Open Public Data and Early Factor Analysis in a Developing Public Health Event

Serge Dolgikh

National Aviation University, 1 Lubomyra Huzara Ave, Kyiv, 03058, Ukraine

Abstract

An important role of timely public availability of data in the early analysis and formulation of hypotheses in developing public health events, such as infectious epidemics has been highlighted in many results. The recent pandemic demonstrated perhaps the first example where serious attempts were made to present consistent and detailed information, on the international scale and in near real-time, for a developing major public health event, with important and numerous implications for the formulation of responses and policies. In this work we analyze characteristics of publicly available data, including timeliness; granularity i.e., level of detail; consistency between different reporting jurisdictions and others; as well as issues and problems with processing publicly available information for early analysis and formulation of early hypotheses. Based on the experience and the analysis in this work we attempt to formulate expectations and conditions for the collection and publication of publicly available data for future public health and more generally, events with potentially high societal impact.

Keywords 1

Data collection, public data, statistical analysis, factor analysis

1. Introduction

Whereas collection and publication of statistics in many areas and aspects of society, politics, demographics and economy have been a known and relatively common occurrence for some time, the recent onset of the Covid-19 global pandemic caused collection, publication and availability of data describing both local and global spread and impacts of the developing epidemic in near real-time. The availability of these data has been invaluable in the early analysis of multiple potential factors of significance, formulation of early hypotheses, analysis, evaluation and feedback on public policy decisions.

In this work we attempted to address the benefits as well as caveats of working with publicly available data; methods and practices of collection and preparation of the data for the analysis, examples where public "ecological" data was used in formulation of early hypotheses; and the need for a thorough and comprehensive follow up by more detailed studies to confirm or reject them. We also discuss possible directions of evolution on the preparation and publication of epidemiological data in the public domain and ways to make it more informative and useful in the analysis.

Here the term "ecological data" signifies, not to be confused with the subject of ecology, the data that is available or can be obtained directly from observations of the system, entity or process being examined, in contrast, for example, to controlled studies with fixed sets of observed subjects to test formulated hypotheses. There are numerous instances and well-established practices of data being collected and published in the open to general public domains and format on a broad range of subjects, including sociology, economics, consumer behavior and others, but to the author's best knowledge, the recent Covid-19 pandemic was one of the first examples in the history where a sustained attempt to

IDDM'2023: 6th International Conference on Informatics & Data-Driven Medicine, November 17 - 19, 2023, Bratislava, Slovakia
EMAIL: sdolgikh@nau.edu.ua (A. 1)
ORCID: 0000-0001-5929-8954

collect and publish consistent data directly related to a major developing epidemiological event have been made.

The sources of this information were both diverse and numerous: from national and local health agencies and offices, to research institutions, media and information services such as Google, Facebook and others. Along with massive volumes of information that were available for early analysis, almost in real-time, it highlighted a number of issues and challenges that will be discussed further in this work.

Generally, a "good" ecological dataset, that is, of high or at least, sufficient quality that can be used without significant modification and / or preparation in statistical analysis, can be characterized by the following features:

- Usability: observable factors are clearly interpretable and straightforward to use in the analysis;
- Representativity: contains a sufficient set of samples, reasonably representative of the real distribution;
- Consistency: reporting of observable factors is consistent between different data points; alternatively, allows obtaining consistent data points via a well-defined process;
- Detailization (granularity): informative observable factors describe observations of the event or phenomena at data points in sufficient detail for different forms of analysis.
- Breadth: a sufficiently large range of data points is presented to support the expectation that the data points in the set describe the entire or most of the variable and value range of the unknown distribution.

Further in this work, the following terminology will be used:

- Distribution: in general, a variable that describes values of certain factors of interest in the domain of interest; for example, the height, weight, immunological, epidemiological, and so on factors in the general population.
- Dataset: a set of data composed of data points each represented by a set, e.g., a numerical vector of observable factors. A supervised dataset additionally associates data points with a (set of) factors of interest.
- Factors:
- Observable factors describe parameters of the data in the set that can be obtained from observations, e.g., by measurement.
- Factors of interest describe certain characteristics of the observed event or phenomena that are of interest in the study.
- Informative factors are associated with the observable factors and allow to establish a clear relationship between the observable factors and the factors of interest.

The problem of *factor analysis* [1] then can be defined as establishing the relationship between the factors of interest and the observable/informative factors, based on the information contained in a representative set of data points.

2. Literature Review

Publication in the open-to-the-public domains of sociological, economic, demographical and other information has been common and regular in recent decades with sites and services such as Statista; Worldometer; Google and many others [2-4] offering access to collected and partially processed data on a wide range of issues and topics. As quoted:

"Worldometer is a provider of global COVID-19 statistics for many caring people around the world. Our data is also trusted and used by the UK Government, Johns Hopkins CSSE, the Government of Thailand, the Government of Pakistan, the Government of Sri Lanka, Government of Vietnam, The Financial Times, The New York Times, Business Insider, BBC, and many others." [2]

"Statista is a German online platform specialized in data gathering and visualization, which offers statistics and reports, market insights, consumer insights and company insights in German, English, Spanish and French". [3]

"The Google Health COVID-19 Open Data Repository is one of the most comprehensive collections of up-to-date COVID-19-related information. Comprising data from more than 20,000 locations worldwide, it contains a rich variety of data types to help public health professionals, researchers, policymakers and others in understanding and managing the virus." [4]

Multiple datasets have been collected over the years and made available in the public domain for independent analysis of sociological, public and economic behavior [5], resulting in active research and multiple publications.

A precipitous onset of the global Covid-19 pandemic stimulated the collection and the publication of information on the development of the epidemic both globally and on the national and subnational level, providing citizens and research professionals with near-real-time view of a developing major public health event. Availability of timely and accurate information was crucial for the formulation of sound and effective responses, advisories and policies by above-national and international bodies like the World Health Organization, EU Center for Disease Control and Prevention [6,7], national health administrations [8,9] and lower-level subnational public health jurisdictions, such as regional, provincial, etc., local and municipal health offices and authorities.

Availability of such data made possible a wide range of research in different aspects and from different perspectives, including: examination and formulation of early hypotheses on potential factors influencing the development and severity of the epidemic [10,11]; investigation into the origins of the epidemics [12]; development patterns and scenarios [13]; evaluation of the effectiveness of policies and responses [14] and many other aspects and directions. All in all, it can be concluded that the availability of such data in the open access was a positive factor in facilitating research and formulation of sound and effective policies.

At the same time, certain challenges and shortcomings in the practice of publication of direct epidemiological information can be mentioned. Accumulating data from many diverse sources it was extremely challenging to maintain a set standard of accuracy; disparities and availability of information from lower-level reporting jurisdictions had significant differences and inconsistencies [15]; granularity and connectedness of data, that is, the ability to navigate between the reporting levels with preservation of some level of accuracy was not, generally, possible with the public data; wide variation in reliability and accuracy of the data between and even within reporting jurisdictions. All these factors may have contributed to the challenges in the early analysis of the data obtained from public sources with the possibility of reducing the confidence and skewing the results of the analysis.

As the initial, challenging phase of the pandemic appears to be over, the time may be right to review both the successes and challenges of the early-stage analysis of public data, including in the collection and publication of the openly accessed epidemiological data. Improvements may include several factors influencing the content; accuracy, consistency and reliability of the published data; level of detail; usability for analysis and other essential characteristics of data made available in public access, with the potential to facilitate early analysis of developing events, formulation of hypotheses, correctness and confidence in the conclusions.

3. Publicly Available Data in Early Factor Analysis and Formulation of Hypotheses

Let us consider a general case of data that can be a sampling of an unknown distribution D, $W = \{ P, F \}$ where $P = \{ p \}$: points of observation, such as individual subjects in medical trials; cities or social groups in sociology etc; $F = \{ f \}$, the observable factors. We would like to investigate and if possible, establish the relation R between certain factor(s) of interest, K and the observable factors of data points, p.

$$K(p) = R(F(p)) \tag{1}$$

The relationship in (1) represents the classical problem of factor analysis [1]: establishing the relationship R between the factor(s) of interest and the observable characteristics of the data points based on the data W.

In this work, we will focus on the methods and approaches to how the sampling of observable data W can be produced in minimal time, accessed and used in early factor analysis research. It is understood that the factor of time can be essential in some situations and applications, of which a developing major public health or social event can be a prime example.

3.1. Sources of Public Ecological Data

The practice of collection and publication of statistical data in the open publicly accessible format has been in place for a considerable time. Notably, information services like Worldometer, Statista and others provided statistical information from multiple national and subnational jurisdictions on a wide spectrum of topics, including economic characteristics, social factors and conditions and others. From the onset of the pandemic in the early months of 2020, a wide range of sources emerged for publicly available data related to the progress and the impacts of the pandemic. Specifically, these sources included:

- International organizations: WHO, EU Health and Food Safety Commission and others;
- National and subnational public health offices, agencies and administrations in most national jurisdictions;
- Local and municipal health units and governments;
- Media companies and services;
- Research institutions, foundations, universities [16,17]
- Specialized statistics and information sites and services [2,3]
- Information and social networks [4]

The composition and content of the information published in open public format is described in Table 1.

Table 1

Composition, content and other characteristics of publicly available Covid-19 data

Type	Example	Scope
International, health	WHO, EHA	Advisories, general information
National, subnational health authorities	CDC, NHS, Health Canada, etc.	National and subnational statistics, advisories, policy
Local and municipal health authorities	Local, municipal health offices	Local statistics, advisories and information, local health policy
Media	Many national, regional and local media sources	National and subnational statistics, advisory, information, stories
Information sites, social networks	Worldometer, Statista, Google	General information, national, subnational and other statistics

Statistical data was published in a variety of layouts and formats. Some of the information was available in formats friendly to statistical analysis such as Excel, csv, xml etc. However, in many cases issues with consistency in the collection and presentation of data were encountered that were notable complicating factors in the analysis of the data.

3.2. Early Factor Analysis with Public Ecological Data

A study that is based on analysis of any data begins with:

- Definition of the objectives of the study;
- Choosing the method or methods to pursue it;
- Establishing the source and obtaining the data;
- Preparing the data for the analysis.

The open availability of the data in the public domain can significantly improve and speed up locating the data that will be used for the study; in comparison to traditional controlled studies, in some applications and types of analysis data can be obtained almost instantly, saving significant lead time in accumulation and collection. However, it in no way means that working with ecological data is free from problems and challenges. Several of them are outlined below.

Research-friendly format: from the outset, many sources of open-to-the-public ecological data are oriented toward informing the public and are not necessarily research-friendly. It can mean that the data is made available in a difficult-to-process format such as HTML, plain text, graphics, etc. and had to be

extracted and compiled manually. Some sources, including public health authorities, research institutions and others (as listed in Section 3.1) may have strived to provide data in research-friendly formats, including csv, XML and other structured data formats.

Consistency: the data is usually obtained from national and subnational public health sources; consistency in collecting and processing data cannot be assured; specifically, it is known that different criteria have been applied by national reporting authorities in the collection of case statistics, and possibly other data.

Accuracy: a consistent accuracy standard cannot be assured with data obtained from different jurisdictions; an examination of different sources of data may be warranted.

Breadth and depth of data: availability of data representing all characteristic groups/regions in the distribution; possibility to trace data to lower-level sources, for example, from national to national, local and municipality levels.

Yet, despite these challenges, publicly available ecological data can be used in the early analysis of the trends in the development of the event and formulation of hypotheses that can be confirmed or ruled out at a later time, when more data has been collected and more comprehensive analysis can be performed. In this section, we will provide some examples of early factor analysis based on public data collected from open sources.

3.2.1. Collecting and Processing Public Ecological Data

The collection of data begins with identifying the source that satisfies the objectives of the study. Some factors of primary importance are:

1. **Representativity** or variance of the data: sufficient for the objectives, for example, to examine the association/correlation of the factor of interest with observable factors among all essential regions of the distribution (for example, groups of population).
2. **Expression**: observable factors describe data points to the level of detail sufficient for the objectives of the analysis.
3. **Accuracy**: the data is expected to be accurate within the constraints of the analysis.
4. **Consistency**: accuracy and other characteristics of collection are expected to be within a reasonable margin of variance, acceptable for the objectives of the analysis.

An example of the last point in the list above, the consistency of data can be given by an ecological analysis of the factor of interest, for example, infection rate among subnational regions. The data published by regional health units can depend on methods of collection of data, reporting standards and practices and so on; and if these procedures are not harmonized or standardized between the regional reporting offices, the consistency of the resulting national data cannot be assured and the accuracy of the analysis may suffer.

Once the data has been accessed and compiled into a set of data points P described by observable factors $F(P)$ possibly, with the recorded values of the factor of interest: $(W(P, F), K(P))$ preparation of the data for the analysis can begin. Rather than entering raw recorded data directly into the selected method(s) of the analysis, it can be essential for the accuracy of the analysis that the observable factors in the dataset are processed and perhaps, transformed to provide a consistent and uniform view of the observation. The methods and objectives of preprocessing are described below.

Scaling (linear and non-linear): statistics of infectious epidemics are often collected and presented in incidence or case count statistics (such as cumulative, interval, etc.). Understandably, comparing total recorded counts in jurisdictions with one million vs. 100 million population would make little sense. A common practice in statistics is to transform the counts into per capita factors that can provide a better basis for comparison, though this practice is not without caveats as briefly discussed in the next section. The basic law of linear scaling is $Cr \rightarrow \frac{Cr}{N_{pop}}$, where C_r: total (raw) count; N_{pop}: population of the region of the recorded total count.

Temporal adjustment: in the analysis of interval values, accumulated over q certain period, it is essential that data points are compared over the same or similar intervals. For this reason, some adjustments in preparation may be needed.

Accuracy and consistency adjustment: where multiple sources of the same or similar factors exist, they can be verified to improve the accuracy and consistency of the data.

Normalization: some standard methods of factor analysis require a standard transformation of the data to produce normalized data as input to the methods of analysis [18].

Feature analysis: covariance analysis of distributions of observable factors can indicate observable factors that are dependent or correlated. Using such factors unintentionally may amplify the significance of these factors and skew the analysis. A decision has to be taken whether to keep such factors in the dataset or exclude them to avoid the amplification effect.

Derived factors: in some problems and studies, invariant factors derived from the observed ones can introduce additional perspectives in the analysis. An example of an analysis with such a derivative factor is given in the next section.

3.2.2. Early Factor Analysis and Formulation of Hypotheses

In this section we will consider some examples of using ecological public data in the early analysis of the factors in the Covid-19 pandemic.

1. **Bacille Calmette-Guérin (BCG) Immunization Correlation with Lower Covid-19 Impact Hypothesis**

In the early days of the pandemic, based on early case and impact statistics, the hypothesis of correlation of lower observed Covid-19 impact, measured in observable factors of incidence; morbidity and mortality was proposed in [19]. Further studies indicated a certain level of statistical significance of the correlation hypothesis in the period up to the introduction of vaccines.

Currently, as a result of several statistical and controlled studies, strong evidence in favor of the correlation hypothesis has not been established. One possible explanation can be surmised as related to the preprocessing, namely the scaling of the data discussed in the preceding section. It can be worth a brief discussion here, even to underline possible caveats in the direct analysis of early ecological data.

In an analysis of the development of an infectious epidemiological event, it is reasonable to assume that the spread of the infectious agent would be proportional to the rate of essential contact in the population (the characteristics determining the effectiveness of the contact are dependent on the nature of the infectious agent). However, the factor measuring the rate or intensity of contact cannot be observed and measured directly; one can conclude that it is not one of the observable factors, that can be derived from the other ones that can be observed directly, such as the total population, population density and such. Then, patterns of distribution of the population geographically can vary significantly among the jurisdictions, and the average density of the population may not be a sufficiently informative factor to describe it.

As a factual example, let us consider three countries: Slovakia, population 5 million; Portugal, approximately 10 million; and Austria, in the same range of population as Portugal. Looking at the concentrations of the population, one can observe that the maximum urban populations are similar between Slovakia and Portugal (circa 0.5 million) whereas it reaches a significantly higher range of values in Austria: 1.9 million. Hence, while the factor of the contact can be expected to be similar between the first two countries, it can be significantly higher in Austria due to the higher concentration of the population. For this reason, the assumption of linearity of the infectivity of the agent relative to the population may not be justified in all cases, and the factor(s) of infectious impact per capita, derived from the impact counts statistics such as total cases, etc. and the population, may not lead to the correct conclusion of the analysis. This example indicates that in many cases of the early factor analysis, preprocessing procedures applied in preparation of the data should be considered themselves as assumptions, to be verified in the subsequent, more detailed studies. This example can be summarized as follows:

Problem: ostensible covariation between early Covid-19 impact statistics and BCG immunization record in the national jurisdictions.

Early result: formulation of the hypothesis of a correlation between a lower Covid-19 impact (in cases and morbidity) and the record of universal BCG immunization.

Benefit to the public: proposing the hypothesis stimulated research into innate immunity and testing the hypothesis with more detailed studies.

Subsequent in-depth analysis and conclusion: hypothesis not confirmed by in-depth data analysis and controlled studies.

2. Policy Effectiveness

An ecological dataset was compiled from publicly available data on national and subnational jurisdictions in Europe and Northern America to examine the effectiveness of public policies aimed at controlling the spread of the epidemics known as "lockdowns". Data on policy in the jurisdictions has been collected and graded by presumed severity of the measure, where maximum value can be associated with physical curfews, closure of some or most services and the lower, with additional information, advice and similar.

The early hypothesis that was tested was a correlation between the severity of the lockdown and reduced epidemiological impact, in incidence and morbidity. The compiled dataset had the following structure:

The range of data (P): national and subnational health jurisdiction, Europe and North America;

Impact (factor of interest, $K(p)$): morbidity, mortality per capita;

Observable factors $F(p)$:

Policy factors: communications; severity; popularity and engagement; targeted response;

Social and cultural factors: average population density; population centers; culture of socialization; international connectivity;

Public health system, condition: general condition; resourcing and capacity; preparedness for major public health event;

Testing the hypothesis with the publicly available data at an early stage of approximately six months after the local onset of the epidemic and before the vaccines with the methods of statistical analysis did not demonstrate strong statistically significant support for the hypothesis.

Summary:

Problem: testing the effectiveness of proposed policy decisions aimed at controlling the spread of the epidemic.

Early result: formulation of the hypothesis of a correlation between the severity of the lockdown policy and lower Covid-19 impact (cases and morbidity). Early analysis did not support the hypothesis.

Benefit to the public: feedback to policy making. Formulation of a variety of methods and approaches in control of the infectious spread, including public information, advice, environmental engineering and other measures [20].

Subsequent in-depth analysis and conclusion: studies continue.

3. Vaccination Level vs Rate of Spread of the New Viral Variants

In this case, a relation between the rate of vaccination and the rate of spread of the Omicron variant of the Covid-19 virus was analyzed with an ecological dataset compiled from publicly available data for a subset of European jurisdictions. For the rate of spread, two variables were selected: *rate_max*, the maximum weekly count of new cases reported in the jurisdiction, per capita; and a novel invariant factor *rate_inv*, defined as the ratio of the maximum case count C_{max} (peak) to the preceding minimum C_{min} (trough), i.e., using exclusively characteristics of the case dynamics specific for the jurisdiction (**Figure 1**). Both types of analysis yielded consistent results.

Summary:

Problem: examining possible relation between the rate of vaccination and the rate of spread of the Omicron variant of Covid-19 virus in European jurisdictions.

Early result: an indication of a possible minor positive correlation, not considered statistically significant; a statistically significant exclusion of a strong negative correlation.

Benefit to the public: developed novel factors describing the rate of spread and methods to analyze the effectiveness of the vaccination policy.

Subsequent in-depth analysis: statistical studies continue.

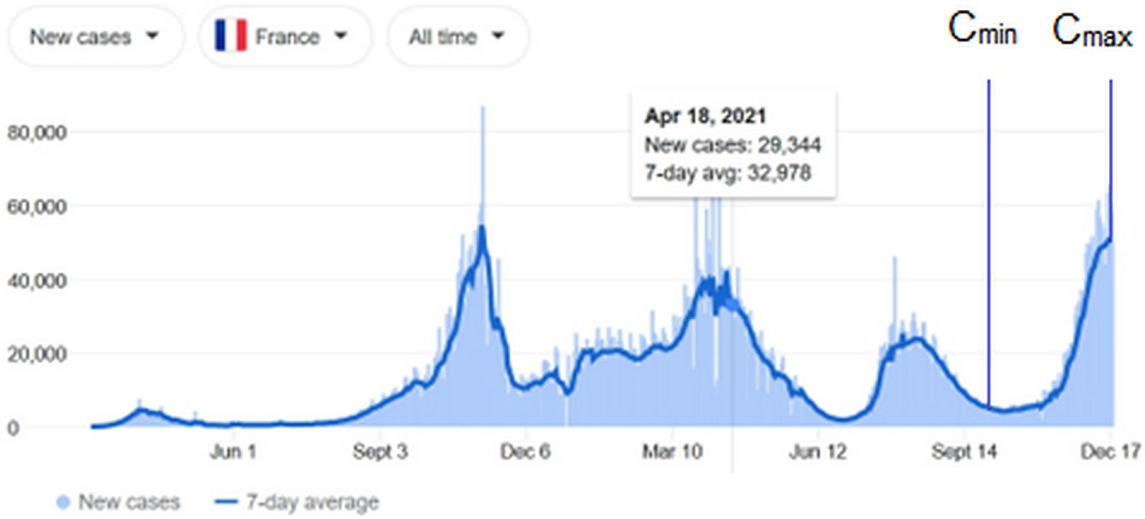

Figure 1: Definition of an invariant rate factor *rate_inv* from the public incidence statistics

The scenarios in the analysis of early ecological public data described in this section illustrate the use of public ecological data in the early formulation of hypothesis and factor analysis, with potential benefits to the public.

4. Methods in Early Factor Analysis

Even a remotely detailed discussion of the methods of factor analysis would need a voluminous dedicated work. Here we will only outline some of the most obvious choices to begin the analysis and obtain the initial results.

4.1. Methods of Statistical Analysis

Methods of statistical analysis can be used to obtain correlation factors (such as correlation coefficient) between the observable factor f and the factor(s) of interest K: $C_f = Corr(W(f), K)$ where $W(f)$: a vertical slice (column) of the dataset W at factor f. Values of C_f approaching 1 by absolute value indicate a strong correlation (positive or negative) whereas those close to zero, an insignificant one.

Additionally, these methods can produce other essential statistical characteristics in the distributions of observable factors and factors of interest, including confidence interval of values of C_f at a given confidence level. Examples of the use of statistical methods in early factor analysis of ecological public datasets were discussed in Section 3.2.2.

4.2. Regression and Multivariate Factor Analysis

Effective methods of regression and multivariate factor analysis have been developed and widely used in practice. These methods produce a projected dependency of the factors(s) of interest on the observable factors and can be both linear (linear regression and interpolation [21]) and non-linear (polynomial and other) in nature.

Some methods, including Random Forest regression, SelectKbest and others [22] can produce in addition, rankings of the observable factors with respect to the influence on the factor(s) of interest. This capability can be essential in the early analysis of the data and formulation of early hypotheses about the possible influencing factors in a developing epidemiological event.

A related aspect of regression and multivariate analysis is the analysis and verification of the produced dependency (trend analysis). It can be instrumental in the evaluation of potential scenarios in the development of the event, risks and preventative policy options. Multivariate factor analysis can be a relatively simple and informative approach in the initial analysis of ecological data.

4.3. Supervised Machine Learning

Methods of supervised machine learning can be seen as a type of multivariate regression, not necessarily of a linear type. They can produce an expected value (prediction) for values of observable factors that were not present in the original data via the process of training:

$$M = T(W(f), K)$$

where *M*: a trained method (predictor); *W*: training dataset that includes known values of the factor of interest (*K*) associated with observations in the set; *T*: training process.

Once the method has been trained, it can produce predictions as: $\tilde{k} = M(\tilde{f})$, where $\tilde{f}$: an arbitrary combination of values of the observable factors, *f*, $\tilde{k}$: the predicted value of the factor of interest.

Many types of supervised methods are used in practice, both linear and non-linear, including ensemble methods, artificial neural networks and many others.

4.4. Unsupervised Learning

Methods of unsupervised learning can be instrumental in the analysis of the structure of the data, expressed in observable parameters, without known association with the factors of interest (i.e., unmarked, unlabeled data). Many different types and flavors of unsupervised methods exist, linear and non-linear.

Some of the simplest ones are linear methods such as Principal Component Analysis (PCA) and more generally, Singular Value Decomposition (SVD). These methods produce representations of data in modified orthonormal (linearly uncorrelated) coordinates, obtained by linear transformation of the observable ones. These methods work well with complex data expressed in a large number of observable factors.

Methods of unsupervised learning were applied in the early analysis of Covid-19 epidemiological data to evaluate the distribution of the datapoints, identify characteristic types or clusters in the data and the informative factors that control the distribution of data points in the characteristic clusters. An illustration of the distribution of an epidemiological impact dataset (Covid-19) in a low-dimensional informative embedding obtained with a generative neural network model is shown in **Figure 2**.

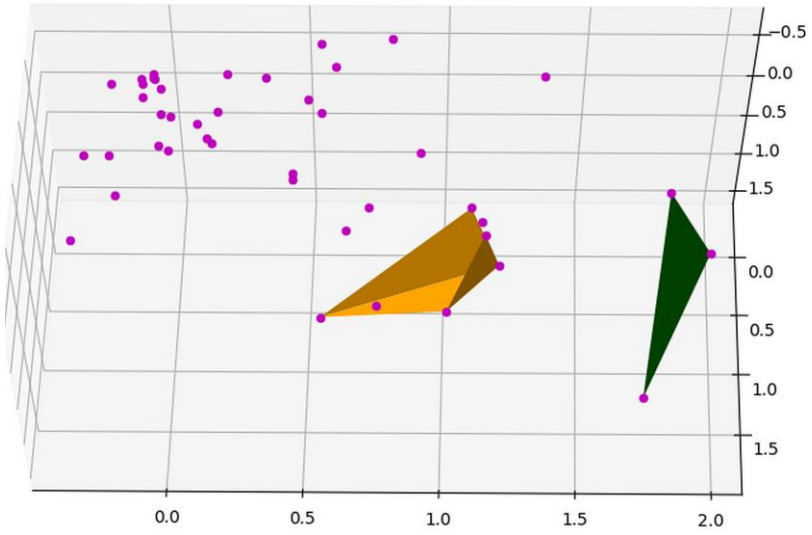

Figure 2: Distribution of Covid-19 epidemiological dataset in an informative three-dimensional embedding (generative ANN, [23]).

Many more methods of non-linear unsupervised learning and dimensionality reduction have been developed and used in practice, including: spectral, t-SN and other linear and non-linear embeddings (manifold learning), generative ANN [24,25] and others [26]. An advantage of these methods is that they have no limitations in expressing the informative factors (that is, embedded coordinates) as linear

combinations of the observable parameters of the data and can produce effective informative low-dimensional representations (embeddings) of complex real-world data.

Using methods of unsupervised learning, both linear and non-linear with complex real-world data may produce insights into the characteristic structure of the data and be instrumental in identifying essential factors from the perspective of the general problem of factor analysis discussed in Section 3.

5. Discussion

As was demonstrated in the discussion in this work, openly available public data can facilitate rapid analysis of a developing public health event, from different perspectives and with a variety of methods thus being instrumental in formulating hypotheses, providing advice and feedback to policy actions and so on.

Taking into account the experience, successes and shortcomings of public data made and being made available during the onset of the pandemic, we suggest the following improvements that can improve the usability of the data, as well as accuracy and confidence in the findings and results.

- Ensure lateral consistency data by harmonizing reported factors, units, collection and reporting practices, standards and formats [15].
- Attempt to provide and ensure vertical traceability and consistency of the data i.e., the ability to navigate from higher to lower levels of reporting without significant loss of accuracy.
- Attempt to identify and provide confidence level in the published statistics.
- Connect and attempt to provide consistency between different observable factors, for example, cases; morbidity; severity; etc.
- Ensure compatibility and consistency between different releases of the data; in the least, it should be possible to convert different releases to the same consistent format.
- Provide a stable permanent repository for the data and easily locatable access point(s).
- Attempt to provide research-friendly formats of the data that do not require significant manual processing.
- As a further perspective, work on harmonization of reporting practices, standards and resulting data between national and international structures.

6. Conclusion

In this work we examined the supposition that the availability of detailed; accurate and current data on the development of the event can be essential for the effectiveness of early factor analysis with methods described in this study and many others that can produce essential early insights into the factors, dependencies and trends in the developing situation; be instrumental in the formulation of hypotheses, testing and evaluation of the effectiveness of the implemented policies and responses. Practical examples of research cases based on the data available in the open public access support this view. Challenges in the use of public data were noted and discussed in detail, and recommendations for the preparation and publication of research-friendly data, availability and access policies formulated for implementation into the practice.

The challenges of the early factor analysis with publicly available data often relate to the trade-off between the shortening of the research cycle such as the formulation of hypotheses and the confidence of the conclusions. The assumptions, hypotheses and trends identified in the early phase will need to be tested and verified with more data and research, including controlled studies. However, it can be expected to have a largely positive effect due to the exchange of ideas and mutual stimulation of research directions in the crucial early phase in the development of public events.

An associated intent of this work has been to stimulate a discussion in the research community. With the improvements in the practice of collection, preparation and publication of the ecological data, as suggested here as well as identified in the follow-up discussions, it can be fully expected that early statistical analysis of public data can become an instrumental and necessary stage in the evaluation of developing public health events.

7. References

[1] Gorsuch, R.L.: Factor Analysis, Chronicle Books 2 ed. (1983).
[2] Worldometers: World Statistics. Online: https://www.worldometers.info/about (2023).
[3] Statista.com: Online Data Platform. Online: https://www.statista.com/aboutus/ (2023).
[4] Google Health: COVID-19 Open Data Repository. Online: https://health.google.com/covid-19/open-data/ (2023).
[5] Henry Bull Library: Advanced Sociological Research. Online: https://hbl.gcc.libguides.com/soci377/data
[6] World Health Organization: Coronavirus Disease. Online: https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (2019).
[7] European Centre for Disease Prevention and Control (ECDC): Covid-19. Online: https://www.ecdc.europa.eu/en/covid-19 (2019).
[8] National Health Service, United Kingdom (NHS) Covid-19 Information and Advice. Online: https://www.nhs.uk/covid-19-advice-and-services/ (2019).
[9] Centers for Disease Control and Prevention, USA (CDC) Covid-19. Online: https://www.cdc.gov/coronavirus/2019-ncov/index.html (2019)
[10] Mukherjee, S., Pahan, K.: Is COVID-19 Gender-sensitive? Journal of Neuroimmune Pharmacology 16, 38–47 (2021).
[11] Sze, S., Pan, D., Nevill, CR et al: Ethnicity and clinical outcomes in COVID-19: A systematic review and meta-analysis. EclinicalMedicine 100630 (2020).
[12] Bloom, J.D., Chan, Y.A., Baric, R.B. et al: Investigate the origins of Covid-19. Science 372 (6543), 694 (2021).
[13] Skegg, D., Gluckman, P., Boulton, G. et al.: Future scenarios for the COVID-19 pandemic. Lancet 397 (10276) 777–778 (2021).
[14] Kim, D.D. and Neumann, P.J.: Analyzing the cost effectiveness of policyresponses for COVID-19: the importance of capturing social consequences. Medical Decision Making, 40 (3), 251–253 (2020).
[15] Simon, S.: Inconsistent reporting practices hampered our ability to analyze COVID-19 data. Here are three common problems we identified. The Covid Tracking project, The Atlantic (2021).
[16] Jons Hopkins Coronavirus Resource Center. Online: https://coronavirus.jhu.edu/map.html (2020).
[17] The Center for Evidence-based Medicine, Oxford University. Online: https://www.cebm.net/oxford-covid-19-evidence-service/ (2020).
[18] Li, B., Wu, F., Lim, S.-N. et al.: On feature normalization and data augmentation. In: Proceedings of CVPR-2021 (2021).
[19] Escobar, L. E., Molina-Cruz, A., Barillas-Mury, C.: BCG vaccine protection from severe Coronavirus disease 2019 (COVID19). Proceedings of the National Academy of Sciences, 117 (44), 27741–27742 (2020).
[20] Dolgikh, S.: Smart-Covid: intelligent solutions for higher risk environments. HAL archives-ouvertes, hal-02915459 (2020).
[21] Wendland H.: Scattered data approximation. Cambridge University Press 9 (2005).
[22] Dolgikh, S. and Mulesa, O.: Covid-19 epidemiological factor analysis: identifying principal factors with Machine. In: 7th International Conference "Information Technology and Interactions" (IT&I-2020) Kyiv Ukraine, CEUR-WS.org 2833 114–123 (2021).
[23] Dolgikh, S.: Unsupervised clustering in epidemiological factor analysis. The Open Bioinformatics Journal 14(1), 63–72, 2021.
[24] Coates, A., Lee, H., Ng, A.Y.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of 14th International Conference on Artificial Intelligence and Statistics 15, 215–223 (2011).
[25] Seddigh, N., Nandy, B., Bennett, D., Ren, Y. et al., "A framework & system for classification of encrypted network traffic using Machine Learning", In: 15th International Conference on Network and Service Management (CNSM) Halifax Canada, 1–5 (2019).
[26] Izonin, I., Tkachenko R., Dronuyk I. et al.: Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method. Math Biosc. Eng, 18 (3) 2599–2613 (2021).