tubes and cameras. These are expensive to operate and are limited in scope. Furthermore, they require operators to make assumptions about which parts of the road network require monitoring. This may mean they may fail to capture important information, for instance where road users choose different routes to what governments are expecting.

Major proprietary telematics providers, including Google and HERE, already develop this sort of information based on the telematics data of their customers, but this data is not open and is expensive to access, meaning industry, government and community cannot make full use of this information. It is also based on a sample that is dominated by passenger vehicles. Freight vehicles, however, have different network usage patterns, behaviour, needs and impacts that require a dedicated source of data.

We therefore decided we would process telematics data to get information about the road network as used by road freight – such as the volume of vehicles and the speed at which they are travelling on different parts of the road network, and how this varies by time of day, week and year. Ideally this information would be made as open as possible whilst protecting the confidentiality of firms who have provided data – for instance by redacting information about local roads near their facilities. This might include a web-based visualisation and the publication of data that other users can analyse and match to an open road database. It may also inform the creation of a regular report like the American Transportation Research Institute's (ATRI) annual freight-significant congestion location report (ATRI 2015), which provides summary information on the top 100 most highly congested freight-significant locations in the U.S.

1.3. Stop information

Governments have invested in many rest area facilities on major routes to enable the regular rest periods necessary for the safe operation of vehicles. However, it has been difficult to evaluate how effective investment has been to date, for instance ensuring facilities that have been built are used; and recognising where drivers have preferred other locations.

As such we decided to produce information on the actual stop locations of vehicles, in particularly commonly used locations, that governments could use to improve their investment in rest area infrastructure.

2. Description of available data including idiosyncrasies

It is important to note that the telematics data was not necessarily collected for any of these purposes, and requires thoughtful processing to get the information users need. The data is effectively a long series of locations reported by a vehicle's GPS unit without other context. The raw data needs to be processed to transform this series of locations into trips, stops and the roads used by the vehicle.

The sample data provided to BITRE came in a variety of file formats. These included formats such as JSON, which is commonly used to stream individual points of data between devices and servers, and table based formats like CSV that are common in statistical analysis.

The data are a series of individual observations of vehicles, or "GPS pings". The frequency of pings differed from firm to firm, ranging from a few seconds to a few minutes, but most vehicles pinged about once per minute. Each firm had differing sets of variables assigned to each ping; for example, some reported events such as engine ignition or information such as altitude whilst others did not. Other variables were common, but expressed in different formats, such as in decimal degrees versus degree minutes second. Nonetheless, the data can be combined and standardised to produce a series of ping events with a limited set of common fields. These are sufficient to meet the needs described earlier. The fields are:

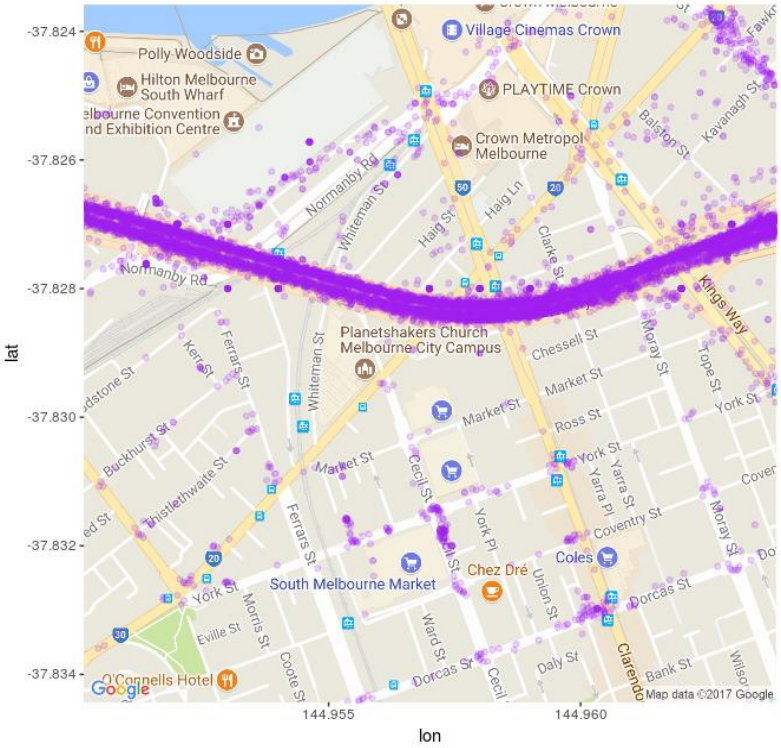
- A unique vehicle identifier, to tie pings together in trajectories
- The time and date of the ping (in Unix epoch seconds)
- Latitude and longitude of the ping
- The recorded speed at the time of the ping
- Azimuth (or bearing, heading, or direction) – present in most but not all firms, improves map matching accuracy.

Table 1: Example data

Vehicle	datetime	lat	lon	Speed	Azimuth
2de73989a3...	1462593757	-37.82637	144.9662	23	*NA*
2de73989a3...	1462593757	-37.82637	144.9662	0	*NA*
2de73989a3...	1462594017	-37.83027	145.0163	80	106
2de73989a3...	1462594137	-37.84092	145.0395	81	168
2de73989a3...	1462594257	-37.85738	145.0587	81	167
2de73989a3...	1462594377	-37.87682	145.0759	95	83
2de73989a3...	1462594497	-37.88625	145.1064	77	100
2de73989a3...	1462594617	-37.89355	145.138	83	98
2de73989a3...	1462594737	-37.91197	145.1651	99	136

All GPS co-ordinates have a degree of error. Generally, the error tolerance is within 7-15 metres, but environmental factors, including tall buildings, can increase it. Figure 1 shows data from the South Bank of Melbourne. Vehicles are clearly on the road network, but a large number of pings are shown just off the side of the motorway and surface streets. We must account for this when processing the data to identify the roads the vehicles used.

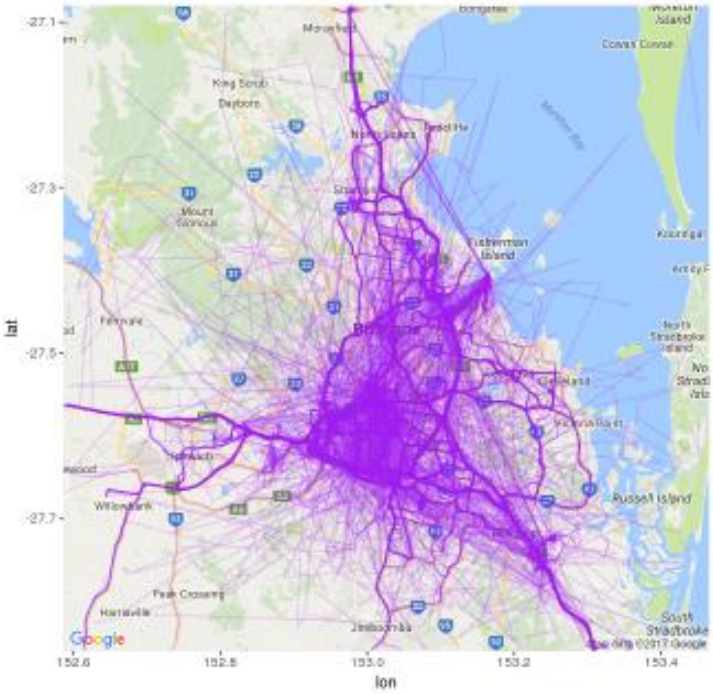
Figure 1: GPS error



There are also other, much larger, errors. For instance, many telematics devices, when they fail to determine a GPS location, will intentionally return impossible GPS co-ordinates. This happens frequently in tunnels where signals from GPS satellites are obscured. Other reported pings are clearly erroneous, for instance they are in the ocean or in areas far from any roads, and the reasons for this are unclear. Lastly, many pings would "echo" previous coordinates, replicating a location the vehicle had been earlier, but far from either the previous or the subsequent ping.

The following image shows data from a firm whose data had an unusually high level of erroneous pings. The lines each represent pings from a single vehicle, connected in chronological order. Whilst most pings are valid enough that major road ways used by the firm are clearly identifiable, it seems implausible these vehicles teleported into the mountains and seas, or jumped from one side of Brisbane to another.

Figure 2: Larger GPS errors



3. Processing

We need to process this data to identify stops and trips by a given vehicle, and to associate the data with the road network. This processing also needs to be consistent and scalable so that we can apply it to a growing pool of data without the need for human intervention. A process that requires us to supervise it constantly would be much more expensive and difficult to scale, and will also be inconsistent.

3.1 Trip/stop identification

The first task is to classify trips and stops. The obvious reasons are to identify the locations of stops, the regions traversed by trips and the time they take. The less obvious reason is to clean the data. Whilst erroneous pings that are outside Australia or are at sea are easily filtered out, removing locations on land, and those that "echo" valid locations, is more complicated. Removing them requires some assumptions about probable vehicle behaviour so an automatic process can remove improbable ones.

Lastly, trip grouping is required for effective map matching, which we will describe later.

We start by identifying stops, because they are much easier to identify than trips. This means that in any approach a "trip" is thus defined as the period between two stops. Trips can later be collated into longer tours as needed.

The most obvious way to identify stops based on the data above is to use the speed field to find sequences of pings where the vehicle is not moving and classify long such sequences as stops. However, in practice this proves unsatisfactory. Firstly, the provenance of the speed variable is unclear and is likely inconsistent across firms. In some cases – especially those when other information like engine ignition is also provided - it is probably the speedometer reading, but in others it seems to be imputed from the GPS locations and time stamps. This means a stationary vehicle can register a speed solely because random GPS error makes it appear it is moving. Speed is also a less important criterion for a stop than location. A truck

that spends time unhitching a trailer at one side of a depot then parks on the other side will register a positive speed whilst moving, but we would not deem that a trip.

The results of a speed based-stop identification are salvageable, but this requires a lot of work adjusting based on judgement calls; for instance changing the definition of "stationary" from "zero" to "near zero". This in turn risks classifying congestion on major roads as stops that we must then filter out. This is labour intensive, not scalable and hard to implement consistently.

Also, because it relies largely on a single variable rather than assumptions about how we expect vehicles to behave, it does not naturally lead to a mechanism to remove erroneous pings.

Another potential approach identifies likely stop locations beforehand and then attributes observations to them. This also is very labour intensive. Some locations are easy to anticipate, such as formal rest stops and firm depots, but these are still very numerous. Other locations, such as customers and informal rest areas—of which the latter is particularly important to policy makers—require a human examination of the data before and repeatedly during the processing. This is impractical to scale and prone to inconsistency. Furthermore, it means we necessarily become aware of information extraneous to government which might be sensitive. A prime example of this is when drivers take their vehicle home at night. An analyst on our end would have to examine the location of a private home to determine if it should be designated a stop location.

The alternative we pursued is based on spatial clustering and does not use the speed variable at all. The algorithms we developed are based on Cich et al (2015). This paper described a way of trip stop identification from smartphone GPS pings to assist the memory of people completing activity surveys, but the underlying logic is also applicable to freight vehicles.

A non-technical description of the basic logic of the algorithm is as follows.

The algorithm initially assumes the earliest ping for a vehicle is part of a stop. It then compares the next ping to it. If the next ping is within a given distance of the first ping, it attributes that ping to the same provisional stop and calculates the centroid (average location) of the stop pings. Then the next ping is compared to this centroid to check if it is also in the same vicinity.

This goes on until the vehicle moves beyond the given distance for determining stops. Then, if the time between the first and last ping in the stop exceeds a given minimum stop time, the algorithm records the stop and the preceding trip (if one was apparent). If not, it adds all the pings provisionally added to the stop to a new trip. If a trip was being recorded before the provisional stop event, the pings are combined with this trip.

In short, the algorithm records trips and stops in pairs only once it has identified the end of a stop.

Sub algorithms account for two problems caused by GPS measurement errors.

The first combines stops which were recorded separately but are effectively in the same location and are separated by only a short period. This can happen, for instance, when two successive GPS pings have strong natural errors in opposite directions and the vehicle has moved a short distance. The combined distance of the two errors and the actual movement can exceed the distance threshold and the algorithm will record two stops. The sub algorithm combines these stops into a single stop.

The second avoids determining a stop has ended when the GPS has returned a clearly erroneous measurement far from the prior stopped position. Such GPS errors are relatively common, as described earlier. The algorithm will discard pings where the minimum possible speed required to travel the distance (that is, in a straight line) is implausibly high.

For each vehicle this algorithm produces a list of stops, a list of trips and a list of residuals. Each stop and trip consists of a series of GPS pings. The residuals are pings that have not yet been assigned to either a stop or a trip. This is because the algorithm only records a trip and a stop at the end of a stop. Since the time period being analysed will always end whilst a trip or a stop is still underway, pings that remain unassigned are saved for the next period.

We then apply a "second pass" algorithm to "clean" erroneous GPS measures based on our assumptions about vehicle behaviour. For each trip, the second ping is compared to the first ping. If the implied speed required to travel the distance between the two in a straight line is implausibly high (for instance over 150 kph), the second ping is assumed to be the result of GPS error, and is removed, and the next ping then compared to the first. If the speed is plausible, the next ping is compared to the second instead. This continues for all pings in the trip, with each point compared to the most recent *valid* ping.

This automatically removes most instances of error, both where a GPS ping is clearly erroneous, for example when it is in wilderness or ocean, and where it "echoes" a valid point from earlier.

Figure 3: Trip grouping process from Cich et al (2015)

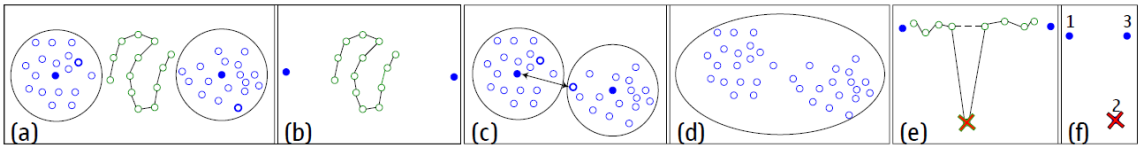


Fig. 1. (a) Stop detection with two stop clusters and one trip. (b) Stop detection with combined stop clusters. (c) Two stop clusters separated by a small distance. (d) A merge of the two stop clusters shown in (c). (e) Trip cleaning. (f) Stop cleaning.

The second path algorithm also separates trips into "subtrips" when the time or valid distance between pings is substantial. This can occur if there is no valid ping for a long period. We initially added this feature to aid us when visually evaluating how well the algorithms were removing erroneous pings, however, we found out it was useful for two other reasons. Firstly, it allows more careful analysis when one is wary about long periods attributed to a trip that we actually know very little about. Secondly, it became essential to the map matching process we used.

All of the parameters in this process, such as minimum stop time and stop distance threshold, can be stipulated by the user. When 19.5 million pings were processed using a minimum stop time of 1 hour and a stop distance of 100 metres, approximately 11 million were assigned to trips, and most of the remainder to stops. The shorter the stipulated stop time, the greater the proportion of pings are attributed to stops.

3.2 Map matching

To turn GPS pings into information about vehicle network use we need to "map match" each ping to a road database. This is more difficult than it appears. As mentioned, GPS pings have an inevitable range of error. This means any given ping may be off to the side of the road on which the vehicle is actually travelling. This is exacerbated by the way roads are stored as data. Most road data records a given road as a two dimensional line; that is a line without any width. As such, unless a GPS measurement records the exact middle of the road way (which

is implausible, especially if drivers are obeying the rules) there needs to be some process for matching the point to the road network.

A "road" as we commonly use the term is not specific enough to store as data, as it can refer to a small portion of a road or a continuous section that can continue for thousands of kilometres. As such, road data splits roads into segments, usually of a few hundred metres long, but often shorter in dense areas. A map matching process attributes a GPS ping to a segment in a set of road data.

3.2.1. Nearest neighbour map matching

The simplest way to map match is a "nearest neighbour" analysis. This takes each ping individually and finds the closest segment in the dataset, usually "correcting" the longitude and latitude to the nearest point on that segment in the process. This is not complex, although it can still be computationally intensive. However, there are substantial limitations of this approach.

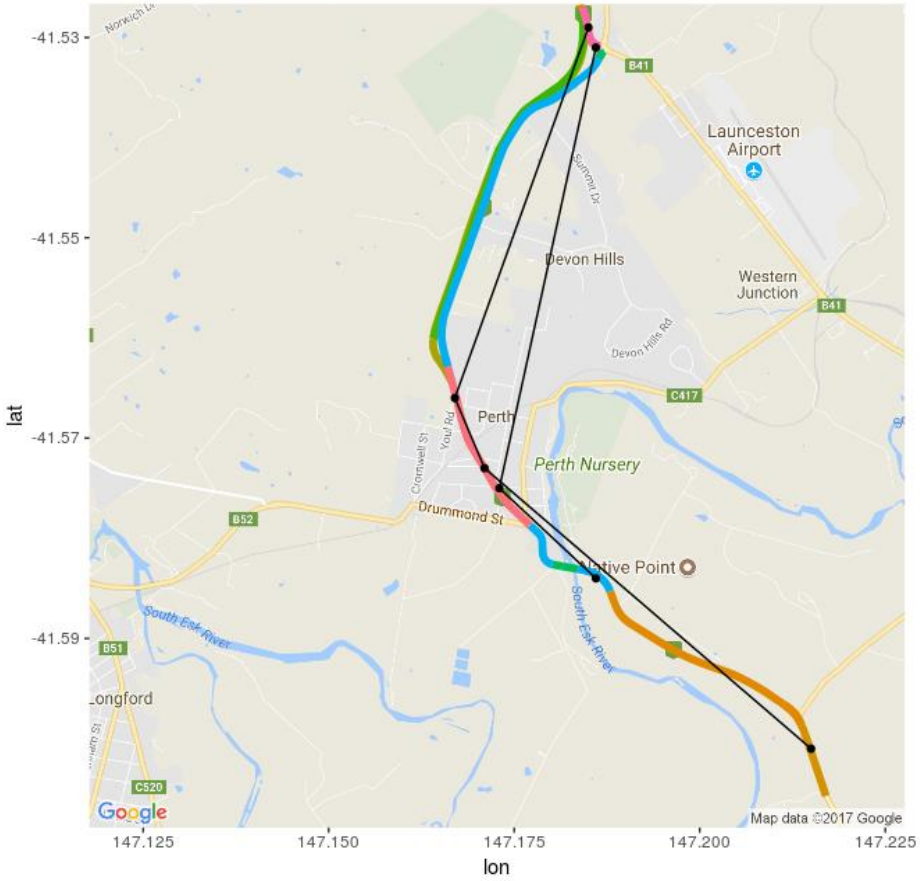
The first is a lack of accuracy. The inevitable GPS error means that the nearest segment to a ping is not always the segment on which the vehicle was travelling. This frequently happens, for instance, when vehicles pass through intersections; where streets are dense, and; where buildings cause larger GPS errors.

Secondly, it provides no information about a segment used by vehicles when GPS signals are unavailable. This occurs routinely in tunnels where direct lines of sight to satellites are obscured. Tunnels are often part of major freight routes and, whilst their operators already monitor traffic closely, it would be remiss to leave them out of this analysis.

The third is this method only induces information about a segment when a vehicle pings whilst on it. This reduces the amount of data available and introduces a bias towards slower vehicles. For instance, imagine a vehicle travelling down a road split into segments. The telematics in the vehicle are pinging at regular intervals, but the vehicle is travelling fast enough to pass over several segments between pings. This means the vehicle will ping on only some of the road segments it uses. If the vehicle is moving faster, it will use a larger number of segments between pings, and the overall proportion of the segments it pings on is even smaller. This means any given segment is more likely to have a slower vehicle recorded on it than a faster vehicle, and inferences based on this data will overstate data from slow vehicles. It also overstates the importance of vehicles that report pings at higher intervals.

Figure 4 illustrates how segments used may not be recorded. It showing two vehicle trajectories on the Midland Highway in Tasmania. Each dot is a GPS ping and they are connected with lines showing the trajectory of the vehicle. Each distinct road segment is coloured separately and overlaid on Google Maps imagery. It is apparent that both trajectories pass over the same set of segments, but each has only pinged on a subset of those segments.

Figure 4: Skipped segments



The limitations of nearest neighbor matching are well known, and geospatial developers have created alternatives. Their aim is usually to improve the accuracy of users' locations shown on in-vehicle displays, however, we have adapted their tools to our own ends.

3.2.2. Hidden Markov Chain map matching

The most common alternative method, and the one we have implemented, is known as Hidden Markov Map-matching (HMM). In short, HMM finds what it considers the most likely match for a ping based on both how far a segment is from the ping and also the probable route from the previous matched road segment. This means, for instance, a ping that appears to be on a local road passing near a freeway will instead be attributed to the freeway if the previous point was also on the freeway. This is because the vehicle is more likely have stayed on the same road than to have taken a circuitous route to get to the local road.

Apart from producing more accurate results, the need to find routes between points means that HMM can also report the other segments the vehicle most likely traversed between pings, including those in tunnels where no GPS was available. As such, inferences about the road network are not only more accurate, but provide ado….you have to ch greater depth of information and redress the bias towards slower vehicles.

We decided to adapt a framework called *Barefoot*[1] developed by BMW and released as open source. Barefoot was designed mainly to correct the displayed location of a vehicle on in-vehicle displays when receiving stream of data from the vehicle to a central server, but it also includes the ability to process batches of recorded data. It was not designed to return results

[1] Barefoot GitHub repository: https://github.com/bmwcarit/barefoot.

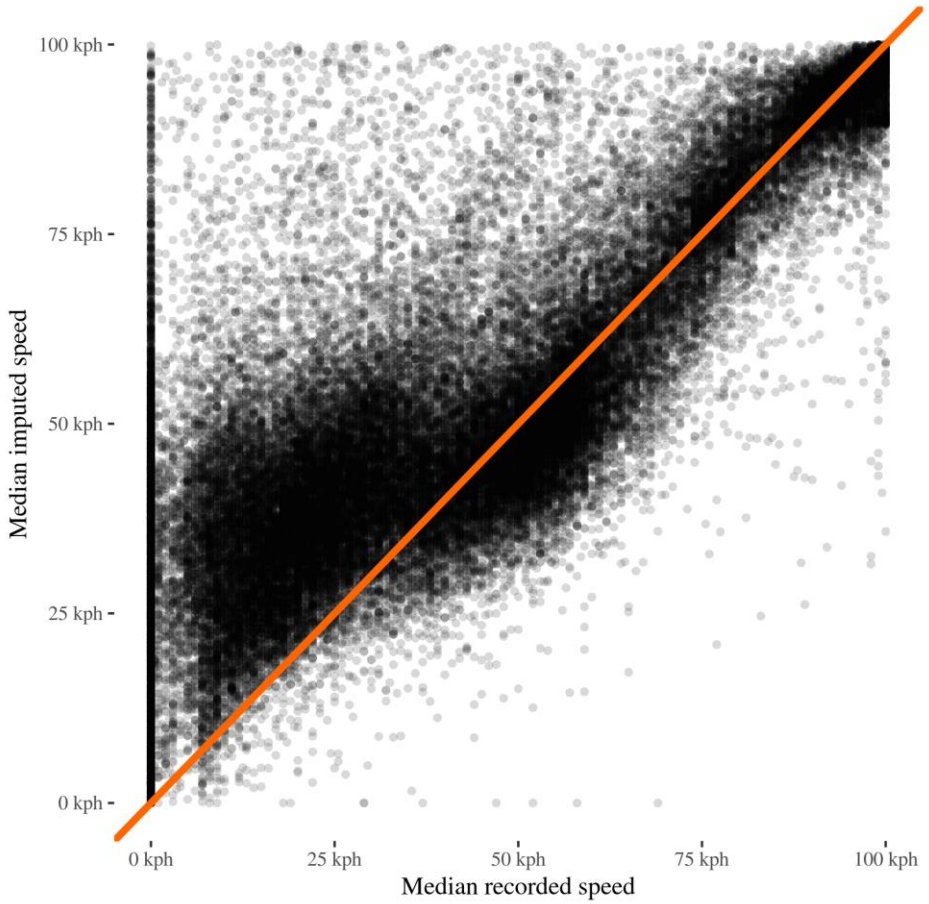
about the road network, but with the assistance of the authors we altered the code so it would do so. Barefoot is also able to use the azimuth variable, if available, to improve accuracy, but does not rely on it.

During this process is became apparent that the "subtrip" grouping created earlier was essential for Barefoot to operate correctly. This was because of checks in the Barefoot code designed to make it run faster. This highlights how adapting tools to new ends means solving unexpected problems.

For each ping provided to it, Barefoot will return a segment ID that it is on, a list of segment IDs used to travel to that ping from the last, and the distance taken on those segments. This also allows us to compute an average *imputed speed* for these intermediate segments if we divide this distance by the time between the two pings. In the initial sample data this meant the roughly 11 million trip pings, each of which with a recorded speed, led to nearly 70 million instances of segments being used, each of which with an imputed speed.

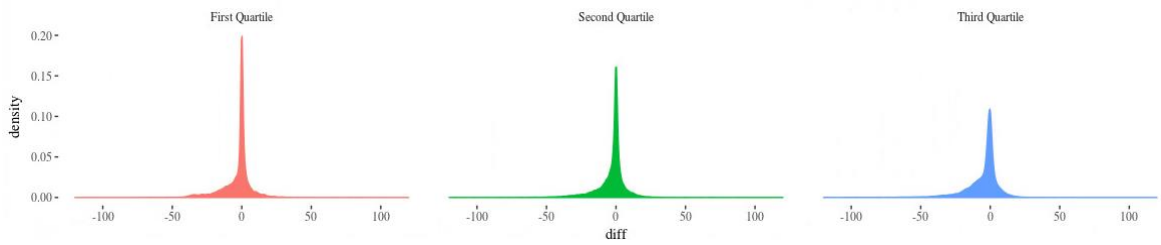
When we compared the *imputed speeds* with the *recorded speed* (when both are available for a segment) they are similar and the most likely difference is zero (see Figures 5 and 6). This suggests the imputation is quite accurate. However, the recorded speeds are more likely to be lower than the imputed speed . We anticipated this because of the bias towards slower vehicles discussed earlier. As more observations are recorded for road segments we anticipate this difference will decrease.

Figure 5: Recorded and imputed speeds (more than 30 observations)



Note the large number of segments with 0 recorded speed. These are largely small segments close to origins and destinations where vehicles are not moving, and which are also of little import to the broader network.

Figure 6: Distribution of the difference between recorded and imputed speeds (segments with more than 100 observations)



3.2.3 Map and road data

To map match we need road data to match to. There is no single authoritative dataset for Australian roads. In part, this is because "the road network" is a somewhat nebulous concept; are carpark entries, rest areas and service roads part of the network for instance? The network is also constantly changing and expanding based on the activities of the federal government, eight state and territory governments and many more local councils and private developers.

Geoscience Australia includes a road dataset as part of the TOPO 250k GIS dataset (GA 2006), but this is badly out of date and excludes recently built bypasses on major freight routes. Private sector firms such as Google and HERE have built proprietary datasets from a variety of sources but these are expensive and their terms potentially limit disseminate of the processed outputs to other licensed users.

Barefoot, by default, was designed to work with Open Street Map (OSM) data, which, as implied by the name, is open and free. We have chosen to keep this as the default so users can relate the data we produce with an open standard to perform their own analysis.

There are inevitable problems with using an 'open' platform map source. Just as the road network is constantly changing as roads are built and closed and altered, the data in OSM also changes as it is updated and corrected by users, who combine and split segments. The full implications of this will only become apparent over time.

The OSM data will not include every trip-related vehicle observation. For example, trips often begin and end in warehouses, yards and other off-road locations, and vehicles will also visit service stations and rest areas. They will record low speeds at these locations, yet this slow movement will be attributed to nearby roads, and these roads will then appear to be more congested than they really are. In most instances this will be in industrial estates and locations of interest only to one or two firms. Where a service station is on a major road, however, we may have to find ways of correcting this error, or rely on the discretion of users.

OSM data can also easily be altered as required. Regardless of the source, errors in the map data can distort the data we produce from map matching. For instance, if a motorway is not connected properly, barefoot will try to route vehicles via nearby surface roads. These will then appear to have absurdly fast average speeds. With OSM we can repair this connectivity easily, but such errors are not so easy to correct when they appear in other sources.

4. Intermediate data

This processing gives us intermediate outputs; the basic ingredients for providing information to inform users, policy makers and planners.

In the sample data, over half the pings were grouped into trips, the remainder were summarized into stops. At the very least this allows us to derive the start and end regions, and the time between them. We can also group trips into "tours" that are distinguished by longer

stop intervals between them, for instance the 8 hours that would be required for a driver to sleep adequately. We also have the opportunity for more detailed analysis using the pings and identified road data for each trip.

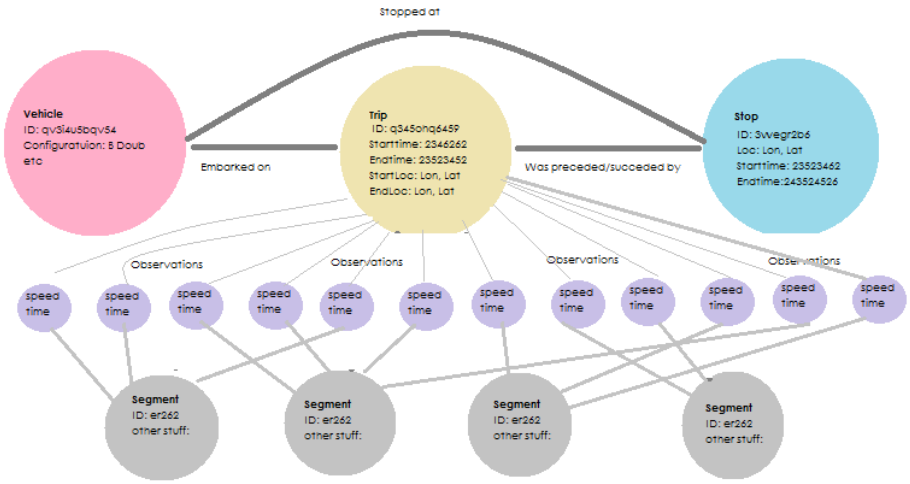
We also have details of a large number of stops, although this is limited to a unique vehicle identifier, the location, and start and end time of each stop. Since further analysis of stops is concerned mainly with rest behaviour we developed a method to exclude irrelevant and potentially sensitive information. We use a clustering algorithm to identify locations where there are many stops. We can then exclude isolated stops that are irrelevant, or locations that are unique to a firm or vehicle which may reveal commercially sensitive or private information.

The process also produces a very large number of road segment observations, each indicating a vehicle has passed through at a given time at an imputed (and sometimes recorded) speed.

5. Database

The intermediate data needs to be stored in a way that lets all the identified user needs, as well as other unforeseen needs, be met. It also needs to be scalable, to deal with rapidly increasing amounts of data as the project continues. We determined the most common form of database – the relational or SQL style – did not meet these criteria, and instead we chose a graph database model. In this type of database "nodes" (in this instance vehicles, trips, road segments etc) are connected to each other by relationships (a vehicle stops at a stop, a trip was observed on a road segment, etc). Figure 7 shows a simplified example of the graph database.

Figure 7: Example graph database



6 Uses

With the data stored in the database we can start calling upon it for analysis.

6.1. Activity indices

From the trip groupings we can create indices of truck activity. These can show changes in overall activity, but can also be interregional, for instance trips between statistical areas (SAs)

or cities. Whilst this data, when combined with survey data and using statistical techniques, can be used to estimate total activity, most industry users more immediately require an indicator of changes in activity. A simple index based on the unrepresentative sample is probably sufficient for this task. We can also produce a vehicle utilisation measure, aggregating the percentage of time in the analysis vehicles were actively on trips.

From this we can publish time series of activity, and also produce trip matrices showing times and volumes between all regions.

The final form of this analysis will respond to input from users.

6.2 Roads

The instances of road segment use allow us to derive information about the road network. For individual segments we can calculate average speeds and variation in speeds, as well as traffic volumes. We can also trace how this average changes by time of day, or day of the week and, in time, over the course of a year.

This can help show areas and times of congestion, for instance when speed and volumes are negatively correlated. It can also show actual experienced speeds to help with trip planning, for example when freight vehicles are slowed by steep inclines. Lastly, it also shows the routes vehicles choose to use, helping to inform planning and improve investment decisions.

Once the data is altered to maintain confidentiality by removing roads used only by one firm, these averages can be published as a simple table. This means users can combine it with OSM data for their own visualisation, network analysis or other purposes as they see fit.

6.3 Map widget

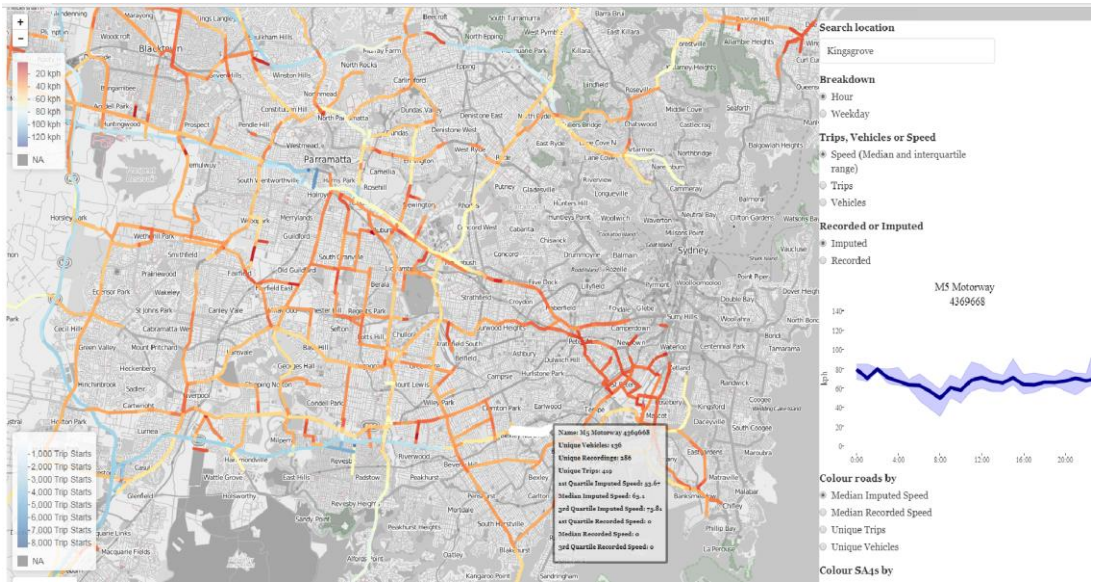
For the purposes of viewing the data, for quality assurance, we have developed an in-house interactive mapping widget to examine the road network, stop locations and interregional movements. It illustrates how the data could be promulgated in future. The map overlays the data we have developed in the project on base map images generated by OpenStreetMap, and users can drag and zoom to investigate areas of interest throughout Australia, or search for locations with a text box.

6.3.1 Roads

Figure 8 shows an example of road visualisation, displaying the Sydney metropolitan area. Users can choose to display roads coloured by imputed or recorded speed, or by volume of trips. Hovering the pointer over a road segment displays summary information, for example for the M5 tunnel in the picture. Clicking on the road allows users to graph speed or volume by time of day or day of week – summary information that is calculated at high speed within the database. This graph includes the 1st and 3rd quartiles for speed, showing the degree of variation about the average.

The summary information for the M5 segment shown in Figure 8, shows some ‘recorded speeds’ of zero, whilst half of the ‘imputed speeds’ are between 54 and 76 kph; speeds that seem much more reasonable. This illustrates one of the advantages of adopting a HMM process.

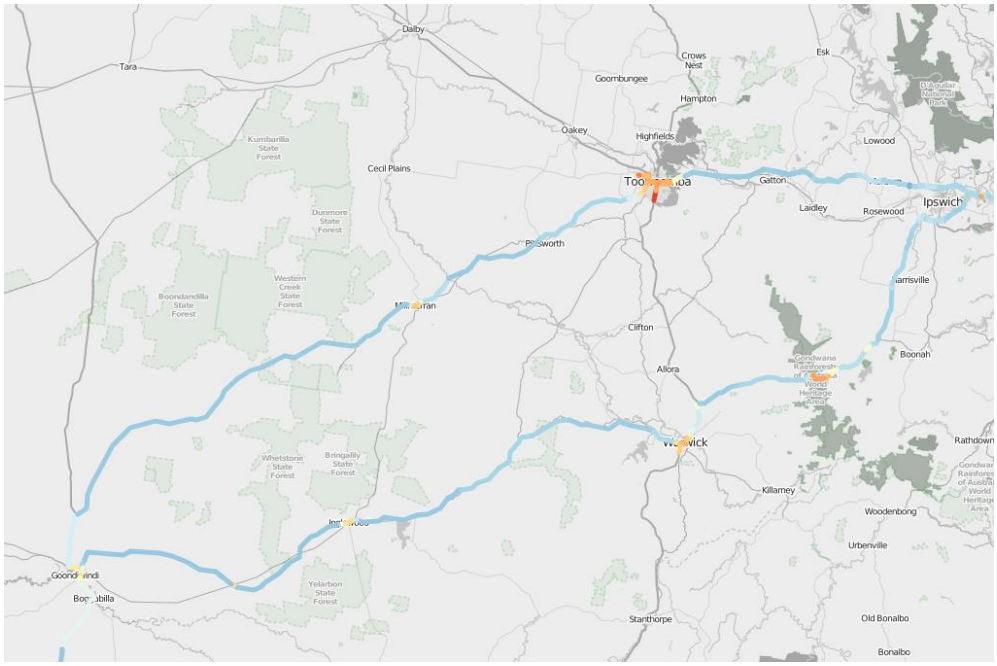
Figure 8: Map widget



Note: Only segments with more than 30 trips.

Figure 9 shows an example of how the data, when mapped, can show route choices by vehicles that might be unexpected by policymakers, even if they are well-known and make sense to industry. It shows that segment use South West of Toowoomba diverges, with roughly two thirds of the observed vehicles travelling via the Cunningham Highway, rather than taking a slightly shorter route via the Gore and Warrego Highways. The Cunningham Highway route also shows slower sections through the town of Warwick and whilst climbing a steep incline in the Gondwana Rainforest.

Figure 9: Route choice and speed near Toowoomba



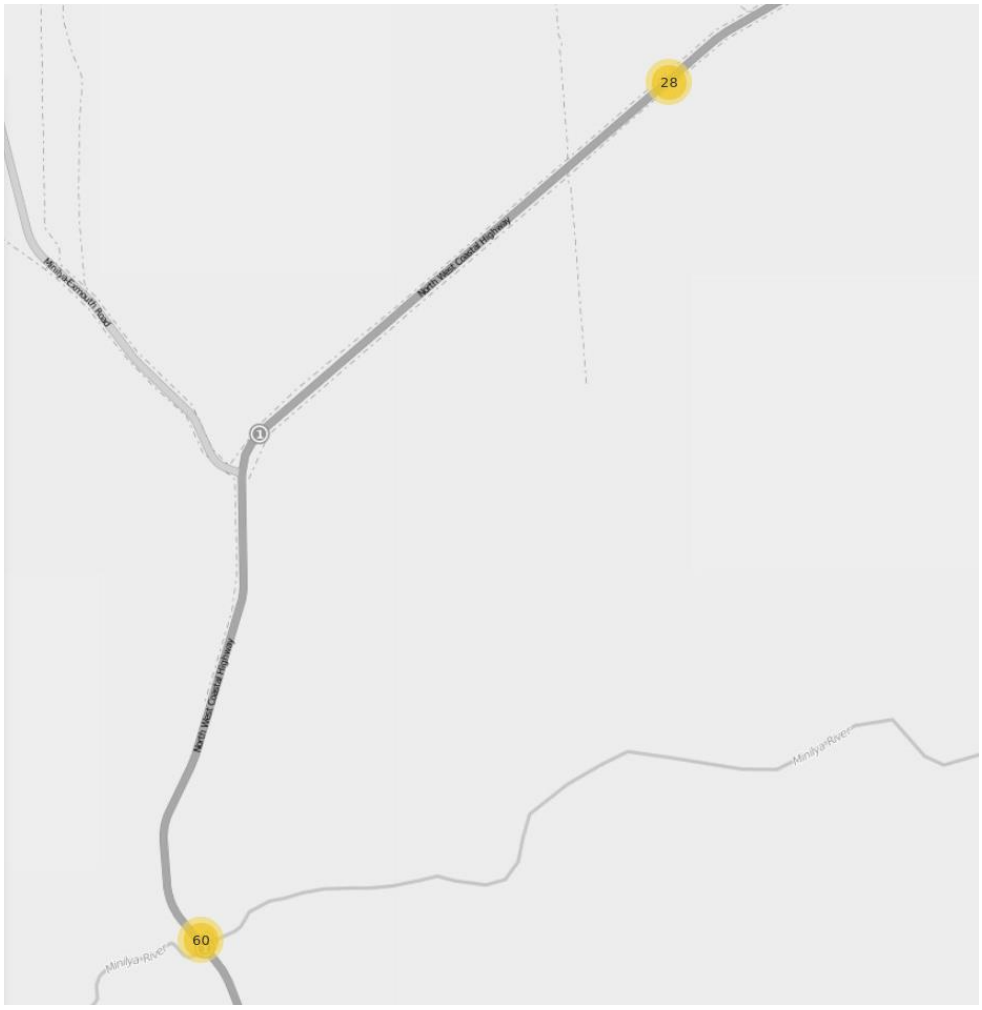
6.3.2 Stops

The map can also show stops at popular stop locations. The map automatically clusters stops at higher zoom levels, and disaggregates these clusters as users zoom in. As discussed, we can filter these locations to protect sensitive data by showing only clusters of stops with

multiple vehicles or firms. Users, in this case governments evaluating rest stops, can see some variation in use and even see individual information by clicking on stops.

Figure 10, shows the benefit of an unsupervised stop identification method, as opposed to a method that identifies locations in advance. For instance, in this image the lower cluster is the Minilya Roadhouse; a business on the North West Coast Highway in Western Australia offering fuel, food and accommodation. Freight vehicles in the sample data typically stop here for periods exceeding 8 hours. However, the upper cluster is located at a place with no features identifiable from satellite imagery, yet in the sample a number of vehicles have stayed there for long periods on the southeast side of the road, usually for around 8 hours. This potentially reflects an unmet need for rest area facilities.

Figure 10: Informal rest areas



6.4 Interregional movements

We can also aggregate trips into tours. This is important if we want to discover information about the volume of traffic between regions, the travel time taken and the routes chosen. The database has been designed so these tours can be defined with a number of criteria depending on the definition of the use. These include

- Regions to and from (including SA2 to SA4, state and Greater Capital City Statistical Areas).
- Minimum stop time to signify the beginning or end of a tour
- Maximum stop time within a tour

- Maximum length of a tour
- Assumptions about route choice (for instance requiring the use of at least some of a list of segments).

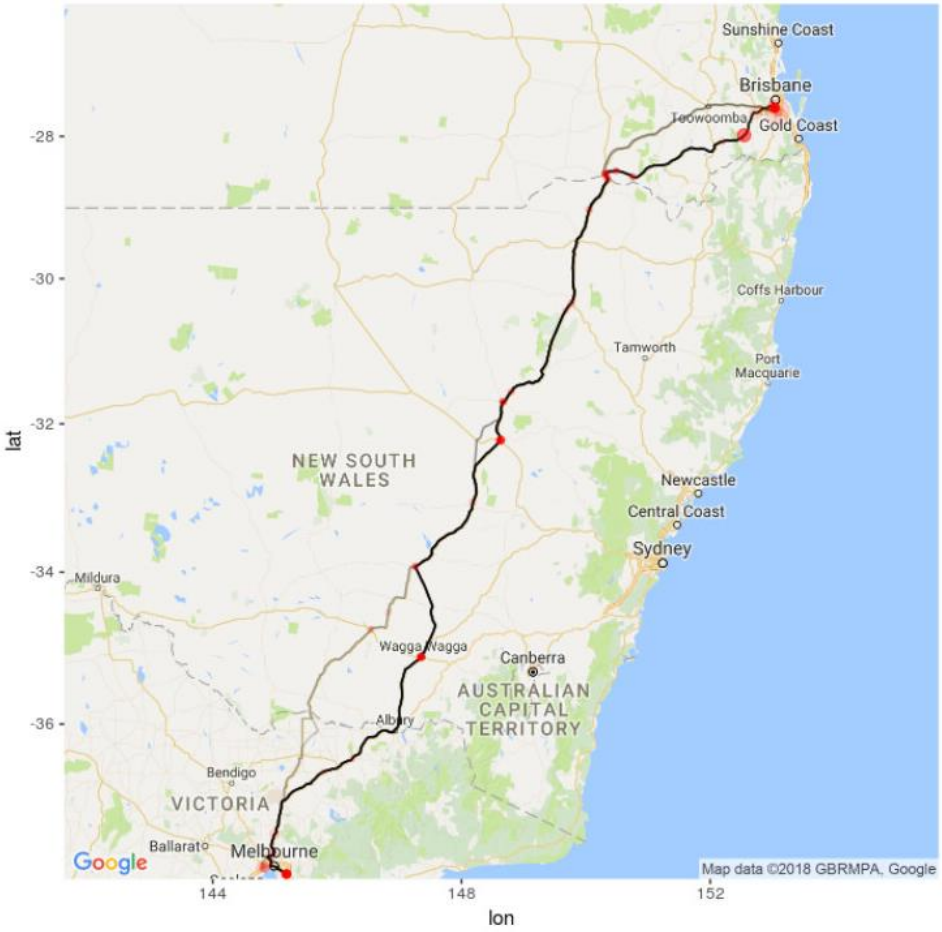
These criteria are crucial. Whilst we can observe vehicle behavior, we have information about intent; a vehicle may leave one city and arrive at another, but this may not have been a "trip" as far as the driver and firm were concerned. They may have considered this a trip from the first city to an intermediate point and then another trip to the second city, or just part of a longer trip to another location entirely.

Figure 11, below, shows the results of a query that creates tours with the following attributes:

- left Melbourne
- arrived in Brisbane
- within 48 hours
- whilst not stopping twice in either city (for more than an hour)
- or stopping for more than 24 hours on route
- does not assume any route choice.

It reveals a number of points where vehicles in the sample traveled via alternate route options, including via Wagga Wagga or Combram (and in the latter via Numurkah or Katamatite), and whether to bypass Dubbo and Toowoomba. However, if we were to change just one criterion, and allow tours up to a week in length, then we would derive more tours from Melbourne to Brisbane that went by Sydney, the Sunshine Coast and even Adelaide. Since users are unlikely to worry about the length of a Melbourne to Brisbane via Adelaide tour, this shows how carefully criteria need to be chosen to answer the question we are asking.

Figure 11: Tours from Melbourne to Brisbane



7. Conclusion and further work

We have developed a process that can transform data from freight providers into information about freight activity, stop locations and the road freight network. The structure of the process is flexible enough to adapt to user needs, and can be extended to other analysis including incorporating weather and other variables; or applying machine learning and other techniques. The underlying framework is also applicable to other modes, including cars, cycles and pedestrians, as long as these modes are restricted to routes in Open Street Map.

8. References

Cich G., Knapen L., Bellemans T., Janssens D., Wets G. 2015 'Threshold settings for TRIP/STOP detection in GPS traces', *Journal of Ambient Intelligence and Humanized Computing*, Volume 7, 2016.

Dash Nelson G, Rae A (2016) An Economic Geography of the United States: From Commutes to Megaregions. *PLoS ONE* 11(11): e0166083.

American Transportation Research Institute (ATRI) 2015, *Congestion Impact Analysis of Freight-Significant Highway Locations – 2015*, ATRI, Arlington, Virginia. URL: www.atri-online.org.

Geoscience Australia (GA) 2006, *GEODATA TOPO 205k Series 3*, Geoscience Australia, Canberra. URL: www.ga.gov.au.