Door-to-door or not door-to-door? A workflow for the characterisation of drivers GPS traces in a stochastic carpooling service

Panayotis Papoutsis, Safa Fennia, Constant Bridon, Tarn Duong

Door-to-door or not door-to-door? A workflow for the characterisation of drivers GPS traces in a stochastic carpooling service

Panayotis Papoutsis*†‡§ Safa Fennia† Constant Bridon† Tarn Duong†

March 31, 2020

Abstract

Carpooling has the potential to be a key component of the transport mix in post-carbon, ecologically sustainable societies. To achieve this potential, it is imperative that the carpooling is transformed from an individualistic, private mode of transport to a hybrid private-public mode. Door-to-door matches are highly convenient for individual, sporadic carpool journeys, but hinder the development of carpooling as a mass transit service. The focus shifts to matching trajectory segments which are highly frequented by both passengers and drivers. Whilst this leads to a loss of personal convenience, as drivers and passengers are constrained to converge at meeting points along the trajectory segments, these meeting points act as aggregators to reach a critical mass of matched passenger demand and driver supply. In a world first, Ecov provides innovative carpooling services where passengers make carpooling requests without a priori matched drivers, and that these requests are stochastically matched in real-time to the driver flow on the pre-selected road segments. The mathematical complexity of stochastic carpooling matching greatly exceeds that of the traditional deterministic services. To address this complexity, we introduce a workflow, comprising of a combination of data science and GIS (Geographic Information Systems), which processes driver GPS traces in order to provide important indicators (e.g. driver participation rate and passenger waiting time) to guide the operational decision-making. We illustrate this workflow on a currently operational carpooling service in the peri-urban region surrounding the city of Lyon, France.

Keywords: data science-GIS, driver flow, driver participation rate, ride-share, travel time, passenger waiting time

1 Introduction

The business model of current market leaders in carpooling, such as Uber, Lyft, Kapten and others, involve constructing large fleets of professional drivers who respond to the passenger requests. This model provides door-to-door carpooling services, where a passenger makes a request at a given time to travel from a given origin to a given destination. This request is then matched deterministically from the database of available drivers. Whilst these door-to-door services possess a high level of convenience as they respond closely to individual travel

*Corresponding author. Email: panayotis@ecov.fr.
†Ecov, Nantes F-44200.
‡Department of Computing and Mathematics, Nantes Central Engineering School, Nantes F-44300
§Jean Leray Mathematics Laboratory, Nantes F-44300

requirements in both time and space. The fact that it satisfies these requirements makes the door-to-door carpooling services incompatible with a large scale utilisation and is not ecological. Thus these market leading carpooling service providers now recognise that large scale utilisation of carpooling depends crucially on relaxing this door-to-door requirement by incentivising both passengers and drivers towards pre-selected meeting points (Stiglic et al. 2015). For example, "Suggested Pickup Points" are proposed by Uber (2019) to passengers so that the drivers avoid taking inefficient and/or infrequent routes. This incentivisation is well-established for public transport services where these predefined meeting points correspond to bus stops or train stations. Thus large-scale utilisation of carpooling requires a paradigm shift from considering carpooling as an exclusively individual means of transport to a closer alignment to mass public transport models, as presaged in Cooper (2007).

In a world first, a range of hybrid public-private carpooling services are proposed by the carpooling provider Ecov (`Ecov.fr`). Further following the public transport model, within Ecov's services, these meeting points and road segments are not defined informally between passengers and drivers, but their placement are decided in consultation with local government authorities so that they respond to the mobility requirements in the local area, which take into account various factors such as aggregated traffic flow, socioeconomic characteristics, pedestrian accessibility, local government regulations, etc. Since they are fixed, physical meeting points, we do not require those matching algorithms which are concerned with the identification of dynamic meeting hotspots between passengers and drivers (Schreieck et al. 2016). These meeting points are marked with fixed, physical structures which are easily visible by drivers on the road, analogous to bus or coach shelters. These meeting points are connected to each other to define route segments, which have a large massification potential, like traditional bus lines. Unlike many of its competitors, the Ecov carpooling services do not cater to densely populated, highly urbanised areas, but rather to peri-urban or rural areas which are often marginalised by transport providers. The lack of transport options in these areas is a key contributing factor in many social issues, like chronic unemployment (Fransen et al. 2019). For these sparsely populated and less well digitally connected areas, the physical meeting points provide local residents access to an economically and ecologically sound transport service.

The other world leading innovation that Ecov carpooling services bring to the market is the shift away from the deterministic matching between passengers and drivers to a stochastic matching. Carpooling usually operates with the individual passenger request being matched deterministically to a particular driver with an agreed departure, destination, and time frame. This deterministic matching requires considerable planning and is well-adapted to infrequent, long distance journeys and/or densely populated areas, e.g. as demonstrated by the market penetration in France of the long-distance carpooling provider BlaBlaCar (`www.blablacar.fr`). On the other hand, for frequent, short distance journeys (from 10 to 40 km roughly) in more sparsely populated areas, which comprises the bulk of home-work commutes, this type of planned carpooling is not adapted. Ecov's carpooling approach removes these pre-planning requirements, as it allows a passenger to make an immediate carpooling request without reservation at a meeting point, since the service subsequently displays the desired destination on an electronic sign on the side of a main road in order to alert the passing drivers of this request in real-time. Since the actual driver who will collect the passenger is not known in advance, but is only known to be drawn from the population of drivers, this is a *stochastic matching*. The innovations proposed by Ecov are only sparsely covered by the recent comprehensive review of the evolution of carpooling services in the past two decades (Wang & Yang 2019).

The physical meeting points provided by Ecov require an integrated infrastructure to facilitate this real-time stochastic matching, as illustrated in Figure 1. Our example is the "Lane" carpooling service (lanemove.com) operated by Ecov, in conjunction with Instant System (instant-system.com), since May 2018 in the south-eastern peri-urban regions around Lyon, France. The orange structure on the right functions like a bus shelter to provide (a) protection from inclement weather whilst the passenger waits, and (b) a prominent visual point of reference for drivers on the road. The passenger makes a carpooling request on the console (the green device with a horizontal yellow stripe) close to the shelter. This request is displayed on the electronic sign on the roadside. In this configuration, the electronic sign is located close to the meeting point, but this can vary considerably according to the local geographical characteristics. A driver who wishes to embark the passenger in response to their request is able to do so safely in the reserved parking place.

Figure 1: Configuration of a physical meeting point for the "Lane" real-time carpooling service. The orange structure functions like a bus shelter. A passenger notifies potential drivers of their carpooling request using the console, which is then displayed on the roadside electronic sign. A driver can safely embark the passenger in the reserved parking place. Reproduced with permission from Ecov.

These meeting points are unable singly to provide a sufficiently high level of service for passengers and drivers for stochastic matched carpooling. To assure this, they are organised into carpooling lines where each line is made of at least two meeting points. The schematic of the carpooling lines in the Lane network is shown in Figure 2. The visual similarities of the schematic of this carpooling service with those associated with bus or train services is designed to induce the perception of Ecov carpooling as a form of public transport. There are 5 physical meeting points, denoted by the circles with the stylised $\mathcal{L}$, which function analogously to bus stops. According to mobility studies in this territory, the coloured lines connect the meeting points that have a sufficient driver flow between them to maintain a carpooling service with stochastic matching. These connected meeting points form a carpooling line, again analogous to a bus line, where carpooling is only available between these pre-selected meeting points.

In this carpooling service, minimal restrictions are placed on the passengers' participation, which consist mostly of arriving at a meeting point during the service operating hours, and

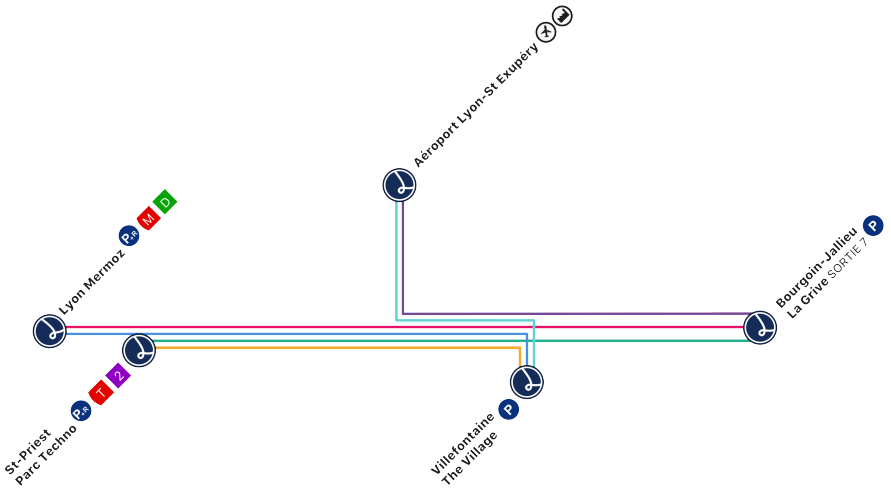

Figure 2: Schematic diagram of the Lane carpooling service, which resembles the geographically restrained trajectories of a public transport service. Reproduced with permission from Ecov.

being prepared to *not* have a fixed departure time. On the other hand, the onus is placed on the population of drivers since they collectively must be ready to respond to a passenger carpooling request in a timely manner. Due to this asymmetry of the involvement of passengers and drivers, according to several empirical studies, the constitution and the retention of a sufficiently sizeable population of participating drivers who can respond to passenger carpool requests in a timely manner is the key element of this stochastic carpooling service (Zhu 2017, 2018). Its feasibility is relying on a comprehensive characterisation of the temporal profiles of the driver flow on the road segments connecting the meeting points.

In this paper, firstly we present the complexities of the stochastic matching and the motivations behind its organisation into pre-selected carpooling lines. Secondly, we propose a data science-GIS workflow to characterise the carpooling driver supply using GPS traces, and elaborate how it is utilised by Ecov to provide information to the passengers concerning their waiting and travel times, for the Lane carpooling service. We end with some concluding remarks.

2 Door-to-door matching as an obstacle to mass carpooling utilisation

Carpooling has seen an explosion of utilisation in recent years. There are many underlying reasons, with concerns ranging from greenhouse gas emissions to road congestion, air pollution, land use, as well as economic costs. It is also attracting intense interest since carpooling is crucial element of almost all developments plans for smart cities (Ghoseiri et al. 2010, Ghoseiri 2012). As alluded in the introduction, door-to-door matching of complete trajectories from the origin to the destination is a structural obstacle to the transformation of carpooling as a mass transit service.

To illustrate the difficulties of spatio-temporal matching for door-to-door trajectories (i.e. passenger-driver matching in space and in time), we can represent it with partition of a 3D cube divided into smaller sub-cubes, where the x-axis is the longitude, the y-axis the latitude and the z-axis the time, as shown in Figure 3. On the left, there are 9 sub-cubes, where each

sub-cube represents the origin/destination of a door-to-door trajectory. The blue sub-cube in the lower left represents all the trajectories whose origins are, say, within a 5 km radius around a residential neighbourhood between 7.00am and 9.00am on Tuesday, and the green sub-cube the trajectories whose destination are within a 5 km radius of the workplace between 8.00am and 10.00am on Tuesday. So for two trajectories to match spatio-temporally in a door-to-door sense, they must share the same sub-cube for the origin, and similarly for the destination: this condition is met only by the 1 pair of green and blue sub-cubes among all possible 27 pairs of sub-cubes. On the right, the conditions for a door-to-door matching are stricter, say the origin is 1 km within the residential neighbourhood during 7.00am to 7.30am, and the destination is 1 km within the workplace during 8.30m to 9.00am. This represents 1 pair out of 125 pairs of sub-cubes. Recall that Ecov's carpooling services comprise non-professional drivers who do not create a trajectory upon a passenger request, but rather mutualise their existing trajectories, so door-to-door matching leads to a combinatorial dilution of spatio-temporal matches. Thus for Ecov, it is crucial to avoid exact door-to-door matching and to move towards matching highly frequented partial road segments of door-to-door trajectories.

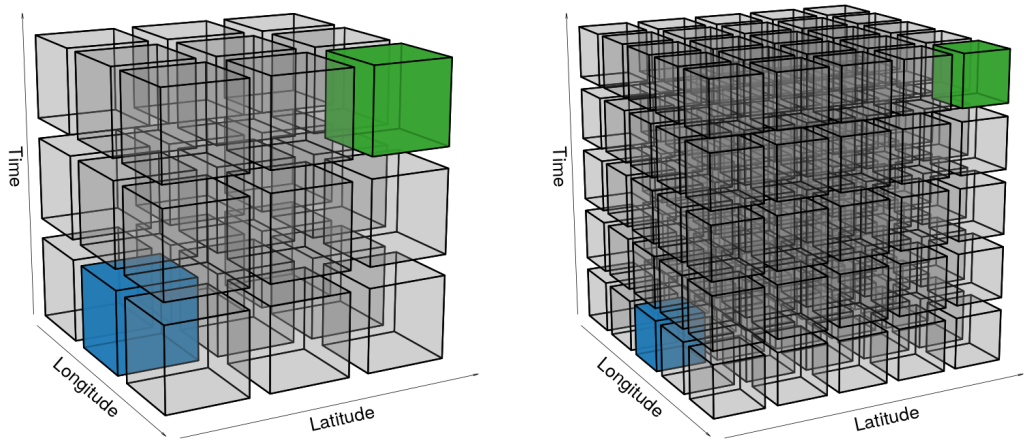

Figure 3: Spatio-temporal door-to-door matching fragments the population of mutualisable trajectories. (Left) Relaxed matching conditions. (Right) Restricted matching conditions. Blue sub-cube represents the origin (residential neighbourhood), green the destination (workplace), and trajectories which share the same origin and destination sub-cubes are considered to be door-to-door matches.

To supplement the heuristic observations for door-to-door matching in Figure 3, we demonstrate that the probability that two users (i.e. a driver and a passenger) share the same origin and destination at the same time decreases rapidly as the spatio-temporal matching conditions become more stringent. For the sake of simplicity, we suppose that the origin and destination for a driver and a passenger are both represented by independent random variables which are uniform over all sub-cubes in Figure 3. If we draw a random sample of 1000 each of drivers and passengers, then the probability of any door-to-door match between these drivers and passengers, as a function of the number of sub-cubes, is given in Figure 4. If there is only 1 sub-cube (i.e. no spatio-temporal constraints) the probability of a match is 1. This probabilistic certainty

decreases rapidly as the spatio-temporal constraints are added: for 27 sub-cubes, this probability is 0.7, and for 125 sub-cubes, it falls to 0.26. Thus it is almost impossible for a carpooling service, if it is based on complete door-to-door spatio-temporal matching, to evolve into a mass transit service. Stiglic et al. (2015), Li et al. (2018) provide more complex synthetic models to affirm that meeting points are essential to the feasibility of the mass carpooling services.

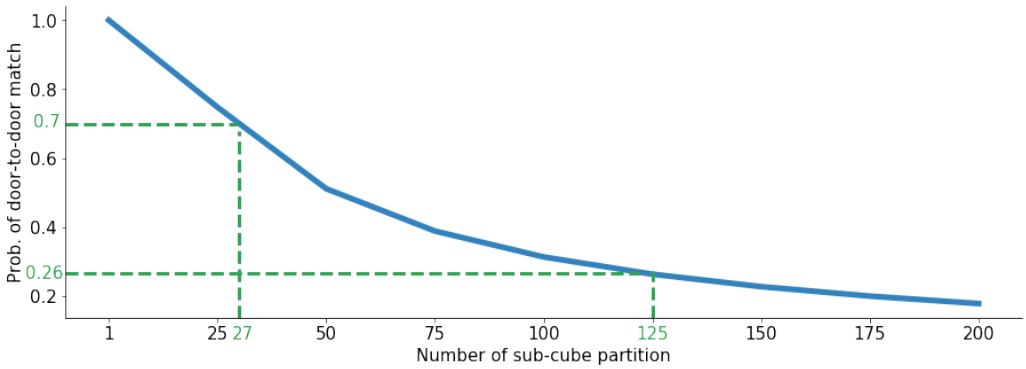

Figure 4: Probability of door-to-door matches for uniformly distributed drivers and passengers, as a function of the number of sub-cube partition classes. Higher number of sub-cubes represent more stringent spatio-temporal matching conditions.

Given that the probability of door-to-door matches diminishes rapidly, apart from increasing the number of avalaible drivers we propose also to enlarge the pool of potential matching by relaxing the spatial conditions. In Figure 5, this is represented by extending the origin and destination to cover 12 sub-cubes each in a horizontal layer, rather than as single sub-cubes. This increases the probability of a match from 0.26 to 0.39.

The previous analysis was based on the uniformly distributed origin and destinations. To offer a more realistic example, we analyse some data generated by an operational Ecov carpooling service. Our main data source is the GPS traces of drivers who are registered with Ecov's services, which can be considered to be a form of crowd-sourced data collection (Lee & Liang 2011). Passenger GPS traces are more difficult to obtain, and as we are not able to replicate exactly the synthetic example of passenger-driver matching above, we use door-to-door matching of driver GPS traces to illustrate the diminishing probabilities. This supplements the results for synthetic data experiments in Stiglic et al. (2015), Li et al. (2018) with empirical results.

Since these GPS traces provide highly detailed spatio-temporal information, we are able to determine the number of empirical door-to-door matches, as well as the effect on the number of matches when matching is carried out between two fixed, physical carpooling meeting points. For an illustrative example in Figure 6, we analyse the $n = 121$ GPS traces of drivers who travelled from the Bourgoin La Grive meeting point (solid black circle labelled B) to the Saint-Priest Parc Technologique meeting point (solid black circle labelled S) in the Lane carpooling service during the morning operating hours (06:30am to 09:00am). We temporarily ignore the location of the carpooling meeting points, and focus on the GPS traces and their origin and destinations. A hierarchical clustering with complete linkage was carried out on the spatial locations of these origins and destinations. The dissimilarity matrix used for this hierarchical clustering is composed of the Euclidean distance between the 4-vector comprising the (origin longitude, origin latitude, destination longitude, destination latitude) of each trajectory. This dissimilarity takes into account both the origin and the destination. On the other hand, it does

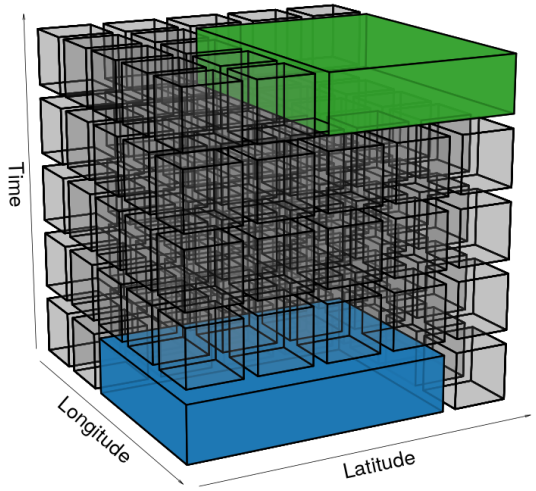

Figure 5: Meeting points aggregate the spatio-temporal matching of mutualisable trajectories. The green sub-cube represents the coverage area of the origin meeting point (residential neighbourhood), the blue the destination meeting point (workplace).

not take into account the intermediate GPS points as these actual route taken is not critical for our purposes. We cut the dendrogram at $h = 6000$ to yield 9 spatial clusters. These clusters are represented with the different colours: the origin and destinations are the diamonds, and the GPS traces are the points. So GPS traces with the same colour can be considered as door-to-door spatio-temporal matches. Upon visual inspection, all 121 GPS traces share the segment between the Bourgoin and St-Priest meeting points, even though the routes on the road network are different. If we take into account the door-to-door matching, then these 121 trajectories are fragmented into 9 groups.

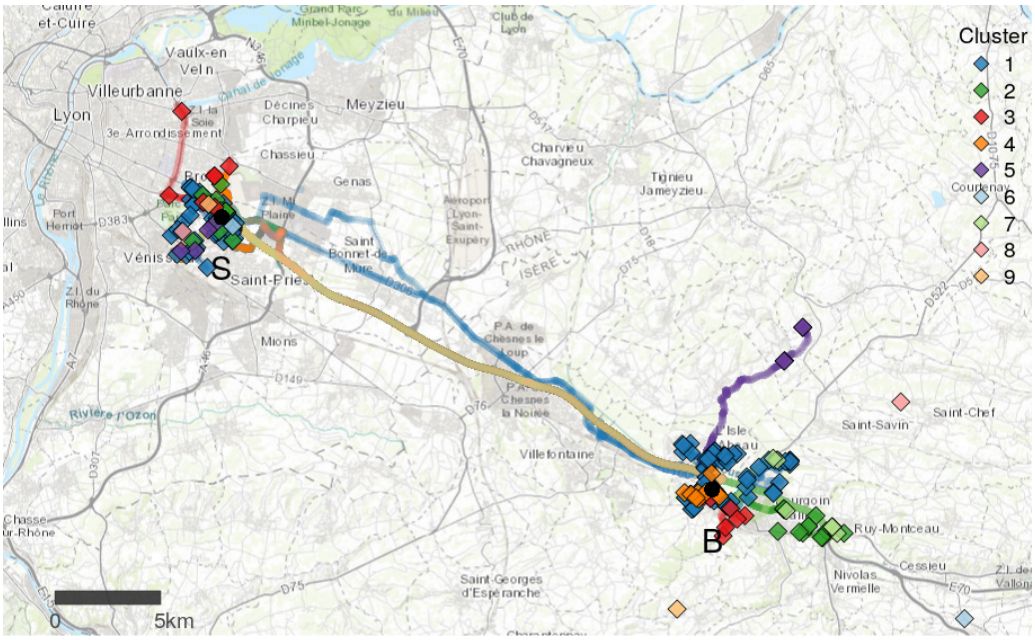

Figure 6: Spatio-temporal door-to-door matching fragments the number of mutualisable trajectories in a carpooling service. The clusters of GPS traces of door-to-door matches are colour coded, with the GPS points as the solid circles, and the origins/destinations as the solid diamonds. The meeting points are the solid black circles: B = Bourgoin, S = St-Priest.

The number of GPS traces per cluster is given in Table 1: whilst cluster 1 contains 75% of the mutualisable traces, so this leaves the other 25% spread sparsely over the other 8 clusters, fragmenting the supply of the carpooling trajectories. Moreover all 121 trajectories share the central Bourgoin to St-Priest segment.

Door-to-door cluster	1	2	3	4	5	6	7	8	9	Total
Number of GPS traces	76	15	7	9	4	1	7	1	1	121

Table 1: Spatio-temporal door-to-door matching fragments the number of mutualisable trajectories in a carpooling service. Door-to-door matches induced by hierarchical clustering of GPS traces. The first line of the table contains the number of the clusters and the second line the number of traces for each cluster.

3 Data science-GIS workflow for a stochastic carpooling service

The Data Science-GIS department of Ecov has developed many workflows in response to the multiple challenges posed by the collection, storage and analysis of numerous, heterogeneous data sources stochastic carpooling services: for brevity we focus on the GPS traces workflow, as illustrated in Figure 7. The left part of the Figure 7 contains the main data sources: the GPS traces, the meeting point locations, the origin-destination matrices, the route finder API and the base maps. The first two are stored as PostGIS SQL databases on a secure server owned by Ecov, the origin-destination matrices are provided by the French national statistical

agency (INSEE 2018), the route finder API is provided by the GPS navigation operator (TomTom 2019), and the base maps are accessed from the cartography provider OpenStreetMaps (OpenStreetMap contributors 2019). There are specialised data wrangling techniques specific to spatial databases, known collectively as *geoprocessing*, and these are carried out, in conjunction with traditional data wrangling, in the central lozenge. The critical geoprocessing concerns the topological simplification of the GPS traces onto its network map of carpooling meeting points and lines. Whilst GPS traces are a rich source of information of driver behaviour, they are voluminous and complex, and whose complexity can be highly variable depending on the GPS technology deployed. Our approach is based on network analysis tools (Guidotti et al. 2017) and complexity reduction/harmonisation algorithms (Douglas & Peucker 2011). This topological simplification is essential to be able to mutualise GPS traces which share common arrival times at the carpooling meeting points. Once these GPS traces are in a suitable format, we are able to produce the required outputs, such as maps of the the geographical extent of the GPS traces, the driver flow per route segment in the carpooling lines and/or per time interval, and subsequently the corresponding waiting times and travel times, as outlined in the right lozenge.

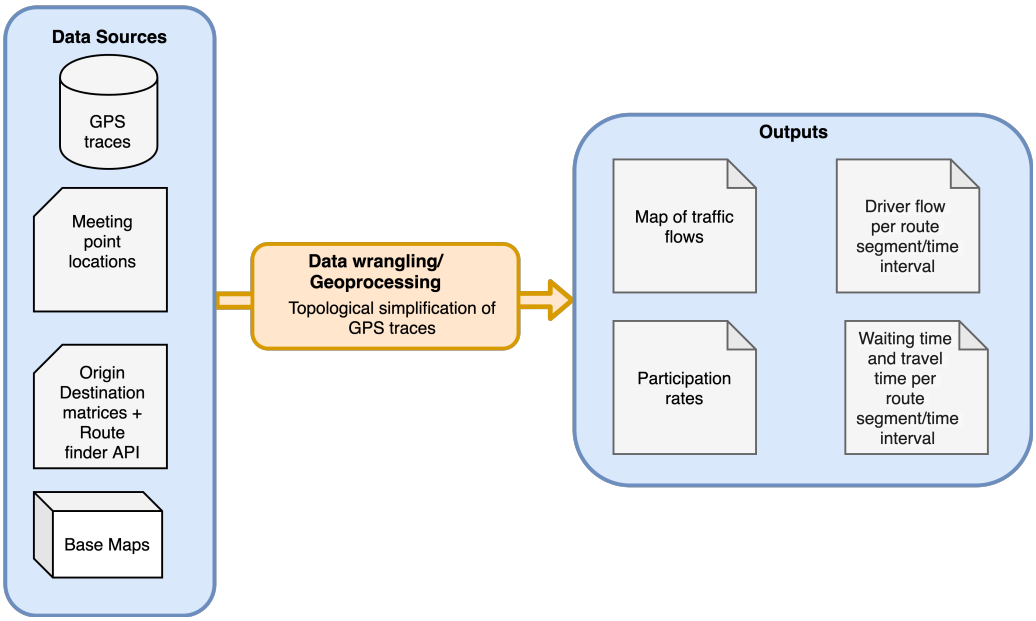

Figure 7: Data science-GIS workflow for the analysis of driver GPS traces in stochastic carpooling service.Left. Spatio-temporal input data sources. Centre. Data wrangling and geoprocessing. Right. Generated outputs.

3.1 Topological simplification of GPS traces on a carpooling line

The basis for the topological simplification of the GPS traces is the network of the Lane carpooling network from Figure 2. This network is represented as a directed graph, where each node is a meeting point and the edge connects two nodes if they form segment of a carpooling line, as shown in Figure 8. For brevity, the node labels have been abbreviated to the first letter of the name of the meeting point, i.e. L = Lyon Mermoz, S = St-Priest Parc Technologique, A = Aéroport Lyon-St Exupéry, V = Villefontaine The Village, and B = Bourgoin La Grive Sortie 7. Since the primary objective is to match passenger and driver trajectories on the arrival

times at the meeting points, then the actual route taken between these two meeting points is of secondary interest so we can represent all routes connecting from one meeting point to the other as a single directed edge. The identification of all GPS traces which share arrival times at the two meeting points, by ignoring the intermediary routes taken, to a single directed edge in the graph, is the mathematical abstraction which facilitates the massification of the driver trajectories which are able to fulfil a passenger carpooling request which respects the latter's spatio-temporal conditions.

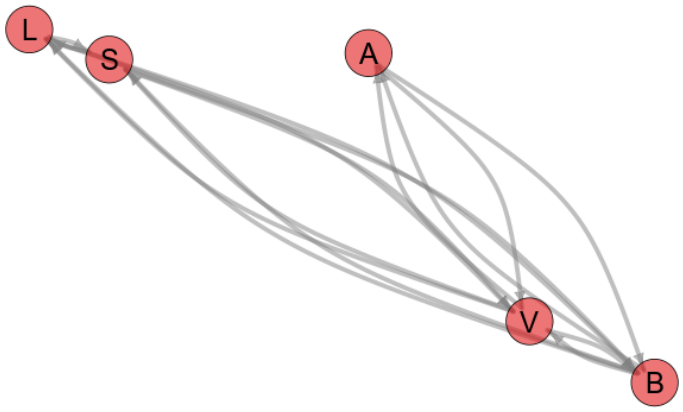

Figure 8: Network of carpooling lines represented as a directed graph. Nodes are the meeting points, edges connect meeting points whenever a carpooling service between them is assured.

A GPS trace is displayed as the sequence of blue circles in Figure 9. It has a complex topology since it is represented by 530 GPS points which follow (more or less) the road network. This complex topology is simplified by retaining a small number of key indicators derived from the complete GPS trace, following Lee & Liang (2011). In addition to the nodes associated with the origin (first GPS point) and destination (last GPS point), we retain only those meeting place nodes if the GPS trace contains a point within a 1 km radius of the nodes. The GPS trace in Figure 9 passes within 1 km of the B, V and S meeting place nodes, so the resulting simplified topology consists of the 5-node sequence: {origin $>$ B $>$ V $>$ S $>$ destination}. These simplified topologies represent a considerable reduction in data complexity, whilst the crucial characteristics with respect to the carpooling service are retained. The individual GPS locations are a secondary detail since it is vastly more important to know if a driver (a) passes by a carpooling meeting point and (b) travels in which direction to which other meeting point(s) in the network.This approach contrasts with Tiakas et al. (2009) who attempt to match driver trajectories along the entire length of the traces.

Recall that a carpooling service is assured between Bourgoin and St-Priest since edge connects the two nodes in Figure 8. So if the directed graph of the simplified topology of a GPS trace contains {B $>$ V $>$ S} or {B $>$ S} as a subgraph, then this driver's GPS trace is able to participate in this carpooling line. This is the case for all 121 GPS traces under consideration. Furthermore, $n = 31$ (out of 121) GPS traces have an arrival time at Bourgoin within 08:00 am to 08:30 am, i.e., around 26% of the GPS traces are a close spatio-temporal match for a passenger request for departure at the Bourgoin meeting point (in a residential neighbourhood) between 08:00 and 08:30 am, with a destination at the St-Priest meeting point (in a neighbourhood with a high employment density). In comparison to the door-to-door matches in Figure 6, for the largest cluster of 76 GPS traces, only 14 of these share a departure time from their origins in the same

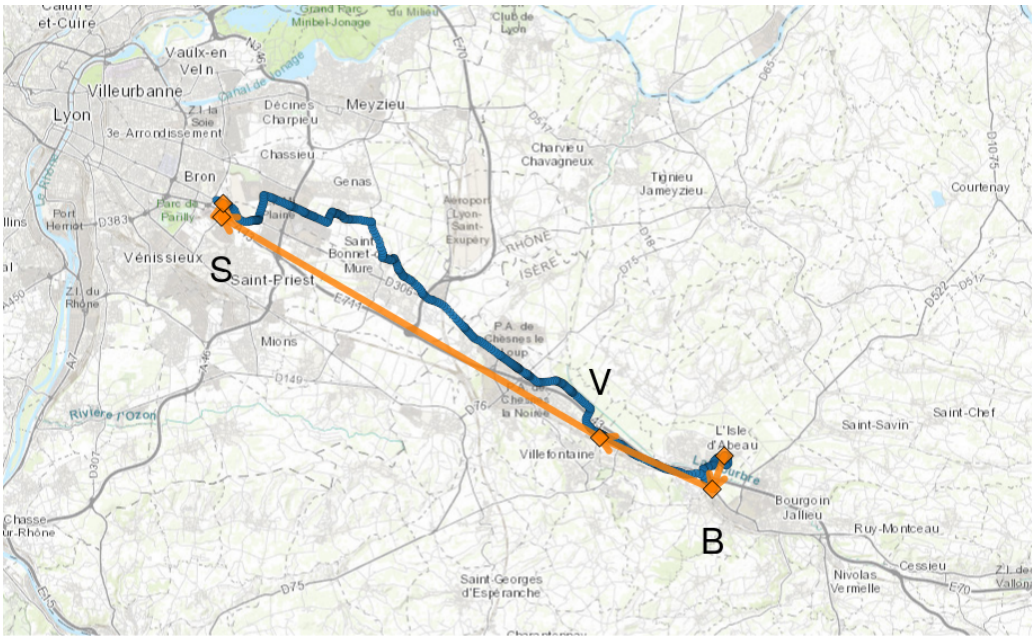

Figure 9: Topological simplification of a GPS trace. The complete GPS trace are the 530 blue circles; the sequence of five nodes, as its simplified topology, are the orange arrows, and the orange diamonds are the origin, carpooling meeting points, destination nodes. The meeting points are S = St-Priest, V = Villefontaine, and B = Bourgoin.

interval of 08:00 am to 08:30 am, i.e. around 12% of the 121 GPS traces under consideration. Moreover, all 14 of these door-to-door matches are a subset of the previous 31 Bourgoin to St-Priest meeting place matches. The simplified traces of these 14 traces which are both door-to-door and meeting place matches are blue arrows and diamonds in Figure 10. This implies that 17 of the meeting place only matches (orange arrows and diamonds) can be added to the 14 door-to-door matches to reinforce the number of potential matches, i.e. allowing for meeting place matches increases the proportion of potential spatio-temporal matches from 12% to 26%.

3.2 Driver flow estimation

These simplified GPS traces with timestamps, in addition to being direct means of aggregating driver GPS traces to augment the number of possible spatio-temporal matches to a passenger carpooling request, they also greatly facilitate the calculation of detailed temporal profiles of the driver flows on each of the segments in the carpooling service. Table 2 displays the average driver flow for 15 minute intervals during 06:30 to 09:00 (morning operating hours). Intervals of 15 minutes correspond roughly to the maximum time that passengers are willing to wait in a real-time stochastic matching carpooling service. Moreover, Smith & Demetsky (1997), McShane & Roess (1990) indicated that these 15 minutes intervals are an optimal choice because the variation in driver flows for shorter intervals is less stable.

3.3 Waiting time prediction

From the passenger point-of-view, a key quality measure of a carpooling service is the driver arrival in a suitable time frame. The estimation of the time of arrival (ETA) is a vast subject

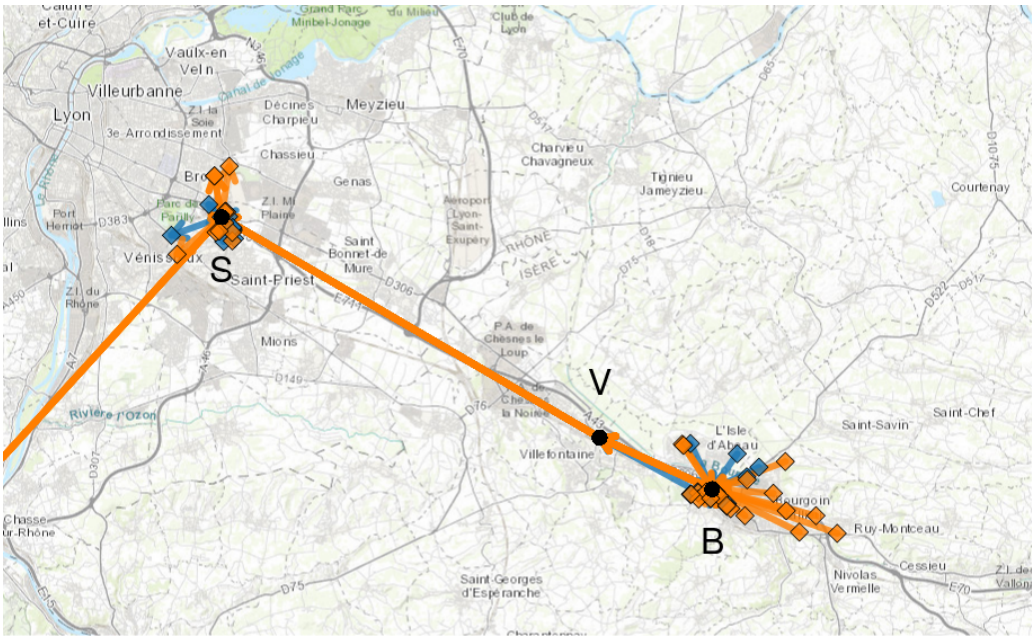

Figure 10: Matching on meeting points increases the number of driver spatio-temporal matches in comparison to door-to-door matching. The orange arrows are the GPS traces which are a meeting point match but not a door-to-door match ($n = 17$), and the blue arrows are the GPS traces which are both meeting point and door-to-door matches ($n = 14$). The diamonds are the origin/destination points. The solid black circles are the meeting points: S = St-Priest, V = Villefontaine, and B = Bourgoin.

	Driver flow					
	06:30-	06:45–	07:00–	07:15–	07:30–	07:45–
Line	06:45	07:00	07:15	07:30	07:45	08:00
B > S	1	1.5	2.5	1.5	3	1.5
	08:00–	08:15–	08:30-	08:45–	09:00–	09:15–
Line	08:15	08:30	08:45	09:00	09:15	09:30
B > S	2	2	2	1	1	1

Table 2: Average driver flow on the Bourgoin > St-Priest carpooling line, per 15 minute intervals, during the morning operating hours 06:30 am to 09:00 am. For sake of simplicity, the average is taken over a period of two weeks.

of active research in itself, see Wang & Yang (2019) for a recent review of these methods within the larger context of the characterisation of carpooling services. For deterministic matching carpooling, the problem of waiting time prediction is the estimation of the travel time of the matched driver to reach the given passenger. For stochastic matching, since a specific driver is not matched to the given passenger, the problem is different since it is the estimation of the arrival time of the first driver from the population of available drivers. Given that we have already established a highly detailed spatio-temporal profile of the average driver flow on the segments between meeting points (in Table 2), with the added hypothesis of driver arrivals as a Poisson point process, then it is straightforward to convert these driver flows into an estimation

of the waiting time. It is a reasonable assumption that the first geolocated driver will pick up the passenger: according to unpublished figures supplied by Ecov, the majority of regular carpooling journeys are assured by motivated drivers who are willing to share their geolocation, and only a minority by unregistered and/or non-geolocated drivers.

Suppose that a passenger makes a carpool request at 08:10 am at the Bourgoin meeting point to travel to St-Priest. Then the expected waiting time is the length of the interval divided by the average driver flow in the interval 08:00 – 08:15, i.e. 7.5 minutes from Table 3.

	Predicted waiting time (mins)					
Line	06:30-06:45	06:45–07:00	07:00–07:15	07:15–07:30	07:30–07:45	07:45–08:00
B > S	15.0	10.0	6.0	10	6.0	5.0
Line	08:00–08:15	08:15–08:30	08:30-08:45	08:45–09:00	09:00–09:15	09:15–09:30
B > S	7.5	7.5	7.5	15	15	15

Table 3: Waiting time predictions for a carpool request on the Bourgoin > St-Priest carpooling line, per 15 minute intervals, during the morning operating hours 06:30 am to 09:00 am for an ordinary day of the week.

More formally let $W(i,t)$ be the waiting time until the first driver arrival for a carpool request made at time t made for carpooling line segment i, $i \in 1,\ldots,n_S$. Assuming a Poissonian driver arrival process, the waiting time and the driver flow are inversely proportional to each other, i.e. $W(i,t)f(i,j) \propto \text{const}$, where $f(i,j)$ is the driver flow for segment i and time interval τ_j. Then $W(i,t) \propto \text{len}(\tau_j)/f(i,j)$ where $j = \{k : t \in \tau_k, k \in 1,\ldots,n_T\}$ implies that t is contained in the time interval τ_j, $\text{len}(\cdot)$ returns the length of a time interval, n_S is the number of carpooling line segments and n_T is the number of time intervals. For simplicity, we set the constant of proportionality to one as this corresponds to the assumption that all geolocated drivers are willing to respond to a carpooling request, then the predicted waiting time until the first driver arrival for a carpool request made at time t made for carpooling line segment $i, i = 1,\ldots,n_S$ is thus calculated as

$$\hat{W}(i,t) = \text{len}(\tau_j)/\hat{f}(i,j)$$

where $j = \{k : t \in \tau_k, k \in 1,\ldots,n_T\}$ and $\hat{f}(i,j)$ is an estimate of the driver flow using the sliding window sample mean as outlined in Table 2.

Since these GPS traces are drawn from an operational carpooling service, we also have access to observed waiting times for roughly 1 500 successful carpooling requests on the Bourgoin > St-Priest carpooling line during a period of one year. So we are able to evaluate the accuracy of these predicted waiting times with respect to these observed ones, as illustrated in Figure 11. Each box plot covers a 15 minute interval during the opening hours with at least one observed waiting time. The blue box plots represent the observed waiting times and the predicted waiting times are the horizontal red lines. During the morning opening hours, the direction of travel is from Bourgoin to St-Priest, whilst in the evening it is the reverse from St-Priest to Bourgoin. The predicted waiting time as the reciprocal of the average driver flow is fairly reliable, especially in the morning operating hours.

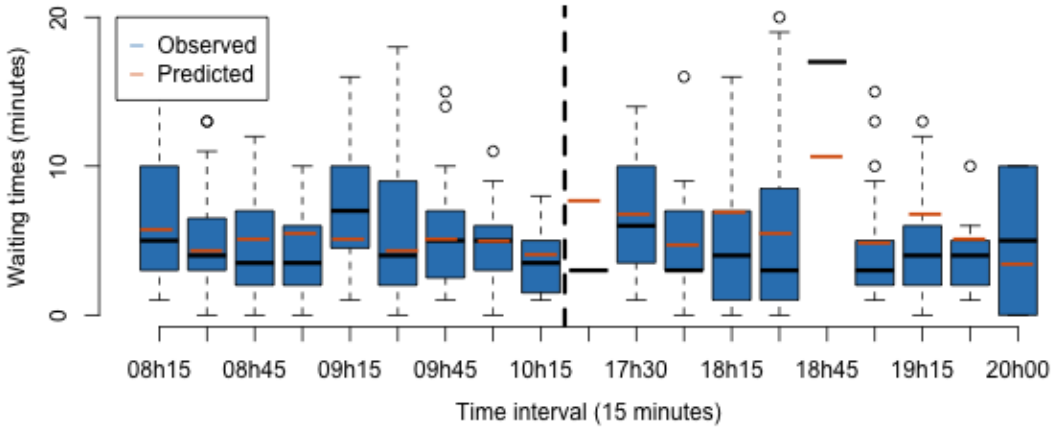

Figure 11: Comparison of predicted and observed waiting times for the Bourgoin <> St-Priest carpooling line. The blue box plots are the observed waiting times, and the predicted waiting times are the horizontal red line, for each 15 minute interval. The black dashed vertical line separates the morning and evening opening hours.

3.4 Travel time estimation

For a passenger, in addition to the waiting time in response to a carpooling request, the travel time from the meeting point to the destination is another important measure of the quality of the carpooling service for a passenger. Since GPS traces contain the timestamps for each of the constituent GPS point, it is straightforward to compute an estimate of the travel time as the sample average of the difference of the origin and destination timestamps, as shown in Figure 12. The box plots reveal that the travel time in the morning peak hours from Bourgoin to St-Priest is longer to than the afternoon peak hours for the return trajectory from St-Priest to Bourgoin. This is due to the fact that these drivers tend to wish to arrive in a restrained time period around 09:00am in the morning peak hour, and thus creating more congestion than in the evening, when the journeys are more dispersed over a longer time interval.

3.5 Driver participation rate estimation

For any carpooling service, the availability of drivers who are able to respond in a timely manner to the passenger requests is crucial. For deterministic matching, this involves dispatching the closest available driver to the passenger's location: since these locations can be anywhere, this usually involves a large fleet of drivers to provide a rapid response. For stochastic matching to a fixed passenger location, the total number of drivers required for a similar response time can be much lower since the closest available driver is drawn from the existing driver flow. A key question for Ecov is what driver participation rate leads to passenger waiting times around 5 to 10 minutes, as observed in the Lane carpooling line in Figure 11?

The driver participation rate is $P = N/N_0$ where N is the total number of the drivers who are motivated to carpool in response to a passenger request, and N_0 is the total numbers of drivers

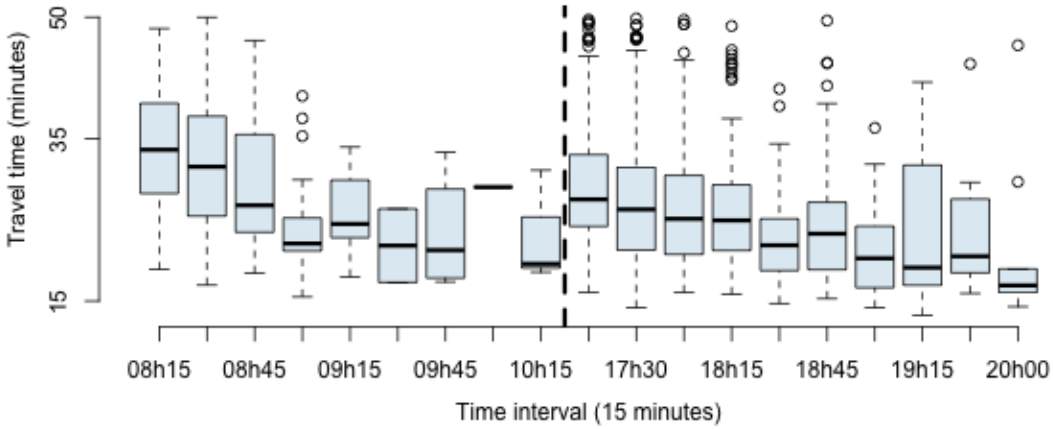

Figure 12: Travel time estimation for the Bourgoin <> St-Priest carpooling line for each 15 minute interval. The black dashed vertical line separates the morning and evening opening hours.

who undertake journeys in the same geographical region as the carpooling service. Both N and N_0 are difficult to define and to estimate precisely. We propose that the number of drivers who share their geolocation with Ecov be our proxy for N, even though Ecov's carpooling service allows for passengers to carpool with unregistered and/or non-geolocated drivers. From anecdotal evidence, the majority of regular carpooling journeys are assured by motivated drivers who are willing to share their geolocation, and only a minority by unregistered and/or non-geolocated drivers. It is also difficult to enumerate those drivers in the same geographical region as the carpooling meeting points, since the GPS traces for all drivers in general are not available. Our proxy is derived from inferring likely trajectories from the reference origin-destination matrix available for home-work journeys. In our case, a county-level origin-destination matrix is provided by the French official statistical agency (INSEE 2018). Since the county level data are insufficiently detailed to decide if the drivers with these origins-destinations travel on the same pre-selected road segments of the carpooling service, we infer likely trajectories. These inferred likely trajectories are determined by the TomTom route finder API (TomTom 2019) as the fastest route starting on Tuesday 8am from the origins (county centroids) to the destinations (county centroids), as shown in Figure 13. We employ a route finder API rather than an explicit model-based methodology, e.g. Tang et al. (2016), to infer these most likely routes. Thus N_0 is the sum of the driver flow from the origin-destination pairs whose likely trajectories includes road segments in the carpooling service. There are $N_0 = 3821$ drivers whose likely trajectories for the Bourgoin > St-Priest carpooling line. From Table 2, there are $N = 20$ between 06:30 and 09:30. This yields a driver participation rate of $P = N/N_0 = 0.52\%$. Even with this low driver participation rate, average waiting times for passengers of 5–10 minutes are observed for these time intervals in Table 3. This demonstrates that a small number of regular drivers assure the punctuality of the carpooling service, and the potential for stochastic carpooling to rival the waiting times proposed by traditional mass transit services does not require infeasible elevated driver participation rates.

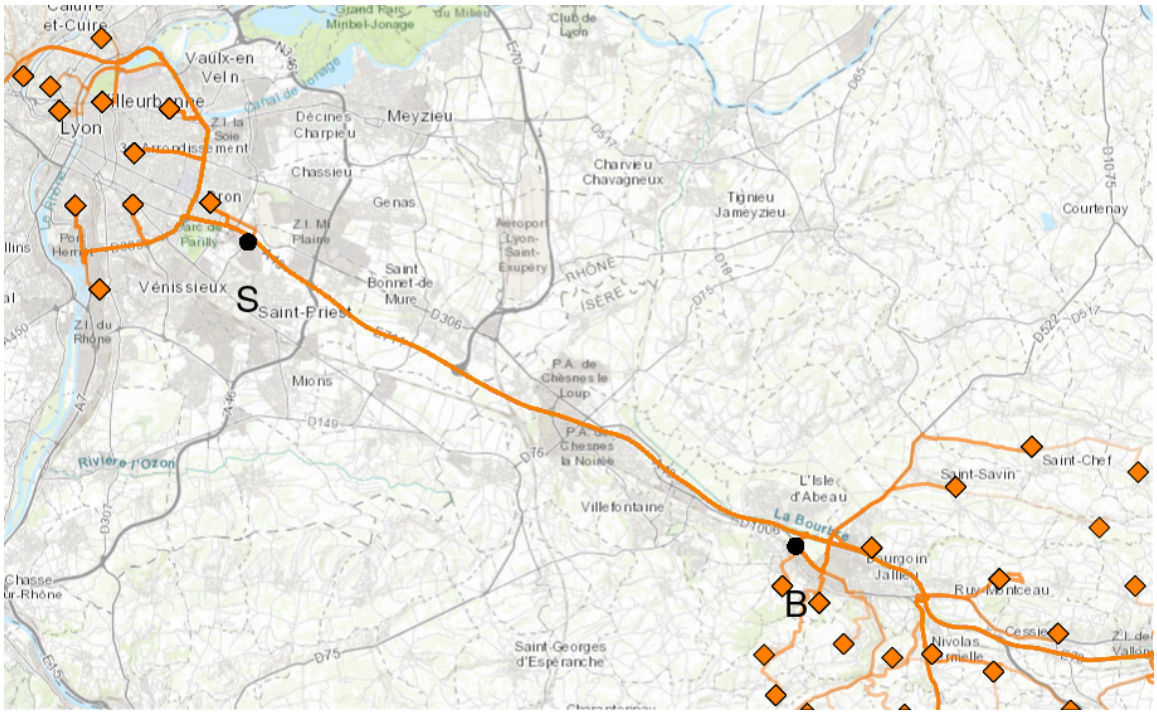

Figure 13: Likely driver itineraries from the TomTom route finder API in the same geographical region as the Bourgoin > St-Priest carpooling line. The origins and destinations (county centroids) are the orange diamonds. The solid black circles are the meeting points: S = St-Priest, B = Bourgoin.

We have only shown the results for the driver flow, the passenger waiting time, the travel time and the driver participation rate, for a single carpooling line (Bourgoin > St-Priest) within the Lane carpooling network, since this the most frequented origin-destination. The results for the return origin-destination (St-Priest > Bourgoin) and the origin-destination pairs involving the other meeting points (Villefontaine, Aéroport, Lyon), defined by the network graph in Figure 8, are produced for internal reporting with in Ecov, but are not shown here for brevity.

3.6 Software

The workflow, since it proposes a novel combination of data science and GIS, draws on software from both of these domains. The GPS traces are stored in the time-aware polyline format (Attam 2016) in a PostGreSQL database. Whilst this format allows for efficient storage of the GPS coordinates and the timestamps, it is not compliant with any standard format for spatial databases. Thus it is transformed into a PostGIS encoding (PostGIS Project Steering Committee 2019) which is one of the standard formats and is fully compatible with PostGreSQL database clients. For the data processing and analysis, e.g. to create the simplified topology of a GPS trace, a combination of R and Python packages are employed. We cite only the main ones which are focused on processing and analysing spatial databases, such as the sf package (Pebesma 2018) in R, and the GeoPandas (GeoPandas Developers 2019) and PyGeos (van der Wel 2019) packages in Python. Whilst these packages are able to handle most geoprocessing tasks, they are not complete, interactive GIS platforms in the strict sense like ArcGIS (ESRI

2019) or QGIS (QGIS Development Team 2019). The data science-GIS platform hosted at Ecov thus is a savvy combination of all of these tools in order to provide the computing infrastructure which is capable of agile responses to the analysis requirements for real-time stochastic matching carpooling services.

4 Conclusions and future work

In this paper we introduced a data science-GIS workflow to facilitate the stochastic matching between drivers and passengers on pre-selected road segments in the real-time carpooling services proposed by Ecov. These differ from competing services which offer door-to-door matching for complete trajectories. Whilst the latter offer a high level of personal convenience in highly urbanised regions, door-to-door matching structurally inhibits a large-scale adoption of carpooling. The mutualisation of high throughput road segments resolves this obstacle, especially in peri-urban and rural regions. The crucial mathematical abstraction in this workflow is to reduce the complexity of driver GPS traces to a graph-based topology which represents the pre-selected road segments of the carpooling service. With the data reduction accomplished, we were able to produce several key performance indicators (KPI), including the driver flow, waiting time, travel time, and participation rate with suitable spatio-temporal resolutions, which then inform the operational strategies to achieve critical market penetration. We illustrated this workflow on a carpooling service currently operated by Ecov in a peri-urban region in south-eastern France. The physical meeting points, by facilitating the convergence of a critical mass of drivers and passengers drawn from a much larger geographical area, forms the foundation of the high level of user satisfaction of a real time carpooling service as designed by Ecov.

This workflow represents only a small fraction of the envisaged, complete data platform. Nonetheless, it details a functioning prototype for the combination of two closely related, but historically separate, disciplines of data science and GIS into a single workflow which responds to the complex questions arising from addressing mobility requirements in previously neglected peri-urban and rural regions. In addition to the Lane carpooling service presented here, Ecov currently operates 6 other carpooling services across France. Ecov has recently been awarded substantial public-private partnerships (PPP) grants by government authorities to facilitate the expansion its operations in France, in Europe and in the rest of the world. Carpooling, as proposed by Ecov, based on stochastic matching of passengers to a driver flow, has a transformative potential within ecologically sustainable economies: to fulfil its potential it requires to be supported by an agile, comprehensive and scalable data platform.

References

Attam, A. (2016), *Time Aware Encoded Polyline for Geospatial Data*. Python package version 0.1.2. https://pypi.org/project/time_aware_polyline.

Cooper, C. (2007), 'Successfully changing individual travel behavior: Applying community-based social marketing to travel choice', *Transportation Research Record* **2021**, 89–99.

Douglas, D. H. & Peucker, T. K. (2011), 'Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature', *Classics in Cartography: Reflections on Influential Articles from Cartographica* **10**, 15–28.

ESRI (2019), *ArcGIS Pro*, Environmental Systems Research Institute, Redlands, USA. Version 2.5. https://pro.arcgis.com.

Fransen, K., Boussauw, K., Deruyter, G. & De Maeyer, P. (2019), 'The relationship between transport disadvantage and employability: Predicting long-term unemployment based on job seekers' access to suitable job openings in Flanders, Belgium', *Transportation Research Part A: Policy and Practice* **125**, 268–279.

GeoPandas Developers (2019), *GeoPandas*. Version 0.7.0. https://geopandas.org.

Ghoseiri, K. (2012), Dynamic rideshare optimized matching problem, PhD thesis, University of Maryland.

Ghoseiri, K., Haghani, A., Hamed, M. et al. (2010), Real-time rideshare matching problem, Technical report, Mid-Atlantic Universities Transportation Center.

Guidotti, R., Nanni, M., Rinzivillo, S., Pedreschi, D. & Giannotti, F. (2017), 'Never drive alone: boosting carpooling with network analysis', *Information Systems* **64**, 237–257.

INSEE (2018), *Mobilités professionnelles en 2015 : déplacements domicile - lieu de travail*, National Institute of Statistics and Economic Studies [INSEE], France. In French. https://www.insee.fr/fr/statistiques/3566477.

Lee, D. W. & Liang, S. H. L. (2011), Crowd-sourced carpool recommendation based on simple and efficient trajectory grouping, *in* 'Proceedings of the 4th ACM SIGSPATIAL International Workshop on Computational Transportation Science', pp. 12–17.

Li, X., Hu, S., Fan, W. & Deng, K. (2018), 'Modeling an enhanced ridesharing system with meet points and time windows', *PLOS ONE* **13**, 1–19.

McShane, W. R. & Roess, R. P. (1990), *Traffic Engineering*, Prentice-Hall.

OpenStreetMap contributors (2019), 'OpenStreetMap', https://www.openstreetmap.osm.

Pebesma, E. (2018), 'Simple Features for R: Standardized Support for Spatial Vector Data', *The R Journal* **10**, 439–446.

PostGIS Project Steering Committee (2019), *PostGIS: Spatial and Geographic objects for PostgreSQL*. Version 3.0. https://postgis.net.

QGIS Development Team (2019), *QGIS Geographic Information System*, Open Source Geospatial Foundation Project. Version 3.8. http://qgis.osgeo.org.

Schreieck, M., Safetli, H., Siddiqui, S. A., Pflügler, C., Wiesche, M. & Krcmar, H. (2016), 'A matching algorithm for dynamic ridesharing', *Transportation Research Procedia* **19**, 272–285.

Smith, B. L. & Demetsky, M. J. (1997), 'Traffic flow forecasting: comparison of modeling approaches', *Journal of transportation engineering* **123**(4), 261–266.

Stiglic, M., Agatz, N., Savelsbergh, M. & Gradisar, M. (2015), 'The benefits of meeting points in ride-sharing systems', *Transportation Research Part B: Methodological* **82**, 36–53.

Tang, J., Song, Y., Miller, H. J. & Zhou, X. (2016), 'Estimating the most likely space–time paths, dwell times and path uncertainties from vehicle trajectory data: A time geographic method', *Transportation Research Part C: Emerging Technologies* **66**, 176 – 194.

Tiakas, E., Papadopoulos, A., Nanopoulos, A., Manolopoulos, Y., Stojanovic, D. & Djordjevic-Kajan, S. (2009), 'Searching for similar trajectories in spatial networks', *Journal of Systems and Software* **82**, 772 – 788.

TomTom (2019), *Routing API and Extended Routing API*. https://developer.tomtom.com/routing-api.

Uber (2019), 'What are suggested pickup locations?'. https://help.uber.com/riders/article/what-are-suggested-pickup-locations?nodeId=9edf05bf-ac3a-4cf8-b08e-76e9ca767f7f.

van der Wel, C. (2019), *PyGEOS*. Version 0.6. https://pypi.org/project/pygeos.

Wang, H. & Yang, H. (2019), 'Ridesourcing systems: A framework and review', *Transportation Research Part B: Methodological* **129**, 122–155.

Zhu, D. (2017), 'More generous for small favour? Exploring the role of monetary and pro-social incentives of daily ride sharing using a field experiment in rural Île-de-France', *DigiWorld Economic Journal* **108**, 77–97.

Zhu, D. (2018), The limit of money in daily ridesharing: Evidence from a field experiment, Technical report, PSL, University of Paris-Dauphine.