THE 1st INTERNATIONAL CONFERENCE ON DATA SCIENCE AND OFFICIAL STATISTICS
ICDSOS

"Strengthening the Contribution of Data Science and Official Statistics to the Society in the Distruption Era"

2021

# Determining the Stopping Point on GPS Data Using Density Based Spatial Clustering of Application with Noise and Gaussian Mixture Model Cluster

**Y A Faeni**

BPS of East OKU Regency

*Corresponding author's e-mail: you@bps.go.id

**Abstract.** GPS data is an interesting thing to research. Various studies have been conducted to find information based on GPS data. In this paper, we propose a novel model for determining the stopping point on a GPS data for cases of human movement without using transportation modes. Further, this information can be used to determines human behavior such as fraud and favorite spot. The GPS data used in this research is the travel data of the SUSENAS survey officers at the time of updating the census block for 27 households. Density Based Spatial Clustering Of Application With Noise (DBSCAN) And Gaussian Mixture Model (GMM) Clustering model is used to create the model. The model made using a flowchart and applied to the GPS data that has been collected. The results of the developed model show that the stopping points generated using the DBSCAN cluster model are better than the stopping points generated using the GMM cluster model**.** Furthermore, the results of this study will be used to make model of surveyor fraud.

## 1. Preliminary

GPS is a satellite system for navigation and positioning using satellites. One of the factors of accuracy of positioning with GPS is data processing strategy. Following the development of GPS needs in human life, GPS observation data can be processed using a variety of software. Data collection carried out by a GPS device to measure the location or research location, distance, absolute location, and land height from sea level.

Several studies have been conducted to determine the stop point of a GPS data starting from the stop point when driving a car to the stop point of the movement of animals [1] (Faeni, 2020). Peter Stoper et al. (2005) describe how GPS devices work and demonstrate the ability of the device to provide accurate data about travel movements. The study describes the software developed to improve the ability to analyze data results (Stopher et al., 2005). The latest GPS devices at that time showed the potential to replace conventional methods of data collection which were not good enough due to known errors and inaccuracies in the data (Gong L et al., 2014). This study developed a procedure that made it possible to infer models and destinations on most of the recorded trip data.

Chaoran Zhou et al (2017) using tracking data obtained from smartphones and internet surveys, drawing conclusions based on data using machine learning methods is used to find out the stops of a trip. In his research, millions of smartphone-based GPS tracking data were used. Various attributes such as speed, distance, aim, etc. used to describe the travel status of smartphone users.

Another research is Jinjun Tang et al. (2017) which uses a two-layer decision framework to model the behavior of taxi drivers looking for passengers in urban areas. The first layer models the taxi driver's decision to pick up passengers from one place. The Huff model is used to describe the attractiveness of the pick-up location (Xiao Z et al., 2017). Furthermore, the Path Size Logit (PSL) model is used in the second layer to analyze route choices based on some information such as path size, path distance, travel time and delay due to intersections. The results show that the proposed Huff model has high accuracy for estimating the driver's pick-up location choice. PSL outperforms traditional multitomial logit in modeling driver's route choice behavior (Tang et al., 2016).

Subsequent research was conducted by Zong fang et al. (2018) designed a process to identify a person's journey and activities based on GPS data. The identification process consists of four stages, namely: determining segment status, detecting activity, identifying trips, and recognizing short-term activities (Fang et al., 2018). By proposing a method for identifying trips and activities from GPS data, this study provides a research scheme for detecting other travel information based on GPS data (Yu Zhang et al., 2018) such as travel mode and travel destination, a sensible decision for transportation planning and management. urban.

From the various literatures above, it can be learned how research related to the use of GPS data or time analysis to determine a person's behavior can be studied. Furthermore, this study uses GPS data and stop time to validate the data collection process in survey activities. Furthermore, the results of the analysis are expected to assist in making a decision whether a data collection process in a survey needs to be repeated or is already valid.

## 2. Research methods

This study uses primary data collected by simulation (Halder, 2006) directly to the field which was carried out in the case of updating (household updating) SUSENAS. The simulation was carried out at RT 03 RW 01, Cibeuying Hegar, Sadang Serang on January 12, 2019. Data was collected using an android application that was made by myself and using a Samsung A50 device.

The development of the stopping point determination model is designed according to the characteristics of the data generated from the results of the simulation data carried out. The system simulation is carried out with the following scenarios:

- Susenas Updating Process Simulation
- Target population : 30 households in RT 03 RW 01, Cibeuying Hegar, Sadang Serang
- Simulation time: 08.00 – 11.00
- Device used:
  - Samsung A50
- The simulation is carried out in one process without a pause
- The simulation is carried out by direct interviews, walking from house to house

The results of the system simulation can be seen in the following table:

**Table 1.** Results simulation.

| No | Parameter | Method |
|---|---|---|
| 1 | GPS Accuracy | 6-14 m |
| 2 | Battery Consumption | 6%/hour |
| 3 | A lot of data collected | 669 records |
| 4 | Data size | 4863 bytes |

The data is processed using R software (Handoyo et all. 2017) using the mclust and fpc libraries. Software is used to create clusters using the DBSCAN (L Ni et all, 2018) and GMM methods. In addition, R software is also used to visualize data and maps.

Density-Based Spatial Clustering of Application with Noise (DBSCAN) is one of the pioneer examples of the development of density-based clustering techniques or commonly known as density-based clustering(Mumtaz ea, 2010). Among the various types of clustering algorithms, density based clustering is more efficient for determining clusters on data with different densities(Matheus ea, 1993). Based on this assumption DBSCAN used in this research.

GMM is a type of density model consisting of components of Gaussian functions. This function component consists of different Thresholds to produce multi-density models. Yihua (2010) used GMM to model the distribution of GPS traces across multiple traffic lanes. The GMM naturally accounts for the inherent spread in GPS data.

In the simulation process, from the target of 30 households, 27 households were successfully met and interviewed. 3 households could not be found because after being visited the household was not at home so that household information was obtained from the nearest neighbor. The model is built using the principle of proximity as a household point determination. Here is a flowchart of the resulting model:

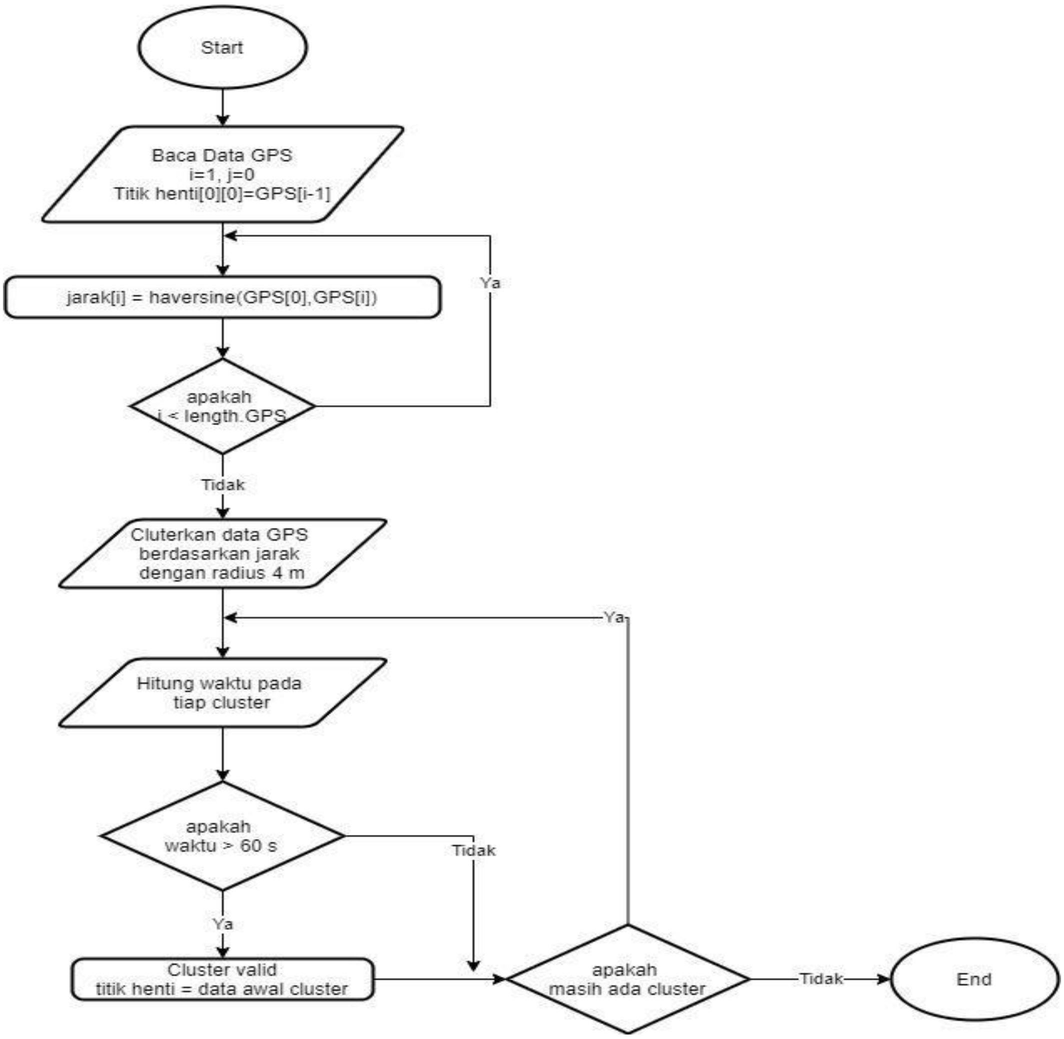


**Figure 1.** Model stop point determination.

Determination of the stopping point using model explained by the flowchart image above. The following steps are carried out in model:

- Calculate the distance between all points and point 0 using the haversine formula. Here is the formula for calculating distance using the haversine formula (Chandra ea, 2020. Saputra ea, 2019):

$$Distance(point1, point2) = ACOS\left(\sin\left(lat1 * \frac{PI(\ )}{180}\right) * \sin\left(lat2 * \frac{PI(\ )}{180}\right) + \cos\left(lat1 * \frac{PI(\ )}{180}\right) * \cos\left(lat2 * \frac{PI(\ )}{180}\right) * \cos\left(lon2 * \frac{PI(\ )}{180} - lon1 * \frac{PI(\ )}{180}\right)\right) * 6371000 \quad (1)$$

Information:

| | |
|---|---|
| lat1 | : latitude coordinates at point 1 |
| lat2 | : latitude coordinates at point 2 |
| long1 | : longitude coordinates at point 1 |
| long2 | : longitude coordinates at point 2 |
| PI() | : constant of magnitude pi (3.14) |

- Clustering data using GPS based on distance so as to get a temporary stop point using density based clustering (DBSCAN) and GMM methods
- Calculate the stop time on each cluster
- If the downtime > 30 seconds then input as a valid break point. If downtime < 30 seconds continue loop
- Continue iterating over the resulting number of clusters

## 3. Results and discussion

### *3.1. Clustering with DBSCAN and GMM*

In this study, researchers compared two clustering methods to get clusters that match expectations. The methods used are DBSCAN clustering (Ester, 1996) and gaussian mixture model (GMM) clustering (Y Li et all, 2020). Figures 2 and figure 3 show the results of two clustering methods on GPS data generated by model.

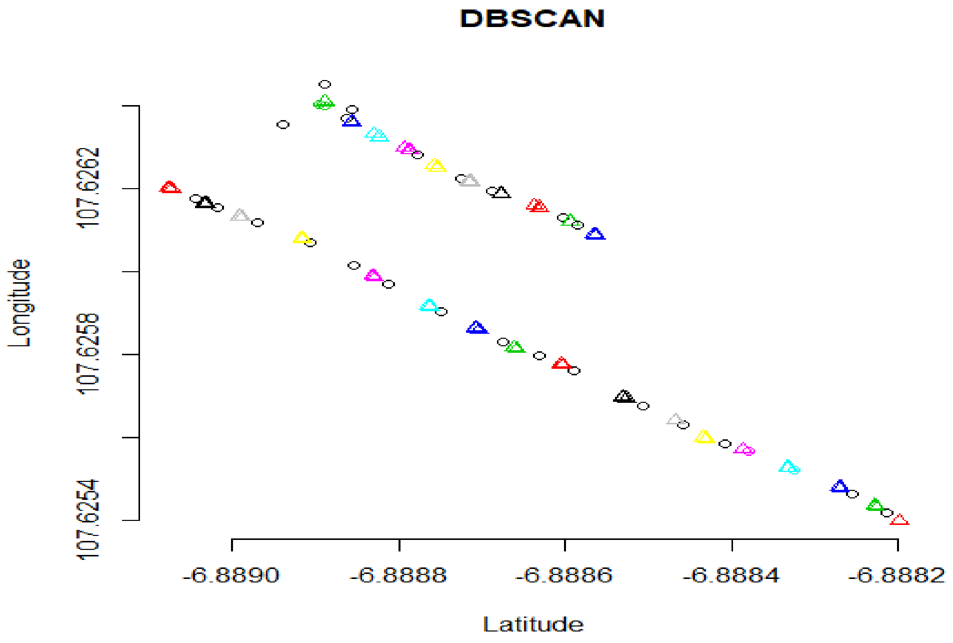


**Figure 2.** Clustering results with DBSCAN.

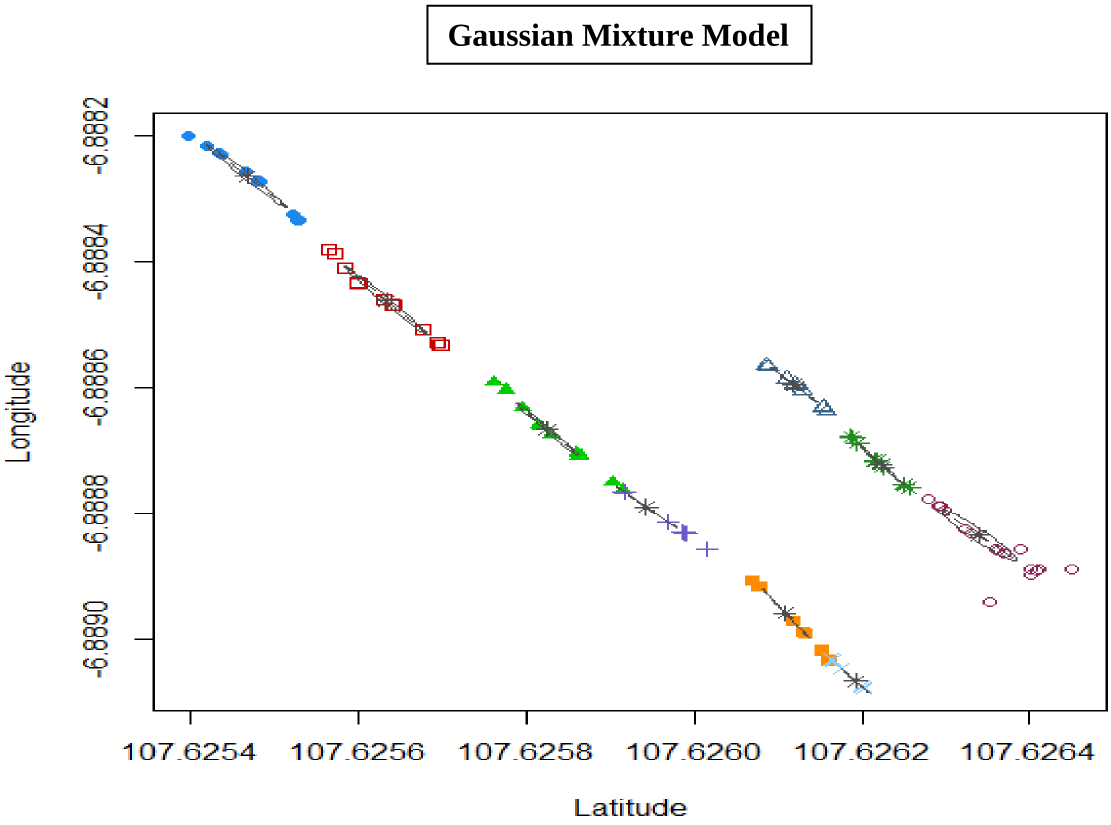


**Figure 3.** The results of clustering with the GMM cluster method.

Figures 2 and figure 3 show the results of clustering using the GMM and DBSCAN methods. The DBSCAN method produces 27 clusters as seen from the different colors of 27 groups. The GMM cluster method produces 9 clusters as seen from the color of 9 groups. The syntax of the R application use mclust library.

From the results of the analysis, it can be concluded that the DBSCAN method is more suitable for this research. Because the distance between clusters can be adjusted according to the characteristics of the distance of the house in a census block. So that the resulting cluster can be more in line with the actual situation.

*3.2. Evaluation of model*

This evaluation is carried out by comparing the average stopping point resulting from the model calculation with the stopping point seen based on the Google Maps application. The evaluation steps are as follows:

Hypothesis

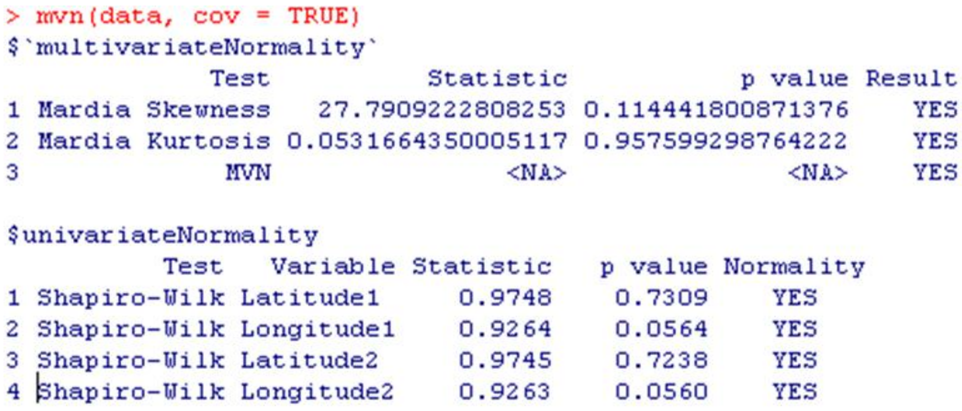
The hypotheses in this evaluation are:

- H0 : : There is no average difference between the variables in the generated GPS data
- H1: The average between the variables in the resulting GPS data is different

If the p-value/ Sig. 0.05 then H0 is accepted.
If the p-value/ Sig. < 0.05 then H0 is rejected

Normality test

Normality test was performed using Shapiro-Wilk (Hinton et al., 2014) on R software. All p-values > 0 ,05. So it can be concluded, all variables with normal distribution are normally distributed and there is no need for data normalization process.



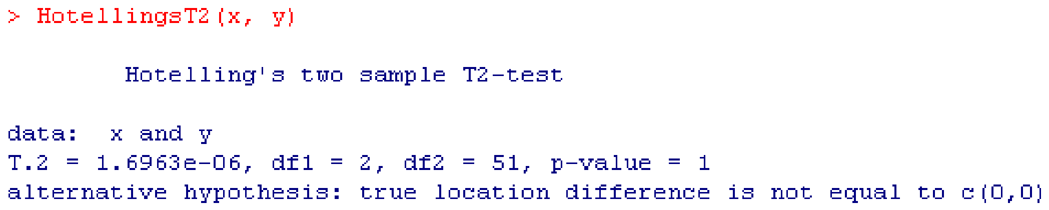
```
> mvn(data, cov = TRUE)
$`multivariateNormality`
             Test          Statistic           p value Result
1 Mardia Skewness   27.7909222808253 0.114441800871376    YES
2 Mardia Kurtosis 0.0531664350005117 0.957599298764222    YES
3             MVN               <NA>              <NA>    YES

$univariateNormality
          Test   Variable Statistic   p value Normality
1 Shapiro-Wilk Latitude1      0.9748   0.7309    YES
2 Shapiro-Wilk Longitude1     0.9264   0.0564    YES
3 Shapiro-Wilk Latitude2      0.9745   0.7238    YES
4 Shapiro-Wilk Longitude2     0.9263   0.0560    YES
```

**Figure 4.** Stop point normality test results with R . software..

Test Multivariate t-test

Because the conditions for the normal distribution assumption are met, then the sample data can then be tested with parametric statistics, namely the paired t test (multivariate) (Kim T, 2015). It can be seen in the figure below that the p-value is 1, which means > 0.05, then accept H0 and H1 is rejected. So it can be concluded at a 95% confidence level that the stopping point calculated by model II is the same as the actual stopping point seen with the Google maps application.



```
> HotellingsT2(x, y)

        Hotelling's two sample T2-test

data:  x and y
T.2 = 1.6963e-06, df1 = 2, df2 = 51, p-value = 1
alternative hypothesis: true location difference is not equal to c(0,0)
```

**Figure 5.** Result of multivariate t-test.

Next, overlay the points on the map using the R application with the leaflet library. There are 2 sets of points that are overlaid on the map, namely the starting point of method II and the stopping point of model . Figure 6 show the results of the point overlay.

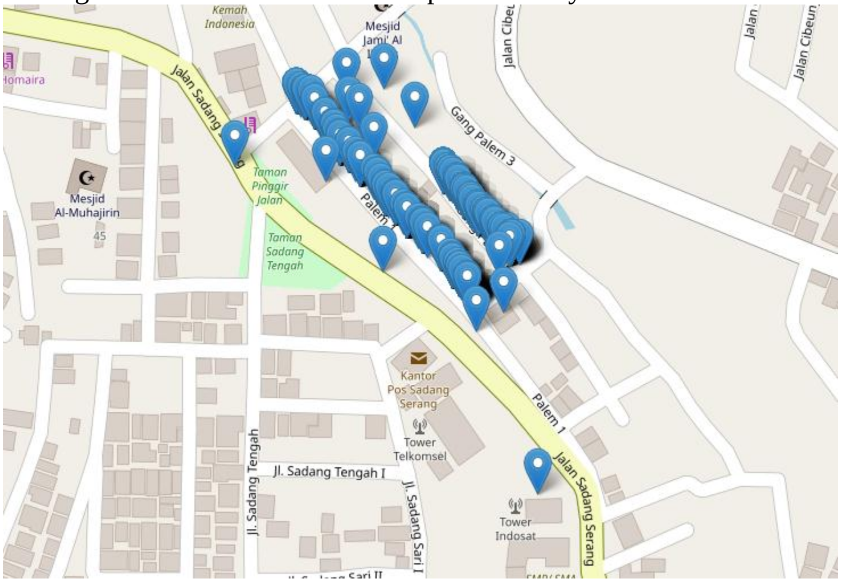


**Figure 6.** The result of overlaying the starting point of model on the map.

## 4. Conclusion

Based on the results and discussion, this study can conclude several things, including:

1. This study succeeded in making a model for determining the GPS stopping point of human activities not using the mode of transportation (walking) and evaluating the model.
2. The DBSCAN Clustering model is more suitable to be used to determine the stopping point in the GPS data of the SUSENAS updating case study.
3. The results of the evaluation of the stopping point show the model for determining the stopping point according to reality compared to the point on google maps with a 95% confidence level.

This research can be used to determine the research stop point using GPS data in cases of movement not using transportation modes. Furthermore, we will use the results of this study to create a surveillance system in a survey that can detect surveyor falsification. We suggest more study to differentiate the activity of humans when they stop it can be for work or just to rest.

## References

[1] Faeni, You Ari,Fadhil Hidayat.2020.Decision-Making Framework for Validation of Data Collection Process in a Survey with GPS Data.MSCEIS.EAI.DOI: 10.4108/eai.12-10-2019.2296545

[2] Esther, Martin,. Kriegel, Hans-Peter., Sander, Jorg., & Xu, Xiaowei. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Germany: University of Munich.

[3] Handoyo, S., Prasojo, APS, & Naba, A. (2017). Applied Fuzzy System with R Software: Universitas Brawijaya Press.

[4] Tang, J., Jiang, H., Li, Z., Li, M., Liu, F., & Wang, Y. (2016). A Two-Layer Model for Taxi Customer Searching Behaviors Using GPS Trajectory Data. IEEE Transactions on Intelligent Transportation Systems, 17, 3318-3324..

[5] Halders. A, "A Study of Petri Nets: Modeling, Analysis, and Simulation," 2006.

[6] Chen, X. et al., “Data-driven prediction system of dynamic people-flow in large urban network using cellular probe data,” J. Adv. Transp., vol. 2019, 2019.

[7] Xiao, ZY Wang, K. Fu, and F. Wu, “Identifying different transportation modes from trajectory data using tree-based ensemble classifiers,” ISPRS Int. J. Geo-Information, vol. 6, no. 2, 2017.

[8] C. Zhou, H. Jia, Z. Juan, X. Fu and G. Xiao, "A data-driven method for trip ends identification using large-scale smartphone-based GPS tracking data", IEEE Trans. Intell. Transp. Syst., vol. 18, no. 8, pp. 2096-2110, Aug. 2017.

[9] Zhou, ZJ Yang, Y. Qi, and Y. Cai, “Support vector machine and back propagation neutral network approaches for trip mode prediction using mobile phone data,” IET Intell. Transp. syst. vol. 12 no 10. pp. 1220–1226, 2018.

[10] Matheus CJ, Chan PK, and Piatetsky-Shapiro G, “Systems for Knowledge Discovery in Databases,” IEEE Trans. Knowl. Data Eng., vol. 5, no. 6, pp. 903–913, 1993.

[11] Mumtaz K, “An Analysis on Density Based Clustering of Multi Dimensional Spatial Data,” Indian Journal of Computer Science and Engineering (IJCSE), vol. 1, no. 1, pp. 8–12, 2010

[12] K. Saputra, N. Nazaruddin, D. H. Yunardi and R. Andriyani, "Implementation of Haversine Formula on Location Based Mobile Application in Syiah Kuala University," 2019 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), 2019, pp. 40-45,

[13] Y. Li, J. Zhang, Z. Ma and Y. Zhang, "Clustering Analysis in the Wireless Propagation Channel with a Variational Gaussian Mixture Model," in IEEE Transactions on Big Data, vol. 6, no. 2, pp. 223-232, 1 June 2020

[14] Stopher, P., Fitzgerald, C., Zhang, J.: 'Search for a global positioning system device to measure personal travel', Transp. Res C. 2008. 16(3).pp. 350–369

[15] Gong Lei, Takayuki Morikawa, Toshiyuki Yamamoto, Hitomi Sato.2014.Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies.Procedia - Social and Behavioral Sciences.Volume 138,Pages 557-565,ISSN 1877-0428.

[16] Fan. L, L. Tang, and S. Chen, “Optimizing location of variable message signs using GPS probe vehicle data,” PLoS One, vol. 13, no. 7, 2018.

[17] Fang. Z, L. Jian-yu, W. Xiao, T. Jin-jun, and G. Fei, “Identifying activities and trips with GPS data,” IET Intell. Transp. syst. vol 12 no. 8. pp. 884–890. 2018.

[18] Wang. L, Y. Zhong, and W. Ma. “GPS-data-driven dynamic destination prediction for on-demand one-way carsharing system,” 2018.

[19] Chandra Husada, Kristoko Dwi Hartomo, & Hanna Prillysca Chernovita. (2020). Implementasi Haversine Formula untuk Pembuatan SIG Jarak Terdekat ke RS Rujukan COVID-19. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 4(5), 874-883.

[20] Yu Zheng, Yukun Chen, Quannan Li, Xing Xie, and Wei-Ying Ma. 2010. Understanding transportation modes based on GPS data for web applications. ACM Trans. Web 4, 1, Article 1 (January 2010).

[21] L. Ni, C. Li, X. Wang, H. Jiang and J. Yu, "DP-MCDBSCAN: Differential Privacy Preserving Multi-Core DBSCAN Clustering for Network User Data," in IEEE Access, vol. 6, pp. 21053-21063, 2018

[22] Kim, TK (2015). T test as a parametric statistic. Korean Journal of Anesthesiology, 68,540–546.https://doi.org/http://dx.doi.org/10.4097/kjae.2015.68.6.540

[23] Yihua Chen and John Krumm. 2010. Probabilistic modeling of traffic lanes from GPS traces. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '10). Association for Computing Machinery, New York, NY, USA, 81–88