Which One is Correct, The Map or The GPS Trace

Abdeltawab Hendawi
University of Rhode Island
hendawi@uri.edu

Sree Sindhu Sabbineni
University of Washington, Tacoma
ssindhu@uw.edu

Jianwei Shen
University of Washington, Tacoma
sjwjames@uw.edu

Yaxiao Song
Microsoft Corporation
yasong@microsoft.com

Peiwei Cao
Microsoft Corporation
peiweic@microsoft.com

Zhihong Zhang
Microsoft Corporation
zhz@microsoft.com

John Krumm
Microsoft Research
jckrumm@microsoft.com

Mohamed Ali
University of Washington
mhali@uw.edu

ABSTRACT

GPS data is noisy by nature. A typical location-based service would start by filtering out the noise from the raw GPS points that are generated by moving objects. Once the locations of the objects are identified, the location-based service is provided. In this paper, we decide not to throw away the noise. Instead, we consider the noise as an asset. We analyze the various noise patterns under different conditions and region characteristics. More specifically, we focus on one example where a lot of GPS noise is experienced; which is urban canyons. We believe that learning the GPS noise patterns in a supervised environment enables us to discover knowledge about new areas or areas where we have little knowledge. This paper is based on the analysis of GPS traces that are collected from the shuttle service within the Microsoft campuses around Seattle, Washington.

CCS CONCEPTS

• **Information systems** → **Spatial-temporal systems**; **Location based services**.

KEYWORDS

Data Cleaning, GPS Traces, Trajectories, Map Visualization

ACM Reference Format:
Abdeltawab Hendawi, Sree Sindhu Sabbineni, Jianwei Shen, Yaxiao Song, Peiwei Cao, Zhihong Zhang, John Krumm, and Mohamed Ali. 2019. Which One is Correct, The Map or The GPS Trace. In *27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19), November 5–8, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3347146.3359099

1 INTRODUCTION

A road network is a core component in location-based services. Accuracy of the underlying road network graph intuitively affects the overall quality of services and applications on top of it. As GPS devices evolve overtime, it might be the time to revisit the existing maps, with road networks as an integral part of it. We revisit the maps to improve the precision of different elements in these road networks, i.e., node and edge locations. GPS trajectories can play an important role in fixing, and hence improving the quality of road networks and location-aware services in general [7–10]. However, sometimes it is not fully guaranteed that these GPS traces are themselves dependable. The outputs of GPS devices are affected by various factors such buildings, tunnels, trees, weather conditions and city architectures.

In a typical location-based service, noise is considered a negative undesirable matter imposed on the clean desired GPS signal. This perception definitely makes sense in a lot of scenarios because noise can distort the true location of the moving object that is asking for a service. Location-based services start by filtering out the noise from the raw acquired GPS locations [12] and, then, map-match these GPS locations to road segments in the road network graph. A variety of map-matching techniques have been proposed to snap the raw GPS locations to road segments [4–6]. These techniques utilize the geometry of the road segments and/or the past trajectories of the moving objects to maximize the likelihood a moving object is on a specific road segment. For example, the algorithm proposed in [13] utilizes a *Hidden Markov model* to provide a map matching technique that is resilient to low sampling rates and to noisy signals.

This paper takes a different approach looking into noise. We consider noise a valuable asset that may reveal interesting patterns related to the moving object, the driving conditions, and the surrounding environment. In this paper, we study the noise pattern of moving vehicles around tall buildings. The study in this paper is based on the GPS data collected from the Microsoft Shuttle Service that transports employees between buildings in the Greater Seattle area. The Greater Seattle area features variations in building heights and tree covers. The shuttle trips vary from short distance trips on inner city roads to long distance trips between far away campuses. The collected GPS points are filtered to only remove outliers but not noise. Outliers are points that are extremely infeasible or unreachable from the shuttle's previous location. Then, the GPS points are map-matched (or snapped) to road segments in the underlying road network graph using the Microsoft Maps *Snap-to-Road* API [2]. The Microsoft Snapping API is based on the technique presented in [13].

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGSPATIAL '19, November 5–8, 2019, Chicago, IL, USA
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6909-1/19/11.
https://doi.org/10.1145/3347146.3359099

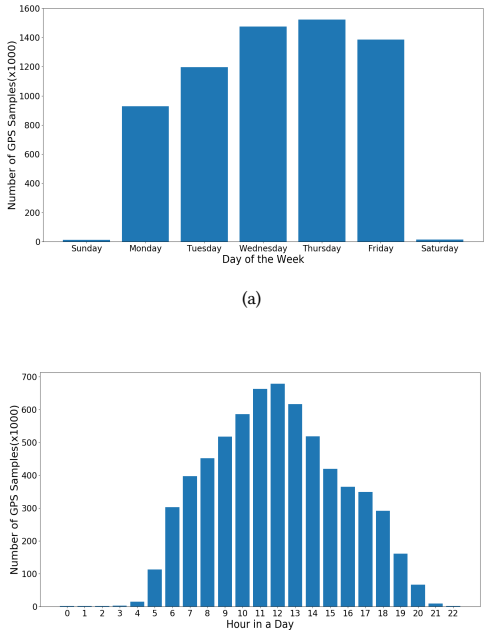


Figure 1: Distribution of GPS sample frequencies over (a) days of the week and (b) hours of the day.

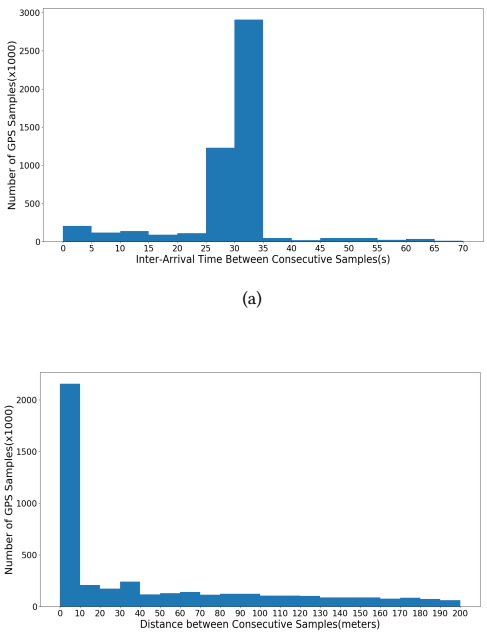


Figure 2: Distribution of the (a) inter-arrival time and (b) distance between consecutive GPS samples.

The distance between the raw GPS point and its snapped counterpart on the road segment can be considered as an indication of the noise in the raw GPS signal (assuming an accurate road network graph).

The rest of the paper is organized as follows. Section 2 describes the GPS data and the experimental setup. Section 3 analyzes the noise patterns. Section 4 concludes the paper.

2 THE DATA SET

The GPS data is collected from 161 shuttles over a three month period. The total number of GPS samples is 6, 536, 702. This gives an average of 74, 280 samples per day and 40, 600 samples per shuttle. Figure 1 shows the distribution of the samples over the days of the week and the hours of the day.

The sampling rate is crucial to many operations including map matching. Low sampling rates (or high inter-arrival time between samples) may pose a risk when map matching the original trace to the road network graph. The average inter-arrival time between two consecutive GPS samples from the same shuttle is 34.05 seconds. However, the standard deviation of 25, which indicates a wide range of sampling rates. To give a better image of the sampling rate, Figure 2(a) illustrates a histogram of the inter-arrival times between consecutive samples. It shows that we have a good amount of GPS traces with 0-25 seconds inter-arrival time between samples. However, the majority of traces show samples that are 25 to 35 seconds apart and we have few traces where the inter-arrival time is more than 35 seconds and stretches to over a minute in some cases. Also, Figure 2(b) illustrates a histogram of the distance between consecutive samples of the same trace. The skewness of the data to the left is attributed to the stop and go nature of the shuttle service and/or the low speed of shuttles while on campus and while boarding/dropping off passengers.

As we mentioned earlier in this paper, we consider noise as an asset. We would like to differentiate between noise that we would consider as an asset and outliers that have to be filtered out before any processing would take place. The deviation of a point from the road segment to the left or to the right is a noise that we are interested in studying and analyzing its pattern in different situations. However, GPS points that are completely far away from other points in the GPS trace are undesirable and are considered outliers. For examples, GPS points in the middle of the ocean or in another continent are clear outliers that we filter out before further analysis takes place. To filter out such outliers, we introduce the concept of *reachability speed*, which is the speed the shuttles needs to go over to reach point number $i + 1$ from point i in the GPS trace. As a quick and early filter, we decide to filter out any GPS points that has to be reached with an over 100 mile/hour speed. This filter removed 18, 531 GPS samples, which is a total of 2.83% of the total GPS samples in hand.

We split GPS samples of each shuttle into several *runs*. The run is a continuous GPS trace of the same shuttle. For example, the GPS trace of a shuttle may be split into runs where each run represents a workday worth of data or a block of consecutive hours where the shuttle has been active. If there is a five minute separation or a one kilometer distance between samples in the trace, the trace is divided into two runs to maintain the continuity and the locality of the shuttle run. After the splitting into runs, we got 59, 699 runs for the 161 shuttles. There are 107 GPS samples for each run on average. The average number of runs per shuttle is 375.

We map-match every run through the *Microsoft Maps Snap-to-Road API*. 58, 200 runs were successfully snapped by the snapping API. The snapping API was not able to find a feasible way of map-matching the remaining runs, which can be attributed to high levels of noise in few traces or to the absence of some parking lots from the underlying maps. The distance between the raw GPS point and its snapped counterpart on the road segment can be an indication of how noisy the GPS signal is, how inaccurate the underlying map is, or both. The distance between the raw GPS point and its snapped counterpart is 10.95 meters on average with a relatively high standard deviation of 13. In the following section, we study the noise patterns in more details.

3 ANALYSIS OF GPS DATA AROUND TALL BUILDINGS

In this study, we consider 63 tall buildings in the down towns of Seattle [3] and Bellevue [1] to generate statistics and histograms of noise. Then, we focus on six tall buildings and six flat areas (with no tall buildings) for deeper analyses and comparisons.

3.1 Statistics on signal noise around tall buildings

In areas of tall buildings, the average distance between raw data points and their snapped counterparts is 18.04 meters with a standard deviation of 14.5. In flat areas where no tall buildings exist, the average distance between raw data points and their snapped counterparts is 7.9 meters with a standard deviation of 7.6. Other than the overall averages and standard deviations, we nail down and classify the raw GPS points into three buckets and recompute the statistics. This classification is based on whether the noise pushed the GPS signal to the left or to the right of the road segment. The rationale behind this classification comes from the intuition that noise (under no external effects) is expected to scatter points randomly and uniformly to the left and right of the road segments. However, buildings and constructions that are on one side of the road may consistently deviate points into one side of the road segment. The three buckets are: (1) points that are within 3 meters of the road segment, and which are considered a perfect match[1], (2) points that are to the right of the road segment by more than 3 meters, and (3) points that are to the left of the road segment by more than 3 meters.

Table 1 shows that in flat areas 30% of the points are considered a perfect match, while only 4.57% is a perfect match in areas of tall buildings. The percentages of points to the left and right of the road segments are 25% and 45%, respectively in flat areas. This is in contrast to the higher percentage of 29.92% and 65.51% in areas of tall buildings. It is clear that high buildings distort GPS signal and cause less points to be a perfect match. We also note that we expect to have more GPS points on the right side of the road segment compared to the left side of the road segment. In the United States, vehicles drive on the right side of the road. Map providers record the median (or the middle) of the road as the geometry representation of the road segments in the road network graph database.

[1]the underlying map provider claims that 3 meters is a good estimate of the road span to the left or to the right of the road geometry recorded in the road network graph.

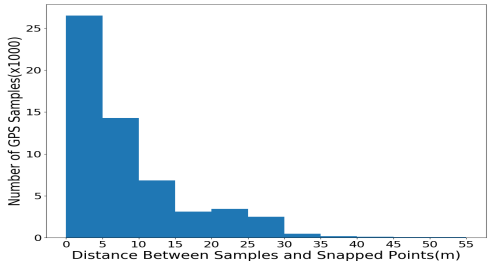
	Flat areas	*Tall buildings*
All points		
Percentage	100%	100%
Average	7.9	18.04
Standard Deviation	7.6	14.5
Points that perfectly match the road segment		
Percentage	30%	4.57%
Average	1.43	1.36
Standard Deviation	.84	.92
Points that are on the left of the road segment		
Percentage	25%	29.92%
Average	10.96	23.76
Standard Deviation	9.54	15.25
Points that are on the right of the road segment		
Percentage	45%	65.51%
Average	10.41	17.73
Standard Deviation	8.92	13.30

Table 1: The average and standard deviation of noise in flat areas and areas of tall buildings.

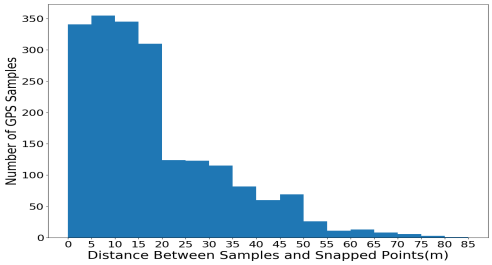
Figures 3 (a) and (b) give the noise histograms in flat areas and in areas with tall buildings, respectively. Figure 3(b) shows that, in some areas of tall buildings, the noise signal goes up to 70 and 75 meters. We also note that, in Figure 3(b), there is an abrupt drop at the value of 20 meters. We suspect that the building heights, in addition to other factors, may have an impact on the noise patterns. Hence, in the following section, we analyze the impact of the building heights and the proximity of the buildings to the roads on the signal noise.

3.2 The effect of the building height on the signal noise

It is intuitive that the taller the buildings are, the higher the noise is in the signal. Figure 4a illustrates the height of buildings against the noise received in the signal. The figure shows six selected tall buildings from the set of tall buildings in Seattle and Bellevue. From the figure, there is no visible trend that shows a correlation between noise and building heights. With a careful look into the locations of the buildings relative to the nearby roads, it is clear that some buildings are immediately on the road, while others are more to the inside and a little far from the roads. Therefore, we decide to weigh the height of the building by the distance from the building to the road. Figure 4b divides the height of the building by the distance to the road (on the x-axis) and visualizes these values against the signal noise (on the y-axis). The figure claims a clear trend where the noise reduces as a function of the building height divided by the building's distance to the road. While the Figure shows the impact of the "distance to the road" on the noise, we believe that this result is only an initial result. The exact effect of the building height and distance to roads needs further investigations. As we collect more data sets, we plan to do advanced curve fitting that can reveal deeper relationships. Other factors that may impact the noise pattern (and that are not studied yet) include: the materials of the building surfaces, the travelling direction of the shuttle relative to the building, the width/speed limit/one-wayness of the road

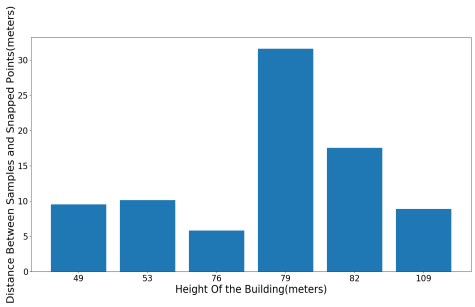


(a) Noise histogram in flat areas

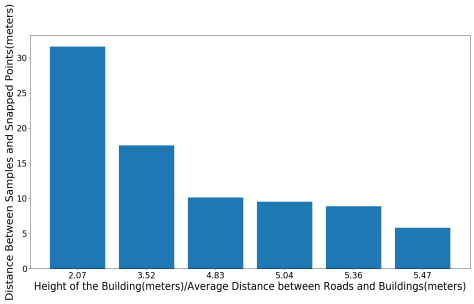


(b) Noise histogram in areas with tall buildings

Figure 3: Histograms that show the noise patterns in (a) flat areas and (b) in areas of tall buildings.



(a) Effect of building height



(b) Effect of building height divided by distance to road

Figure 4: The effect of the building height on the noise of the GPS Signal.

segments where these tall buildings lie. We also refer the reader to the study in [11] where the authors investigate the signal to noise ratio that occurs when buildings obstruct the line-of-sight between GPS receivers and the satellites.

4 CONCLUSION

In this paper, we studied the noise patterns of the GPS signal under various conditions. Our definition of noise is the distance between the raw GPS point and the point's map-matched location on the road network graph. We provided statistics and histograms of the noise patterns near tall buildings. This paper opens the door for two research directions: (1) identifying the environmental, weather and driving conditions from the noise patterns, and (2) identifying inaccuracies in the underlying map from mismatches between the raw GPS points and their map-matched counterparts. This paper dealt with noise as a valuable asset and presented our experience in analyzing the noise patterns. We expect future work to apply machine learning techniques for a deeper analysis to the noise patterns.

REFERENCES

[1] Bellevue tall buildings. https://en.wikipedia.org/wiki/List_of_tallest_buildings_in_Bellevue,_Washington. Accessed: 2019-06-01.
[2] Bing maps snap to road api. https://www.microsoft.com/en-us/maps/snap-to-road. Accessed: 2019-06-01.
[3] Seattle tall buildings. https://en.wikipedia.org/wiki/List_of_tallest_buildings_in_Seattle. Accessed: 2019-06-01.
[4] H. Aly, A. Basalamah, and M. Youssef. Map++: A crowd-sensing system for automatic map semantics identification. In *2014 Eleventh Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 546–554. IEEE, 2014.
[5] H. Aly and M. Youssef. semmatch: Road semantics-based accurate map matching for challenging positioning data. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 5. ACM, 2015.
[6] C. Y. Goh, J. Dauwels, N. Mitrovic, M. T. Asif, A. Oran, and P. Jaillet. Online map-matching based on hidden markov model for real-time traffic sensing applications. In *2012 15th International IEEE Conference on Intelligent Transportation Systems*, pages 776–781. IEEE, 2012.
[7] A. M. Hendawi, M. Khalefa, H. Liu, M. Ali, and J. A. Stankovic. A vision for micro and macro location aware services. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 12. ACM, 2016.
[8] A. M. Hendawi, A. Rustum, A. A. Ahmadain, D. Hazel, A. Teredesai, D. Oliver, M. Ali, and J. A. Stankovic. Smart personalized routing for smart cities. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 1295–1306. IEEE, 2017.
[9] A. M. Hendawi, A. Rustum, A. A. Ahmadain, D. Oliver, D. Hazel, A. Teredesai, and M. Ali. Dynamic and personalized routing in prego. In *2016 17th IEEE International Conference on Mobile Data Management (MDM)*, volume 1, pages 357–360. IEEE, 2016.
[10] A. M. Hendawi, E. Sturm, D. Oliver, and S. Shekhar. Crowdpath: a framework for next generation routing services using volunteered geographic information. In *International Symposium on Spatial and Temporal Databases*, pages 456–461. Springer, 2013.
[11] T. S. D. A. M. K. Kihwan Kim, Jay Summet and I. Essa. Localization and 3d reconstruction of urban scenes using gps. In *IEEE International Symposium on Wearable Computers*.
[12] W.-C. Lee and J. Krumm. Trajectory preprocessing. In *Computing with spatial trajectories*, pages 3–33. Springer, 2011.
[13] P. Newson and J. Krumm. Hidden markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 336–343. ACM, 2009.