Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Using Network Analysis and Machine Learning to Identify Virus Spread Trends in COVID-19

Carlos Andre Reis Pinheiro *, Matthew Galati, Natalia Summerville, Mark Lambrecht

SAS Institute Inc., 100 SAS Campus Drive, Cary, NC 27513, USA

ABSTRACT

The outbreak of Coronavirus Disease 2019 (COVID-19) has infected and killed millions of people globally, resulting in a pandemic with enormous global impact. This disease affects the respiratory system, and the viral agent that causes it, SARS-CoV-2, spreads through droplets of saliva, as well as through coughing and sneezing. As an extremely transmissible viral infection, COVID-19 is causing significant damage to the economies of both developed and lower- and middle-income countries because of its direct impact on the health of citizens and the containment measures taken to curtail the virus. Methods to reduce or control the spread of the virus and protect the global population are needed to avoid further deaths, long-term health issues, and prolonged economic impact. The most effective approach to reduce viral spread and avoid a substantial collapse of the health system, in the absence of vaccines, is nonpharmaceutical interventions (NPI) such as enforcing social containment restrictions, monitoring overall population mobility, implementing widespread viral testing, and increasing hygiene measures. Our approach consists of combining network analytics with machine learning models by using a combination of anonymized health and telecommunications data to better understand the correlation between population movements and virus spread. This approach, called location network analysis (LNA), allows for accurate prediction of possible new outbreaks. It gives governments and health authorities a crucial tool that can help define more accurate public health metrics and can be used either to intensify social containment policies to avoid further spread or to ease them to reopen the economy. LNA can also help to retrospectively evaluate the effectiveness of policy responses to COVID-19.

© 2021 Elsevier Inc. All rights reserved.

1. Introduction

In December 2019, cases of pneumonia of unknown cause were reported in the city of Wuhan, in the Hubei province of China. In January 2020, a previously unknown virus was identified and later named the 2019 novel coronavirus. In February 2020, this novel coronavirus was named Coronavirus Disease 2019 (COVID-19) by the World Health Organization (WHO). The virus that causes COVID-19 is known as SARS-CoV-2. In March 2020, WHO declared the coronavirus disease 2019 a global pandemic.

A pandemic is defined as a disease occurring over a wide geographic area and infecting a high proportion of the population [1].

As of March 11, 2021, there are more than 118 million reported positive COVID-19 cases in the world and more than 2.5 million deaths. A major global effort is needed to stop the spread of the virus while vaccines are administered, and treatments are developed.

Although several research studies are under way, the details of human-to-human transmission are still not completely clear. But evidence indicates that the disease can be spread through large respiratory droplets [2] and direct or indirect contact with infected secretions [3]. Because of the modes of transmission, WHO recommends a set of preventive measures to avoid or reduce the transmission rate, including a series of hygiene-related precautions and social distancing of approximately 2 meters between people not from the same household.

Most of the recommendations are based on individual actions, such as frequent handwashing and wearing a mask. One specific measure (social distancing) is collective. Based on that series of recommendations from WHO, one of the best preventive measures to avoid mass infection is social containment. This study aims to correlate mobility behavior, or how people move between geographic areas, with virus spread, and then to prioritize geographic locations for deeper epidemiological analysis and identification of at-risk populations.

*Corresponding author.

E-mail addresses: Carlos.Pinheiro@SAS.com (C.A. Reis Pinheiro), Matthew.Galati@SAS.com (M. Galati), Natalia.Summerville@SAS.com (N. Summerville), Mark.Lambrecht@SAS.com (M. Lambrecht).

https://doi.org/10.1016/j.bdr.2021.100242
2214-5796 © 2021 Elsevier Inc. All rights reserved.
The common approach of policy enforcement during an infectious disease pandemic is mostly reactive. Public health officials track changes in active cases, identify hot spots by the number of positive cases found, and enforce containment policies primarily based on geographic proximity. This study proposes a proactive approach, based on mobility behavior in geographic areas and its correlation with virus spread over time. Population movements across geographic regions are evaluated over time, and a series of correlation analyses are performed to identify locations that play a key role in the flow of people and how these flows affect the way the virus spreads over time and across different locations.

In the past, some researchers and practitioners have combined network analytics with epidemiological studies to model mobility and predict disease spread in humans. Pierson et al. integrated mobility networks with susceptible, exposed, infected, and recovered (SEIR) models to fit the trajectory of infection [4]. Shah et al. analyzed interactions between livestock herds during simulated disease outbreaks [5], whereas Alvarez et al. applied this approach specifically to dairy herds [6]. Morris focused on data collection and analysis by using a network formulation of infectious diseases [7].

Our study uses a combination of telecommunications and public health data. Telecommunications data are primarily used in the form of geolocation positioning of anonymized subscribers over time. These data are turned into mobility information. Anonymized subscribers’ positionings are computed over time to create vectors of movement between distinct geographic locations. This information can be aggregated into different subsets, such as neighborhoods, municipalities, counties, or cities. In our study, the public health data are in the form of the number of new positive COVID-19 cases, per time (daily), and per geographic location (municipalities). The mobility data are aggregated daily at the municipality level to match the granularity of the available public health data.

A series of network analytics were performed on the combined data so we could better understand population movement behaviors and its correlation with the spread of the virus over time and by geography. Understanding this mobility behavior and how it correlates with the virus spread enables public health authorities to make more proactive decisions about social containment policies.

2. Data and methodology

In this section, we describe the data used in the study and the construction of a series of networks that represent population movement. We briefly describe the analytical methodologies we used to understand the structure of these networks over time. In addition, we describe the methods used for ranking key locations and correlating these with the change in positive COVID-19 cases over time.

2.1. Data description

The movements of telecommunication service subscribers were identified by the positioning of distinct cell towers within a one-hour interval. All movements were then aggregated by origin and destination cell towers. The Department of Health (DOH) in the Philippines officially publishes the number of positive cases daily by municipality. The mobility data provided by the telecommunications company are then aggregated by municipality every day to match with the positive cases provided by DOH. This study considers 1,551 municipalities and seven months of mobility data.

Fig. 1 shows the number of movements by month, varying from 3.4 to 4.6 million movements, and the number of positive COVID-19 cases, varying from 618 to 10,659 occurrences.

2.2. Methodology

Network analysis is the study of connected data. Every industry, in every domain, has information that can be analyzed in terms of linked data. Network analysis can be applied to understand the spread of influence through social networks [9]. Typical examples are churn and product adoption in industries such as telecommunications and entertainment, service consumption in retail, fraud in insurance, and money laundering in banking, among others.

Location network analysis (LNA) focuses on analyzing networks by using georeferencing data over time. This method evaluates people’s movements over time across different geographic areas. Our study applied LNA to correlate the overall population movements with the spread of the coronavirus over time by considering multiple geographic areas. The telecommunications data used in this study were provided by a major telecommunications carrier in the Philippines, and the health data were provided by local Philippine health authorities. The data were anonymized and aggregated.

There is no individual subscriber information or individual positive COVID-19 case information in the data that we analyzed in this study.

The main goal of the study is to predict and identify specific geographic areas to target for social containment policies, either to better define shelter-in-place measures or to gradually evaluate which locations are ready for containment measures to be relaxed.

A key aspect of the spatiotemporal analysis is subscriber geolocation over time. This information is anonymously collected to transform the sequence of subscribers’ geolocations into subscribers’ vectors of movements, as shown in Fig. 2.

The common scenario of spatiotemporal analysis is described by Fig. 2. An anonymous subscriber is at geolocation g1 at time t1. At time t2, that subscriber is at geolocation g2. We can assume that a movement by that subscriber took place from geolocation g1 at time t1 to geolocation g2 at time t2. Because subscriber data are anonymized and aggregated based on specific geolocations (municipalities) to comply with data privacy regulations, this subscriber’s movements are computed in terms of the sum of all individual subscribers’ movements in a one-hour interval during the day. These data are then aggregated at the daily level.

In spatiotemporal analysis, geolocations can have different levels of granularity—for example, a specific coordinate, a polygon, a neighborhood, an administrative region, etc. In our study, each geolocation is represented by the latitude and longitude of one of the 1,551 municipalities in the Philippines.

Every geolocation that we analyzed becomes a node in the mobility network. This network is defined as a directed graph because the direction of subscribers’ movements is relevant to the analysis. Every aggregated vector of movement, or a set of people’s displacements, becomes a link in the mobility network. All correlation analysis and spatiotemporal evaluation are performed on the basis of that mobility network graph, which considers a series of networks over time.

The telecommunications data include, in addition to the subscribers’ locations over time, the penetration of cell phones in each geographic location, the market share of the phone service provider in that region, and the total population of all geographic regions that we analyzed. Based on that additional information, we can extrapolate from subscribers’ movements to population movements. The extrapolation is defined according to the following equation:

\[ p_i^j = v_i^j \times \frac{p_l}{m_l} \]

where \( v_i^j \) is the number of subscribers who move from location \( i \) at time \( t - 1 \) to location \( j \) at time \( t \); \( p_l \) is the total population of
K-core decomposition K-core decomposition is an alternative method of community detection to find cohesive subgroups within a network. A subgraph is a k-core if every node in the subgraph has degree (connectivity) greater than or equal to k and if that subgraph is the maximum subgraph with this property [11].

K-cores have numerous practical applications—for example, describing social networks, visualizing complex graphs, determining roles in biological protein networks, and studying viral spread in epidemiology [12].

2.2.2. Centrality metrics

Network centrality metrics help to rank locations according to their mobility traffic [13]. The following subsections briefly describe the relevant centrality metrics we calculated in this study [14]. For details about how each network centrality metric is calculated, see Newman [15].

Degree centrality Degree centrality identifies the amount of mobility traffic directly into or out of a particular node (geographic location).

Influence centrality Influence centrality is a generalization of degree centrality that considers two levels of mobility. The first-order influence metric is like degree centrality but is normalized by the population of the origin location. The second-order influence metric for a particular location L considers the number of people flowing into and out of the locations directly connected to L. In simple terms, the second-order influence metric is the sum of the first-order influence metrics across locations connected to location L.

In this study, the first- and second-order influence centrality presented the highest correlation with virus spread. Locations with high influence centrality are the ones that have a higher probability of spreading the virus throughout multiple geographic regions.

Closeness centrality Closeness centrality computes the average shortest distance between one specific location and all other locations within the network. It shows key locations that can influence the speed of transmission of the virus over time across multiple geographic regions.

Betweenness centrality Betweenness centrality counts the number of times a specific location is on a shortest path between two other locations. It shows which locations control the population traffic flow across multiple geographic regions, thus indirectly affecting viral spread over time.

Locations with high betweenness centrality are often called gatekeepers. Gatekeeper locations do not necessarily have many

location i; and m_j is the mobile penetration of the service provider at location i. Mobile penetration is the number of mobile phone numbers activated by the provider in a specific location divided by the total population in that location.

From this point onward, the collected data are processed daily to create a time series of network data, which contains population movements over time across geographic locations. Based on this series of network data, we used a set of network algorithms to compute various centrality measures and to extract relevant topological structures as described in the next section.

2.2.1. Topological extraction

Topological extraction is a standard technique in network analysis. Two specific methods that present a high correlation with the spread of the virus over time are community detection and k-core decomposition.

Community detection Community detection partitions a network into groups of nodes (locations), where the links (number of people flowing in and out between the locations) within the communities are more densely connected than the links between communities.

In this study we used the Louvain algorithm, which partitions nodes into communities by heuristically optimizing modularity [10].

Fig. 1. Number of movements and positive COVID-19 cases by month.

Fig. 2. The map shows an example of an anonymized subscriber’s geopositioning over time and the person’s vector of movements.

Fig. 2.
positive cases, but they move a great number of people between geographic areas. The number of gatekeeper locations is relatively low, but this is a key predictive metric because these locations move a substantial number of people across different regions and therefore cover a larger percentage of the country.

Local clustering coefficient centrality Local clustering coefficient centrality computes the connectivity of the neighborhood of a specific location, or how the locations directly connected to that specific location are connected to each other. It correlates with how easily the virus can spread locally.

Hub and authority centrality Hub centrality computes a specific location’s importance on the basis of the importance of the locations connected to it and the number of people traveling from it to the other connected locations. Locations with high hub centrality have a high amount of outgoing traffic and can indicate locations that the virus can move more easily and more quickly spread from.

Authority centrality computes a specific location’s importance on the basis of the importance of the locations connected to it and the number of people traveling to it from other connected locations. Locations with high authority centrality have a relatively high amount of incoming traffic, making them particularly susceptible to virus spread.

PageRank centrality PageRank centrality creates a rank of the nodes that is based on the probabilities associated with their being reached by other nodes through the existing links.

2.2.3. Pearson correlation

Pearson product-moment correlation is a parametric measure of a linear relationship between two variables [16]. Define the set \( \Omega \) as the following network metrics: community, k-core, degree, influence, closeness, betweenness, local clustering coefficient, hub, authority, and PageRank. For each network metric \( C \in \Omega \), at each geolocation \( g \), and time period \( t \), we calculate two Pearson correlation coefficients.

The first correlation coefficient \( (\rho_{CG}) \) considers, for each geolocation \( g \), how the change in network metric \( C \) relates to the change in the sum of the number of positive cases across locations connected from \( g \) between time \( t \) and \( t-1 \). For example, \( g \) might change from high-velocity spread at time \( t-1 \) to medium-velocity spread at time \( t \), based on closeness centrality. This change in the network metric is then correlated with the change in the number of positive cases at connected locations at time \( t \). This coefficient focuses on the direct destinations of geolocation \( g \) and will be used to understand which network metrics are more indicative of viral spread from each location.

The second correlation coefficient \( (\mu_{CG}) \) considers, for each geolocation \( g \), how the change in network metric \( C \) relates to the change in the sum of the number of positive cases across locations connected to \( g \) between time \( t \) and \( t-1 \). This coefficient focuses on the origins directly connected to geolocation \( g \) and will be used to categorize risk at each location.

3. Results discussion

In this section we describe the results of our study. To visually show the results found in the study, we have created various dynamic maps to describe how LNA identifies key locations that can affect viral spread throughout geographic areas over time.

3.1. Topological extraction

Topological extraction is a standard practice for understanding how nodes in a network relate to each other. In this study, nodes are geolocations and the links between them are weighted by population movements. In the following section, we look at communities and cores to better understand how topological patterns in the network relate to potential viral spread.

3.1.1. Community detection

The community detection algorithm groups together locations according to the density of the volume of movements among them. Fig. 3 shows the clustering of locations defined by community detection based on the number of people flowing between them. Most of the communities identified are geographically close to each other. This result might indicate that most people tend to travel to nearby locations or that eventually they need to pass through other locations to reach their final destinations.

Notice that we are not analyzing commuting patterns, but instead we are accounting only for the origin and destination locations of travel. We are analyzing all movements performed by the population over time. Long commutes can be represented by a series of small trips. In the study of population mobility, we are interested in all these movements, because they can all contribute to the spread of the virus along the way, throughout multiple geographic locations.

In terms of virus spread, information about locations within the same community can be quite relevant. If one location turns out to be a hot spot, other locations in the same community are at a higher risk of viral spread in the near future.

3.1.2. K-core decomposition

Using k-core decomposition, we cluster locations on the basis of similar levels of interconnectivity across regions [17]. We are interested in the most cohesive cores within the network. The most cohesive cores present a high level of interconnectivity between the locations within the core.

Locations within a core are not necessarily close geographically. Instead, cohesive cores show how the locations are highly connected to each other with respect to population flow [18]. One of the most important outcomes of k-core decomposition is the high correlation with wider spread of the virus. When cores are identified, specifically the more cohesive ones, social containment policies can be made more proactive in identifying groups of locations that should be quarantined together, rather than simply being based on geographic proximity to the current hot spots. This explains the spread of the virus over time throughout locations geographically distant from each other but close in terms of interconnectivity, as shown in Fig. 4.

3.2. Centrality metrics

Centrality metrics can play a key role in explaining the spread of the virus. For example, closeness centrality can identify key locations that, according to the flow of people, contribute most to the velocity of virus spread across different geographic locations. Betweenness centrality can identify locations that serve as gatekeepers, which are locations that do not necessarily have a high number of positive cases but serve as bridges connecting multiple geographic locations, leading to wider virus spread.

To explore this further, we created a series of networks, based on the population movements between geographic regions and the health information about positive cases in these regions over time. This set of network centralities and topologies is computed for each individual network, which represents one day of analysis in our time frame.

As described earlier, we calculate, for each network metric \( C \) and each location \( g \), the Pearson correlation coefficient \( (\rho_{CG}) \) of the change in metric \( C \) to the change in the number of positive cases
Fig. 3. Communities consist of locations in geographic proximity. When the virus hits one location, the other locations within the same community are at higher risk.

Fig. 4. Core decomposition group locations based on the level of interconnectivity (flows of people) between them. The maps side by side show the correlation between virus spread and most cohesive core. On the left, locations in dark blue are in the most cohesive core. On the right, these locations are the ones where the number of new positive COVID-19 cases (shades of red) increased. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

between time \( t - 1 \) and \( t \). These coefficients provide evidence supporting the hypothesis that the way the network evolves over time, in terms of mobility pattern, is directly correlated with the way the number of positive cases changes over time across geographic areas.

Next, we derive a normalized network metric, \( \bar{\rho}_{g/\ell}^C \), for each network metric, at a particular location and at a specific time. Finally, a combined network metric is calculated for each location and time period that considers all the standardized network metrics weighted by their respective correlation coefficients, as shown in the following equation:

\[
W_{g/\ell} = \sum_{C \in \Omega} \rho_{g/\ell}^C \bar{\rho}_{g/\ell}^C
\]

In Fig. 5, the map on the left highlights (in yellow) locations with a particularly high combined network metric \( W \) at some speci-
fied time $t$. The arrows between locations represent large volumes
of population movement. In this case, the larger volumes repre-
sent major public transportation systems, going from the central
area of the country southward. This public transportation system
is an important commuting line between different geographic re-
genous within the country, moving a substantial number of people
over time. This representative number of people traveling across
those geographic locations increases the likelihood that the virus
will spread throughout a wider area, particularly to the locations
along the transportation line. As time goes by, an increase in the
number of positive cases can be seen in the map on the right
(darker shades of red) along the regions connected to the key loca-
tions (in yellow in the map at left).

The shades of red get darker and wider at time $t + 1$ and $t + 2$
(one and two weeks later). The positive cases lagged the key loca-
tions, indicating that the network metrics and topologies provide
good predictive behavior of future spread.

3.3. Calculating risk level

For each geolocation $g$, by using the correlation ($\rho_{gt}^C$) between
network metrics and the number of positive cases at destination
locations of $g$, we have defined an overall metric, $W_{gt}$, for iden-
tifying key locations that correlate with virus spread. In a similar
manner, we can use the correlation ($\mu_{gt}$) between network met-
rics and the number of positive cases at locations for the origin
locations of $g$ to categorize risk, as follows:

$$R_{gt} = \sum_{C \in \Omega} \mu_{gt}^C$$

Risk level is binned into five groups, using a standard bucket binn-
ing algorithm [19]. The first bin has about 1% of the locations;
it is considered high risk. All locations in this bin have the highest
risk for the number of positive cases to increase over time because
of their incoming connections. The second bin has about 3%–4% of
the locations and is considered medium-high risk. The third bin has
about 5% of the locations and is considered medium risk. The fourth
bin has about 40% of the locations and is considered medium-low
risk. Finally, the fifth bin has about 50% of the locations and is con-
sidered low risk. All groups are shown on the map in Fig. 6. The risk
level varies from light shades of green for low risk to dark shades
of red for high risk.

3.4. Using machine learning models to predict new outbreaks

As previously described, the chosen network metrics correlate
mobility behavior with virus spread over time. Because of this
clear correlation, we have used this set of network measures as
dependent variables in a set of supervised machine learning clas-
sification models. In addition to the original variables (network cen-
tralities and extracted topologies), we have computed a new
set of additional variables to better describe how the network evolves
over time and affects the number of positive cases across multiple
locations. Most of the derived variables are based on ratios of net-
work metrics over time. For example, for some network metrics, we
considered the first and last day of the week, the last day of
the week, and the average value for the week by using ratios of
maximum and minimum values, ranges, and standard deviations,
among others. The main idea of these ratios is to describe relevant
trends in the network’s evolution in terms of mobility behavior
over time and changes in the number of positive cases for multi-
ple geographic locations.

The set of supervised machine learning models that we trained
includes logistic regression, decision tree, random forest, gradient
boosting, neural network, and support vector machine models. For
each (weekly) time period $w$, all models were trained using the
network metrics, considering the previous 4 weeks ($w - 4, ..., w -
1$). The classification model’s binary target value is defined as

$$T = \begin{cases} 1, & \text{if } C_{gw} > C_{gw-1} \\ 0, & \text{otherwise} \end{cases}$$

where $C_{gw-1}$ is the number of positive cases in week $w - 1$ for
a location $g$, and $C_{gw}$ is the number of positive cases in week $w$
for the same location $g$. The binary target $T$ for the classification
model is 1 if the number of positive cases in the current week is
greater than the number of positive cases in the previous week,
and 0 otherwise. This model is trained to classify the following
week at time $w + 1$.

Assume that the current week $w$ is week 24, starting June 7
and ending June 13. Also assume that on June 14 (week 25), the
number of positive cases in week 24 would already be compiled,
allowing us to start training all models. The models will classify as
1 all locations where the number of positive cases at the end of
week 25 (starting June 14 and ending June 20) is greater than the
number of positive cases at the end of week 24. During training,
when the current week is over, the previous 5 weeks of data are
fetched—the full current week with the number of positive cases, plus the previous 4 weeks with the number of positive cases and the network metrics.

The classification model predicts the target as a decision (1 or 0), foreseeing whether or not a location g will have an increase in the number of positive cases. The decision is made on the basis of the predictive probability, which defines the likelihood of a location to experience, or not experience, an increase in the number of positive cases. If the predictive probability is greater than 0.5, the decision is made to be a 1 (that is, an increase in the number of positive cases is observed). If the predictive probability is less than or equal to 0.5, the decision is made to be a 0 (that is, an increase in the number of positive cases is not observed).

Two approaches were used to assess and compare the accuracy of all supervised models. The first method compares all models by using the receiver operating characteristic (ROC) curve. This method shows the ability of a binary classifier to discriminate the target as the predictive probability varies across the population. The second method compares all models by using the lift fit statistics. This approach measures all models’ ability to classify the cases against a random response, or the average response rate.

Fig. 7 shows the ROC curve for all the supervised models that were trained.

Fig. 8 shows the lift chart for the same models.

As can be seen in Figs. 7 and 8, the gradient boosting model was slightly better than the other models in overall accuracy. The ROC curve in Fig. 7 shows the gradient boosting model performing slightly better than the random forest and neural network models, better than the decision tree model, and significantly better than the logistic regression and support vector machine models. Because the performance is similar for the gradient boosting, random forest, and neural network models, we calculated the area under the curve (AUC) to select the best model based on the ROC curve. A similar interpretation about the comparison of these models can be extracted from the lift chart shown in Fig. 8. The lift chart shows the models’ performance across the population (geographic locations), ranked by predictive probability. We can see that gradient

boosting performs better than the other models, particularly considering the top 50% of the population. Based on the ROC curve, we can see that the gradient boosting model shows the greatest accuracy, considering the true positive and true negative rates (correct classifications for locations experiencing an increase in the number
Fig. 8. Lift chart considering logistic regression, decision tree, random forest, gradient boosting, neural network, and support vector machine models.

of positive cases and locations not experiencing such an increase, respectively. Based on the lift chart, we can see that gradient boosting shows the best performance, particularly considering the higher predictive probability values (the likelihood of an increase in the number of positive cases). Based on both methods, we selected the gradient boosting model as the best overall model and deployed it in production to classify locations that potentially experience an increase in the number of positive cases in subsequent weeks.

Local authorities can use the predictive probability to determine the level of risk associated with each geographic location to help establish effective social containment measures. The higher the predictive probability, the more likely the location will experience an increase in the number of positive cases; this suggests a need for stricter social containment. For example, locations $g_1$ and $g_2$ can both present a predicted target as 1, indicating that they are likely to experience an increase in the number of positive cases. However, location $g_1$ might have a predictive probability of 0.9, and location $g_2$ might have a predictive probability of 0.6. This means location $g_1$ is more likely than location $g_2$ to experience an increase in the number of positive cases in the following week at time $w+1$.

Fig. 9 shows (on the left) the results of the gradient boosting model in classifying the locations that are predicted to experience an increase in the number of positive cases the following week. The model's performance on average is about 92%–98%, with an overall sensitivity around 90%. The overall accuracy measures the true positive rates as well as the true negative, false positive, and false negative rates. The sensitivity measures only the true positive rates, which can be more meaningful for local authorities in determining what locations to apply social containment policies to.

Network centrality measures and topologies represent 18 original features of the supervised machine learning model. An instrumental approach in building the machine learning model is to extract new features from the original ones, considering how they evolve over time—for example, the average closeness centrality over the training period (4 weeks), and some ratios of the average closeness centrality based on different time frames, such as the value for the last 2 weeks over the value for the last 4 weeks. In total, 95 features were generated to feed the supervised machine models. Five features stood out from the input set as relevant predictors for the models:

- the average closeness centrality for the period
- the ratio of the closeness and betweenness centrality from the latest week to the average for the period
- the frequency of being in the most cohesive k-core across the period
- the ratio of the maximum influence centrality from the latest 2 weeks to the average of the previous 2 weeks
- the ratio of the latest authority centrality to the average for the period

4. Conclusions

A spatiotemporal analysis of mobility behavior can reveal important trends about the spread of a virus throughout geographic regions over time. Location network analysis enables health authorities to understand the impact that population movements have on the spread of the virus and ultimately to predict possible new outbreaks in specific geographic locations. The correlation between mobility behavior and virus spread allows local authorities to identify groups of locations to be put in social containment together, as well as the level of severity required. Location network analysis provides accurate information for government agencies about the pattern of the virus spread and its common paths, enabling them to make good decisions about implementing shelter-in-place policies, planning public transportation services, and most important, allocating medical resources to locations where the mobility behavior indicates a substantial increase in the number of positive cases. Mobility behavior analysis can also be used to identify geographic locations at lower risk of contagion so that authorities can start easing social distancing restrictions and allow economic activity to resume.
This study has shown that health authorities can use LNA outcomes to control the virus spread significantly and proactively, as locations are closely monitored by using anonymized telecommunication data and infection information over time. These techniques can also be used to retrospectively study government measures and evaluate their impact on population movements and virus spread. Monitoring the level of population movements over time allows local authorities to better define actions to control virus spread, identifying the level of risk in opening or closing specific geographic locations throughout the country.

The analysis of mobility behavior can be used for any type of infectious disease, evaluating how population movements over time can affect virus spread in different geographic locations. This methodology can be an important tool for health and local authorities to use in defining more accurate social measures to contain the spread of any infectious disease and globally reduce transmission rates among multiple regions.

Declaration of competing interest

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

Acknowledgements

The technology used in this study is provided by SAS Institute Inc. SAS® Visual Data Mining and Machine Learning software was used for the data preparation tasks, network analytics, and supervised machine learning models. The dynamic geographic maps were generated using SAS® Viya®, an in-memory distributed computing engine tailored for analytical tasks.

Uncited references

[8]

References

[1] C. Wang, P.W. Horby, F.G. Hayden, G.F. Gao, A novel coronavirus outbreak of global health concern, Lancet (24 January 2020).

[2] R. Zhang, Y. Li, A.L. Zhang, Y. Wang, M.J. Molina, Identifying airborne transmission as the dominant route for the spread of COVID-19, in: Proceedings of the National Academy of Sciences, 2020.

[3] L.T. Phan, T.V. Nguyen, Q.C. Luong, H. Nguyen, T.V. Nguyen, Importation and human-to-human transmission of a novel coronavirus in Vietnam, N. Engl. J. Med. 382 (27 February 2020) 872–874.

[4] E. Pierson, S. Chang, P.W. Koh, G. Jaline, B. Redbird, D. Grusky, J. Leskovc, Mobility model outcomes of COVID-19 explain inequities and inform reopening, Nature (2021) 82–87.

[5] N. Shah, M. Malensek, H. Shah, S. Pallickara, S. Lee, Scalable network analytics for characterization of outbreak influence in voluminous epidemiology datasets, Concurr. Comput., Pract. Exp. 31 (7) (2019) e4998.

[6] L. Alâvarez, C. Webb, M. Holmes, A novel field-based approach to validate the use of network models for disease spread between dairy herds, Epidemiol. Infect. 139 (12) (2011) 1863–1874.

[7] M. Morris, Data driven network models for the spread of infectious disease, in: Epidemic Models: Their Structure and Relation to Data, Newton Institute, Cambridge, 1995.

[8] J.J. Nalluri, Network Analytics for the miRNA Regulome and miRNA-Disease Interactions, Ph.D. Diss., Virginia Commonwealth University, 2017.

[9] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the spread of influence through a social network, Theory Comput. (2015) 105–147.

[10] A. Lancichinetti, S. Fortunato, Community detection algorithms: a comparative analysis, Phys. Rev. E 80 (5) (November 2009) 056117.

[11] V. Batagelj, M. Zaversnik, An O(m) algorithm for cores decomposition of networks, Comput. Res. Repos., arXiv.org, arXiv/0310049v1 [cs.DS], 25 October 2003.

[12] M. Eidsaa, E. Almaas, 5-core network decomposition: a generalization of k-core analysis to weighted networks, Phys. Rev. E 88 (30 December 2013) 062819.

[13] S. Seidman, Network structure and minimum degree, Soc. Netw. (1983) 267–287.

[14] L.C. Freeman, Centrality in social networks conceptual clarification, Soc. Netw. 1 (1978) 215–239.

[15] M.J.F. Newman, Networks: An Introduction, Oxford University Press, Oxford, 2010.

[16] P.C. Coefficient, in: W. Kirch (Ed.), Encyclopedia of Public Health, Springer, Dordrecht, 2008.

[17] Y. Liu, M. Tang, T. Zhou, Y. Do, Core-like groups result in invalidation of identifying super-spreaders by k-shell decomposition, Sci. Rep. 5 (2015).

[18] D. Chen, L. Lu, S. Ming-Sheng, Y.-C. Zhang, T. Zhou, Identifying influential nodes in complex networks, Phys. A, Stat. Mech. Appl. (2012) 1777–1787.

[19] B. Hemada, K.S.V. Lakshmi, A study on discretization techniques, Int. J. Eng. Res. Technol. 2 (2013).