Big Data in Indonesia

Using mobile phone data to estimate SDG indicators

May 2021

Contents

Figures

Tables

1. Background

Since 2016, Indonesia has adopted the 17 Sustainable Development Goals (SDGs) and integrated them into national development plans. Since 2017, Statistics Indonesia (BPS) has published the Indicators of Indonesia's Sustainable Development Goals, including province-level SDG indices and regency-level indices at the subnational level. All SDG indicators are monitored to the regency (LAU2) level, including those relevant to information and communication technology (ICT) development (SDG 5.b.1, 9.c.1, and 17.8.1).

ICT development indicators are primarily derived from household surveys (the National Socio-Economic Survey, SUSENAS) and administrative data from the Ministry of Communication and Informatics (MCI). Some indicators are not yet available at the regency level. For SDG indicator 9.c.1, BPS relies on MCI data, but only national aggregates are provided and the timeliness is limited—data are typically available with a six-month lag due to mobile network operators' (MNOs) audit cycles.

Indicators 5.b.1 and 17.8.1 come from SUSENAS, conducted annually. In recent years the sample has expanded to roughly 330,000 households, enabling estimation at regency level[1]. However, household surveys can under-represent higher-income groups due to non-response, potentially biasing estimates toward middle- and lower-income populations. The survey's strength is its detailed questionnaires, allowing cross-tabulation by gender, age, income and other characteristics. However, as with most household surveys, timeliness is a weakness considering the significant processing time required from data collection to obtaining the final result.

In addition, the Indonesian Internet Service Providers Association (APJII) runs a survey that generally yields a higher share of individuals using the Internet than SUSENAS. The need for comparable alternative data motivated Indonesia's participation in the second phase of the ITU project on using big data to measure the information society. Mobile positioning data can help address coverage gaps and provide a basis for triangulation across sources.

Population data are also required to measure SDG and ICT indicators. This project compares population estimates from WorldPop (1×1 km grid) with BPS population projections and mobile positioning–derived home locations.

The analysis was conducted and led by Badan Pusat Statistik (BPS) Indonesia, with support from the International Telecommunication Union (ITU). The Ministry of Communication and Informatics (MCI) provided reference data and valuable insights throughout the implementation, contributing to the successful execution of the pilot.

2. Data access

Statistics Indonesia (BPS) collaborates with mobile network operators (MNOs) under Memorandum of Understanding (MoUs) grounded in Statistics Law No. 16/1997. Although the law does not yet grant BPS direct statutory access to operator data, the MoUs—implemented through annual contracts—provide the legal and operational basis for data sharing. The MoUs set clear objectives and minimum output expectations aligned with current and future statistical needs. They

[1] Indonesia has 34 provinces that consist of 514 Kabupaten/Kota (regency) break down to 7.217 Kecamatan (districts).

have enabled the production of several official statistics, including inbound tourism at border-crossing points (since 2016), domestic tourism and commuting statistics (since 2017), and outbound tourism statistics (since 2019).

BPS data scientists access a secure MNO sandbox either on-site or remotely via a virtual private network (VPN). Within this environment, they query and transform datasets for statistical use under strict confidentiality safeguards. Personally identifiable information (e.g., phone numbers) is encrypted or hashed, and all staff with access are bound by non-disclosure agreements (NDAs). This access model supports a controlled, auditable data pipeline—from secure sandbox access, through encrypted processing by authorized personnel, to the release of independent statistical outputs—while upholding privacy and ethical standards.

3 Data sources

This project uses mobile positioning data (MPD) from one of Indonesia's largest mobile network operators, with an average market share of about 60%. The dataset spans the full year 2019 and covers all 514 municipalities, supporting the development of a home-anchor model. To ensure robustness and reliability, the model is augmented with additional data sources and methods.

Administrative boundary data—provided through collaboration with the Ministry of Internal Affairs and other stakeholders and typically derived from population censuses - add essential geographic context for analysis. The MPD structure includes key variables such as MSISDN (a unique user identifier), geographic coordinates (latitude/longitude), cell ID (specific location identifier), and technology (infrastructure type). These attributes strengthen georeferencing and illuminate technology-use patterns across regions.

Integration draws on multiple reference datasets: a Digital Elevation Model (DEM), population projections, population data at 1×1 km grid level from WorldPop[2], and Indonesia's National Socioeconomic Survey (SUSENAS). A blend of techniques, chosen for data compatibility and analytical coherence, was used to combine these sources, enabling a richer, more nuanced understanding of the landscape and a comprehensive methodology.

4 Home-anchor model

This study uses the AMDA-Home algorithm for home detection—an advanced modification of the Anchor model proposed by Ahas (2010). AMDA-Home was chosen for its robustness with location data and reliability in identifying home locations. The algorithm first standardizes raw timestamps over a 24-hour window: if an hour value exceeds 12, it is transformed using a defined mathematical conversion; if it is 12 or below, the value is retained. After this timestamp transformation, the algorithm computes the frequency of each subscriber's presence across local administrative units (LAUs), specifically at the LAU2 regency level. These frequency counts provide quantitative evidence of habitual presence and are used to infer the likely home location within the 24-hour observation frame.

[2] https://www.worldpop.org/

5 Quality assurance

Quality assurance (QA) is the systematic process that ensures data used for analysis are fit for purpose—accurate, complete, and consistent. Emphasizing completeness, accuracy, and pattern identification supports effective decision-making and provides a solid foundation for subsequent analytical stages. For mobile positioning data (MPD), high data quality is paramount to producing reliable, meaningful insights; therefore, a stringent QA protocol is integral to the workflow. This protocol confirms that the data are suitable and sufficient for downstream analysis.

Key QA metrics for MPD include: the number of missing values (a direct measure of completeness); daily counts of records and unique subscribers (indicators of data volume and user activity that require regular monitoring); and the average number of records per hour (useful for understanding daily patterns and diagnosing issues at specific times). Table 1 describes the checks in more detail.

Table 1. Examples of quality assurance checks

Indicators	Level of Importance	Description
Records / events checks		
1	Critical	**Missing values in mandatory fields** (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	**Daily count of records and subscribers** (Number of records and unique subscribers per day group by lbs source/type)
3	Critical	**Monthly count of records and subscribers** (Number of unique days, records, and subscribers per month)
4	Important	**Average weekly records and subscribers** (Calculate average numbers of rows and unique subs)
5	Critical	**Subscriber per AMDA steps**
Cell Data (to master cells data)		
6	Critical	**Number of unique cell locations per month**
7	Critical	**Coverage of kabupaten / kecamatan per day**
8	Critical	**Cells with coordinates outside national boundaries** (validated against administrative geodata)
9	Critical	**Cells with records in domestic dataset**
10	Critical	**Cross-checks between cell master registers and observed cells in domestic dataset**
Other pattern checks		
11	Critical	**Share of days with subscriber presence across the period** (the share of days domestic subscribers are present for all days in the period).
12	Critical	**Diurnal distribution** (average number of records by hour, 0-23)
13	Nice to have	**Average time gap between subsequent events.**

Prior to the calculation of ICT indicators for SDGs, input raw data checking is conducted. In general, all quality checking passed the standard. Figure 1 shows a steady flow of data from three sources: CHG (charging) data—i.e., call detail records (CDR); UPCC (Unified Policy and Charging Controller) data; and LBA (location-based advertising) data. Figure 2 illustrates that these sources are complementary—the more sources combined, the better the coverage. In LBA and UPCC, subscribers with 28–30 active days per month are more prevalent, while in CDR the distribution is skewed toward 1–2 active days. Figure 3 shows the diurnal distribution of records in Indonesia, with pronounced peaks around 09:00–10:00 and 18:00–19:00.

Figure 1. Daily data flow across sources (QA #2)

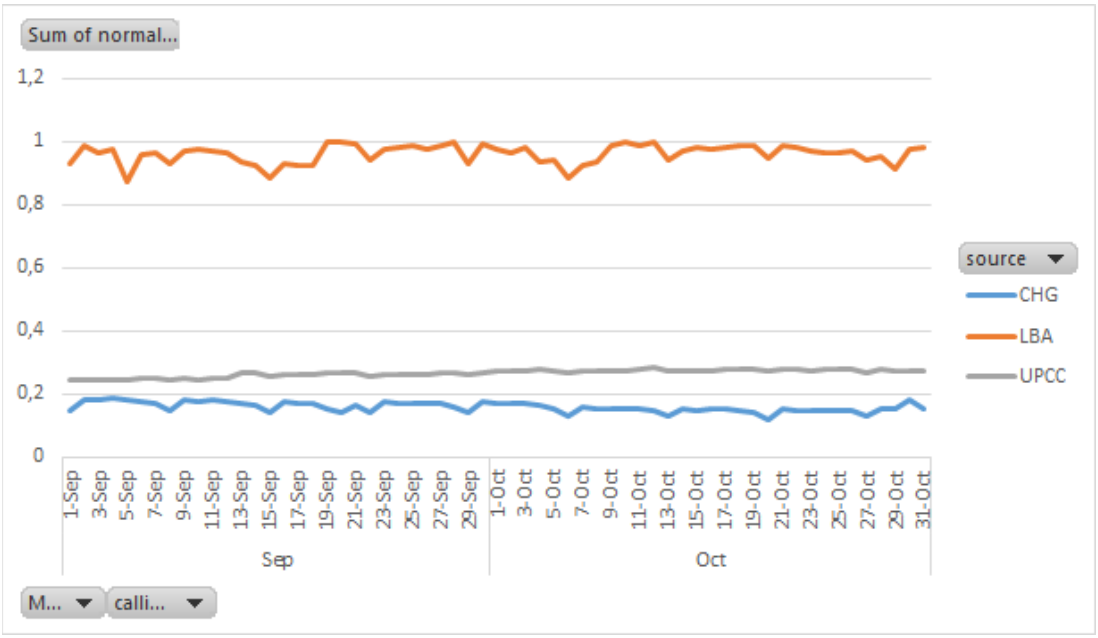


Figure 2. Share of days with subscriber presence (QA #11)

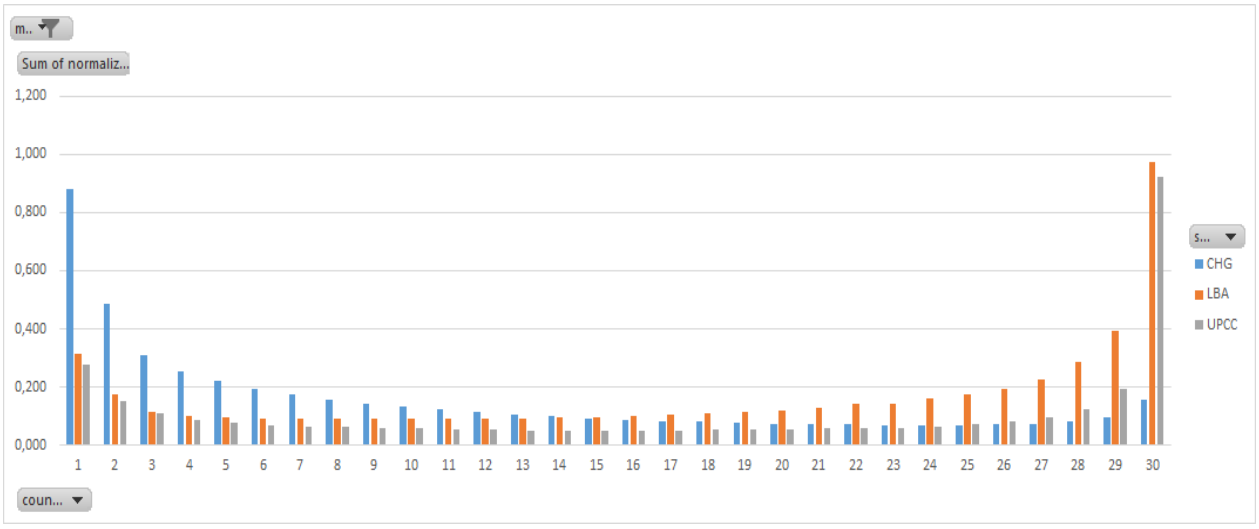
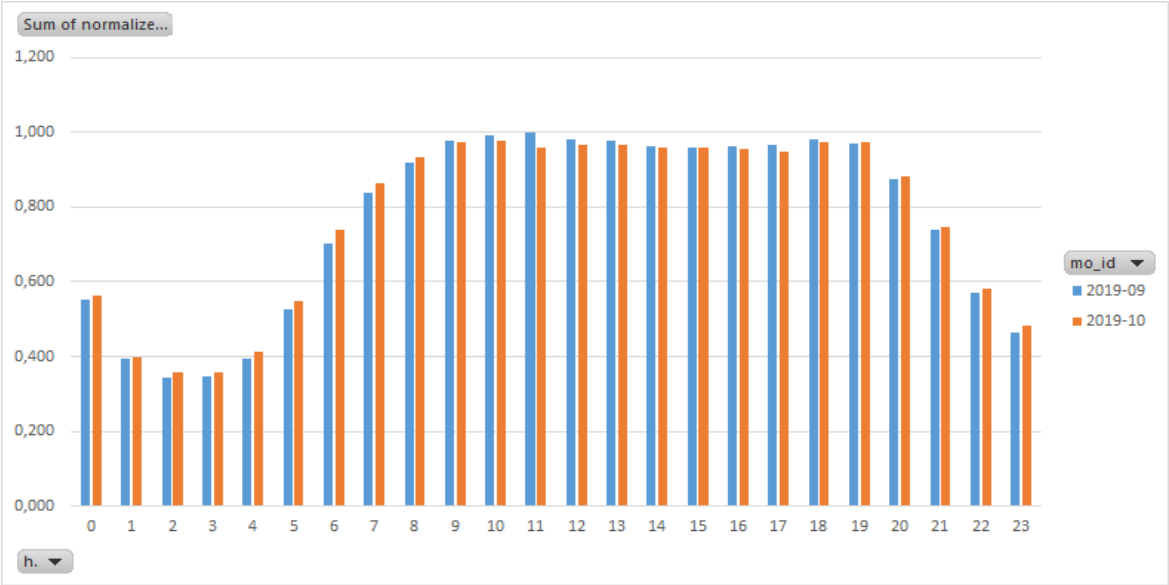


Figure 3. Diurnal distribution of records (QA #12)



6 Reference data

Two LAU levels are used for Indonesia: 34 provinces (LAU1) and 514 municipalities/cities (LAU2). Administrative boundaries at the municipality level were obtained from BPS, together with digital elevation map (DEM) at 30m resolution.

MCI maintains a registry of cellular sites submitted by MNOs. These were compared with OpenCellID and found to be concordant for the areas assessed. Others reference data that also included were Population Projection and WorldPop population data at 1x1 km grid population.

Figure 4. WorldPop 1x1 km grid population (national)

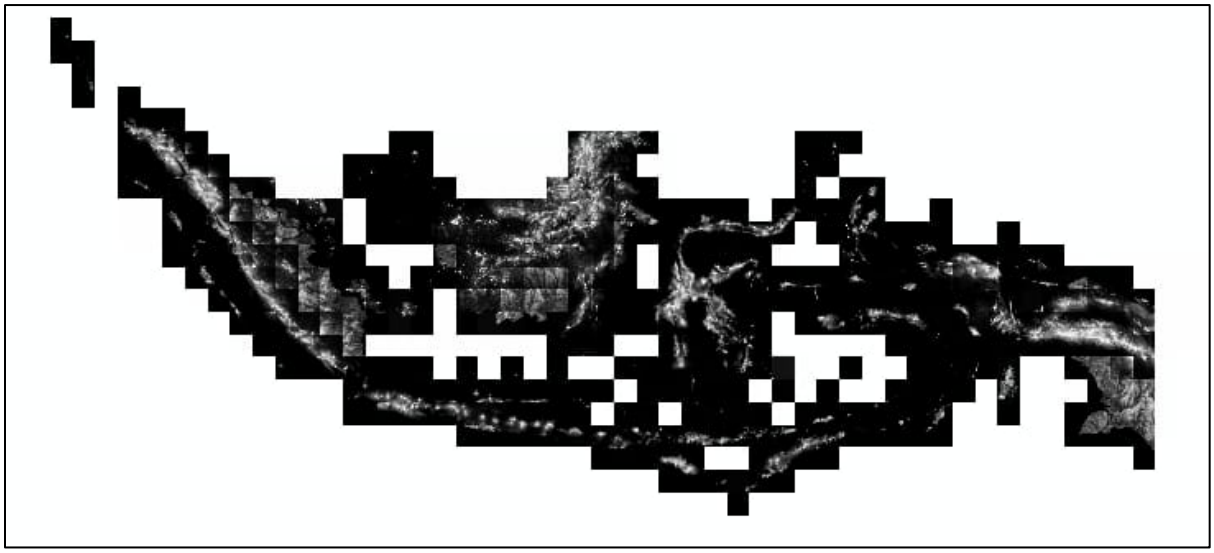


Figure 5. Digital Elevation Model (Bali example)

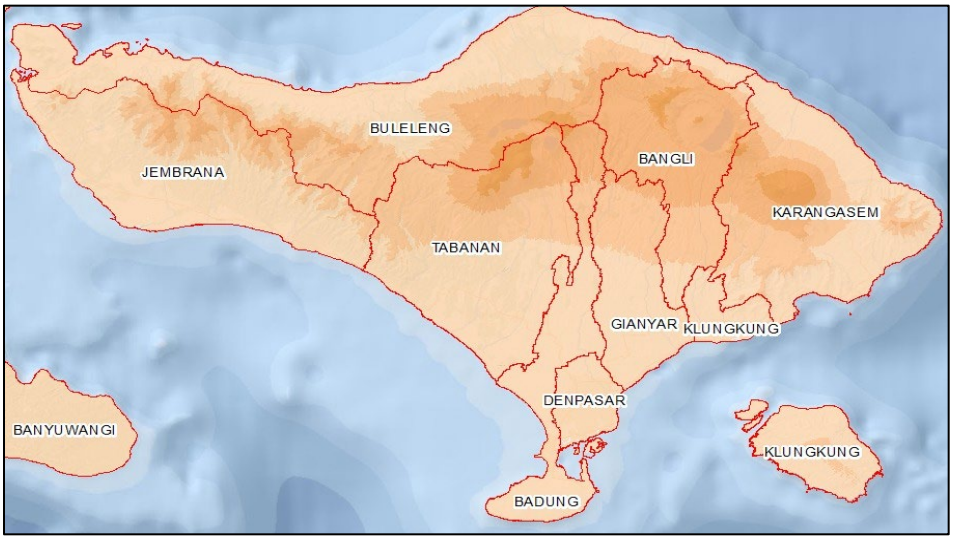


Figure 6. Cellular sites from MCI and OpenCellID

7 Results

7.1 SDG indicator 9.c.1 — Population within range of a mobile-cellular signal

The indicator measures the proportion of inhabitants within range of a mobile-cellular signal in a given area. Using the methodology described earlier—with Indonesia-specific parameters informed by local studies—the assumed coverage radii were 10 km (2G), 5 km (3G), and 3 km (4G), with a nominal antenna height of 50 m.

At the national level, Table 2 shows the estimated shares of the population covered were 77.70% (2G), 72.34% (3G), and 67.83% (4G). Coverage is highest on Java and lowest in Papua, reflecting population density and network deployment patterns. Spatial coverage maps (2G/3G/4G) are available for all provinces.

Table 2. The percentage of the population covered by a mobile cellular network, by technology

Province Code	Province Name	SDG 9.c.1		
		2g (10 km)	3g (5 km)	4g (3 km)
(1)	(2)	(3)	(4)	(5)
11	ACEH	65.17	57.57	56.9
12	SUMATERA UTARA	73.49	66.94	65.65
13	SUMATERA BARAT	69.66	61.34	58.81
14	RIAU	56.65	48.53	45.62
15	JAMBI	54.85	48.61	46.32
16	SUMATERA SELATAN	61.25	54.84	51.28
17	BENGKULU	60.54	51.64	49.18
18	LAMPUNG	68.94	64.08	59.6
19	KEPULAUAN BANGKA BELITUNG	63.97	58.52	52.56
21	KEPULAUAN RIAU	79.69	76.18	71.85
31	DKI JAKARTA	99.93	99.83	99.82
32	JAWA BARAT	90.99	88.01	82.55
33	JAWA TENGAH	86.63	81.75	74.05
34	DI YOGYAKARTA	91.15	88.07	83.98
35	JAWA TIMUR	86.05	81.33	74.05
36	BANTEN	90.85	89.85	86.51
51	BALI	96.22	94.47	92.37
52	NUSA TENGGARA BARAT	86.42	81.35	78.67
53	NUSA TENGGARA TIMUR	61.38	42.23	40.51
61	KALIMANTAN BARAT	49.17	42.52	37.46
62	KALIMANTAN TENGAH	33.95	27.27	23.36
63	KALIMANTAN SELATAN	71.8	66.29	62.5
64	KALIMANTAN TIMUR	64.27	60.56	56.06
65	KALIMANTAN UTARA	44.84	43.4	41.19
71	SULAWESI UTARA	74.68	66.53	62.72
72	SULAWESI TENGAH	46.27	35.83	34.64
73	SULAWESI SELATAN	73.39	66.9	63.98
74	SULAWESI TENGGARA	51.42	41.47	39.95
75	GORONTALO	65.46	59.24	57.55
76	SULAWESI BARAT	42.35	37.14	35.37
81	MALUKU	69.03	36.02	38.4
82	MALUKU UTARA	74.35	26.95	31.39
91	PAPUA BARAT	26.01	25.3	11.29
94	PAPUA	15.79	12.31	6.71
	INDONESIA	**77.70**	**72.34**	**67.83**

Figure 7. The percentage of the population covered by a 2G mobile cellular network

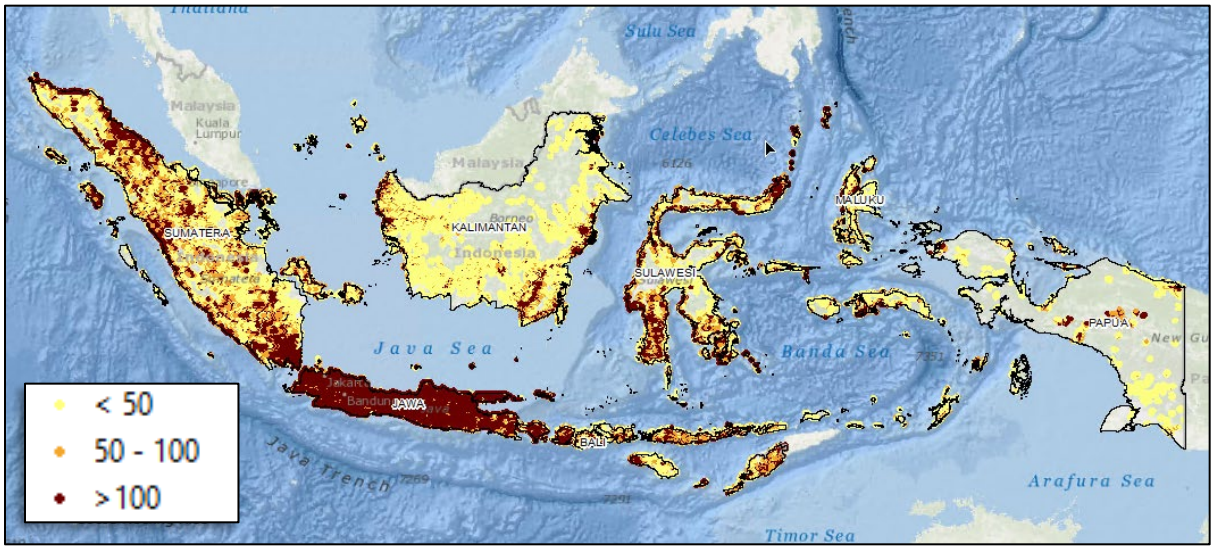


Figure 8. The percentage of the population covered by a 3G mobile cellular network

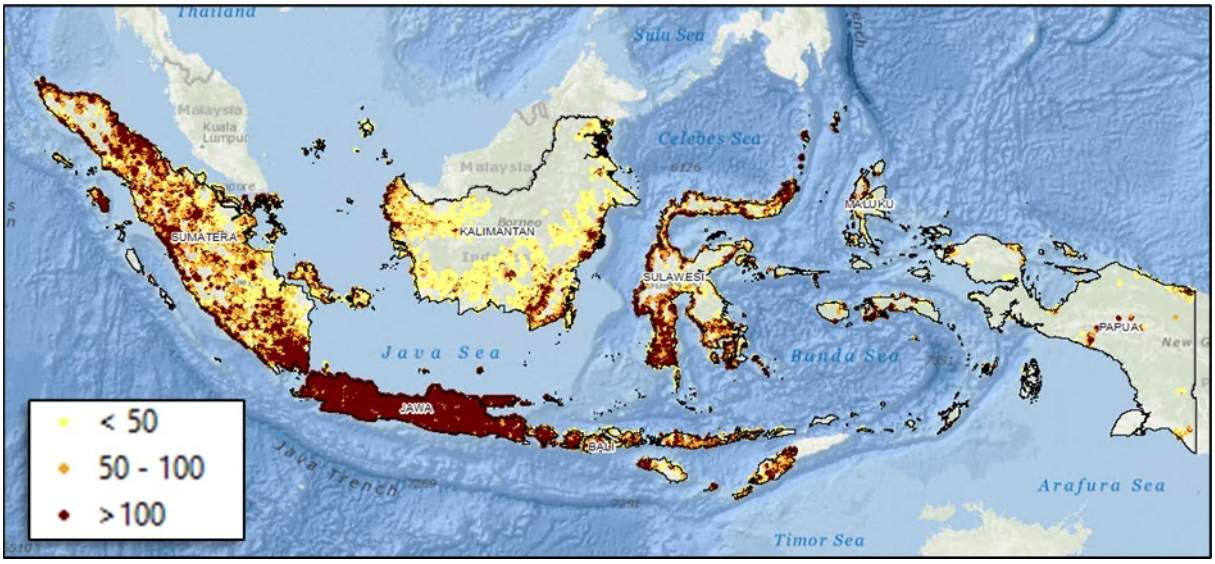
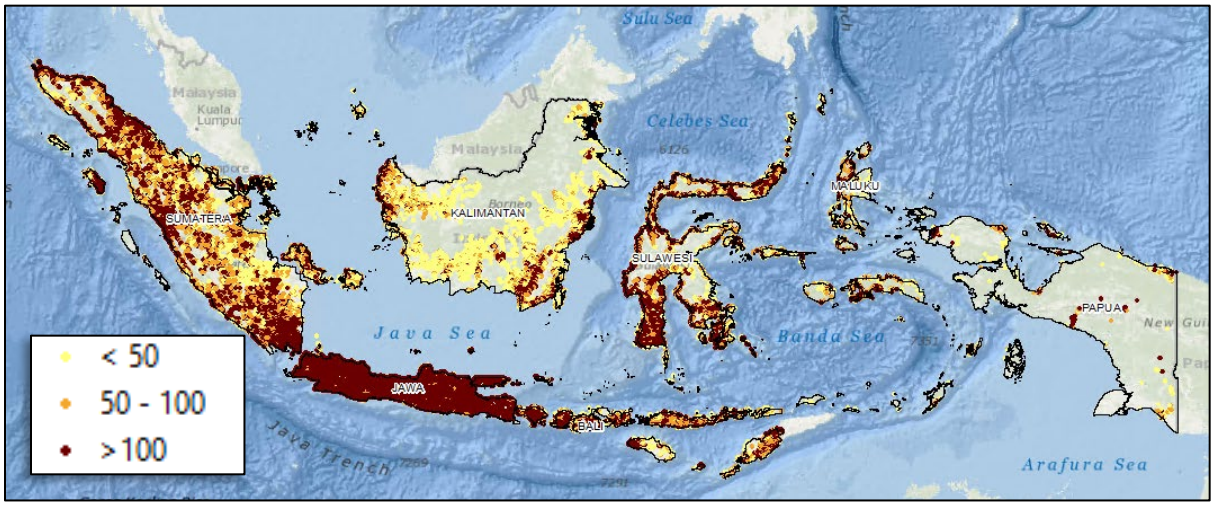


Figure 9. The percentage of the population covered by a 4G mobile cellular network



The table below presents detailed, municipality-level results for Bali Province: the percentage of the population covered by a mobile-cellular network. Denpasar, the provincial capital, has the highest coverage across all technology generations, while Klungkung—the smallest municipality in Bali—has the lowest coverage.

Table 3. The percentage of the population covered by a mobile cellular network, by technology, Bali province

Municipal ID	Municipal Name	SDG 9.c.1		
		2g (10 km)	3g (5 km)	4g (3 km)
(1)	(2)	(3)	(4)	(5)
5101	JEMBRANA	92.41	88.22	85.19
5102	TABANAN	96.28	93.69	91.06
5103	BADUNG	99.02	98.91	98.41
5104	GIANYAR	98.13	98.57	98.31
5105	KLUNGKUNG	85.25	87.27	83.67
5106	BANGLI	91.35	88.14	82.24
5107	KARANGASEM	93.76	84.66	77.95
5108	BULELENG	93.15	90.56	87.07
5171	DENPASAR	100.00	100.00	100.00

Figure 10. The percentage of the population covered by a 2G mobile cellular network, Bali province

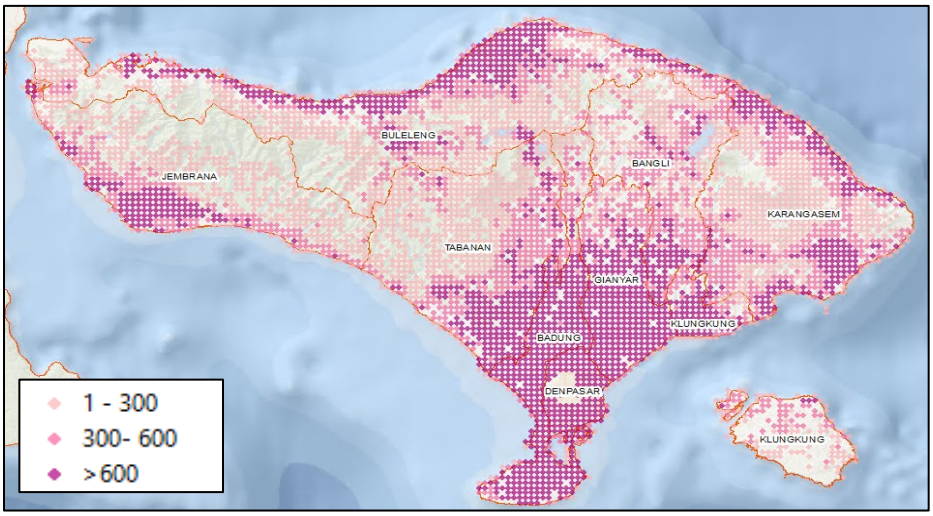


Figure 11. The percentage of the population covered by a 3G mobile cellular network, Bali province

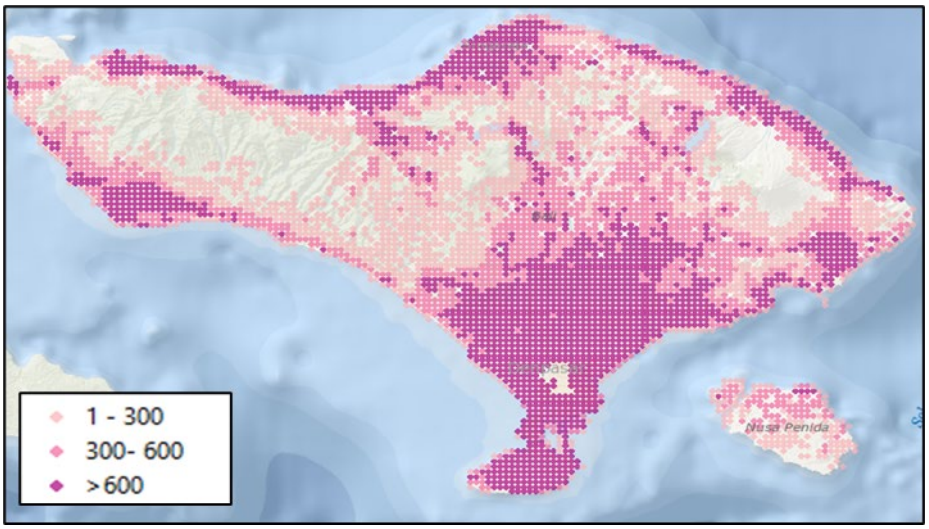
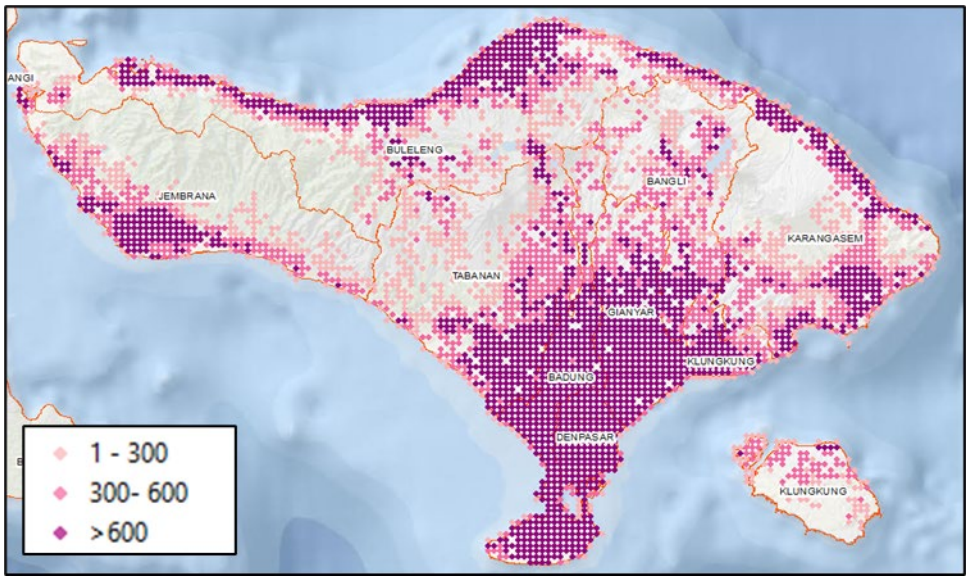


Figure 12. The percentage of the population covered by a 4G mobile cellular network, Bali province



7.2 SDG indicator 17.8.1 — Individuals using the Internet

Traditionally, SDG 17.8.1 is estimated from SUSENAS. In this project, BPS intended to combine MNO-provided home counts by technology generation with SUSENAS to refine estimates; however, only total home counts were available. As a pragmatic approximation, the proportion of individuals using a cellular phone from SUSENAS was applied to total homes to derive Internet-user estimates. Two alternative methods were tested, both yielding higher levels than SUSENAS alone, consistent with the expectation of some over-coverage (e.g., 2G-only devices lacking Internet access).

Table 4. Comparison of Household Survey (Susenas) with Mobile Data, Bali Province (Method 1)

Municipalities/Cities	SDG 17.8.1	
	SUSENAS	MOBILE DATA
(1)	(2)	(3)
5101. JEMBRANA	45,86	60,98
5102. TABANAN	48,06	57,56
5103. BADUNG	67,62	69,12
5104. GIANYAR	54,81	62,52
5105. KLUNGKUNG	47,67	55,48
5106. BANGLI	37,38	51,62
5107. KARANG ASEM	35,06	57,72
5108. BULELENG	38,66	55,75
5171. DENPASAR	74,04	77,13
Total	**54,08**	**64,89**

Figure 13. Comparison of Household Survey (Susenas) with Mobile Data, Bali Province (Method 1)

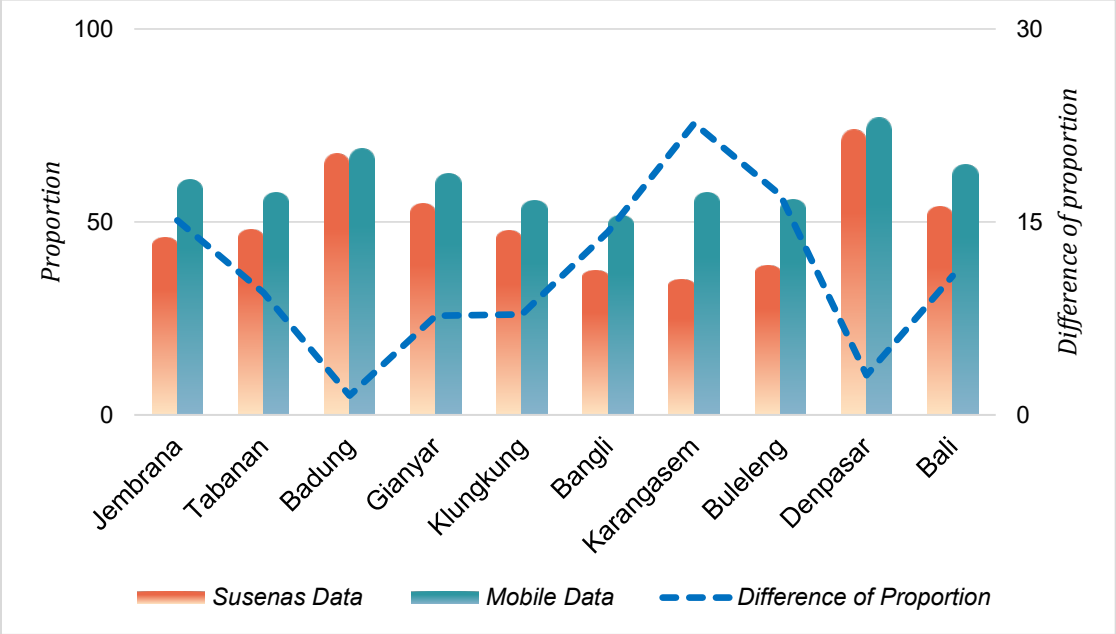


Figure 14. The percentage of the population using the Internet, Bali province (method 1)

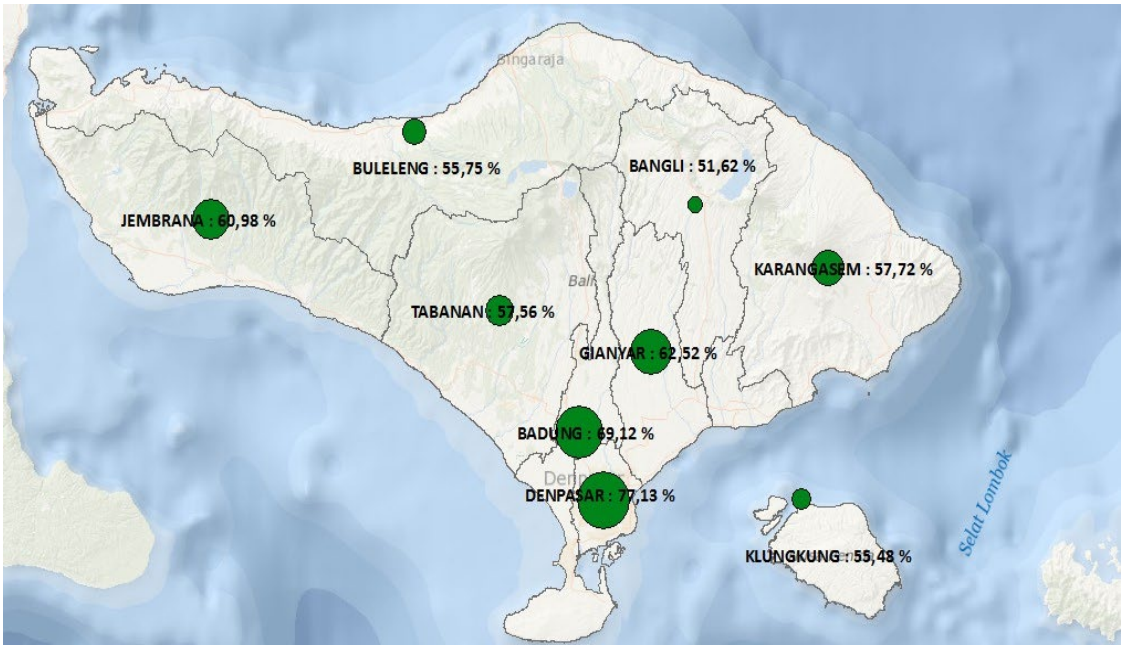


Table 5. Comparison of Household Survey (Susenas) with Mobile Data, Bali Province (Method 2)

Municipalities/Cities	SDG 17.8.1	
	SUSENAS	MOBILE DATA
(1)	(2)	(3)
5101. JEMBRANA	45,86	56,52
5102. TABANAN	48,06	66,45
5103. BADUNG	67,62	82,61
5104. GIANYAR	54,81	70,80
5105. KLUNGKUNG	47,67	67,79
5106. BANGLI	37,38	56,05
5107. KARANG ASEM	35,06	46,20
5108. BULELENG	38,66	52,53
5171. DENPASAR	74,04	83,35
Total	**54,08**	**68,73**

Figure 15. The percentage of the population using the Internet in Bali province (method 2)