Simple Techniques to Predict the Onset of Pandemics

Sunitha Suresh
Vellore Institute of Technology
Tamil Nadu 632014
India

Rajan Chattamvelli
Vellore Institute of Technology
Tamil Nadu 632014
India

ABSTRACT

Accurate identification and prediction of underlying factors of pandemics (like COVID-19) are research priorities, but require elaborate data analysis on a global scale. Due to the large number of mutations of coronavirus that have already ensued during a short span of around 3 years, it is absurd to assume that further mutations will cease to exist. The place and time of future mutations are unpredictable. However, simple visualization techniques can sometimes reveal data peculiarities and provide quick answers to the onset of new mutations, thereby avoiding an expensive analytics run. This paper is aimed at exploring how visualization plots play an essential role to expose hidden relationships among a multitude of variables involved, and how it can be effectively used to predict new waves in selected geographic regions.

General Terms

Analytics, pandemic, vaccination immunity

Keywords

Acute respiratory distress syndrome, Andrew plots, coronavirus, COVID-19, Google Trends, serial correlation

1. INTRODUCTION

COVID-19 pandemic is caused by a virus, code-named SARS-CoV-2, which is a single-stranded RNA genome that has diameter in the range 60-140 nm. After its first appearance in Wuhan, China in December 2019, it has killed around 6.5 million people worldwide as of September 2022. It is expected to be a major killer for months to come due to new variants of it being reported in various countries.

As efficacy of several vaccines decrease drastically after 6 months of vaccination, booster doses are recommended to healthcare and frontline workers, as well as to senile with co-morbidities. These booster shots could become a regular feature of the lives of seniles in developed and developing countries for some more time. Several research studies are underway to ascertain enhanced protection of vaccines or combinations thereof against new variants, efficacy against hospitalization and death, identifying various comorbidities on patient’s death, adjusted dosage levels of vaccines, optimal booster dose repetition intervals, alternate drug delivery routes (like oral or nasal route, 3D printed computer controlled intradermal vaccine patches) and so on.

Medical researchers use several methods and models to check the spread and containment of epidemics and pandemics. WHO has warned that nothing precludes new COVID-19 viral evolutions, as the conditions in several countries are conducive to emergence of such new variants like Omicron and its sub variants. Some of the deadly variants that are yet to come could work against immune response of even fully-vaccinated people.

Complex data visualization techniques can be used to detect new waves, emergence of new virus variants, waning of vaccination immunity among different age-groups, efficacy of booster doses, etc. An advantage of this approach is that it can provide important clues with minimal expenses and ease of work, which can be used as a basis for intricate and all-inclusive epidemiological research studies.

Some of the COVID-19 virus variants like Omicron.ba.2 and BF.7 have the ability to surpass the immunity of even twice (or more) vaccinated persons. Accurate and highly effective procedures are needed to ascertain whether excess deaths are caused by virulentvariants or other factors like air pollution, climatic conditions, and so on. However, this paper demonstrates that simple multivariate visualization techniques can be used as a first-tool to detect abnormalities as and when they occur. This drastically reduces the expense outlay for COVID-19 management and control.

2. RELATED WORKS

The COVID-19 pandemic has caused sweeping changes in the healthcare sector in an incredibly short span of time. It increased the strain on the healthcare system and even brought it to the brink of collapse in some countries. These challenges have helped healthcare management in optimal resource sharing, establishing rapid communication and transport networks and protocols, using real-time monitoring, and providing prompt response to case-load fluctuations. As per some evolutionary biologists, the COVID-19 pandemic started in Huanan seafood wholesale market in Wuhan, China during December 2019 through wild mammals concealing the virus[1]. Nobody assumed that it could become a pandemic and kill millions of people through numerous mutations of the coronavirus. It took several weeks for WHO to warn that it could become a pandemic. Even after it was declared a pandemic, it took several months for the pharma industry to produce an effective medicine. Some of the initial vaccines had low efficacy.

Even those with high potency became dubious after the emergence of novel viral mutations. WHO has already warned that nothing precludes further viral mutations to appear. In fact, conditions are conducive to such mutations in various countries. New “highly infectious” Omicron sub-variants BF.7 (B.1.1.529.5.2.1.7) and BA.5.1.7 were reported in Yantai and Shaoguan cities of Inner Mongolia region recently. As it has high transmission rate and immunity evasion, it is expected to become a new dominant variant and may pose a greater risk. Some of the remote pacific island nations are witnessing the first and second waves now, and are expecting to have more future waves.

Some of the future mutations could even surpass vaccineinduced immunity in old people and those with co-morbidities. Catching such future mutations require expensive laboratory tests, which are lacking in several countries. But it maybe possible to catch such mutations easily by using multivariate visualization techniques. These are inexpensive, fast, reliable and usable in any part of the world. Distinct peaks in such data visualization is an indication of new waves.

Predicting such new waves before they arise is the key to successful control of the pandemic. A progressive growth or an increasing trend in the Andrew plot could be an indication of such a new wave. Further statistical and epidemiological studies can then be initiated.

Andrew plots introduced by D.F. Andrews (1972) is a useful tool to visualize multivariate datadependency in an intuitive way. It has since been successfully applied in a variety of fields ranging from business, finance and stock markets, healthcare and epidemiology, data communication and networks of various sorts. It has gained considerable attention and recognition due to the prevailing COVID-19 pandemic because it can reveal intricate dependencies inherent in quantitative data. In addition, many abnormalities in the data can be detected easily by the plot as well as peculiarities over time brought to the attention. The application of these plots in visualizing COVID-19 pandemic related data could prove to be helpful to medical practitioners, vaccine manufacturers, healthcare agencies and policy makers, to name a few. This can be used to monitor and restraint the spread of COVID-19 variants like delta and omicron. As new avatars of the SARS COVID-19 virus continue to appear due to genetic mutations in various countries, existing vaccines and medications that are mass produced now may also have to be continuously modified structurally to make it effective against the upcoming variants. This has great financial implications as huge wastages can be reduced or totally prevented if timely and critical decisions are made on an ongoing basis. Suppose that there is a k-dimensional data point $x^J = (x_1, ..., x_k)$ for which a function can be defined as follows:

$$f_x(t) = x_1/\sqrt{2} + x_2 \sin t + x_3 \cos t + x_4 \sin(2t) + x_5 \cos(2t) + ... \quad (1)$$

This function is then plotted over the range $-\pi < t < \pi$. Multivariate Andrew Plots (MAP) preserves Euclidean distance between data points in the original space. Thus two points that are close together appear in close proximity in the plot as well for all t values. Any abnormalities in the data can be detected by the plot as a set of data points appear as a set of lines drawn across the plot. Even though this is not the only data visualization technique that can be used for multi-dimensional data, its salient features itself is a motivation for choosing Andrew plots in medical studies. Certain variables might be either spurious or have a real-time influence to the results. Such peculiarities in the data are brought to the attention of the researcher by MAP. This allows policy makers to make timely and critical near-future decisions.

Andrews (1972) discussed its characteristic features which help us in determining what kind of data are to be visualized using it. In [3] Herzberg used MAP to detect time variations in model parameters and to detect outliers. He considered the parameters in a regression model. Through the example of Canadian unemployment figures from 1956 to 1975, he was able tosuccessfully draw conclusions from the generated MAP. Herzberg also established that MAP can be used as a graphical method not only to examine changes over time in the parameters, but also to detect abrupt changes in the observations reflected by changes in the parameters ofthe model over time.

Khattree et. al. (2002) [5], and Chattamvelli (2016) [10] suggested new improvements to Andrew plots and discussed some applications, as also some of the shortcomings. The paper tries to establish that Andrewsplot is a powerful graphical technique for multivariate data. It is not only useful in traditional clustering or outlier detection problems, but can also be very important tool in analyzing experimental data from robust designs as well as in other discrete or descriptive multivariate analysis problems, such as contingency tables and correspondence analysis.

Grinshpun (2016) [2] suggested few more applications of Andrews plots to visualizing multi- dimensional data. In the article, he depicts the possibility of using Andrews plots in big- data. It implies that Andrews plots preserve the information about mean values, distance and dispersion and produce a large number of one-dimensional projections onto the vectors (2 1⁄2, *sin(t)*, *cos(t)*, ...)($-\pi < t < \pi$). Since the distances between Andrews plots are a linear reflection of the distances between data points, two plots that are closer to each other correspond to two points that are close as well. This property proves very useful when representing big datasets like COVID-19 pandemic data.

Motivated by the ongoing global COVID-19 pandemic, which was first identified in December2019 and has caused more than 6.45 million deaths subsequently, we have collected and visualized the data related to major search trends given by Google Trends (GT) that are related to coronavirus of the state of Kerala in India. It matches with the data released by the state from direct RT-PCR and RAT test results.

GT is a publicly available source of online Google search trafficking data[2], which allows users to visualize changes in time series related to the general public's online interest in certain keywords. It's been utilized for the study of epidemiological trends of certain disease outbreaks such as the Middle East Respiratory Syndrome (MERS) epidemic and the Ebola outbreak.The potential use of GT to predict COVID-19 cases or deaths with regard to GT trends and keyword searches of "COVID-19" or any of its symptoms has already been worked upon by many researchers since the outbreak of the pandemic in early 2020[6].

In this paper, we explore how Andrew plots can be used to visualize the data related to COVID-19, and to surmise on possible new waves. We then explore the possibility of serial correlation with an appropriate lag to detect new waves. The second wave started in India during mid-February 2021, after the first wave somewhat subsided in Sep 2020; and the third wave in January 2022, which waned in February-end. Data from Kerala was used for our study because Kerala had the highest number of COVID-19 infections for a long time. We are also comparing the time-series data forecast by the state on positively tested COVID-19 cases and the data obtained from GT using a set of COVID-19 related keywords. Relevant observations are also noted in this paper.

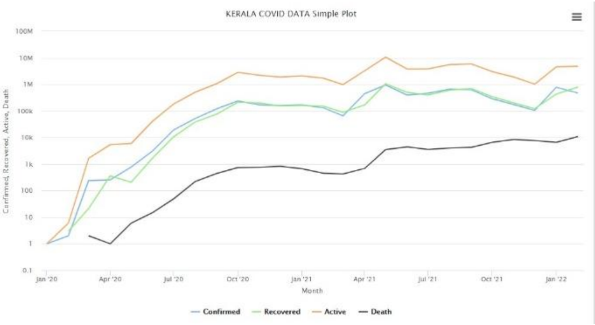


Figure 1: Simple plot for Kerala

The symptoms of COVID-19 starts to appear from 2 to 14 days after infection with a right skewed distribution. A small duration is observed in elder people, especially with co-morbidities. The median time is 4 to 5 days. Most people approach a hospital after one week with a median of 9 days from the onset of symptoms. These people either recover in one week to 10 days' time or progress to the severe stage with acute respiratory distress syndrome (with a median of 11 days for ICU admit). At this stage, the lungs may fail to deliver enough oxygen to vital body parts. Death may occur if the patient is not put on ventilator. Hence the time from infection to death varies between 14 to 20 days in most cases. Thus a serial correlation with lag 10 may be used to detect new pandemic waves and a higher lag used to detect deadly new waves. This of course could vary from country to country or region to region.

The challenge is to identify deaths due to COVID-19 complications from deaths due to other causes like heart attacks, stroke, cancer and other opportunistic illnesses. Age can be used as a critical variable because most deaths due to opportunistic illnesses are observed among the elderly, whereas COVID-19 related deaths occur among middle-aged persons, and to a lesser extent among youth as well (roughly 25% of the deaths occurred in the age-group 0 to 39; and 50% deaths in the age-group 0 to 70). An Andrew plot can throw insights to distinguish deaths due to various illnesses.

3. SOURCE OF THE DATA SET

Inspired by the data visualization in [4], we try to visualize the data obtained from Google trends and the test positivity rate of a particular southern state in India. The state of Kerala consists of 14 districts. The data visualization of a comparison between the district-wise RT- PCR testing results published by the state government and TPR has been depicted in figure 1. Subsequently, we also worked on identifying the keywords related to COVID-19, which could have possibly gained popularity at different stages of the pandemic through GT. A time-series data was obtained initially which had to be worked upon in order to use it for data visualization.

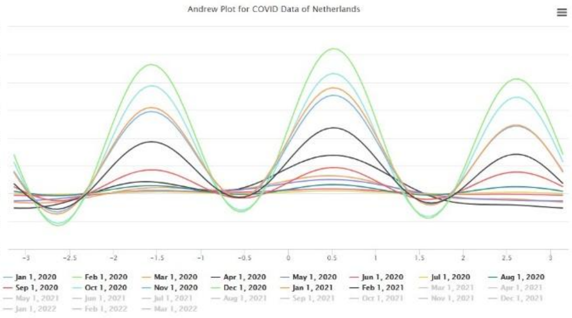


Figure 2: Andrew plot for Netherlands

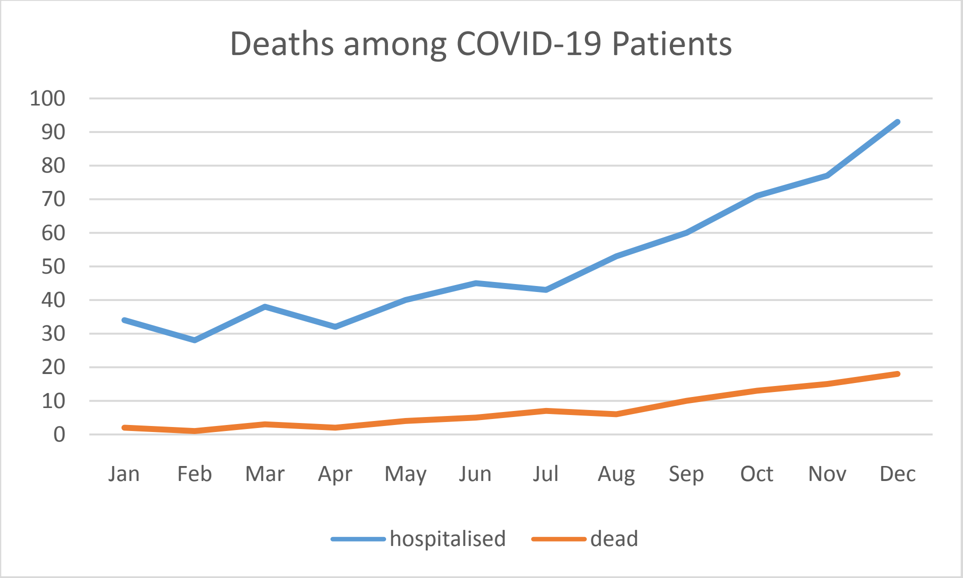
4. PROPOSED SCHEME

Andrew's plot is used for data visualization, as it can easily reduce multi-dimensional datainto two dimensions and visualize it. The nominal attribute is each of the 14 districts of Kerala. There are three more attributes, say, percentage of contribution from RAT (Rapid Antigen Test), percentage of contribution from RTPCR and TPR for a period of 30 days. The figure 1 has each of these attributes plotted. Figure 2 shows the waves observed in Netherlands.

5. CONCLUSION

Visualizing the data related to COVID-19 pandemic over a period of time can easily spot distinct waves in selected geographic regions. Hospital admission data provides a simple and inexpensive method to detect the emergence of new waves, which could be confirmed using elaborate laboratory tests.

Further epidemiological studies could ensue to understand the root causes that lead to new waves.



6. REFERENCES

[1] Andreadis,S., Antzoulatos,G. et.al.,A social media analytics platform visualising the spread of COVID-19 in Italy via exploitation of automatically geotagged tweets, Online Social Networks and Media, Volume 23, 2021, 100134, ISSN 2468-6964, https://doi.org/10.1016/j.osnem.2021.100134.

[2] Andrew D.F, Plots of high dimensional data, Biometrics, 28, 125--136, 1972.

[3] Grinshpun, V. Application of Andrew's Plots to Visualization of Multidimensional Data, International Journal of Environmental and Science Education, 11(17), 2016, 10539-10551

[4] Herzberg A. M., An Example of the use of Andrew's Plots to Detect Time Variations in Model Parameters and to Detect Outlying Observations, Journal of Applied Statistics, 9:2, 146-154, 1982 DOI: 10.1080/02664768200000016.

[5] Jane K. L. Teh,David A. Bradley, et.al., Multivariate visualization of the global COVID-19 pandemic: A comparison of 161 countries, Plos One, 2021, https://doi.org/10.1371/journal.pone.0252273

[6] Khattree, R., Naik, D.N, Andrew plots for multivariate data: some new suggestions and applications, Journal of statistical planning and inference, 100, 411--425, 2002.

[7] Sato, K., Mano, T., Iwata, A. et al., Need of care in interpreting Google Trends-based COVID-19 infodemiological study results: potential risk of false-positivity. BMC Med Res Methodology, 21, 147, 2021. https://doi.org/10.1186/s12874-021-01338-2

[8] Spencer, N. Investigating Data with Andrews Plots, Social Science Computer Review, 2003, 21. 10.1177/0894439303021002010.

[9] Cholette, Pierre A., and Norma B. Chhab. Converting Aggregates of Weekly Data into Monthly Values, Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 40, no. 3, 1991, pp. 411--422, https://doi.org/10.2307/2347521.

[10] Chattamvelli, R. Data Mining Methods, Alpha science international, Oxford, UK, 2016.