Visualising and predicting the COVID-19 outbreak in Malaysia using network analysis and support vector regression

Siti Nurhidayah Sharin

Universiti Putra Malaysia

Mohamad Khairil Radzali

Universiti Putra Malaysia

Muhamad Shirwan Abdullah Sani (✉ shirwansany@iium.edu.my)

International Islamic University Malaysia <https://orcid.org/0000-0002-8202-1491>

Research Article

Keywords: Coronavirus, COVID-19, Network analysis, Machine learning, Support vector regression

Posted Date: November 17th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1087977/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Coronavirus disease 19 (COVID-19) was first discovered in December 2019 in Wuhan, China and spread quickly throughout the world, affecting the economy, social disruption, and public health. Concerning confirmed and death COVID-19 cases in Malaysia, the correlation of states and prediction model using support vector regression (SVR) associated with COVID-19 in Malaysia are yet to discover. Hence, the proposed works employ network analysis and SVR from July 2020 (Q3 2020) until June 2021 (Q2 2021) based on given data by the Ministry of Health Malaysia (MoH) (i) to correlate and visualise the COVID-19 pandemic spread between the states and (ii) to predict the cumulative number of COVID-19 confirmed and death cases. Network analysis was employed using Spearman rank coefficients and revealed an increasing degree of connectedness between different states, thus pinpointing key actors of transmission. Meanwhile, the proposed SVR predictive model could forecast the future COVID-19 cases and deaths (July 2021 to December 2021), with an excellent regression score ($R^2 = 0.829$) and low mean squared error ($MSE = 0.171$), as well as root mean square error ($RMSE = 0.413$); hence, making this model reliable enough. Current data demonstrate that network analysis and the SVR model provide insightful and potential information to minimise COVID-19 transmission.

1. Introduction

World Health Organization (WHO) reported that 44 cases of unknown pneumonia aetiology were detected in Wuhan City, Hubei Province of China between 31st December 2019 to 3rd January 2020 [1, 2]. On 7th January 2020, a new type of coronavirus associated with the unknown pneumonia aetiology was then identified by Chinese authorities, later known as coronavirus disease 2019 (COVID-19) on 12th January 2020 [1, 3]. WHO declared this communicable disease a global pandemic on 11th March 2020 due to the rapid dissemination of the worldwide outbreak [1, 3–5].

In Malaysia, the first COVID-19 case was confirmed on 25th January 2020 and continues to disseminate fast in the country [6]. Between July to September 2020, Malaysia recorded a successful story in flattening the curve with less than 100 daily cases due to strict Movement Control Order (MCO) effectiveness [7]. Nonetheless, Malaysia is once more hit by the third wave of the outbreak in late September 2020 up until now [1]. To prevent this sporadic and communicable disease from worsening, Malaysian health sectors and enforcement authorities (police and military), academicians, and statisticians are collaborating to manage this issue [3]. The most recent standardised operating procedures (SOP) revised are needed as the guidelines for the public since the situation of COVID-19 changes over time. Therefore, many academicians and statisticians are willing to work with the Malaysian government to provide reliable data through data insights to control this communicable disease [7].

To date, with the emerging programming languages using Python and R in data visualisation and prediction, public knowledge is expanding rapidly, and people are getting a greater understanding of curbing the COVID-19 disease. Various statistical methods have been employed to visualise and predict

the COVID-19 cases worldwide in addressing the dissemination of this disease. Network analysis and support vector regression (SVR) are some of the visualisation and prediction techniques that have been utilised over the years in many sectors such as science, finance, economy, tourism, social, and health system [6]. Network analysis is a simple yet powerful method to evaluate the pandemic risk by visualising the correlation among various regions based on real-time and historical data [8]. Meanwhile, SVR is one of the famous and influential prediction tools that has been employed recently, especially in predicting COVID-19-related cases [9]. SVR is a supervised machine learning technique generalised from a support vector machine (SVM) [10]. The SVR algorithm is commonly used to predict discrete values by finding the best-fitted line for both types of linear and non-linear regression [5].

Estimating pandemic risk based solely on the confirmed cases gives restricted information regarding pandemic patterns [8]. Recently, the relationship among states in Malaysia associated with COVID-19 resulting in higher confirmed cases is still unknown, and the prediction model of COVID-19 spread in Malaysia using support vector regression is not yet discovered. Therefore, this study utilised network analysis and the SVR model to understand better the number of confirmed cases and deaths and the states' relationship on the increment of COVID-19 in Malaysia. These statistical techniques are essential complementary tools to obtain reliable visualisation and prediction of Malaysia's forthcoming confirmed cases and deaths. These techniques can also give an early insight and understanding in preventing and curbing this COVID-19 disease from rampantly spreading in Malaysia.

2. Methodology

2.1 Data source

In this study, three .csv files (cases_states, cases_malaysia and deaths_malaysia) were retrieved from an open-source website provided by the MoH via <https://github.com/MoH-Malaysia/covid19-public>. These three .csv files entailed COVID-19-related information from July 2020 to June 2021 (365 days).

2.2 Bar plot and Spearman's analysis

Prior to building the bar plot and Spearman's analysis using Python 3.3, the datasets of cumulative confirmed cases (cases_malaysia) and deaths (deaths_malaysia) in Malaysia were combined into a .csv file dataset. In this analysis, a total number of 365 days was observed (3rd quarter of the year 2020, 4th quarter of the year 2020, 1st quarter of the year 2021 and 2nd quarter of the year 2021; $n = 365$). A new column named 'days' was created in the same combined dataset to represent the date into particular days in 2020 and 2021 (365 rows $\times$ 4 columns). A bar plot was built to observe the general COVID-19 cumulative confirmed cases and deaths in Malaysia from July 2020 to June 2021. Based on the skewness (confirmed cases = 1.14, deaths = 2.31), kurtosis (confirmed cases = 0.57, deaths = 4.52) and Shapiro-Wilk test at p-value (confirmed cases) = 3.34×10^{-17} , p-value (deaths) = 5.57×10^{-28} , the data was assumed as non-parametric (not normally distributed). Therefore, Spearman's analysis at p-value < 0.05 was conducted to determine the correlation strength and significant difference between cumulative confirmed cases and deaths in Malaysia.

2.3 Network analysis

A dataset of daily confirmed cases of states in Malaysia (cases states) was employed in this analysis consisting of 5840 observations (16 states $\times$ 3rd quarter of the year 2020 $\times$ 4th quarter of the year 2020 $\times$ 1st quarter of the year 2021 $\times$ 2nd quarter of the year 2021, $N = 5840$). A new column named 'quarter' is created in the dataset to classify the months into the respective quarter of 2020 and 2021, while the original columns were remained (5840 rows $\times$ 4 columns).

Network analysis was carried out to study the connections between states in Malaysia in response to the COVID-19 spread. Basic network properties such as Spearman rank correlation and degree of interaction were computed using the iGraph R package (V 3.5.1) [11] before being fed into Cytoscape (V 3.8.2) for visualisation. The degree of interaction was determined based on the correlation (edge) formed between the states. By default, the states' correlation was considered a strong co-occurrence if the Spearman correlation coefficient is > 0.5 [12] and statistically significant if the computed p-value < 0.05 . Additionally, the NetworkAnalyzer Cytoscape plugin was applied to calculate the number of significant nodes and edges that denoted the networks' topological properties. In this study, nodes represented the states, whereas edges represented the correlations (nodes). Additionally, connected components were the maximum group of nodes connected by edges in a path, while network density is the ratio of observed edges over possible edges in a given network.

2.4 Support vector regression (SVR)

In this prediction modelling, the COVID-19 cases in Malaysia in the 2nd quarter of 2021 showed a high correlation among all states based on network analysis. Therefore, the same dataset as in Section 2.2 was used to forecast cases for the whole Malaysian country. The steps of building the SVR model were adapted from Parbat & Chakraborty (2020). In this study, the SVR model was developed using a 70% training dataset via the Radial Basis Function (RBF) kernel with epsilon = 0.1. The developed SVR model was validated with a 30% testing dataset, respectively. Based on the SVR model, predictive values of future confirmed cases and deaths in Malaysia was computed by employing the Python 3.3 command `y_pred = scy.inverse_transform(regressor.predict(scx.transform([[100]])))`. These values were manually input in Microsoft Excel to construct the SVR forecast. The performance of the SVR model was evaluated based on statistical goodness-of-fit criteria, e.g., mean squared error (MSE), root mean square error (RMSE), regression score (R^2) and accuracy.

3. Results And Discussion

3.1 General insight of COVID-19 confirmed cases and deaths in Malaysia

This study visualised the trend of COVID-19 cumulative confirmed cases and deaths in Malaysia from July 2020 to June 2021 via Fig. 1. A remarkable increasing trend was observed on cumulative confirmed

cases in Malaysia starting from Day 70 (8th September 2020) with a triple-digit number of 100 cases (Fig. 1a). The similar increasing pattern was observed on accumulated confirmed deaths on day 84 (22nd September 2020) with 3 cases (Fig. 1b). The confirmed deaths associated with COVID-19 on 22nd September 2021 involved individuals aged 48- and 54-year-old in Sabah who showed symptoms of COVID-19 on Day 76 (14th September 2020) and Day 80 (18th September 2020), respectively. Another confirmed death involved an asymptomatic 72-year-old individual from Alor Setar, which was found positive on 19th August 2020 (Day 50). Generally, the signs of COVID-19 usually appear after 1 – 14 days of the incubation period but commonly occur after five days [13]. Based on the estimation of serial interval and incubation period, it was estimated that 44% of transmission probably had occurred before symptoms appeared [14, 15]. In the previous study, it was also reported that there was a significant relationship between viral load and incubation period, in which the initial viral load begins to increase within 5 to 6 days before the first symptoms appeared [14, 16, 17]. The incubation period becomes shorter when the viral loads are high, corresponding to low Cycle Threshold (Ct) values. Since the viral loads evolve, the high viral loads are probably the primary cause of transmission [16, 17].

There are significant increments (p -value < 0.05) on daily confirmed cases and deaths in Malaysia from July 2020 to June 2021 (Table 1). An excellent correlation between the number of confirmed cases and deaths was also observed (0.907, p -value < 0.05), in which high cases influenced a high mortality rate (Table 1). Previously, Malaysia has successfully curbed the first and second waves of the outbreak by lowering the confirmed cases in July until early September 2020 (Day 1 – Day 69), with less than 100 cases per day (Fig. 1a) [1]. However, an increment of confirmed cases is observed in the fourth week of September 2020 (Day 85 – Day 92), commencing the third epidemic wave in Malaysia [1]. The increasing of confirmed cases occurred right after the state election in Sabah on 26th September 2020 (Day 88) [1]. Many cases are associated with high-risk areas in Sabah, which led to 29 clusters located in Sabah, and 26 clusters had an index case with travel history to Sabah mainly from the east of Sabah, including Lahad Datu, Semporna, Tawau, Kunak and Sandakan areas [18]. Despite the increment number of confirmed cases in Sabah, the control of people movements over the country was not restricted. The swab tests were not mandatory before travelling among the states, resulting in the confirmed cases being continuously escalated from single-digit to thousands per day [19].

Furthermore, the condition in Sabah has become worse due to lack of awareness about COVID-19 and its symptoms, especially among people who live in rural areas, failure to comply with the instructions by health officers, as well as the paucity of healthcare workers in Sabah's hospitals which had caused 10 400 backlogged COVID-19 test samples [20]. Based on the Department of Statistics Malaysia Official Portal in 2021, Sabah was one of the top three states with the highest population composition of 11.7%, preceded by Selangor with 20.1% and followed by Johor with 11.6% [21]. However, the population density of 99 people per one square kilometre in Sabah (52/km²) is not relatively high as in Federal Territory (FT)

Kuala Lumpur (7188/km²), FT Putrajaya (2354/km²), Selangor (674/km²), and Johor (174/km²) [19, 21]. Although the population density in Sabah is not densely high as in Peninsular Malaysia, the majority of 3.83 million people in Sabah are settling along the Sabah's coastline instead of the interior mountainous part, which caused the spike of the COVID-19 cases in those areas after the state election of Sabah [22]. Besides that, irregular and undocumented migrants in Sabah have caused the COVID-19 situation in this state to become more challenging to COVID-19 screening tests and contact tracing since they were at risk of detention or deportation if found, resulting in difficulty in getting robust and reliable data [23].

Based on Figure 1(a), the increment of confirmed cases in Day 70 – Day 215 (8th September 2020 – 31st January 2021) were not steep as compared to confirmed cases in Day 280 – Day 340 (6th April 2021 – 5th June 2021). The commencing of triple-digit COVID-19 cases was observed on Day 70 (8th September 2020) during the recovery movement control order (RMCO) and later exponentially increased during the conditional movement control order 2.0 (CMCO 2.0) in Day 106 – Day 196 (14th October 2020 – 12th January 2021) [24]. The exponential increment in COVID-19 cases during CMCO 2.0 was due to the emergence of new clusters right after the Sabah state election held on 26th September 2020. Malaysia government was then decided to implement the movement control order 2.0 (MCO 2.0) again on 13th January 2021 (Day 197) after observing worrying COVID-19 numbers that reached thousands per day [25]. During Day 197 – Day 247 (13th January 2021 – 4th March 2021), MCO 2.0 successfully showed a decreasing trend in COVID-19 cases per day. However, the implementation of MCO 2.0 was not last long. The government was once again announced for the third CMCO on 5th March 2021 for the safety of Malaysia's economy [26]. Although the MCO 2.0 execution was not stricter than MCO 1.0 and allowed most businesses to operate, Malaysia still recorded a RM 600 million loss per day since most businesses were struggling in recovering phase, and investors remain pessimistic [27].

During the CMCO 3.0 and MCO 3.0 (Day 280 – Day 340), the spike in COVID-19 cases was observed higher than CMCO 2.0 and MCO 2.0 (Day 70 – Day 215) due to the mass testing in Selangor and Penang, failure to comply with the standard operating procedures (SOPs) by the societies, as well as the emergence of new coronavirus variants with higher infection rates comprising of United Kingdom variant (Alpha Variant B.1.1.7), South African variant (Beta Variant B.1.351), and Indian variant (Delta Variant B.1.617.2) [28]. Social gathering activities and the concentration of people in crowded spaces are the primary causes of COVID-19 spreading due to societies' difficulty in complying with the SOPs. In Selangor, the government state decided to fully utilise the antigen rapid test kit (RTK-Antigen) method during the mass testing since the testing results can be obtained in the same day as compared to the reverse transcription-polymerase chain reaction (RT-PCR) method, which the testing results can take up to three days and cause backlog [29]. The purpose of mass testing using RTK-Antigen was to promptly detect and isolate the silent carriers and understand the positivity rate and hotspots better. Therefore, the

expectation of COVID-19 cases to spike higher than the previous was unsurprised. The increasing number of COVID-19 cases has also caused an overburden on the healthcare system, particularly in highly affected states such as Selangor, Sarawak, Penang, Kelantan and FT Kuala Lumpur, leading to the escalating of COVID-19 deaths [18].

3.2 Correlation among states using network analysis

In the current study, network analysis was constructed to determine the relationship of states in Malaysia associated with confirmed COVID-19 cases. COVID-19 pandemic risk can be assessed and visualised using correlation and network analysis. States that were densely connected with others will exhibit higher complexity of edges in the network graph suggesting the critical centre of virus transmission throughout the networks [8]. In this study, Spearman's rank coefficient was used to measure the polarity (-1 to 1) of correlation between states based on daily confirmed cases. A positive value of the Spearman rank correlation represents co-existence, whereas a negative value indicates opposition between two states. The starting point of a timeframe in the current study was set for quarter 3 (Q3) of 2020 (July – September), quarter 4 (Q4) of 2020 (October – December), quarter 1 (Q1) of 2021 (January – March 2021), and quarter 2 (Q2) of 2021 (April – June) as daily confirmed cases fluctuated, prompting this study to investigate the correlation between states that led to the spiked number of cases. Table 2 summarises the number of nodes and edges and the analysis time of these quarters of time frame. The correlations that were significantly different (p -value < 0.05) were discussed in this section.

In Q3 of 2020, Sabah and Kedah were highly correlated ($r = 0.329$) despite having a weak positive correlation compared to Perak with Perlis ($r = 0.322$) and Malacca with Selangor ($r = 0.326$) (Figure 2a). Sabah and Kedah had 1505 and 270 confirmed cases, respectively, throughout the entire quarter, yet no reports linking the COVID-19 transmission between these two states. Sabah reported the first cluster on 1st September at Lahad Datu District Police Headquarters lock-up, accounting for 74.7% of the total new cases between 7th to 13th September 2020 [23]. As for Kedah, the earliest positive COVID-19 cases were contributed by the PUI Sivagangga cluster and spread to Perlis and Penang. Several factors linked to COVID-19 transmission in Kedah included lack of physical distancing, family gathering who flouted standard operating procedures (SOP), and hospital visits [18]. Generally, the MoH expressed an alarming concern of COVID-19 spread as most respiratory viral tract infections were reported during rainy seasons in tropical regions [30]. Wan Nik et al. (2019) stated that two monsoon seasons with rapid wind speed faced by Malaysia: late May to September and November to March in Southwest and Northeast Malaysia, respectively, might contribute to the transmission of COVID-19 within this time frame.

Surprisingly, the total confirmed COVID-19 cases increased from 2594 to 101786 from Q3 to Q4 of 2020. Network analysis revealed a total of nine states including FT Kuala Lumpur, Johor, Perak, Selangor, Kelantan, Pahang, Negeri Sembilan, Pulau Pinang and Sabah that were significantly correlated in which Johor and FT Kuala Lumpur had the highest degree of interaction among others based on the visualisation (Figure 2b). Of the nine states, FT Kuala Lumpur and Selangor had a strong positive

correlation ($r = 0.765$), followed by Johor and Selangor ($r = 0.756$). The increasing number of COVID-19 cases might have been contributed by geographical factors such as high population density and population movement, especially in urban centres [32]. Additionally, the confirmed COVID-19 cases in FT Kuala Lumpur and Selangor were also contributed by manufacturing industries [33]. Johor was also positively correlated with FT Kuala Lumpur ($r = 0.755$), Pahang ($r = 0.674$), Perak ($r = 0.607$), and Kelantan ($r = 0.595$). Other correlations (r values and degree of interaction) are summarised (Supplementary 1).

However, Johor and Sabah showed a negative correlation ($r = -0.530$), suggesting strategic implementations in Sabah that might reduce the spread of COVID-19 in Johor. Several comprehensive implementations in Sabah including limited non-essential services, implementation of Targeted Enhanced Movement Control Order (TEMCO), increasing of healthcare equipment (beds, ventilators, etc.) capacity, medical personnel mobilisation, point-of-entry testing, maximum daily RT-PCR testing capacity, mandatory 14-day quarantine at designated centres, quarantine centres for undocumented migrants, stringent border control, and more. Apart from that, Johor was placed under Conditional Movement Control Order (CMCO) and MCO, closing worship places, opening COVID-19 Quarantine and Low-risk Treatment Centres, and enforcing SOPs [34].

In Q1 of 2021 (January to March 2021), Figure 2c showed more complex networks in which 11 states formed 42 significantly strong positive correlations. This finding was supported by shifting the National Transmission State Assessment from Stage 2 (Localised Community Transmission) to Stage 3 (Large-scale Community Transmission – low confidence). Based on the visualisation, Johor, Kedah, Sabah, Selangor, Terengganu, and FT Putrajaya exhibited a similar and highest degree of interaction associated with COVID-19 transmission. Out of these states, Sabah and FT Putrajaya have the strongest positive correlation ($r = 0.834$), followed by Johor and FT Kuala Lumpur ($r = 0.754$). Additionally, Johor also exhibited positive correlations with Selangor ($r = 0.715$), Sabah ($r = 0.646$), Terengganu ($r = 0.637$), FT Putrajaya ($r = 0.574$), Melaka ($r = 0.566$), Kedah ($r = 0.552$), Pahang ($r = 0.536$), and Negeri Sembilan ($r = 0.535$). Other correlations (r values and degree of interaction) are also summarised (Supplementary 2).

The increase of COVID-19 spread was potentially due to inter-state travel during holiday celebrations, mainly in FT Kuala Lumpur, Selangor, Johor, Penang, Sabah, Kedah, Perak, Negeri Sembilan, and Malacca [18]. A few festive seasons (Q1 2021) that applied to these states, including New Year's Day (1st January 2021), Thaipusam (28th January 2021), and Chinese New Year (12th - 13th February 2021), hence might lead to an increase in population movement within the time frame. In addition, data from Google Mobility Report also indicated a surge of cumulative population movement (workplace, retail and recreations, parks, grocery and pharmacy, and transit stations) for Johor, Kedah, Sabah, Selangor, Terengganu, and FT Putrajaya within the quarter (Supplementary 4) [35], suggesting potential factor of COVID-19 spread [36].

The second quarter (Q2) of 2021 revealed the most complex network in the current finding (Figure 2d). All 16 states significantly correlated in COVID-19 transmission nationwide and exponentially increased the

number of confirmed and death cases (Figure 1). All four states, including Selangor, Pahang, Malacca, and Kedah, had the highest degree of interaction (12 edges), among others. The National Transmission Stage Assessment was consecutively changed within this quarter from Stage 3 (Large-Scale Community Transmission – low confidence) to Stage 3 (Large-Scale Community Transmission – moderate confidence) effective on 26th April 2021, which further shifted to Stage 3 (Large-Scale Community Transmission – high confidence) on 10th May 2021. Kedah and Selangor remained the states with the highest degree of interaction from 9 to 12 correlations (edges) from Q1 to Q2 of 2021, respectively. During the time frame, the surge cases in Kedah and Selangor were linked to densely populated areas and those who contracted the virus at factories [37].

Additionally, Selangor and FT Kuala Lumpur had a strong positive correlation ($r = 0.886$). Subsequently, Melaka exhibited the highest positive correlations with Selangor ($r = 0.883$), Negeri Sembilan ($r = 0.860$), and Pahang ($r = 0.854$). Both Selangor and FT Kuala Lumpur consistently reported a high proportion of confirmed cases due to the burden of the healthcare system apart from Sarawak, Penang, Johor, and Kelantan. Moreover, multiple hospitals across FT Kuala Lumpur and Selangor struggled with surged admission of critically ill COVID-19 patients requiring oxygen support during this period [38]. Other correlations (r values and degree of interaction) are also summarised (Supplementary 3).

A total of 132673 and 11873 confirmed COVID-19 cases in Selangor and Malacca had been reported in the current quarter. However, no reports between Malacca and Selangor were found despite having a strong positive correlation ($r = 0.883$), and we inferred the transmission might be due to inter-state travel and rapid spread of COVID-19 within the local community, educational institutions, and places of worship. Considering the rise of population movement (residential, grocery and pharmacy) (Supplementary 5), the asymptomatic carriers and the emergence of new COVID-19 variants in Q1 of 2021 could potentially cause the virus to be more transmittable throughout the states [39]. In addition, several national festive seasons in Q2 of 2021 (April – July 2021), including Labour Day (1st May 2021), Eid Fitr (13-14th May 2021) and Wesak Day (26th May 2021), might link to the increase of population movements.

3.3 Prediction of confirmed cases and deaths in Malaysia using support vector regression model

In this study, SVR was employed to observe the reliability of this model in predicting the future number of confirmed cases and deaths in Malaysia. All SVR models constructed using a 70% training set of confirmed cases vs days, confirmed deaths vs days, and confirmed cases vs confirmed deaths obtained the best R^2 values with 0.846, 0.859 and 0.829, respectively (Table 3). High R^2 values (near to 1) together with low MSE and RMSE (near to zero) indicated that all SVR models are considered excellent and reliable predictive models [40]. Besides, low MSE and RMSE values also influence the high accuracy of SVR models. Meanwhile, the R^2 values of 30% testing set for confirmed cases vs days, confirmed deaths

vs days and confirmed deaths vs confirmed cases in Figure 3a, Figure 3b and Figure 3c are 0.855, 0.909 and 0.836, respectively (Table 3).

Based on Figure 3a and Figure 3b, it was observed that the predicted values of daily confirmed cases and deaths from Day 1 until Day 365 (July 2020 to June 2021) were lower but almost similar to the actual reported cases. This finding indicated that the SVR was a reliable and robust prediction method to briefly predict the impending number of daily infections and mortality rates. Nevertheless, in this study, the prediction of the SVR model was solely based on historical data and did not take into account the reproduction number (R_0). The R_0 is the estimated number of cases that an infected individual causes in spreading the disease to other individuals who are not yet infected. The R_0 was utilised to determine the potential for a disease to spread in that population [6]. Recently, the determination of the R_0 value is vital since this value is able to indicate the severity rate of the outbreak to spread among an individual [41]. Since our current aim only focuses on observing the SVR model's reliability in predicting the forthcoming COVID-19 cases, the R_0 value may be proposed together with the SVR model for future study. Figure 4 forecasts the future number of daily infection and mortality rates was predicted from July 2021 until December 2021. It was observed that the number of confirmed cases and deaths in Malaysia will spike around July until August 2021, and a downward trend was expected to start in September 2021 (Figure 4) provided that the MoH and Malaysia government remain the similar intervention to curb the COVID-19 transmission. However, it was stressed that this forecast was merely based on daily confirmed cases and deaths variables, and more variables are needed to observe the influence of other variables on the COVID-19 trend in Malaysia.

4. Conclusion

Our approach demonstrated visualisation of COVID-19 pandemic risk through interactions between states in Malaysia through network analysis despite depending on reported confirmed COVID-19 data only. The connection of states increased within the study time frame, and a few states with a higher degree of interaction were identified as the potential key of transmission. The future number of COVID-19 cases and deaths could predict using the SVR model. The reliability of the SVR model with low MSE and RMSE values and excellent regressor scores is somewhat comparable with other predictive models; thus, it can be proposed in further prediction study. This study could deduce that the SVR model was equivalent to other prediction models such as logistic regression, autoregressive integrated moving average (ARIMA), and long short-term memory (LSTM) in predicting COVID-19 cases in other study fields.

Nevertheless, our current findings are only limited to daily confirmed cases and deaths variables. Hence, more variables are needed to observe the influence of other variables on the COVID-19 trend in Malaysia. This study helps to understand the spreading of the virus among the communities and gives early knowledge in preparing to mitigate the daily confirmed cases.

Declarations

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Acknowledgement

The authors self-financially supported the present work.

References

1. Hashim JH, Adman MA, Hashim Z et al (2021) COVID-19 Epidemic in Malaysia: Epidemic Progression, Challenges, and Response. *Front Public Heal* 9:1–19
2. COVID-19 - China. In: World Heal. World Health Organization, Organ (2020) <https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON229>. Accessed 13 Nov 2021
3. Elengoe A (2020) COVID-19 outbreak in Malaysia. *Osong Public Heal Res Perspect* 11:93–100
4. Gallego V, Nishiura H, Sah R, Rodriguez-Morales AJ (2020) The COVID-19 outbreak and implications for the Tokyo 2020 Summer Olympic Games. *Travel Med Infect Dis* 34:1–4
5. Parbat D, Chakraborty M (2020) A python based support vector regression model for prediction of COVID19 cases in India. *Chaos, Solitons and Fractals* 138:109942
6. Alsayed A, Sadir H, Kamil R, Sari H (2020) Prediction of epidemic peak and infected cases for COVID-19 disease in Malaysia, 2020. *Int J Environ Res Public Health* 17:1–15
7. Aziz NA, Othman J, Lugova H, Suleiman A (2020) Malaysia's approach in handling COVID-19 onslaught: Report on the Movement Control Order (MCO) and targeted screening to reduce community infection rate and impact on public health and economy. *J Infect Public Health* 13:1823–1829
8. So MKP, Tiwari A, Chu AMY et al (2020) Visualizing COVID-19 pandemic risk through network connectedness. *Int J Infect Dis* 96:558–561
9. Herlawati H (2020) COVID-19 spread pattern using support vector regression. *PIKSEL Penelit Ilmu Komputer, Syst Embed Log* 8:67–74
10. Tamhane R, Mulge S (2020) Prediction of COVID-19 outbreak using machine learning. *Int Res J Eng Technol* 7:5699–5702
11. Csardi G, Nepusz T (2006) The igraph software package for complex network research. *InterJournal Complex Syst* 1–10
12. Akoglu H (2018) User's guide to correlation coefficients. *Turkish J Emerg Med* 18:91–93
13. Hu B, Guo H, Zhou P, Shi ZL (2021) Characteristics of SARS-CoV-2 and COVID-19. *Nat Rev Microbiol* 19:141–154

14. He X, Lau EHY, Wu P et al (2020) Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med* 26:672–675
15. World Health Organization (2020) Transmission of SARS-CoV-2: Implications for infection prevention precautions. *Sci Br* 1–20
16. Cornelissen L, André E (2021) Understanding the drivers of transmission of SARS-CoV-2. *Lancet Infect Dis* 21:580–581
17. Robertson S (2021) Viral load the main driver of SARS-CoV-2 transmission. In: *Med. Life Sci.* <https://www.news-medical.net/news/20210203/Viral-load-the-main-driver-of-SARS-CoV-2-transmission.aspx>. Accessed 13 Nov 2021
18. COVID-19 in Malaysia situation report 14. In: *World Heal. World Health Organization, Organ* (2020) <https://www.who.int/malaysia/internal-publications-detail/covid-19-in-malaysia-situation-report-14>
19. Pang NTP, Kamu A, Mohd Kassim MA, Ho CM (2021) Monitoring the impact of Movement Control Order (MCO) in flattening the cumulative daily cases curve of Covid-19 in Malaysia: A generalized logistic growth modeling approach. *Infect Dis Model* 6:898–908
20. COVID-19 in Malaysia situation Report 22. In: *World Heal. World Health Organization, Organ* (2020) <https://www.who.int/malaysia/internal-publications-detail/covid-19-in-malaysia-situation-report-22>
21. Press release: Current population estimates, Malaysia, 2018-2019. In: *Dep. Stat. Department of Statistics Malaysia, Malaysia* (2019) <https://www.dosm.gov.my/v1/index.php?r=column/pdfPrev&id=aWJZRkJ4UEdKcUZpT2tVT090Snpydz09#:~:text=Malaysia's population in 2019 is,1.1 per cent in 2018>. Accessed 13 Nov 2021
22. Lim JT, Maung K, Tan ST et al (2021) Estimating direct and spill-over impacts of political elections on COVID-19 transmission using synthetic control methods. *PLoS Comput Biol* 17:1–15
23. World Health Organization (2020) COVID-19 in Malaysia situation report 17. In: *World Heal. Organ*
24. *New Straits Times* (2020) CMCO in all but 3 Peninsular states. In: *New Straits Time.* <https://www.nst.com.my/news/nation/2020/11/639067/cmco-all-3-peninsular-states>. Accessed 13 Nov 2021
25. Rodzi NH (2021) Malaysia to impose MCO for 2 weeks from Jan 13 in several states to curb Covid-19 cases: Muhyiddin. In: *The Straits Times.* <https://www.straitstimes.com/asia/se-asia/malaysia-to-impose-mco-for-2-weeks-from-jan-13-in-several-states-to-curb-covid-19-cases>. Accessed 13 Nov 2021
26. *Malaysiakini* (2021) Selangor, KL, Johor, Penang under CMCO strating March 5. In: *Malaysiakini.* <https://www.malaysiakini.com/news/564983>. Accessed 13 Nov 2021
27. Afrina Arfa (2021) How has MCO affected the Malaysian economy? In: *Taylor's Univ.* <https://university.taylors.edu.my/en/campus-life/news-and-events/news/how-has-mco-affected-the-malaysian-economy.html>. Accessed 13 Nov 2021
28. *Astro A* (2021) COVID-19: 4 varian membimbangkan termasuk dari UK, India dikesan di Malaysia. In: *Astro Awani.* <https://www.astroawani.com/berita-malaysia/covid19-4-varian-membimbangkan-termasuk-dari-uk-india-dikesan-di-malaysia-297096>. Accessed 13 Nov 2021

29. Hassandarvish M (2021) Mass screening: Here's what Selangor's mass Covid-19 testing tells us about the pandemic. In: Malay Mail. <https://malaysia.news.yahoo.com/mass-screening-selangor-mass-covid-075639686.html>. Accessed 13 Nov 2021
30. Price RHM, Graham C, Ramalingam S (2019) Association between viral seasonality and meteorological factors. *Sci Rep* 9:929
31. Wan Nik WB, Ahmad MF, Ibrahim MZ et al (2019) Wind energy potential at east coast of Peninsular Malaysia. *Int J Appl Eng Res* 4:9–16
32. Aw SB, Teh BT, Ling GHT et al (2021) The covid-19 pandemic situation in malaysia: Lessons learned from the perspective of population density. *Int J Environ Res Public Health* 18:. <https://doi.org/10.3390/ijerph18126566>
33. Supriya S (2020) Covid-19 cuts Top Glove both ways. In: Edge Malaysia. <https://www.theedgemarkets.com/article/covid19-cuts-top-glove-both-ways>. Accessed 13 Nov 2021
34. World Health Organization (2021) COVID-19 in Malaysia situation report 45. In: World Heal. Organ
35. Google LLC (2021) Google mobility report. In: Google LLC. <https://www.google.com/covid19/mobility/>. Accessed 13 Nov 2021
36. Zhao S, Lin Q, Ran J et al (2020) Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *Int J Infect Dis* 92:214–217
37. Anand R, Rodzi NH (2021) KL to place six Selangor districts under first-tier MCO, avoiding widespread Covid-19 lockdowns. In: The Straits Times. <https://www.straitstimes.com/asia/se-asia/malaysia-to-place-six-selangor-districts-under-first-tier-mco-avoids-widespread#:~:text=KUALA LUMPUR - Malaysia will place,despite surging Covid-19 figures>. Accessed 13 Nov 2021
38. Zainuddin A (2021) Klang Valley hospitals on the brink of collapse. In: CodeBlue. <https://codeblue.galencentre.org/2021/07/07/klang-valley-hospitals-on-the-brink-of-collapse/>. Accessed 13 Nov 2021
39. Johansson MA, Quandelacy TM, Kada S et al (2021) SARS-CoV-2 Transmission from People without COVID-19 Symptoms. *JAMA Netw Open* 4:1–8
40. Chicco D, Warrens MJ, Jurman G (2021) The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput Sci* 7:1–24
41. Chu J (2021) A statistical analysis of the novel coronavirus (COVID-19) in Italy and Spain

Tables

Table 1: Spearman correlation coefficient of confirmed cases vs days confirmed deaths vs days and confirmed cases vs confirmed deaths in Malaysia from July 2020 until June 2021 (365 days)

Correlation type	Data	r-value	p-value
Spearman's correlation	Confirmed cases vs days	0.902	1.868 e-134
	Confirmed deaths vs days	0.863	1.862 e-109
	Confirmed cases vs confirmed deaths	0.907	1.260 e-138

Table 2: Summary statistics for each network analysed using the NetworkAnalyzer Cytoscape plugin

Summary Statistics	Quarter 3 2020 (July – September 2020)	Quarter 4 2020 (October – December 2020)	Quarter 1 2021 (January – March 2021)	Quarter 2 2021 (April – June 2021)
Number of nodes (states) ¹	6	9	11	16
Number of edges (correlation)	3	16	42	67
Analysis time (s)	0.106	0.008	0.014	0.029

Note: ¹The node represents states in Malaysia, whereas the edge represents a correlation between the states.

Table 3: Model performance of support vector regression for confirmed cases, confirmed deaths and days

Prediction model	Data	70% training dataset ²			30% testing dataset ²		
		MSE	RMSE	R ²	MSE	RMSE	R ²
Support vector regression ¹	Confirmed cases vs days	0.154	0.393	0.846	0.151	0.388	0.855
	Confirmed deaths vs days	0.141	0.376	0.859	0.108	0.329	0.909
	Confirmed cases vs confirmed deaths	0.171	0.413	0.829	0.196	0.443	0.836

Note: ¹Execution of support vector regression was based on radial basis function (RBF) kernel with epsilon value 0.1.

²MSE = Mean squared error; RMSE = Root mean square error; and R² = Regression score.

Figures

Figure 1

The trend of COVID-19 (a) confirmed cases and (b) confirmed deaths from July 2020 until June 2021 (365 days)

Figure 2

Co-occurrence networks based on daily confirmed state cases for (a) Q3 of 2020, (b) Q4 of 2020, (c) Q1 of 2021, and (d) Q2 of 2021. The node size is proportional to the number of connections (i.e., degree), and the edge width (i.e., weight) is proportional to the r value. The edge is coloured blue and red for positive and negative correlation, respectively.

Figure 3

SVR models of (a) confirmed cases vs days, (b) confirmed deaths vs days, and (c) confirmed deaths vs confirmed cases from July 2020 until June 2021 (365 days)

Figure 4

Forecast of COVID-19 cumulative confirmed cases and deaths for the next 184 days (starting from July 2021 until December 2021)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [1stdraftlistofsupplementary.docx](#)