Bayesian network-based spatial predictive modelling reveals COVID-19 transmission dynamics in Eswatini

Wisdom M. D. Dlamini • Sabelo P. Simelane • Nhlanhla M. Nhlabatsi

Received: 19 August 2021 / Revised: 24 October 2021 / Accepted: 26 October 2021 / Published online: 19 November 2021 © Korean Spatial Information Society 2021

Abstract The first case of COVID-19 in Eswatini was first reported in March 2020, posing an unprecedented challenge to the country’s health and socio-economic systems. Using geographic information system (GIS) data comprising 15 socioeconomic, demographic and environmental variables, we model the spatial variability of COVID-19 transmission risk based on case data for the period under strict lockdown (up to 8th May 2020) and after the lockdown regulations were gradually eased (up to 30th June 2020). We implemented and tested 13 spatial data-driven Bayesian network (BN) learning algorithms to examine the factors that determine the spatial distribution of COVID-19 transmission risk. All the BN models performed very well in predicting the COVID-19 cases as evidenced by low log loss (0.705–0.683) and high recall values (0.821–0.836). The tree-augmented naïve (TAN) model outperformed all other BN learning algorithms. The proximity to major health facilities, churches, shopping centres and supermarkets as well as average annual traffic density were the strongest predictors of transmission risk during strict lockdown. After gradual relaxation of the lockdown, the proportion of the youth (15–40 years old) in an area became the strongest predictor of COVID-19 transmission in addition to the proximity to areas where people congregate, excluding churches. The study provides useful insights on the spatio-temporal dynamics of COVID-19 transmission drivers thereby aiding the design of geographically-targeted interventions. The findings also point to the robustness of BN models in spatial predictive modelling and graphically explaining spatial phenomena under uncertainty and with limited data.

Keywords COVID-19 • Bayesian network • Eswatini • Spatial predictive modelling • Risk mapping

1 Introduction

Since its reported first detection in December 2019, the highly contagious coronavirus disease 2019 (COVID-19) has spread throughout the world. As of 10 June 2020, over 7 million people were confirmed to have been infected coupled with over 400,000 fatalities. The disease has also since spread to many African countries [1, 2]. The spread of the disease to developing countries presents a unique challenge considering the resource requirements for disease control and prevention. Since the disease is novel, there is still a lot of uncertainty on the drivers of its spread. Other observed challenges include the limited testing equipment, low detection efficiency, and the delay in policy-making and implementation of proper prevention and control measures during the early stage of the epidemic [2]. Despite this uncertainty, clinical and epidemiological evidence suggests human-to-human transmission of the novel coronavirus is the most dominant form of transmission [3, 4]. In the absence of an approved vaccine, non-pharmaceutical interventions are crucial in containing the

Wisdom M. D. Dlamini mwdlamini@gmail.com
Sabelo P. Simelane sabeloaussy@gmail.com
Nhlanhla M. Nhlabatsi ntini0064@gmail.com

1 Department of Geography, Environmental Science and Planning, University of Eswatini, Kwaluseni, Eswatini
2 Central Statistical Office (CSO), Mbabane, Eswatini
3 Epidemiology and Disease Control Unit (EDCU), Ministry of Health, Mbabane, Eswatini
spread of the virus [5]. Such interventions require an understanding of the dynamics of geographic spread of the disease and the underlying factors. Early forecasting of the potential outbreaks is key in making effective infection prevention and control strategies [6]. Combating and proper management of COVID-19 spread, therefore requires an understanding of the key determinants of its morbidity [7].

Factors such as human mobility, economic activities and other social factors have been observed to play a key role in the transmission of COVID-19, although their impact varies from one area to another [8, 9]. Many of these factors have a spatial dimension which makes the understanding of COVID-19 a geographical problem [10, 11]. Therefore, analysis and modelling of COVID-19 requires the use of spatial analytical techniques. Geographic information systems (GIS) are increasingly being used to provide insight on the locations and distribution patterns of COVID-19 [12, 13]. Moreover, the identification of areas where COVID-19 is likely to spread and the influencing factors is equally important for health authorities to make targeted prevention and control measures for specific areas.

A review by Fatima et al. [14] found that most studies using GIS in the analysis of COVID-19 were centred on Asia and the Americas and notably very few from Africa. Hence, there is a need for more comparable spatial studies using geographically fine-scaled data in other areas of the world, particularly in Africa. Even within the countries where studies have been conducted, COVID-19 has been observed to exhibit within-country variations driven by various influencing factors. This, therefore, points to the need for locally specific interventions, in accordance with local conditions, to increase the effectiveness of response strategies. Understanding the spatiotemporal incidences of COVID-19 at the national level is extremely important to elucidate how the pandemic progresses over space and time. For a small country such as Eswatini, this is highly dependent on the availability of geographically fine-scaled data to reveal the dynamics of the disease’ spread. The spatial epidemiological approach may also enable early assessments of local variants and of the effectiveness of COVID-19 vaccines [14].

The application of spatial analysis tools and techniques provides a wide range of benefits. These include more intuitive visualization, tracking of confirmed and reported case numbers, contact tracing, spread direction, and also, to identify the hotspots to limit the dispersion and community spread [12]. The spatial analysis of public-health related issues has been undertaken through numerous studies in the past, some predating the modern GIS era which was born in the mid-1960s. GIS has recently been extensively used in analyzing, visualizing, and detecting patterns of disease and recent reviews (e.g. [14, 15]) indicate that it is increasingly being used in many epidemiological studies.

Mapping disease outbreaks can also assist in understanding their key determinants [7]. Statistical analysis of spatial patterns is more effective as it involves assessing statistical significance of clustering or dispersion and modelling spatial relationships. As a result, the key factors or drivers contributing to the diseases incidence can be determined. Many of the drivers are interrelated and, although significant progress has been made in developing models that use traditional statistical approaches to understand the relationships between predictor variables and COVID-19 morbidity, most of the traditional models assume that the underlying relationships are in a steady state. This assumption might not be true in the case of COVID-19 transmission which has been observed to be complex thereby having a major effect on the disease dynamics [16]. Further, it is possible that the transmission is driven by unobserved factors, i.e., variables that we do not have data on. Thus, it is recommended that COVID-19 models develop richer non-mechanistic appreciation of variable interactions across space and time due to changing conditions at different geographical scales. Therefore, to proceed in a pragmatic manner to better understand COVID-19 transmission from environmental, demographic and socio-economic perspectives, we adopt a relatively novel predictive and non-mechanistic technique. However, many of the commonly used techniques used in epidemiology do not easily allow for the modelling of complex, non-linear data that is characteristic of most geographic and epidemiological data.

Bayesian networks (BNs) [17, 18] are alternative machine learning techniques that are capable of inferring functional network structures, capturing nonlinear, dynamic and arbitrary combinatorial dependency relationships [19, 20]. BNs are directed acyclic graphical models that are used for modelling uncertain relationships amongst variables in a complex system and reasoning under uncertainty, where nodes represent random variables (discrete and/or continuous) and arcs (or links) represent direct causal and informative connections between them. The advantage of BNs is that few assumptions are made about the data and, in addition, complex spatially varying interactions can be recovered from data. This is highly useful in epidemiological analyses where complex variable interactions change in time due to changing socio-economic conditions and demographic patterns. The capability of BNs to predict from incomplete observations, through the use of Markov blankets, enables predictions to be made at an early stage of an event or with fewer observations [20], employing only the information that is present at any given time such as early COVID-19 detections. The predictions can subsequently be updated as new information, such as
new COVID-19 case data, becomes available. In addition to helping us learn about and visualize causal relationships, the BN enables us to understand a problem area and to predict the consequences of any intervention. The BN also provides an efficient approach for avoiding the overfitting of data [19].

This study, therefore, utilizes BNs to model the potential spread of the disease using data for the first 300 cases (up to 6 June 2020), as well as available data on potential influencing factors in Eswatini (formerly Swaziland). BNs are probabilistic graphical models that also incorporate the relationships between variables and commonly used in creating decision support systems and developing models under uncertainty. BNs have the advantage of being able to represent uncertainty and handle missing data, in addition to the ability to represent nonlinear relations, among others [21]. Hence, BNs are gaining popularity for use in developing predictive infection risk maps [22, 23].

In this analysis, we use BNs linked to GIS data to produce a predictive risk map of COVID-19 infection in Eswatini. This study differs from previous studies because it includes more complex relationships between bio-physical, demographic and socio-economic components of COVID-19 spatial epidemiology. Thus, the study graphically and spatially illustrates the complexity of the disease spread. Moreover, this study emphasizes the advantage of BN in overcoming limited data, which is a common problem in many developing countries such as Eswatini. Through the development and testing of the BN in our case study, we are able to identify important gaps in data and beyond that should be addressed to improve model accuracy and facilitate further development of spatio-temporal models. We specifically aim to answer the research question: Can BNs be used to conceptualize the complex interaction of factors related to COVID-19 transmission in order to assist decision making? We surmise that BNs are particularly suited to investigate the influence of geographic factors on the occurrence of COVID-19 incidents in situations where data is scarce. We also examine different BN structure learning algorithms to examine how these perform in predicting the risk of COVID-19 spread in Eswatini.

2 Materials and methods

2.1 Data collection and integration

In this study, the case data was obtained from the Ministry of Health’s Epidemiology and Disease Control Unit (EDCU) wherein each confirmed case is reported with, among other things, location (name of the community). The first case of COVID-19 in Eswatini was recorded on the 11th March 2020. The COVID-19 data for Eswatini, derived from testing and contact tracing, indicates that while only the initial 10 cases were imported from Italy, the United States of America and South Africa, the rest of the cases were local transmission. Subsequently, as at end of June 2020, a total of 690 cases had been confirmed together with 3 fatalities. The locality information was converted into a point layer by geocoding it using enumeration area information and point of interest data for Eswatini (see Fig. 1).

COVID-19 transmission factors were determined using existing domain knowledge (literature review) as well as best available data. Since the transmission of the disease is through person-to-person contact, we considered several factors related to human demography, behaviour, socio-economic status and human interaction. Subsequently, fifteen (15) variables (Table 1) were collated at census enumeration area level for the analysis. Distances were measures using the Euclidean distance. Eswatini has a total of 2326 enumeration areas of varying sizes. Example of the maps of the variables used are shown in Fig. 2.

![Fig. 1 Location of cases reported as of the 30th June 2020 in Eswatini. Also shown are the major roads, cities and major towns.](image)
2.2 Data pre-processing

Some of the derived explanatory datasets were collated as 1 km² raster data within the GIS for spatial analysis. Furthermore, the mean of each of the collated raster datasets was estimated for each enumeration area. For the BN models, the continuous variables were discretized into a finite number of discrete values or intervals using the minimal description length (MDL) criteria \[24\]. The MDL algorithm dynamically searches for the number of states and cut-off points that minimize the class variable entropy given each predictor by selecting bin boundaries based on the minimization of the class information entropy. This is the advantage of BNs in that the data need not be transformed as BNs can handle data of varying scales. The requirement for complete observations also does not hold for BNs which can handle incomplete and/or missing data through applying the Markov blanket over every variable \[25\].

2.3 Bayesian network modelling

A BN is essentially a graphical representation of a probability distribution over a set of variables \(X = \{X_1, X_2, \ldots, X_n\}\), \(n \geq 1\). The BN generally consists of two parts, \(B = (G, \Theta)\), where \(G\) is a directed network structure in the form of a directed acyclic graph (DAG), and \(\Theta\) is a set of the local probability distributions for each node/or variable, conditional on each value combination of the parent variables. The graphical component, \(G\) makes BNs a class of probabilistic graphical models for reasoning under uncertainty, where the nodes represent variables (which can be discrete and/or continuous) and the arcs (links) represent direct (and sometimes causal) connections and dependencies between the linked variables. Those variables that are not linked directly in the graph are conditionally independent of each other. The second part of the network, \(\Theta\) represents the conditional probability distributions, which model the quantitative strength of the connections or dependencies between variables. These are represented through conditional probability tables (CPTs), allowing probabilistic beliefs to be updated automatically as new data becomes available. The local probability distributions contain a parameter \(\theta_{x_i|\Pi_{x_i}} = \Psi(x_i|\Pi_{x_i})\) for each possible value \(x_i\) of \(X_i\), given each combination of the direct parent variables of \(X_i\), \(\prod_{x_j} \prod_{x_i}\), where \(\prod_{x_i}\) denotes the set of direct parents of \(X_i\) in \(G\). The network \(G\) then represents the joint probability distribution (JPD):

\[
P_G(X_1, \ldots, X_n) = \prod_{i=1}^n P_G(X_i|\Pi_{x_i}) = \prod_{i=1}^n \theta_{x_i|\Pi_{x_i}}
\]

(1)

This is the chain rule, which states that for a given set of variables \(X\), the BN specifies a unique JPD given by the product of all the CPTs specified in the BN.

By combining probability and graph theory to express mutual relationships between variables, BNs provide a robust mathematical modelling framework for reasoning about uncertainty. BN modelling generally comprises of structure and parameter learning. Structure learning can be done using expert/domain knowledge, data-driven (machine learning) approaches or a hybrid of the two. For this study, we used a data-driven approach considering the uncertain and limited information on the causal factors of COVID-19 transmission. As such, we tested a total of 13

| Variable                                      | Data source                                      |
|-----------------------------------------------|--------------------------------------------------|
| Household density                             | Derived from [37]                                |
| Percentage of youth (15–40)                   | Derived from [37]                                |
| Household size                                | Derived from [37]                                |
| March–May temperature                         | Data from the Eswatini Meteorological Service    |
| Poverty headcount index                       | Derived from [37]                                |
| Population density                            | Derived from [37]                                |
| Percentage of elderly (55 +)                  | Derived from [37]                                |
| Proximity to supermarkets                     | Points of interest (POI) from Google Maps/Google Earth |
| Proximity to major referral health facility   | Data from the Ministry of Health (Government of Eswatini) |
| Annual average daily traffic                  | Data from the Ministry of Public Works and Transport (Government of Eswatini) |
| Proximity to churches                         | Points of interest (POI) from Google Maps/Google Earth |
| Ratio of people to buildings                  | Derived from [47]                                |
| Proximity to all shopping centres             | Points of interest (POI) from Google Maps/Google Earth |
| Building density                              | Derived from [47]                                |
| Employment rate                               | Derived from [37]                                |

Table 1 List of variables used in this study
BN learning algorithms (Table 2) available on the WEKA data mining software [26]. The resultant network topologies and search strategies differ in the trade-offs made between network structure, computational complexity and structural richness.

One of the main tasks of this study was to identify those variables which influence COVID-19 transmission. A hybrid approach which iteratively alternates between filter ranking and wrapper feature subset selection was used in order to speed up the computation process. This approach used an evaluation heuristic which prefers subsets of features that are highly correlated with the response variable while having low inter-correlation. The Incremental Wrapper Feature Subset Selection with Naive Bayes (IWSS-NB) [27] was used for this purpose. This method combines the linear time complexity of filter algorithms...
and advantages of wrapper methods. The IWSS-NB technique attempts to get best ranking variables and once a variable is included, it is maintained until end of the search. A naïve Bayes classifier allows for the addition of new variables or interchange of new variable with that which was previously added.

A Markov blanket [28] was then applied over the response variable to ensure that every node or variable in the network is a parent or child of sibling of the response variable. In order to guarantee impartiality of the structure learning process, the number of parents for any node was set to a value of 100,000 which imposes no restrictions on the number of parents a variable could have. After learning the structure, the parameters or probabilities were learned based on the structure and the data using the Bayesian estimator [29] which directly computes the relative frequencies of the associated combinations of the attribute values in the training data. The Bayesian estimator also assumes that the conditional probability of each node corresponding to its parent instantiation conforms to the Dirichlet distribution with local parameter independence [28]. The alpha parameter, which is the initial count on each value, was used in the estimation of the conditional probability table for each of the BN models (except for the NB). This parameter was set to 0.5 to avoid bias.

### 2.4 Model performance evaluation

A stable and efficient result was obtained through 10 replicates and the cross-validation method that all the samples were randomly divided into ten folds, and each fold, in turn, used for test data. As a result, two BN models were developed: the first one using data for the period from 11th March to 8th May 2020 when a strict lockdown was enforced throughout the country and the second model covering the period from 9th May to 30th June 2020 when the lockdown restrictions were gradually eased. The models were evaluated using logarithmic loss and recall values [30]. The logarithmic loss was chosen so as to quantify model accuracy by penalising false COVID-19 predictions based on the posterior probabilities from the BN models. Hence, high log loss values indicate higher penalty and inferior model performance and vice versa. The recall, on the other hand, measured the proportion of COVID-19 cases that were correctly identified. These two metrics were deemed suitable for the problem at hand considering the need to accurately detect and predict COVID-19 cases. In addition, the mutual information or entropy reduction [31] between the predictor and response variable was estimated as a measure of variable influence on the spatial distribution of COVID-19 cases. Given a probability distribution $p$ defined over two sets of variables $X$ and $Y$, the mutual information between $X$ and $Y$, which is measured in bits, is:

$$I(X, Y) = \sum_{x,y} p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right),$$  

where $p(x)$ and $p(y)$ are the probability densities of $x$ and $y$, and $p(x,y)$ is the joint density.

### 3 Results

Table 3 provides the performance outcomes of each BN learning algorithm. Generally, all the BN algorithms performed relatively well. However, the TAN model outperformed all other algorithms with a mean log loss of 0.705.
and recall value of 0.821. This is followed by the GS algorithm while the kNB had the lowest overall performance. Hence, the TAN model was selected for modelling COVID-19 cases for both the period before 8th May 2020 and after.

Of all the variables initially presented to the models, only five variables were selected through the variable selection method used. The selected variables include proximity referral health facilities, churches and shopping facilities. The BN model explains the underlying process by which the observed COVID-19 cases might be explained. The links between the variables in the model (nodes in graph) indicate the conditional dependence on each other and not necessarily a causal relationship (Fig. 3). For instance, daily traffic is dependent on proximity to supermarkets. Similarly, proximity to supermarkets is conditionally dependent on proximity to shopping centres.

The predicted probabilities of COVID-19 transmission show very similar spatial patterns as the reported cases in Eswatini (Fig. 4). All the areas with reported cases were correctly predicted with high posterior probabilities from the BN model. The first model, which was based on the earliest epidemic data, possessed comparable accuracy rates for the high to very high transmission risk zones, covering 98% of reported cases after 8th May 2020. The high transmission risk areas shown on the maps are in densely populated urban and peri-urban areas centred around the Manzini-Mbabane corridor where there is high human mobility. Other potentially infectious areas were the regional towns and surrounding areas.

A total of five variables were also selected for model covering the period after 8th May 2020 (Fig. 5). The selected variables are very similar to those of the first model for the period when there was a strict lockdown. Notably, the proximity to churches in the model for the earlier period is replaced by the proportion of youth as an important predictor of COVID-19 infection, a variable that is also conditionally dependent on proximity to supermarkets.

The spatial prediction of risk of COVID-19 transmission shows similar patterns as the earlier period (before 8th May 2020). However, there is a notable increase in the level of risk which spreads laterally from the hotspot areas to other parts of the country (Fig. 4). When comparing Figs. 4 and 6, it is evident that the risk of transmission is spreading geographically following the key driving variables such as

---

**Fig. 3** The Bayesian network model for the period up to 8th May 2020 (during strict lockdown)

**Fig. 4** The map showing posterior probabilities of COVID-19 transmission derived from the BN model for the period up to 8th May 2020
human (traffic) movement and proximity to referral health facilities and amenities such as shopping facilities and supermarkets.

The proximity to major referral health facilities had the largest influence followed by proximity to churches, shopping centres and supermarkets (Fig. 7). Annual average daily traffic was also an important predictor albeit with lower mutual information. In contrast, the model developed from data after 8th May 2020 case indicates a shift in the ordering of the drivers of COVID-19 spatial distribution.

Traffic density, a proxy for human mobility, becomes the strongest predictor followed by proximity to major health facilities. These temporal variations in both the magnitude and ordering of variable influence possibly point to changes in the drivers of COVID-19 transmission over time as determined by the changes in the non-pharmaceutical intervention strategies.
4 Discussion

The COVID-19 infection rate in Eswatini has been steadily increasing since the first imported case in early March. Most infections were spatially clustered in the major urban centres of Manzini and Mbabane and the surrounding peri-urban areas (see Fig. 1). Thereafter, the infections steadily spread to other populated areas of the country as driven by a number of factors.

Our modelling indicates that the risks of COVID-19 outbreak (also referred to as the ‘transmission risk’) are higher in the major urban areas where there is a high number of shopping facilities, health facilities and large churches. These high-risk areas can extend to areas in close proximity to the major transportation routes between the country’s major towns. The risk of the disease slowly spread to rural areas following proximity to the major urban areas driven by increased mobility as lockdown measure were relaxed. Altogether, the Manzini-Mbabane corridor and surrounding areas appear to be most at risk of a major outbreak. The findings may be useful since they highlight parts of the country in which subsequent COVID-19 cases were most likely to be detected.

Our findings show that during the strict lockdown (before 8th May 2020), a combination of proximity to shops, supermarkets, major referral health facilities and churches together with the average annual daily traffic explained the relatively high variability of the disease distribution in Eswatini. Upon close examination of the case data, it was evident that health facilities played an important role in the spread of the disease. Faced with limited personal protective equipment and uncertainty on the disease incidence, a number of health workers were infected and in turn infected other patients and vice versa [32]. Similarly, supermarkets were also identified as sources of super spreader events, which the models captured. Although churches have been reported to be sources of super spreader events elsewhere [33, 34], there were no such reports or cases to that effect in Eswatini despite the model identify them as an important factor. This, however, does not exclude such places as potential sources of infection during the earlier stages of the pandemic in the country.

However, after gradual easing of the lockdown restrictions (from 9th May 2020), the proportion of youth became an important factor and replaced proximity to churches as one of the selected variables. This is an notable outcome because the youth are predominantly more mobile, asymptomatic and perhaps also more carefree, considering the earlier reports that they were least affected by the virus [35]. The youth has been observed to be the slowest to adherence to social distancing measures [36]. Eswatini has a predominantly young population, 37% of which is within the ages of 15–40 years [37]. Low income countries, such as Eswatini, are characterized by a relatively young population, a factor that is predicted to limit the severity of the disease [35, 38]. Conversely, the younger population may also perpetuate the spread of the disease to vulnerable individuals through closer inter-generational contact [38]. Additionally, the working young population is more likely to be attracted to the highly populated urban areas, thereby increasing the transmission risk. The youth and working-class population should, therefore, be the key focus of the COVID-19 containment strategies.

The relationship between human movement and transmission risk is well established [39, 40]. Several studies have indicated that these densely populated urban environments and the high levels of mobility therein could increase the potential spread of COVID-19 [34, 40]. It is no surprise that daily traffic volume is also an important predictor in Eswatini. This is evidenced by the high clustering of cases and high probabilities in major urban and peri-urban areas that are characterized by high traffic volumes. It is expected, therefore, that with increased traffic and re-influx as a result of relaxation of the lockdown regulations, there will be rapid increases in the number of confirmed cases after 30th June 2020. In particular, the increasing traffic implies greater probability of contact between people, thereby increasing the risk of COVID-19 spread. Therefore, it is imperative that social distancing policies and other non-pharmaceutical interventions are continually implemented and enforced.

Climate, specifically temperature, did not present as one of the most influential variables despite previous links between climate, especially temperature, and COVID-19 transmission [41–43]. This may be attributed to insufficient variation across the country for temperature to be an influential predictor. Our study demonstrates the significant influence of socio-economic variables on COVID-19 incidence. Dowd et al. [44] emphasize the importance of considering demographic variables in infection control efforts. Continued monitoring of the identified factors can aid in understanding COVID-19 spread dynamics.

The ability of the model based on the earlier reported cases to accurately predict the incidence of the subsequent cases implies that BN models can forecast the spatial patterns of the disease. This is one of the reasons for the growing awareness of the importance of BNs for data analysis [45]. BNs also reduce the assumption of independence commonly used in regression models by allowing relationships between the variables to be accounted for in the model. The ability to handle data from disparate sources and types, handling of missing values, robustness to outliers, the ability to extract non-linear combinations of features, as well as the observed predictive power and interpretability makes BNs appealing over other modelling
techniques. The developed BN models can then help improve our understanding of both the dynamics of COVID-19 spread and the interrelationships between the predictor variables. The findings provide a useful reference for implementing targeted containment efforts in areas with increasingly community transmission.

However, one of the limitations of this study was data availability. Some of the potentially useful variables such as the incidence of other co-morbidities and behavioural variables (e.g., smoking and use of alcohol) were not readily available. Human mobility could also have been better presented using anonymized and aggregated cellular phone data, but that data was not available at the time of undertaking the study due to privacy concerns. Notably, other mobility data such as the Google mobility reports were also not available for Eswatini.

Another caveat is related to the COVID-19 testing strategy which is focused on individuals that are either contacts of individuals found to have the disease (contact tracing) or already hospitalized with significant symptoms. Such testing, according to [46], biases the identification of the extent to which COVID-19 is present in people with mild or no symptoms. BNs, therefore, provide a framework for explaining of the observed data to avoid highly misleading conclusions about infections [46]. The findings of the study, particularly the understanding of the drivers of COVID-19 infection, may also be used to explore the social and economic impacts of the pandemic in the affected parts of the country.

5 Conclusions

The BN models used in this study effectively modelled COVID-19 for identifying its potential geographic spread and the underlying influencing factors in Eswatini. The findings of this study suggest that the disease was spreading, driven by increasing human interaction through increased mobility and shopping. Whilst earlier infection intervention strategies may have made notable contributions to the containment of the disease, this study shows that easing of the lockdown regulations resulted in more widespread risk of transmission driven by young people and spreading from the major urban areas through peri-urban areas and potentially to rural areas. Our study implies that mitigation strategies should focus on the predominant influencing factors particularly in the areas with increasing infections. With limited research on the country-level spatial modelling of COVID-19 in Africa and Eswatini, this study provides a basis for improved containment of the disease. Furthermore, the findings of the study may further contribute to the exploration of the social and economic impacts of the pandemic. The robustness of BNs in spatial predictive modelling is also demonstrated even under conditions of uncertainty and limited data.

Acknowledgements

Epidemiological data was provided by the Ministry of Health’s Epidemiology and Disease Control Unit (EDCU) whilst demographic and socio-economic data was provided by the Central Statistics Office (CSO).

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit organizations.

Declarations

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.

References

1. Adekunle, I. A., Onanuga, A. T., Akinola, O. O., & Ogunbanjo, O. W. (2020). Modelling spatial variations of coronavirus disease (COVID-19) in Africa. Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2020.138998
2. Zhao, Z., Li, X., Liu, F., Zhu, G., Ma, C., & Wang, L. (2020). Prediction of the COVID-19 spread in African countries and implications for prevention and control: A case study in South Africa, Egypt, Algeria, Nigeria, Senegal and Kenya. Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2020.138959
3. Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., & Siddique, R. (2020). COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. Journal of Advanced Research. https://doi.org/10.1016/j.jare.2020.03.005
4. Chan, J. F. W., Yuan, S., Kok, K. H., To, K. K. W., Chu, H., Yang, J., et al. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. The Lancet, 395(10223), 514–523. https://doi.org/10.1016/S0140-6736(20)30154-9
5. Ferguson, N., Laydon, D., Nedjati Gilani, G., Imai, N., Ainslie, K., Baguelin, M., et al. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf. Imperial Ac.Uk, (March), 1–20. https://doi.org/10.25561/77482
6. McBryde, E. (2020). The value of early transmission dynamic studies in emerging infectious diseases. The Lancet Infectious Diseases. https://doi.org/10.1016/S1473-3099(20)30161-4
7. Hassaan, M. A., Abdelwahab, R. G., Elbarky, T. A., & Ghazy, R. M. (2021). GIS-based analysis framework to identify the determinants of COVID-19 incidence and fatality in Africa. Journal of Primary Care & Community Health, 12, 21501327211041210. https://doi.org/10.1177/21501327211041208
8. Zhao, W., Zhang, J., Meadows, M. E., Liu, Y., Hua, T., & Fu, B. (2020). A systematic approach is needed to contain COVID-19 globally. Science Bulletin. https://doi.org/10.1016/j.scib.2020.03.024
9. Huang, R., Liu, M., & Ding, Y. (2020). Spatial-temporal distribution of COVID-19 in China and its prediction: A data-driven modeling analysis. Journal of Infection in Developing Countries, 14(3), 246–253. https://doi.org/10.3855/jidc.12585
10. Zhou, C., Su, F., Pei, T., Zhang, A., Du, Y., Luo, B., et al. (2020). COVID-19: Challenges to GIS with big data. *Geography and Sustainability, 1*(1), 77–87. https://doi.org/10.1016/j.geosust.2020.03.005

11. Mollalo, A., Vahedi, B., & Rivera, K. M. (2020). GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Science of the Total Environment*. https://doi.org/10.1016/j.scitotenv.2020.138884

12. Franch-Pardo, I., Napoletano, B. M., Rossete-Verges, F., & Billa, L. (2020). Spatial analysis and GIS in the study of COVID-19. A review. *Science of the Total Environment, 739*, 140033. https://doi.org/10.1016/j.scitotenv.2020.140033

13. Kamel Boulos, M. N., & Geraghty, E. M. (2020). Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: How 21st century GIS technologies are supporting the global fight against outbreak. *International Journal of Health Geographics, 19*(1), 8. https://doi.org/10.1186/s12942-020-00202-8

14. Fatima, M., O’Keefe, K. J., Wei, W., Arshad, S., & Gruenbner, O. (2021). Geospatial analysis of COVID-19: A scoping review. *International Journal of Environmental Research and Public Health*. https://doi.org/10.3390/ijerph18052336

15. Lyseen, A. K., Nehr, C., Sørensen, E. M., Gudes, O., Geraghty, E. M., Shaw, N. T., et al. (2014). A review and framework for categorizing current research and development in health related geographical information systems (GIS) studies. *Yearbook of Medical Informatics, 9*(1), 110–124. https://doi.org/10.15265/2014-0008

16. Mour Mouris, P., Tzelves, L., Roidi, C., & Fotsali, A. (2021). COVID-19 transmission: A rapid systematic review of current knowledge. *Osong Public Health Res Perspect, 12*(2), 54–63. https://doi.org/10.24171/j.phrp.2021.12.2.02

17. Pearl, J. (1988). Probabilistic reasoning in intelligent systems. *Probabilistic reasoning in intelligent systems*. San Mateo: Morgan Kaufmann.

18. Puga, J. L., Krzywinski, M., & Altman, N. (2015). Bayesian networks. *Nature Methods, 12*(9), 799–800. https://doi.org/10.1038/nmeth.3550

19. Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning, 20*(3), 197–243. https://doi.org/10.1023/A:1022623210503

20. Mihaljević, B., Biezl, C., & Larrañaga, P. (2021). Bayesian networks for interpretable machine learning and optimization. *Neurocomputing, 456*, 648–665. https://doi.org/10.1016/j.neucom.2021.01.138

21. Haddawy, P., Hasan, A. H. M. I., Kasantikul, R., Lawpoolsri, S., Sa-anchari, P., Kwaenkungwai, J., & Singhhasivanon, P. (2018). Spatiotemporal Bayesian networks for malaria prediction. *Artificial Intelligence in Medicine, 84*, 127–138. https://doi.org/10.1016/j.artmed.2017.12.002

22. Ho, S. H., Speldewinde, P., & Cook, A. (2017). Predicting arboviral disease emergence using Bayesian networks: A case study of dengue virus in Western Australia. *Epidemiology and Infection, 145*(1), 54–66. https://doi.org/10.1017/S0950268816002090

23. Mayfield, H. J., Smith, C. S., Lowry, J. H., Watson, C. H., Baker, M. G., Kama, M., et al. (2018). Predictive risk mapping of an environmentally-driven infectious disease using spatial Bayesian networks: A case study of leptospirosis in Fiji. *PLoS Neglected Tropical Diseases, 12*(10), e0006857. https://doi.org/10.1371/journal.pntd.0006857

24. García, S., Luengo, J., Sáez, J. A., López, V., & Herrera, F. (2013). A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering, 25*(4), 734–750. https://doi.org/10.1109/TKDE.2012.35

25. Vogel, K., Riggelsen, C., Korup, O., & Scherbaum, F. (2014). Bayesian network learning for natural hazard analyses. *Natural Hazards and Earth System Sciences, 14*(9), 2605–2626. https://doi.org/10.5194/nhess-14-2605-2014

26. Frank, E., Hall, M. A., & Witten, I. H. (2017). The WEKA workbench. *Data mining* (4th ed., pp. 553–571). Morgan Kaufmann. https://doi.org/10.1007/978-1-280-29271-5.00024-6

27. Bermejo, P., Gámez, J. A., & Puerta, J. M. (2014). Speeding up incremental wrapper feature subset selection with Naive Bayes classifier. *Knowledge-Based Systems, 55*, 140–147. https://doi.org/10.1016/j.knosys.2013.10.016

28. Chen, Y. C., Wheeler, T. A., & Kochenderfer, M. J. (2017). Learning discrete Bayesian networks from continuous data. *Journal of Artificial Intelligence Research, 59*, 103–132. https://doi.org/10.1613/jair.5371

29. Bouckaert, R. R. (1994). A stratified simulation scheme for inference in Bayesian belief networks. *In R. L. de Mantaras & D. Poole (Eds.), Uncertainty proceedings 1994* (pp. 110–117). Berlin: Elsevier. https://doi.org/10.1016/b978-1-55860-332-5.50019-5

30. Monaco, R. M. (1992). Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis. *International Journal of Forecasting, 8*, 119–120. https://doi.org/10.1016/0169-2070(92)90021-z

31. Laskey, K. B. (1995). Sensitivity analysis for probability assessments in Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics, 25*(6), 901–909. https://doi.org/10.1109/36.400525

32. Wang, J., Zhou, M., & Liu, F. (2020). Reasons for healthcare workers becoming infected with novel coronavirus disease 2019 (COVID-19) in China. *Journal of Hospital Infection*. https://doi.org/10.1016/j.jhin.2020.03.002

33. Yong, S. E. F., Anderson, D. E., Wei, W. E., Pang, J., Chia, W. N., Tan, C. W., et al. (2020). Connecting clusters of COVID-19: An epidemiological and serological investigation. *The Lancet Infectious Diseases, 20*(7), 809–815. https://doi.org/10.1016/S1473-3099(20)30273-5

34. Shim, E., Tariq, A., Choi, W., Lee, Y., & Chowell, G. (2020). Transmission potential and severity of COVID-19 in South Korea. *International Journal of Infectious Diseases, 93*, 339–344. https://doi.org/10.1016/j.ijid.2020.03.031

35. Diop, B. Z., Ngom, M., Pougué Biyong, C., & Pougué Biyong, J. N. (2020). The relatively young and rural population may limit the spread and severity of COVID-19 in Africa: A modelling study. *BMJ Global Health*. https://doi.org/10.1136/bmjgh-2020-002699

36. Neyens, T., Faes, C., Vranckx, M., Pepermans, K., Hens, N., van Damme, P., et al. (2020). A spatial model to optimise predictions of COVID-19 incidence risk in Belgium using symptoms as reported in a large-scale online survey. *MedRxiv*. https://doi.org/10.1101/2020.05.18.20105627

37. Central Statistical Office. (2018). *The 2017 Population and Housing Census: Preliminary Results*. Mbabane.

38. Walker, P. G. T., Whittaker, C., Watson, O. J., Baguelin, M., Winskill, P., Hamlet, A., et al. (2020). The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries. *Science, 369*(6502), 413–422. https://doi.org/10.1126/science.abc0335

39. Lee, H., Park, S. J., Lee, G. R., Kim, J. E., Lee, J. H., Jung, Y., & Nam, E. W. (2020). The relationship between trends in COVID-19 prevalence and traffic levels in South Korea. *International Journal of Infectious Diseases, 96*, 399–407. https://doi.org/10.1016/j.ijid.2020.05.031
40. Jia, J. S., Lu, X., Yuan, Y., Xu, G., Jia, J., & Christakis, N. A. (2020). Population flow drives spatio-temporal distribution of COVID-19 in China. Nature, 582(7812), 389–394. https://doi.org/10.1038/s41586-020-2284-y
41. Byass, P. (2020). Eco-epidemiological assessment of the COVID-19 epidemic in China, January–February 2020. Global Health Action. https://doi.org/10.1080/16549716.2020.1760490
42. Méndez-Arriaga, F. (2020). The temperature and regional climate effects on communitarian COVID-19 contagion in Mexico throughout phase 1. Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2020.139560
43. Ma, Y., Zhao, Y., Liu, J., He, X., Wang, B., Fu, S., et al. (2020). Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2020.138226
44. Dowd, J. B., Andriano, L., Brazel, D. M., Rotondi, V., Block, P., Ding, X., et al. (2020). Demographic science aids in understanding the spread and fatality rates of COVID-19. Proceedings of the National Academy of Sciences of the United States of America, 117(18), 9696–9698. https://doi.org/10.1073/pnas.2004911117
45. Aronow, P. M., & Sävje, F. (2020). The book of why: The new science of cause and effect. Basic Books. https://doi.org/10.1090/noti1912
46. Fenton, N. E., Neil, M., Osman, M., & McLachlan, S. (2020). COVID-19 infection and death rates: The need to incorporate causal explanations for the data and avoid bias in testing. Journal of Risk Research. https://doi.org/10.1080/13669877.2020.1842902
47. Center for International Earth Science Information Network (CIESIN) Columbia University, & Novel-T. (2020). GRID3 Benin settlement extents version 01, Alpha. Palisades. https://doi.org/10.7916/d8-7etg-x863
48. Duda, R. O., & Hart, P. E. (1973). Pattern classification and scene analysis. Wiley, New York, 1973, 482 pages, $22.50. Artificial Intelligence, 4(2), 139–143. https://doi.org/10.1016/j.artint.1973.08.0004-0
49. Friedman, N., Geiger, D., & Goldszmit, M. (1997). Bayesian network classifiers: overfitting and underfitting with machine learning algorithms (no date). Machine Learning, 29(2/3), 131–163. https://doi.org/10.1023/A:1007465528199
50. Buntine, W. (1990). A guide to the literature on learning probabilistic networks from data. IEEE Transactions on Knowledge and Data Engineering, 8(2), 195–210. https://doi.org/10.1109/69.494161
51. Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 9(4), 309–347. https://doi.org/10.1023/A:1022649401552
52. Abramovic, M., Neubach, M., & Holland, A. (2008). Competing fusion for Bayesian applications. In Proceedings of IPMU (pp. 378–385). Malaga. http://www.gimac.uma.es/IPMU08/proceedings/papers/051-HollandEtAI.pdf
53. Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. Science, 220(4598), 671–680. https://doi.org/10.1126/science.220.4598.671
54. Glover, F. (1990). Tabu search—Part II. ORSA Journal on Computing, 2(1), 4–32. https://doi.org/10.1287/ijoc.2.1.4
55. Verma, T., & Pearl, J. (1992). An algorithm for deciding if a set of observed independencies has a causal explanation. In D. Dubois, M. P. Wellman, B. D’Ambrosio, & P. Smets (Eds.), Uncertainty in artificial intelligence (pp. 323–330). Elsevier. https://doi.org/10.1016/b978-1-4832-8287-9.50049-9
56. Lambot, S., Slob, E. C., Den Van Bosch, I., Stockbroeckx, B., & Vancluchooster, M. (2004). Modeling of ground-penetrating radar for accurate characterization of subsurface electric properties. IEEE Transactions on Geoscience and Remote Sensing, 42(11), 2555–2568. https://doi.org/10.1109/TGRS.2004.834800
57. Cooper, G. F., Hennings-Yeomans, P., Visweswaran, S., & Barouda, M. (2010). An efficient bayesian method for predicting clinical outcomes from genome-wide data. AMIA ... Annual Symposium proceedings/AMIA Symposium, 2010, 127–131.
58. Goldberg, D. (1989). Genetic algorithms in search, optimization, and machine learning. Choice Reviews Online, 27(02), 27-0936-27-0936. https://doi.org/10.5860/choice.27-0936
59. Zhang, H., Jiang, L., & Su, J. (2005). Hidden naive Bayes. In Proceedings of the national conference on artificial intelligence (Vol. 2, pp. 919–924). Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.