Understanding the buses delay caused by the geometry of the roads using unsupervised learning

Muhammad Naeem[a], Mehdi Katranji[b*], Guilhem Sanmarty[b], Sami Kraiem[b],
Mahdi Zargayouna[c] and Fouad Hadj Selem[b]

[a] VEDECOM, 23 bis Allée des Marronniers, 78000 Versailles, France
[b] Entropy, 23 bis Allée des Marronniers, 78000 Versailles, France
FirstName.LastName@entropy.sc
[c] Univ Gustave Eiffel, COSYS-GRETTIA, 14 Boulevard Newton, 77420 Champs-sur-Marne, France
mahdi.zargayouna@univ-eiffel.fr
* Corresponding author

Extended abstract submitted for presentation at the 11th Triennial Symposium on Transportation Analysis conference (TRISTAN XI)
June 19-25, 2022, Mauritius Island

April 5, 2022

Keywords: performance, indicator, path analysis, public transport, pointwise delays

1 INTRODUCTION

Public transport is a part and parcel of urban life, and the improvement in the quality of public transport plays an essential role in the quality of urban mobility. Limiting delays in the public transport is at the heart of quality control metrics. Indeed, a key parameter useful to observe the bus lines quality of service is the accurate estimation of time delay. Limiting delays assists in maintaining an adequate service quality with a reasonable certainty. Low reliability, slow speed and increased load of passengers accumulate towards a situation which ends up in a critical situation for the bus operator (Ryus, 2003). Determining the delay or early arrival at a bus station may not be too complex with the current technological environment. The continuous monitoring over a long period of time enables to provide an immediate analysis to transport operators of delays or early arrivals at a specific bus station. However, public transport operators are interested beyond this point. There are numerous types of delays: delays caused by traffic congestion, passenger boarding, bus station stop, road delays resulted by specific geometry of the route. It is very important for the operators to find the implicit root cause for a delay at a single or many bus stations other than the obvious reasons at designated bus station. It was reported that on average 60% of travel time in public transport in Beijing China is claimed by delays at intersection (Zhang *et al.*, 2011) and (Liu *et al.*, 2013). It is an interesting problem to find the correlation lurking between the point geometry and absolute delay status at that point independent of its previous delay (or relative delay). This interesting problem turns into a challenge when the bus data is insufficient to explicitly calculate the relationship between point geometry and timing given a certain number of conditions such as busy or idle hours. In this paper, we investigate this correlation with a data-oriented approach. In the following sections, we present our methodology and results before to conclude.

2 METHODOLOGY

2.1 Data and main difficulty

The input data of our algorithm is made of: i) The map of the territory (e.g. OSM) ii) the GPS tracking data of several buses on a local network. Each bus produces the GPS coordinates every nine to twenty seconds; iii) the theoretical transit time of each bus stop, which refers to the time usually required by a bus to travel between two points. Our objective is to identify the points on the network where the bus concedes a delay that is uniquely related to the geometry of this part of the road. It is important to note that this is not a cumulative delay since the beginning of the trips but purely related to the local geometry. In order to achieve this, it is necessary to have enough crossing points on a short stretch of each road so that the statistical power is sufficient. By observing the tracking data, we noticed a non-synchronization of the recording points of the GPS tracks, so that the cumulative points form almost a continuous cloud of points obtained by several different buses, for tracks in different directions. Trips in different directions are really problematic when the delay must be calculated separately for each direction. This is not a simple partition between a return trip of each line. Indeed, the same point can be used by several buses and in different directions (outbound and return at the same time). To obtain a homogeneous partition of these points, one can think of a simple clustering but in fact, such an approach would not respect the shape and topology of the road network and will mix points on different roads. Hence, we propose a topology based on the cumulative distance traveled by each bus from a starting reference point. With a clustering on this new variable, we obtain homogeneous clusters, i.e. allowing to measure a delay purely related to the geometry. This will be explained in the next section.

2.2 Main steps

The vehicle profile is able to estimate the intersection delays whether they are “running time delays”, “acceleration”, “deceleration” or “stop delays”. For a single vehicle, this is possible by examining its route profile. However, on single route, different vehicles face different situations over a range of various time slots. Hence, a single route is not sufficient to provide strong estimation as it is not consistent enough to consider for a long period of time.

Given a bus trajectory along with its start and end station configured under particular timing schedule over many days, there are n number of points in between the two stations such that a route profile can be defined as a function of set of GPS points P, various time delay and speed profile sp. The objective is to mark the GPS sequence that is potentially responsible for substantial delay. To do so, we have devised the algorithm described in the following. First, the distance between two consecutive points is calculated. This helps in finding the cumulative distance. Then mark out the “out of route” GPS coordinates resulted in map-matching[1]. Then we calculate the distance and the trajectory between two points. If the distance between two points is too large then it means the GPS coordinates need to be shredded. This way, we remove the outlier or incorrectly recorded GPS points especially on circular paths and crossing points to avoid ambiguity. The algorithm can be summarized into two steps, corresponding to two sub-problems as follows:

- Cluster the n points into m clusters.
- Perform statistics on each cluster to identify whether the delay in this cluster is significant or not.

While formulating the cluster, we have available attributes including place names, original GPS points, distance between two GPS points, and time required to move between two consecutive GPS points. The time was originally available for a complete trajectory. However, our

[1] We use OSMnx to find the nearest point available on the road

purpose is to analyse the route over large number of days for a given range of time. Once the bus is passed by a GPS coordinate, it is not guaranteed that the bus will pass by this point again in its next trajectory. This means we need to cluster the nearby points. In this context, we use the cumulative distance as an input parameter for clustering. The rationale behind the choice of this parameter is that the k-means algorithm is based on Euclidean distance. If we use the "point to point" distance or "point to point" time then the cluster will be composed of random points from the trajectory: the nearest distance points will fall under same cluster. Such a cluster will have no practical value for interpretation. A continuous function is a good candidate for clustering. For this purpose, we derive a new feature of cumulative sum of the distance. As the bus starts from its route, its distance naturally increases from the starting point until the end point. This continuous function is also marked by sharp or minor additive value to increase the inter-cluster and decrease the intra-cluster distance. Moreover, it is a known fact that selecting appropriate value of k is a scientific problem for which techniques like Silhouette score are well known computational techniques to find optimized value of k. We adopt Silhouette score and BIC (Hamerly, 2002) in our methodology. After many experiments, we noticed that in overwhelming situations, BIC was more helpful in finding the exact value of k in k-means clustering. Finally, we statistically studied the significance of the delay at each accumulation point.

Note that the empirical validation carried out in this study is independent of any assumption of particular distribution (normal or non normal distribution). However, we know that the student t test is a parametric test. It means that we are supposed to assume that input in the t test follows a normal distribution. The samples with size greater than thirty are resilient in t test as even if there is a violation of normal distribution, the test remains efficient thanks to the CLT theorem. Hence the shape of the data has no effect in case of large sample size as the central limit theorem also supports the same assumption. However, it is important to ascertain the degree of deviation from the normality of the underlying data. For this purpose, we use the Shapiro-Wilk test. If the null hypothesis (assumption of data normality) is rejected in Shapiro-Wilk test then we adopt the wilcoxon test as a non parametric test.

3 RESULTS

3.1 Main observations

All our experiments have shown that the point delay accumulates around 23.6% of overall delay. Recall that Nielsen *et al.* (1998) pointed out that the intersection delays account for 17% to 35% of the total travel time . This is due to the fact that our proposed geometry point delay is more generalized to cover the intersection delay as well. Moreover our delay estimation explains to pinpoint specific paths and geographic coordinates with high importance in terms of non trivial delay. This new methodology is independent of other delays by neglecting out the effect of bus stop delays and is not tied to any particular distribution of data. It automatically examines the distribution of the data and then selects the parametric or non parametric test accordingly.

3.2 Example

We present an example in Figure 1. If we analyze the map, we can notice that before the roundabout, there is weak or minor evidence of significant delay because it is a straight section of the road unlike the traffic circle. However, beyond the roundabout, we identify significant delay. Therefore, it can be said that an unusual or unexpected delay is observed beyond the roundabout. A possible explanation is that there are some public offices such as a bank, a mini-mart and a medical center. The delay may be caused by the slow movement of passengers (getting off or on the bus), or by a crowd of passengers. However, in fact, there is no bus stop exactly at this location.

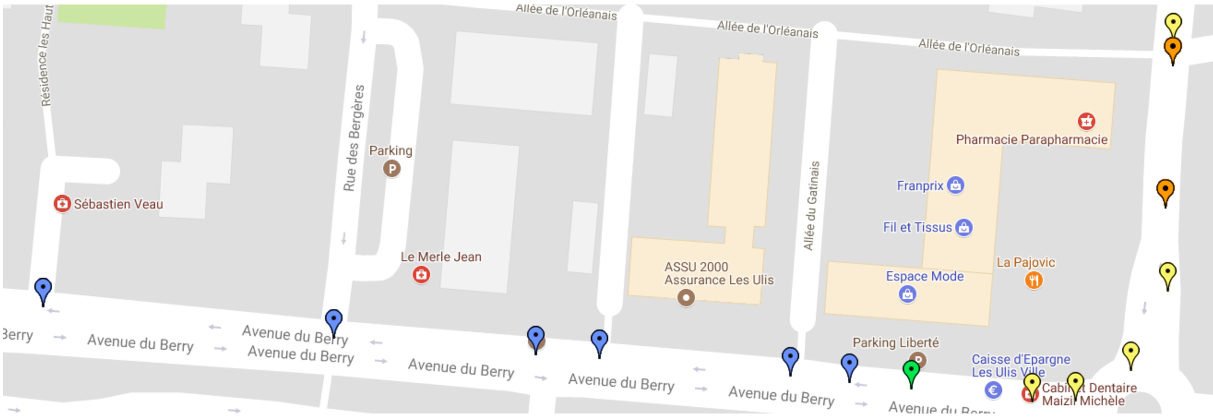

Figure 1 – *Marking of the points with respect to the significance of the delay :*
blue*: strong significant,* ***green****: significant,* ***yellow****: weak evidence,* ***orange****: little or no evidence*

4 CONCLUSION

This study investigates the reliability of travel time to build a new key performance indicator of public transport network. We have introduced a suitable point wise clustering followed by an adapted statistical significance analysis. The outcome is an estimation of absolute delay at geometrical points independent of the delay at bus station. This outcome serves as an incremental delay towards overall delay. Our investigation suggests that this novel metric of delay time contributes around 23.6% in overall delay. For the future extension, we propose to estimate the passengers leaving out the bus at any stop so this flow can also be accounted towards the explanation of the overall and point wise delay.

References

Hamerly, Greg. 2002. Learning the K in K-means. *Advances in neural information processing*, 281–288.

Liu, Xiliang, Lu, Feng, Zhang, Hengcai, & Qiu, Peiyuan. 2013. Intersection delay estimation from floating car data via principal curves: a case study on Beijing's road network. *Frontiers of earth science*, **7**(2), 206–216.

Nielsen, Otto Anker, Frederiksen, Rasmus Dyhr, & Simonsen, Nikolaj. 1998. Using expert system rules to establish data for intersections and turns in road networks. *International Transactions in Operational Research*, **5**(6), 569–581.

Ryus, Paul. 2003. *A Summary of TCRP Report 88– A Guidebook for Developing a Transit Performance-measurement System.* Transportation Research Board.

Zhang, Hengcai, Lu, Feng, Zhou, Liang, & Duan, Yingying. 2011. Computing turn delay in city road network with GPS collected trajectories. *Pages 45–52 of: Proceedings of the 2011 international workshop on Trajectory data mining and analysis.* ACM.