# Detecting Stops from GPS Trajectories: A Comparison of Different GPS Indicators for Raster Sampling Methods

Y. Wang*[1] and D. P. McArthur[2]

[12] Urban Big Data Centre, University of Glasgow, Glasgow, G12 8RZ, United Kingdom
*Email: yang.wang@glasgow.ac.uk; david.mcarthur@glsagow.ac.uk

## Abstract

With the increasing prevalence of GPS tracking capabilities on smartphones, GPS trajectories have proven to be useful for an extensive range of research topics. Stop detection, which estimates activity locations, is fundamental for organizing GPS trajectories into semantically meaningful journeys. With previous methods overwhelmingly dependent on thresholds, contextual information or a pre-understanding of the GPS records, this paper addresses the challenge by contributing a 'top-down' raster sampling method which samples pre-calculated GPS indicators and clusters the raster cells with significantly different values as stops. We report a comparison of a set of pre-calculated GPS indicators with two baseline methods. By referencing a ground truth travel dairy, the raster sampling method demonstrates good and reliable capabilities on producing high accuracy, low redundancy and close proximity to the ground truth in three distinct travel use cases. This further indicates a good generic stop detection method.

**Keywords:** Stop Detection, Raster, GPS, Semantic Trajectory

## 1. Introduction

GPS trajectories, track records with latitude/longitude and timestamps, provide new opportunities to capture human activity patterns, improving transportation planning methods and models, modelling the spread of disease and analysing and clustering individuals for meaningful semantic recommendations on social networks. Stay point detection is now recognized as an important phase to better infer activities that are conducted at certain locations, to enable segmentation of the whole trajectory into separate travel purposes and to indicate travel mode interchange points.

Many current stop or stay point detection methods such as density-based approaches (Schoier and Borruso 2011, Hinneburg and Keim 1998, Ankerst et al. 1999, Campello et al. 2013), threshold-based approaches (Ashbrook and Starner 2003, Schuessler and Axhausen 2009, Srinivasan et al. 2009, Spaccapietra et al. 2008, Yan et al. 2008, Yan 2010) and locations of interest (Alvares et al. 2007) scan GPS records while applying both temporal and spatial constraints for stop detection. Contextal information, thresholds and parameters are prerequisites for good results. A raster sampling based method is proposed as a 'top-down', rather than 'bottom-up', approach which employs a unified raster template to sample some pre-calculated GPS indicators. Our approach is different from existing raster approaches such as kernel density (Thierry et al. 2013, Lei et al. 2011) as we are not sampling the density of GPS records but rather taking GPS inferred information such as total dwelling time, frequencies of visits and travel time into account. The proposed raster method with minimal threshold and parameter settings is demonstrated to be fast and accurate by comparing the experiment result

with the ground truth travel dairies deliberately collected for three use cases with significant different travel behaviours within diverse transport environment.

# 2. Data Preparation and Method

## 2.1 Data Cleaning and Ground Truth

Our data is a collection of three user's travel activities: one in a suburban area of Glasgow, one inside city of Glasgow and one in London, harvested from the Catch! smartphone journey planning app. Figure 1 illustrates the spatial distribution of their activities in Kernel Density Estimation (KDE) maps. The GPS records are cleaned to have no duplicated timestamps and to ensure the inferred speed is always under 200km/h. The analysis is performed on day-by-day GPS episodes. The ground truth is a travel dairy containing geographical locations of stops in chronological order.

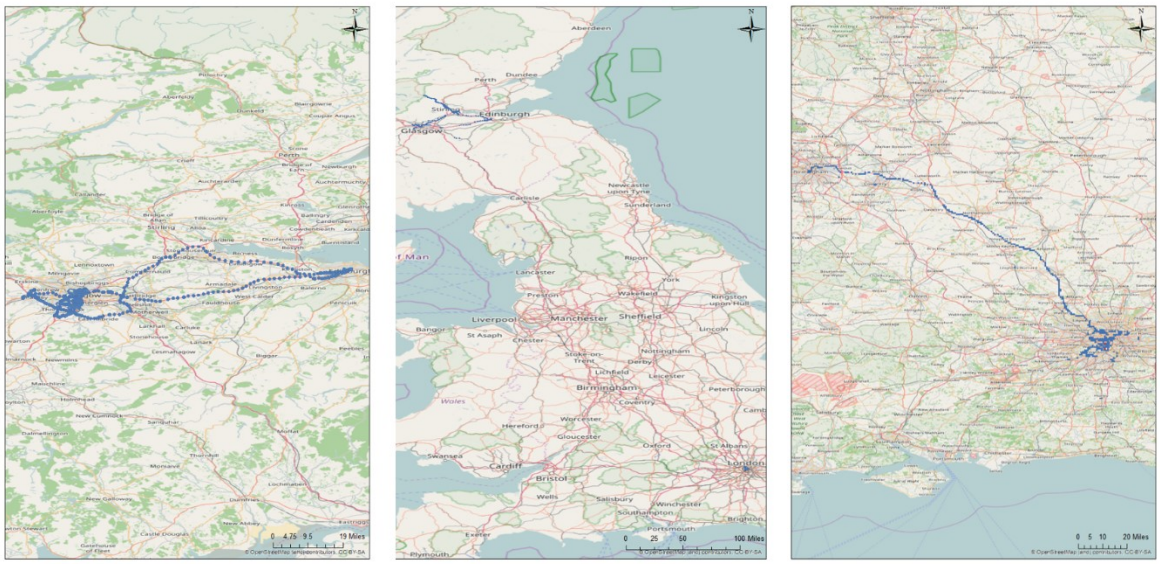


Figure 1. KDE maps for use cases 1, 2 and 3 from left to right respectively.

## 2.2 Method

The proposed stop detection method takes a top-down sampling process. With a given grid cell, $< r_j, c_k >_{j \in rows, k \in columns}$, certain GPS values are sampled. We report the results of our experiment for four pre-calculated GPS indicators to compare their ability to infer stops accurately. They include

(a). $dur\_lag_i$ dwelling time as a time difference between two consecutive GPS tuples;
(b). $actual(dur\_lag_{(j,k)})$ as a dwelling time deducing the travel time,

$$actual(dur\_lag_{(i)}) = dur\_lag_i - est(travelTime)_i \quad (1)$$

The estimated travel time is an estimation based on the speed before and after a GPS record.

$$est(travelTime)_i = {dur\_dist_i}\Big/{mean(speed_{window(i)})},$$
$$\text{where } window(i) \text{ is} < foreNeighour(5), afterNeighbour(5) > \quad (2)$$

(c). $dur\_lag_{(j,k)} / freq\_visits_{(j,k)}$ as an estimation of single trip GPS dwelling time at a given cell with arbitrary definition of trips when consecutive timestamps have a gap of more than 300 seconds;

(d). $actual(dur\_lag_{(j,k)}) / freq\_visits_{(j,k)}$ as an estimation of actual dwelling time (b) per visit.

Finally, we use Natural Break ($Jenks$) (https://pypi.python.org/pypi/jenkspy) to group sampled raster values into classes where the average deviations to the mean is minimized inside and maximized outside the classes. To avoid pre-setting the number of classes, goodness of variance fit (over 0.8) is adopted. We further select the raster cells with values higher than the 25% quantile of the clustering result as stops (Note that the data clustering algorithm is not the main contributions of the paper, other data clustering method such as $k$-means is also applicable). A threshold method (e), using thresholds to select stops with higher GPS dwelling time and method (f), detecting stops less 'bounded' with the road network through a map matching process, are chosen as baselines for comparison. Figure 2 demonstrates a typical detection results of one trip produced by all the methods.

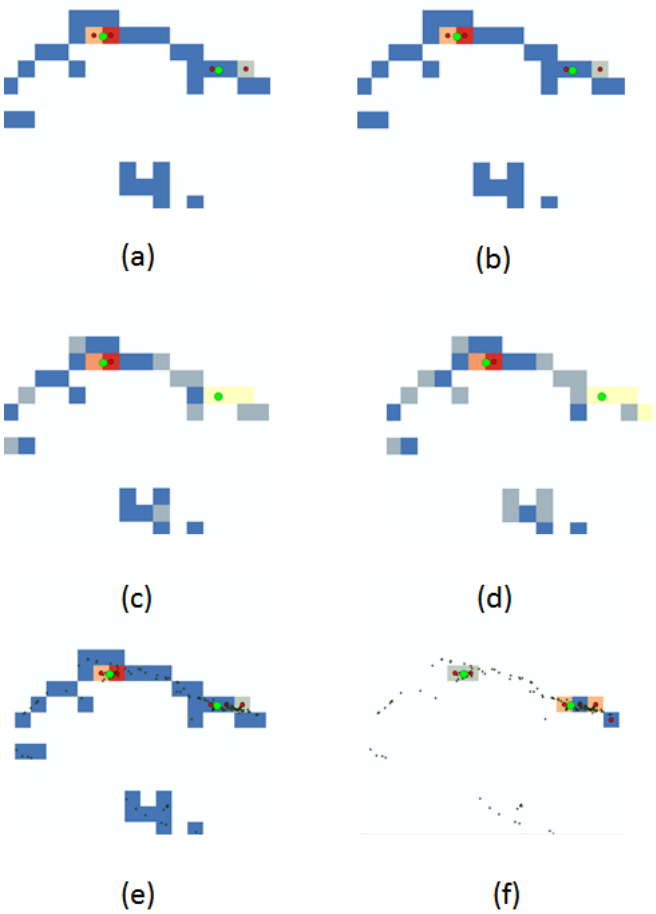


Figure 1. An example of one day output of (a) – (f) methods. Ground truth stops are denoted in green large points whilst detected stops in smaller red points against sampling raster. We convert results of baselines into raster format for comparison purposes.

# 3. Experimentation Results

For comparing the accuracy of detected stops with the ground truth stops, we measure the distances between a ground truth stop to the nearest detected stop cell's centroid for accuracy. Three distances tolerance levels, 100, 200, 300 meters, are tested. Precision/recall rates for individual methods in each use case are calculated as

$$Recall = \frac{Accurate(Stops)_{distanceTolerence}}{Total(GroundTruthStops)} \quad (3)$$

$$Precesion = \frac{Accurate(Stops)_{distanceTolerence}}{Total(DetectedStops)} \quad (4)$$

Figure 3 shows the precision/recall plots for the collected detection accuracy of methods (a)-(f) in three use cases. Each indicator has a distinct colour and shape where the darker the colour, the larger the distance tolerance. It is found that indicators (a) and (b) perform closely with higher precision/recall scores in all the use cases. Indicator (b) (in purple dots) is identified as a consistently strong indicator and marginally better in use case 1 and 3. Indicators (c) and (d) also produce similar but lower recall rates although the precisions are higher than for indicators (a) and (b). Indicators (e) and (f), on the contrary, are likely to generate higher recall but low precision rates.

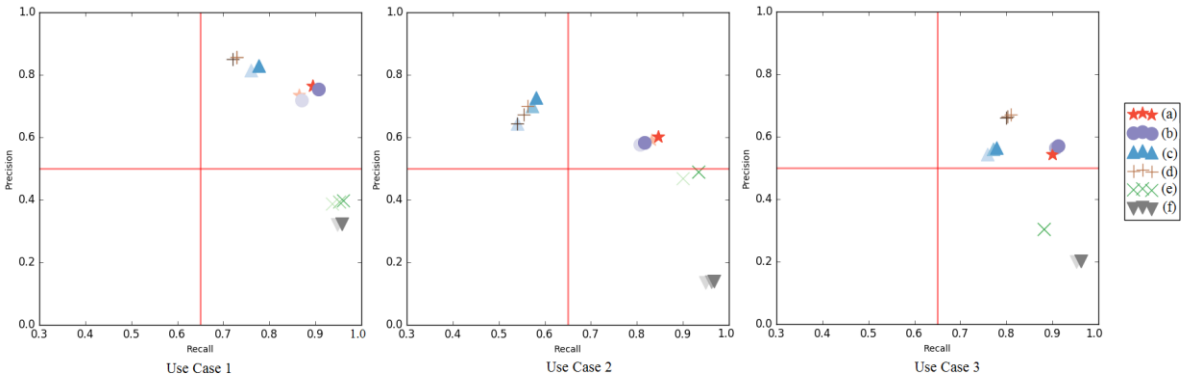


Figure 3. The precision/recall plots for the collected detection accuracy of the GPS indicators (a)-(d) and baselines (e) and (f) in three use cases with three distances tolerance settings.

As a sensitivity test, the performance of different methods under various cell sizes is shown in Table 1 among which the better performing indicators (a) and (b) are illustrated in Figure 4 with (a) in red (b) in green. The darker the colour, the larger the distance tolerance. It is shown that precisions/recall for (a) and (b) are again similar. Both rates are stable around 0.8, especially for use cases 1 and 2. Precision drops to around 0.6 while recall is high at around 0.9 for use case 3 where the user has a multi-modal travel pattern within an extremely complex travel environment. A slight decreasing trend can be observed when cell size increases.

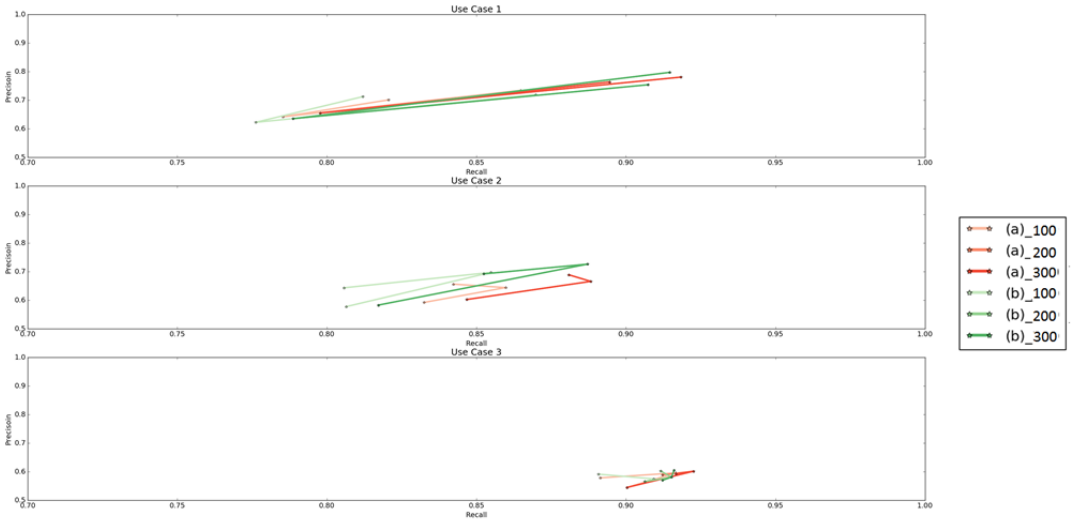


Figure 5. A comparison of precision/recall rates for raster sampling method (a) and (b) with an adjustment of cell sized from 0.00091 decimal degrees under WGS_1984 to one and half and double size (approximately 60, 85 and 110 meters).

The detected stops, as raster cells, can be measured w.r.t their proximity to the ground truth. We collect the distances from every ground truth stop to their nearest detected stops to examine which method from (a)-(f) has a higher probability of detecting stops closer to the ground truth. Taking use case 1 as an illustration, Figure 6 shows that except for the two baselines, method (b) takes a clear lead among all the indicators.

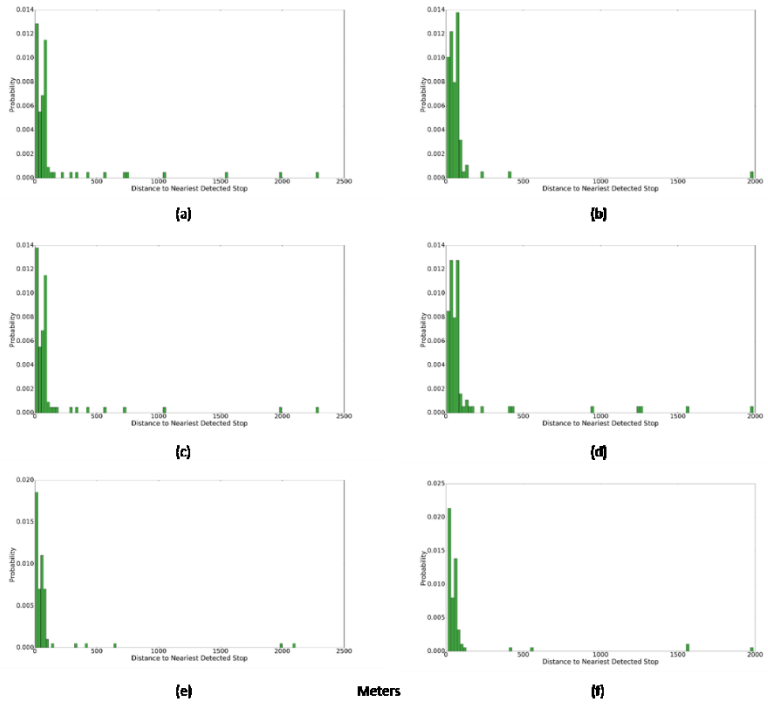


Figure 2. Probability Distribution of Collected Distances from Ground Truth Stops to the Nearest Detected Stops for (a) - (f) Methods.

# 4. Conclusion

This paper describes our effort towards designing a raster sampling based stop detection method with different indicators extracted from the GPS trajectories. The accuracy of detected stops is justified in three different use cases where users have distinct travel behaviour within different travel environment. Results show that raster methods in general produce satisfactory stops from the perspective of high and more stable precision/recall and proximity to detected stops and ground truth in the three use cases, with slight variations when adjusting the cell sizes. Although the accuracy reduces in more complicated use cases, with future combinations of smoothing functions, dwelling time (a) and dwelling time with reduced travel time (b), are strong indicators for a generic raster based stop detection method.

# 5. Acknowledgements

The authors want to thank TravelAI for their support and assistance with data collection, and to Innovate UK for funding the project (102426/53001-404133).

# 6. Reference

Alvares, L.O., Bogorny, V., Kuijpers, B., de Macedo, J.A.F., Moelans, B. and Vaisman, A., 2007, November. A model for enriching trajectories with semantic geographical information. In Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems (p. 22). ACM.

Ankerst, M., Breunig, M.M., Kriegel, H.P. and Sander, J., 1999, June. OPTICS: ordering points to identify the clustering structure. In ACM Sigmod record (Vol. 28, No. 2, pp. 49-60). ACM.

Ashbrook, D. and Starner, T., 2002. Learning significant locations and predicting user movement with GPS. In Wearable Computers, 2002.(ISWC 2002). Proceedings. Sixth International Symposium on (pp. 101-108). IEEE.

Campello, R.J., Moulavi, D. and Sander, J., 2013, April. Density-based clustering based on hierarchical density estimates. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 160-172). Springer Berlin Heidelberg.

Hinneburg, A. and Keim, D.A., 1998, August. An efficient approach to clustering in large multimedia databases with noise. In KDD (Vol. 98, pp. 58-65).

Lei, P.R., Shen, T.J., Peng, W.C. and Su, J., 2011, June. Exploring spatial-temporal trajectory model for location prediction. In Mobile Data Management (MDM), 2011 12th IEEE International Conference on (Vol. 1, pp. 58-67). IEEE.

Schoier, G. and Borruso, G., 2011, June. Individual movements and geographical data mining. Clustering algorithms for highlighting hotspots in personal navigation routes. In International Conference on Computational Science and Its Applications (pp. 454-465). Springer Berlin Heidelberg.

Schuessler, N. and Axhausen, K., 2009. Processing raw data from global positioning systems without additional information. Transportation Research Record: Journal of the Transportation Research Board, (2105), pp.28-36.

Spaccapietra, S., Parent, C., Damiani, M.L., de Macedo, J.A., Porto, F. and Vangenot, C., 2008. A conceptual view on trajectories. Data & knowledge engineering, 65(1), pp.126-146.

Srinivasan, S., Bricka, S. and Bhat, C., 2009. Methodology for converting GPS navigational streams to the travel-diary data format. Department of Civil and Coastal Engineering, University of Florida.

Thierry, B., Chaix, B. and Kestens, Y., 2013. Detecting activity locations from raw GPS data: a novel kernel-based algorithm. International journal of health geographics, 12(1), p.14.

Yan, Z., 2010, November. Traj-ARIMA: a spatial-time series model for network-constrained trajectory. In Proceedings of the Second International Workshop on Computational Transportation Science (pp. 11-16). ACM.

Yan, Z., Macedo, J., Parent, C. and Spaccapietra, S., 2008. Trajectory ontologies and queries. Transactions in GIS, 12(s1), pp.75-91.

| | Cell Size=0.00091≈60 meters | | | | | | Cell Size*1.5=0.00091*1.5≈85 meters | | | | | | Cell Size*2=0.00091*2≈110 meters | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Use Case 1 | | Use Case 2 | | Use Case 3 | | Use Case 1 | | Use Case 2 | | Use Case 3 | | Use Case 1 | | Use Case 2 | | Use Case 3 | |
| | recall | precision | recall | precision | recall | precision | recall | precision | recall | precision | recall | precision | recall | precision | recall | precision | recall | precision |
| **(a) 100** | 0.864744 | 0.734441 | 0.832520 | 0.592305 | 0.900298 | 0.544765 | 0.785256 | 0.64185 | 0.859756 | 0.644164 | 0.916667 | 0.595773 | 0.82062 | 0.70098 | 0.842276 | 0.656388 | 0.891369 | 0.578366 |
| **(a) 200** | 0.894551 | 0.763578 | 0.846748 | 0.602468 | 0.900298 | 0.544765 | 0.797756 | 0.655037 | 0.888211 | 0.665941 | 0.922619 | 0.601726 | 0.918376 | 0.780836 | 0.880894 | 0.689005 | 0.912202 | 0.589279 |
| **(a) 300** | 0.894551 | 0.763578 | 0.846748 | 0.602468 | 0.900298 | 0.544765 | 0.797756 | 0.655037 | 0.888211 | 0.665941 | 0.922619 | 0.601726 | 0.918376 | 0.780836 | 0.880894 | 0.689005 | 0.916667 | 0.592255 |
| **(b) 100** | 0.869872 | 0.719689 | 0.806452 | 0.577611 | 0.906250 | 0.565407 | 0.776175 | 0.622436 | 0.854839 | 0.697427 | 0.909226 | 0.574943 | 0.812073 | 0.712576 | 0.805691 | 0.643438 | 0.890774 | 0.591288 |
| **(b) 200** | 0.907372 | 0.754029 | 0.817204 | 0.582988 | 0.906250 | 0.565407 | 0.788675 | 0.635256 | 0.887097 | 0.726229 | 0.915179 | 0.580896 | 0.914637 | 0.797405 | 0.852439 | 0.692218 | 0.911607 | 0.602597 |
| **(b) 300** | 0.907372 | 0.754029 | 0.817204 | 0.582988 | 0.912202 | 0.569872 | 0.788675 | 0.635256 | 0.887097 | 0.726229 | 0.915179 | 0.580896 | 0.914637 | 0.797405 | 0.852439 | 0.692218 | 0.916071 | 0.605574 |
| **(c) 100** | 0.759722 | 0.813782 | 0.539815 | 0.644444 | 0.758170 | 0.543129 | 0.761538 | 0.755311 | 0.60098 | 0.692157 | 0.816993 | 0.633442 | 0.764103 | 0.794414 | 0.535185 | 0.686111 | 0.784314 | 0.621102 |
| **(c) 200** | 0.777030 | 0.828205 | 0.572222 | 0.700000 | 0.771242 | 0.559469 | 0.796474 | 0.799267 | 0.635294 | 0.75098 | 0.823529 | 0.63671 | 0.812179 | 0.852106 | 0.613889 | 0.792593 | 0.820261 | 0.652801 |
| **(c) 300** | 0.777030 | 0.828205 | 0.581481 | 0.727778 | 0.777778 | 0.564371 | 0.796474 | 0.799267 | 0.635294 | 0.75098 | 0.823529 | 0.63671 | 0.816987 | 0.864927 | 0.613889 | 0.792593 | 0.820261 | 0.652801 |
| **(d) 100** | 0.720726 | 0.849817 | 0.539815 | 0.644444 | 0.799020 | 0.659633 | 0.731197 | 0.739744 | 0.620588 | 0.695098 | 0.854575 | 0.726751 | 0.726496 | 0.827564 | 0.516667 | 0.681481 | 0.825163 | 0.686368 |
| **(d) 200** | 0.728419 | 0.855311 | 0.553704 | 0.672222 | 0.802288 | 0.664535 | 0.761325 | 0.78141 | 0.654902 | 0.753922 | 0.861111 | 0.730672 | 0.774573 | 0.882051 | 0.613889 | 0.806481 | 0.851307 | 0.709788 |
| **(d) 300** | 0.728419 | 0.855311 | 0.562963 | 0.700000 | 0.808824 | 0.671071 | 0.761325 | 0.78141 | 0.654902 | 0.753922 | 0.861111 | 0.730672 | 0.774573 | 0.882051 | 0.613889 | 0.806481 | 0.851307 | 0.709788 |
| **(e) 100** | 0.937073 | 0.387613 | 0.899187 | 0.469742 | 0.952381 | 0.199362 | 0.912714 | 0.359623 | 0.808943 | 0.607877 | 0.872024 | 0.457775 | 0.873504 | 0.446626 | 0.708537 | 0.553262 | 0.830357 | 0.44365 |
| **(e) 200** | 0.951496 | 0.393489 | 0.933740 | 0.489080 | 0.961310 | 0.201594 | 0.936752 | 0.373191 | 0.846341 | 0.639177 | 0.885417 | 0.46715 | 0.922329 | 0.472641 | 0.781707 | 0.619928 | 0.857143 | 0.457114 |
| **(e) 300** | 0.961111 | 0.397763 | 0.933740 | 0.489080 | 0.961310 | 0.201594 | 0.936752 | 0.373191 | 0.846341 | 0.639177 | 0.885417 | 0.46715 | 0.922329 | 0.472641 | 0.781707 | 0.619928 | 0.857143 | 0.457114 |
| **(f) 100** | 0.947650 | 0.321370 | 0.949187 | 0.136018 | 0.881845 | 0.305179 | 0.925214 | 0.300621 | 0.934959 | 0.159497 | 0.952381 | 0.279078 | 0.890598 | 0.419322 | 0.860569 | 0.169531 | 0.925595 | 0.265302 |
| **(f) 200** | 0.957265 | 0.322909 | 0.961382 | 0.138429 | 0.881845 | 0.305179 | 0.944444 | 0.306497 | 0.945122 | 0.162478 | 0.96131 | 0.281629 | 0.952457 | 0.439694 | 0.922764 | 0.181867 | 0.946429 | 0.267947 |
| **(f) 300** | 0.957265 | 0.322909 | 0.967480 | 0.139863 | 0.881845 | 0.305179 | 0.944444 | 0.306497 | 0.945122 | 0.162478 | 0.96131 | 0.281629 | 0.952457 | 0.439694 | 0.928862 | 0.183899 | 0.955357 | 0.270923 |

Table 1. Precision/recall rates in summery using 100,200 and 300 meter distance tolerance for (a) – (f) method with cell sizes adjusted as 0.00091, 0.00091*1.5, 0.00091*2 in WGS_1984.