Time-Series Clustering of the Regencies Hotel Room Occupancy Rate in Indonesia after the COVID-19 Pandemic

L Busaina[1]**, S B Panuntun**[2]**, S Pramana**[1,*]

[1]Politenik Statistika STIS, Jakarta, Indonesia
[2]BPS Statistics Indonesia, Jakarta, Indonesia

*Corresponding author's email: setia.pramana@stis.ac.id

Abstract. After COVID-19 pandemic, Indonesia entered the recovery era. The government provides incentives for the tourism industry's recovery. This policy was created because the impact of COVID-19 pandemic on tourism industry in each regency or city is different. This study investigates a different recovery pattern in regencies and cities across Indonesia. The data of this study consist of the room occupancy rate (ROR) from BPS Statistics Indonesia and from web scraping monthly data from the Agoda website between January 1, 2021, and August 1, 2023. The regencies and cities are clustered by ROR category using the dynamic time-warping method. According to the findings of the study, there is a difference in the rate of recovery of the tourism industry across Indonesia, with fast, medium, and slow being the most common. This could be the result of different policies in each regency or city to respond to the COVID-19 pandemic in their tourism industry.

1. Introduction

After the first case of COVID-19 reported at the beginning of March 2020, there are 1,528 cases of COVID-19 and 136 deaths because of COVID-19 in Indonesia by March 31, 2020. The nation's case fatality rate (CFR) is also at high number, at 8.9% [1]. These number are much higher than People's Republic of China, which is 4%. Then, the number of COVID-19 cases has been increasing significantly all over Indonesia. At the beginning of February 2021, the number of confirmed COVID-19 cases reached 1.2 million and 33.596 deaths [2]. After around three years, the president set rules that state the end of the COVID-19 pandemic, which can be found in Presidential Decree No. 17 at 2023.

The COVID-19 pandemic had an impact to employment in Indonesia. There is a job turnover to informal from formal sector. This is the result of layoff of employee. Based on Population Research Center-BRIN), there is 15.6% of workers in Indonesia experienced layoffs during the initial implementation of social restrictions, even 13.8% of them did not receive any benefits severance pay [3].

Tourism is also affected by the COVID-19 pandemic. Based on the Central Statistics Agency's data, there was a decline in the number of foreign tourist visits to Indonesia in July 2020 compared to July 2019, which was 159,763 visitors from 1,468,173 visitors. Compared to the number of foreign tourists visiting cumulatively from January to July 2019, which was 9.18 million visits, the number of foreign tourist visits to Indonesia decreased 64.64%, which was 3.25 million in the same period in 2020 [4].

That is the impact of the Tourism-supporting sector, such as hotels. In July 2020, the occupancy rate of star-classified hotel rooms in Indonesia decreased by 28.66 points compared to July 2019 (56.73%), which only reached an average of 28.07%. The average length of stay for foreign and domestic guests in star-classified hotels also declined compared to 2019. During July 2020, the average length of stay for foreign and domestic guests in star-classified hotels decreased by 0.14 points, which was recorded at 1.66 days [4].

Besides that, the COVID-19 pandemic also had an impact to the accommodation and food and drink provision sector of Bali. The results of the Covid-19 Impact Survey on Business Actors which was carried out in July 2020, the accommodation and food and drink provision sector was the business sector most affected by Covid-19 with 92.47 percent of business actors in this sector experiencing a decrease in income. This decline was greatly felt by the people of Bali. This is because 80 percent of Balinese people depend on the tourism sector, According to the Head of the Bali Association of Indonesian Travel Companies (ASITA) [5].

The UN World Tourism Organization (UN-WTO) expects that tourism industry of the world activities crash by around 73.9% in 2020. The UNWTO also estimated a loss of approximately 1.1 billion international tourist arrivals, with a loss of US$ 910 billion to US$ 1.1 trillion in export revenues due to the COVID-19 pandemic [6].

For developing and developed countries like Indonesia, the tourism industry has the biggest contribution to the economy. This industry is also one of the world's fastest growing economic sectors. In 2018, 1.4 billion international tourist arrivals worldwide were identified. In the amount of US$ 2.9 trillion for Gross Domestic Product (GDP) was accounted in 2019 from tourism sectors. This is the highest contribution by sector to the GDP [6].

In many countries, the pandemic impacted the tourism very hard and government interventions are required for tourism sector revitalization [7][8][9][10]. The focus of recovery is not only for large scales industries, but also for medium and small enterprises [10]. Japan's governments provide assistance including coaching, subsidies, and financial support to add value to the tourism industry in several areas [11][12].

After the COVID-19 pandemic, the tourism and creative economy sector in Indonesia continues to grow, prioritizing nature, ecotourism, wellness tourism, and adventure tourism. The government of Indonesia has launched several incentives for tourism sector via a national economic recovery program.

The government introduced an incentive program for the tourism and creative economy to the provincial governments and tourism operators in most of regencies or cities in Indonesia. This study aims to study the recovery of tourism sector in regencies and cities in Indonesia clustered by the room occupancy rate (ROR) to conceive recovery pattern after the COVID-19 pandemic.

2. Methods and Materials

2.1. Data

The data collection method employed in this research involved web scraping techniques applied to the Agoda website. The web scraping process was conducted by accessing the available API using the Request library in the Python programming language. Data collection was performed daily, gathering accommodation data for that specific day. The data period is from January 2021 until August 2023, monthly.

The Agoda website was chosen as the data source due to its status as one of the world's largest online travel booking platforms. Agoda has expanded its reach by offering a global network of over 2 million properties in more than 200 countries and regions worldwide. Additionally, according to similarweb.com, Agoda ranks as the second most visited accommodation and hotel category website in many Asian countries.

Numerous pieces of hotel information can be extracted from various pages on the Agoda website. For example, as shown in Figure 1, some of the information that can be obtained includes: hotel name, hotel rating, hotel location, hotel facilities, hotel reviews, hotel photos or images, and etc.

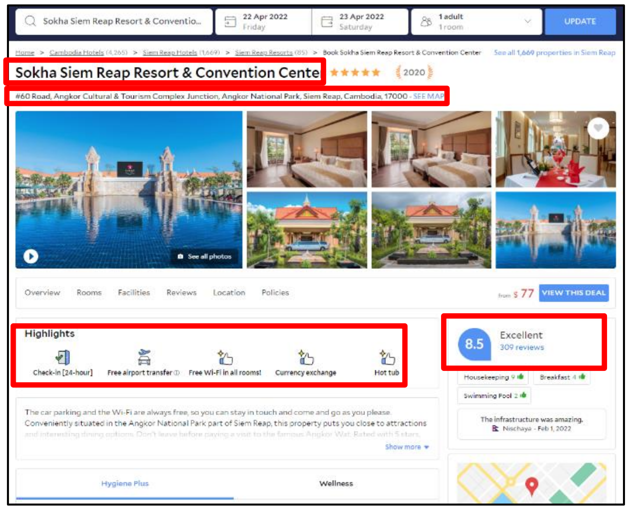

Figure 1. Hotel Information Page on the Agoda Website

There are three main steps, as shown in Figure 2, in collecting accommodation data from the Agoda website: accessing the Agoda API service endpoint, extracting information from the JSON-formatted response data, and storing the tabulated data results in a database [13]. These three steps are then repeated every day.

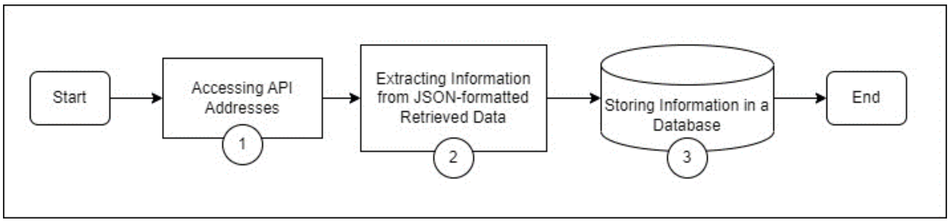

Figure 2. Data Collection Stages

2.1.1 *Accessing API addresses*. In this stage, the API service address used by the Agoda website will be accessed using the Request library in the Python language. The structure of the Agoda website's API address is as shown in Figure 3.

```
GetSecondaryData?finalPriceView=1&isShowMobileAppPrice=false&cid=1891460&&adults=2&checkIn=2021-03-02&isCalendarCallout=false&childAges=&numberOfGuest=0&missingChildAges=false&travellerType=1&showReviewSubmissionEntry=false&currencyCode=USD&isFreeOccSearch=false}&isCityHaveAsq=false&tspTypes=5%2c-1&los=1&hotel_id={hotelid}&all=false&price_view=1&pagetypeid=7
```

Figure 3. Agoda API Address Structure

From the above API address, it can be seen that there are several parameters that need to be inputted to collect accommodation data from the site. These input parameters include:

- The hotel code was obtained from the hotel directory on the Agoda website. This code typically consists of 8 digits. Hotel codes can be found in the Agoda site's sitemap. Currently, 49,534 hotel codes are used for daily data collection throughout Indonesia.Hotel Rating
- Check-in date and time or the data to be retrieved in the format YYYY-MM-DD. This parameter will be filled in according to the data collection date.

2.1.2 Extracting information from JSON. After accessing the API address, the Agoda server will return JSON-formatted data containing comprehensive information about the requested hotel. Below is an example of the API request result:


```
"protocol": "https:",
"translations": {...},
"hotelId": 19889600,
"culture": "en-us",
"origin": "ID",
"gtmId": "GTM-5TXL8JK",
"debug": false,
"browserInfo": {...},
"imageParams": {...},
"searchbox": {
  "defaultTab": -1,
  "searchCriteria": {
    "searchText": "Bobobox Pods Juanda",
    "checkIn": "2023-01-26T00:00:00",
    "checkOut": "2023-01-27T00:00:00",
    "occupancy": {
      "rooms": 1,
      "adults": 2,
      "children": 0,
      "checkbox": {
        "type": 0
      },
      "travellerType": -1
    },
    "datelessLanding": false
```


Figure 4. JSON Data from Agoda

In each acquired accommodation JSON, the necessary information is then collected, including: hotel name, hotel type, star rating, address, price, review score, number of reviews, room type, number of room types, number of available rooms, number of floors, number of restaurants, total number of rooms, year built, latitude and longitude coordinates, city or region, and hotel facilities

2.1.3 Storing information to database. Every piece of information collected from one hotel code is then inputted into a MySQL database. To date, the accumulated hotel accommodation data exceeds 12 million rows.

4.2. Data Processing Method

After the daily accommodation data is collected, referred to as raw data, it will be processed to obtain the daily Room Occupancy Rate (ROR) value for each hotel, which will then be aggregated by city, province, or nationally in Indonesia. There are four steps in processing hotel data: data preprocessing, calculating the number of occupied rooms, calculating the hotel's ROR, and finally aggregating it to a higher-level regional level. The data processing process is performed using the Pandas library in the Python programming language.

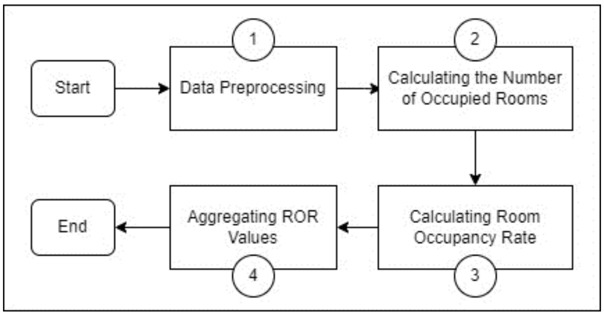

Figure 5. Data Processing Stages

2.2.1 *Data preprocessing*. In this stage, the raw data from the database will be "cleaned" before further calculations are performed. Data preprocessing involves:

- Removing incomplete hotel data: The collected data sometimes contains one or two incomplete pieces of information, such as hotel location, room count, or others. This can be due to errors during the data collection stage or because the hotel information is no longer available on the Agoda website.
- Eliminating duplicate data: Duplicate data is also sometimes found in the collected data, meaning that there are two pieces of information with the same hotel code on the same date. This can be due to errors during the data collection stage. Hotel data with identical hotel codes and scraping dates will then be removed.

2.2.2 *Calculating the number of occupied rooms*. The cleaned hotel data will then undergo a calculation of the number of occupied rooms. This information on occupied rooms is required as one of the parameters for calculating the room occupancy rate. This information can be obtained by calculating the difference between the total number of rooms and the number of available rooms.

2.2.3 *Calculating room occupancy rate (ROR)*. The hotel Room Occupancy Rate (ROR) is a comparison between the number of occupied rooms and the total number of rooms. A high ROR value indicates that the hotel is frequently visited by tourists. Conversely, a low ROR value indicates that the hotel is less attractive and therefore less frequented by tourists.

$$ROR = \frac{Number\ of\ Occupied\ Rooms}{Total\ Number\ of\ Rooms} \times 100 \tag{1}$$

2.2.4 *Aggregating ROR values*. Finally, the ROR data per hotel will be aggregated by region (city or province) by grouping it based on the city location information of the respective hotels.

2.3 Dynamic Time Warping Cluster

The dynamic time warping (DTW) is carried out by clustering the regencies/cities by ROR pattern across the month in three years. DTW is a class of pattern-matching algorithms for equalizing temporal sequences. Although two sequences have dynamic spatiotemporal differences, DTW intend to provide an optimal similarity-based match between them. By minimizing cumulative point-to-point distances, DTW will search for the optimal path. DTW is also flexible for describing the cumulative point-to-point distances to optimize tier performance for different datasets. DTW has become popular in speech recognition under varying speaking speeds, gesture recognition, and time series clustering because of its flexibility [14].

Given two time series $X = (x_1, x_2, \dots, x_N)$ and $Y = (y_1, y_2, \dots, y_M)$ The matrix of distance $C \epsilon R^{N \times M}$ describe all pairwise distances between X and Y. This matrix is known as the local cost matrix for the tier of two sequences X and Y:

$$C_l \epsilon R^{N \times M} : c_{i,j} = \|x_i - y_j\|, i \in [1:N], j \in [1:M] \tag{2}$$

The tier path made by DTW is a point's sequence $p = (p_1, p_2, \dots, p_K)$ with $p_l = (p_i, p_j)$ $[1:N] \times [1:M]$ for $l \in [1:K]$.

The function of cost related to a warping path contained in the local cost matrix (which describes all pairwise distances) is defined as follows:

$$c_p(X, Y) = \sum_{l=1}^{L} c(x_{nl}, y_{ml}) \tag{3}$$

The optimal warping path is the one with the minimal cost associated with alignment. It is called P^*. Every path between X and Y must be tested to find one optimal warping path. Here is the function of DTW distance:

$$DTW(X, Y) = c_{p^*}(X, Y) = \{c_p(X, Y), p \epsilon P^{N \times M}\} \tag{4}$$

where $P^{N \times M}$ is the set of all possible warping paths and builds the accumulated cost matrix or global cost matrix.

In general, DTW is a method that calculates an optimal match between two given sequences (ex: time series) with the following restrictions and rules:

- Every index of the first sequence should be matched with one or more indices from the other sequence, and vice versa.
- The first sequence's first index should be matched with the other sequence's first index (but not necessarily the only match).
- The first sequence's last index should be matched with the other sequence's last index (but not necessarily the only match).
- The association between first sequence's indices and other sequence's indices should increase monotonically, and vice versa, i.e., if $j > i$ is the first sequence's index, there should not exist two indices $l > k$ in another sequence, this index i match with index l and index j match with index k, etc.

We obtained the DTW distance by summing the elements of the path with the minimum cumulative distance. This is the root of the sum:

$$d_{DTW}(X,Y) = min\sqrt{\sum_{k=1}^{K} w_k} \tag{5}$$

where w_k is the distance that corresponds to the k-th element of the warping path W. This distance is the same as the Euclidean distance for the case where $n = m$ and only the diagonal of the LCM is traversed [15].

In this study, clustering was performed using the time series k-means available in the tslearn Python package with the dtw metric. tslearn is a Python package that provides machine learning tools for the analysis of time series. This package builds on (and hence depends on) the scikit-learn, numpy, and scipy libraries [16]. The steps are:

2.3.1 *Determining The Number of Clusters.* The elbow method is a method to define the optimal clusters number. This method used the total within-cluster sum of squares (WSS) as a function of the clusters number. The total WSS counts the clustering compactness and sets it to be as smaller. The 'elbow' or a corner of a line chart indicates the cluster number fits the model in a good way. The cluster number after 'elbow' can be said to be meaningless because adding the cluster's number does not improve the data's model.

2.3.2 *Performing The Cluster.* The ROR data are clustered using the time series k-means function dtw metric in Python with tslearn package. The K-means algorithm need the K-clusters number that we want to build from the dataset. With the aim of getting the clusters, the algorithm will be run on our dataset. It starts with randomly initializing K centroids and then calculating the 'Euclidean' distance between each point and both the centroids. This way, it defines the points to the closest centroid, thus making clusters [17].

2.3.3 *Cluster Similarity.* Phase synchrony is the moment-to-moment synchrony between two pattern or signals. The two pattern or signals from the data are selected subjectively by the researcher's interest. That selection must be for theoretical reasons. To calculate phase synchrony, the phase of the signal will be extracted by splitting the signal into its phase and power. This method called the Hilbert transform. This lets researcher know if two signals are in phase (moving up and down together) or out of phase.

3. Results

3.1 Determining The Number of Clusters
As shown in Figure 6, the 'elbow' is at three until six. In this study, we used six clusters to divide the regencies/cities based on the ROR pattern across the months in three years.

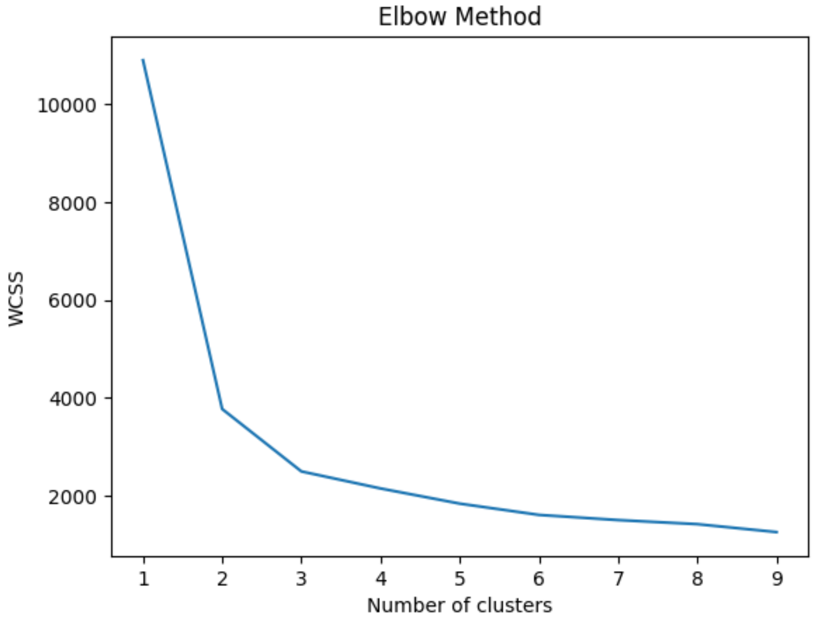

Figure 6. Elbow Plot

3.2 The Cluster of Regencies/Cities
The regencies/cities in Indonesia are divided into six group based on their monthly ROR over three years, from 2021 until 2023. Cluster 1 is cluster with regencies/cities in Indonesia that have pretty fast recovery from outbreaks. The regencies/cities in this cluster are not much effected by the COVID-19 pandemic. The ROR of regencies/cities in this cluster has been quite stable for the past three years. The regencies/cities are Batam, Balikpapan, Banda Aceh, etc.

Different from cluster 1, regencies/cities in cluster 2 have not yet recovered after the COVID-19 pandemic. As shown in Figure 8, the graphic of cluster 2 is very low and just went up at the end of 2023. The regencies/cities are Aceh Tenggara, Batang, Batu Bara, etc. Cluster 3 is quite similar to cluster 1. But regencies/cities in cluster 2 did not recover right away after the COVID-19 pandemic. There is a decline before going up. This cluster's ROR is also not as stable as cluster 1. The regencies/cities in this cluster can be said have quite a good recovery after the COVID-19 pandemic. The regencies/cities are Aceh Barat, Aceh Besar, Agam, etc.

Cluster 4 have a very volatile graphic compared to the five others. At first, the start, regencies/cities in this cluster couldn't handle the effect of the COVID-19 pandemic on their tourism sectors. This can be seen from the decline at the start. Along with time, the tourism sector in the regencies/cities in this cluster can adjust to the COVID-19 pandemic. This can be seen from the increase after the decline and after that, stability until the end of period. The regencies/cities are Brebes, Bulungan, Banda Aceh, etc.

Cluster 5 is quite similar to cluster 3. But, the graphic of cluster 5 is lower than cluster 3. The graphic also more volatile than cluster 3. The regencies/cities in this cluster can be said to have quite a good recovery after COVID-19 pandemic, but not as good as the regencies/cities in cluster 3. The regencies/cities are Bandar Lampung, Bangka Tengah, Banjar Baru, etc. Last, cluster 6. This cluster has a good recovery like cluster 1, but this cluster is better than cluster 1. This is can be seen by the stable of ROR's regencies/cities graphic. The regencies/cities in this cluster are not much effected by the COVID-19 pandemic. The regencies/cities are Aceh Tengah, Bangka Barat, Banyuasin, etc.

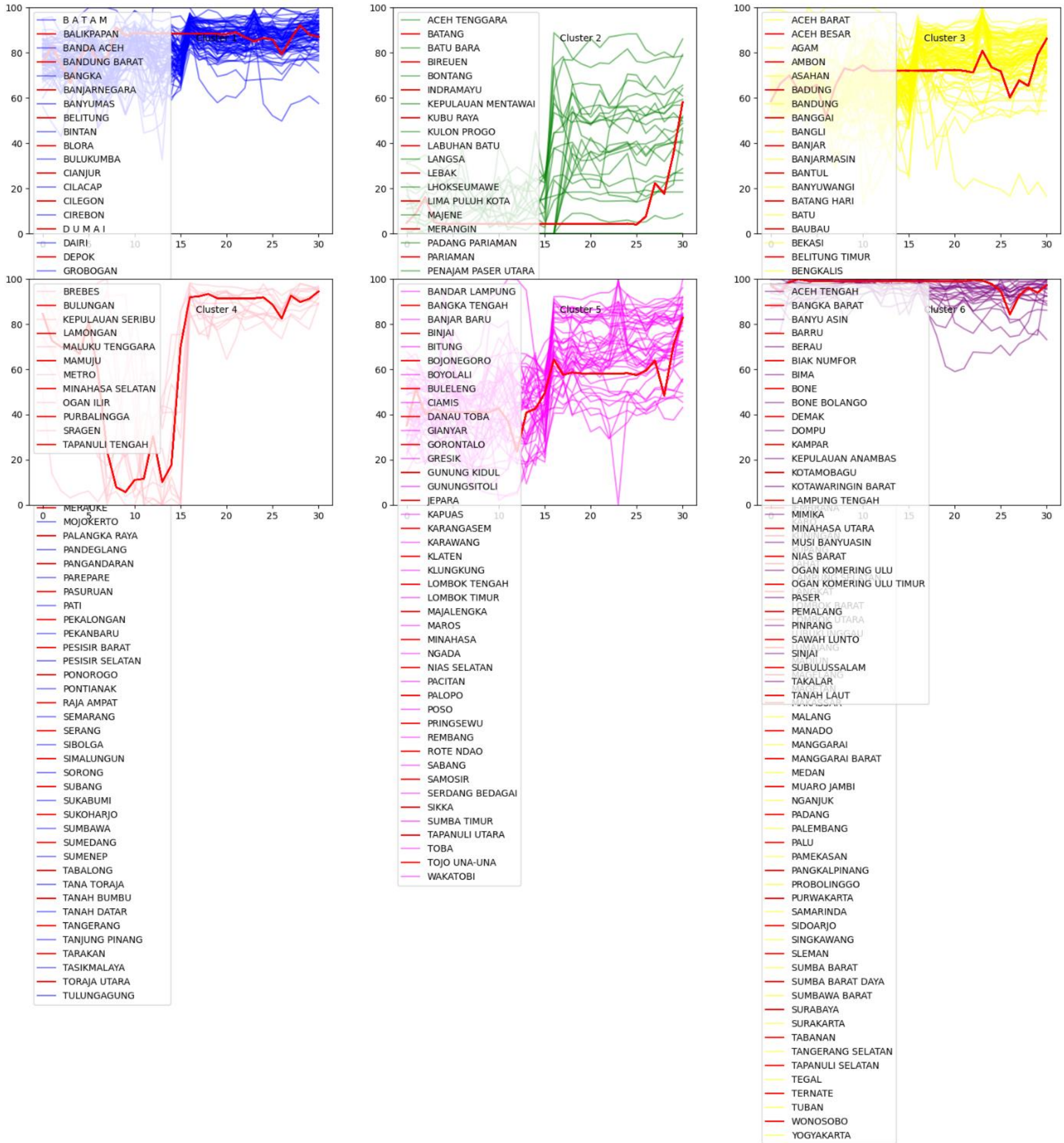

Figure 7. Visualization of Six Cluster Regencies/Cities in Indonesia

3.3 Cluster Similarity

We want to see the similarity of two of regencies/cities in the same cluster. For instance, Batam and Balikpapan are in the cluster 1.

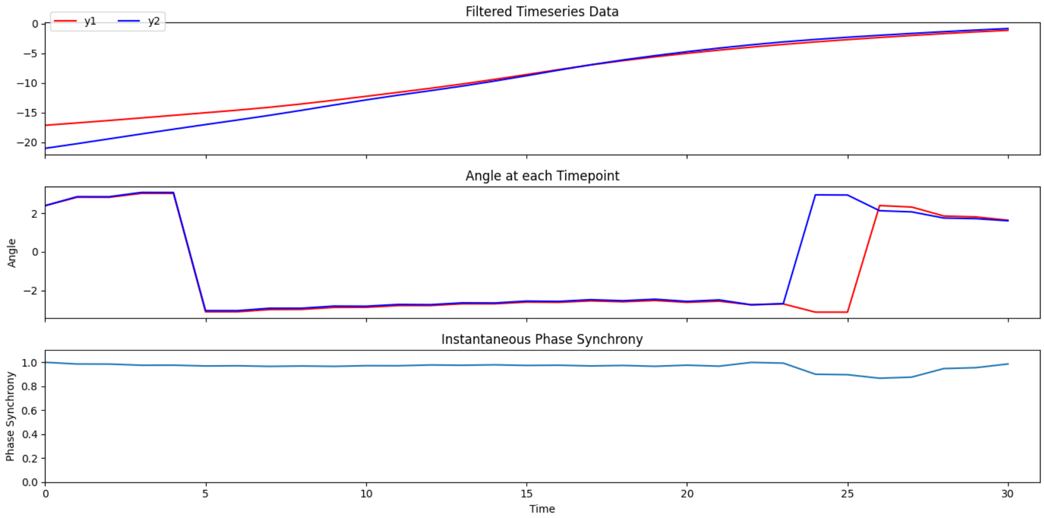

Figure 8. The Phase Synchrony of Batam and Balikpapan

As we can see, the phase synchrony is quite high between these 2 regencies/cities as they have a similar pattern of ROR, which was rightly clustered together using dtw. The phase synchrony mean of two regencies/cities are pretty high, at 0.9627.

We want to see the dissimilarity of two of regencies/cities in a different cluster. For instance, Batam is in the cluster 1 and Aceh Barat is in Cluster 3.

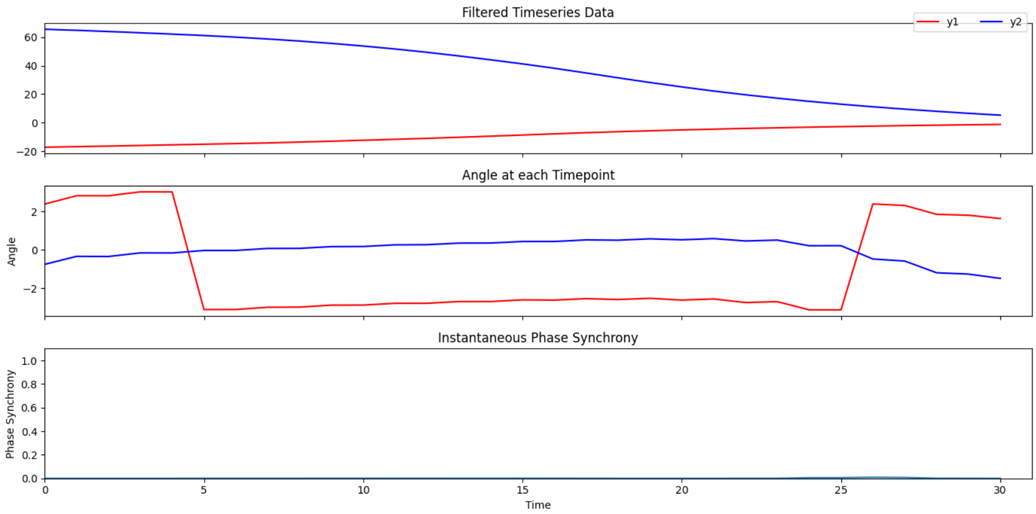

Figure 9. The Phase Synchrony of Batam and Aceh Barat

As we can see, the phase synchrony is quite low between these 2 regencies/cities as they have different pattern of ROR, which were not clustered together using dtw. The phase synchrony mean of two regencies/cities is very low, which is 0.0013.

4. Conclusion

The COVID-19 pandemic has an impact on the tourism sectors and the related industries all over Indonesia. However, the impact is different in every regency or city. The tourism industry in some regencies/cities managed to bounce back the situation progressively, such as Aceh Tengah and Batam or gradually, such as Aceh Barat and Brebes, but not for some regencies/cities such as Aceh Tenggara and Bandar Lampung. This could be the result of differences in policy in each regency/city. If the government of regency/city can respond well, the tourism industry in their regency/city can managed the impact of the COVID-19 pandemic very well. But if not, the tourism industry in their regency or city needs time to recover after the COVID-19 pandemic.

References

[1] S. Setiati and M. K. Azwar, "COVID-19 and Indonesia," no. April, 2020.

[2] S. Pramana, D. Y. Paramartha, G. Y. Ermawan, N. F. Deli, and W. Srimulyani, "Impact of COVID-19 pandemic on tourism in Indonesia," *Curr. Issues Tour.*, vol. 25, no. 15, pp. 2422–2442, 2022, doi: 10.1080/13683500.2021.1968803.

[3] A. P. S. Prasojo, "Pandemi Covid-19 Dan Turnover Ke Pekerjaan Informal," *J. Apl. Stat. Komputasi Stat.*, vol. 14, no. 2, pp. 49–62, 2022, doi: 10.34123/jurnalasks.v14i2.360.

[4] Y. Kristiana, R. Pramono, and R. Brian, "Adaptation Strategy of Tourism Industry Stakeholders During the COVID-19 Pandemic: A Case Study in Indonesia," *J. Asian Financ. Econ. Bus.*, vol. 8, no. 4, pp. 0213–0223, 2021, doi: 10.13106/jafeb.2021.vol8.no4.0213.

[5] M. Z. Ahmad and E. T. Astuti, "Dampak Pandemi Covid-19 terhadap Pendapatan Sektor Penyediaan Akomodasi dan Makan Minum Provinsi Bali," *J. Apl. Stat. Komputasi Stat.*, vol. 15, pp. 1–14, 2023.

[6] Z. Zhu, X. Gao, and R. Chen, "Impact of Covid-19 on Tourism Industry," *Proc. 2021 3rd Int. Conf. Econ. Manag. Cult. Ind. (ICEMCI 2021)*, vol. 203, no. 102834, 2022, doi: 10.2991/assehr.k.211209.419.

[7] D. Chiawo, C. Haggai, V. Muniu, R. Njuguna, and P. Ngila, "Tourism Recovery and Sustainability Post Pandemic: An Integrated Approach for Kenya's Tourism Hotspots," *Sustain.*, vol. 15, no. 9, pp. 1–13, 2023, doi: 10.3390/su15097291.

[8] R. B. Recio *et al.*, "1. Navigating the impacts of COVID-19 pandemic in Asia: state measures, grassroots responses and implications for recovery," *Reg. Stud. Policy Impact Books*, vol. 5, no. 1, pp. 19–29, 2023, doi: 10.1080/2578711x.2023.2196210.

[9] L. Plzáková and E. Smeral, "Impact of the COVID-19 crisis on European tourism," *Tour. Econ.*, vol. 28, no. 1, pp. 91–109, 2022, doi: 10.1177/13548166211031113.

[10] M. Roman, M. Roman, E. Grzegorzewska, P. Pietrzak, and K. Roman, "Influence of the COVID-19 Pandemic on Tourism in European Countries: Cluster Analysis Findings," *Sustain.*, vol. 14, no. 3, 2022, doi: 10.3390/su14031602.

[11] A.-G. Baranyi, "Strategic management in Japanese tourism industry during the COVID-19 pandemic," *Proc. Int. Conf. Bus. Excell.*, vol. 15, no. 1, pp. 1127–1134, 2021, doi: 10.2478/picbe-2021-0105.

[12] N. Yagasaki, "Impact of COVID-19 on the Japanese travel market and the travel market of overseas visitors to Japan, and subsequent recovery," *IATSS Res.*, vol. 45, no. 4, pp. 451–458, 2021, doi: 10.1016/j.iatssr.2021.11.008.

[13] Y. Adhinugroho *et al.*, "Development of online travel web scraping for tourism statistics in indonesia," *Inf. Res.*, vol. 25, no. 4, p. 1, 2020, doi: 10.47989/irpaper885.

[14] Y. Jiang *et al.*, "EventDTW: An improved dynamic time warping algorithm for aligning biomedical signals of nonuniform sampling frequencies," *Sensors (Switzerland)*, vol. 20, no. 9, 2020, doi: 10.3390/s20092700.

[15] S. Pramana, Y. Yuniarti, D. Y. Paramartha, and S. B. Panuntun, "Mobility Pattern Changes in Indonesia in Response to COVID-19," *Econ. Financ. Indones.*, vol. 67, no. 1, p. 75, 2021, doi: 10.47291/efi.v67i1.924.

[16] T. Romain, "tslearn's documentation," tslearn's documentation, https://www.scribbr.com/ieee/ieee-website-citation/#:~:text=To%20write%20an%20IEEE%20reference,or%20removed%20in%20the%20future. (accessed Aug. 30, 2023).

[17] H. M. Gawai, "Project Nottingham: a Study of User Behavior on Mobile Investing Applications," no. August, 2022.