**new abstract**

**Complete:**

**[Background]Covid-19 is one of the most widespread pandemic diseases in human history, which has today an enormous global impact all over the world. From the perspective of the government, they need to consults the opinions about the progression of the situation of experts of various disciplines – from medicine to data science – to make their decisions. In this background, predictions about Covid-19 cases and their impact on society will help to better allocate medical resources, and facilitate the decision-making process for policymakers. This paper aims to develop a Covid-19 case prediction algorithm to model the pandemic evolution.**

**[Data source]Our main dataset is secondary data from COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (**[**https://github.com/CSSEGISandData/COVID-19**](https://github.com/CSSEGISandData/COVID-19)**). It provides cleaned time series data (one case report per day) for covid-19 confirmed, covered, and deaths in Switzerland since Feb 14, 2020, till now. The data is relatively reliable since it is scrapped from the Switzerland Federal Office of Public Health FOPH(**[**https://www.covid19.admin.ch/en/overview**](https://www.covid19.admin.ch/en/overview)**). But there are still some potential issues: the absolute number may be underestimated since it is based on the laboratory-confirmed data, and those patients who have mild symptoms and do not go for a test are unreported; besides, as the prevalent virus variants changes, for example, from Delta to Omicron, most infected persons changes from having severe symptom to having mild symptom and thus their tendency to get a test may also change. But we could still use these data to get a relatively accurate approximation of the outbreak, turning points, the maximum number of infections, etc. If later we want to increase model complexity, for example, taking vaccination (%people with at least one dose, partially vaccinated, fully vaccinated, with booster vaccination, age range), hospital capacity (%intensive care units occupied by covid-19 patient, % total hospital capacity occupied by the covid-19 patient), and virus variants into account, we can try to scrape and gather our time-series data from the FOPH website.**

**[Methods]In this research, we adopted multiple kinds of models for the prediction of the Covid-19 pandemic trend in Switzerland. To predict the trend from different aspects, we use models from traditional medical models with expertise, hybrid time series models, and statistical machine learning models. Medical models, such as the SEIR have good interpretability and intuitive dynamical processes. Time series models, such as the Fbprohet gets close to the specific form of original data and are suitable to reflect the macro tendency and short-term periodical fluctuation. Statistical machine learning models go from data and always give accurate predictions. We will compare the results and get the conclusion thoroughly.**

**Besides prediction, we also plan to analyze the impact of the Covid-19 pandemic on society through different perspectives. We are interested in its influence on trading, industry, and economy now. As our research progresses, we are open to utilizing more advanced methods and more meticulous data analysis ideas.**

**[Social impact]With our project we aim to evaluate traditional and more modern statistical modeling techniques to identify suitable methods to reliably predict the future development of the pandemic, while also giving a reliable forecast of the future pandemic situation in Switzerland.**

**123**

Covid-19 is one of the most wide-spread pandemic diseases in human’s history, which has today an enormous global impact all over the world. From the perspective of the government, they need to consults the opinions about the progression of the situation of experts of various disciplines – from medicine to data science – in order to make their decisions. In this background, predictions about Covid-19 cases and its impact on society will help to better allocate medical resources, and facilitate the decision making process for policy makers. The aim of this paper is to develop a Covid-19 case prediction algorithm to model the pandemic evolution.

**4**

Our main dataset is secondary data from COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (<https://github.com/CSSEGISandData/COVID-19>). It provides cleaned time series data (one case report per day) for covid-19 confirmed, covered, and deaths in Switzerland since Feb 14, 2020 till now. The data is relatively reliable since it is scrapped from the Switzerland Federal Office of Public Health FOPH(<https://www.covid19.admin.ch/en/overview>). But there are still some potential issues: the absolute number may be underestimated since it is based on the laboratory-confirmed data, and those patients who have mild symptoms and do not go for a test are unreported; besides, as the prevalent virus variants changes, for example, from Delta to Omicron, most infected persons changes from having severe symptom to having mild symptom and thus their tendency to get a test may also change. But we could still use these data to get a relatively accurate approximation of the outbreak, turning points, maximum number of infections etc. If later we want to increase model complexity, for example, taking vaccination (%people with at least one dose, partially vaccinated, fully vaccinated, with booster vaccination, age range), hospital capacity (%intensive care units occupied by covid-19 patient, % total hospital capacity occupied by covid-19 patient), and virus variants into account, we can try to scrap and gather our own time-series data from the FOPH website.

**5**

In this research, we adopted multiple kinds of models for the prediction of Covid-19 pandemic trend in Switzerlands. To predict the trend from different aspects, we use models from traditional medical models with expertise, hybrid time series models, and statistical machine learning models. Medical models, such as the SEIR have good interpretability and intuitive dynamical processes. Time series models, such as the Fbprohet gets close to the specific form of original data and are suitable to reflect the macro tendency and short-term periodical fluctuation. Statistical machine learning models go from data and always give accurate predictions. We will compare the results and get the conclusion in a thorough way.

Besides prediction, we also plan to analyze the impact of Covid-19 pandemic to the society through different perspectives. We are interested in its influence on trading, industry and economy now. As our research progresses, we are open to utilize more advanced methods and more meticulous data analysis ideas.

**6**  our study's impact on society

With our project we aim to evaluate traditional and more modern statistical modeling techniques to identify suitable methods to reliably predict the future development of the pandemic, while also giving a reliable forecast of the future pandemic situation in Switzerland.

**Papers summary:**

P. Wang & cie, Prediction of epidemic trends in COVID-19 with logistic model machine learning technics, Chaos, Solitons and Fractals, Elsevier, 2021.

> Logistic model fit the cap of epidemic trend and feed the cap value into >

> FbProphet model based on time series

M. Zivkovic & Cie, COVID-19 cases prediciton by using hybid machine learning and beetle antennae search approach, Sustainable Cities and Society, Elsevier, 2021.

> current time-series predictions based on

>> machine learning, adaptive neuro-fuzzy inference system

>> & nature-inspired algorithms, enhanced beetle antennae search swarm intelligence metaheuristics

> tested and validated against a wider set of benchmark functions and proved that it substantially outperforms original implememntation

> evalutated using the WHO official data

> compared against several existing state-of-the-art approaches tested on the same datasets

> test robustness by validate against two different datasets of weekly influenza

> proofs of quality with simulation results and comparative analysis

J. C. Clement & Cie, A survey on mathematical, machine learning and deep learning models for COVID-19 transmission and diagnosis, IEEE Reviews in biomedical engeneering, EMB, 2022.

> mathematical models: compartement models, statistical models

> data driven models - ML, DL, natural language processing, SEIR , SIR, sentimental analysis

W.-X. Tang & Cie, Causal Analysis of impact factors of COVID-19 in China, Elsevier 2022.

> Parameters: mobility, group awareness, temperature/weather

> DAG (directed acyclic graph)

> regression of toral causal effect among explanatory variables, negative binomial regression

> comprehensive causal analysis

F. Rustam & Cie, COVID-19 future forecasting using supervised machine learning models, IEEE Access, 2020.

> Future forecasting mechanisms

>> LR (Linear regression)

>> LASSO ( Least absolute shrinkage and selection operator)

>> SVM (support vector machine)

>> ES (exponential smoothing)

> adjusted R^2 score

> supervised ML

S. Khalilpourazari, & Cie, Designing a hybrid reinforcement learning based algorithm with application in prediction on the COVID-1 pandemic in Quebec, Annals of operations research, Springer, 2020.

> Novel hybrid reinforcement learning-based algorithm capable of solving complex optimization problems

> combined with the most recent mathematical model for Covid-19 pandemic prediction

> Reinforcement learning

> SIDARTHE

> Machine Learning

Z.-K. Zhang, Dynamics of information diffusion and its applications on complex networks, Physics reports, Elsevier, 2016.

> comparision of methods

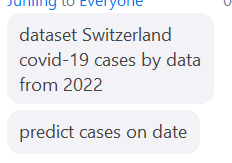
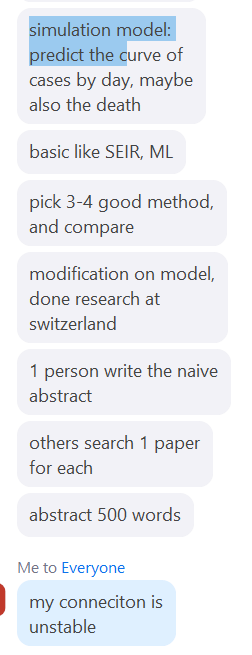
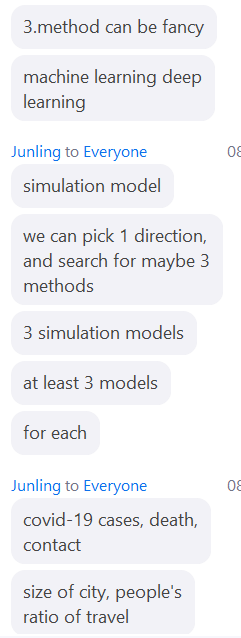
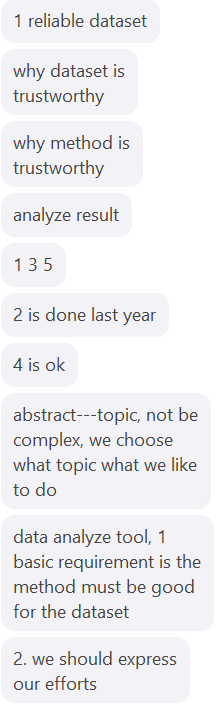
Z. Yang & Cie, Modified SEIR and AI prediction of COVID-19 in China under public health interventions, Journal of thoracic disease.

> population migration data & epidemiological data

> SEIR (Susceptible-Exposed-Infectious-Removed)

> AI

**What have been discussed during our meeting:**



**FINAL Abstract:**

**Draht abstract (600words max):**

***DATA CHALLENGE.***

***Background***

Covid-19 is one of the most wide-spread pandemic diseases in human’s history, which has an enormous global impact all over the world. From the perspective of the government, they need to **consults the opinions about the progression of the situation of experts of various disciplines – from medicine to data science – in order to make their decisions.**

**The daily life of every single citizen is directly impacted by policy measures which are themselves based on the prediction of the pandemic evolution.**

In this background, predictions about covid-19 cases and its impact on society will help to better allocate medical resources, and facilitate the decision making process for policy makers.

**This study focuses on the Covid-19 pandemic issues, which is We chose the nowadays most important and questionable health issue: the covid pandemic evolution.**

**In fact, every government consults the opinions about the progression of the situation of expert**s of various disciplines – from medicine to data science – in order to maketheir decisions. Furthermore, everyone asks himself how the future will be if new covid waves will occur. The literature shows that a tremendous amount of work has been done in this field, however, the situation is far from being completely managed yet. This is why research on the prediction of new covid cases must **be pursued**, and why we aim to add our forces to resolve this challenge.

**comment: data source, no need of attribute,**

***DATASET.*** The dataset will consist of the number of covid cases **and** covid related deaths per day. Those data will permit us to plot the curves (number of covid cases/deaths per day VS time) representing the pandemic trends. Added to that, one could add parameters as for instance the number of people vaccinated, recovered after 10 days, the size, weather and temperatures of the city, people’s ratio of travel, the number of contacts per person, etc. … Such further information could be used to develop parallel mathematical models that will permit more robust predictions. The area of focus will first be Switzerland and could be later on extended to other countries, even globally. The data will be collected from various reliable sources (e.g. BFS … , …) in order to cross-validate, as well as to detect falsified measures. Ancient data from 2019-2021 are important to learn from the previous trends, moreover, it is crucial to collect the latest epistemological daily data as well, in order to provide the most up-to-date predictions.

***comment: aml process consider specific situation.***

***DATA CLEANING.*** Data cleaning plays a significant role in the modeling process. We will first detect outliers through several outlier detection methods. After that, we will s for every dataset, it will contain imperfections. After the cross-validation, the remaining outliers will be removed using confidence intervals. Furthermore, the true-false and false-true covid cases must be taken care of, since antigen, PCR tests are not 100% reliable. In order to identify possible outliers we will run several outlier detection methods.

**SEIR, LASSO, Fbprophet(Machine learning), neural network**

***DATA MODELLING: METHODS, MODELS AND TOOLS*.**

**In order to find the optimal balance between model complexity and explainability, we will use a model selection algorithm such as, for instance, forward stepwise selection using a predefined model choice criteria.**

First we will create our model based on the latest litterature. Combining mathematical and data-based models seems us the best way to approach the problem of forecasting the evolution of the pandemic.

Using supervised machine learning, we will verify – (with f.ex. p-values?) – how the predicted data fits the real data as time goes in order to reinforce the algoritm and make the model more trustworthy. The danger will be to have the right sensibility, the good balance in minimizing testing and training errors,i.e. a reasonable fit. —— To prevent overfitting, a too complex model should be avoided. To avoid underfitting, we will … .

{

- adding noise for better results?

- Reducing variance by assuring a low bias.

Possible methods from the course:

* Gradient descent?
* Newton method?
* …
* see below methods in papers
* …
* Time series – NN Networks… (see if we add specific tools/methods from the lecture 5 of the course on it)

}

> some Models we could try:

>> LR (Linear regression)

>> LASSO ( Least absolute shrinkage and selection operator)

>> SVM (support vector machine)

>> ES (exponential smoothing)

> adjusted R^2 score

> supervised ML

***DATA INTERPRETATION.***

Understand peaks, effect of policies… variants, mutation of the virus…

Extend to other countries…,

Compare the effect of each policy…

***EVALUTATION OF THE PROCESS.***

To evaluate the method, we will compare our simulation results with real data as with the outcomes of other existing state-of-the-art approaches tested on the same datasets. For this purpose, we will use classical comparative anallysis, for instance mean square errors, variance and biais.

***RESOLVING THE CHALLENGE.***

…

**Examples of abstract:**

Example 1: Our project will be based on the Air Quality Forecast challenge.

The main object of the Air Quality Forecast challenge is to predict the air quality levels in Zurich, using data collected with two environment monitoring sensors, connected to a Raspberry Pi. There are two of these sensors : the SPS 30 from Sensirion and the DHT22. SPS 30 is an optical particle matter sensor, which provides us with the pollution particle density in the atmosphere, and DHT22 is a temperature and humidity sensor.

We are first planning on studying the particle density variations by themselves, using for example the Fourier transform (either discrete or continuous based on the sampling theorem). Additionally, we can assume the time series to have time dependent properties with values correlated to the day of the week and the time of the day. From these properties and latest measures we should be able to make first predictions on future air quality.

Moreover, having the possibility to measure humidity and temperature, we will try to infer statistical properties and correlation between temperature and the density of pollution particles in the atmosphere, using for instance Dynamic Time Warping (DTW) algorithms. If successful, we should be able to use it to strengthen our particle density prediction based on previous temperature measures (concept of Transfer Entropy). Optionally, more complete measures and weather forecast predictions from the city of Zurich could bring even more information for our prediction model.

To avoid overfitting the data and predicting on misleading values, we will implement a few outlier detection models. As we don’t have a lot of parameters we thought of starting our outlier detection by using a simple Z-score algorithm or by using box plots. If the results are not satisfactory enough we will use stronger algorithms : Isolation trees or Density-based spatial clustering of applications with noise (DBScan).

For the missing values that may occur during our measurements, we will go with the solutions of trying to predict what should have been the value. We will first test linear interpolation, but we suppose that it may be a too simple model to accurately predict the missing values. Furthermore, if the number of missing values is fairly small compared to the whole dataset we could consider ignoring them. Another technique consists in replacing these values with the average over a certain time window. We will test these methods and will select the one that provides the best results in average.

Finally, we will try various machine learning approaches to the prediction task. Our ideas for now are to use SARIMA (Seasonal Autoregressive Integrated Moving Average) which seems to deal well with our kind of problem. We will also try neural network models such as a Long short-term memory network, on which we can additionally compare the results with or without providing our temperature and humidity measures. Moreover, we will use and test different error functions and select the one that fits our model the best.

Example 2

===================================

The SARS-CoV-2 pandemic calls for new means and methods to reliably predict the development of infections and evaluate existing countermeasures to battle the spread of the virus. As part of the epidemic datathon in the context of the “Data Science in Techno-Socio-Economic Systems” lecture, our goal is to apply different prediction methods and evaluate, whether they are a suitable choice in the context of prediction of epidemic evolution. To do so, we compare three different approaches to forecast the development of the pandemic infection for any given moment since the outbreak. All models aim to predict the number of daily cases and deaths, in different locations, over the next thirty days following each training phase. The models we evaluate are:

• Simple linear model: Simple linear model that fits a function to the data. This model can serve as a baseline for evaluation performance and as a prototype for a mathematically simple and explainable model.

• Feed-forward neural network: Comparable to an n-gram model in natural language processing, this model takes a fixed number of timesteps and makes a prediction for the infection rate for each of the following 30 days.

• Recurrent neural network: Similar to the feed-forward neural network but with the difference that an arbitrary number of timesteps can be taken into account.

While the data provided in the context of the datathon only contains newly registered cases and new deaths in all available countries, the pandemic is influenced by many more factors than just its previous development. To account for this, we use additional data such as information about lockdowns, mask-wearing-mandates or school closings. The data used is provided by the University of Oxford through the “Our world in data” project (<https://ourworldindata.org/coronavirus-source-data>). We perform different preprocessing and normalization steps, for example logarithmization of large values, or applying a Savitzky-Golay filter to smooth the data and reduce variance in order to facilitate the prediction by the different algorithms. For each of the four prediction intervals that are part of the pandemic datathon, we exclusively use training data prior to the prediction data to prevent data leaking. As a consequence, we suspect that our prediction precision will improve with increasing prediction date as more and more training data is available. While this should not be problematic when fitting a curve to the data, training neural network heavily relies on large number of training data. We therefore aim to measure the impact and finally make a statement, whether a useful application of neural networks is possible in a context like ours. Finally, we make predictions for the predefined time intervals to take part in the Pandemic Datathon Challenge.