Covid-19 Time Series Data Quality Analysis (Case study at COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University)

Yuyun Hidayat, Dhika Surya Pangestu, Titi Purwandari
Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Indonesia.
yuyun.hidayat@unpad.ac.id; dhikasurya.surya7@gmail.com; titipurwandari@yahoo.com

Sukono
Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Indonesia.
sukono@unpad.ac.id

Abdul Talib Bon
Department of Production and Operations, University Tun Hussein Onn Malaysia, Malaysia
talibon@gmail.com

Abstract

Covid-19 is a contagious disease caused by SARS-CoV-2, which is a type of coronavirus. Since its appearance at the end of 2019, until 2 August 2020, there have been more than 17.7 million infected people worldwide. During that time, various studies appeared to study the Covid-19 pandemic. One of them was research on the development of the number of Covid-19 cases. One of the many datasets used to study the development of Covid-19 cases is data from the "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University". The purpose of this study is to analyze the quality of the data and detect errors that occur in the data so that researchers who will use it know the quality of the data before using them. This study uses a statistical quality control approach. The methods used in this study are acceptance sampling and control charts. 30% of the data will be issued using a control chart and investigated for any possible errors. After that, the data is corrected according to the error that occurred. This process is repeated until there are no more errors in the data. After several iterations, we found errors in the Covid-19 data in this study. The errors found are data input errors, decreasing value, confirm data is less than recover, confirm data is less than death, zero confirm on the first date, not zero recovers on the first date, and not zero death on the first date. It is recommended for those who will use the data from this source to check and correct them first before using them.

Keywords:
Covid-19, Data Quality, Statistical Quality Control

1. Introduction

Covid-19 is a contagious disease caused by SARS-CoV-2, which is a type of coronavirus. Since its appearance on December 17 2019 in the city of Wuhan, Hubei, China, until August 2, 2020, there have been more than 17.7 million people in the world who have been infected, with a death toll of more than 682,000. This disease is spread in 188 countries in the world, making it a world pandemic (Wikipedia, 2020) . Indonesia is one of the countries currently being hit by the Covid-19 pandemic. Covid-19 was confirmed to have appeared for the first time in Indonesia on March 2, 2020. At that time there were two people who were infected with Covid-19 due to contact with Japanese citizens. This is known after a Japanese citizen was declared infected with the coronavirus after leaving Indonesia and arriving in Malaysia (J.Akbar, 2020).

Since the beginning of the emergence of the development of the number of COVID-19 cases in Indonesia, it has continued to increase, until August 1, 2020, there have been 109,936 people infected by COVID-19

(Indonesia's Task Force of COVID-19 Rapid Response, 2020). Based on the Worldometer, Indonesia is ranked 16th in the World and 4th in Asia for active cases of Covid-19 (Worldometers, 2020). This is of course very worrying. Various studies have appeared to study the Covid-19 pandemic and one of them is research on the development of the number of Covid-19 cases. One of the many datasets used to study the development of the number of Covid-19 cases is data from the "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University" (Dong E, 2020). The data from John Hopkins University is claimed to have high accuracy and good quality. However, as researchers, we need to examine the quality of existing data first, before using them. This is done to avoid the GIGO (Garbage in Garbage out) phenomenon. This phenomenon needs to be watched out for in a study, because if the data used in the study are wrong, then the conclusions of the study will also be wrong (John, 2009).

2. Literature review

Data Quality

The *New Oxford American Dictionary* defines data first as "facts and statistics collected together for reference or analysis." The American Society for Quality (ASQ) defines data as "A set of collected facts. And the International Standards Organization (ISO) defines data as "re-interpretable representation of information in a formalized manner suitable for communication, interpretation, or processing" (International Standards Organization, 2004). Data enable us to understand facets of reality by abstracting (separating) and representing them in summary form (an abstract) (Peirce, 1955). Data are always representations. Their function is primarily semiotic. They stand for things other than themselves (Chisholm, 2010). These things can be objects (people, places, things) or events or concepts.

Data quality talks about whether a data meets the expectations of the data users and how well it represents the objects, events, and concepts it is created to represent. To measure whether data meets expectations or is "fit for use," expectations and uses need to be defined. Therefore, the discussion of data quality depends on the interests of data users (Sebastian-Coleman, 2013). Data quality is defined as follows: data has quality if it satisfies the requirements of its intended use. It lacks the quality to the extent that it does not satisfy the requirement. In other words, data quality depends as much on the intended use as it does on the data itself (Olson, 2003). So the quality of data is directly related to the reasons why the data was created. Data with high quality better meets the expectations of its users than data with low quality.

Acceptance Sampling

Acceptance Sampling involves collecting and analyzing a relatively small number of measurements to make "accept or reject" decisions about a relatively large number of units (Allen, 2006) . Statistical evidence is generated about the fraction of the units in the lot that are acceptable. Acceptance sampling plans have been developed for application to a group or class of attributes and variables that are of about equal criticality to product (or process, or service) acceptability (Baillie, 1990). The method of acceptance sampling provides an economical way to evaluate the acceptability of characteristics that might otherwise go uninspected. There are three aspects to consider regarding acceptance sampling (Montgomery, 2009) :

1) The purpose of acceptance sampling is to accept or reject a product, not to estimate the quality of the product.
2) The acceptance sampling plan does not provide a direct form of Quality Control. Acceptance sampling is only accepting, or rejecting, of a product. Even if all products are of the same quality, this sampling will accept and reject some.
3) The best use of acceptance sampling is not to "check the quality of a product" but as an audit tool to ascertain whether the process conforms to the specified criteria.

Since not all units are inspected in acceptance sampling, acceptance sampling unavoidably involves risks. The method of "complete inspection" involves using one measurement to evaluate all units relevant to a given situation. A complete inspection might naturally be expected to be associated with reduced or zero risks.

Control Chart

Control Chart is an on-line monitoring technique used to detect the emergence of assignable causes from changes in a process so that it can be handled immediately. The purpose of a control chart is to detect a change in the performance of a process (Oakland, 2003). Apart from monitoring, control charts can also be used to estimate the parameters of a process and based on this information to find out what the capabilities of a process are. Control charts are excellent tools to reduce variability in a process (Montgomery, 2009).

The control chart consists of a centerline (CL) which represents the average of the quality characteristics being monitored. Two horizontal lines, namely, upper control limit (UCL) and lower control limit (LCL). This control limit is determined to ensure that if a process is in control, then the sample is plotted between the control

limits. In general, the model of the control chart is, for example, w is the statistic of the sample that shows the quality characteristics to be measured, where μ_w is the average of w and the standard deviation of w is σ_w. Then the centerline (CL), upper control limit (UCL), and lower control limit (LCL) are defined as

$$UCL = \mu_w + L\sigma_w \quad (1)$$

$$Center\ line = \mu_w \quad (2)$$

$$LCL = \mu_w - L\sigma_w \quad (3)$$

L is the distance between the control limit and the centerline, which is expressed in terms of standard deviation (Montgomery, 2009). As long as the data is plotted within the control limit, the process is assumed to be in control, and no action needs to be taken. However, if there is data that is plotted outside of control limits, then it is interpreted as evidence that the process is out of control. If that happens, investigations and corrective actions are needed to find and eliminate the cause of the incident.

Shewhart Individual Control Chart

Shewhart individual control chart is a simple but very useful tool in statistical quality control. The usefulness of Shewhart's control charts has been proven in various applications (Hart M.K. & Hart R.F., 1992). Shewhart individual control chart is used to monitor the process with sample units, n = 1. In its application, the individual control chart uses a moving range to predict the variability of a process (Montgomery, 2009). The moving range on the Shewhart individual control chart is defined as

$$MR_i = |x_i - x_{i-1}| \quad (4)$$

In general, the parameters of the Shewhart individual control chart are, for example, x is the statistic of the quality characteristics to be measured, where $\bar{x}$ is the average of x, and the average of MR is $\overline{MR}$. Then the centerline (CL), upper control limit (UCL), and lower control limit (LCL) are defined as

$$UCL = \bar{x} + 3\frac{\overline{MR}}{d_2} \quad (5)$$

$$Center\ line = \bar{x} \quad (6)$$

$$LCL = \bar{x} - 3\frac{\overline{MR}}{d_2} \quad (7)$$

Where d_2 is the factor for the centerline used in the variable control chart.

Out of Control Action Plan

Out of control action plan is a flow chart or description of a series of activities that must be carried out when there is out-of-control data on the control chart. The OCAP consists of *checkpoints*, which are potential assignable causes, and *terminators*, which are actions taken to resolve the out-of-control condition, preferably by eliminating the assignable cause. It is very important that the OCAP specify as complete a set as possible of checkpoints and terminators, and that these be arranged in an order that facilitates process diagnostic activities. Often, analysis of prior failure modes of the process and/or product can help design this aspect of the OCAP. Furthermore, an OCAP is a *living document* in the sense that it will be modified over time as more knowledge and understanding of the process is gained. Consequently, when a control chart is introduced, an initial OCAP should accompany it. Control charts without an OCAP are not likely to be useful as a process improvement tool (Montgomery, 2009).

3. Research Methodology

This research uses COVID-19 data from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The data is converted into COVID daily data or (CDD). CDD data consists of the number of confirmed, recovered, and deaths from people with Covid-19. Data is compiled in daily format from 22 January 2020 to 1 August 2020. With an acceptance sampling of 30% of the data. The data is output using the Shewhart Control chart method. So that 30% out of control data will be obtained. Then the Out of Control (OOC) data will be checked according to the out of control action plan (OCAP, Figure 1) that has been made.

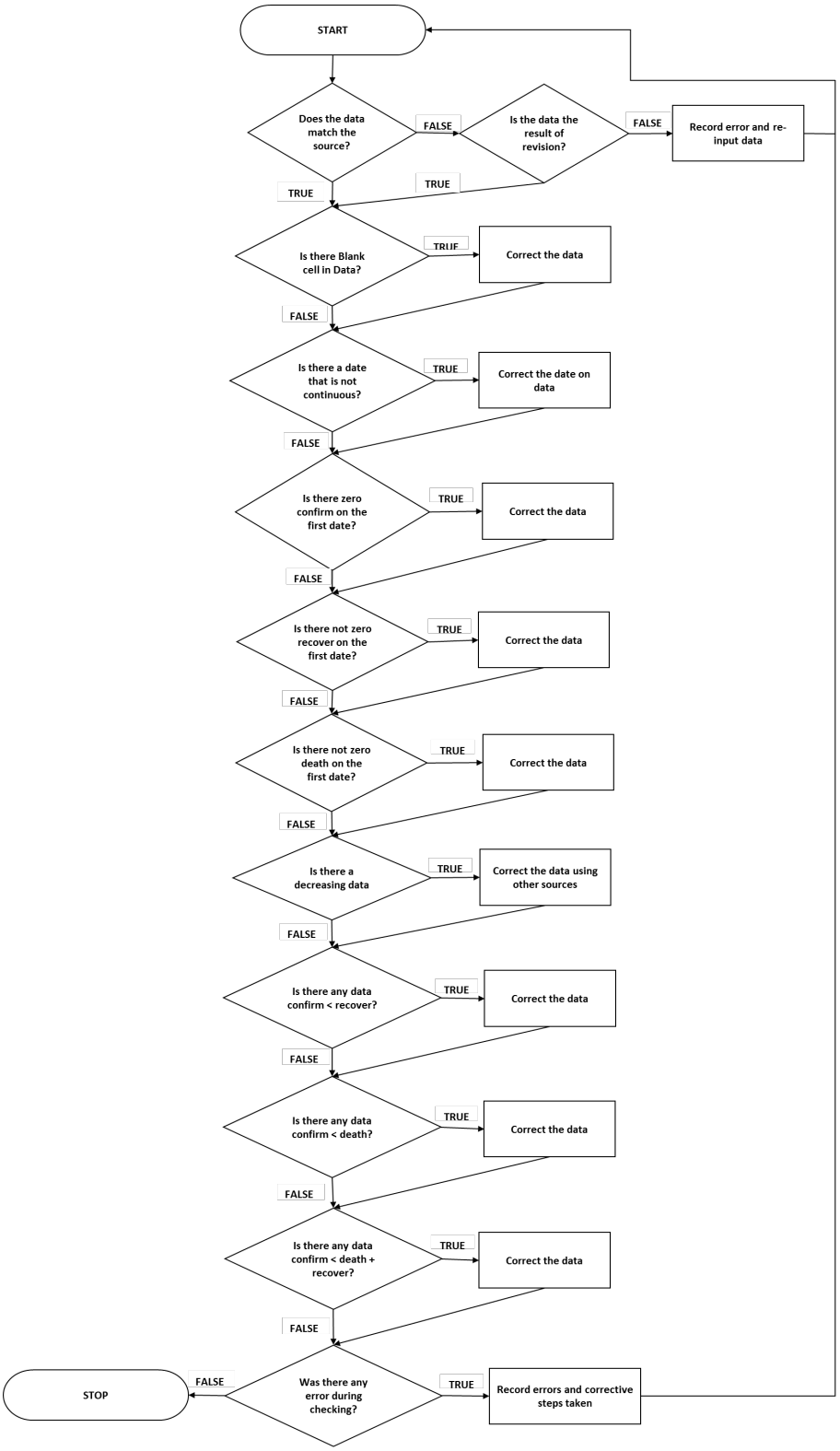

Figure 1. Out of control action plan in this research

4. Results and Discussion

The data used in this study is the COVID-19 time-series data from January 22 to August 1, 2020. The research process was carried out using the method previously discussed. Out of control data is issued using it to check the presence of errors that may occur. The following is the number of out of control data that was checked in this study.

Table 1. Amount and percentage of out of control data

	Iteration			
	1	**2**	**3**	**4**
Out of control data	10938	8724	8320	8353
Total data	36091	28237	27730	27729
% OOC data	0.303067	0.308956	0.300036	0.301237

Based on the Table 1, it can be seen that in this study there were 4 iterations carried out. The amount of data checked from the first iteration to the fourth iteration is decreasing. This is due to the data corrections performed at each check iteration. Table 1 does not show the number of errors that occurred in the data. Table 1 shows the amount of data that is out of control and needs to be checked for any errors that may exist in the data. To find out the existence of an error, out of control data will be checked and corrected according to the out of control action plan that has been made. The following is the result of checking the error in the out of control data.

Table 2. Data errors found in out of control data

Error	Iteration			
	1	2	3	4
1. Data input error	2540	0	75	78
2. Blank cell	0	0	0	0
3. Not continuous date	0	0	0	0
4. Zero confirm on first date	0	26	0	0
5. Not zero recover on first date	0	15	0	0
6. Not zero death at first date	0	16	3	0
7. Decreasing data	0	33	2	0
8. Confirm < recover	0	0	0	0
9. Confirm < death	0	0	0	0
10. Confirm < death + recover	0	0	0	0
Total	2540	90	80	78

- In the first iteration, there were 2540 errors in the Data that did not match the input, because the first iteration stopped at the initial check. Repair data by re-inputting data.
- The second iteration, using the results of the improvements in the first iteration. In the second iteration, there were no errors in data input. So that the existing data is following the sources used. But in the second iteration, even though the data is following the source, an error is still found. The errors found were zero confirmation on the first date, non-zero recovered on the first, non-zero deaths on the first date, and data decline. In the second iteration, data correction is done with unneeded data, and replacing the wrong data with another source, or if another source has the same error the data is made the same as the previous data.
- The third iteration is done using the data corrected in the second iteration. In the third iteration, 75 data were found that were not following the JHU CSSE COVID-19 data, but because 75 of these data were corrected data, according to the out of control action plan, the expression was checked. There are fewer errors found in the third iteration than in the previous iteration. In this iteration, it was found that there were errors, not zero deaths on the first date and data decline. Corrections made in this iteration are unnecessary data and replace the wrong data with another source, or if another source has the same error the data is made the same as the previous data.
- The fourth iteration is done using the correction results in the third iteration. 78 data were not following the JHU CSSE COVID-19 data, but because the 78 data were corrected data, according to the action plan Out of control, the expression was checked. In the next check, there were no more errors in the data out of control. So checking the error data stops at the fourth iteration.

In addition to the number and types of errors that occur in out of control data. We also found that there was a change in the distribution of out of control data for each iteration that was carried out. The following is a graph of the frequency distribution of the out of control data from the first to fourth iterations.

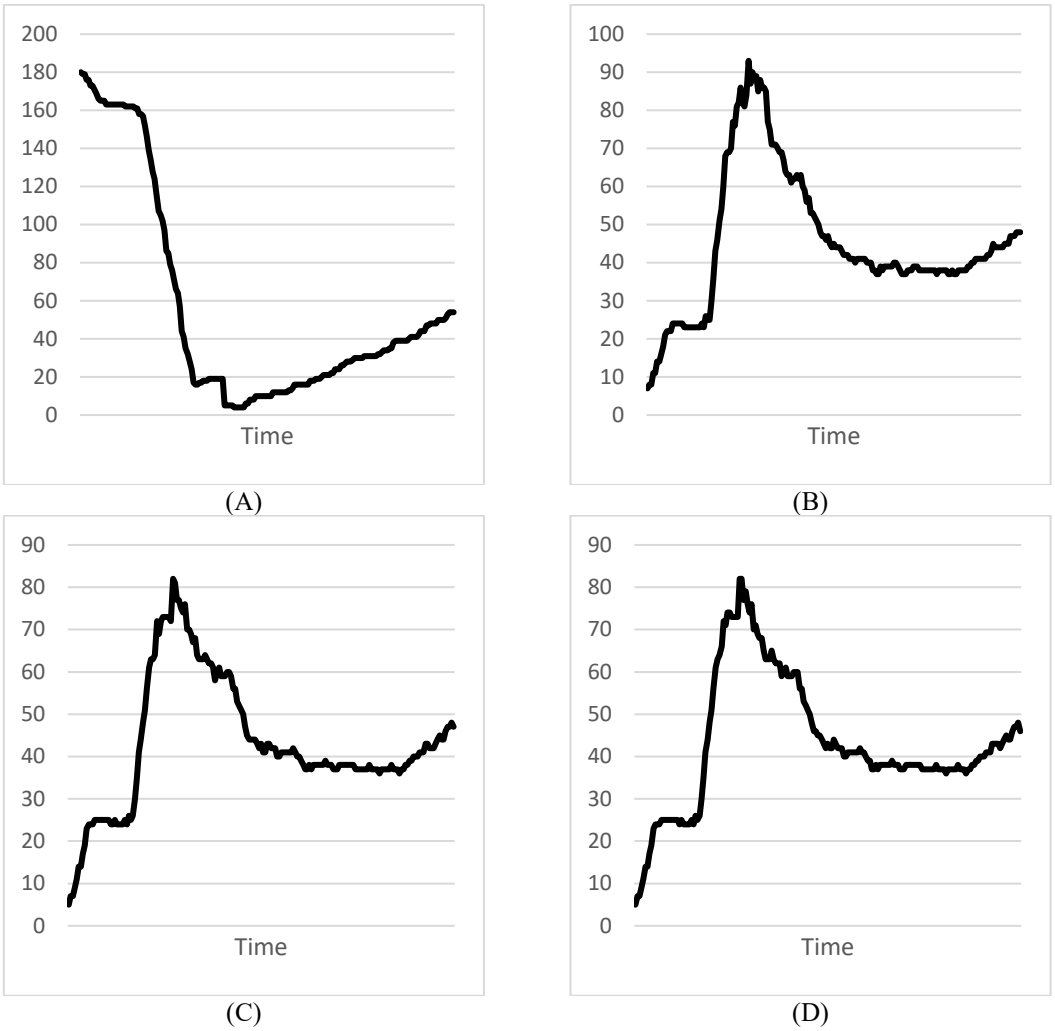

Figure 2. Out of control data frequency distribution, with the y-axis out of control frequency and the x-axis time (A) Iteration 1 (B) Iteration 2 (C) Iteration 3 (D) Iteration 4

Based on Figure 2, the distribution from the first to fourth iterations changes. The first iteration with the highest number of errors has the most different out of control data frequency distribution compared to other iterations. As for the other iterations, it has an identical out of control distribution, with slight differences between the iterations.

5. Conlusions

With the COVID-19 pandemic still going on, and showing no signs of ending, various studies are emerging to study it. At the same time, a wide variety of existing datasets are being used to study the Covid-19 pandemic. As previously stated, this research aims to ensure the quality of the COVID-19 data originating from the "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University". Of the four iterations carried out, the first iteration found the most errors, and the cause was input errors. However, after entering the second iteration, when the data matched the source, errors were still found. The important thing to remember is that the errors found in this study are errors found in the out of control data. Even though until the fourth iteration, there are no more errors in the data, it does not mean that the data is 100% error-free. The method for extracting 30% of the data using a control chart is not a sampling method. This method is used to detect errors that may occur in the data. So, there is still the possibility of errors occurring in data that were not checked in this study. This needs to be known and be aware of by those who will use the data. In this study, we found errors in the data, and there is still the possibility of errors in the out of control data we checked. If allowed and used immediately, there is a concern that it will affect the results of the research and

cause garbage in garbage out (GIGO). To avoid this, we recommend that those who use this data check the data thoroughly before using it.

Reference

Allen, T. T. (2006). *Introduction to Engineering Statistics and Six Sigma.* London: Springer.

Baillie, D. (1990). Attriables Acceptance Sampling Plans. *Frontiers in Statistical Quality Control*, 3-33.

Chisholm, M. D. (2010). *Definitions in information management: A guide to the fundamental semantic metadata.* Canada: Design Media.

Dong E, D. H. (2020). An interactive web-based dashboard to track COVID-19 in real time. *20*(5), 533-534. doi:10.1016/S1473-3099(20)30120-1

Hart M.K., & Hart R.F. (1992). Shewhart Control Charts for Individuals with Time Ordered Data. *Frontiers in Statistical Quality Control, 4*. doi:10.1007/978-3-662-11789-7 9

Indonesia's Task Force of COVID-19 Rapid Response. (2020). Retrieved August 3, 2020, from https://covid19.go.id/peta-sebaran

International Standards Organization. (2004). Information technology–metadata registries (MDR) Part 4 formulation of data defnitions. ISO/IEC.

J.Akbar. (2020). *Kompas*. Retrieved August 3, 2020, from https://surabaya.bisnis.com/read/20200527/531/1245132/naik-turun-positif

John, A. (2009). *The Art of Creative Thinking : How to be Innovative and Develpo Great Ideas.* Kogan Page Publishers.

Montgomery, D. C. (2009). *Introduction to Statistical Quality Control* (6th ed.). NJ: Wiley: Hoboken.

Oakland, J. S. (2003). *Statistical Process Control* (5th ed.). Oxford: Butterworth Heinenmann.

Olson, J. E. (2003). *Data Quality : The Accuracy Dimension.* San Fransisco: Morgan Kauffman.

Peirce, C. (1955). *Philosophical writings of Peirce* (. ed.). New York: Dover Publications.

Sebastian-Coleman, L. (2013). *Measuring Data Quality for Ongoing Improvement.* Waltham: Morgan Kauffman.

Wikipedia. (2020). *Wikipedia*. Retrieved August 3, 2020, from https://en.wikipedia.org/wiki/Coronavirus_disease_2019

Worldometers. (2020). *Reported Cases and Deaths by Country, Territory, or Conveyance*. Retrieved August 3, 2020, from https://www.worldometers.info/coronavirus/#countries

Biographies

Yuyun Hidayat is a lecturer in the Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Pdjadjaran, Indonesia. He received his Ph.D. in the University of Malaysia, Trengganu, Malaysia. He has published more then 20 papers in national and international journals and participated in many national and international conference. Currently initiated accurate predictions about Covid-19 case data in Indonesia

Dhika Surya Pangestu, is an undergraduate student. Currently, he is studying in the Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran.

Titi Purwandari, is a lecturer in the Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Pdjadjaran, the field of applied statistics, with a field of economic statistics.

Sukono is a lecturer in the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran. Currently as Chair of the Research Collaboration Community (RCC), the field of applied mathematics, with a field of concentration of financial mathematics and actuarial sciences.

Abdul Talib Bon is a professor of Production and Operations Management in the Faculty of Technology Management and Business at the Universiti Tun Hussein Onn Malaysia since 1999. He has a PhD in Computer Science, which he obtained from the Universite de La Rochelle, France in the year 2008. His doctoral thesis was on topic Process Quality Improvement on Beltline Moulding Manufacturing. He studied Business Administration in the Universiti Kebangsaan Malaysia for which he was awarded the MBA in the year 1998. He's bachelor degree and diploma in Mechanical Engineering which his obtained from the Universiti Teknologi Malaysia. He received his postgraduate certificate in Mechatronics and Robotics from Carlisle, United Kingdom in 1997. He had published more 150 International Proceedings and International Journals and 8 books. He is a member of MSORSM, IIF, IEOM, IIE, INFORMS, TAM and MIM.