Big Data Continual Assessment

Analyzing and visualizing google location history

Manuel Fidalgo Fierro

Institute of technology Tralee, Co. Kerry, Ireland

Abstract

In this paper, I analyze the GPS data obtained from an android device during one year and a half. The aim of this paper is to determine whether the kmeans algorithm is suitable to divide locations into countries. I also create a predictive model to see if it's possible to predict timestamps for a given location. This data is presented in a map of Europe where I represent the flights taken during this period, the clusters previously calculated and the errors detected in the GPS analyzing the speed between locations.

Keywords: location; history; google; takeout;

Nowadays everyone has one device in his pocket at all the times. This device is continually recording different types of data such as voice, accelerometer data, network status, locations etc.

The GPS data in android mobiles is recorded by google, which stores this data in its servers and uses the information to enhance the advertisement service. The company provides to the user the opportunity to see and delete its data through the page https://takeout.google.com/settings/takeout. Here the user can download many types of information, but in this case, it's needed the relative to the location service. This file is a JSON composed by the following fields

Field	Example value	Detail
timestampMs	150444868XXXX	The number of millisecond from epoch
latitudeE7	42558XXXX	The latitude in E7 format
longitudeE7	-5590XXXX	The longitude in E7 format
accuracy	20	Estimated precision
activity	STILL	Contains information relative to the activity in this timestamp

Table 1, JSON description

1. Literature review

1.1. How google extracts patterns from location history

Google compute the user’s commonly visited places (including home and work) and commute patterns. They use the google latitude history dashboard. A user's location history can be used to provide several useful services. They can cluster the points to determine where he frequents and how much time he spends at each place. The raw data taken from the phones may contains errors due to software or hardware bugs. Google applies some filter in order to solve this problem.

- Reject any points that fall outside the boundaries of international time zones over land.
- Reject any points with timestamps before the known public launch of the collection software.
- Identify cases of “jitter”, where the reported location jumps to a distant point and soon returns.
- If a pair of consecutive points implies a non-physical velocity, reject the later one.

(Andrew, Tushar, & Pablo, 2017)

1.2 Jitter, the concept

Jitter is defined as a variation in the delay of received packets. At the sending side, packets are sent in a continuous stream with the packets spaced evenly apart. Due to network congestion, improper queuing, or configuration errors, this steady stream can become lumpy, or the delay between each packet can vary instead of remaining constant. (https://www.cisco.com, 2016)

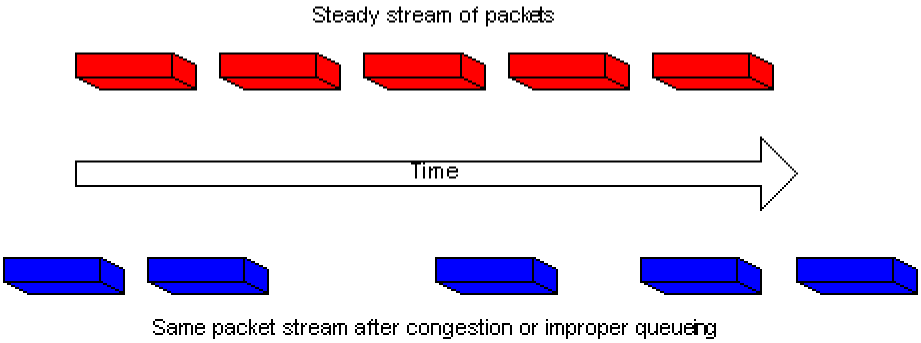

Figure 1. Jitter representation

1.3 Identifying jitter algorithm

Google algorithm to identity jitter. Having a list of consecutive positions {P1, P2, …, Pn} where the following conditions hold:

- P1 and Pn are within a small distance threshold D of each other.
- P1 and Pn have timestamps within a few hours of each other.
- P1 and Pn have high reported accuracy.
- P2, …, Pn-1 have low reported accuracy.
- P2, …, Pn-1 are farther than D from P1.
-

In such a case, we conclude that the points P2, …, Pn-1
are due to jitter, and discard them

1.4 GPS issues when flights are analyzed

- GPS location can be inaccurate. Not always spot-on. We could filter out inaccurate data points, but GPS doesn't have to be that far off to break our criteria. Think about it - if we're sampling location once per minute then all it would take is to be off by 200kph/60min or 3.3 km *(Hartley, 2014)*

- Your phone collects GPS data mid-flight. The airplane mode deactivates WIFI and cellular data, but the smartphone is sending and receiving GPS data. *(Hartley, 2014)*

- Using speed assumes no delays. It's possible to turn off the phone for a flight only to sit on the tarmac for 2 hours. Or, the flight may be in a holding pattern before landing. Either scenario would dramatically decrease my computed "speed" by artificially increasing the time between airplane-mode on and off *Section headings (Hartley, 2014)*

1.5 Earth distances

In order con calculate the distance between two locations it's necessary to use the "haversine" formula, wich calculated the grat-circle distance between two points in the surface of a sphere. *(de Mendoza y Ríos, 1795)*

$$a = sin^2(\Delta\varphi/2) + cos\,\varphi 1 \cdot cos\,\varphi 2 \cdot sin^2(\Delta\lambda/2)$$

$$c = 2 \cdot atan2(\sqrt{}a, \sqrt{}(1-a))$$

$$d = R \cdot c$$

Where φ is latitude and λ is longitude, both in radians. R is earth's radius (mean radius = 6,371km);

2. Results

The accuracy variates with the time, not in every moment the accuracy it's the same, but during this period of time and after removing the outliers the average of the accuracy is 23.435 m and the median and the mode is 20 m.

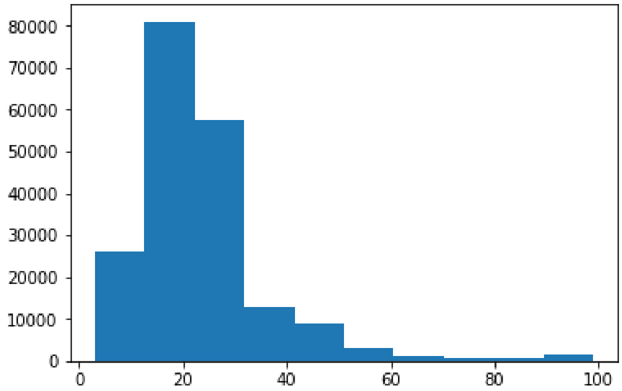

Figure 3.1 Accuracy histogram.

The aim of using clustering in this project is to determine the grade of correlation between the clusters generated by kmeans algorithm and the countries. As we can observe in the plots below, using 5 clusters the algorithm divides between locations in Spain and locations in Portugal but it does not recognize the difference between Hungary and Romania. I was not expecting this result because the distance between Romania and Hungary is higher than Spain a Portugal. This might be caused because there are more locations in the Iberian Peninsula than in any other region.

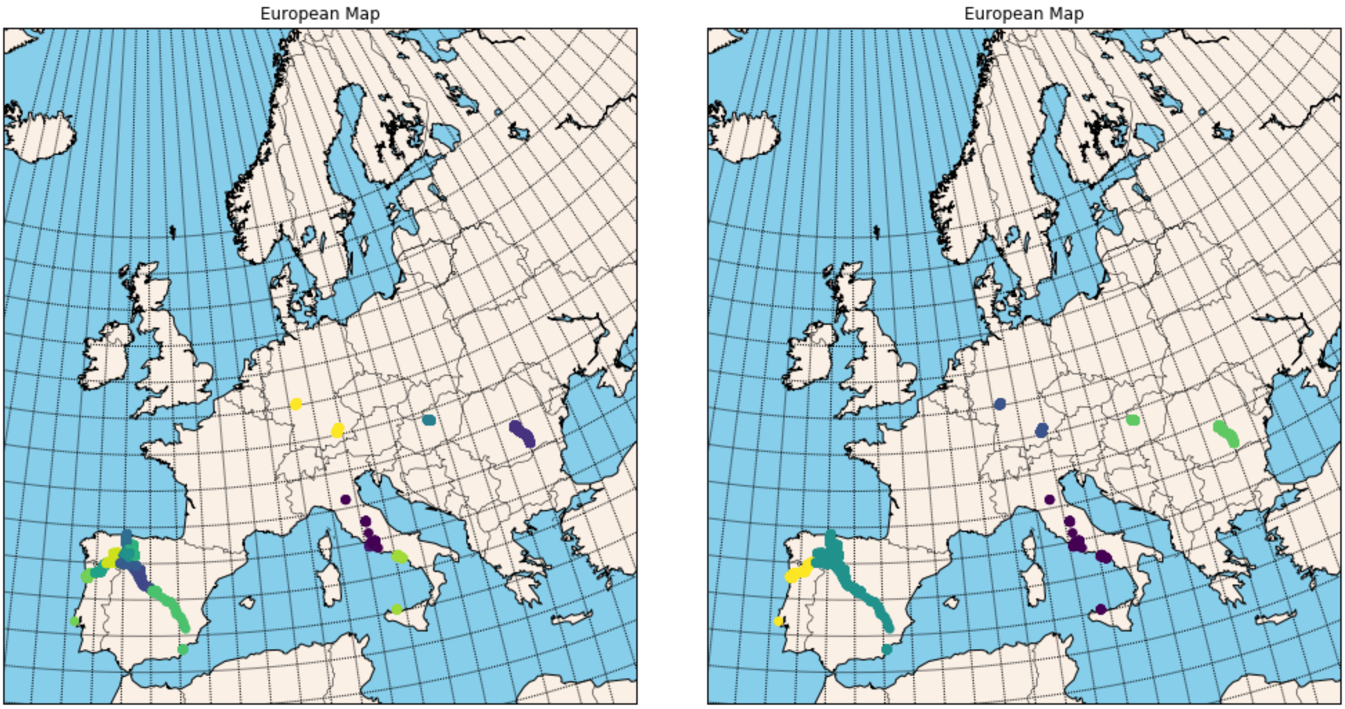

Figure 3.2 (a) 15 clusters (b) 5 clusters

In order to display the flighs taken and due to the previous research, The algorithm used to extract the flights is based on the distances instead of the speed. In the example blow, there is a representation of the flights taken during this period of time. As a curiosity, there is an error in this map because I’ve never taken a flight from Porto to Munich.

According to lit review, the gps data may contain errors, so the second image represents the errors based on the speed between locations. It’s observable that there are many errors around all the map, and there are two locations, (Sicilia and Lisbon) where I’ve never been.

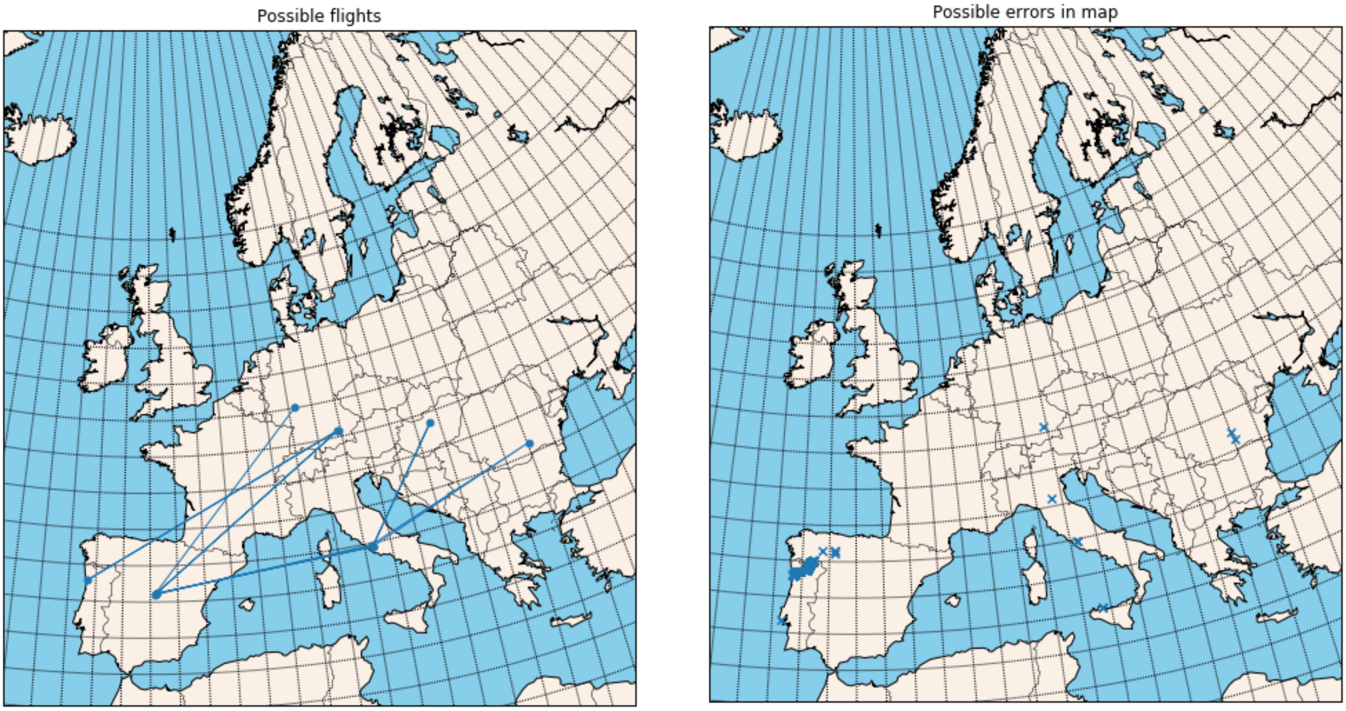

Figure 3.3 (a) Flights (b) GPS errors

The predictive model.
It’s used a stats model to predict a timestamp having the latitude and the longitude, the values of the coefficients are

R-squared:	0.109
Adj. R-squared:	0.109
F-statistic:	1.190e+04
Prob (F-statistic):	0.00

The F-static in high but the rest are really low values
In the plot below is shown the difference between the model and the original dataset, concluding that this linear regression is not valid and It’s not possible to predict a timestamp from a location.

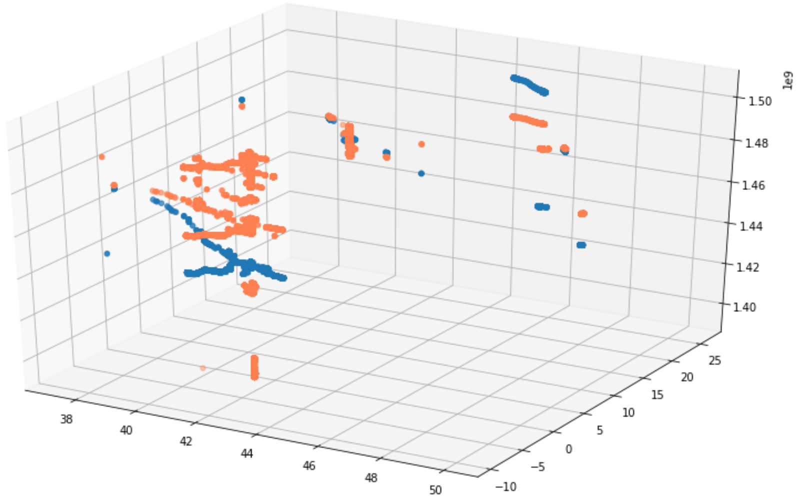

Figure 3.4 Predictive model. In blue the dataset, in coral the predictive model

3. Conclusion

After this study I've discover tons of error in the GPS tracking system. I cleaned the data because the original data has altitudes under -1000 m and over 5000m, I've also discover that the gps sometimes "jumps" and shows fake locations. It's possible to use the dataset to provide information about the flights, but the algorithm used is not the most accurate because it shows fake flights. This study is really interesting because provides an interactive experience with GPS data. This assessment also has a deep learning component of geometry, data mining and statistical concepts.

References

Andrew, K., Tushar, U., & Pablo, B. (2017). *Extracting Patterns from Location History*. Mountain View, California: google.inc.

de Mendoza y Ríos, J. (1795). *Memoria sobre algunos métodos nuevos de calcular la longitud por las distancias lunares: y aplication de su teórica á la solucion de otros problemas de navegacion.* Madrid.

https://www.cisco.com. (2016, February 2). Retrieved from https://www.cisco.com: https://www.cisco.com/c/en/us/support/docs/voice/voice-quality/18902-jitter-packet-voice.html