LPEM-FEBUI Working Paper - 045
March 2020

ISSN 2356-4008

MODERNIZING OFFICIAL STATISTICS WITH BIG DATA: A CASE ON PODES

Chaikal Nuryakin
Nandaru Annabil Gumelar
Muhammad Dhiya Ul-Haq
Riefhano Patonangi
Andhika Putra Pratama

LPEM-FEB UI Working Paper 045

Chief Editor : Riatu M. Qibthiyyah
Editors : Kiki Verico
Setting : Rini Budiastuti

Institute for Economic and Social Research
Faculty of Economics and Business
Universitas Indonesia (LPEM-FEB UI)

Salemba Raya 4, Salemba UI Campus Jakarta, Indonesia 10430
Phone : +62-21-3143177
Fax : +62-21-31934310
Email : lpem@lpem-feui.org
Web : www.lpem.org

LPEM-FEB UI Working Paper 045, March 2020
ISSN 2356-4008

Modernizing Official Statistics with Big Data: A Case on PODES

Chaikal Nuryakin[1,★], Nandaru Annabil Gumelar, Muhammad Dhiya Ul-Haq, Riefhano Patonangi, & Andhika Putra Pratama

Abstract
This study serves as an example of how Indonesia can improve its official data by using big data. In this case, we compare village potential data (PODES) published by Statistics Indonesia (BPS) with Google Places API data and ministerial data. We use the number of hospitals, high schools, and public health centers within Jakarta province as the variables. The result shows that despite counting for the same thing, there are discrepancies between all three sources with a varying margin for each variable. We discuss our findings and give suggestions in the hope of improving official data in Indonesia, which could be helped by utilizing big data as this study exemplified.

JEL Classification: C8; D8; E2; R1

Keywords
official statistics — big data — official data

[1] *Lembaga Penyelidikan Ekonomi dan Masyarakat (LPEM), Fakultas Ekonomi dan Bisnis, Universitas Indonesia*
[★] **Corresponding author**: LPEM FEB UI. Gd. Ali Wardhana, Kampus UI Salemba, Jalan Salemba Raya No. 4, Jakarta Pusat 10340. Email: chaikall@gmail.com.

1. Introduction

Technology has been evolving rapidly in the past few decades. An essential part of that rapid development is the digitization of analog technology. With digitization, a whole slew of information is now easier to track and record. Furthermore, as technology progress further and swiftly penetrates the daily life of the populous from updated business systems, social networks to the internet of things, more and more data by-product is generated into what commonly known as big data. The plethora of information provided through big data could tremendously benefit policymaking by providing information in a timelier manner, by generating different insights, and as an alternative source of data to official statistics. (Cornelia et al., 2017).

Making an informed decision is an indispensable part of life and it is especially critical for a public policy where the livelihood of many people is at stake. In most cases, policymakers use official data as the basis or at least to substantiate their decision making. In Indonesia, one of the official data used by policymakers is *Statistik Potensi Desa* (PODES) or village potential statistics. PODES is a series of statistical data published by *Biro Pusat Statistik* (BPS), the country's statistic institution. PODES is the only source of regional data that contains myriads of in-depth information in regards to regional development in Indonesia that are not done through the household approach. Other government bodies like the ministries do gather their own data, which are done through various methods. However, most if not all of those data are relatively limited, and much data is only available in PODES. PODES are done primarily for the government to use as their primary source of information to plan and evaluate regional development all over Indonesia. Furthermore, PODES has also been used either as the source of primary data or substantiating source of data in a plethora of academic and non-academic studies. Thus, the importance of PODES cannot be understated.

"*Data is now available faster, has greater coverage and scope, and includes new types of observations and measurements that previously were not available.*" (Einav & Levin, 2014). With the rapidly increasing data production on top of the ongoing swift progression in big data analytics throughout the past decade, more and more people opted to use big data for sources of information for researches. Some people start to question the quality and validity of big data with its more modern information-gathering methods compared to a more conventional method currently being used by BPS for official statistics data such as PODES.

However, making use of big data is not particularly easy as it requires an additional set of skills and tools compared to the operational requirement around traditional official data; "*Multidisciplinary teams will be needed to make big data speak*" (Cornelia et al., 2017). Thus, before any institution in Indonesia starts on a potentially costly endeavor, this study serves as an example of how Indonesia can improve the strengths and weaknesses of its official statistics in the form of the village potential statistics.

This study aims explicitly to compare official data, which is PODES in this case, with big data and other official data from different sources. By comparing them, not only can we assess the reliability of PODES, but we can also validate whether it is possible to update the infrequently published but frequently used PODES. It should be acceptable just to update PODES with other official data when possible. Nevertheless, official data for districts are not always available; then, we could rely on big data to update. The variables we chose to compare are the number of public facilities such as high schools, hospitals, and public health centers. This study will compare those variables between PODES 2018, corresponding ministerial data as the other official data source, and google data collected through google places API as the representative of big data in this study.

2. Official Statistics, PODES, and Big Data

In this section, we will focus on the in-depth literature review of the Indonesian official statistics, big data and its usage, and Indonesia's Central Bureau of Statistics (BPS). First, we will cover the overview of current conditions as well as limitations that official statistics and PODES face. Secondly, we review the plentiful literature of big data with its advantages and how it is used. Lastly, we will provide a general overview of how big data can help or augment the official statistics ability, whether in an increase in accuracy or better credibility.

2.1 Indonesia Official Statistics

The purpose of official statistics is to provide ‘an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an equal basis by official statistical agencies to honor citizens' entitlement to public information’ (United Nations Statistics Division, 2014: Principle 1). However, history proved that it was never the case.

During the colonial era of 19^{th} century Indonesia, the very first attempt to collect statistical data was conducted in Java mainly to acquire a taxable base in the mainly plantation-based economy of the Javanese. The Dutch-Indies colonial administration statistical enumeration failed to provide a comprehensive statistical overview of Java mainly because of the difference in the compilation method. Moreover, the census and extensive survey data were compiled by the local colonial administrator and not by a statistician, leaving the accuracy of the data to be dubious, if not questionable.

Today, the method of data collection has significantly improved. Since its expansion from 1971, the number of personnel working in BPS has swollen, allowing more survey areas to be covered, such as development and labor force. Also, BPS initiatives can publicize their findings from industry data to the Human Development Index (HDI).

Each Ministry and agency have their departments responsible for obtaining data for policy review and planning. The Ministry of Health, for example, thorough exhibit information regarding health topics in Indonesia in their annual publication Indonesia Health Profile. The report includes a wide array of health-related topics such as the number of community health centers in Indonesia, the number of hospitals, the percentage of accredited community health centers, to the ratio of hospital beds per 1,000 population.

The Central Bureau of Statistics has collected village-level data since 1980 in conjunction with the population census 1980. The data collection is done three times every ten years, alongside with Population Census, Agricultural Census, and Economic Census. However, since 2008 PODES is done independently from other census programs.

PODES has been used as the basis of researches amongst study discipline. Parmanto et al. (2008) used the village statistics data to conduct spatial and multidimensional analysis for community health assessment. Czaika and Kis-Katos (2009) used it to identify the determinants of displacement behavior of the forced migration during Aceh 1999–2002 Conflict. Gatto et al. (2017) used it to randomly select five sub-districts to evaluate the effects of oil palm contracts that involve smallholder farmers on rural economic development. PODES is also used to gain insight on community mental health; the study suggests that various factors such as cash transfer can reduce the suicide rate and supports a vital role for policy intervention (Christian et al., 2018).

In ensuring the accuracy and validity of the data, BPS implements two additional questionnaires; the Subdistrict Supplement Questionnaire (PODES08-Kec) and the District Supplement Questionnaire (PODES08-Kab/Kota). Furthermore, each enumerator that was tasked to collect data could go to each region up to three times to ensure completion of the surveys on top of necessary direct physical calculations when necessary. Besides, the enumerator and their supervisors would double-check the collected data with previous PODES and other official data if deemed necessary.

2.2 Big data

Although the word became the talk of the century, big data has a somewhat vague definition (Stephens-Davidowitz, 2017), nor is it universally accepted (Mayer-Schonberger & Cukier, 2013). A general definition by Tam and Clarke (2015) describes big data as potentially everything from traditional sources and more modern sources that became more accessible from the internet. The European Commission (2014) defined big data as ”*large amounts of data produced very quickly by a high number of diverse sources.*” While Doug Laney (2001) provided the term 3 ’Vs. ’ definition of big data, high-volume, high-velocity, and high-variety information assets that require cost-effective, innovative forms of information processing that facilitate better insights, decision making, and process automation.

As our dependence on technology became more and more apparent, leaving a monumental amount of digital footprint, potentially everything will be a source of data. From google searches, movies watched, songs listened to online transactions, big data are being generated by a multitude of sources at an astonishing rate, as one could not fathom how the digital age has ushered this birth of data generation. Approximately there are 3.5 billion Google searches, 8 billion Snapchat stories (Aslam, 2015), and 500 million tweets sent (Krikorian, 2013) every day. In 2015 over 227.1 billion global debit/credit card purchase transactions were made (The Nilson Report, 2019). Moreover, approximately 1.25 million trades were made on the New York Stock Exchange (2018). All of these data generate the ever-growing big data, which drives the trend of analyzing big data for a multitude of purposes.

As new sources of data arise, so does the types of data; ’*the days of structured, clean, simple, survey-based data are over. In this new age, the messy traces we leave as we go through life are becoming the primary sources of data*’ (Stephens-Davidowitz, 2017). The data we encounter today contains not only numbers but also other objects such as texts, sounds, images, or even movements. The adage ’Everything is data’ is becoming more relevant.

2.3 Ways to Incorporate Big Data into Official Statistics

There are certainly promising ways to improve official statistics. According to the Big Data Project Inventory (United Nations Statistics Division, 2018) compiled by the United Nations Global Working Group on Big Data, 34 national statistical systems from around the world have conducted 109 separate big data projects. National statistics agencies are attempting to use a wide array of sources to complement their datasets such as data scraped from websites, mobile phone data, social media, criminal records, satellite imagery, aerial imagery, health record, and public transport usage. These include: compiling mobility, transport and tourism statistics; indicators on crime and corruption; population density and migration, health and well-being measures; and labor market statistics. Figure 1 shows the sources of big data and project topics by national and international organizations (i.e., the World Bank and the United Nations Global Pulse).

The projects put light on the prospect that big data to be the new cost-effective or timely method of compiling statistics or easing the burden of national surveys. Big data also shows the potential to be able to generate more granular data or disaggregated statistics, paving ways for more detailed and comprehensive analyses.

Moreover, big data may hypothetically be better than survey data. The content of social media posts, likes, and dating profiles is no more nor less accurate than that of survey responses. As noted from Everybody Lies (Stephens-Davidowitz, 2017), ’*the trails we leave as we seek knowledge on the internet are tremendously revealing. In other words, people’s search for information is, in itself, information*’. This implies the very search from our google search history could be the ’digital truth’ that might not be captured by surveys as there are possibilities we do not response truthfully to the question. Hand (2015) also notes that big data are transaction data; they are closer to social reality than traditional surveys and census data that are heavily reliant on opinions, statements, and recall.

Finally, big data could also be used to enhance timeliness. In order to better serve the people, and more importantly, the working class, policymakers need not only long-term structural information but also up-to-date, real-time information – particularly when a nation is hit by a natural disaster or other catastrophic events. Official statistics have laid a good and sound framework for general analysis but fails to provide any meaningful insight when an emergency arises. This criticism is laid out by the Data Revolution Group (2014:22) ’*Data delayed is data denied. . . The data cycle must match the decision cycle.*’ The real-time nature of big data can be a breakthrough and critical importance for policymakers to react quickly to a variety of events, whether it is a pandemic, crisis, and other unforeseeable events.

3. Methodology

This paper aims to measure and compare the reliability of public facilities data from PODES by comparing them with ministerial data as an alternative source of official data. Google Maps data acquired through Google Places API, as the representation of big data. There are three types of public facilities, which are analyzed in this paper: Senior High School (SMA and MA), Public Health Centre (Puskesmas), and Hospital. For the alternative of official data sources, we will use official data from the Ministry of Health (Kemenkes) and the Ministry of Education (Kemendikbud). This paper will focus solely on DKI Jakarta data since it is the region with the complete data when compared to any other region. Furthermore, we assume that as the capital city, its data is the most reliable and valid compared to those of other cities. Thus, if we find any discrepancies or inaccuracies on DKI Jakarta data, then we may conclude that the inaccuracy occurs to all provincial data. All of those data are then utilized to employ descriptive analysis, hypothesis testing, and calculating error measurement.

3.1 Google Places API

This study employs Google Places API to find out the number of selected public facilities in Jakarta. Various queries of all the items are requested to the Google Maps server. The Google Maps server responds to the queries by sending a data frame of the location name, the type of location, and the decimal coordinates of the location. The data frame then plotted to a Geographical Information System (GIS) environment to automatically calculate the district level RFEI as the measure to the variable of interest of this study.

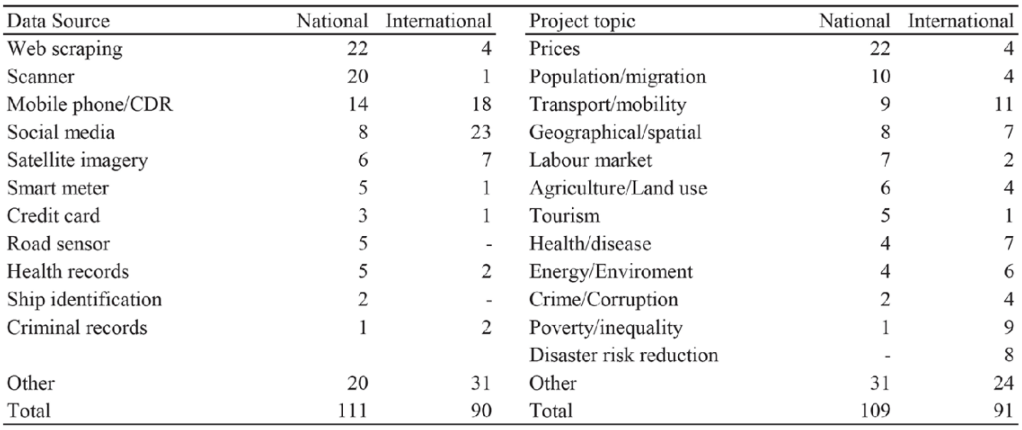
However, there is some potential criticism of this method of monitoring the presence of public facilities in this study. First, there is a chance that some portion of public facilities still not yet recorded on the Google Places server. This not supposed to be much of a problem in our study, which focused on Jakarta. However, if we try to replicate this study to a rather remote area or try to upscale this study to a national study, there is a fair chance that the number will be inaccurate. Second, the created time of each point of the place was not exposed by Google. If our study is going to add a time dimension into account, we cannot track how much of the selected place point at the given area in a specific time frame.

All of the criticism aside, this geolocation method is still the best method possible to obtain places data if no other sources are available. That is because a large-scale survey would be needed to accurately measure the actual number of places within every given region. In the end, if we compare the costs and the effort spent between those two methods, the geolocation method is still more efficient and, therefore, preferable.

3.2 Official Data

For this research, on top of PODES 2018, we are using three sources of official data that we sourced from each relevant government body in order to update PODES. We picked the sources of official data that have timestamp related to them so we can add or subtract the number from PODES which were gathered in 2017.

We gathered the official data on community health center (Puskesmas) from the epidemiology surveillance site run by the Social Health Subdivision of the Jakarta Health Agency. Each community health center is needed to register



Data Source	National	International	Project topic	National	International
Web scraping	22	4	Prices	22	4
Scanner	20	1	Population/migration	10	4
Mobile phone/CDR	14	18	Transport/mobility	9	11
Social media	8	23	Geographical/spatial	8	7
Satellite imagery	6	7	Labour market	7	2
Smart meter	5	1	Agriculture/Land use	6	4
Credit card	3	1	Tourism	5	1
Road sensor	5	-	Health/disease	4	7
Health records	5	2	Energy/Enviroment	4	6
Ship identification	2	-	Crime/Corruption	2	4
Criminal records	1	2	Poverty/inequality	1	9
			Disaster risk reduction	-	8
Other	20	31	Other	31	24
Total	111	90	Total	109	91

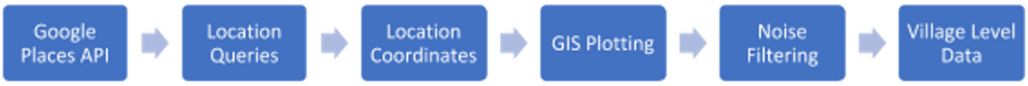


Figure 3.1. Steps of Google Places Data Processing

and give reports to the province's health agency, and the data presented on the site are gathered from there. Included in the site database is information on the address of every single community health center within Jakarta province.

In the case of hospitals, we sourced our data from RS ONLINE database service that is run by the directorate general of health services under The Ministry of Health. Similar to the case of the community health center, each hospital is required to register itself and give reports to the ministry through the "RS ONLINE" service.

Moreover, Indonesia's Ministry of Education and Culture has a site called "*Sekolah Kita*" where the public can check the profiles of schools all around Indonesia. The information shown through the sites is a part of the mandatory registration and reports that each school needs to do for the ministry.

Those various web pages were scraped to obtain the attribute of each facility. The leading information we aim to obtain are name, address, and launching year. The addresses of those facilities were then geocoded to find the exact coordinates of each facility. The coordinates are then plotted in order to identify which village-level region each of them belongs. Afterward, we separated the facilities that officially started their operation from 2018 onwards. The numbers of newer facilities are then added on to the number of each type of facility for each village-level region acquired through PODES 2018. Thus, we obtain the updated total number of each facility on each village-level region to compare with the google maps data.

3.3 Variables and Sources

The breakdown of the chosen variables is shown in Table 3.1, with an in-depth explanation of the sourcing process elaborated in the previous part of this chapter.

There are two main reasons why the variables are chosen. The first one is that those three variables have data available and accessible from other official data sources and big data sources. As mentioned before, a large number of variables in PODES are not readily available nor accessible from other sources, and we need variables that can be sourced through other official data sources and the big data sources to be able to do this study. Secondly, the chosen variables are of public facilities, which numbers tend to change from time to time barely. This should minimize the differences in numbers between sources due to time lag.

4. Results

4.1 Descriptive Analysis: Distribution of Margins

For the first stage of analysis, we calculate the difference of selected public facility numbers between each data source with official data. The margins were simply calculated by subtracting the count value of public facilities in each village within PODES or Google Places Data with the value within Official Data. Thus, the value can be considered to be accurate if the margin in the village equals to zero. The value of margin (deviation) larger than zero is a signal of overestimation.

In Figure 4.1, we can see the histogram of calculated margins between PODES and Official Data on three different public facility objects. The margins are rather evenly distributed and dominated by zero in all of the three cases of public facilities. Moreover, the further the value of margins differ from zero, the frequency of occurrence will be smaller in the histogram.

Even though the margins are dominated by zero, the count of accurate value varies across objects. We can see that Puskesmas and Rumah Sakit are having more than 150 occurrences of zero margins while SMA has under 125 occurrences of zero margins. The range of margins in SMA also happens to be the largest with the interval between -5 to 5. Furthermore, the nature of margins also differs between public facility objects. For Puskesmas and SMA, the frequency of overestimations is somewhat similar to underestimation. While for Rumah Sakit, the frequency of overestimation outnumbers the frequency of underestimation.

More interesting patterns emerged when the distributions were visualized as a map. Even though the patterns are not consistent, we can see that the villages near the border of the province have the tendency to have larger margins rather than villages near the center of the province. The mar-

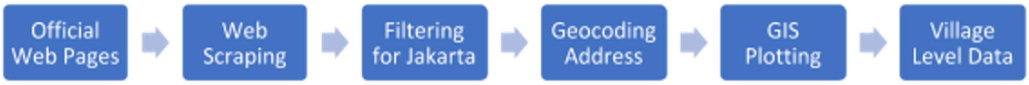


Figure 3.2. Steps of Official Data Processing

Table 3.1. Operational Definition from Various Data Sources

Source	Senior High School	Hospital	Puskesmas
PODES	Total of ”SMA/MA Negeri” and ”SMA/MA Swasta”	Total of ”Rumah Sakit” and ”Rumah Sakit Bersalin”	Total of ”Puskesmas dengan Rawat Inap”, ”Puskesmas tanpa Rawat Inap” and ”Puskesmas Pembantu”
Google Places	Processed results from search query ”Sekolah Menengah Atas”	Processed results from search query ”Rumah Sakit”	Processed results from search query ”Puskesmas”
Official Data	Count of SMA and MA within the Village	Count of Rumah Sakit within the Village	Count of Puskesmas within the Village

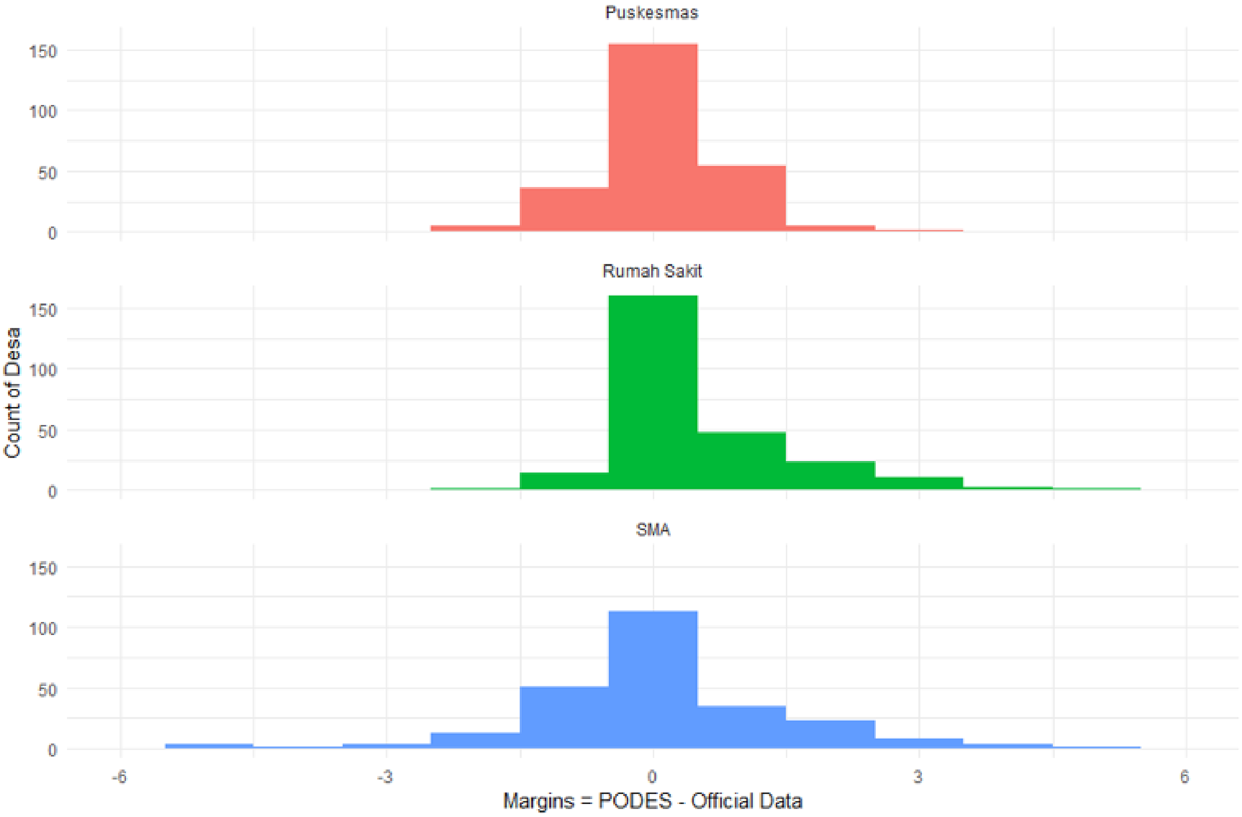


Figure 4.1. Distribution of Margin between PODES and Official Data

gins for Puskesmas are relatively smaller than Rumah Sakit and SMA, and it has occurred in fewer villages than the others. The margins of PODES and Official Data are more visible for Rumah Sakit and SMA, with a higher difference in villages further away from the city center and each district center. It may indicate that the more synchronized data have occurred in the city center, which has better access and smaller village area than the urban fringe.

We then move on to examine the accuracy of Google Places Data. From Figure 4.3, we can see that Google Places data are having a similar occurrence of zero margins for Puskesmas and SMA. On the other hand, the number of zero margins for Rumah Sakit is less than PODES with less than 100 occurrences. While the occurrence of zero margins still dominates the histogram. The tendency of overestimation and underestimation differs across public facility objects. Puskesmas and Rumah Sakit tends to be overestimated where SMA tends to be underestimated. Moreover, the range of margins in Google Places data is also higher than PODES with the interval of (-6, 8).

However, when plotted in maps, we can see the margins are rather randomly distributed across villages, as we can see in Figure 4.4. Although the difference is less vivid for Puskesmas and Rumah Sakit, the difference for SMA is highly visible. There are several villages in urban fringe that has a higher number of SMAs in the official data than in the Google Places data. Assuming the government has better SMAs data, this may indicate that several villages in the urban fringe have a less reliable mobile signal or less accurate in locating high schools. Thus, the fewer number of SMAs recorded in Google Places. Another possibility to explain this phenomenon is that the official data has lagged in updates compared to the Google Places data, which more regularly updated. Hence, we have overestimated official data for SMAs compared to the Google Places data.

To examine the difference in nature between PODES and Google Places data, we also plot the comparison between both sources of data. From figure 4.5 and 4.6, we

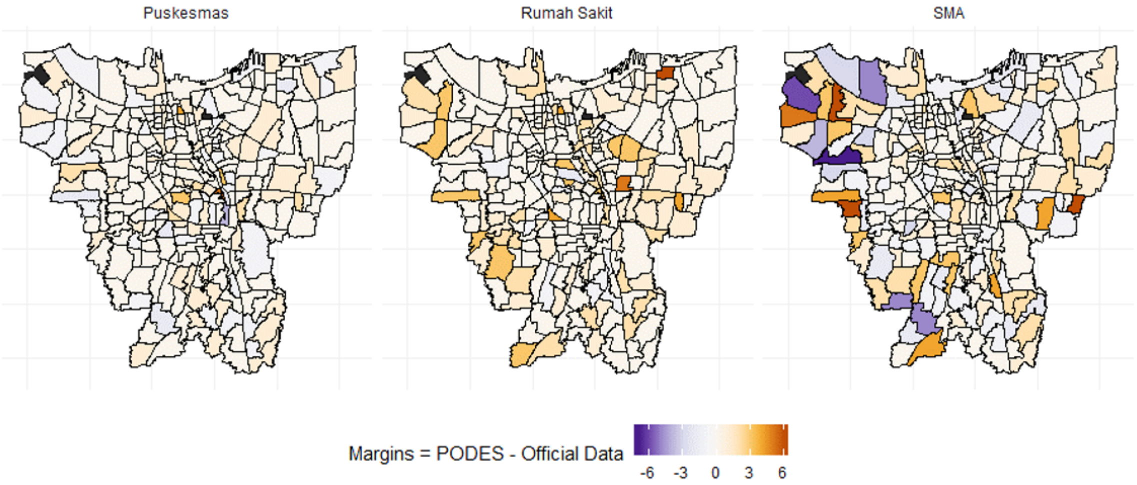


Figure 4.2. Map of Margin between PODES and Official Data

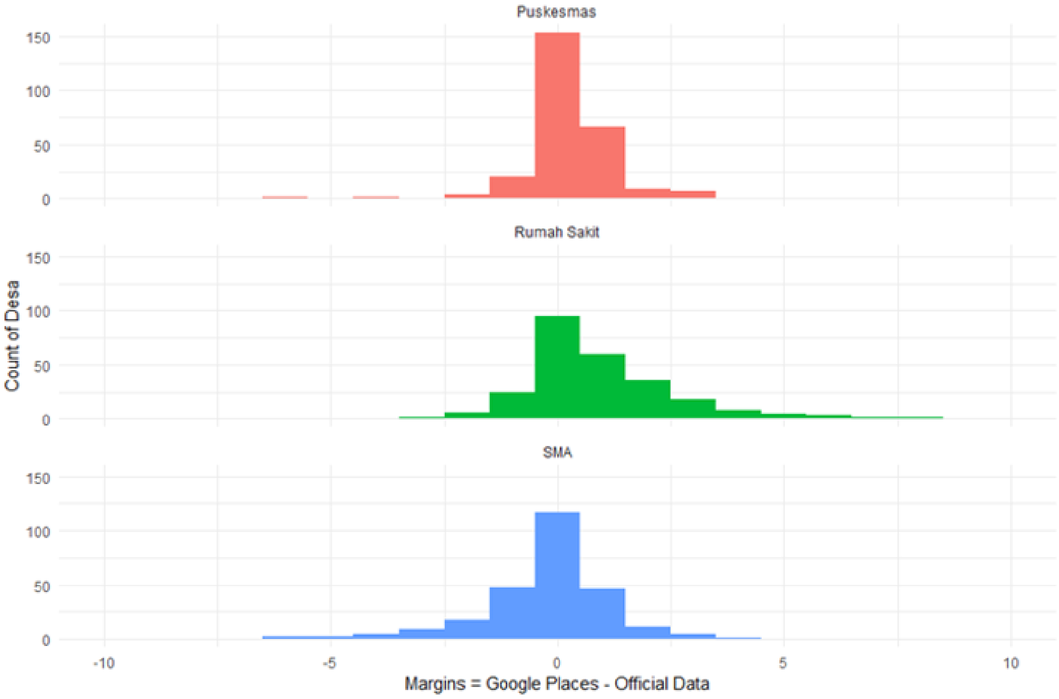


Figure 4.3. Distribution of Margin between Google Places and Official Data

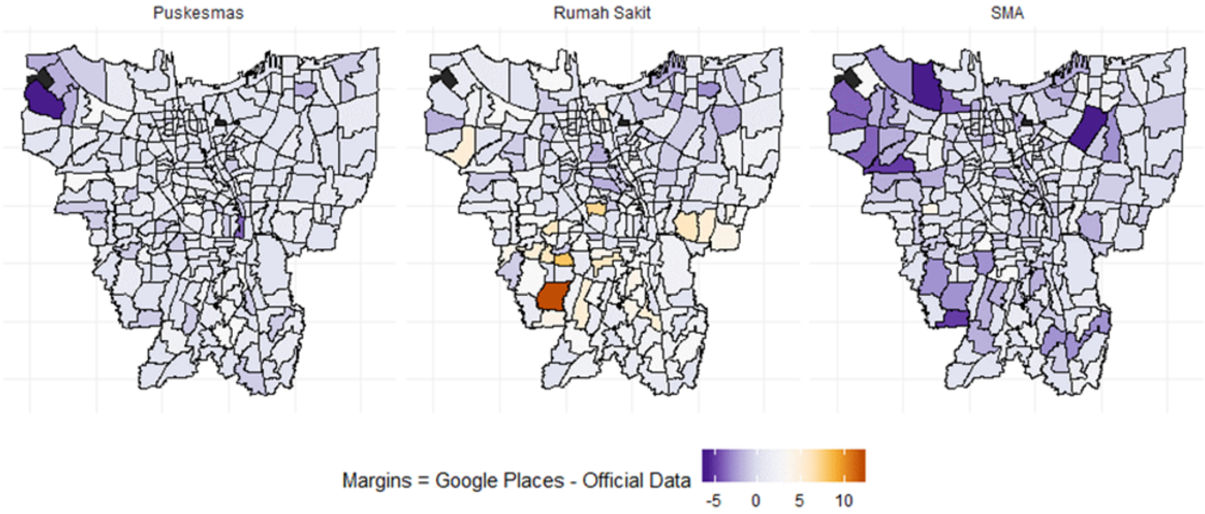


Figure 4.4. Map of Margin between Google Places and Official Data

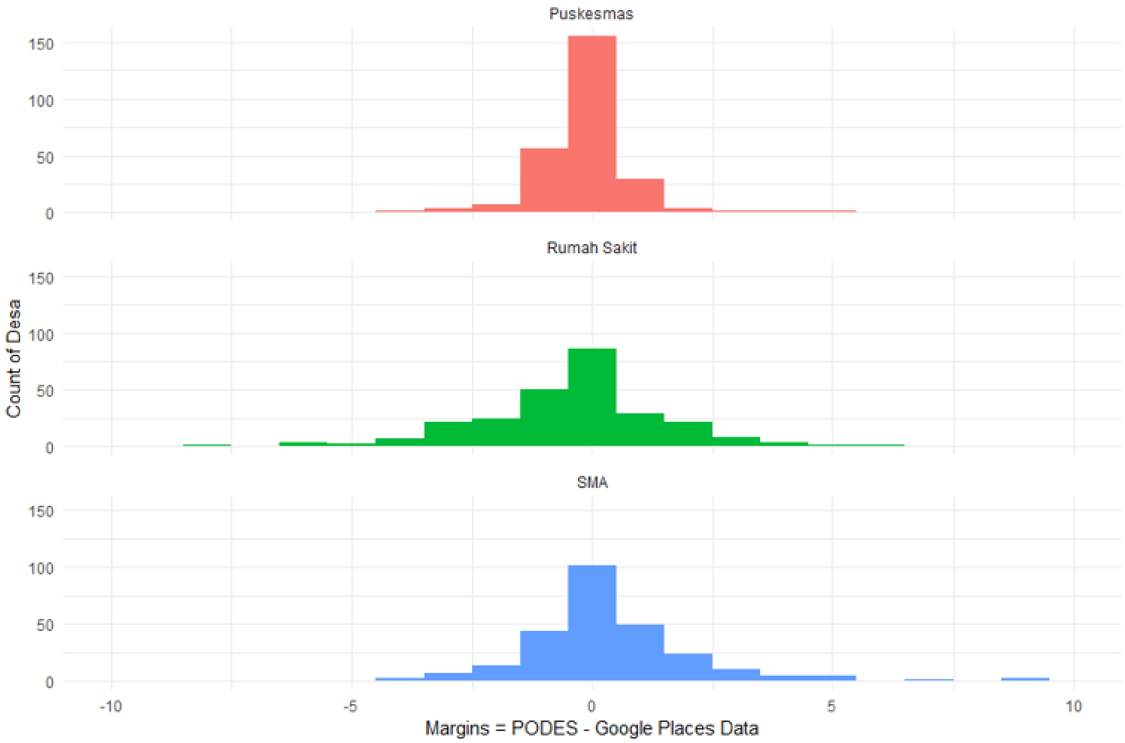


Figure 4.5. Distribution of Margin between Google Places and Official Data

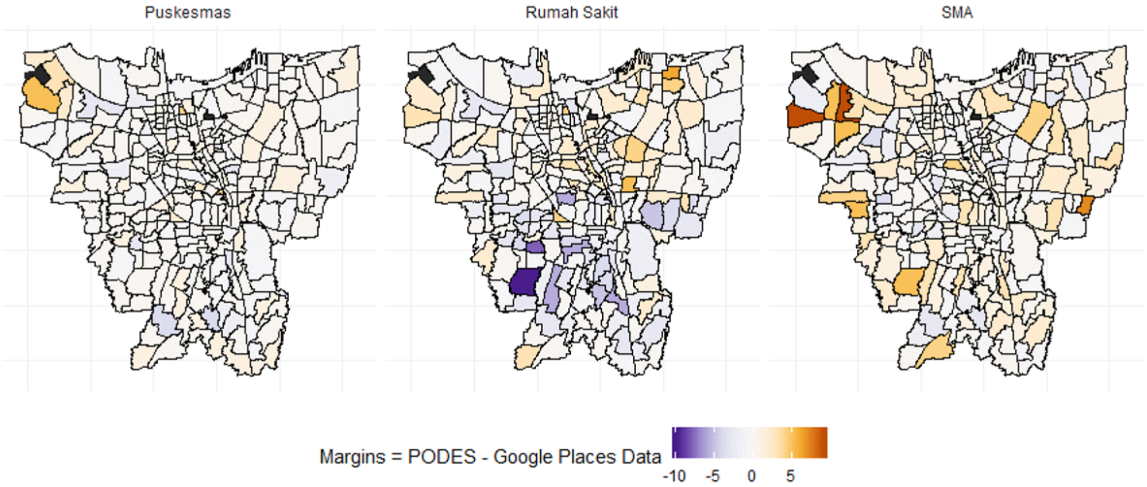


Figure 4.6. Map of Margin between Google Places and Official Data

can see that not only PODES and Google Places differs inaccuracy, the stated number of each public facility object are also differed both in number and by locations. Moreover, with the interval of (-9, 9), the range of margins between PODES and Google Places is even wider than in the case of PODES and Google Places versus Official Data. This phenomenon might be related to the crowd-sourced nature of Google Places data where the location points recorded by users' mobile phone rather than an interview-based survey employed for PODES.

4.2 Hypothesis Testing: T-Test for Mean Comparison in Two Data Sources

To extend the descriptive analysis above, we then compute the t-test to test whether the difference of mean value between two different data sources in all three public facility objects are statistically significant or not. This analysis will show how bad the inaccuracy exists in PODES and Google Places Data with Official Data as reference.

From Table 4.1, we can see that on average, Google Places Data underestimates the number of SMA while PODES overestimates the number of SMA. Meanwhile, for two other public facility objects, both PODES and Google Places data overestimates the number of Rumah Sakit and Puskesmas. Moreover, in all three cases, the mean absolute difference between Google Places and Official Data outnumbers the mean absolute difference between PODES and Official Data.

The overall result of the t-test is consistent with the descriptive analysis: Google Places Data is less accurate than PODES. In addition to bigger mean absolute difference, the mean number of two public facility objects on Google Places Data were found to differ with the mean on Official Data significantly. The difference is significant at $\alpha < 0.01$ for Rumah Sakit and $\alpha < 0.05$ for Puskesmas. Meanwhile, on PODES, only one public facility object's mean number

Table 4.1. Mean Comparison between PODES, Google Places and Official Data

	PODES		Google Places		Official Data		PODES – Official Data		Google Places – Official Data	
	mean	sd	mean	sd	mean	sd	b	t	b	t
SMA	2.38	2.3	2.05	1.8	2.29	2.2	0.09	(-0.45)	-0.25	(-1.40)
Rumah Sakit	1.17	1.4	1.63	2	0.66	0.9	0.51***	(-4.94)	0.97***	-7.08
Puskesmas	1.36	0.8	1.5	1	1.24	1	0.11	(-1.51)	0.26**	-2.96
Observations	261		261		261		522		522	

Table 4.2. Mean Comparison between PODES and Google Places Data

	PODES		Google Places		T-Test	
	mean	sd	mean	sd	b	t
SMA	2.38	2.26	2.05	1.78	0.33	(-1.87)
Rumah Sakit	1.17	1.38	1.63	2.01	-0.46**	-3.07
Puskesmas	1.36	0.75	1.5	1	-0.14	-1.83
Observations	261		261		522	

found to differ with the mean on Official Data significantly.

Meanwhile, the difference between PODES and Google Places Data, which displayed on the descriptive analysis, turned out to have a different result based on the t-test. If measured by mean rather than plotting the individual distribution, only one public facility object's mean number found to differ between PODES and Google Places significantly. The difference is also only significant at $\alpha < 0.05$.

4.3 Error Measurement: Root Mean Square of Error

So far, we have examined the distribution of margins with distribution plots such as histogram and choropleth. We also have examined how bad the inaccuracy of PODES and Google Places by conducting a t-test. To complete the analysis, we measure the error with the assumption of PODES and Google Places Data as predicted value and Official Data as the actual value. We then compute Root Mean Square of Error as Error Measurement. It turned out that while in overall PODES is more accurate than Google Places Data, the accuracy differs across cases.

Table 4.3. Root Mean Square Error for Official Data

Object	PODES	Google Places
SMA	165.050	142.232
Rumah Sakit	118.257	203.231
Puskesmas	0.88841	0.98456

From the RMSE measurement above, we can see that PODES outperformed Google Places in terms of accuracy in the case of Rumah Sakit and Puskesmas. Meanwhile, Google Places outperformed PODES in the case of SMA. This result might indicate that the accuracy of PODES and Google Places are differing across cases. If we replicate expand this analysis on another object, we can reasonably expect different performance results on each case.

5. Conclusion

As shown by the result above, despite the zero margin majority in most villages for all three variables when the three sources are compared, there are still quite a large number of villages with variances. If all three sources are accurate, there should be zero margins across the board. In the case of comparison with Google data, we do expect a slight discrepancy since the data is mined in late 2019, while the other two sources gathered their data in 2017. However, the expected discrepancy is low, especially since the variables are time-insensitive. Public facilities such as schools, health centers, and hospitals are not built in a short time-frame. The closing down and new opening of such facilities are still something that could still happen within the time differences, but for a village-level region to have a difference of more than three facilities in a couple of years is pretty much unprecedented. Furthermore, since there is no comparison between two sources that yield completely zero margins, this means that at least two sources are inaccurate with a possibility that all three sources are inaccurate.

As a solution, soon, we are hoping to progress this study by conducting a field survey where we would go to some of the regions throughout DKI Jakarta and physically count the number of existing and operational public facilities. By doing so, we would have the actual number for the variables which we could use as base point comparison to find out which ones that are or at least close to being accurate and which ones are inaccurate.

Although we cannot confirm whether PODES is accurate in this study, but it is not befitting of the country's official statistic institution if it were to publish such a vital data inaccurately. We do acknowledge that gathering entirely accurate regional data throughout Indonesia is not only costly, but it is also an arduous task to shoulder. In the short run, we do suggest for BPS to do more quality control for PODES, at least by making sure that it is in-line with other available official data. BPS could also remodel their methods to not focus too much on one source for each village. In other words, BPS need to reinforce their processes in producing PODES, especially with how important it is. In the long run, though, we suggest BPS to at least consider augmenting their process by using big data.

Lastly, we hope that this study could help highlight the weaknesses of the currently available official data in Indonesia and trigger more effort to improve that highly crucial information through making the most of big data if and when it is possible.

References

Aslam, S. (2015, 7 October). Snapchat by the numbers: Stats, demographics & fun facts. *Omnicore*. Retrieved from: https://www.omnicoreagency.com/snapchat-statistics/ (January 10th, 2020).

Christian, C., Hensel, L., & Roth, C. (2018). Income shocks and suicides: causal evidence from Indonesia. https://www.briq-institute.org/wc/files/people/chris-roth/working-papers/income-shocks-and-suicides.pdf.

Czaika, M., & Kis-Katos, K. (2009). Civil conflict and displacement: Village-level determinants of forced migration in Aceh. *Journal of Peace Research, 46*(3), 399-418. doi: https://doi.org/10.1177/2F0022343309102659.

Cornelia, L. H., Diane, C. K., & Gabriel, Q. (2017). Big Data: Potential, Challenges and Statistical Implications. IMF Staff Discussion Notes.

Data Revolution Group. (2014). *A world that counts: mobilising the data revolution for sustainable development*. UN Data Revolution - The UN Secretary General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development. Retrieved from: https://www.undatarevolution.org/report/ (January 10th , 2020)

Einav, L., & Levin, J. (2014). The Data Revolution and Economic Analysis. In J. Lerner & S. Stern (eds), *Innovation Policy and the Economy, Volume 14* (pp. 1-24), University of Chicago Press - National Bureau of Economic Research. https://www.nber.org/chapters/c12942.

European Commission. (2014). *Big data*. Retrieved from: https://ec.europa.eu/digital-single-market/en/big-data (January 10th, 2020).

Gatto, M., Wollni, M., Asnawi, R., & Qaim, M. (2017). Oil palm boom, contract farming, and rural economic development: Village-level evidence from Indonesia. *World Development, 95*, 127-140. doi: https://doi.org/10.1016/j.worlddev.2017.02.013.

Hand, D. J. (2015, March). 'Official Statistics in the New Data Ecosystem. *PowerPoint Presentation*. Retrieved from: https://ec.europa.eu/eurostat/cros/system/files/Presentation%20S20AP2%20%20Hand%20-%20Slides%20NTTS%202015.pdf (September 10th, 2019).

Krikorian, R. (2013, August 16). New Tweets per second record, and how!. *Twitter blogs*. Retrieved from: https://blog.twitter.com/2013/new-tweets-per-second-record-and-how (September 10th, 2019).

Laney, D. (2001, 6 February). 3D data management: Controlling data volume, velocity and variety. *META Group: Application Delivery Strategies, file 949*. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Mayer-Schönberger, V., & Cukier, K. (2014). Big Data: A Revolution That Will Transform How We Live, Work, and Think. American Journal of Epidemiology, 1143-1144.

New York Stock Exchange. (2018). *Transactions, statistics, and data Library*. Retrieved from: https://www.nyse.com/data/transactions-statistics-data-library (January 10th, 2020).

Parmanto, B., Paramita, M. V., Sugiantara, W., Pramana, G., Scotch, M., & Burke, D. S. (2008). Spatial and multidimensional visualization of Indonesia's village health statistics. *International Journal of Health Geographics, 7*(1), 30. doi: https://doi.org/10.1186/1476-072X-7-30.

Stephens-Davidowitz, S. (2017). *Everybody lies: big data, new data, and what the Internet can tell us about who we really are*. New York: Dey St./William Morrow.

Tam, S.-M., & Clarke, F. (n.d.). Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics. Methodology and Data Management Division, Australian Bureau of Statistics.

The Nilson Report. (2019, February 20). *General Purpose Cards—U.S. 2018*. Retrieved from: https://nilsonreport.com/mention/321/1link/ (January 10th, 2020).

United Nations Statistics Division. (2014). *Principle 1 - Relevance, impartiality and equal access*. Retrieved from: https://unstats.un.org/unsd/goodprac/bpaboutpr.asp?RecId=1 (January 10th, 2020).

United Nations Statistics Division. (2018). *Big data project inventory*. Retrieved from: https://unstats.un.org/bigdata/inventory/ (January 10th, 2020).

Gedung LPEM FEB UI
Jl. Salemba Raya No. 4, Jakarta 10430
Phone : +62-21 3143177 ext. 621/623;
Fax : +62-21 3907235/31934310
Web : http://www.lpem.org/category/publikasi/workingppers/

9 772356 400001