Mining Spatial Data from GPS Traces for Automated Map Generation

Fernanda Lima[1,2], **Michel Ferreira**[1,2]

[1] LIACC, Artificial Intelligence and Computer Science Laboratory, Universidade do Porto, Portugal
[2] Departamento de Ciência de Computadores, Universidade do Porto, Portugal
{flins,michel}@dcc.fc.up.pt

Abstract

GPS receivers have become ubiquitous devices. They are present in PDAs, cell phones, jogging watches or even key holders. Virtually every one of these devices is able to log the received data to a file, producing a huge amount of spatial information. As is known, map generation, refinement and update are very expensive if done in the conventional way. With the constant bandwidth increase in wireless networks, tracking companies are now transmitting position reports "in-raw", i.e. as received by the on-board GPS unit, with a point every second. Such detail allows processing the tracking information from a large number of vehicles to produce vectorial maps of the road network, in an inexpensive, accurate and permanently up-to-date manner. In this paper we describe an algorithm for the automated generation of highly detailed and accurate vectorial road maps from GPS traces. Our algorithm is implemented using spatial SQL queries to aggregate data from multiple traces to produce a weighted-mean geometry of road axles, diluting GPS errors. Further layers are possible to extract from the GPS data, such as traffic-lights location, parking information, road classification and points of interest of a particular vehicle. Using our algorithm, we have produced vectorial road maps from two very different portuguese counties, Arganil, a low-populated rural area, and Porto, an urban area comprising the second largest city in Portugal. Our results show a highly accurate overlapping with existing maps in all areas where a sufficient number of GPS traces have been collected.

1. Introduction

Portable GPS receivers have become more integrated in our daily lives. Many of them are installed in vehicles and are coupled to a communication device, such as a GPRS modem or a wi-fi network adapter. Such hardware combination is the basis of real-time tracking services, widely used by trucking companies and rapidly expanding to all types of vehicle fleets. Depending on the type of receiver and certain other conditions, such as the number of buildings in the neighbourhood of the vehicle, it is possible to achieve real-time position accuracies within meters or even centimetres. Even for off-line devices, devoided of any communication technology, the ability to log the received data to a file is available in virtually every one of these GPS receivers, producing a huge amount of spatial information that can later be consulted and uploaded to a website. In addition to the proliferation of in-vehicle GPS receivers, wireless technology is also advancing very rapidly, offering many attractive advantages over traditional wired networks, such as mobility, flexibility,

scalability and reduced long-term cost in rapidly changing environments. A visible metric of this evolution is the dramatically falling *cost per byte* over the wireless medium in the past years. As a result, tracking companies are now able to transmit position reports *"in-raw"*, i.e. as received by the on-board GPS unit, with a point every second. This creates a large-scale online network of moving GPS receivers, enabling the back-end server to store valuable spatial information. This information can be used for several purposes, including the generation of real-time traffic reports, the analysis of mobility patterns or, as we address in this paper, the automated construction and updating of vectorial road maps. Such inexpensive and automated construction is particularly important for developing countries, where such information is incomplete or non-existent. Even in highly industrialized nations it is of paramount importance to have inexpensive methods of keeping road maps permanently up-to-date. A good example is currently found in Beijing, which for the last couple of years has had to update its city map every three months.

Several projects have been developed in order to take advantage of spatial data produced by GPS devices. Most of them allow users to create their own vectorial maps from GPS traces in an assisted way. One of the most important projects developed in this research area is the well known OpenStreetMap. It is a collaborative project to create free editable maps using data from portable GPS devices and other free sources. The idea of generating vectorial road maps from GPS traces is not new but very little has been reported about the map generation without the requirement of an initial input map (BrÄuntrup et al, 2005). Earlier works have focus on systems that refine and update available maps (Rogers et al, 1999; Schroedl et al, 2004). In this paper we describe an algorithm for the automated generation of highly detailed and accurate vectorial road maps from GPS traces without requiring any geometric editing from the users and any available base map. Our input data is obtained from vehicles equipped with a GPS unit travelling through their usual routes. The paper is organized as follows: the next section describes the process of collecting and filtering GPS data. Section 3 presents the map generation algorithm. Section 4 reports the results of experiments with the automated map generation process. Section 5 introduces a technique for inferring road classification. We end by outlining some conclusions.

2. Data Collection and Filtering

Data collection and filtering are very important stages in the map generation process. If done in a correct way, they can guarantee the quality of the input data. Our data set consists of offline GPS logs, in NMEA format, and compact real-time logs from a tracking company, both with a one point per second detail. Each fix position is given as longitude, latitude and altitude. GPS receivers also provide some more important information like the Horizontal Dilution of Precision (HDOP), the number of satellites being tracked, the speed over the ground and the track angle.

In order to accurately represent road information, a great number of GPS traces is required. To evaluate our algorithm we have used two very different portuguese counties, Arganil, a low-populated rural area, and Porto, an urban area comprising the second largest city in Portugal. A total of 12.603,014 tracked kilometres have been used for the generation of Arganil map, while 3.485,119 tracked kilometres have been used for the Porto map.

Unfortunately, GPS data is not precise due to uncertainty in the location fixes when the GPS satellite signal is obscured. In order to guarantee the integrity of our data set we first have to extract and filter the relevant data from GPS logs. We assume that points collected with a very low speed (less than 6 km/h) are not accurate enough for map generation. So 15.31% of the collected points were discarded. The velocity distribution of our data set is shown in Fig. 1.

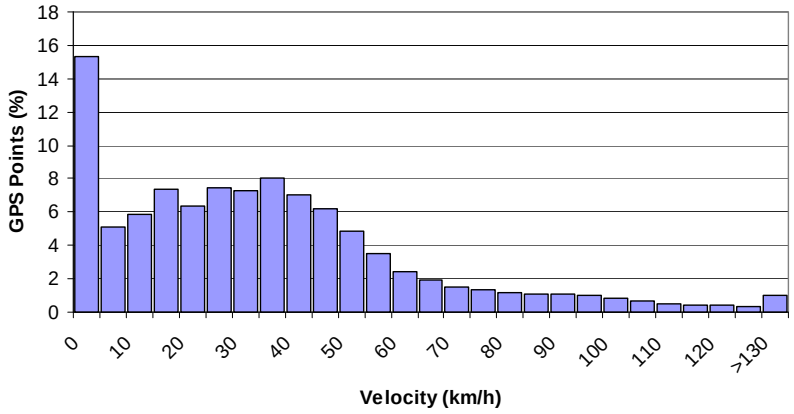


Figure 1. Velocity distribution

Gaps in the GPS receiver signal can be another error source for road representation. GPS traces with gaps greater than seven seconds between two adjacent measured points had to be split into two or more different traces in the pre-processing stage of our work. We have used two more kinds of filters based on the number of satellites used to calculate the coordinates of a point and the HDOP value. Then we have applied a line simplification algorithm in the pre-processed GPS Traces. Results are shown in the following subsections.

2.1 HDOP Filter

HDOP is a parameter that allows to more precisely estimate the accuracy of GPS horizontal (latitude/longitude) position fixes by adjusting the error estimates according to the geometry of the tracked satellites. A low HDOP value represents a better GPS positional accuracy due to the wider angular separation between the satellites used to calculate positions. The HDOP distribution of our data set is shown in Fig. 2.

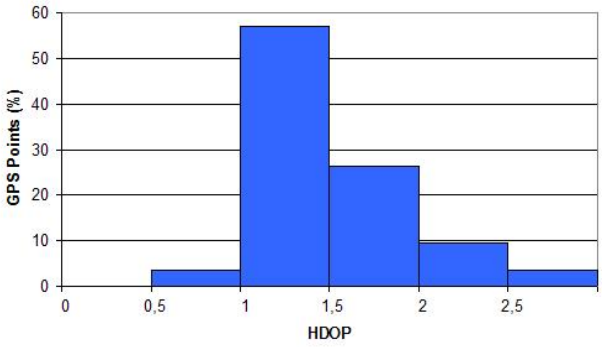


Figure 2. HDOP distribution

In our HDOP filter, points that have a Horizontal Dilution of Precision higher than 2 were discarded. The results obtained with this filter are shown in Table 1. It is possible to realize that HDOP value is lower in rural areas like Arganil, and higher in urban areas like Porto.

Table 1. HDOP filter

Tracked Area	HDOP Average	Discarded Points
Arganil	1.504	11,34 %
Porto	1.663	22,48 %

2.2 Filtering based on the number of satellites been tracked

The number of tracked satellites has a large influence on the strength of the satellite configuration for positioning. When more than four satellites are tracked, the redundant satellites can be used to detect erroneous measurements, facilitating data refinement. The GPS Points distribution according to the number of tracked satellites is shown in Fig. 3. Our filter has discarded points that have been fixed by less than 5 satellites. Results are shown in Table 2.

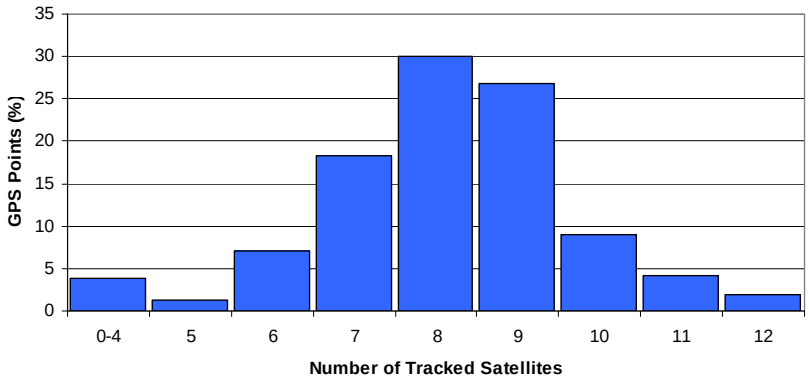


Figure 3. GPS Points distribution according to the number of tracked satellites

Table 2. Filtering based on the number of satellites been tracked

Tracked Area	Tracked Satellites Average	Discarded Points
Arganil	8	3,92 %
Porto	7	4.18 %

2.3 Line Simplification

Line simplification is a map generalization technique that replaces a linear map feature with a less complex representation of the same feature. It is an important step of our algorithm that discards a significant percentage of the points collected by the GPS units. It improves the performance of our map generation process and minimizes the required memory space. In this work, GPS traces are simplified by Douglas-Peucker algorithm that is the accepted standard for simplifying polylines. The classic Douglas-Peucker polyline simplification algorithm (Douglas and Peucker, 1973) selects

a subset of vertices from which it builds a new simplified polyline which lies within a predefined distance of the original polyline. The results obtained are shown in Tables 3 and 4.

Table 3. Percentage of eliminated points with Douglas-Peucker line simplification algorithm

	GPS Traces	Simplified Traces	Eliminated Points
Total Number of Points	2.324.624	255.709	89%

Table 4. Elimination error

	GPS Traces	Simplified Traces	Elimination error
Km tracked	24.946	24.809	0,55%

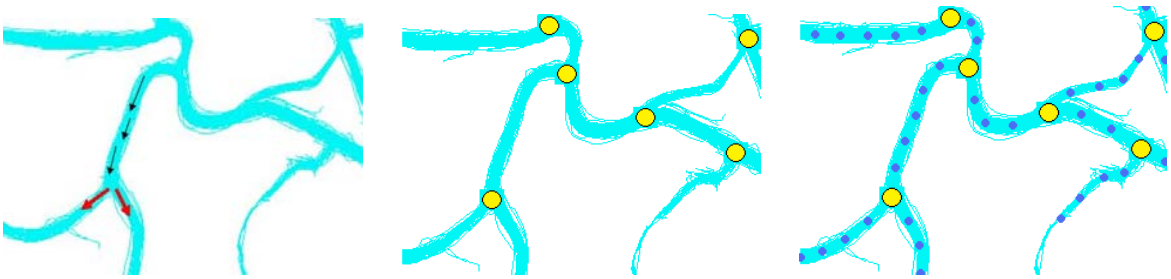
3. Map Generation Algorithm

The main purpose of our map generation algorithm is to construct a directed graph where edges represent road segments and nodes represent junctions. In order to guarantee the quality of our map, each road of the tracked area must be covered by multiple GPS traces. Our algorithm is implemented using spatial SQL queries to aggregate data from multiple traces to produce a weighted-mean geometry of road axles, diluting GPS errors. The first step of our algorithm is to identify intersection areas. To capture the places where an intersection occurs we have developed a method that tries to find points of a single trace around which parallel traces diverge from each other. After that, we can define an intersection node by calculating the average of the coordinates of GPS points that are within an intersection area. Then, we are able to apply a modified clustering algorithm for each set of GPS points situated between two intersection nodes. Clustering is a discovering process that groups a set of data objects in a way that maximises the similarity between the objects inside a cluster, and minimise the similarity between different clusters (Kaufman and Rousseeuw, 1990). It is considered a data mining technique and an unsupervised learning technique since the user has no influence in the discovery process. Our modified clustering algorithm adds k clusters with a constant distance from each other. It groups points according to their similarity to a cluster center that is updated when new points are inserted in the cluster. In order to calculate the cluster _centers coordinates we can use a simple SQL query like:

```
SELECT SUM(ST_X(p.geometry))/COUNT(p.point_id) AS x, SUM(ST_Y(p. geometry))/COUNT(p.point_id) AS y,
SUM(p.velelocity)/COUNT(p.point_id) AS velocity FROM points p, cluster_centers c
WHERE ST_DWithin(p. geometry, c. geometry, d)  GROUP BY c.cluster_center_id
```

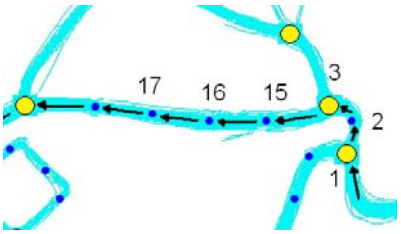
We assume that a cluster center is one point of the road centreline which can be viewed as a weighted average trace. Figure 4 shows these three steps of our map generation algorithm.

Besides the road axle geometry layer, our algorithm also produces a topological connectivity layer of the road axles intersections. It is done by an adapted map-matching module that does not require an initial base map. The connectivity between two adjacent points of a road centerline is established when these points are reached by a GPS trace. Each adjacent cluster center should be linked to each other cluster center for which connecting traces exist (see Fig. 5). The topological connectivity layer describes traffic rules, such as the driving direction of a road segment or the allowed turns at each intersection.



(a) (b) (c)

Figure 4. Steps of the map generation algorithm: (a)Intersection areas identification, (b)Intersection nodes definition, (c)Cluster centers inference



CONNECTIVITY		
Id1	Id2	Number_of_parallel_traces
1	2	54
2	3	54
3	15	47
15	16	47
16	17	47

Figure 5. Topological connectivity establishment

The input of our algorithm is the GPS data that is stored in three tables of our spatial database (vehicles, traces and points). The output of our algorithm are two more tables in our spatial database, containing cluster centers geometry and the topological connectivity between them with an attribute that represents the number of traces used to establish the connectivity (see Fig. 6).

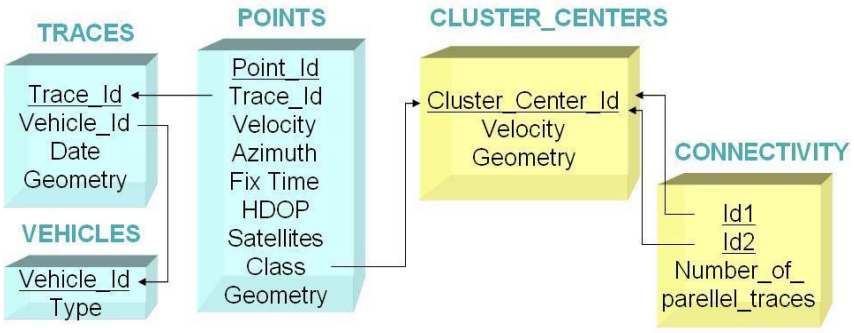
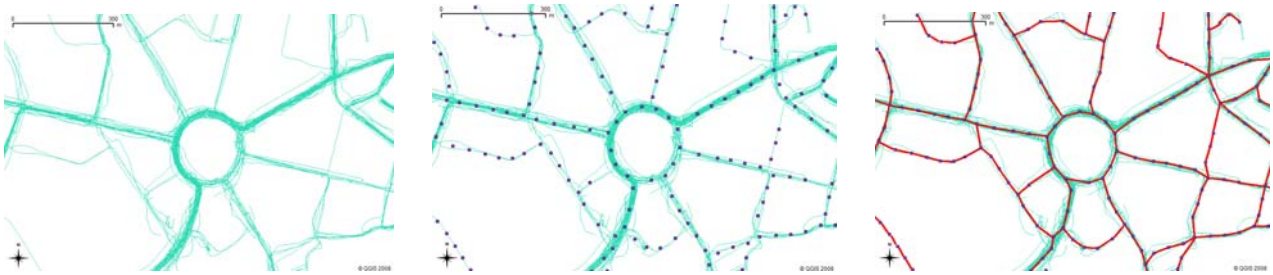


Figure 6. Spatial Database Structure

4. Experimental Results

In this section, we describe the results we obtained when applying our map generation algorithm. Using our algorithm, we have produced a vectorial map from the cities of Arganil and Porto, the last one being the second largest in Portugal, and compared their geometric and topological layers with vectorial maps produced by the City Hall and maps from Google. We do not infer lane-precise maps due to the lack of more precise differential GPS information. However our results show a highly accurate overlapping between the maps in all areas where a sufficient number of GPS traces have been collected. Figure 7 shows the directed graph created by our algorithm and Figure 8 shows an overlapping area between our generated map and a map from Google.



(a) (b) (c)

Figure 7. Directed graph generation: (a)GPS logs, (b)Clustering, (c)Topological connectivity

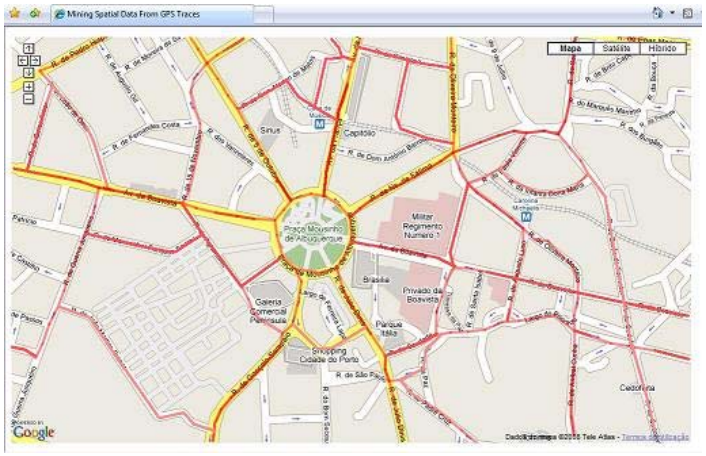


Figure 8. Overlapping area between our generated map and a map from Google

6. Road Classification

GPS logs provide much more information than just the geometry conveyed in latitude and longitude coordinates. An useful example of such information is the travelling speed of the tracked vehicle that, if used in real-time, can provide the basis for real-time traffic information systems. Figure 9 shows a speed plot of a tracked vehicle in the area of Porto.

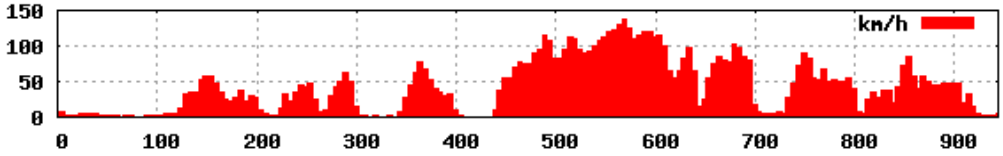


Figure 9. Speed plot of a tracked vehicle in the area of Porto

We have used such speed information to derive an automatic classification of roads. Has it can be seen from Fig. 9, there are clear areas with very different speed patterns. A spatial aggregation of such speed data from multiple logs is able to derive quite accurately the road classification type based on such speed information. Figure 10 shows our experimental results over the data collected in Porto, where we clearly identify all the highways around the city (in red), together with the structural roads inside the city (blue and green). Looking at Fig. 9 we can also identify momentary stops in the urban travel of the vehicle. Again, the spatial aggregation of several GPS logs and their respective speed patterns should be able to derive the location of traffic lights based on the aggregation of this information.

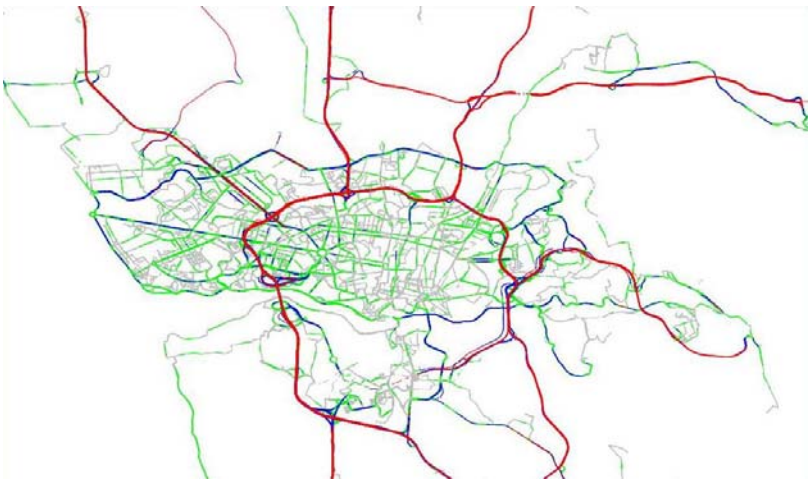


Figure 10. Road classification inference

7. Conclusions and Future Work

In this paper we have described a clustering algorithm to automatically generate a vectorial road map from GPS traces. The important advantages of our approach are that no initial map is needed, we can infer maps even in unknown terrains and new road segments can be integrated to an already generated map. Using our algorithm, a web site where users can upload GPS logs in standard formats, such as NMEA, is able to provide open and free vectorial road maps for download, without requiring any geometric editing from the users. This study focused on the map generation and road classification processes, but data mining over position traces can yield more types of knowledge, such as traffic-lights location, parking information and points of interest of a particular vehicle.

Acknowledgments

This work has been partially supported by the Portuguese Foundation for Science and Tecnology (FCT) under project JEDI (PTDC/EIA/66924/2006) and by funds granted to LIACC through the Programa de Financiamento Plurianual and POSC.

References

BrÄuntrup, R, Edelkamp, S, Jabbar, S, & Scholz, B 2005, 'Incremental map generation with GPS Traces', *Proceedings of IEEE Intelligent Transportation Systems*, Vienna, Austria.

Douglas, DH & Peucker, TK 1973, 'Algorithms for the Reduction of the Number of Points Required to Represent a Line or its Caricature',*The Canadian Cartographer*, vol.10, pp.112-122.

Kaufman, L & Rousseeuw, PJ 1990, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons Inc., New York.

Rogers, S, Langley, P & Wilson, C 1999, 'Mining gps data to augment road models', *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA: ACM Press, pp. 104–113.

Schroedl, S, Wagstaff, K, Rogers, S, Langley, P & Wilson, C 2004, 'Mining GPS traces for map refinement', *Knowledge Discovery and Data Mining*, vol. 9, pp. 59-87.