Data Input Quality Metrics on Mobile Positioning Data (MPD)

A R M N S P Munaf[1], A P Putra[2], W O Z Madjida[1], I A Setyadi[1], A R S Nugroho[1]

[1]Directorate of Statistical Information System, BPS - Statistics Indonesia - Jakarta, Indonesia
[2]Directorate of Statistics Methodology, BPS - Statistics Indonesia - Jakarta, Indonesia

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

Abstract. Statistics Indonesia (BPS) has been using Mobile Positioning Data (MPD) to support official statistics since 2016. As a source of big data, MPD also has veracity characteristics, indicating uncertainty in the data. Therefore, it is necessary to check that the data are good enough to allow further analysis and the quality assurance process. Currently, there is no established international standard for quality assurance of MPD. This paper describes the quality matrix used by BPS in examining data from mobile operators. BPS uses thirteen indicators in conducting quality assurance, where the inspection uses several different methods, such as setting a threshold, checking data completeness, and checking the form of data distribution. Exploratory Data Analysis is carried out to determine whether the data meets the requirements for further analysis. We conducted this research on a mobile network operator data for June - July 2020 as the basis for MPD analysis in 2021. Based on the inspection during this period, BPS can cooperate with this cellular operator to conduct data analysis in 2021. However, the operator must repeat the calculation of the required matrix as quality assurance every month.

1. Introduction

As the National Statistics Office, Statistics Indonesia (BPS) has the mandate to build, maintain, and develop various statistical products. In this role, BPS is obliged to carry out statistical quality assurance to ensure security and convince various parties to implement and use statistical data.

Laney (2001) on Patgiri and Ahmed [1] explains that big data has three main characteristics: Volume, Velocity, and Variety. Ellars [2] explained that at the Big Data Innovation Summit in Boston, Inderpal Bhandar, Chief Data Officer at Express Scripts, added a fourth V, namely veracity. The most critical point of veracity is to ensure that the source of big data that we use is good enough for further analysis [3].

BPS has been using Mobile Positioning Data (MPD) as one of the data sources since 2016, initially in the project of counting foreign tourists who enter through cross-border [4]. The use of MPD must meet the eligibility standards following the framework used by BPS. Therefore each quality dimension is used as the basis for implementing quality assurance which is carried out in a quality inspection scheme. For the MPD analysis in 2021, Statistics Indonesia uses data from June - July 2020 to determine quality assurance at the data collection stage.

2. Mobile Positioning Data

The increase in the use of cellular telephones, both number and area, and the continuously generated data can further analysis in research involving human movement [5]. For example, in collaboration with the largest cellular operator in Indonesia, BPS has been using mobile positioning data in several studies such as tourism [4, 6], commuter, and metropolitan area delineation [7].

Cellular phones produce data with the dimensions of time and geographical location, called Mobile Positioning Data. Cellular operators record the geographic location of cellular activity in Location Area Coordinates (LAC) and Cell Identity (CI). LAC identifies each location area within the Global System for Cellular Communications (GSM) Public Land Mobile Network (PLMN), while Cell Identity (CI) represents the radio network element in a cellular operator [8].

After translation to LAU, another critical step in analyzing MPD data is determining the usual environment. BPS uses the Anchor Mobility Data Analytic (AMDA) algorithm to determine a person's usual environment. This algorithm pays attention to the consistency of a person's presence in the data every day, week, and month. Therefore, completeness of information, availability of data in each period becomes essential for the successful establishment of AMDA. In ensuring that the available data meets the needs of BPS, a data inspection process is carried out as quality assurance as to the initial stage of data acceptance.

This study used three types of data from cellular phones, as described in Table 1.

Table 1. Three basic types of data sources generated from LBS

Source Type	Description	Alias
CHG	Billing domain log, which stores successful charging transaction records such calls, messaging, and other activity.	calls, SMS, and MMS logs
LBA	The technology is used to pinpoint consumers' locations and provide location-specific advertisements on their mobile devices.	signaling data
UPCC	It provides policy, service, subscription, quota, and bearer resource management functions and admission control for internet data usage.	internet data usage

These three data types are complementary, and LBA is the data type that has the highest number of rows of data. Therefore, in conducting data analysis, these three data are combined and considered a single data unit. BPS collaborates with the largest Mobile Network Operator (MNO) in Indonesia for data provision [9].

3. Definition and dimension of quality

The dimensions of quality as outlined in the BPS module "Statistical Quality Assurance Framework" consist of the following dimensions [10]:

3.1. Relevance

Relevance refers to the conformity of activity outputs with user needs, targeting the primary needs and several derivatives in the same context.

3.2. Accuracy

Accuracy refers to the output of statistical data that can accurately describe the conditions or phenomena in the observed object being measured. Traditionally, accuracy is usually described in statistics as sampling error and non-sampling error.

3.3. Punctuality and timeliness

Timeliness refers to the time difference between when statistics are collected and published. The shorter the time interval, the more time it is and the more value it adds to the results obtained. Punctuality refers to the difference between when data is first released and the target time scheduled for release as announced in the official release calendar or another similar term.

3.4. Interpretability

Interpretability reflects the extent to which the output of statistical activities can be presented clearly and easily understood by users. Interpretability can be determined by the availability of metadata, additional information, and support services for users to ask questions to gain a complete understanding and use statistical outputs effectively.

3.5. Accessibility

Accessibility refers to the ease with which users can access statistical output data. Accessibility includes the ease of users in using tools that can check the availability of the expected data, the suitability of the form as a data access medium, access fees, and the availability of various access options that users can use.

3.6. Coherence

Coherence refers to the domains/levels of statistical output at different levels but can be integrated and comprehensively described phenomena.

3.7. Comparability

Comparability can be equated with coherence but refers to output containing the same data items but differs in the period, region, or other relevant domain.

3.8. Trustworthiness

Trustworthiness is the level of trust of data users in recognizing and understanding the statistical output generated efficiently. Trustworthiness is also related to the image of the BPS institution as a data producer, which is the key to trust from users who want to take advantage of the output of statistical activities produced.

Viewed from the quality dimension, MPD as one of the uses of Big Data for official statistics, in this case, has several advantages, including the dimensions of relevance, accuracy, timeliness, and reliability. These advantages are obtained because MPD can produce more timely statistical outputs than conventional surveys. In addition, MPD also has a good guarantee of accuracy because it describes the actual movement conditions of the observed subject. On the other hand, the use of MPD shows that BPS has innovated by utilizing modern alternative data sources to enhance the reputation of BPS as a statistical institution that is adaptive to technological developments.

4. Quality Checking Scheme

The principle of quality assurance is to ensure that the data to be used is adequate for further analysis. Things of concern are the number of records per unit time, check the completeness of the data for each variable, and data patterns. This quality measurement metric guide is based on BPS's experience in conducting an assessment of MNO data that has been carried out since 2018. Details are described in the following table.

Table 2. List of indicators, level of importance, and description of indicators as quality assurance requirements

Indicators	Level of Importance	Description
1	Critical	Checking missing values (Out of total records, how many NULL values are in the dataset per mandatory field. Mandatory fields are mno_cell_id, lon, lat (or x, y). E.g., how many lac values are missing out of total records; how many latitude values are missing out of total records))
2	Critical	Records & subscriber per day (Number of records and unique subscribers per day group by lbs source/type)
3	Critical	Records & subscriber per month (Number of unique days, records, and subscribers per month)
4	Important	Average records & subscriber per weekly (Calculate average numbers of rows and unique subs every week)
5	Critical	Subscriber per AMDA (Anchor Mobility Data Analytic) steps
Check Cell Data (to master cells data)		
6	Critical	Count no. of unique cell locations per month
7	Critical	Count no. of unique kabupaten and kecamatan per day
8	Critical	How many cells have incorrect coordinates (are out of the country)? BPS map is used as a reference
9	Critical	How many of the cells have records in the domestic dataset
10	Critical	How many of the cells are missing from the cells table? (Master cells as reference)
Check Pattern		
11	Critical	On how many days domestic subscribers are present out of all days in the period.
12	Critical	The average number of records per hour (0-23).
13	Nice to have	The time gap between subsequent events.

Quality assurance checks are carried out in several ways, setting thresholds, checking data completeness, and paying attention to the distribution shape. The first two indicators are the basis for forming AMDA using thresholds. First, data were normalized for each category of observed variables. Then, outliers are checked. Finally, the input data needs further examination and explanation if it is below median minus three times interquartile range and is considered unreasonable (rejected) if it is below median minus five times interquartile range.

The first indicator describes the completeness of geographic information in the daily data. As we know, the primary information from this MPD includes time and geographical location. Human movement is highly dependent on place and time, so the missing value in this information must be controlled. However, not all LAC-CI can be translated into geographic coordinates. Based on the assessment of the data for the last two years, the completeness of information for this variable is 84%. Below 84%, operators should be given a warning that there is a problem with data retrieval. If the completeness of the data is below 77%, then this data cannot be used for analysis.

The second quality assurance indicator describes the daily attendance of subscribers. Normally, the number of subscribers every day is not the same because this depends on the cell phone user's activity. However, fluctuations in numbers must be controlled. The examination is carried out separately for each type of data, considering the unique Mobile Subscriber Integrated Services Digital Network Number (MSISDN) and the number of data rows each day. Observations were made on daily data by first

normalizing it to the number of customers at the beginning of the year. We get the results as shown in Table 3. Average daily subscriber attendance based on the number of unique MSISDN for data types LBA 0.92, CHG 0.87, and UPCC 0.92; while the threshold for data rejection is for LBA if < 0.88, CHG if < 0.83, and UPCC if < 0.88. The average daily data rows for the LBA data type are 0.88, CHG 0.56, and UPCC 0.83, while the threshold for assessing the data is for LBA if <0.82, CHG <0.38, and UPCC if < 0.77. This examination is applied to weekly data to check the fourth indicator.

Table 3. Threshold for Indicator 2

Data	Data Source	Warning	Reject
Records	CHG	0.56	< 0.38
	LBA	0.88	< 0.82
	UPCC	0.83	< 0.77
Subscriber	CHG	0.87	< 0.83
	LBA	0.92	< 0.88
	UPCC	0.92	< 0.88

The third indicator requires the availability of daily data for each data type. We do not calculate the third and fourth indicators threshold because it is already done on daily data. However, it is necessary to ensure that there is complete data available for all dates. Therefore, if any daily data is missing, this month's data is rejected.

Each stage of AMDA reduces data. In the fifth indicator, an examination is carried out on the proportion of data that make up the anchor. The proportion of data that forms the anchor to the initial data amount should not be less than 0.75.

In the sixth to tenth indicators, the position of the geographic coordinates in each data line is checked according to the location reference table (master cell) that MNO should maintain. In addition, the sixth and seventh indicators set a threshold for the number of locations in the data. For example, the sixth indicator requires at least 115,000 unique locations, and the seventh indicator requires that the number of sub-districts cannot be less than 6,718, and the number of sub-districts cannot be less than 514.

The eighth to tenth indicators explain the availability of LAC-CI to be translated into local coordinates, which are then used to determine Local Administration Units (LAU) consisting of Provinsi, Kabupaten, and Kecamatan. The NULL value in the data for this indicator cannot be more than 5%.

The eleventh indicator checks the completeness of the MPD data. The criteria for good data completeness are customers who have complete data (available every day of the month) more than those who do not. In addition, the top of the chart should be at the highest number of days in each month. For example, the top of the chart in June should be at 30 days.

The twelfth indicator aims to check the average number of data lines per hour for one month. This graph shows the pattern of subscriber activity each month that forms a unique pattern: the elephant curve [11]. The graph rises from eight in the morning to eight in the evening, when subscribers are usually active. Then it slowly descends into the wee hours of the morning when regular subscribers take a break and then rises back up towards activity time.

Of the thirteen quality assurance indicators, the critical indicators that determine whether the existing data sources are good enough to be analyzed are 1-12. The thirteenth indicator has nice-to-have criteria, providing more insight into the data, but is not mandatory.

5. Result and Discussion

The following are the results of quality assurance checks on MPD data for June-July 2020, which BPS and cellular operators carried out.

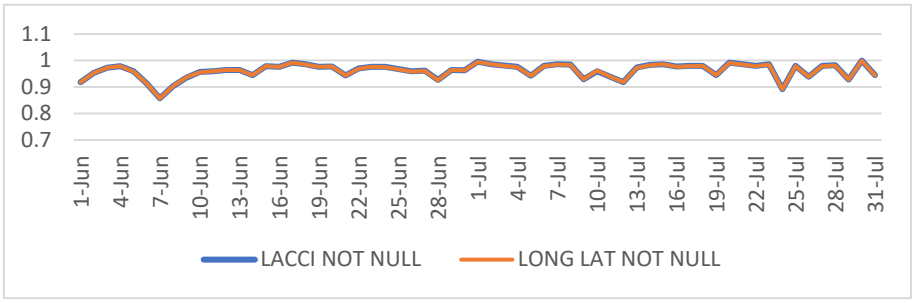
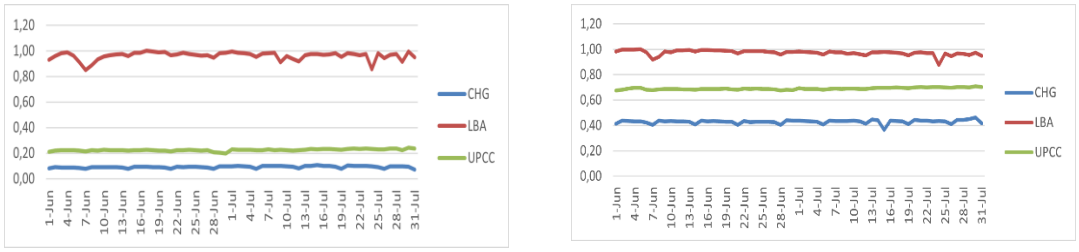


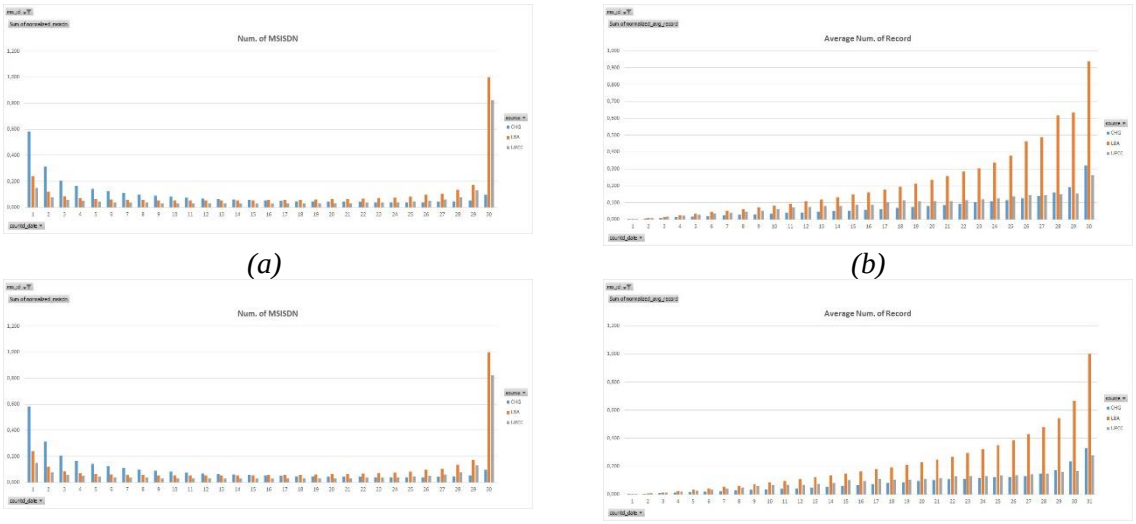
Figure 1. Normalization of the number of data with LACCI and LONG LAT is not null (First Indicator)

In general, the completeness of the data is good, with a daily average of 0.96. However, there is a point where the completeness of the data reaches 85.74% (7 June). This figure is close to the lower limit, at which the operator must explain the decrease in data.



(a) *(b)*

Figure 2. Normalization of the number of data (a) and the number of the subscriber (b) based on data source type (Second Indicator)



(a) *(b)*

(c) *(d)*

Figure 3. Normalization of unique MSISDN (June (a), July (c)), and average number of records per number of days (June (b), July (d)) in 1 month per source type (Third Indicator)

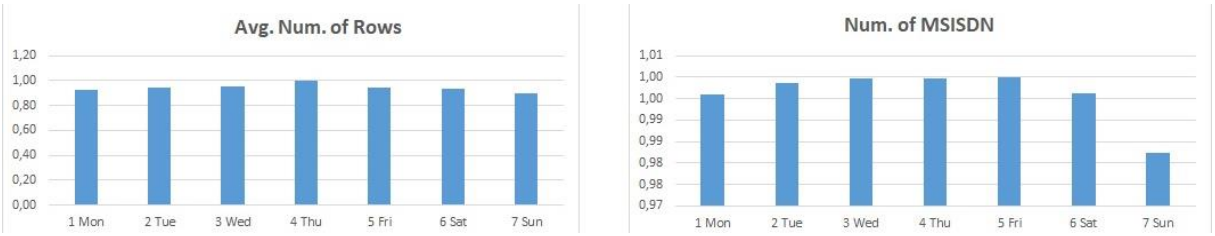


Figure 4. Normalization of the number of data and number of the subscriber on weekly data (Fourth Indicator)

The number of records and subscribers every day tends to be stable. The image above is when normalization is performed on all data types. We can see that from Figure 2, the LBA data source dominates the data availability. For unique MSISDN, there are different patterns based on the existing data source type (LBA, CHG, UPCC). The unique MSISDN from CHG shows more subscribers who only make active transactions (call or SMS) a few times a month, while for LBA, more subscribers have complete data (Figure 3). The three data types are complementary, so the more data we can use, the better. The weekly data shows that the number of records is also relatively stable, while the number of unique subscribers on weekends tends to be lower (Figure 4). However, the fourth indicator is only used as an insight because the daily data is already above the threshold.

Not all customer data can form the usual environment. For June and July 2020, the percentage of subscribers who have the usual environment is in the range of 0.78. This figure is still above the data rejection threshold (below 0.75). The following vital point is location; from mobile location data, we check how many unique location points, sub-districts, and districts are recorded is checked. Based on the experience of previous studies, the threshold is set at a minimum of 115,000 location points and 6,718 Kecamatans. The selection of MNOs invited to cooperate by BPS requires the presence of MNOs in all districts in Indonesia; logically, the number of locations does not decrease. The examination results showed that in June 2020, there were 132,215 location points and 6,718 Kecamatans, while in July, there were 132,754 locations, 6,718 Kecamatans.

Completeness of data is essential in analyzing MPD. The criteria for good data completeness are customers who have complete data (available every day of the month) more than those who do not. The top of the chart should be at the highest number of days in each month. From Figure 5, we can see that the chart's peak is at 30 days for June, while for July, it is at 31 days.

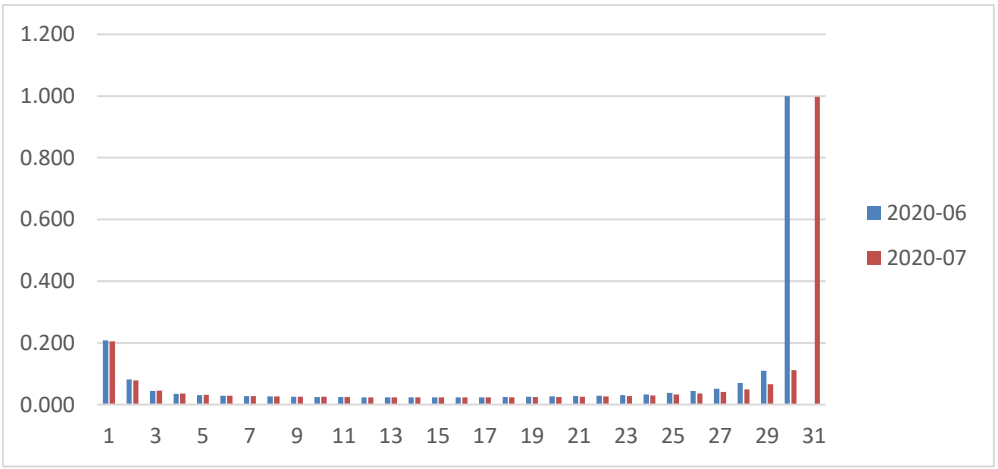


Figure 5. Normalization of the number of days domestic subscribers are present out of all days (Eleventh Indicator)

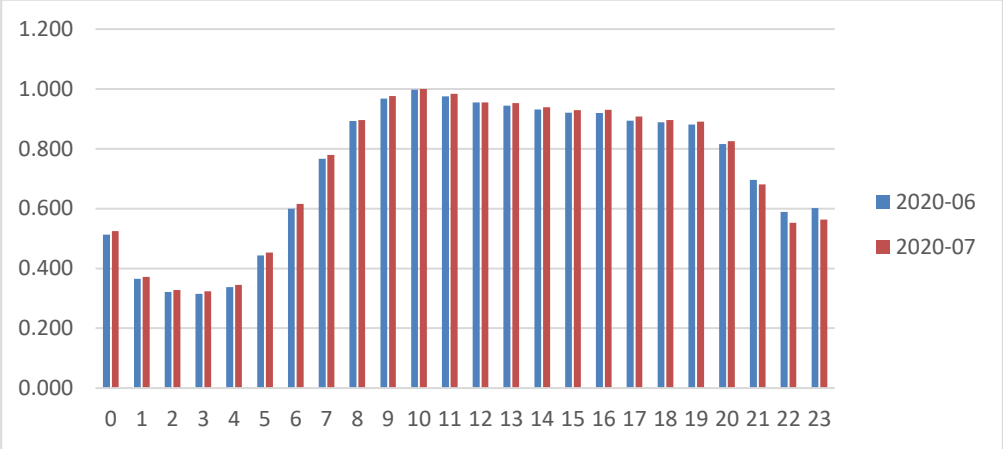


Figure 6. Normalization of the average number of records per hour (Twelfth Indicator)

Indicator 12 can also be said as the natural hourly rhythm of cellular phone users in Indonesia. The distribution of diurnal records shows that peak activity in the morning occurs at 9-10 and in the afternoon occurs at 18-19. And as we can see in Figure 6, the shape of the distribution follows the elephant's curve. The process of checking data quality in June - July 2020 is required for further cooperation in 2021. Overall, based on the results of an examination of 12 of the 13 indicators that are critical (the thirteenth indicator is not available), the quality of MNO data is good and meets the requirements for further analysis.

6. Conclusions and suggestions

The results of quality assurance checks on MPD data for June – July 2020 indicate that the data can be used for further analysis. However, quality assurance checks must be carried out every month to ensure that the data condition is maintained and according to the standards that have been previously set.

BPS has plans to collaborate with other MNOs, and it is necessary to adjust the indicators and their thresholds based on the type of data that the MNO has.

References

[1] Patgiri R, Ahmed A. Big data: The v's of the game changer paradigm. 2016 IEEE 18th International Conference on High-Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)2016. p. 17-24.

[2] Ellars S. insideBIGDATA [Internet]2013. [cited 2021]. Available from: https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/.

[3] Reimer AP, Madigan EA. Veracity in big data: How good is good enough. Health Informatics J. 2019;25(4):1290-8.

[4] Lestari TK, Esko S, Sarpono, Saluveer E, Rufiadi R. Indonesia's experience of using signaling mobile positioning data for official tourism statistics. 15th Global Forum on Tourism Statistics; Cusco, Peru: OECD; 2018.

[5] Ahas R, Aasa A, Silm S, Aunap R, Kalle H, Mark Ü. Mobile positioning in space–time behavior studies: Social positioning method experiments in Estonia. Cartography and Geographic Information Science. 2007;34(4):259-73.

[6] Widyasanti AA, Munaf AR, Esko S, Tiru M, Lestari TK. The use of mobile positioning data to measure visitors of multisport events: A case study of Asian games 2018 in Indonesia. 2020 Asia-Pacific Statistics Week; 15 - 19 June 2020; Bangkok, Thailand: United Nations ESCAP; 2020.

[7] Prabawa P, Soblia H, Amin Y, Albertha W, Setiawan E. The use of mobile positioning data (mpd) to delineate metropolitan area in indonesia: Case study in cekungan bandung. 2020 Asia-Pacific Statistics Week; 15 - 19 June 2020; Bangkok, Thailand: United Nations ESCAP; 2020.
[8] ETSI. European digital cellular telecommunications system (phase 2); numbering, addressing, and identification (gsm 03.03 version 4.10.1). France: European Telecommunications Standards Institute 2000. Contract No.: ETSI EN 300 523 V4.10.1 (2000-11).
[9] Nugroho ARS, Munaf AR, Madjida WOZ, Putra APP, Setyadi IA. Home and work identification process using mobile positioning data. CONFERENCE OF EUROPEAN STATISTICIANS: Expert Meeting on Statistical Data Collection: UNECE; 2021.
[10] Statistical quality assurance framework (stat-qaf). In: Building PMUISC, editor. Jakarta, Indonesia: Statistics Indonesia; 2011.
[11] Piela P. Non-traditional data sources in social statistics of Statistics Finland. Non-traditional data sources in the National Statistical Systems, 17th Meeting of ECLAC; 1 - 2 October 2018; Santiago de Chile: ECLAC; 2018.

Acknowledgments
The study of mobile positioning data as the data source of official statistics was carried out by BPS and Telkomsel. This cellular operator is a subsidiary of Telkom, a State-Owned Enterprise.