Identifying Explosive Cases with Unsupervised Machine Learning

Serge Dolgikh[1][0000-0001-5929-8954]

National Aviation University, Kyiv

Abstract. An analysis of a combined dataset of Wave 1 and 2 cases, aligned at approximately Local Time Zero + 2 months with unsupervised machine learning methods such as PCA and deep autoencoder dimensionality reduction allows to clearly separate milder background cases from those with more rapid and aggressive onset of the epidemics. The analysis and findings of the study can be used in evaluation of possible epidemiological scenarios and as an effective modeling tool to design corrective and preventative measures to avoid developments with potentially heavy impact

Keywords: Infectious epidemiology, Covid-19

1 Introduction

A possible link between the effects of Covid-19 pandemics and a number of epidemiological factors including universal immunization program against tuberculosis with BCG vaccine was proposed and investigated in [1-3]. Here we provide an analysis of the combined dataset of Wave 1 and Wave 2 cases adjusted and aligned at the time period of approximately 2 months after the first local exposure.

The intent of this work is to analyze the distribution of case data points in the most informative parameter spaces identified by unsupervised machine learning methods and to attempt identification of the cases with the heaviest impact. Evaluating the combination of the parameters that can identify such cases would allow to evaluate and predict the potential risks of heavy impacts in the population proactively with the potential to make necessary corrections before the explosive onset that may result in a heavy cost to the society.

The methodology is based on processing the input data expressed as a set of observable parameters that were identified and described in the study with unsupervised machine learning methods to identify and extract a smaller set of the most informative components. In many cases, evaluating distributions of data in the representations of informative components such as principal components in PCA or dimensionality reduction with neural network autoencoder models allows to identify and separate classes in the data by essential characteristics that can be linked to the outcome.

2 Data

2.1 Combined Case Dataset

The zero time of the start of the global Covid-19 pandemics was defined in [2] as:

$$TZ = 31.12.2019$$

Clearly, the time of local exposure to the epidemics is one of the defining parameters of the impact, so the case data was adjusted and aligned at a similar phase in the development, chosen based on the availability of data at approximately LTZ + 2 months, i.e. 2 months after the first local exposure to the infection. That translates to the beginning of April, 2020 for Wave 1 cases (LTZ in January, 2020) and beginning of May for Wave 2 (LTZ end of February to early March, 2020).

A combined dataset of 43 cases was thus constructed based on the conditions outlined in [2], essentially, certain level of reliability of the reported data and exposure to the epidemics.

2.2 Data

The dataset was constructed from the current data on the epidemics impact per case, i.e., reporting jurisdiction as described in [2]. In addition to the current measure of impact, several initial or observable parameters were recorded as described further in this section with the hypothesis of a certain level of correlation between the parameter set and the severity of the outcome, measured by a combination of perceived severity and the actual reported impact in mortality per capita (m.p.c.), per million of population. On the relative scale of m.p.c. by jurisdiction, the “explosive” cases were normally identified as those with relative m.p.c. (relative to the maximum among all reporting jurisdictions worldwide, New York City in both groups of cases) of around and above 0.5. This subgroup of cases included all commonly reported cases of high epidemics impact.

In evaluation of distribution in the coordinates of principal components two higher impact clusters of cases were identified by relative impact: explosive cases with relative M.p.c. above 0.8 group included the widely commented first wave cases: Italy; Spain and New York with the highest impact worldwide observed to date. In the second group were six somewhat milder-impact cases, namely: United Kingdom; France; Belgium; Netherlands; Ireland and Quebec (Canada), with relative M.p.c. in the range from 0.6 to 0.8.

2.3 Observable Parameters

The examples of extra factors can include, among others: genetic differences; population density, social traditions and cultural practices, past widespread public policy such as immunization; smoking habits and of course the epidemiological policy of the jurisdiction aimed at controlling the spread of the disease.

In addition to the factors of population density, smoking level and BCG immunization practice described in [2,3] a number of other factors with potential impact on the

severity of the epidemics pattern were considered in this study as described in this section. A common comment for several of them is that due to limitation of time and resources, a rating scale approach was chosen for those factors that cannot or would be challenging to measure directly. Granted, such an approach can be influenced by subjective perceptions; however, we believe that more robust and objective techniques can be developed over time improving the quality of the analysis and the resulting conclusions.

Connections hub: this band parameter defined in the range 0 – 1.0, was intended to measure the intensity of communications related to the case, on multiples levels for example, international, inter and intra-continental, regional and so on; clearly more intensive connection hubs could be expected to have higher exposure to the pandemics increasing the probability of a heavier impact.

Social proximity: a band parameter, range 0 – 1.0, intended to reflect the closeness of inter-personal connections in the case, again in multiple spheres and domains, for example: family connections; socializing practices and traditions; the intensity of business connections; lifestyle practices; social events and others. Again, as was commented previously [1] modeling such a complex factor as a single value parameter may open the analysis to the vulnerability of subjectiveness; yet we believed that it could be important for the analysis and improvements to make its evaluation, by case more objective and accurate are possible in the future studies.

We also used three rating parameters intended to measure the policy of the jurisdiction as relates to the response to the pandemics. They are: 1) epidemiological preparedness of the health case system to an intensive and rapid development of an epidemics; 2) the effectiveness of the policy response; and 3) the timeliness of the response.

Epidemiological preparedness: a band parameter, range 0 – 1.0, intended to measure the preparedness of the health care system to handle a rapid onset of a large-scale epidemics (specifically, not the general state of the health care system, its technological level, funding and so on).

Policy response: rang**e** 0 – 1.0, intended to indicate both the effectiveness of the policy in controlling the epidemics based on available scientific data at the time; as well as its clarity to the general population and its preparedness to participate. While some concerns can be expressed that this factor can be influenced by post-impact considerations with potential post-factum correlated with the outcome, we believe that with the accurate approach these risks can be minimized. For example, it is clear that an unclear or misleading policy message could be highly detrimental to the intended effect of the policy and one doesn't need the outcome to judge such policy parameters objectively at the time the decision is taken and before the outcome is recorded.

Policy timeliness: measures the relative timing of introduction of the control policy to the local exposure to the epidemics. Range 0 – 1.0.

The resulting dataset of 42 identified cases with 8 observable parameters and the "explosiveness" label, proportional to the relative impact of the epidemics is presented in Table 1, Appendix.

2.4 Machine Learning Methods

To evaluate the hypothesis of the correlation between the identified parameters and the epidemiological outcome in the case, several well-known machine learning methods were used:

1. Linear regression
2. Principal Component Analysis and identification of principal informative factors
3. Unsupervised deep neural network-based dimensionality reduction and selection of principal informative factors

The first method produces a best fit linear approximation of the resulting effect series with a total deviation (error) from the trend [4].

PCA [5] produces a linear transformation of the dataset to the coordinates with the highest variation and does not use the resulting effect labels.

A deep neural network autoencoder (method 3) produces a non-linear compression (i.e., dimensionality reduction) of the observable data to the lower-dimensional representation with the most informative features [6]. The structure of the deep neural network model used in this work is described in detail in [7]. The diagram of the model is given in Fig. 1.

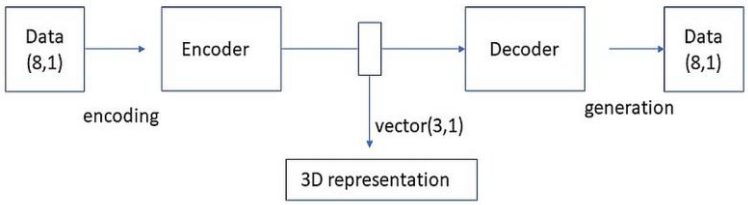

Fig. 1 Autoencoder with dimensionality reduction

In the unsupervised training phase, the model is trained to reproduce the input data with a good accuracy and thus does not require labels marked with the outcome, same as PCA. Achieving an improvement in the accuracy of reproduction, that can be measured by a number of training metrics indicates that the model has learned some essential characteristics of the initial distribution. Once trained, the unsupervised model can perform two essential transformations: the encoding one, from the observable data creating a representation image; and the generative one, from the representation to the observable data space. The aim of unsupervised learning is thus to minimize the deviation of the original training sample from its regeneration created by the model.

The models were implemented in Python, Keras, Tensorflow with a number of common machine learning and data processing packages

3 Results

3.1 Linear Regression

Linear regression with 8 identified input parameters produces a trend with a strong match score to the label of *0.9* of out 1.0 maximum. The most influential factors in the regression trend are shown in Table 1.

Table 1 Linear Regression Analysis

Factor	Linear Regression	Correlation
Policy, Time	0.534	0.906
Connection hub	0.196	0.697
Policy, Effectiveness	0.094	0.856
BCG immunization	0.092	0.686
Social proximity	0.078	0.794

Policy factors were expected to have a strong influence on the outcome of the case that is confirmed by the results of the linear regression analysis. As well, the importance of other factors such as connection intensity, social proximity culture, BCG immunization and smoking were also observed.

3.2 Principal Component Analysis

PCA identified three principal components with overall influence of 96% as described in Table 2. The highest influence factors in the PCA analysis are mostly aligned with the earlier results: policy-time, connection hub, social proximity, BCG and the smoking rate.

PCA transformation is inherently unsupervised method of learning, meaning that the prior known outcome labels are not required to learn the principal components as well as representation of the input data in the coordinates of identified principal component eigenvectors. By plotting thus transformed dataset in the coordinates of the principal component vectors, interesting result can be observed by identifying the highest impact cases

Table 2 Principal Components

Eigenvector	Main factors	Relative Weight
Axis 1	Policy-time, BCG	0.570
Axis 2	BCG, smoking	0.166
Axis 3	Connection hub, social proximity	0.127

Shown in Fig.1 are the visualizations of the distribution of the dataset in the coordinates of principal components identified by PCA analysis. The cases and the resulting region of the highest-impact cluster is shown in blue; whereas the milder cluster of 6 cases, in magenta.

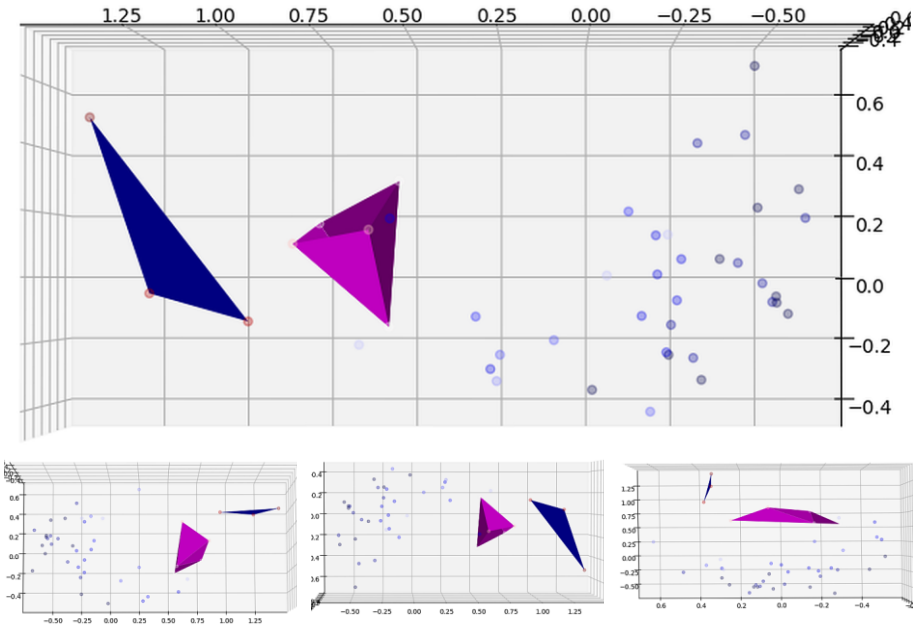

Fig. 1 High-impact clusters in the PCA representation

In the visualizations of the case clusters in the principal component representation one can observe a clear separation of the higher-impact case clusters from the general background cases. Such a clear separation allows to identify the region where the cases with potentially higher impact including the "explosive" pattern can be located, in the coordinates of principal component representation as well as in the initial, observable parameter space, with the possibility to identify the combinations of the observable parameters that can be linked to higher impact outcomes.

3.3 Unsupervised Autoencoder Model

A similar approach and results can be demonstrated with an unsupervised neural network autoencoder model that reduces the number of parameters by compressing the observable date into a lower dimensional representation while unsupervised training process is aimed at improving the accuracy of regeneration form the compressed representation to the observable space.

The dimensionality of the unsupervised representation for the models in the study that is defined by the size of its central encoding layer was chosen as 3 based on the results of the principal component analysis in the previous section, indicating three most informative dimensions.

Presented in Fig.3 are direct visualizations of the distributions of data in the unsupervised representation created by a trained autoencoder model.

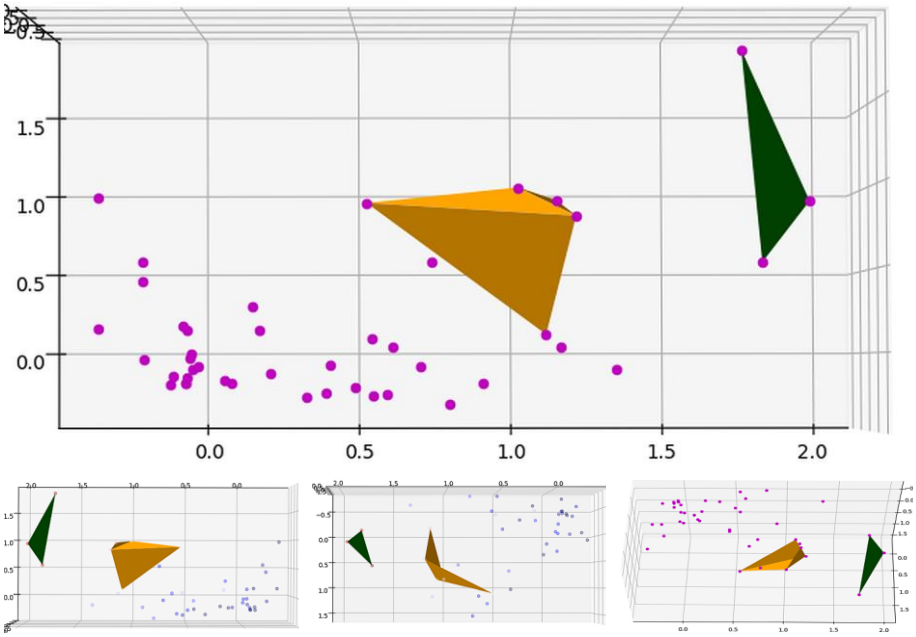

Fig. 3 High-impact clusters in the autoencoder representation

The highest impact cluster (3 cases) is shown in Fig.3 in green whereas the milder one (6 cases), in orange. Again, a similar pattern of clear separation of higher-impact cases from general background can be observed with these models as well.

It is worth noting that as with PCA, though essentially non-linear, autoencoder models also allow to identify the higher-impact regions in the coordinates of the unsupervised representation as well as in the observable parameters, by forward-propagating through the model the identified set of points defining the region of interest in the representation coordinates to the observable space.

4 Conclusion

The methods of unsupervised machine learning often allow to identify and separate the most informative components in complex general data. In this case, two different methods unsupervised learning methods applied independently demonstrated good separation of cases with higher impact from general background.

Further studies can lead to more precise and confident measurement of the observable parameters resulting in higher confidence in the results.

The analysis and the findings of the study can be used in evaluation of possible epidemiological scenarios and as an effective modeling tool to identify the areas of potential risk and design corrective and / or preventative measures to avoid the developments with potentially heavy impact.

References

1. Miller A., Reandelar M-J., Fasciglione K., Roumenova V., Li Y., Otazu G.H. Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19: an epidemiological study, medRxiv 2020.03.24.20042937 2020.
2. Dolgikh S., Further evidence of a possible correlation between the severity of Covid-19 and BCG immunization, preprint MedRxiv, https://www.medrxiv.org/content/10.1101/2020.04.07.20056994v1 April 2020.
3. Dolgikh S., Possible Covid-19/BCG Correlation: Factor Analysis. Preprint ResearchGate April 2020.
4. Freedman D., Statistical Models: Theory and Practice. Cambridge University Press, 2005.
5. Jolliffe I.T., Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002.
6. Bengio Y., Learning deep architectures for AI, Foundations and Trends in Machine Learning, vol.2, no.1, pp. 1–127, 2009.
7. Dolgikh S., Categorized representations and General learning, Proceedings of the 10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions (ICSCCW-2019), vol.1095, pp.93-100, 2019.
8. Zwerling A., Behr M.A., Verma A., Brewer T.F., Menzies D., Pai M. The BCG World Atlas: A Database of Global BCG Vaccination Policies and Practices PLOS Medicine • https://doi.org/10.1371/journal.pmed.1001012, 2011.
9. BCG World Atlas http://www.bcgatlas.org/
10. Coronavirus map Google https://www.google.com/covid19-map/ (4.04.2020).
11. World smoking prevalence https://ourworldindata.org/smoking
12. Word population data https://www.worldometers.info/world-population

Appendix Case Dataset

Table 1 Combined Wave 1, 2 Case Factor Dataset, at LTZ + 2m

Case	*Policy*			*Conn*	*Bcg*	*Smo*	*Den*	*Soc*	*Label*	*R.mpc*
	p-hc	*p-eff*	*p-tme*							
Taiwan	0	0	0	0.1	0	0.34	0.2	0.2	0	0.001
Japan	0.1	0.1	0	0.6	0	0.674	0.1	0.2	0.1	0.002
Singapore	0	0	0	0.4	0	0.33	0.5	0.3	0	0.004
Australia	0.2	0.2	0	0.2	0.3	0.298	-0.5	0.3	0	0.005
S.Korea	0.1	0.2	0	0.2	0	0.996	0.1	0.2	0.1	0.013
Finland	0.3	0.2	0.1	0.1	0.3	0.418	-0.2	0.2	0.1	0.017
Canada	0.4	0.2	0.2	0.3	0.8	0.354	-0.5	0.4	0.2	0.023
Ontario (Canada)	0.4	0.2	0.25	0.3	0.8	0.258	-0.25	0.4	0.25	0.025
Germany	0.3	0.2	0.2	0.5	0.2	0.608	0	0.4	0.25	0.052
Sweden	0.3	0.3	0.3	0.1	0.6	0.412	-0.2	0.3	0.3	0.148
UK	0.5	0.7	0.7	0.7	0.8	0.398	0	0.5	0.55	0.248
France)	0.5	0.5	0.6	0.7	0.6	0.596	0	0.7	0.60	0.371
Belgium	0.5	0.4	0.5	0.7	1	0.53	0	0.5	0.65	0.429
Spain	0.8	0.7	0.8	0.5	0.8	0.584	0	0.8	0.9	0.965
Italy	0.8	0.8	0.9	0.7	1	0.566	0	0.8	0.9	0.969
USA	0.5	0.5	0.5	0.3	1	0.39	-0.2	0.4	0.4	0.095
New York City (US)	0.8	0.8	0.9	1	1	0.25	0.5	0.8	1.0	1.000
California (US)	0.5	0.3	0.2	0.5	1	0.226	0.1	0.4	0.2	0.040
Slovakia	0.2	0.2	0.2	0	0	0.794	0	0.2	0	0.016
Argentina	0.4	0.3	0.3	0	0	0.478	-0.25	0.3	0	0.019
Chile	0.2	0.2	0.1	0	0	0.76	-0.1	0.2	0.15	0.050
Ukraine	0.6	0.4	0.3	0	0	0.94	0	0.4	0.1	0.027
Kyiv	0.5	0.3	0	0.1	0	0.7	0.2	0.5	0.1	0.024
Poland	0.3	0.2	0.1	0.2	0	0.648	0	0.3	0.15	0.066
Moldova	0.6	0.4	0.3	0	0	0.56	0	0.4	0.2	0.125
Czechia	0.3	0.2	0.1	0.1	0	0.766	0	0.25	0.15	0.082
Croatia	0.3	0.2	0.1	0	0	0.74	0	0.25	0.15	0.068
Albania	0.3	0.2	0.1	0	0	0.8	0	0.25	0.1	0.038
Greece	0.2	0.1	0	0.4	0	1	0	0.5	0.1	0.049
Israel	0.1	0.1	0.1	0.4	0.15	0.382	0.25	0.2	0.1	0.094
Peru	0.5	0.4	0.3	0.1	0	0.096	0	0.3	0.2	0.125
Prairies[1] (Canada)	0.3	0.2	0.2	0	0.6	0.292	-0.4	0.2	0	0.016
Quebec (Canada)	0.6	0.4	0.5	0.3	0.8	0.304	-0.1	0.5	0.7	0.912
Ecuador	0.6	0.4	0.4	0	0.6	0.28	0	0.3	0.3	0.309
Norway	0.2	0.2	0.2	0.2	0.2	0.452	-0.2	0.25	0.2	0.138
Denmark	0.2	0.2	0.2	0.1	0.3	0.352	0	0.25	0.3	0.303
Switzer-land	0.2	0.2	0.2	0.2	0.3	0.51	0.1	0.25	0.5	0.603
Austria	0.2	0.2	0.2	0.2	0.25	0.704	0	0.25	0.3	0.238
Portugal	0.3	0.3	0.3	0.3	0	0.63	0.1	0.5	0.3	0.355
Ireland	0.4	0.3	0.5	0.4	0	0.444	0	0.6	0.6	0.653
Nether-lands	0.3	0.4	0.4	0.5	1	0.524	0.25	0.25	0.65	0.774
Iran	0.7	0.6	0.8	0.1	0.1	0.6	0	0.7	0.3	0.265

[1] Manitoba and Saskatchewan provinces, Canada

Case factors

Policy

p-hc: health care preparedness, band

p-eff: response measures, band

p-tme: response timing, band
Conn: connection intensity, band
Bcg: BCG immunization record
Smo: smoking rate
Den: population density, band
Soc: social proximity, band
Label: relative mortality per million capita, band
R.mpc: Mortality per million capita, relative to world's highest