Evaluating Visual Analytics Methods for Epidemiological Application: A review

Ahmed Arshad, Edoardo Santilli, Eleni Barmpopoulou, Ilektra-Makrina Papazoglou, Konstantinos Stathoulopoulos, Sergio Montero Castanares, Tommaso Buranelli

City University London, 2016

Abstract— We conduct a literature review to determine the utility of visual analytics in epidemiological applications. We find that the key components for such systems to be effective and useful are spatial-temporal flexibility with rich set of analytical methods acting as a back end allowing for direct manipulation of data and for a wide and often disparate datasets being ingested, processed and merged as input to these systems. To gain an appreciation of these methods, we conduct a systematic literature review, in which we explore the application of visual analytics methods in analysing geographical time series data and within this context we look more specifically at tools and techniques utilised in epidemiology. Within the epidemiological space, we consider tools and techniques that allow professionals to understand disease trends over time and have been applied in a practical context to aid in policy and decision making. We look at some specific methods and case studies where some or all of these methods have been used together in a unified visual analytics framework to derive insight and deliver value. We highlight two dominant approaches to the visual analytics approaches to such datasets 1) GIS/Map based and 2) Network based.

INTRODUCTION

Visual Analytics is an approach that combines human intuition and the science of mathematical intuition to perceive patterns and extract knowledge and insight from the data. It is about the formation of visual metaphors that complements the human information driven storytelling discourse, which enables pattern recognition within vast and dynamically changing information spaces [1].

In this multidisciplinary research field different visualization experts closely cooperate with researchers from analytical disciplines, such as statistical analysis and modelling, machine learning and data mining, and geographical analysis and modelling, on developing new approaches to solve complex analysis. In order to get insights about the complex dynamics which steer the modern society, people need to model and understand global processes, like demography, economy, environment, energy, epidemic, international relatioship, to mention some. A fundamental step of this learning process consists in discovering how characteristics of these fenomena change and relate in time and space. Geovisual analytics (or geospatial visual analytics) provides tools to address those kind of problems involving geographical space and various objects, events, phenomena, and processes populating it with a particular attention on revealing spatial and spatio-temporal patterns. Geovisual analytics (or geospatial visual analytics) deals with problems involving geographical space and various objects, events, phenomena, and processes populating it. Since most of the things populating space occur or change in time, geovisual analytics must give proper attention to time and relationships between space and time [2].

In this context we explore the variety of tools and techniques that have been proposed that could potentially help public health professionals in understanding disease trends and aid in policy making [3].

The paper is organised as follows: In section 2, we perform a review of interesting applications of such methods in literature, section 3 we present case studies followed by our discussion in section 4. Section 5 consists of our conclusions.

2 CURRENT WORK

In the last 20 years, an increasing focus on the need of informatics and analytics in public health has been observed [4], [5] and several systems have been elaborated to help public health professionals to analyse the complex data used in infectious disease control[6], [7]. Approaches suggested for visualizing datasets within an epidemiological context can be broadly summarised into 2 broad categories 1) Geographic Information System and Map based methods (GIS) 2) Network based modelling and visualisation approaches [3]. It is worth noting that a wide range of methods such as data integration, parsing, filtering and statistical analysis is required as a pre-processing step before these methods could be applied.

A general discussion and framework for visualizing time oriented data is presented in [8]. They note the importance of interaction with time oriented data analysis. Users must be able to explore their data and directly manipulate it, linking and brushing is common in visualisation frameworks but the ability to switch between different time slices (e.g. daily, weekly, monthly) within the same environment is rather uncommon. They also suggest a time, data and representation based categorization criteria for thinking about appropriate time oriented visualisations. The Seasonal Trend decomposition based on locally-weighted regression (STL) technique and its effectiveness in analysing serially correlated time series data is discussed in [9]. They note the techniques robustness and flexibility in parametrisation of trend variations and seasonal time series components being highly conducive to visual analytics approaches.

In [10] a practical application is demonstrated which incorporates a visual analytics architecture that ingests real time data from a variety of sources such as hospitals, clinics and weather satellites, which is pre-processed, merged and then fed into machine learning models and the outputs of which are piped into heat maps, heights and choropleth maps to enable clinicians and health officials to understand disease and prepare for disease outbreaks at a country level. The merits of a specific visual analytics platform for modelling and exploration of pandemic influenza is discussed in [11]. They identify the limitations of current methods utilised in planning for potential mass casualty events such as pandemic flue and highlight two key limitations. Firstly, they note the often highly complex modelling required by officials responsible for making plans for such events which require supercomputers and second, which is the opposite case the drastic simplifications that are made to models to make them computationally feasible. Both [11], [10] incorporate the GIS/Map based presentation with the back end statistical analysis to present the user with a flexible environment to explore multiple scenarios and decisions. This is exemplified in Fig 2.

In [12] a unified package in the open source statistical software R, to do multivariate, multidimensional analysis and visualisation of disease outbreak data. Through a unified data structure the user is able to use different types of data such as individual metadata (e.g. age, sex), time correlated observations such as swab results, contacts between patients, DNA sequences of pathogens, phylogenetic trees, and contextual data at the population level. An example of the visual analytics made possible by such an approach is shown in Fig 1.

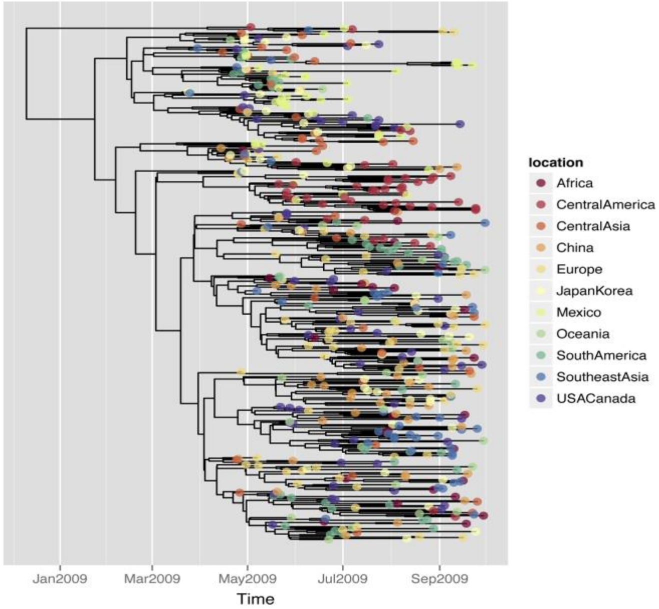


Fig 1: Dendogram shows Phylogeny of pandemic influenza H1N1 sequences and their prevalence in countries. This is an example of network based visualisation.[11][12]

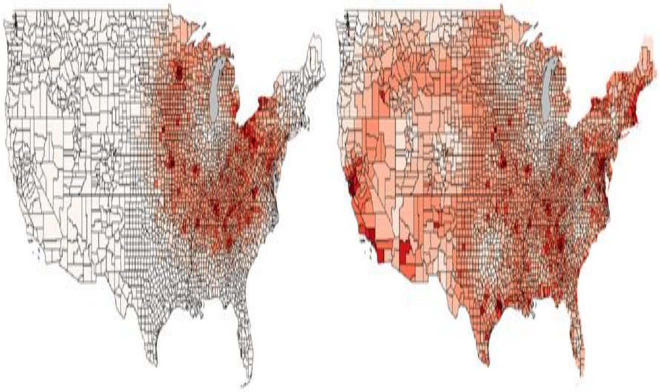


Fig 2: Map showing the results of modelling a pandemic spread originating in Chicago. (Left) The effects of an outbreak after 40 days using a single source point spread model. (Right) The effects of an outbreak after 40 days including air travel between the 15 largest United States airports [11]

In [13] the increasing interest within the epidemiological research community in utilising social network analysis (SNA) techniques and agent based models (ABM). They note the suitability of network models in simulating the dynamics of social contagion within populations but also recognise its potential limits with regards to generalizability and causal inference. ABM's are also discussed for their suitability and ability to model health determinants at multiple levels with potential couplings with social interaction networks. In [14] a very successful application of SNA techniques complemented with whole genome sequencing. They were able to identify two genetically distinct lineages of tuberculosis with identical genotypes suggesting parallel outbreaks. Through SNA techniques they were able to identify key members of the high risk social network which were not possible with traditional techniques.

3 Case Studies

In this section we look at some case studies which exemplify the GIS and network based approaches we identified earlier. The first two case studies are of the GIS case. Also we present a brief description of network terminology in context of epidemiology before presenting the network case study.

3.1 Hand, foot and mouth disease: spatiotemporal transmission and climate

Hand-Foot-Mouth Disease (HFMD) [15] is a very common gastrointestinal infectious disease in China. The highest risk factor groups are children less than 5 years because of the weakness of their immune system at this age. This virus is problematic due to its strong transitivity causing large and fast epidemics. Furthermore, HFMD is correlated to population density and communication. Buffer areas between urban and rural ones are the most affected one. Finally, there is the issue of regional health services not being unable to identify the virus. Hence, there is a need to understand the factors which affect disease transmission and predict outbreaks to both inform control strategies and improve public health outcomes.

The possible correlations between climate and HFMD are investigated through the BME-S method. This is a combination between Bayesian Maximum Entropy (BME) theory and the Self-Organized Map (SOM) technique. Descriptions of BME are beyond the scope of this paper and the reader is referred to references provided. BME method for spatial analysis and mapping provides definite rules for incorporating prior information from different sources[16]. This allows for both the incorporation of 'hard' data i.e. observations, 'soft' data such as interval observations, higher-order moments, probability characteristics, physical laws, experience and prior knowledge[16]. BME in this case allowed incorporation of core knowledge in terms of epidemic laws, when available, in addition to site-specific information in terms of case numbers.

The SOM is a special type of competitive neural network. It has the property of topology preservation and can be used for projection of multivariate data, density approximation, and clustering [17]. In this case the SOM is used for dimensionality reduction of the high dimensional data into 2 dimensions and clustering. The SOM possesses two layers: an input and output layers. The input neurons receive the input information simulating the retina. The output layer simulates the cortex and its neurons establishing a neighborhood structure.

The advantages for this approach include the ability to analyze large volumes of multi-dimensional data due to effective dimensionality reduction through SOM. The topology of the SOM helps it outperform mature methods such as hierarchical, k-means clustering. Some disadvantages include the fact that only 11months of data are analyzed.

However, this method allows us to detect the composite space-time domain in which HFMD varies rather than a purely spatial and purely temporal variation. It is found that cases are geographically clustered and linked to precipitation patterns shown in Fig 3. Thus precipitation patterns can be used for prediction.

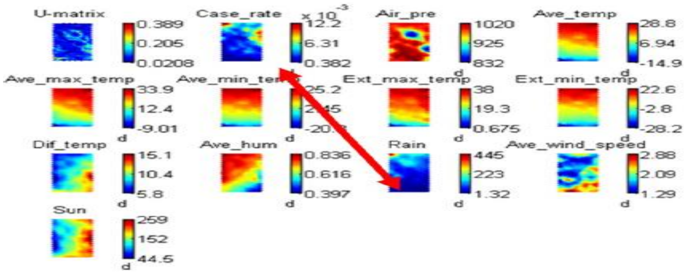


Fig 3: SOM component plains for the HFMD and meteorological indicators.

3.2 ID-Viewer: A visual analytics decision support system for infectious diseases surveillance in Pakistan

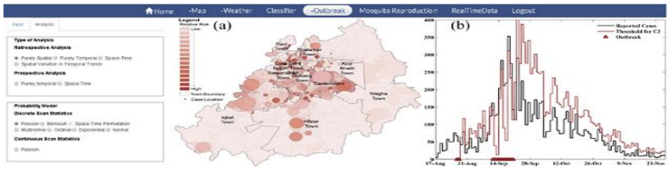
Developed countries use disease surveillance systems in order to accurately detect possible outbreaks of infectious diseases (ID) and tackle this threat immediately[10]. However, Pakistan does not have the infrastructure, nor the data, to use such systems. To make matters worse, the existing surveillance systems cannot be adapted by Pakistan, since they are robust only for countries where the climate variation between seasons is not extreme. For this reason, the authors suggested a novel visual analytics framework which could improve the analysis of multivariate spatiotemporal data and help health care officials to gather streaming ID data and make predictions based on them.

ID-viewer system contains the following elements:

- Web services which automatically gathered data for different cases of ID.
- Statistical data analysis tools which relied on dynamic syndromic classification and could track the spatio-temporal spread of ID and predict disease outbreaks.
- An interactive visual analytics environment that includes various analysis techniques and improves the decision making process regardless of the size and complexity of the dataset.

This case study like the previous one utilizes GIS based methods that we identified in section 2. The interactive visual platform of ID-Viewer is displayed on LCD monitors that enable the user to allocate resources effectively during an ID outbreak. Moreover, several spatio-temporal visualizations can be produced, such as heat-maps for syndromic clusters, real-time diagrams which identify the spread pattern of a disease and choropleths that map how an epidemic expands. It should be mentioned that ID-viewer allows the user to focus on a specific area on the GIS-map and examine it in detail.

Streaming data are passed as input to Visual Infrastructure Management (VIM). VIM includes an information renderer, preprocessing filters, visualization methods and geospatial routines. The real-time data stream is processed, organized and analyzed through the VIM while the information renderer sets the visual aspects of the processed data. Colors, shading effects and textures as well as cartographic and typographic problems are addressed at this step. VIM passes the refined elements to Visualization Control Room (VCR) which manages the visual framework.



Type of Analysis
Retrospective Analysis
● Purely Spatial ○ Purely Temporal ○ Space-Time
○ Spatial Variation in Temporal Trends
Prospective Analysis
○ Purely temporal ○ Space-Time

Probability Model
Discrete Scan Statistics
● Poisson ○ Bernoulli ○ Space-Time Permutation
○ Multinomial ○ Ordinal ○ Exponential ○ Normal
Continuous Scan Statistics
○ Poisson

Fig 4: a) Scan statistics for spatial clusters b) Temporal outbreak detection for constitutional/hemorrhagic fever

VCR allows the user to interact and experiment with different features and create epidemic clusters while it builds geographic visualizations such as choropleths and heat maps, to locate potential epidemic threats and determine possible solutions. To summarize, the always online, interconnected nature of ID-Viewer can reveal unexpected ID outbreaks and hidden disease spread patterns which could not be predicted with any other system in countries like Pakistan.

3.3 Contact Network Epidemiology

In [18] some of the models used in mathematical epidemiology, its assumptions such as mass-action and limitations are discussed. They introduce the network approach as a more powerful analytical technique in overcoming the shortcomings of this assumption. The mass-action assumption states number of new cases of disease in a time interval is proportional to the product of numbers of infected and susceptible hosts in the previous time interval.

This assumption leads to the commonly used SIR model, which compartmentalizes the population into being in either the Susceptible (S), Infected (I) or Recovered (R) state. In this model, a host upon infection are assumed to be immediately infectious and remain in that state until they transition to the recovered state. Also, these hosts are assumed to have disease causing contacts with random individuals from the population defined by a Poisson process. Thus transmission occurs if and only if the individual is susceptible when in contact with a host. This is the mass-action assumption of the model.

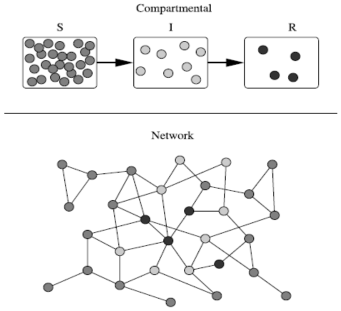


Fig 5: Shows the difference between a traditional SIR epidemiological model and a contact network model. SIR assume that all individuals in a group are equally likely to become infected, while contact network epidemiology considers diverse contact patterns that underlie disease transmission.

A limiting case of such models is identified in the SARS outbreak and estimates of its transmission rates based on the mass action assumption. The initial estimate for the spread of SARS were based to a great extent on the outbreak data from a hospital and a crowded apartment building, where there was an unusually high level of close contacts among individuals. This scenario did not hold in the wider population therefore these models greatly overestimated the transmission efficiency of this disease due to its inability to consider heterogeneous interaction patterns.

The contact network approach overcomes this by building a realistic network model of interaction patterns at a suitable temporal and spatial scale. This model is then used to predict the spread of diseases through the population based on pathogen characteristics and structural properties of the network. The network can then be used to model the effectiveness of control strategies and allow epidemiological impact of such changes.

Contact Network Models

These captures interaction patterns that could facilitate transmission of a disease by means of a graph representation. In this graph a person or location can be represented as a node and contacts as edges between nodes. An important network measure that is relevant here is the Degree of a node. The Degree of a node is total number of edges incident on a node and total number of edges emanating from a node. This can represent potential contacts that can lead to disease propagation. Hence, the degree distribution is crucial for the spread of a disease through a population. The mass action assumption can in the network case be represented as a random graph with a Poisson degree distribution. If the network does not show this structure traditional methods will be insufficient to model such phenomenon.

Some commonly observed network types include small world networks, which are distinguished by high levels of local clustering and global connectivity. Scale free networks are characterized by a power law degree distribution with a small fraction of highly connected components or hubs.

Contact networks usually have a more complex structure and do not fit easily into the network architectures described shown in Fig 6.

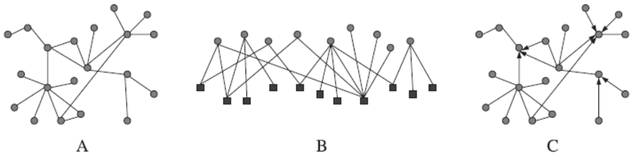


Fig 6: Contact networks A) Undirected Network B) Bipartite Network C) Semi-directed network

Bipartite networks with two node types can be used to represent asymmetric transmission probabilities between healthcare professionals and patients in a healthcare environment. Semi-directed networks where some contacts can be reciprocal and other one directional can model situations in which a person may infect another but the reverse is not possible.

Disease Dynamics Prediction

If we assume that an ID first appears at a random edge in the contact network, the disease can spread according a compartmental model with the Poisson process being replaced by the network structure. This node represents patient zero. This node will remain infected for some time and retain the ability to spread the infection to each of its contacts. In the secondary cases, the subsequent nodes can propagate the infection to their contacts and so on. Therefore the spread or percolation of the disease through a network depends on the level of contagion and structure of the contact network.

To understand the evolution and the fate of diseases in a network we can use Probability Generating Functions (PGF) which describe probability distributions and summarize salient information regarding the structure of the contact network. When the disease is introduced into a network it will traverse some but not all edges of the network with some average transmissibility. The edges that contract the infection can be thought of as being occupied. After the epidemic is over, the edges that link other nodes to the initial node with a continuous chain represents the outbreak. It also suggested that these interactions be modelled as bond percolations from statistical physics. Percolation Theory describes behavior of connected groups of edges in a random graph. This then allows us to characterize the size and distribution within this infected cluster.

Small outbreaks in a network

For a fixed network there typically exists a threshold value on the transmission rate below which only small finite sized outbreaks occur and above which large scale outbreaks covering the whole network occur.

Probability and size of a large scale epidemics

If we consider the case when the disease transmission rate breaches the threshold value for the transmissibility then we have an epidemic in the network. Here the transmission becomes so frequent that loops in the transmission chains become more likely. Given that an epidemic is occurring, we can estimate the probability of this happening and fraction of infected individuals. For an undirected network these quantities will be roughly equal. It is essentially the giant component defined by the occupied edges.

The SIMID case study presented next will build on the concepts we have expressed here. We should note that the power of network based modelling comes through when we consider the range of scenarios it can capture in its network structure and topology that are not possible in more traditional analysis.

3.2 SIMID VISUALIZATION TOOL

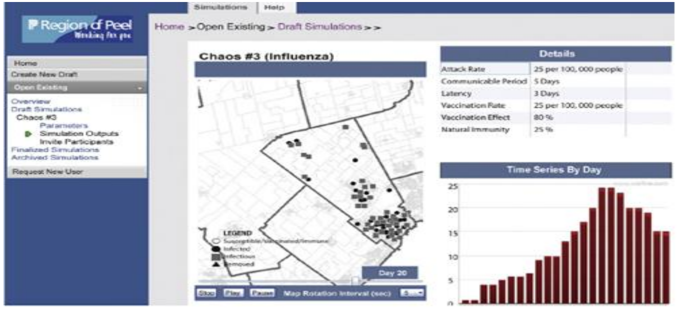
It is essential to detect gastrointestinal and respiratory infectious diseases promptly since they tend to spread and transmit rapidly throughout the population of an area. Dynamic spatiotemporal data can be used in order to monitor and predict outbreaks of severe influenza pandemic [19], however this information is not widely used by healthcare professionals.

SIMID is a novel, simulation-based visualization surveillance system that utilizes state of the art random network methodology and is established on R and GIS frameworks. The suggested network model is an extension of the design for networks of social contacts that was introduced by Newman M., E., J. [20]. This approach enabled the authors to build a population network of contacts that consisted of several subnetworks. To be more specific, this structure expresses the connections that may result in disease transmission and is illustrated as a probability distribution for the number of connections that each affected individual has with other people in an area.

An important characteristic of this network methodology is that it can accurately generalize some epidemic models while its hypotheses (infection rate, infectious period and homogeneity of the population) are not as strict as in other models. Moreover, SIMID can compute probability distributions for the sum of people who are infected by a disease, while it considers the following control measures:

- Mass vaccination that was done before the outbreak
- Ring vaccination and isolation that are done during the spread of a disease

The probabilistic character of SIMID enables the user to form an ensemble of probable scenarios for infectious outbreaks and allow the healthcare professionals to create optimal policies.



Chaos #3 (Influenza)

Details	
Attack Rate	25 per 100, 000 people
Communicable Period	5 Days
Latency	3 Days
Vaccination Rate	25 per 100, 000 people
Vaccination Effect	80 %
Natural Immunity	25 %

Fig 7: Shows the visualization of an outbreak scenario and temporal evolution in the SIMID tool.

How it works

To start with, the population that we are interested in has to be illustrated as a network whose peaks are single persons or groups and whose edges serve as the type of contact that "an infected agent can use for transmission from an infective to a susceptible individual" [18]. Then, the algorithm that will generate the different scenarios has to be implemented. Two models can be used by SIMID:

- SIR: The examined person is susceptible to the infectious agent, so he can be easily get infected and transmit the disease. After this stage, it either recovers or is removed. This method assumes that the individual cannot be affected by the virus again because he is either immune or dead.
- SEIR: The examined person is susceptible to the infectious agent and after the infection, he is categorized as exposed. This means that he is caring the disease but cannot infect other individuals yet. After a certain period of time of the exposure, he can infect other people and then he stops caring the disease.

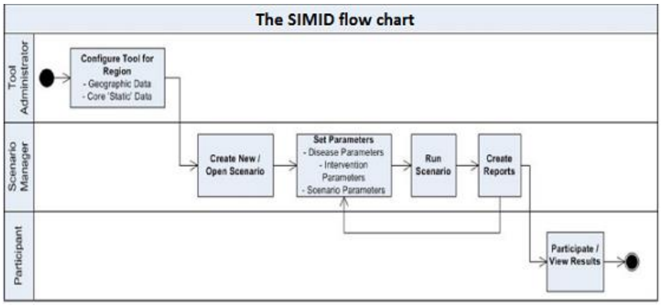


Fig 8: Flow chart showing a SIMID workflow

Once the simulation of the scenarios has finished, the generated output is used to produce maps that represent these probable outbreaks. Furthermore, these graphs are accompanied by a time-slider which allows the user to observe the spatiotemporal characteristics of the simulated spread of various diseases.

To summarize, SIMID has a simple and parametrized structure that combines probabilistic models with spatiotemporal visualizations. This software can improve the decision making process of healthcare professionals and it can also be used as a tool which will measure the state of readiness against an outbreak of an infectious disease.

4 Discussion

We have presented the GIS and Network paradigms of visual analytics in the context of epidemiology. The key component of both paradigms are the fact that there is an interplay between the computation and visualization aspects. This enables for insight and for effective decision making.

By evaluating several previous publications, the main focus of such visualisations was understood to be threefold in the context of epidemiology:

- Timely awareness of disease outbreaks
- Forecasting disease patterns & trends
- Classifying disease features & affected populations based on commonalities

From case studies 1 and 2 we observe that one of the key components of both systems were high volume information ingestion from a variety of sources. The data ingestion and preprocessing enabled application of sophisticated machine learning techniques. A combination of supervised and unsupervised methods are used and their results presented on a map to gain understanding of the spatial and temporal aspects of the problem. Supervised methods are those that involve learning probability distribution from labelled data. For a subsequent input the data then be classified as being of one category or the other or a function will be approximated from the data which enables the prediction of a continuous value. Classifiers such as Support Vector Machines are utilized in case 1. Case 2 makes use of an SOM to for unsupervised learning to perform dimensionality reduction and classification. These again highlight the interplay where the visualization informs the computation. From the SOM the authors infer the precipitation dependence and regions with such precipitation patterns are then selected for control measures.

The key analytical questions we see they were trying to answer with these techniques were:

- Which areas are going to experience a potential outbreak?
- What might be the scale of this outbreak?
- What factors influence such outbreaks?

Combination of the computation techniques with GIS based presentation enables the authors to effectively answer these questions.

In case studies 3, 4 we the graph paradigm of Visual Analytics. We observe that a graph representations allows for a much richer and complex hypothesis representations which are not otherwise possible with non-graph methods. Case 3 highlights how epidemiological quantities can be calculated from graph structure and degree distributions. Case 4 shows a more practical example of utilizing the flexible representation allowed by a graph structure to utilize more complex models. In [21] we see more advanced application network metrics. They use similarity measures of disease graphs to gain insight into disease relationships.

Some of the key analytical questions pertaining to the network paradigm can be summed up as:

- How will a certain disease spread through a network?
- What will be the scale of the epidemic and what are the probabilities associated with this?
- What impact will control measures have on this network?

The results of these investigations for both paradigms can be presented as a map allowing spatial context to be understood.

Spatial Data

According to [22], in the field of epidemiology the use of spatial methods of visualizing data is frequent, with a 4% of all epidemiology publications utilizing such methods. Recent publications cited throughout this review confirm the notion that spatial information is used widely in this field.

Georeferencing has recently become a popular selection of visualising data, using coordinates on observations. In the present context this method proved effective to a certain extent, when the focus was on regions of a map. However, due to data quality issues, the particular method should be used with care when the objective is to identify locations on an address or street level [20].

Location data in the context of epidemiology has been used for proximity calculations. This method has been particularly useful in assessing the toxicity of factors such as air or water pollutants in specific regions [21]. Most studies referencing such measures use linear distances (~80% of proximity articles) [19]. This distance measure would not be always recommended, considering that human behavioural and interaction patterns would require a more complicated measure that is reflective of accessibility per location.

Another common technique in epidemiology data visualisations was the aggregation of statistical measures on distance-based clusters [19], [22]. Classical statistics is frequently used in healthcare research to compare samples. In the case of spatial data, spatial clusters of sub-populations were used to compare different factors that may be causal for disease. While forming distance based clusters to define the geospatial prevalence of a disease could be an easy way of sampling, apart from the spatial commonalities, a lot of noise should be expected in the data. A good example of this would be the genetic variability of populations of metropolitan areas. Therefore such aggregations based on location must also be put to context and carefully edited according to the question in place.

When evaluating spatial analyses, it is clear that the main disadvantage lies with not satisfying the forecasting potential required by the mining of infectious disease data. Thus, although informative and rich, spatial data alone are not enough to provide actionable insight via visual exploration in epidemiological applications.

Temporal Data

The size of the regions examined in infectious disease research are expected to be large. However the temporal aspect does not usually include multiple time series and time is not necessarily examined in a low level of granularity in such studies [23]. The lowest level of granularity is commonly a 24 hour interval [24]. This was deemed natural due to the nature of the studies and the fact that patients are diagnosed and monitored on a higher temporal level. In order to model the temporal features in the data autoregressive models were most commonly used in the articles cited [24].

Thus, the challenges involved in time series analysis for epidemiological data are not similar in magnitude as they are for other analyses of human behavioural patterns [25]. However, computational algorithms should still be applied to aggregate the temporal data and create meaningful visualisations of time series in epidemiology. An important application in this field would be to successfully identify peaks and pits in disease prevalence patterns. Such computational methods were not identified in the majority of studies related to epidemiology, yet an interesting algorithm was formerly described by [25] and applied to a different field of study.

Spatiotemporal Data

While spatial clustering was common in the past, a more recent trend has been to utilise spatiotemporal data to monitor disease prevalence. Several techniques have been recently used to capitalise on spatiotemporal data [19]. In infectious disease research, the Knox technique is widely used to cluster together patients on the common grounds of time and location [26]. This was used in the formulation of the dynamic continuous-area space-time system (DYCAST), which is used to analyse spatial time series in order to monitor high risk areas [26]. Another commonly used method is the K-function which yielded better results in a comparative study on seasonal cancer-patient data from several regions [23]. The SOM clustering technique can be used for multivariate pattern identification & clustering, a relatively young technique not found to have been referenced within the field of epidemiology yet [27]. Many more algorithms and clustering techniques exist for spatiotemporal applications, which indeed have greater potential in both predicting disease outbreaks and forecasting the progress of on-going epidemics. However, as mentioned by [23] such spatiotemporal clustering techniques are affected by differential population changes during the duration of the study and are mainly affected by population growth rates.

It can therefore be concluded that whilst great potential for data exploration and spatiotemporal aggregation exists within the area of epidemiology, the techniques used can be improved to take advantage of higher volumes of data without giving up interpretability, and become more robust to the noise that is inherent in such multivariate and dynamic data sets.

Maps vs Graphs

Maps are a very popular tool for visualizing spatial & spatiotemporal data. However, several things may go wrong when reaching conclusions based on exploratory analysis of mapped data. Maps can easily lead to misinterpretation of the data if what is charted is not explicitly identified. Data pre-processing alterations are not easily communicated on map visualization. A second realization throughout this review is that while maps are a very useful tool for plotting spatiotemporal data, they do not directly provide information that is suggestive of relationships among the data. Thus causal effects are not enhanced through this visualization method. Instead, the generated view simply provide information in a way that is concise and informative, in the same way that a table would. In this sense, it cannot be denied that maps are a user-friendly tool that directly communicates a lot of information at once, especially when temporal factors are included. Nonetheless, knowing the data and the field of study, in combination with other tools is usually what leads to insightful hypotheses.

In order to overcome this problem the users should be encouraged to adopt tools that allow for interactive data interrogation with the use of maps. This would enable them to apply techniques such as the interactive grouping and clustering of observations in time and space within a single application. Moreover the users should familiarise themselves with more sophisticated methods of group assignment, such as the Voronoi tessellation, convex and concave hulls in addition to machine learning algorithms [25]. Being able to visualise the implications and results of such methods of more advance data analysis would enhance the insight provided by datasets in the field of epidemiology and could, in fact, reveal actionable conclusions and solutions that would affect many.

Graphs and trees are more abstract visual techniques that can be easily read by a user. The "ease" lies within the fact that the data representation, if properly designed, will focus on the relationships among the data. In contrast to maps the commonalities or connections among data points and variables are made prominent this makes them easier to read when many points are involved, or rather, makes the message they are conveying more clear. However these network graphs can also be biased in that they may only communicate a certain type of relationship, which may in fact prove misleading. In comparison to maps, these types of visualisation methods were considered to guide the user towards observing certain relationships, whereas maps provide the raw information, without favouring a particular conclusion. Therefore networks manage to compress information about the data relationships and effects, but should be carefully approached, as their quality relies heavily on the skills and understanding of the field that their creator has developed.

5 CONCLUSION

A large issue with epidemiology data nowadays is the volume of the data, combined with the need to interrogate them in order to unmask causal relationships. The latter hinders the applications of dimension reduction, as the data would need to be finally used by medical experts and policy makers rather than statisticians. This increases the need for interpretable visualisations, but also condensed ones. In order to explore multidimensional data and efficiently process data streams from a visual analytics perspective, the approach of multiple-linked views would be a good suggestion in epidemiology, as it would enable scientists and policy makers to manipulate the data without compressing dimensionality [28]. Only few and recent reports have utilised this as a means of exploring spatiotemporal epidemiology data.

In conclusion, a combination of the two aforementioned methods would be favourable. An even better approach would be the use of a tool that allows for dynamic interaction with views and graphs in order to overlay different types of networks on top of actual spatiotemporal maps. By examining both in parallel, a user would be able to interrogate the data and the relationships found amongst them more thoroughly, without ignoring the broader landscape of information and its changes in the time and space dimensions.

REFERENCES

[1] P. C. Wong and J. Thomas, “Visual analytics,” *IEEE Comput. Graph. Appl.*, vol. 24, no. 5, pp. 20–21, 2004.

[2] G. Andrienko, N. Andrienko, D. Keim, A. M. MacEachren, and S. Wrobel, “Challenging problems of geospatial visual analytics,” *J. Vis. Lang. Comput.*, vol. 22, no. 4, pp. 251–256, 2011.

[3] L. N. Carroll, A. P. Au, L. T. Detwiler, T. chieh Fu, I. S. Painter, and N. F. Abernethy, “Visualization and analytics tools for infectious disease epidemiology: A systematic review,” *J. Biomed. Inform.*, vol. 51, pp. 287–298, 2014.

[4] B. Reeder, D. Revere, R. a Hills, J. G. Baseman, and W. B. Lober, “Public Health Practice within a Health Information Exchange: Information Needs and Barriers to Disease Surveillance.,” *Online J. Public Health Inform.*, vol. 4, no. 3, 2012.

[5] A. Friede, H. L. Blum, and M. McDonald, “Public Health Informatics: How Information-Age Technology Can Strengthen Public Health,” *Annu. Rev. Public Health*, vol. 16, no. 1, pp. 239–252, 1995.

[6] S. C. Edberg, “Global Infectious Diseases and Epidemiology Network (GIDEON): a world wide Web-based program for diagnosis and informatics in infectious diseases.,” *Clin. Infect. Dis.*, vol. 40, no. 1, pp. 123–6, 2005.

[7] A. Friede, J. A. Reid, and H. W. Ory, “CDC WONDER: a comprehensive on-line public health information system of the Centers for Disease Control and Prevention.,” *Am. J. Public Health*, vol. 83, no. 9, pp. 1289–1294, Sep. 1993.

[8] W. Aigner, S. Miksch, W. M??ller, H. Schumann, and C. Tominski, “Visualizing time-oriented data-A systematic view,” *Comput. Graph.*, vol. 31, no. 3, pp. 401–409, 2007.

[9] M. J. Sanchez-Vazquez, M. Nielen, G. J. Gunn, and F. I. Lewis, “Using seasonal-trend decomposition based on loess (STL) to explore temporal patterns of pneumonic lesions in finishing pigs slaughtered in England, 2005-2011,” *Prev. Vet. Med.*, vol. 104, no. 1–2, pp. 65–73, 2012.

[10] M. A. Ali, Z. Ahsan, M. Amin, S. Latif, A. Ayyaz, and M. N. Ayyaz, “ID-Viewer: a visual analytics architecture for infectious diseases surveillance and response management in Pakistan,” *Public Health*, Feb. 2016.

[11] R. Maciejewski, P. Livengood, S. Rudolph, T. F. Collins, D. S. Ebert, R. T. Brigantic, C. D. Corley, G. a. Muller, and S. W. Sanders, “A pandemic influenza modeling and visualization tool,” *J. Vis. Lang. Comput.*, vol. 22, no. 4, pp. 268–278, Aug. 2011.

[12] T. Jombart, D. M. Aanensen, M. Baguelin, P. Birrell, S. Cauchemez, A. Camacho, C. Colijn, C. Collins, A. Cori, X. Didelot, C. Fraser, S. Frost, N. Hens, J. Hugues, M. H??hle, L. Opatowski, A. Rambaut, O. Ratmann, S. Soubeyrand, M. A. Suchard, J. Wallinga, R. Ypma, and N. Ferguson, “OutbreakTools: A new platform for disease outbreak analysis using the R software,” *Epidemics*, vol. 7, pp. 28–34, 2014.

[13] A. M. El-Sayed, P. Scarborough, L. Seemann, and S. Galea, “Social network analysis and agent-based modeling in social epidemiology.,” *Epidemiol. Perspect. Innov.*, vol. 9, no. 1, p. 1, 2012.

[14] J. L. Gardy, J. C. Johnston, S. J. Ho Sui, V. J. Cook, L. Shah, E. Brodkin, S. Rempel, R. Moore, Y. Zhao, R. Holt, R. Varhol, I. Birol, M. Lem, M. K. Sharma, K. Elwood, S. J. M. Jones, F. S. L. Brinkman, R. C. Brunham, and P. Tang, “Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak,” *N. Engl. J. Med.*, vol. 364, no. 8, pp. 730–739, 2011.

[15] J. Wang, Y.-S. Guo, G. Christakos, W.-Z. Yang, Y.-L. Liao, Z.-J. Li, X.-Z. Li, S.-J. Lai, and H.-Y. Chen, “Hand, foot and mouth disease: spatiotemporal transmission and climate.,” *Int. J. Health Geogr.*, vol. 10, no. 1, p. 25, 2011.

[16] G. Christakos and X. Li, “Bayesian Maximum Entropy Analysis and Mapping: A Farewell to Kriging Estimators?,” *Mathematical Geology*, 1998. [Online]. Available: http://link.springer.com/article/10.1023/A%3A1021748324917\nhttp://link.springer.com/article/10.1023%2FA%3A1021748324917?LI=true#page-1\nhttp://link.springer.com/content/pdf/10.1023%2FA%3A1021748324917. [Accessed: 02-Mar-2016].

[17] J. Mao, “Why artificial neural networks?,” *Communications*, vol. 29, pp. 31–44, 1996.

[18] L. A. Meyers, “Contact network epidemiology: Bond percolation applied to infectious disease prediction and control,” *Bull. Am. Math. Soc.*, vol. 44, no. 1, pp. 63–86, 2007.

[19] L. L. Ram??rez-Ram??rez, Y. R. Gel, M. Thompson, E. de Villa, and M. McPherson, “A new surveillance and spatio-temporal visualization tool SIMID: SIMulation of Infectious Diseases using random networks and GIS,” *Comput. Methods Programs Biomed.*, vol. 110, no. 3, pp. 455–470, 2013.

[20] M. E. J. Newman, “Spread of epidemic disease on networks,” *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, vol. 66, no. 1, 2002.

[21] K. Sun, J. P. Gonçalves, C. Larminie, and N. Przulj, “Predicting disease associations via biological network analysis.,” *BMC Bioinformatics*, vol. 15, p. 304, 2014.

[22] A. H. Auchincloss, S. Y. Gebreab, C. Mair, and A. V Diez Roux, “A review of spatial methods in epidemiology, 2000-2010.,” *Annu. Rev. Public Health*, vol. 33, pp. 107–22, 2012.

[23] P. A. Zandbergen, T. C. Hart, K. E. Lenzer, and M. E. Camponovo, “Error propagation models to examine the effects of geocoding quality on spatial analysis of individual-level datasets,” *Spat. Spatiotemporal. Epidemiol.*, vol. 3, no. 1, pp. 69–82, Apr. 2012.

[24] B. Davis and C. Carpenter, “Proximity of fast-food restaurants to schools and adolescent obesity,” *Am. J. Public Health*, vol. 99, no. 3, pp. 505–510, 2009.

[25] E. Namosha, B. Sartorius, and F. Tanser, “Spatial Clustering of All-Cause and HIV-Related Mortality in a Rural South African Population (2000-2006),” *PLoS One*, vol. 8, no. 7, pp. 1–8, 2013.

[26] M. P. W. A. Houben, J. W. W. Coebergh, J. M. Birch, C. C. Tijssen, C. M. Van Duijn, and R. J. Q. McNally, “Space-time clustering patterns of gliomas in the Netherlands suggest an infectious aetiology,” *Eur. J. Cancer*, vol. 41, no. 18, pp. 2917–2923, 2005.

[27] K. Bhaskaran, A. Gasparrini, S. Hajat, L. Smeeth, and B. Armstrong, “Time series regression studies in environmental epidemiology,” *Int. J. Epidemiol.*, vol. 42, no. 4, pp. 1187–1195, 2013.

[28] G. Andrienko, N. Andrienko, M. Mladenov, M. Mock, and C. Pölitz, “Discovering Bits of Place Histories from People ’ s Activity Traces,” pp. 59–66.

[29] C. N. Theophilides, S. C. Ahearn, S. Grady, and M. Merlino, “Identifying West Nile virus risk areas: The dynamic continuous-area space-time system,” *Am. J. Epidemiol.*, vol. 157, no. 9, pp. 843–854, 2003.

[30] D. Guo, J. Chen, A. M. MacEachren, and k Liao, “ A Visualization System for Spatio- Temporal and Multivariate Patterns (VIS-STAMP),” *IEEE Trans. Vis. Comput. Graph.*, vol. 12, no. 6, pp. 1461–1474, 2006.

[31] S. Li, S. Dragicevic, F. Anton, M. Sester, S. Winter, A. Coltekin, C. Pettit, B. Jiang, J. Haworth, A. Stein, and T. Cheng, “Geospatial Big Data Handling Theory and Methods: A Review and Research Challenges,” 2015.