

The Art of Geographical Analysis of COVID-19-Related Data



Dimitris Kavroudakis , Sofia Zafeirelli, Panagiotis Agourogianis, and Marios Batsaris

Abstract Almost all sectors of our life have been influenced by COVID-19 pandemic. Informed decision-making regarding pandemic is essential and can be based on credible spatial data. Available COVID-19 spatial data reflect administration areas and in most of the cases used for comparison between countries and regions. These comparisons are tricky and ask for careful consideration of a number of country aspects, especially when the variable in question is dynamic and changes happen very often. Comparison between countries should consider relative numbers (e.g., incidents per capita) and should include information regarding spending for health services. Also, medical provision and climate-related aspects of each country are also important when comparing between countries. Finally, age structure of population is also crucial and needs to be examined. This work illustrates the difficulties when comparing country-data related to COVID-19 pandemic and presents the CRISTINA project. We argue that country-level COVID-19 data ask for standardization in terms of population and geography as well as that correlation of data with country-related characteristics does necessarily not imply direct causation. Finally, this work presents a number of relevant logical fallacies that should be considered when analyzing spatial data of COVID-19 pandemic.

Keywords Spatial analysis · Spatial data · COVID-19 · Logical fallacies · Big data

D. Kavroudakis (✉) · S. Zafeirelli · P. Agourogianis · M. Batsaris
Department of Geography, University of the Aegean, 81100 Mytilini, Lesvos, Greece
e-mail: dimitrisk@aegean.gr

S. Zafeirelli
e-mail: s.zafeirelli@aegean.gr

P. Agourogianis
e-mail: geod17010@geo.aegean.gr

M. Batsaris
e-mail: m.batsaris@aegean.gr

Introduction

COVID-19 pandemic has influenced almost all sectors of social life. Informed decision-making regarding pandemic is essential and can be based on credible data which are geo-referenced in most cases. COVID-19 data reflect administration areas and in most of the cases are used for comparison between countries. These comparisons are tricky and ask for careful consideration of a number of country aspects, especially when the variable in question is dynamic and changes happen very often.

Comparison between countries should consider relative numbers (e.g., incidents per capita) and should include information regarding spending for health services. Also, medical provision and climate-related aspects of each country are also important when comparing between countries. Finally, age structure of population is also crucial and needs to be examined.

This work illustrates the difficulties when comparing country-data related to COVID-19 pandemic. We argue that country-level COVID-19 data ask for standardization in terms of population and geography as well as that correlation of data with country-related characteristics does necessarily not imply direct causation.

Globally, from early 2020, with some billions of people under house restrains, in almost all countries and territories of the world, there have been millions of confirmed cases of COVID-19, including thousands of deaths, reported to the World Health Organization (WHO). The COVID-19 virus, which was described as an “enemy of humanity” by the World Health Organization, has infected about 2.3 millions of people. The number does not reflect reality in the absence of extensive diagnostic tests. The effects of the pandemic are incalculable, and thousands of billions of euros or dollars will be allocated by the governments of the world to deal with the deep economic recession that is coming.

The UN is calling the pandemic the worst crisis humanity has faced since 1945. It combines a deadly disease and an economic recession unprecedented in the recent past. While in China the province of Hubei and its capital, the city of Wuhan, are coming out of quarantine, Italy, the country with the most deaths, has crossed the barrier of 15,000 deaths. And people are starting to wonder about the next day, after the easing of restrictive measures. Is there a risk of a possible future wave of this epidemic? Were governments slow to react? How close to reality is the death toll announced by China (just over 3,000)? What will be the dimensions of the economic crisis?

Global tourism industry has been significantly impacted by various types of crises, particularly the pandemic and terrorist crises. These crises have caused travelers to be wary of visiting new places due to the threat of quarantine, fear of using airports, airplanes, restaurants, museums and archaeological sites, anxiety of not knowing what to do in the event of illness in a foreign country, need for cross-border medical insurance, and difficulty of changing tickets and reservations at both hotels and airlines. Media plays an important role in shaping public perception of such situations and can have a large effect on the global tourism market. For example, fear is often cultivated and escalated by news outlets regarding the severity of diseases, health

effects and death rates, which has resulted in many people canceling their trips and avoiding travel for long periods of time even after the pandemic has passed. Examples of this include the SARS coronavirus (Holmes 2003), avian influenza A (H5N1) virus (Taubenberger and Morens 2006) and pandemic influenza A virus (H1N1) (Zimmer and Burke 2009, p. 1) all of which caused widespread concern and alerted public health services to the risk of rapidly spreading respiratory viruses with pandemic potential.

Digital Epidemiology and Big Data

The increase in the number of electronic mass media, the massive use of the Internet and electronic social media by most of the world's population, as well as the widespread use of smartphones in recent years have led to the creation of new data sources. Also, new algorithmic techniques have enabled the creation of new tools for data processing such as artificial intelligence (AI) (Hamet and Tremblay 2017), machine learning (ML) (Rajkomar et al. 2019) and natural language processing (NLP) (Locke et al. 2021). All of the above contributed to the emergence of a new branch of epidemiology called digital epidemiology. Digital epidemiology embraces the goals of clinical epidemiology but takes a different approach to their implementation. Instead of relying only on data from the health sector, it makes use of these new data sources. These new sources of data, such as electronic social media, are also called sources of "big" data (Big Data) and are characterized by very large volumes of data which have a complex structure and show great heterogeneity. The big challenge facing digital epidemiology is finding the right tools to process and analyze these data, avoid wrong conclusions and misrepresent the right information as a wrong result can have unintended consequences (Salathe et al. 2012; Salathé 2018; Park et al. 2018).

Nowadays, the term "*big data*" has a double meaning that sometimes refers to the data itself and the sources from which they come, while on the other hand sometimes it refers to the processing methods of these data. We use the term "big data" referring to the actual data itself that is produced by the daily life of people who use the internet, the devices or the sensors for various reasons and activities. These reasons include buying/selling, transferring, using electronic services, as well as social media and location-based services. Today, there are eight characteristic words used to describe big data known as "The 8 V's." The V's can be divided into two groups of which the first contains three which are the general characteristics of the nature of "big" data, while the second contains the characteristics that "big" data acquire once entering an information system. More specific, the basic features are:

Volume: Refers to the very large volume of Big Data referring to the difficulty of collecting and processing large amounts of data. **Velocity:** It refers to the speed with which the Big Data are produced from various electronic sources. **Variety:** Refers to the different types of Big Data (such as image, sound, and electronic receipts) and their different structure.

The characteristics they acquire after entering a digital information system include the following topics. Value: It refers to the value that the exploitation of Big Data can offer in various sectors. Veracity (Validity): Refers to the validity of the results that are produced by the utilization of “big” data and the reliability of these data. Variability: Refers to the different forms they can be transformed into, the different models they can be processed with, and the different associations they can be made after entering a system. Virality: Refers to how quickly they can spread through a network to different users. Viscosity: Refers to how much resistance-delay can be observed in the flow-transmission of a certain volume of big data.

“Big” data can also be characterized by the ever-increasing speed of its collection and use in most scientific research today. It is very important to emphasize that the profit of exploiting the “big” data depends entirely on the time interval between its creation and the moment when its use will have given the desired results. The shorter this interval, the more the value of the information derived from it increases. This is true for almost all uses of “big” data and even more so for the field of digital epidemiology, since the time from the outbreak of an epidemic to the moment when the first measures are taken to deal with it is crucial in limiting it. To effectively reduce this time period, the systems that undertake the collection and correlation of these data must be properly designed to achieve the highest possible speeds. Although the health field has long distanced itself from the use of “big” data, the new possibilities that arise with its proper exploitation are enormous. In digital epidemiology, not all sources of “big” data are equally useful, so in the following we will describe the most important of them in terms of their usefulness in monitoring and predicting disease outbreaks, and in terms of the difficulties involved in utilization of data from them (Ali 2019; Bansal et al. 2016; Park et al. 2018; Saecker and Markl 2012).

COVID-19 Data and Countries Comparison

Since the COVID-19 pandemic start, there are numerous sources of official data sharing across the web. There are a number of official sources with statistics, research data and other information about coronavirus (SARS-CoV-2), the disease it causes (COVID-19), the global pandemic and its economic. There are some well-established sources of official COVID-19-related data. One of the most notable source of these data is the World Health Organization, which has posted a special website with the purpose of providing the necessary information on everything related to SARS-CoV-2 (WHO 2022). It also includes international epidemiological statistics, precautionary advice, good practices, debunking relevant widespread fake news, etc. Additionally, the European Center for Disease Prevention and Control (ECDC) has also set up its own special website for information on the novel coronavirus. It also provides international statistics of confirmed cases and deaths (ECDC 2022). Finally, the last official source of COVID-19 information source regarding Greece, is the Greek National Public Health Organization (EODY) has the official information on the

evolution of the pandemic in Greece, useful information, instructions for citizens and businesses of health interest (EODY 2022).

Apart from the above sources of official data, there are a number of related sources about International Epidemiological Statistics. The American Johns Hopkins University has created and daily updates an open data repository with international analytics on the SARS-CoV-2 pandemic. This repository, which is the primary source of statistics for all pandemic-related analyses conducted worldwide, is freely accessible for access (CSSEGIS and Data 2022). In addition, Johns Hopkins University has an interactive map of the pandemic, with data from the same database (Johns Hopkins University 2022). Finally, another source of very informative data related to COVID-19 pandemic is the website “Our World in Data” (Mathieu et al. 2020) which includes a plethora of statistics and informative charts.

There is a need for countries comparison, especially when we need to identify over-spread or under-spread of COVID-19 cases. In a very globalized economy, and with full traveling potentials, we need to understand the various country-related COVID-19 statistics. The research project “CoRona vIrus SpaTial aNalysis” (CRISTINA) (Kavroudakis 2022) examines the time-series progress of COVID-19 events in 177 countries of the world. We collect daily data for events as well as government interventions and associate them with geographical time-series datasets in order to evaluate the rate of change. We also estimate future events based on assumptions of previous events in a time window of 5 days. Future projections are based on assumptions and should only be used for educational reasons. The webpage of this project is the following: www.dimitrisk.gr/covid19.html. The interventions are grouped by type such as isolation measures, transportation measures and economic activity measures.

Data sources for this project include daily COVID-19 data for 177 countries (confirmed cases, deaths, recovered). Also, government measures for 177 countries (lockdown, business measures, transportation measures, education lockdown, etc.). Population data (age groups, sex, population density), health-related data (hospitals, facilities, intensive care units, doctors, health funds) and finally economic data (GDP, economic sectors, etc.). The main research questions of Christina Project are the following:

- What are the effects of stay-home lockdowns on flattening the curve of confirmed COVID-19 cases by country for 177 countries?
- How did similar lockdown measures affect the curve of confirmed COVID-19 cases by country for 177 countries?
- What is the relationship between: % of GDP for health and % of confirmed population above 65 years old by country for 177 countries?
- What is the contribution of various measures by country: home lockdown, business closure, transportation restrictions 1 and 2?
- What is the contribution of average temperature on COVID-19 cases per week, month?

The results of this project can be found on the website. Some of the most notable results include the following points. The following results are very suitable on the

basic understanding on comparison between countries and the avoidance of possible logical fallacies. The following Fig. 1 depicts the relationship between: the ration of deaths over confirmed cases by time. We use Day since first event by country as the main time unit. It is clearly visible that Italy, UK and Spain lines are rising after 58th day, much more than any other country. Germany, on the other hand, shows very low ratios on the same period. On this basis, Fig. 2 depicts confirmed cases of COVID-19, since the day of 1st event. Germany's numbers are among the top countries for this variable. It is now more than obvious that we can use various COVID-19 data sources to focus on different top countries. Germany's numbers are very small in the first figure while they are very high at the second figure. Variable selection can be a form of "cherry-picking" approach while COVID-19 data analysis, for countries comparison. Selection of specific variables which fit our assumptions can be quite confusing if not deceiving. To avoid such misconceptions, we argue on favor of multiple variable evaluation on country comparisons. This is the use of all available variables and then use some sort of voting system before ranking country's progress.

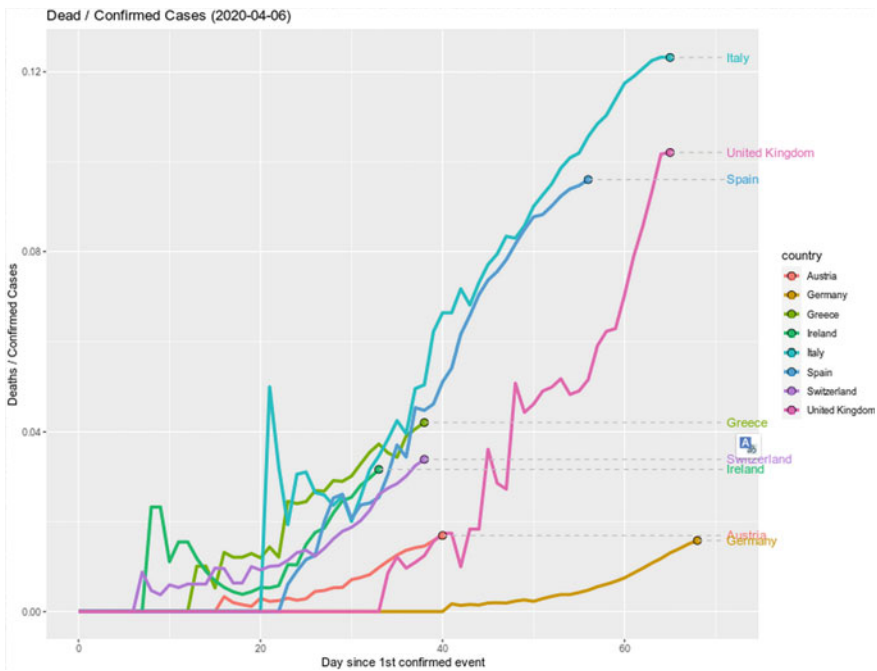


Fig. 1 Deaths over confirmed cases of COVID-19, since the day of 1st event. Germany is almost in the bottom of all other countries

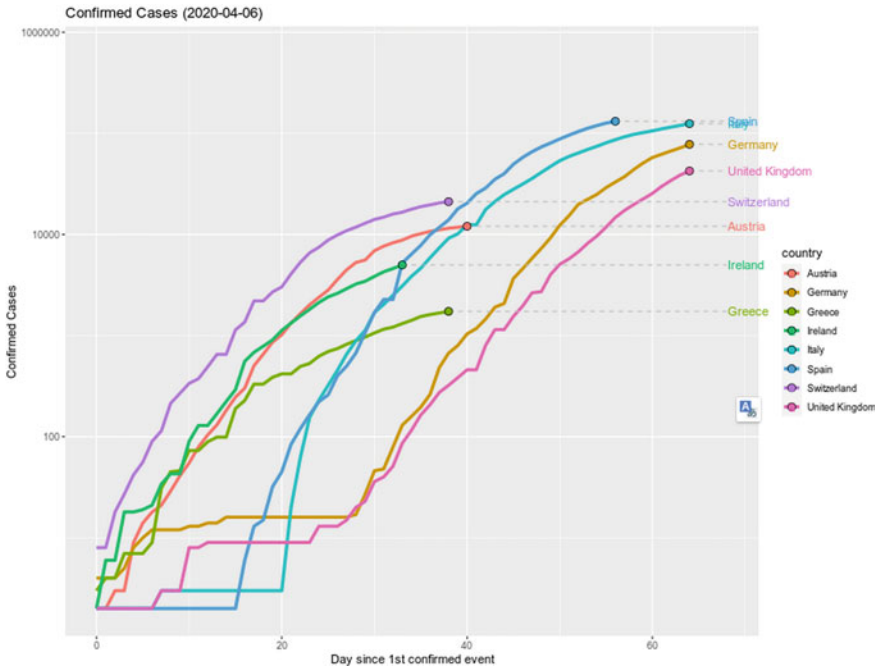


Fig. 2 Confirmed cases of COVID-19, since the day of 1st event. Germany is the 3rd country

Excessive Deaths by Age Group

Another very interesting topic from the CHRISTINA project is the analysis of excessive deaths by age group. This is the evaluation of death counts against previous non-COVID-19 years for each country. This approach can offer a more reliable way of comparison between countries especially when we consider temporal variations. Following Fig. 3 depicts excessive deaths by age group comparison between Greece (left) and Italy (right). Green and blue periods (bars in x axes) represent the two lockdown events in 2020. Black line indicates average deaths by week for the years 2005–2019. Red line is showing deaths by week for 2020. Finally, green and blue periods (x axes): the two lockdown events in 2020. It is more than clear that Italy (right column) has shown some excessive deaths across all age groups. The baseline for the excessive deaths calculations is the deaths of the last 10 years, for each age group. This measure is a comparative measure against the actual numbers of each country for the previous years. It is a consistent measure that is not influenced by temporal variations and is using a baseline of numbers before the COVID-19 pandemic.

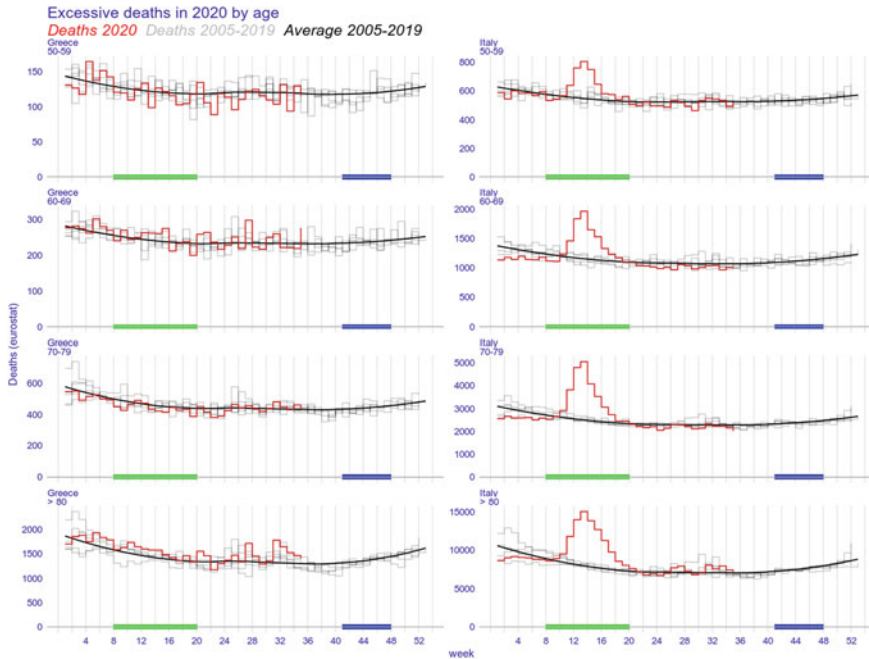


Fig. 3 Excessive deaths be age group comparison between Greece (*left*) and Italy (*right*). Green and blue periods (bars in *x* axes): the two lockdown events in 2020. *Gray lines*: deaths by week for the years 2005–2019. *Black line*: average deaths by week for the years 2005–2019. *Red line*: deaths by week for 2020

Logical Fallacies

Some of the most important logical fallacies when comparing COVID-19 data between countries are presented here in this part of the chapter. One of the most significant is *Cherry Picking*. It refers to selecting results that fit your claim and excluding those that don't. It is also related to selecting specific variables to present according to the argument we need to support. *Data Dredging* is another very interesting fallacy which refers to the act of repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance. When comparing data between countries, a very common logical fallacy is *False Causality* (spurious correlation) which is falsely assuming when two events appear related that one must have caused the other. Also, when referring to COVID-19 sample data, *Sampling Bias* should also consider as one of the most essential logical fallacies. It is referring on drawing conclusions from a set of data that isn't representative of the population you're trying to understand.

Another significant misconception regarding probabilities of events is the *Gambler's Fallacy*, which is mistakenly believing that because something has happened more frequently than usual, it's now less likely to happen in future (and vice

versa). This is quite relevant to COVID-19 comparisons when we falsely believe that because some rises and falls of the numbers (positive cases) are less likely to happen in the future. Also, *Simpson's Paradox* is very relevant in countries and regions comparisons. More specific when comparing intra-countries events, we may sometimes misunderstand the big picture for the total country. This paradox is referring to the case when a trend appears in different subsets of data but disappear or reverse when the groups combined. Finally, equally important is the *Publication Bias* which is also prominent in COVID-19-related studies. More specific, this is referring to the fact that interesting research findings are more likely to be published, distorting our impression of reality.

Cherry-Picking in Geography

Cherry-picking logical fallacy regarding COVID-19 geographic data can be a problematic practice as it can result in biased (or even incomplete) data that misrepresent the true state of the pandemic in a geographical area. Suppose a researcher wants to cherry-pick COVID-19 data for a specific city in the United States, such as, Los Angeles. The researcher can find a reliable source of COVID-19 data, such as the website of the Centers for Disease Control and Prevention¹ or the World Health Organization.² Once the researcher has located the data source, he can navigate to the page that provides data on COVID-19 cases in the United States and look for the data for the state of California. Next, he can locate the data for Los Angeles County, which is the largest county in California and includes the city of Los Angeles. The researcher can find data on the number of confirmed COVID-19 cases, deaths and other relevant metrics for Los Angeles County. However, cherry-picking this data for Los Angeles alone could give a misleading picture of the state of the pandemic in the city, as it does not take into account the wider context of the county, state, or even country. It is important to analyze the geographical data for the city in relation to the data for the broader geographic region and to consider factors such as population density, demographics and/or other social and economic factors that may affect the spread of the COVID-19 virus. It is therefore important to take into account the broader context of the pandemic to avoid misleading conclusions.

Data Dredging in Geography

Data dredging is the practice of selectively analyzing spatial data to find spatial patterns that appear significant, but are actually due to chance. This can lead to false conclusions and incorrect interpretations of the data. Suppose a researcher wants to

¹ <https://www.cdc.gov>.

² <https://www.who.int>.

determine if there is a correlation between temperature and the number of COVID-19 cases in Brazil. He starts by collecting data on the average temperature and the number of COVID-19 cases for each state in Brazil. The researcher then analyzes the data and finds that there appears to be a negative correlation between temperature and the number of COVID-19 cases, meaning that as the temperature increases, the number of cases decreases. However, this conclusion is based on a selective analysis of the spatial data and ignores many other factors that could be driving the spread of the virus, such as population density in Brazil, demographics, public health measures and other social and economic factors. It is also possible that the apparent correlation is simply due to chance, as correlations can appear by random chance in any dataset. In general, in order to avoid data dredging and draw valid conclusions, it is important to use a rigorous and systematic approach to spatial data analysis and to take into account all relevant factors that may affect the spread of the virus in a country. This can involve using multiple regression analysis to control for confounding variables and conducting sensitivity analyses to test the robustness of the results to different assumptions and model specifications.

False Causality in Geography

False causality, in geospatial sciences, is a type of error that occurs when a correlation between two spatial variables is assumed to indicate a causal relationship, even though there may be other spatial factors that are responsible for the observed relationship. Suppose we want to investigate the effect of air pollution on the number of COVID-19 cases in Athens, Greece. We collect data on air pollution levels and the number of COVID-19 cases for the city over a period of several months and analyze the data to find a positive correlation between air pollution levels and the number of cases. We may conclude that air pollution is causing an increase in COVID-19 cases in Athens, and suggest that reducing air pollution levels could help to mitigate the spread of the virus.

However, this conclusion is false, as there may be other factors that are responsible for the observed correlation, such as population density, demographics and/or public health measures. In order to avoid the pitfalls of false causality, it is important to use a rigorous and systematic approach to data analysis and to consider all relevant factors that may affect the relationship between variables. While it is important to investigate the relationship between COVID-19 and environmental factors such as air pollution, it is important to use caution when drawing causal inferences from observational data and to consider all relevant factors that may be responsible for the observed relationship.

Sampling Bias in Geography

Sampling bias is a type of error that occurs when samples are not representative of the population being studied. Sampling bias could affect the spatial analysis of COVID-19 data. For example, if we want to investigate the relationship between COVID-19 cases and income levels in a city, we first collect data on the number of COVID-19 cases and the median income for each neighborhood in the city. Then, we analyze the data to find that there is a negative correlation between “income levels” and COVID-19 cases. However, this conclusion may be biased, as the sample of neighborhoods we selected may not be quite representative of the entire city. It is possible that we may have selected only neighborhoods with higher income levels that have lower population densities and better access to healthcare facilities, which in turn may be responsible for the observed negative correlation. To avoid sampling bias, it is important to use a representative sample of data that accurately reflect the population being studied. This can be achieved using random sampling techniques or stratified sampling methods to ensure that all segments of the population in the city are represented in the sample.

Gambler’s Fallacy in Geography

The gambler’s fallacy is a type of cognitive bias that occurs when individuals assume previous random events will affect the outcome of future events, even though two events are statistically independent. For example, when analyzing the number of COVID-19 cases in a municipality over time, we may notice that there have been several consecutive days of increasing cases, and assume that this trend will continue into the future. However, this assumption is a form of the gambler’s fallacy, as each day’s COVID-19 case count is independent of the previous day’s count and there is no statistical basis for assuming that the trend will continue. It is therefore important to use statistical methods to analyze data and account for the effects of randomness and variability. This can involve using time-series-analysis techniques to model trends and possible seasonal patterns in spatial data as well as conducting hypothesis tests to determine the statistical significance of observed patterns.

Simpson’s Paradox

Simpson’s paradox is a type of statistical paradox occurring when a trend appears in different groups of data, but disappears (or reverses) when groups are combined. For example, when investigating the relationship between COVID-19 cases and ethnicity in a city, initially we collect data on the number of cases and ethnic composition of each neighborhood. Then, we may find that in each neighborhood, the number

of cases is higher among a specific ethnic group compared to residents of other ethnicities. However, when combining data across all neighborhoods, we may find that the opposite trend appears: the overall number of cases is higher among residents of other ethnicities compared to this specific ethnic group. This reversal of trend is a form of Simpson's paradox, as the relationship between ethnicity and COVID-19 cases changes when the data are aggregated at different spatial level. It is important to consider the underlying factors that may be driving the observed patterns in the data, and to use appropriate statistical methods to control for confounding variables. This can involve using regression analysis to model the relationship between ethnicity and COVID-19 cases while controlling for other variables such as age, income and access to health care.

Publication Bias

Publication bias is a type of bias occurring when the availability or publication of research results is influenced by their statistical significance. Suppose we conduct a systematic review of the relationship between COVID-19 cases and air pollution in Athens, Greece. We identify several studies that have investigated this relationship, but notice that most of the studies have reported significant positive associations between air pollution and COVID-19 cases, while only few studies have reported non-significant or negative associations. This pattern may be a form of publication bias, as studies with non-significant (or even negative) findings may be less likely to be published or included in the review, leading to an over-representation of only positive findings in the literature. To avoid publication bias, we should conduct a comprehensive search of the literature and include all relevant studies, regardless of their statistical significance. This may involve using search strategies that are not limited by language, geography or publication status, and using methods to assess the risk of bias in individual studies. It is important to use caution when interpreting the results of systematic reviews or meta-analyses, and to consider the potential impact of publication bias on the observed patterns in the spatial data.

Conclusions

Digital epidemiology is a cutting-edge field of epidemiology that is essential for improving people's health and quality of life. It is still in its early stages and may take some years to become fully realized. Currently, systems that use large amounts of data for digital epidemiology are being developed and can only detect and track the spread of contagious diseases; they have very limited forecasting abilities. Additionally, these systems typically rely on receiving confirmed data from other sources, so they cannot provide any warnings before the data are verified.

The continuous development of technology, the creation of faster computing systems as well as improvements in the field of artificial intelligence, within the next few years, for sure will offer new possibilities in the field of digital epidemiology. Thus, it will become strong the creation of new systems that can process and confirm information autonomously and faster than other epidemiological agencies. Also, it is possible to create forecasting systems that will be able to predict the advent of a new epidemic, even from the first cases, while there will also be possibility of creating spatial models that will predict its spread. By using such systems, health agencies will now be able to limit and deal with outbreaks of new communicable diseases in their very early stages, before they get out of control. An example which points to the need to create advanced systems in the field of digital epidemiology is the coronavirus pandemic, which it could be dealt with much faster and with better methods if we had a spatial model for predicting of its spread. The development of the field of spatial epidemiology and, in general, of the field of processing and analysis of “big” spatial data can lead to many improvements in the health and quality of life of the world’s population. However, there are always some risks when collecting and processing data from electronic sources.

Logical fallacies can be found in almost any scientific field, including geography. Geographic research often involves complex spatial data, analysis and interpretation, which can make it vulnerable to a variety of cognitive biases and logical errors. This work illustrated some of the most prominent logical fallacies in geospatial science and more specific when dealing with COVID-19-related spatial data. It is important for geographers to recognize the limitations and uncertainties of spatial analysis of data and to be open to revising their conclusions as new evidence emerges.

References

- Ali AH (2019) A survey on vertical and horizontal scaling platforms for big data analytics. *Int J Integr Eng* 11(6):138–150. Retrieved from <https://publisher.uthm.edu.my/ojs/index.php/ijie/article/view/2892>
- Bansal S, Chowell G, Simonsen L, Vespignani A, Viboud C (2016) Big data for infectious disease surveillance and modeling. *J Infect Dis* 214:S375–S379. <https://doi.org/10.1093/infdis/jiw400>
- CSSEGIS and Data (2022) COVID-19 data repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.
- ECDPC (2022) European Centre for Disease Prevention and Control COVID-19 [WWW Document]. European Centre for Disease Prevention and Control. <https://www.ecdc.europa.eu/en/covid-19>. Accessed 29 Nov 2022
- EODY (2022) Greek National Public Health Organization Covid-19 [WWW Document]. <https://eody.gov.gr/neos-koronaivos-covid-19/>. Accessed 29 Nov 2022
- Hamet P, Tremblay J (2017) Artificial intelligence in medicine. *Metabolism* 69:S36–S40. <https://doi.org/10.1016/j.metabol.2017.01.011>
- Holmes KV (2003) SARS coronavirus: a new challenge for prevention and therapy. *J Clin Investig* 111(11):1605–1609. <https://doi.org/10.1172/JCI18819>
- Johns Hopkins University (2022) Web map of COVID-19 [WWW Document]. <https://gisanddata.maps.arcgis.com/apps/dashboards/bda7594740fd40299423467b48e9ecf6>. Accessed 29 Nov 2022.

- Kavrouidakis D (2022) CRISTINA project [WWW Document]. <https://www.dimitrisk.gr/covid19.html>. Accessed 29 Nov 2022
- Locke S, Bashall A, Al-Adely S, Moore J, Wilson A, Kitchen GB (2021) Natural language processing in medicine: a review. *Trends Anaesthesia Critical Care* 38:4–9. <https://doi.org/10.1016/j.tacc.2021.02.007>
- Mathieu E, Ritchie H, Rodés-Guirao L, Appel C, Giattino C, Hasell J, Macdonald B, Dattani S, Beltekian D, Ortiz-Ospina E, Roser M (2020) Coronavirus pandemic (COVID-19). Our world in data. Retrieve from: <https://ourworldindata.org/coronavirus#citation>
- Park H-A, Jung H, On J, Park SK, Kang H (2018) Digital epidemiology: use of digital data collected for non-epidemiological purposes in epidemiological studies. *Health Inform Res* 24(4):253–262. <https://doi.org/10.4258/hir.2018.24.4.253>
- Rajkomar A, Dean J, Kohane I (2019) Machine learning in medicine. *New Engl J Med* 380(14):1347–1358. <https://doi.org/10.1056/NEJMra1814259>
- Saecker M, Markl V (2012) Big data analytics on modern hardware architectures: a technology survey. In: Aufaure MA, Zimányi E (eds) *Business intelligence. eBISS 2012. Lecture Notes in Business Information Processing*, 138. Springer, Berlin, Heidelberg, pp 125–149. https://doi.org/10.1007/978-3-642-36318-4_6
- Salathé M (2018) Digital epidemiology: what is it, and where is it going? *Life Sci Soc policy* 14:1–5. <https://doi.org/10.1186/s40504-017-0065-7>
- Salathe M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, Campbell EM, Cattuto C, Khandelwal S, Mabry PL (2012) Digital epidemiology. *PLoS Comput Biol* 8(7):e1002616. <https://doi.org/10.1371/journal.pcbi.1002616>
- Taubenberger JK, Morens DM (2006) 1918 influenza: the mother of all pandemics. *Emerg Infect Dis* 12(1):15–22. <https://doi.org/10.3201/eid1201.050979>
- WHO (2022) Coronavirus disease (COVID-19)—World Health Organization [WWW Document]. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>. Accessed 29 Nov 2022
- Zimmer SM, Burke DS (2009) Historical perspective—emergence of influenza A (H1N1) viruses. *New Engl J Med* 361(3):279–285. <https://doi.org/10.1056/NEJMra0904322>