Report

UNIVERSITY
OF SKÖVDE

Covid-19 Measures Impact Under Parametric Uncertainty

The complete report of the final project for the DDDM Course. Submitted on 27-10-2020

Data Driven Decision Making (IT728A)
Autumn 2020

Filotas Theodosiou

Contents

1 | Introduction

In the past 6 months, humans' way of life has changed dramatically due to the pandemic. As over 200 countries have been affected by the COVID-19 outbreak with numerous cases, deaths, and hospitalizations, this once-in-the-century crisis has resulted in several novel measures taking place. Examples include global curfews or mass flight cancellations during summertime. While people are hesitantly trying to return back to a modified version of normality, the whole scientific community investigates the mechanism of COVID-19 and how it operates under different scenarios. The aim is to guide governments in ensuring a safe return to normality.

Despite the coordinated effort of the scientific community, many questions about the novel coronavirus still remain unanswered. Many studies present conflicting evidence due to survey biases and sampling or measurement errors (ECDC 2020). What is the incubation period or the infectivity before the first appearance of symptoms? What are the seasonal characteristics of the virus, or which is the specificity of the virus for certain population groups? In addition, the different approaches countries implement to report their daily number of cases and deaths, result in biased data full of inconsistencies. All the aforementioned, create extra pressure on policy-makers and governments. They turn measures and interventions taken to tackle the pandemic, like school closures or wearing masks in public, uncertain.

According to the official guidelines of World Health Organization(WHO) (2020) regarding uncertainty during the pandemic, uncertainty leads to fear and loss of trust in policymakers. As a result, managing uncertainty by being consistent and communicating each action taken is of extreme importance. Most importantly, during times of crisis, it is fundamental to be transparent and validate any measures taken, by providing explanations in the form of visualizations.

This project presents a Data-Driven Decision-Making Framework visualizing the uncertainty presented in the scientific research community regarding important factors responsible for the spread of the virus. (For example incubation period, the proportion and the effect of asymptomatic people, the probability of severe cases or deaths, and the patient's hospitalization time). The aim of the framework is to test the impact different measures have on preventing the spread of Covid-19 while considering the uncertainty regarding its mechanisms. The framework consists of two sub-models. First, a knowledge extraction NLP model is designed to handle the huge volume of Covid-19 literature. Its goal is to extract the conflicting results researchers found for structural features regarding the spread of the virus. Then an epidemiological forecasting model combines the values of the extracted features with country-specific attributes in order to forecast the number of new cases, deaths, and hospitalizations in a country. Different values for the virus mechanics would result in different forecasts and hence another strategy to tackle the spread.

Users of the framework will be able to test the effect of different measures in the generated forecasts. In simple terms, they will be able to tackle different scenarios produced by different parameter combinations. A utility function is defined to translate the government's actions to stop the spread of the pandemic. Users will be able to select the situation which they would like to explore by picking the desired initial reproduction number R_0. The reproduction number

R_0 reflects how much has the virus been spread in the community. Its initial value reflects the scenario users want to approach and is related to the current situation in a country. Afterward, by manipulating the values of the utility function, they will be able to find the optimal combination of measures in order to stop the spread. All of these, while also considering the impact of the parametric uncertainty found in the literature. Users will also be able to choose different values for each feature representing a core mechanism of the virus. The range of the available values for each feature is extracted from the Covid - 19 literature corpus.

This framework is specifically designed for policymakers responsible for applying extra measures for slowing the spread of the virus. Different approaches for handling specific scenarios would be comparable, as different values on the respective parameters are mapped into different measures. The impact of the uncertainty of fundamental Covid-19 dynamics is also taken into consideration. Due to the differently specified Covid-19 dynamics, the different measures proposed for different scenarios have by nature high uncertainty values. For this reason, demonstrating the impact of different parametric combinations found in literature would highly support decision-makers (ECDC 2020).

For example, by exploring the impact the different proportions of asymptomatic (and/or their infectious period) have, the public use of masks or the reduction of the average number of a person's contracts per day, under specific scenarios (different initial R_0), would be justified. The goal of the designed framework is to assist policymakers in finding the optimal measure combinations, under different scenarios, while considering the visualized parametric uncertainty. The aim is to ensure the long-term gain, which for the purposes of this work, is defined as an even and smooth spread of the virus for the span of 2-3 months. Most importantly though, policymakers should focus on ensuring that the hospital system does not collapse

2 | Related Work

This project is based on the suggestions by ECDC (2020) regarding the effect of parametric uncertainty. Two main themes are covered in this work. A forecasting model and a knowledge extractor. To my understanding, their combination is a novel approach as existing literature focuses on either one of the two themes.

2.1 Epidemiological Models

Mathematical epidemiological models based on differential equations are very common in modeling epidemics and pandemics due to their ability to perform simulations based on different scenarios. They split the population of a country into different groups. The simplest model splits the population into Susceptible, Infected, and Recovered(SIR model). The rate of change of the total number of people belonging at each epidemic stage at every time step t, is given by a differential equation specifically designed for this group. Solving the system of differential equations after specifying the total population, the initial number of infections, and recoveries, along with the total forecasting horizon, results in the number of people belonging to each group at every time step.

The size of the differential equations system is relative to the total number of groups the population is split upon. Each group requires its own differential equation. By increasing the number of equations, complexity is also increased. In turn, the modeling of the pandemic becomes more realistic. In (ECDC 2020) the authors defined a realistic, but complex model, with a system of over 20 differential equations. The proposed model is officially considered by several countries. The general Susceptible - Exposed - Infected - Recovered(SEIR) model is consisted of four differential equations and is widely applied over the past year (He et al. 2020, Mwalili et al. 2020, Carcione et al. 2020). The simple Susceptible - Infected - Recovered (SIR) model is also applied to model the spread of the virus (Joby Mackolil & Mahanthesh 2020).

Selecting how many and which epidemic stages to include, is a critical procedure and depends on the specific task the model is used for. For example, on Paiva et al. (2020), the addition of asymptomatics, hospitalized and deceased points out the effect of asymptomatics, in the selected community. This results in a model simulating events more realistically. In this work, the picked epidemiological model is a balanced selection between complexity and realism. It has a similar structure to SEIR - HCD (hospitalized - critical - deceased) defined on Unlu et al. (2020) and Ghamizi et al. (2020). The simple implementation and the transparency of such models have many advantages over more complex deep learning techniques like in Ghoshal & Tucker (2020).

The independent variables of each differential equation are parameters regarding specific dynamics of Covid-19. Some examples include the mortality rate or the proportion of the asymptomatics. The values of these variables are either given manually when designing the model, or they are calibrated (optimized on actual real data) by minimizing a loss function. These

independent variables highlight the impact some core mechanisms of Covid-19 have on the produced forecasts. This is an advantage over other transparent statistical models such as ETS (Petropoulos & Makridakis 2020) or ARIMA (Verma et al. 2020).

2.2 Knowledge Extraction

When it comes to extracting information from unstructured text, several techniques exist for a small number of articles or papers. Approaches such as rule-based information extraction (Chiticariu et al. 2013) or relation extraction after applying part of speech tagging (Btoush et al. 2016) are some of the most popular ones. However, when it comes to answering multiple queries on huge volumes of unstructured papers things are more complex. The general idea of the knowledge extraction model is a system that automatically analyzes a huge corpus of papers and returns sentences that contain relative information about the given query.

In general, this is a very simple but effective approach. The main challenge is the computation expense of analyzing over 150.000 papers. Another challenge is the discovery of quality answers from a high amount of sentences which the system would return. For example, over 40.000 papers contain the keyword combination "incubation period" and thus, over 40.000 possible results would be returned to be evaluated. The algorithms presented focus on the procedure used to decrease the volume of the corpus and pick the most relative papers to answer a user's query.

Before applying any text mining techniques, preprocessing of the unstructured text is necessary. Here, we give some basic definitions and explanations of the applied techniques. Stop words refer to commonly used words which contain no valuable information. These words vary from context to context, as in scientific literature words such as, “preprint” and “copyright”, or "pandemic" and "covid" are treated as noise and should be removed. Afterward, converting the unstructured text into something meaningful for computers is performed through vectorization. CountVectorizer returns a vector where each row describes the number each word appears on every document within the corpus. On the other hand, the Term frequency-inverse document frequency(TF-idD) vectorizer returns the Tf-iDF score of every term. The TF-iDF formula for a term t in a document d given a document collection D is given in Equation 2.1

$$tfidf = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} * log \frac{N}{|\{d \in D : t \in d\}|} \tag{2.1}$$

It can be viewed as the product of a term's frequency in a document with the document's inverse frequency. The document's inverse frequency is the logarithm of the total number of documents in the collection, divided by the number of documents which include the specific term t (Qaiser & Ali 2018)

Principal component analysis(PCA) is the process of computing the principal components of a dataset and using them for dimensionality reduction. The aim is to reduce the computation expense (Mishra et al. 2017). Moreover, text clustering refers to an unsupervised approach used to classify documents into groups with no predefined categories. It is based on the hypothesis that relevant documents are more similar to each other than non-relevant ones (Liu et al. 2018). Finally, the last technique implemented in this work is LDA. LDA is a generative probabilistic model where documents are weighted vectors over hidden topics characterized by a distribution of words. It is used to extract the main theme of a paper or a papers' collection (Anupriya &

Karpagavalli 2015). The most representative keywords for every topic, found within a collection of papers are extracted.

2.3 Uncertainty Visualization and Decision Making

The aim of the framework is to visually present the uncertainty found in Covid-19 literature. In this section, a range of techniques applied for equivalent tasks is discussed. To visualize the parametric uncertainty found in the COVID-19 scientific literature, Dong & Hayes (2012) used bar plots. However, such plots often make the users incorrectly interpret distributional data as categorical (Joslyn & LeClerc 2013). Instead and according to Padilla et al. (2020), the use of violin and gradient plots can facilitate people in better interpreting the uncertainty in the visualized data. A widely applied alternative to gradient and violin plots is box plots. Ghoshal & Tucker (2020)(Ghoshal and Tucker, 2020) used them effectively for a similar task.

In Padilla et al. (2020), the authors present proof that the use of icon arrays is useful for communicating probabilities. In addition, they state that the usage of quantile plots should be preferred over probability density plots. In general, frequently framing the uncertainty (1 out of 10) rather than probabilistically can be easier interpreted. Quantile dot plots have successfully been used on Fernandes et al. (2018), to improve transit decision making. As for the interactive framework to test the effect of measures in stopping the spread of the virus, Noll et al. (2020) presents a multi-dimensional tool with multiple interventions and parameters. The users can test the effect different combinations of measures have in halting the spread of the pandemic.

Finally, similarly with Dong & Hayes (2012), the project will be based on multi-criteria-decision-making methods(MCDM). The authors in Dong & Hayes (2012) present a visualization framework allowing decision-makers to simultaneously compare the aggregated uncertainty of multiple alternatives. MCDM techniques are included in a large family of normative decision-making methods (Klein et al. 1993). Policymakers can choose different alternatives in the form of qualitative or quantitative criteria (Saaty 1980). This work follows a similar idea as users can manipulate multiple parameter values at the same time. The proposed framework is designed to support normative decision-making. Policymakers should find the optimal trade-off between preventing the hospital system from collapsing, keeping the virus loads inside the community low. They should also make sure no extreme and prolonged harsh measures are taken, so abysmal socioeconomic consequences are prevented.

3 | Data

The first picked dataset includes over 280.000 scholarly articles, with over 150.000 full texts, about COVID-19, SARS, and related coronaviruses. It is part of an attempt to support the efforts on understanding and haltering the spread of the novel virus and is publicly available on Kaggle. Out of the 150.000 full-text articles, only those which address topics related to Covid-19 and are written in English are selected. The entire pre-processing approach will be later described in Section 4.

It is natural that several papers included in the corpus discuss the same Covid-19 dynamics. This implies high data variance. A simple example regards the mortality rate of different countries. Naturally, due to several issues such as survey bias, measurement errors, or unique country characteristics, the results obtained by researchers differ. In addition, articles are published during different time periods, and thus, results vary. In the early stages of the pandemic, medical staff did not have the current knowledge or means to take care of their patients. As a result, a different average hospitalization time per patient is found in earlier research papers. Another source of uncertainty within Covid-19 literature concerns measurement errors. Due to the different number of tests among countries, the accurate percentage of asymptomatic and patients with mild symptoms varies across the different research papers. This results in the probability of a patient developing severe symptoms to also be uncertain. The uncertainty of the aforementioned features is visualized and the users are able to manipulate the input features values given to the forecasting model. Under these conditions, users are able to test the forecasting results based on different parametric values.

In addition, demographic characteristics of each country are required to initialize the parameters of the epidemiological model. These features include the country's population, the number of initial infections, and the total hospital beds. Demographic data are also extracted from Kaggle. Lastly, the total number of ICU beds per country is taken from ECDC and Simon Rozendaal. (2020). It should be pointed out, that for the purpose of this work and the correct implementation of SEIR - HCD, the population of each country has been assumed to remain stable during the forecasting period, with the number of new births being equal to the number of those deceased. For each country, the total number of ICU beds may vary depending on the current status of the outbreak. In this work, it has been assumed to remain stable. Finally, to test the performance of the forecasting model, time series data regarding the number of new infections per day over different countries is also considered. The performance of SEIR -HCD is evaluated in countries with different characteristics.

4 | Methodology

In this section, the various steps taken for the design of the proposed framework are described. The section is split into three parts, one for every component of the framework. The first component regards the Knowledge Extraction Model. It is used for extracting the parametric values of specific features found within the COVID literature. Next SEIR - HCD is discussed. Finally, in the last subsection, the design of the final interactive framework is explained.

4.1 Knowledge Extraction Model

First of all, the selected dataset contains a vast amount of scientific articles being directly or indirectly related to Covid-19. This huge volume of data would create plenty of computational problems. This might result in some methods failing to be implemented. For this reason, several pre-processing techniques based on realistic and unrealistic assumptions are utilized to reduce the size of the literature's corpus. Firstly, as the novel coronavirus first appeared in late 2019 and early 2020, papers written before 2020 are assumed to contain non-relevant information and hence, they are removed. In addition, some papers are addressing issues such as pneumonia, SARS, or drugs and vaccine-related issues. These concepts are not directly related to Covid-19 mechanisms. The second assumption is that in order for a paper to contain relative information, a representative word for the novel coronavirus should be included either on the abstract or the title of the selected papers. Examples of such representative words include keywords like Covid-19 and SARS-CoV-2.

Furthermore many papers are not written in English. This might create problems with the implementation of some of the techniques. As a result, non-English papers are removed. Lastly, after testing different tools like Spark for parallel processing of BigData and computers with additional RAM, exploring all the remaining 62.230 papers at once, is not possible. For this reason, the third assumption is that papers before April have high uncertainty levels and their findings are not as accurate compared to those after April. The assumption is based on the fact that after April the pandemic has been expanded worldwide and the whole scientific community has been alerted. Thus, the dataset is split into two periods, those before and after April. It should be noted that this assumption is not realistic and does not imply that articles before April of 2020 do not contain valuable information. It is an assumption used for simplifying the computations.

Next, the following steps describe the pre-processing techniques applied to the documents, before clustering. Stop words, both common and scientific ones, along with some handcrafted ones, are removed and documents are vectorized using TF-iDF. To deal with potential sparsity issues, PCA is used with the dimensions of the vector being reduced to 90 %. The reason for applying PCA is to slightly decrease the dimensions by removing some noise on the sparse feature vector. Then K-Means with Euclidean distance is applied. The total number of clusters is picked

using the elbow method as described in Figure 4.1. It should be noted that clustering is a total of 3 times to deal with potential validity threats.

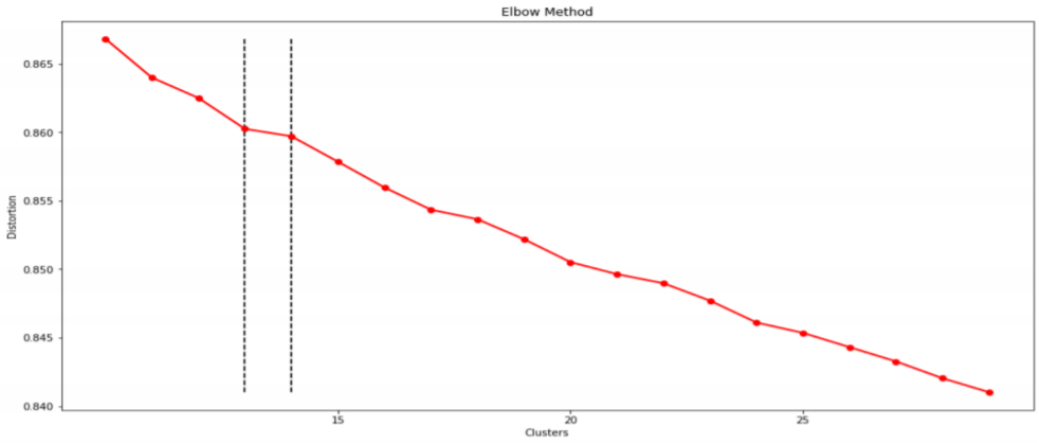

Figure 4.1: The results of the Elbow Method. On k = 14 a linear decay starts to take place. A total of 14 clusters is selected

Then on every cluster, Latent Dirichlet Allocation (LDA) is applied to extract the topic of each cluster. LDA has been successfully implemented on several occasions (Liu et al. 2018, Bastani et al. 2019). Its good overall performance is the main reason for its selection. Despite its wide usage and its good overall performance, a limitation of LDA is that it is highly dependent on the clustering results. For example, if the used clustering algorithm does not group relevant papers together, then the keywords extracted for every topic might not be accurate. For example, if papers talking about the reproduction number R_t are not clustered together then the keyword "reproduction" will not appear for any topic on any given cluster. This resulted in many queries not returning any representative keywords. Thus, a query needed to be repeated several times. On the other hand, if the query matches with a keyword then there is a strong connection between the particular cluster and the made query. The total number of topics extracted from every cluster is optimized by applying Grid-Search. The number of topics that maximizes the log-likelihood is picked (Subeno et al. 2018). A total of 8 topics are extracted for every cluster. For each of the 8 topics found within each cluster, the 10 most representative keywords are extracted. As a result, a total of 80 keywords are generated to describe each cluster of documents. If a user-made query is matched with a topic keyword then the cluster is automatically picked for further analysis.

In addition, for every set of clusters, the aggregated Tf-IDF vector is generated. For every word found within the cluster of papers, its total aggregated Tf-IDF score is stored. These vectors are used for finding the most representative cluster for each query. A challenge with this approach is that meaningless words, like "pandemic", had extremely high scores. Equivalent words in analogous occasions are treated as stopwords and are removed. The whole described procedure is repeated and similar keywords are added as stop words. For each query performed by users, the two clusters with the highest aggregated TF-iDF are selected for further exploration.

Finally, a manually designed knowledge extraction algorithm is implemented. Users are asked to perform two queries. One is an abstract, more general query and the second is a more specific one aiming at extracting the relevant information from the most representative cluster. For the abstract query, the two clusters with the highest aggregated TF-iDF are selected. In addition, if the query matches with a topic keyword of a cluster generated with LDA, then this cluster

is also selected. This approach identifies the most representative clusters for each query. Then, the assumption taken is that in order for an article to discuss a specific query, the words of the query should be included in the abstract. With this assumption, only the papers with a mention of the query are kept. On papers included in the selected clusters, a more detailed query is used. Sentences including the query words are recommended by the proposed model.

The described model is implemented for every core mechanism of Covid-19 affecting the performance of the forecasting model. Values that accompany features such as "incubation period" and "proportion of asymptomatic" are generated and are stored in vectors. An example of the usage of the knowledge extraction algorithm is given in Figure 4.2. It should be mentioned that an option on whether to present results from both databases of papers (before and after April), is also included. For queries, where a small range of values is returned, the results of the database including papers before April, are automatically considered.

```
general = 'incubation'
detailed = 'incubation period days'
zz = knowledge_extraction_both(general,detailed,to_print=True)
Next Paper: incubation period and serial interval of covid-19 in a chain of infections in bahia blanca (argentina)
____
The estimated median incubation period in this study was 5.8 days for general transmissions; the estimated mean incubation wa
s 6.9 days which is 33% longer than the previously frequently adopted value-5.2 mean days as reported by Li (16)


The random-effects meta-analysis using restricted maximum likelihood (REML) was used to summarize the median incubation perio
d (days) and the corresponding 95% confidence interval (95%CI)


____
Next Paper: a systematic review and meta-analysis reveals long and dispersive incubation period of covid-19
____
The median incubation period 46 of COVID-19 is estimated as 5 to 6 days (1-4), while that of influenza A and B and SARS-CoV-1
are 1.4, 0.6 47 and 4.0 days, respectively (5)


____
Next Paper: estimation of the incubation period of covid-19 using viral load data
____
```

Figure 4.2: General query “incubation” searches the Tf-IDF vectors of every cluster and returns the clusters with the highest score. In addition, if the query is the topic of a cluster, this particular cluster is also included. Papers not containing the general query on their abstract are discarded. Detailed query “incubation days” search the full text for mentions of the two words side by side. In the “databases” option, the value “both” could be given where a search would take place on both databases of papers

4.2 SEIR - HCD Model

A wide extend of Covid-19 literature regards mathematical epidemiological models. These models are specifically designed to mathematically model a pandemic and simulate different scenarios. The model’s simple implementation, transparency, and ability to highlight the importance of core virus mechanisms are the main reasons for their popularity. Similar models have also been designed for other pandemics like SARS.

Epidemiological models use differential equations, estimating the rate of change of people at each pandemic stage during different times. The total number of differential equations along with their formulas depend on the scenario to be simulated. In general, the fundamental mechanisms of Covid-19 like incubation and infectious period, in addition to reproduction number, are the most important features of the core equations on such models. Examples of other features which

can be included are the percentage of asymptomatic, the probability to develop severe symptoms, the average hospitalization days, and the mortality rate.

Several variations of such models exist in literature. The simplest model, SEIR with 4 population groups (Susceptible, Exposed, Infected, Recovered) does not account for the hospitalized patients. On the other hand, the model used in ECDC (2020) is a very complex model with over 30 differential equations. An important goal policymakers should accomplish when compiling a strategy to tackle the spread of a pandemic is to ensure that the hospital system does not collapse. As the aim of this work is to assist policymakers with such decisions, to simulate a real-world scenario the SEIR-HCD model was picked. It also accounts for the number of Hospitalized citizens, patients in Critical condition, and the number of Deceased. The differential equations measuring the rate of change the number of people at each group is updated is given in the following Table. Differential equations are extracted from [1] and [2].

Name	Description	Differential Equation	Explanation of Features
Susceptible	Population not immune to the virus	$\frac{dS}{dt} = -\frac{R_t}{t_{inf}} \cdot IS$	Rt-> reproduction number t_inf -> infectious period
Exposed	Population currently in incubation	$\frac{dE}{dt} = \frac{R_t}{t_{inf}} \cdot IS - \frac{1}{t_{inc}} \cdot E$	t_inc -> incubation period
Infectious	Number of infections actively circulating	$\frac{dI}{dt} = \frac{1}{t_{inc}} \cdot E - \frac{1}{t_{inf}} \cdot I$	
Recovered	Population no longer infectious due to recovery	$\frac{dR}{dt} = \frac{m}{t_{inf}} \cdot I + \frac{1-c}{t_{hosp}} \cdot H$	m -> prob of having mild or no symptoms t_hosp -> days in hospital before recovering or moving to ICU
Hospitalized	People who developed severe symptoms and moved to hospital	$\frac{dH}{dt} = \frac{1-m}{t_{inf}} \cdot I + \frac{1-f}{t_{crit}} \cdot C - \frac{1}{t_{hosp}} \cdot H$	f -> prob of passing out t_crit -> days in ICU(critical condition)
Critical	Hospitalized people whose condition got critical and moved to ICU	$\frac{dC}{dt} = \frac{c}{t_{hosp}} \cdot H - \frac{1}{t_{crit}} \cdot C$	c -> fraction of severe cases that turn critical
Deceased	People who passed out after being in critical condition	$\frac{dD}{dt} = \frac{f}{t_{crit}} \cdot C$	

Figure 4.3: The differential equations composing the SEIR-HCD model

In order for SEIR-HCD to be successfully implemented some critical assumptions are required. First, the total population is assumed to remain stable at each time step and be equal to the total sum of people in each group. Next, people with mild symptoms and asymptomatics are assumed to have identical properties and do not require hospitalization. In addition, all people with severe symptoms are assumed to move to a hospital. Then they either recover or their condition becomes critical and they move to an ICU. Finally, the last assumption states that people who pass out have developed severe symptoms, moved to a hospital, then to an ICU and then they eventually passed out. The transmission from one group to another is given in Figure 4.4.

[1] https://github.com/gabgoh/epcalc/blob/master/src/App.svelte
[2] https://github.com/neherlab/covid19_scenarios

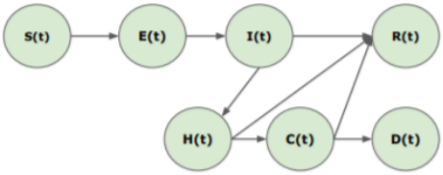

Figure 4.4: The transmission from one group to another in SEIR-HCD model.

Regarding the values for the constants in the differential equations two different options exist. They can either be calibrated or be given manually. Calibration can be viewed as a fitting procedure where the model fits on the input data and identifies the parameter combinations that minimize a given loss function. For the purpose of this work, the second approach is selected. The range of values these features can take are extracted from literature with the knowledge extraction model. For the model to be initialized, the initial states of each group along with the total population are given as input. For example (N , S(t) , E(t) , I(t) , R(t) , H(t) , C(t) , D(t)) = (Population , Population - Initial-infections, Initial-Infections, 0, 0, 0, 0, 0). Then by solving the system of differential equations, the number of people at each group at each time t is produced.

4.3 Parametric Uncertainty Visualization

As described, different visualizations are used for different parametric features extracted from Covid literature. On probabilistic features (including the probabilities of developing mild or no symptoms, the probability of one's condition becoming critical after severe symptoms, and the mortality rate after severe symptoms) according to Padilla et al. (2020) common approaches like probability density plots might create difficulties for users to accurately interpret the probability under the curve. Studies have pointed out that quantile dot plots are more useful for decisions under risk and give a more accurate description of uncertainty (Padilla et al. 2020). The user simply counts the dots and understands the area under the curve. As a result, in this work, quantile dot plots are preferred.

In addition, icon arrays offload the cognition of users when comparing probabilities. As a result, comparisons are easier and faster. Moreover, studies have shown that people trust icon arrays more than other common techniques (Hawley et al. 2008). For this reason, to visualize the uncertainty in probabilities, a combination between quantile dot plots and icon arrays has been selected. Each dot is replaced by an icon array of a person. Each icon represents a 10 % probability. An example is given in Figure 4.5

Figure 4.5: Visualising the uncertainty of mortality rate after developing critical symptoms

In the given example, the most probable outcome is that 5 out of 10 people with critical symptoms will eventually pass out. To interpret the example, the user simply has to count how many icons in a single column cover the most area under the density plot. In this example, five icons lie under the center of the curve. This can be translated into 5 out of 10 people being more likely to pass out. By adding or removing icons the equivalent probability decreases. These visualizations will be used for all probabilistic features.

Regarding hospitalization days, either with severe symptoms or in critical condition, the range of values depends on each country's unique characteristics. Examples of such characteristics include the quality of the hospital system or the number of drugs. For example, a country with high-quality ICUs could host a patient for more days, as they have the equipment to take care of them for a longer time. Another example is that patients hospitalized (not in ICU) might recover and leave the hospital faster if the country has higher quantities of drugs available. These examples can be viewed as less likely occasions. Other less likely occasions are countries with very bad hospital systems. To visualize these features, error bars that highlight less likely values, are picked to visualize this specific category of features.

As far as incubation and infectious days are concerned, these features represent some general dynamics of Covid-19 and they do not depend on countries' characteristics. For this reason, outliers are not so important to be pointed out. Following Padilla et al. (2020) suggestions, when a simple interval is to be shown and the most representative information is to be displayed, violin and gradient plots stand as a good option. By mapping the probability that a feature takes a specific value to the width at each position, users' attention is turned to values closer to the mean. In other words, they do not focus on less probable values found on the far left or far right. In this work, violin plots are picked, with extra width being given to the values closer to the mean of the range.

4.4 Data Driven Decision Making (DDDM) Framework

The goal of the Data-Driven Decision Making (DDDM) is to allow users to simulate the spread of the pandemic under different scenarios. Users are able to test the effect of different measures by forecasting the number of people in each group defined in Section 4.2. In addition, they can apply different combinations of hypothetical measures to halt the spread in different scenarios. In the following section, the core components of the framework are discussed.

4.4.1 Reproduction Number R_t

The reproduction number R_t plays a crucial role in estimating the spread of the virus inside a community. It reflects the changes in the disease transmission rate over time. In simple terms, R_t represents how many individuals does an infected person transmits the virus to. An R_t around (or below) 1 indicates that the spread is under control. A number well over 1 indicates that the situation is critical and gets out of control. Policymakers and public health officials use R_t to assess the effectiveness of interventions.

In this work, the initial reproduction number R_0 reflects the particular scenario a user would like to simulate. A R_0 over 2.5 indicates that the virus is well spread inside the community. In this scenario, users should identify a combination of measures, such as a lockdown or the obligatory use of masks, so the spread is reduced as soon as possible. A R_0 between 1 and 1.5 indicates that the situation is under control. Users should ensure the situation remains stable without applying intense measures. Before exploring the different mechanisms of the framework, users are recommended to pick an initial reproduction number R_0. This initiates the scenario to be simulated.

4.4.2 Interventions

In the next step, after picking the desired scenario and setting the values for the parametric features, a forecast is produced. An example of a scenario where R_0 is over 3 indicating high spread within the community is given in Figure 4.6

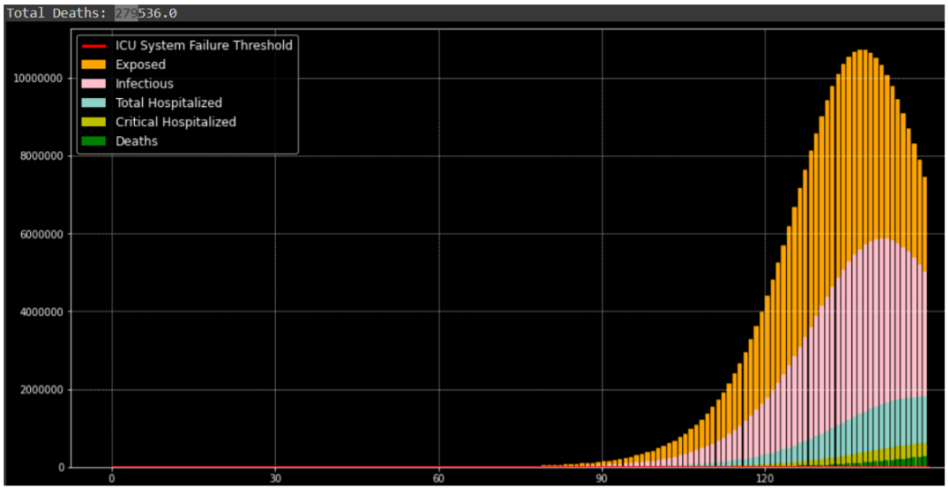

Figure 4.6: A nightmare scenario in Spain. Over 10,000,000 get infected, the hospital system collapses and almost 280,000 citizens pass out.

In order to prevent this catastrophic scenario, the policymaker should intervene and reduce the reproduction number R_0 of 3. Appropriate measures should be applied. With the designed framework, users are able to pick the date they would like to intervene. In addition, they can pick the duration of the applied measures. This results in a variety of measure combinations available for testing. Except for identifying the optimal date and duration, users are asked to pick the combination of the correct measures.

4.4.3 Utility Function - R_t Reduction Rate

As described on Gostic et al. (2020) and Syakur et al. (2018), the different approaches of policymakers aim at reducing the reproduction number. To do that, they should apply different measure combinations. In this work, two key assumptions are considered. Under these assumptions, the R_t Reduction Rate (RtRR) formula is defined. It reflects the measures applied by policymakers to stop the spread of the pandemic into the reproduction number R_t. The first assumption states that every action applied by policymakers has an impact on the reproduction number. In accordance with Gostic et al. (2020) and Syakur et al. (2018), such actions aim at reducing the existing R_t.

The second key assumption is that the virus is solely transmitted from a person to a person after they have been in close contact. The formula for the R_tRR is inspired by ECDC (2020) and Noll et al. (2020). A limitation of the defined formula is that it is constructed and evaluated empirically. It has no theoretical background. However, in practice, it produces realistic results. The formula assumes that the transmission rate and the average number of contacts per day are the only mechanisms that can be tune in order to tackle the spread of the pandemic. In addition, the strictness of the measures determines the degree to which measures are applied. It should also be mentioned that too high a value for strictness for a long period of time would result in public disappointment. The formula for the R_tRR is given below:

$$R_tRR = (1-b)^{c*a}, \tag{4.1}$$

where b is the transmission rate, c the average contacts per day, and a the government's strictness.

The default values for the above parameters are 0.9 for a, 13 and 0.9 for c and b respectively. A lower value for a points out to a more intense application of the selected measures. It should be mentioned that governments' strictness can not drop lower than 0.5. The new R_0 number is given by :

$$NewR_0 = R_0 - R_tRR * R_0 \tag{4.2}$$

Users can pick different values for a, b, and c that reflects different combinations of measures. Some examples are given in Table 4.1. Different measure combinations are useful in different scenarios.

Measures description	Transmission Rate	Contracts per day	Strictness, a
• Total Lockdown • obligatory masks (everywhere)	b = 0.1 Usage of masks reduces the transmission rate	c = 3. People will only contract their families	a = 0.5 . Governments will be extremely strict
• Work from home • Indoors Masks • Light Recommendations for measures application	b = 0.2 Masks only outdoors reduces transmission rate	c = 8 People don't meet colleagues but can meet friends in restaurants	a = 0.7 Governments give light recommendations
• Work from home • Closed Bars • Indoors Masks(and crowded places) • Heavy Recommendations	b = 1.5 Masks indoors and on crowded places but not outdoors	c = 5 People work from home, can't go to bars but they can still visit friends' houses	a = 0.6 Governments give heavy recommendations and possible fees.

Table 4.1: Possible combinations of measures. Users can either pick from a predefined list of recommendations are use their combinations

It should be added that when users have picked a date and the type of intervention, the transition from the new to the old R_0 is not instant. This is not realistic. Another key assumption taken is that all measures require at least 14 days before they are reflected in the new forecasts. As a result, and by considering that R_t is time-varying, the following linear decay function is manually defined.

$$R_t = R_0 - \frac{(R_0 - NewR_0)}{14} * (t - t_0), \tag{4.3}$$

where t_0 is the intervention date. The formula is used to determine the value of R_t at each time step t. When the complete duration of the measures has passed the R_t is assumed to remain constant and equal to NewR_0. This assumption is not realistic and creates issues with the proper application of some measures. Users are recommended to initiate a new scenario with a new initial R_0.

5 | Results and DSS Prototype

In this chapter, the final DSS is presented and its usage is demonstrated in three different scenarios. The final design of the framework is still under construction so all features are included. For the purpose of this report, the presented figures are from the rough draft of the framework.

5.1 Layout of the Framework

A potential layout format for the framework is given in Figure 5.1. Users are asked to pick a country and an initial reproduction number R_0 to initiate the framework. Then they can modify the given parameters and pick a different scenario.

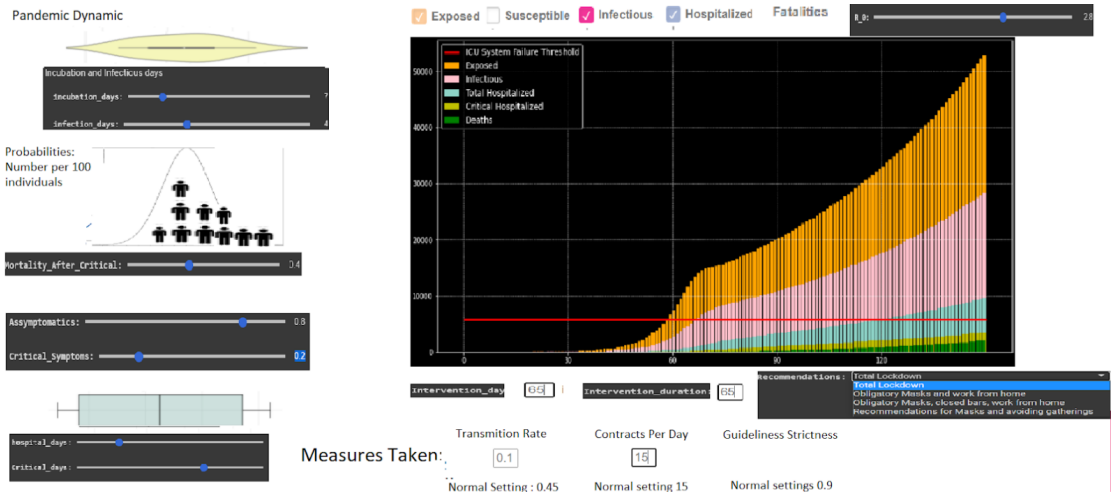

Figure 5.1: Rough draft of the DSS network. On the left side, the parametric uncertainty is visualized. On the right side the measures menu along with the forecast simulations are given. On the top right, users pick their initial scenario by manipulating the R_0 slider

5.2 Usage of the Framework

We demonstrate three different usage scenarios. In this first example, we assume that the virus has already been spread in the community. For this scenario, the initial reproduction number is set to 3. For this example, Spain is picked as the simulated country. The country has a

total population of 60,000,000 and a total of 5,900 ICU beds. The values for the parametric mechanism are set to the recommended mean values. The critical situation the country might face, if no intervention takes place, is given in Figure 4.6. In the first example, we assume that the government of Spain tests the effect of the second proposed option of Table 4.1 on day 50 for 30 days. The results are given in Figure 5.2.

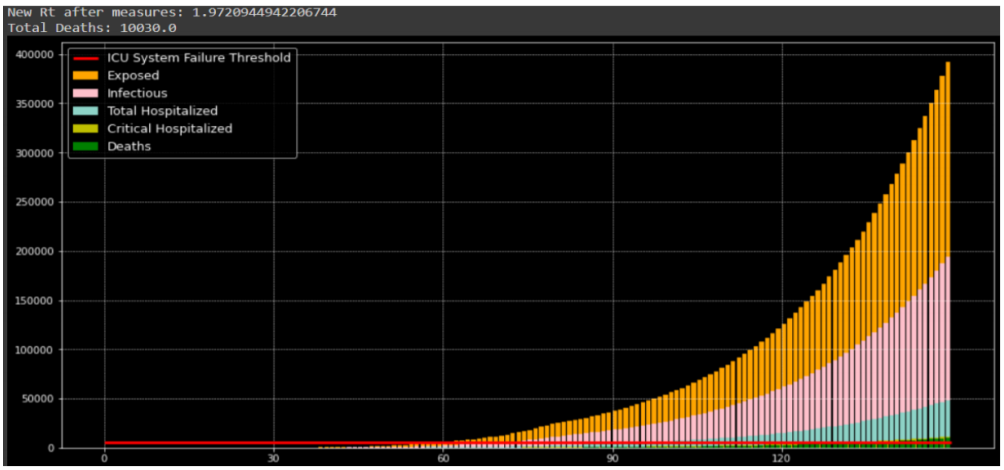

Figure 5.2: Example of recommended option 2, applied on day 50 for a total of 30 days.

Despite that, the worst-case scenario is avoided, the measures do not produce the expected results. The pressure on the ICU system is very heavy and over 10.000 eventually pass out. A faster intervention on day 30 would have produced better results as only 1474 people would have died and the pressure would have been much softer.

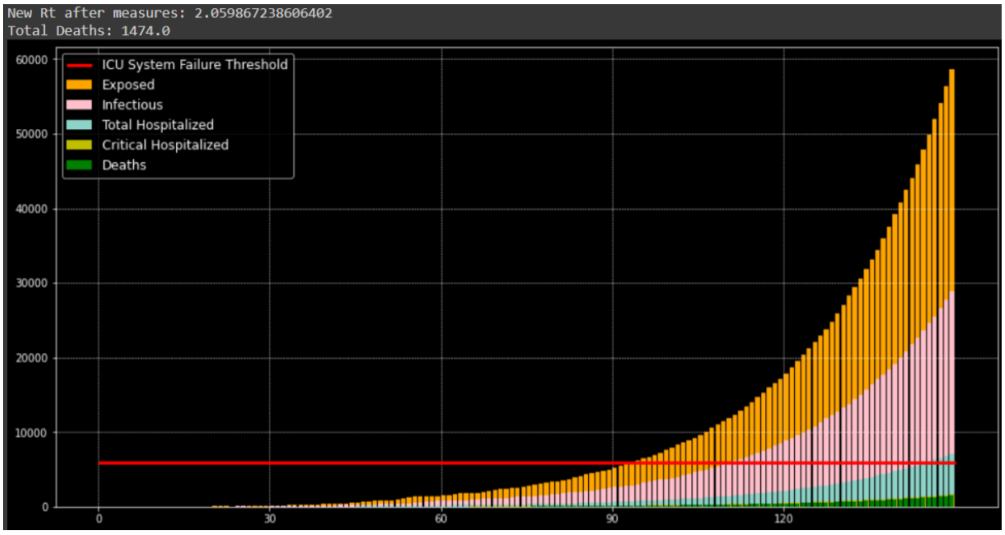

Figure 5.3: Example of recommended option 2, applied on day 30 for a total of 30 days.

Another decision for this scenario is option 1. A total lockdown for a total of 14 would have

halted the spread.

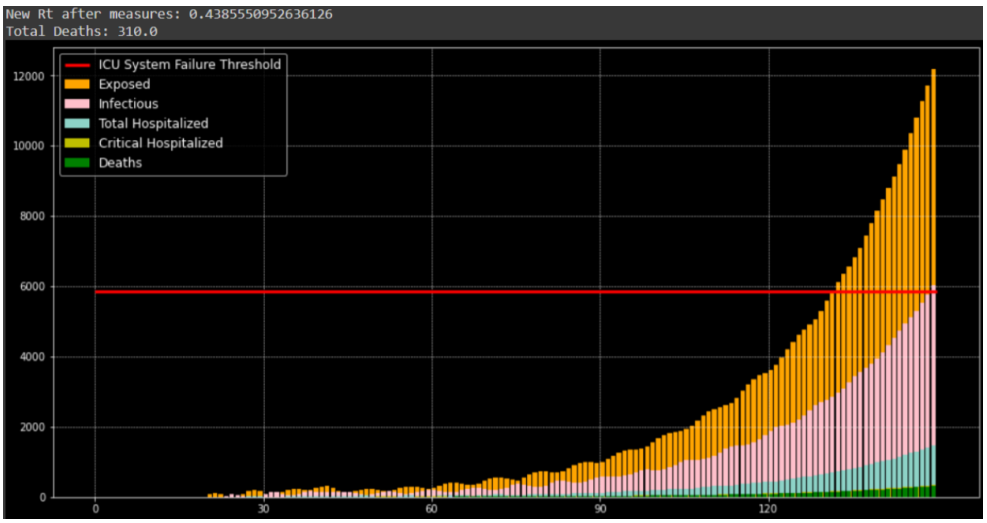

Figure 5.4: The results of option 1 applied on day 30 for a total of 14 days

A total of just 310 people would have passed out. In addition, the spread of the virus would have stopped and the new R_t would end up being below 0.5. Another example with an initial reproduction number standing between 1.5 and 2 indicates that the spread within the community is not as high. In this case, option 2 would have given better long-term results as presented in Figure 5.5.

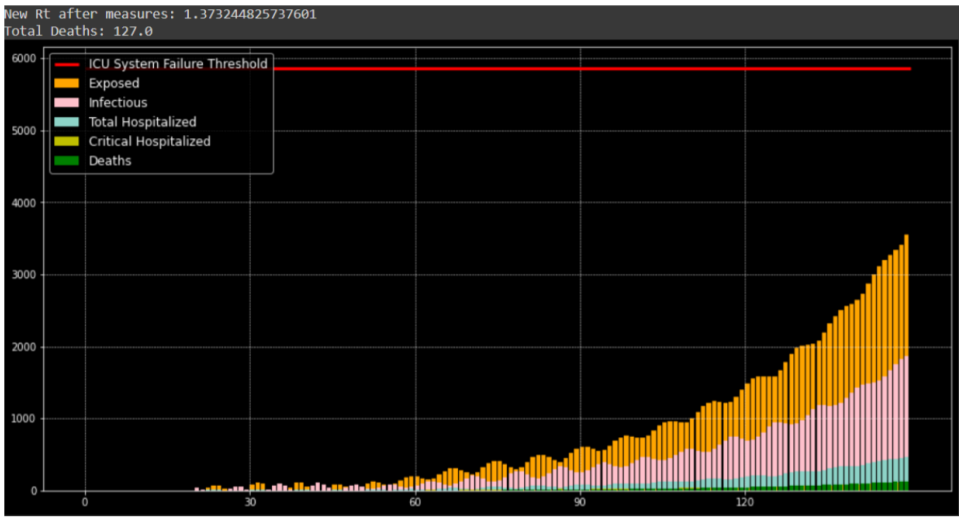

Figure 5.5: Option 2 applied on a new scenario with an initial reproduction number of 2

Option 2 seems to work better when the virus load is not as high. Pressure on the ICU would not be high and the virus would evenly spread for the course of two months. As the aforementioned examples show, different interventions at different times with different duration

periods produce different results. As a result, for each picked scenario, a different combination of measures should be applied.

5.3 Discussion: Decision Making

For each scenario, the goal of the decision-makers using the framework is to find the combination of the optimal measures while accounting for the parametric uncertainty. This is a multi-criteria decision-making process. Policymakers have to fulfill different goals. They should ensure the long-term gain by "flattening the curve" and make sure that the virus is evenly spread in the community without getting out of hand. Most importantly though, they should avoid a hospital system failure. Finally, they should not reduce the max strictness of the applied measures for a long period of time to prevent socioeconomic consequences.

To make sure that every goal is successfully fulfilled, policymakers could follow different strategies. For example, during the first wave of the pandemic, Sweden followed a normative decision-making approach. They focused on the first goal and aimed for long-term gain. They did not pay as much attention to short-term losses as other countries. In the meantime, they made sure the hospital system was able to cope up with their strategy. On the other hand, Greece followed a different approach. Their approach is closer to descriptive decision-making theory. Avoiding any losses (defined as the number of deaths) was prioritized over the long-term gain reflected in an even spread of the virus.

As multi-criteria decision making is a subsection of normative decision making (Klein et al. 1993), users of the framework are suggested to follow a normative decision making approach. They should aim at optimizing the utility function (R_tRR function) for the selected scenario. In addition, users should make sure that the hospital system of each country can withstand the picked measures. It should also be mentioned that in some scenarios, despite that strict measures are viewed as the most rational decision, abusing them would result in failing the third goal. Socioeconomic consequences should be avoided. From my point of view, prioritizing the minimization of losses in the short term does not ensure long-term gains. The virus would still remain spread in the community and further action might be needed in the future. In this case, policymakers would struggle to use strict measures in the near future as the public might react. For this reason and from my point of view, the most rational decision would be a smooth spread of the virus in the community. The aim is to not allow the hospital system to collapse and avoid multiple losses. Under these conditions, normative decision-making approaches are suggested.

6 | Conclusion

To conclude, this project describes a DM Framework designed for Covid-19 policymakers to tackle the spread of the pandemic. The aim is to test different scenarios and find the appropriate set of measures while considering the uncertainty of fundamental Covid-19 dynamics. Current literature includes inconsistencies regarding the core mechanisms of the virus. These inconsistencies can lead to accuracy defects as forecasting models might be negatively affected. The proposed framework visualizes the described uncertainty. This work suggests that users should try different set-ups and modifications to deal with the potential parametric uncertainty.

The project is based on several realistic and unrealistic assumptions. It should also be noted that due to limited time and computational resources, not all concepts related to the spread of the virus have been taken into account. In some future work the current SEIR - HCD would be updated. The goal is to host more parameters and to also consider the unique demographic characteristics of each country. In addition, more than one intervention would be added, so the long-term spread of the virus is estimated. Finally, the assumption that R_t remains constant and equal to NewR_0 after the duration of the intervention has passed should also be addressed. It is unrealistic and might create simulation issues with the application of some measures.

Due to space limitations, the impact different parametric values have on the final forecasts is not demonstrated. However, by exploring the differential equations described in Figure 4.3 the importance of the parametric input values is quickly highlighted. The uncertainty levels within Covid-19 literature are pretty high and as a result, the final predictions are far from perfect. Thus, users are encouraged to try different combinations of values and test all the different hypothetical plots for every scenario. This was they will be able to deal with unexpected surprises. The framework is also suggested for non-experts to explore the impact different parameters have on the spread of the pandemic

The designed framework can be implemented in any country and can be calibrated, as described in Section 4.2. Under this set-up, the uncertain parameters are optimized with regard to a country's current situation. In addition, parameters can take any initial values. As a result, the framework can be implemented to test the impact of measures at any time during the course of the pandemic. Moreover, the framework can be generalized and be used for any disease or critical situation within a country. Updating the equations of the SEIR-HCD model described in 4.3 is the only modification necessary.

Finally, depending on the decision-making process policymakers follow, they can try different combinations of methods for each scenario. As described, both normative and descriptive decision-making theories are applicable. However, the framework is designed to simultaneously fulfill different goals. Hence, the usage of descriptive theory would be problematic as it focuses entirely on reducing the number of deaths. As a result, normative decision-making is suggested. The aim is to identify the most rational measures for each scenario while considering the effect of the parametric uncertainty. Multiple simulations of the same scenario under different parametric setups are suggested to tackle the high uncertainty levels.

Bibliography

Anupriya, P. & Karpagavalli, S. (2015), LDA based topic modeling of journal abstracts, *in* '2015 International Conference on Advanced Computing and Communication Systems', pp. 1–5.

Bastani, K., Namavari, H. & Shaffer, J. (2019), 'Latent dirichlet allocation (lda) for topic modeling of the cfpb consumer complaints', *Expert Systems with Applications* **127**, 256–271.
URL: *https://www.sciencedirect.com/science/article/pii/S095741741930154X*

Btoush, M., Alarabeyyat, A. & Olab, I. (2016), 'Rule based approach for arabic part of speech tagging and name entity recognition', *International Journal of Advanced Computer Science and Applications* **7**.

Carcione, J., Santos, J., Bagaini, C. & Ba, J. (2020), 'A Simulation of a COVID-19 Epidemic Based on a Deterministic SEIR Model', *Frontiers in Public Health* **8**, 230.

Chiticariu, L., Li, Y. & Reiss, F. (2013), Rule-based information extraction is dead! long live rule-based information extraction systems!, *in* 'EMNLP'.

Dong, X. & Hayes, C. (2012), 'Uncertainty Visualizations: Helping Decision Makers Become More Aware of Uncertainty and Its Implications', *Journal of Cognitive Engineering and Decision Making* **6**, 30–56.

ECDC (2020), 'Baseline projections of COVID-19 in the EU/EEA and the UK: update'.
URL: *https://www.ecdc.europa.eu/en/publications-data/baseline-projections-covid-19-eueea-and-uk-update*

Fernandes, M., Walls, L., Munson, S., Hullman, J. R. & Kay, M. (2018), Uncertainty displays using quantile dotplots or CDFs improve transit decision-making, *in* 'CHI 2018 - Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems: Engage with CHI', Association for Computing Machinery.
URL: *https://www.scholars.northwestern.edu/en/publications/uncertainty-displays-using-quantile-dotplots-or-cdfs-improve-tran*

Ghamizi, S., Rwemalika, R., Cordy, M., Le Traon, Y. & Papadakis, M. (2020), Pandemic Simulation and Forecasting of exit strategies:Convergence of Machine Learning and EpidemiologicalModels, Technical report, University of Luxembourg.
URL: *https://orbilu.uni.lu/handle/10993/43166*

Ghoshal, B. & Tucker, A. (2020), 'Estimating Uncertainty and Interpretability in Deep Learning for Coronavirus (COVID-19) Detection', *arXiv:2003.10769 [cs, eess, stat]* . arXiv: 2003.10769.
URL: *http://arxiv.org/abs/2003.10769*

Gostic, K. M., McGough, L., Baskerville, E. B., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J., De Salazar, P. M., Hellewell, J., Meakin, S., Munday, J., Bosse, N. I., Sherrat, K., Thompson, R. N., White, L. F., Huisman, J. S., Scire, J., Bonhoeffer, S., Stadler, T., Wallinga, J., Funk, S., Lipsitch, M. & Cobey, S. (2020), 'Practical considerations for measuring the effective reproductive number, Rt', *medRxiv : the preprint server for health sciences* .
URL: *https://europepmc.org/articles/PMC7325187*

Hawley, S. T., Zikmund-Fisher, B., Ubel, P., Jancovic, A., Lucas, T. & Fagerlin, A. (2008), 'The impact of the format of graphical presentation on health-related knowledge and treatment choices', *Patient Education and Counseling* **73**(3), 448–455.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0738399108003431*

He, S., Peng, Y. & Sun, K. (2020), 'SEIR modeling of the COVID-19 and its dynamics', *Nonlinear Dynamics* **101**(3), 1667–1680.
URL: *https://doi.org/10.1007/s11071-020-05743-y*

Joby Mackolil & Mahanthesh, B. (2020), 'Mathematical Modelling of Coronavirus disease (COVID-19) Outbreak in India using Logistic Growth and SIR Models'.
URL: *https://www.researchsquare.com/article/rs-32142/v1*

Joslyn, S. & LeClerc, J. (2013), 'Decisions With Uncertainty: The Glass Half Full', *Current Directions in Psychological Science* **22**(4), 308–315. Publisher: SAGE Publications Inc.
URL: *https://doi.org/10.1177/0963721413481473*

Klein, G. A., Orasanu, J. & Caldenwood, R. (1993), *Decision Making in Action: Models and Methods*, Praeger, Norwood, N.J.

Liu, Q., Wang, J., Zhang, D., Yang, Y. & Wang, N. (2018), Text features extraction based on tf-idf associating semantic, *in* '2018 IEEE 4th International Conference on Computer and Communications (ICCC)', pp. 2338–2343.

Mishra, S., Sarkar, U., Taraphder, S., Datta, S., Swain, D. P., Saikhom, R., Panda, S. & Laishram, M. (2017), 'Principal component analysis', *International Journal of Livestock Research* p. 1.

Mwalili, S., Kimathi, M., Ojiambo, V., Gathungu, D. & Mbogo, R. (2020), 'SEIR model for COVID-19 dynamics incorporating the environment and social distancing', *BMC Research Notes* **13**.

Noll, N. B., Aksamentov, I., Druelle, V., Badenhorst, A., Ronzani, B., Jefferies, G., Albert, J. & Neher, R. (2020), 'COVID-19 Scenarios: an interactive tool to explore the spread and associated morbidity and mortality of SARS-CoV-2', *medRxiv* p. 2020.05.05.20091363. Publisher: Cold Spring Harbor Laboratory Press.
URL: *https://www.medrxiv.org/content/10.1101/2020.05.05.20091363v2*

Padilla, L., Kay, M. & Hullman, J. (2020), Uncertainty Visualization, Technical report, PsyArXiv. type: article.
URL: *https://psyarxiv.com/ebd6r/*

Paiva, H. M., Afonso, R. J. M., Oliveira, I. L. d. & Garcia, G. F. (2020), 'A data-driven model to describe and forecast the dynamics of COVID-19 transmission', *PLOS ONE* **15**(7), e0236386. Publisher: Public Library of Science.
URL: *https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0236386*

Petropoulos, F. & Makridakis, S. (2020), 'Forecasting the novel coronavirus COVID-19', *PLOS ONE* **15**(3), e0231236. Publisher: Public Library of Science.
URL: *https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0231236*

Qaiser, S. & Ali, R. (2018), 'Text mining: Use of tf-idf to examine the relevance of words to documents', *International Journal of Computer Applications* **181**, 25–29.

Saaty, T. (1980), 'The analytic hierarchy process'.

Simon Rozendaal. (2020), 'The variability of critical care bed numbers in Europe prepare for covid-19?'.
URL: *https://www.covid-19.no/critical-care-bed-numbers-in-europe*

Subeno, B., Kusumaningrum, R. & Farikhin, F. (2018), 'Optimisation towards latent dirichlet allocation: Its topic number and collapsed gibbs sampling inference process', *International Journal of Electrical and Computer Engineering (IJECE)* **8**, 3204.

Syakur, M., Khotimah, B., Rohman, E. & Dwi Satoto, B. (2018), 'Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster', *IOP Conference Series: Materials Science and Engineering* **336**, 012017.

Unlu, E., Leger, H., Motornyi, O., Rukubayihunga, A., Ishacian, T. & Chouiten, M. (2020), 'Epidemic analysis of COVID-19 Outbreak and Counter-Measures in France', *medRxiv* p. 2020.04.27.20079962. Publisher: Cold Spring Harbor Laboratory Press.
URL: *https://www.medrxiv.org/content/10.1101/2020.04.27.20079962v1*

Verma, P., Khetan, M. & Dwivedi, S. (2020), 'Forecasting the covid-19 outbreak: An application of arima and fuzzy time series models'.

World Health Organization(WHO) (2020), 'Communicating and Managing Uncertainty in the COVID-19 Pandemic: A quick guide'.