University of Southampton
Faculty of Physical and Applied Sciences
Electronics and Computer Science

Causal Reasoning in Machine Learning

Author: Pier Paolo Ippolito

Project Supervisor: Srinandan Dasmahapatra
Second Examiner: Age Chapman
September 16, 2020

A dissertation submitted in partial fulfilment of the degree of
MSc in Artificial Intelligence

“Fortunate is he, who is able to know the causes of things”

— Virgil, Verse 490 of Book 2 of the ”Georgics” (29 BC).

Acknowledgements

I would like to thank my supervisor Dr. Srinandan Dasmahapatra, for giving me the opportunity to do this project. This experience enabled me to broaden my knowledge on a variety of topics and skills which will play a crucial role in my upcoming career.

Finally, I express my gratitude to my second examiner, Dr. Age Chapman, for her support throughout the project.

Statement of Originality

– I have read and understood the ECS Academic Integrity information and the University's Academic Integrity Guidance for Students.

– I am aware that failure to act in accordance with the Regulations Governing Academic Integrity may lead to the imposition of penalties which, for the most serious cases, may include termination of programme.

– I consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to verify whether my work contains plagiarised material, and for quality assurance purposes.

I have acknowledged all sources, and identified any content taken from elsewhere.

During the course of this project I followed different online tutorials and courses. All of them have been acknowledged in the body of the report and referenced in the bibliography (such as: [12], [23], [24], [26], [39]). Any minor additional source used, has been acknowledged in the code files available in the Project Design Archive as a comment.

I did all the work myself and I have not helped anyone else.

The material in the report is genuine, and I have included all my data/code/designs.

Part of the content from assignments in the "ELEC6211 Project Preparation" module was reused in this report according to the instructions outlined on the "COMP6200 MSc Project" assignment page.

My work did not involve human participants, their cells or data, or animals.

Abstract

Nowadays Machine Learning models, are able to learn from data by identifying patterns in large datasets. Although, humans might be able to perform a same task after just examining a few examples. This is possible thanks to the inherit humans ability to understand causal relationships and use inductive inference in order to assimilate new information about the world. Creating models able to demonstrate causal reasoning would therefore open a whole new world of opportunities in Artificial Intelligence research. Causality arises naturally in our daily life every time we ask ourselves any type of interventional or retrospective question (eg. What if I take this action? What if I acted differently?).

Causality has been researched and used for many years in statistics but not in great depth in Artificial Intelligence. Identifying useful connections between these two different ambit could therefore play a vital role in making a breakthrough towards creating intelligent systems. Enabling Machine Learning models to be more easily examinable to gain insights of their decision making processes could in fact facilitate adoption of these kind of technologies in fields such as medicine, surveillance and recruitment.

List of abbreviations

AI = Artificial Intelligence

ML = Machine Learning

PPS = Predictive Power Score

MAE = Mean Absolute Error

Covid-19 = Coronavirus

SCM = Structural Causal Models

I.Q. = Intelligence Quotient

IID = Independent and identically distributed

BBN = Bayesian Belief Networks

ODE = Ordinary Differential Equation

PDE = Partial Differential Equation

NLP = Natural Language Pre-processing

VADER = Valence Aware Dictionary and sEntiment Reasoner

RMSE = Root Mean Squared Error

L-BFGS = Limited-memory BFGS

ARIMA = Auto Regressive Integrated Moving Average

AIC = Akaike Information Criterion

LSTM = Long-Short-Term-Memory

RNN = Recurrent Neural Network

T-CNN = Temporal Convolutional Neural Network

MAPE = Mean Absolute Percentage Error

PDF = Probability Density Function

CDF = Cumulative Density Function

CI = Continuous Integration

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Motivation

Thanks to recent advancements in Artificial Intelligence (AI), we are now able to leverage Machine Learning and Deep Learning technologies in both academic and commercial applications. Although, relying just on correlations between the different features, can possibly lead to wrong conclusions as correlation does not necessarily imply causation.

Three of the main limitations of Machine Learning and Deep Learning models are:

- **Robustness**: trained models might not be able to generalise to new data and therefore would not be able to provide robust and reliable performances in the real world.

- **Explainability**: complex Deep Learning models can be difficult to analyse in order to clearly demonstrate their decision making process.

- **Data Dependency**: Deep Learning models efficiency is highly dependent on the amount and quality of data available.

Developing models able to identify cause-effect relationships between different variables might ultimately offer a solution to solve these problems. This idea has also been supported by researchers such as Judea Pearl and Jonas Peters, whom advocated having models able to reason in uncertainties could not be enough to enable researchers to create machines able to truly express intelligent behaviour [1].

1.2 Going Beyond Correlation

Paradoxes are a classes of phenomena which arise when, although starting from premises known as true, we derive some sort of logically unreasonable result. One of the most common form of paradox in Data Science is **Simpson's Paradox**.

As an example, let us consider a thought experiment: we carried out a research study in order to find out if doing daily physical exercises can help or not reduce Cholesterol levels (in mg/dL) and we are now starting to examine the obtained results. First, we divide our population sample into two main categories based on the individuals age (under/over 60 years old) and then we plot their cholesterol levels against the number of hours the subjects exercised per day. By examining the results in the first two plots of Figure 1.1, we can then infer that exercising for more hours a day can then lead to an overall reduction in our cholesterol levels. This hypothesis can then also be reinforced by examining the overall trend of the best fit line inferred through Linear Regression and the quite strong negative Person Correlation scored in both cases. At this point, reassured by our derived results, we can then try to repeat this same analysis taking into consideration this time the whole population sample (rightmost plot in Figure 1.1). In this case, we are faced with a completely contradictory scenario and a positive correlation implying that more exercise can lead to increased cholesterol levels.

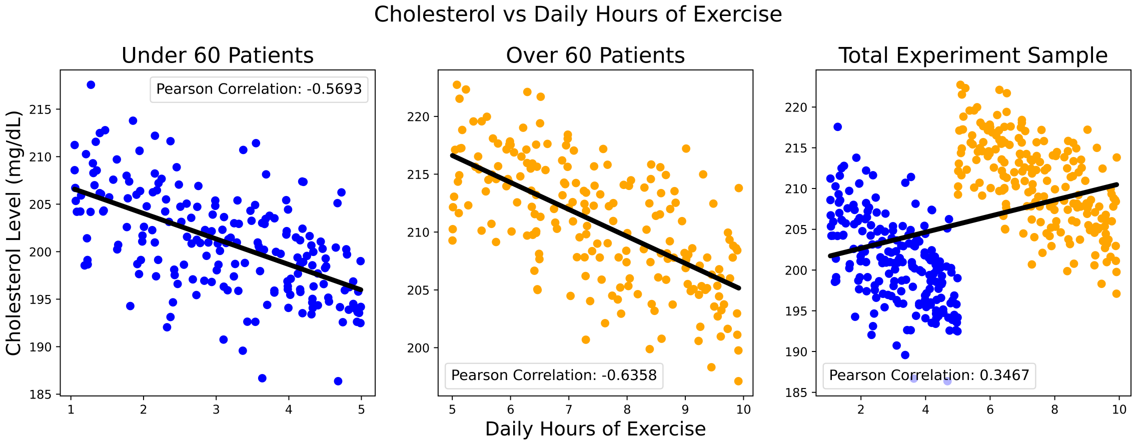


Figure 1.1: Cholesterol vs Daily Hours of Exercise

This type of scenario is commonly known as Simpson's Paradox and takes place everytime we have some form of correlation which points in a direction when considered in a sub-group and points instead in the opposite direction if considered as part of the whole group. In order to unveil the reasons behind this type of mechanism, we need to try to go beyond the provided data and think about how our data was generated in the first place to **cause** this outcome (e.g. what unknown missing variable might be preventing us to see the full picture?).

In this simple scenario, our missing component could be any potentially influential variable such as: individual's comorbidities, diet and age. We decide then to take a closer look on how cholesterol levels vary with greater age (Figure 1.2). Repeating the same analysis done in Figure 1.1, we can then clearly see how cholesterol levels are strongly positively correlated to individual's age.

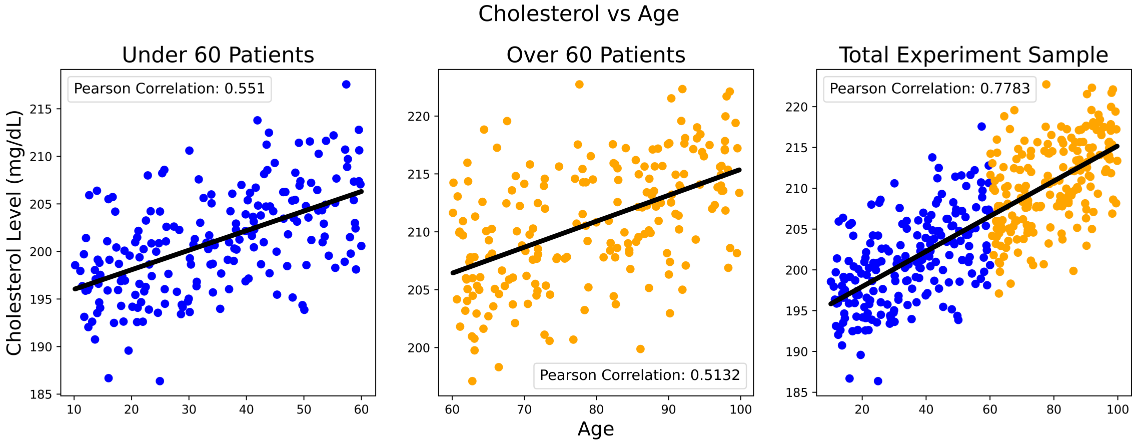


Figure 1.2: Cholesterol vs Age

From these results, we can then deduce that cholesterol levels are more likely to increase with aging and lack of exercise (there is a cause effect relationship between the three variables). Therefore, in order to try to quantify the benefits of exercising in reducing cholesterol levels and overcome the Simpson Paradox, we should then make sure to run our experiment while having a fixed value for the age of the subjects (**controlling the variable**).

During the course of the last century, the Simpson Paradox occurred in many statistical studies such as: UC Berkeley gender bias, Kidney stone treatment and Racial disparity in the death penalty [2]. Other common examples of statistical/mathematical paradoxes are the Monty Hall Problem, the Berkson's Paradox and the Accuracy Paradox.

Additional information about technical limitations of correlation and possible alternative metrics which have been designed in order to overcome this type of problems is available in Appendix B.

1.3 Causality vs Explainability

One of the major trade-offs in modern day Machine Learning is model performance against complexity. In fact, complex Deep Learning architectures are usually able to perform better in a wide variety of tasks compared to traditional linear classifiers and regression techniques. This trade-off has been analysed in-depth in the 2016 publication ”Why should I trust you?” by Ribiero et. al. [3] and led a new trend in AI to focus on interpretability.

Complex and more accurate models are referred to as **Black-boxes**. These type of models working progresses are more difficult to comprehend and they are not able to estimate the importance of each feature and how they are related to each other. Some examples of Black Boxes models are neural networks and ensemble models.

On the other hand, simpler and less accurate models such as decision trees and linear regression are instead regarded as **White-boxes** and can be much more interpretable. Two of the main measures which can be used in order to estimate the explainability of a model are the linearity and monotonicity of a model response function [4].

One of the key differences between Explainable AI and Causal AI is that the former aims just to understand how a model might come to a prediction by weighting the provided features while the latter is designed undercover the process governing the system we are analysing to create insights. In this way, Causal AI can be used in order to answer common retrospective and system design types of questions, providing vital business value to organizations (e.g. EU’s General Data Protection Regulation, right to explanation clause) [5].

1.4 Foundations of Modelling and Simulations

Modelling and Simulations is a branch of mathematics which aim to be able to imitate real-world processes over a period of time. In this way, artificially generated historical data can be easily created and used in order to make inference in real-world applications. Simulation models are usually based on a series of simplifying assumptions (of the real-world environment) which can then be expressed in a mathematical or symbolic notation [6].

There are two main types of programmable simulation models:

- **Mathematical Models**: make use of mathematical symbols and relationships in order to summarise processes. Compartmental Models in Epidemiology are a typical example of mathematical models.
- **Process Models**: are based on a list of steps handcrafted by the designer in order to represent an environment (e.g. Agent Based Modelling).

Modelling and Simulations, are used in many different fields such as finance (e.g. Monte Carlo Simulations for Portfolio Optimization), medical/military training, epidemiology and threat modeling [7, 8].

Some of the main uses of simulations is to verify analytical solutions, experiment policies before creating any physical implementation and understand the connection and relative importance of the different variables composing a system (e.g. by modifying input parameters and examining the results). As a result, these properties makes the Modelling and Simulations paradigm a **white-box** approach to predict future trends.

1.5 Objectives

As part of this research study, it will be outlined the main principles of Causal Reasoning, different application approaches (e.g. Bayesian Belief Networks, Time Series Analysis) and an Epidemic Modelling case study concerning COVID-19 (Coronavirus).

The Novel Coronavirus, is a new type of RNA virus which is able to infect humans potentially causing respiratory infections. The 2020 Novel Coronavirus outbreak started in the late 2019 in Wuhan, China and, as of August 2020, it is believed the virus is mainly able to spread by air through sneezing and coughing.

The proposed Compartmental Models, will be based on paradigms defined in the Epidemiology literature [9] and the Imperial College of London COVID-19 report studies used by the UK government in order to handle the outbreak [10]. The Agent Based Models have instead been handcrafted in order to provide an alternative approach to traditional mathematical model implementations including different elements of stochasticity due to non-linear interactions at a population level.

Due to the design of the proposed models, they can potentially provide greater help for decision makers (compared to traditional Machine Learning approaches), if and only if, the decision makers in question have the necessary understanding of epidemiology and its key control metrics.

Finally, these models have been exclusively designed for educational and research purposes and are not to be applied in any other ambit (e.g. commercial, governmental).

Chapter 2

Background Theory

2.1 Concepts of Causality

Current supervised Machine Learning techniques are designed to exploit possible relationships/correlations between features and labels in order to produce reliable estimates. Use of this kind of technologies in sectors such as medicine, finance and law is now raising increasing concerns due to the lack of ability in such systems to correctly identify causal relationships and provide explanations about their decisions. One possible solution in order to overcome these type of problems is by taking into account causal relationships.

Causality arises naturally in our daily life every time we ask ourselves any type of interventional or retrospective question (eg. What if I take this action? What if I would have acted differently?).

As shown in Figure 2.1, Causal Reasoning can be divided into three different hierarchical levels (Association, Intervention, Counterfactuals). At each level, different types of questions can be answered and in order to answer questions at the top levels (eg. Counterfactuals) it is necessary to have a base knowledge from the lower levels [11]. In fact, in order to answer retrospective questions, we would expect to first be able to respond to intervention and association type of questions.

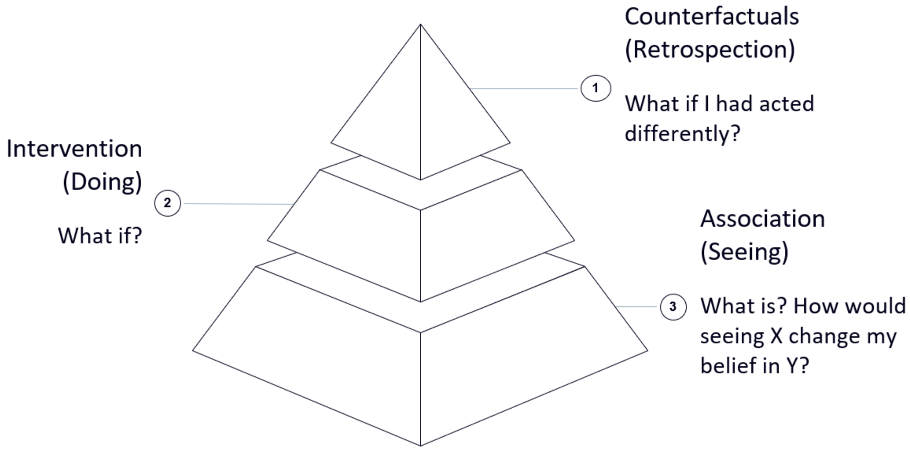

Counterfactuals
(Retrospection)
1
What if I had acted differently?
Intervention
(Doing)
2
What if?
Association
(Seeing)
3
What is? How would seeing X change my belief in Y?

Figure 2.1: Causality Hierarchy

Currently, Machine Learning models are only able to answer the probabilistic type of questions related to the Association level. Thanks to the rising interest in this topic, a mathematical framework able to represent causal relationships has been constructed (Structural Causal Models (SCM) [11]). Using this type of framework, causal expressions can then be formulated and used in conjunction with data in order to make predictions.

This type of framework can then be divided into two main parts: causal diagrams and a symbolic language. The causal diagrams can be used in order to summarise our knowledge about the topic, while the symbolic language can be used to express what we are aiming to find out.

As an example, let us consider the diagram shown in Figure 2.2. Using this type of representation, the arrow directions indicate how the different variables effects each other.

In this example, a survey is carried out between individuals of age 3-20 in order to find out if there is any correlation between height and individuals' Intelligence Quotient (I.Q.) Scores. Although the study might result in a positive correlation between the two different variables, a more in depth analysis might instead show how height does not directly cause higher I.Q. Scores but these two variables are instead dependent on a third hidden variable (Confounder). In fact, as children grow up, over time both their I.Q. Scores and height tends to increase due to their improved education and greater age.

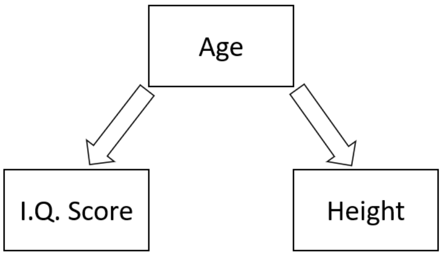


Figure 2.2: Causality Diagram

In case we want to query additional information from what we currently have available [i], we can then make use of the symbolic language in order to advance questions

[i] So that to move from the association to the intervention level in the causality hierarchy.

such as: What is the probability (P) that a student will get an higher I.Q. score (S) if he studies an additional amount of time (T)? This question could then be formulated in a symbolic form such as $P(S|do(T))$ [ii]. When formulating these type of questions, we are then implying that we are not anymore passively observing possible results but instead actively intervening in order to find out about possible consequences. This type of approach is known as an Interventional Study and is in contrast with traditional Observational Studies.

Finally, in order to create a full Causal Inference Engine, an architecture like in Figure 2.3, might be necessary [12]. Following this type of approach, three inputs are needed and three outputs are produced. Our three inputs are: any given assumption made about the model (**Assumptions**), any questions we are trying to answer (**Queries**) and any data which can be used in order to fuel our engine (**Data**). The model will then output a Boolean value to show if is able or not to answer the given queries, and if so, it would provide as second output a mathematical formula which can be used in order to answer the queries. Finally, a numerical prediction specific to the given input data is provided (this might contain some form of uncertainty in the estimation given the amount of data provided and assumptions made).

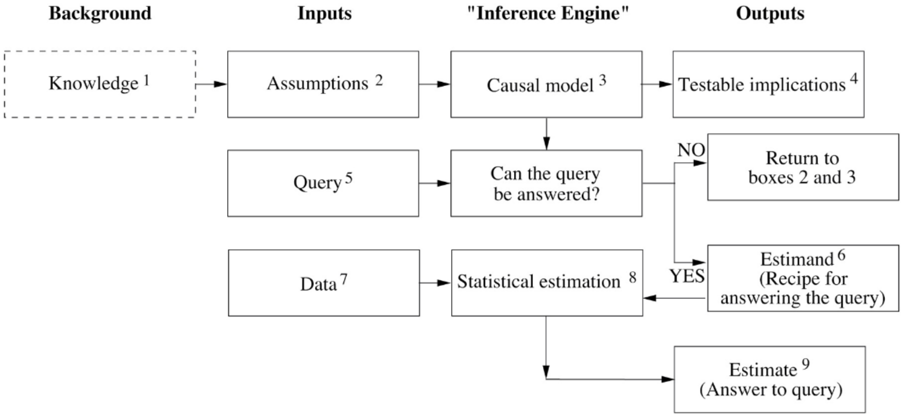


Figure 2.3: Causal Inference Engine (Image reproduced from: [12])

Using this type of paradigm, can ultimately make our model much more flexible than contemporary Deep Learning models (our model is now much less dependent on data and more focused about intrinsic relationships and connections).

[ii] Although, Do-Calculus notation might look similar to the one of conditional probability, the two have two different meanings.

2.2 Linear and Non-Linear Causality

Causality is divided into two main types: linear and non-linear (Figure 2.4) [13]:

- In linear causality, connections between the variables can be in a single direction and every effect can be originated by a limited number of causes. Causes always linearly precede effects (time precedence).
- In non linear causality, connections between variables can be bi-directional and effects can possibly be originated by an unlimited number of causes.

Linear causation systems are characterised by proportional relationships between cause and effects variables (e.g. Deterministic Systems). Instead, in non-linear causation systems, disproportionate effects can take place (e.g. Non-deterministic Systems). For example, small changes in input conditions would then result in different consequences (e.g. "Butterfly Effect").

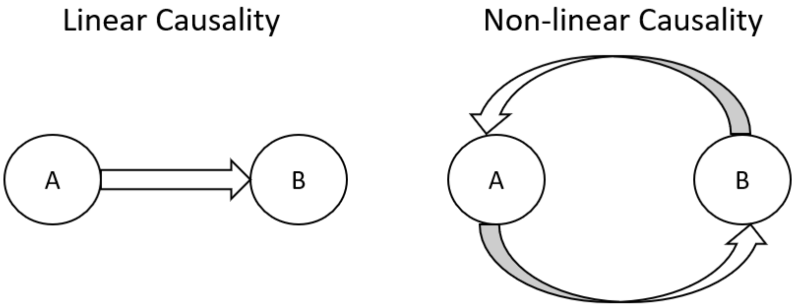


Figure 2.4: Linear vs Non-Linear Causality

From an external point of view, each causal systems can then be characterised as a composition of events, which might be regulated by a series of hidden trends and rules. Being able to correctly identify how these different constituent forces are interconnected to each other (grasping any reciprocal causal mechanism), would then allow us to make any system much more predictable.

The causal analysis of any dynamical system can then be summarised by the following workflow (Figure 2.5).

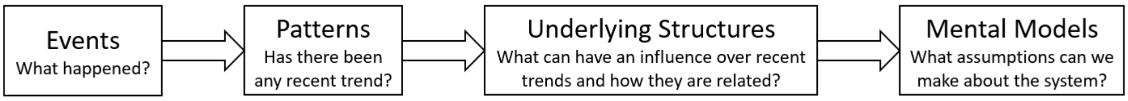
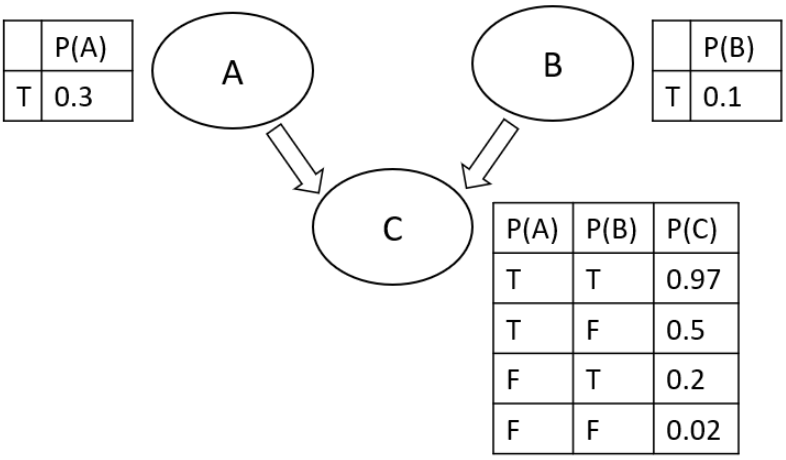

Events
What happened?
Patterns
Has there been any recent trend?
Underlying Structures
What can have an influence over recent trends and how they are related?
Mental Models
What assumptions can we make about the system?

Figure 2.5: Dynamical Systems Analysis (Adapted from [13])

2.3 Bayesian Belief Networks

Bayesian Belief Networks (BBN) are a type of probabilistic model which makes use of simplifying assumptions to reliably define connections between different elements and calculate their probabilities relationships efficiently. By analysing interactions between the different elements, we can finally make use of these type of models in order to discover causal relationships.

In a Bayesian Network, nodes represent variables while edges report the probabilistic connections between the different elements. A simple example of a three variables Bayesian Belief Network is available in Figure 2.6.



	P(A)
T	0.3

	P(B)
T	0.1

P(A)	P(B)	P(C)
T	T	0.97
T	F	0.5
F	T	0.2
F	F	0.02

Figure 2.6: Bayesian Belief Network

Bayesian Belief Networks led later to the development of Causal Networks. In fact, they can also be considered as a Causal Network in some specific cases. For this reason, Bayesian Belief Networks are considered to be one of the main techniques which can be used in Machine Learning in order to move from the Association to the Intervention level in the Causality Hierarchy (Figure 2.1).

Bayesian Belief Networks are able to express both conditional dependent and independent variables connections. These type of networks follow additionally the Markov condition [14] (provided the parents of every node in a network, each node is conditionally independent of their nondescendent nodes). Finally, using Bayes probabilistic approach (Equation 2.1), we can update the connection probabilities iteratively based on new gathered evidence.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{2.1}$$

where:

A, B	Events
$P(B\|A)$	Likelihood
$P(A)$	Prior
$P(B)$	Normalizing Constant
$P(A\|B)$	Posterior

Complex BBNs can be constructed by starting from three basic types of junctions: Chains (e.g. $A \Rightarrow B \Rightarrow C$), Forks (e.g. $A \Leftarrow B \Rightarrow C$) and Colliders (e.g. $A \Rightarrow B \Leftarrow C$). Making use of the three types of junctions and of a technique called d-separation, it can then be possible to reach the Counterfactuals level in the Causality Hierarchy [12]. D-separation, allow us in fact to understand, in Causal Diagrams, if a set of variables is independent of another set when given a third one.

What distinguishes Bayesian statistic from the classical frequentistic approach is that we allow to incorporate some level of subjectivity in our model[iii] (by combining prior knowledge with evidences). Additionally, in Bayesian statistics, the weight of our prior belief gradually vanishes as more data is provided (therefore converging to the frequentistic approach if given an unlimited amount of data). This case doesn't instead hold true when talking about causality analysis.

Great research focus by companies such as DeepMind is currently put into using Bayesian Belief Networks as a starting point in order to create Causal Bayesian Networks (CBN) [15]. Causal Bayesian Networks, are used in order to visually identify and quantitatively measure unfairness patterns in datasets (elements in the data which can lead to Machine Learning models biased towards specific subcategories). Additionally, researches also demonstrated the possibility to use Causal Bayesian Networks in order to identify if not just the recorded data but also the Machine Learning models itself are biased or not towards specific classes [16].

[iii] Frequentistic probability aims instead for complete objectiveness.

2.4 Intervention

What allows us to talk about cause and effect are experiments. Experiments are a set of procedures carried out under controlled conditions so that to test an hypothesis and try to undercover causes. Controlling the conditions of an experiment can allow us in fact to eliminate any alternative explanations which we might have about how a phenomena works. When creating an experiment, we need to make sure that different treatments are applied and that they are randomly assigned [17]. In an experiment, a treatment is defined to be as a change imposed from us to the environment of the experiment. Treatments should additionally be randomly assigned in order to make up for any variability in the environment space (e.g. different individuals/objects might have different characteristics). Finally, if working with just a sample from a population, it is necessary to make sure that the available sample size is large enough to be representative of the whole population. If all these characteristics are provided, then we can be able to accurately discover causal relationships even under uncertainty. Therefore, experiments play a key role in order for us to move between the different levels in the Causality Hierarchy (Figure 2.1). If any of the conditions doesn't instead hold true, we might end up accidentally adding some form of bias in our experiment. Without experiments, data driven decisions can just be backed by correlations and domain knowledge, but no pure evidences.

One of the main differences between Causal Diagrams and Bayesian Belief Networks is that the former are able to deal with interventions, while the latter works only with observations. Interventions allows us to use Causal Inference, while observations allows us to make predictions. What makes possible for Causal Diagrams to work with interventions is the Do Operator. From the Do Operator, it is then possible to create the Do Calculus which has been demonstrated to be complete as a technique (if Do Calculus is not able to identify an effect, then this cannot be identified anywhere else). On the other hand, what makes it difficult to use Causal Diagrams is making sure they are designed correctly. In fact, different Causal Diagrams can potentially be proposed in order to describe a process and depending from different point of view, it can be difficult to understand which one is able to better capture the underlying dynamics of a process.

For example, let's imagine we have carried out a study in order to find out if a diet can bring a positive effect (above the average) on the overall well-being of a sample in a population. Therefore, we divide the participants of this experiment into two groups: one which will strictly follow the diet and the other one which will instead follow common used diet-lifestyles. From this study, it might result that following the diet causes a positive effect to the individuals' well-being (Figure 2.7).

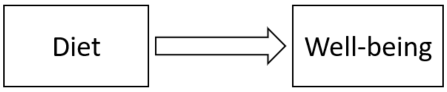


Figure 2.7: Simple Causal Diagram

Although, at this point we might start having some doubts about if there could be any potential pitfall in our analysis. One possible approach in order to test for the presence of confounding variables, is to take measurements about the suspected hidden variable (in this way, we can be able to fill any missing piece in our diagram). Carrying out a controlled experiment, we will be able to deconfound any true and "spurious" effect this hidden variable might cause. In the case of our example, we could for instance notice that the group which followed the diet had an overall younger age than the other group. Therefore, age could be considered as a possible factor which has skewed our early results and that we should therefore take into account (Figure 2.8).

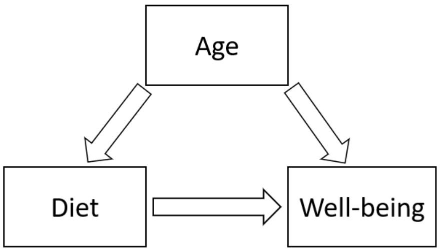


Figure 2.8: Improved Causal Diagram

Our confounding variable (Age) could then be deconfounded by comparing the two different groups of the experiment for different age groups, averaging the results and weight the different age groups sections by their percentage presence in the population composing the experiment. This procedure is commonly referred as **adjusting/controlling for a variable** [12]. Examining the results of our analysis, we could then think there might be some other variables missing and repeat this process again to improve our representation of the system. A confounding can therefore be defined to be as anything making $P(A|B)$ different from $P(A|do(B))$.

In a causal diagram information can flow through links from one vertex to another in two different directions (causal and non-causal). In this setting, information flowing in the non-causal direction can then lead to the creation of confoundings. In order to avoid this problem, information flowing in the non-causal direction can be blocked by using either of the following three approaches:

- Accurate controlled experiment randomization.
- Statistical variable adjustments.
- Applying the Do Operator on a variable, can stop the flow of information in the non-causal direction (this procedure can also be applied when working with observational data).

Finally, information flow in causal diagrams is regulated by the same type of junctions introduced in Section 2.3. Controlling for different variables in these types of junctions (to prevent presence of confoundings), would then lead to the following results:

- $A \Rightarrow B \Rightarrow C$: controlling B would stop information to flow between A and C.
- $A \Leftarrow B \Rightarrow C$: like in the previous case, controlling B would stop information to flow between A and C.
- $A \Rightarrow B \Leftarrow C$: controlling B would in this case enable information to flow between A and C (without intervention, no information would have been able to flow).

Making use of this set of information, it can then be possible to create and work with a multitude of different causal diagrams.

2.4.1 A/B Testing

A/B Testing is one of the most common forms of experimentation in Computer Science. Companies make common use of A/B Testing, for instance when shipping a new feature in production. For example, a company might come up with a new design for a section on an App. To a randomised half of the users it is going to be proposed the new version (Treatment Group), while to the other half it is going to be proposed the original version (Control Group). After collecting users behaviours and feedback, it could then be possible to observe the causal effects of our interventions. This type of approach is analogous to how vaccines/drugs are evaluated in clinical trials.

2.5 Counterfactuals

Carrying out experiments can be difficult and expensive in different real-life situations, in which case, we can perform just observational studies (we don't exercise any control on our independent variable). When working with observational data, changes between the control and treatment groups become **counterfactual**, therefore making difficult to uncover causal effects. Due to these limiting circumstances, it is then necessary to make some assumptions in order to create an approximation model able to make predictions. Counterfactual analysis focuses then on how different types of interventions could have retrospectively led to different outcomes (imagining alternative ”worlds”).

Some examples of Counterfactuals types of questions are: Would the patient have survived without taking any medication? How likely it is that a political party would have won the elections if it had proposed a more liberal policy than the advertised one?

Because of the nature of these types of questions, it can be difficult to provide answers with full certainty. Therefore, some form of probabilistic mechanism needs to be incorporated (e.g. We are 90% confident that choosing a more liberal policy would have increased the chances of winning the elections by 7%).

One of the main applications of Counterfactuals is **mediation analysis**. Mediation analysis is based on the concept of mediating variables (variables used to influence an outcome based on the effect of an applied treatment). An example of a mediating variable can be $VitaminA$ in $EatingCarrots \Rightarrow VitaminA \Rightarrow ImprovedEyeSight$. The aim of mediation analysis in this setting would then be to understand if the mediating variable is able to capture all the effects caused by the treatment variable ($EatingCarrots$) or not. Effects can then either reach our outcome variable directly or indirectly (through the mediating variable). Direct effects could then be represented in our example as moving through the following diagram without the need of mediating variables: $Eating\ Carrots \Rightarrow Improved\ Eye\ Sight$. Using mediation analysis we could then be able for example to find out if eating carrots is possible to improve our eye sight just by the increase of Vitamin A resources or through any other effect.

What makes mediation analysis an interesting field of research is the fact that the total effect exercised by a variable is not simply equal to the sum of its exercised direct and indirect effects in the case of third party variables interactions.

Examples of other techniques ideated during the course of the last decade in order to try to solve Counterfactual problems are [18]:

- Ordinary Least Squares with Confounding Variables
- Propensity Score Matching
- Instrumental Variables
- Difference in Differences

2.6 Causality in Machine Learning

The core principles which characterise Machine Learning are closely related to Statistics, in which we aim to undercover properties from an unknown distribution by sampling independent and identically distributed (IID) random variables from it. What makes Causal Inference more difficult to apply is the assumption that direct dependencies in the data exists (inferring a causal structure from just data can therefore be a challenging task). On the other hand, causal tools can allow us to go beyond mere statistical associations, providing us with more insightful and powerful models. Some example applications areas of Causality in Machine Learning are [19]:

- Semi-supervised learning.
- Half-Sibling Regression.
- Time Series Analysis (e.g. Granger causality).
- Reinforcement Learning.

2.6.1 Case Study: Recommendation Systems

One of the main weakness of most Machine Learning models is the assumption that the data fed in is independent and identically distributed. When this assumption holds, convergence to the lowest possible loss is achievable but when this constrain is violated, the model might perform poorly even when attempting simple tasks (e.g. poisoning attacks) [20]. As an example, let us consider an e-commerce recommendation system. Nowadays, systems are able to offer recommendations mainly based on products correlated to the ones we are planning to buy, although this cannot always lead to accurate estimates. For instance, we might have recently bought a new phone and we are now looking for a phone case. While browsing for phone cases, although our recommendation system might try to suggest us other items such as phones (just because they are correlated) instead of more cause-effect related items like screen protectors.

Chapter 3

Epidemic Modelling

3.1 Motivations

Modelling is a technique commonly used in order to approximate an environment/system to gain insights and better understand possible outcome scenarios especially in situations when we don't have data available about the topic. Some examples of applications in which modelling is commonly applied are: climate change, military defense, designing cities, infecting diseases development and testing financial policies.

Using modelling simulations can be of great help when trying to answer different types of causal questions about our research topic (e.g. varying the different simulation parameters, it can be possible to see how these are related each other and how they effect the overall outcome).

As part of this study, an interactive online web application [i] has been developed in order to quickly analyse in real time COVID-19 developments and simulate different scenarios and approaches which can be taken in order to mitigate the consequences of the outbreak. Most of the provided models have been designed so to be flexible enough to model any other type of possible future infectious disease.

Additionally, a secondary website has been created using GitHub Pages in order to share additional notebooks and animations in Python and Julia [ii] (Appendix D).

3.2 Introduction to Epidemiology

3.2.1 Different Classes of Diseases

Infectious diseases can mainly be classified into three different categories depending on their characteristics [21]:

1. **Endemic**: an endemic is a health concern which is constantly present at a low

[i] Web Application Link: http://3.22.240.181:8501/

[ii] GitHub Pages Website Link: https://pierpaolo28.github.io/Epidemics-Modelling/

rate within a population (its presence doesn't either substantially increase or decline). Some examples of endemics are Malaria and Chicken Pox.

2. **Epidemic**: an epidemic is a health issue which can cause a fast and unforeseen increase in cases within a population. An example of epidemic, can be considered to be the seasonal flu, which can lead to a sharp increase in the number of infected at specific times of the year.

3. **Pandemic**: epidemics can finally later become pandemics if they manage to spread around the world and affect a great number of people. Some examples of pandemics are the Spanish Flu and COVID-19.

3.2.2 Exponential vs Logistic Growth

Pandemics usually develop due to a disease's ability to spread at an exponential rate. In the case of COVID-19, the number of cases from one day to the next was in fact equal to the number of current cases multiplied by some constant between 1.25-1.5 (depending on factors such as population density and restrictions in place). The change in the number of cases from a day to another, can then be defined by the following equation [22]:

$$\Delta N_d = E \times p \times N_d \tag{3.1}$$

Where E represents the average number of people we are exposed to every day, p represents the probability that an exposure might lead to an infection and N_d is the number of cases as of today. Therefore, in this type of situation, the only possible way to try to slow down our exponential trend is by decreasing E and p. In order to make this possible, different techniques such as track and trace, social distancing and travel restrictions can be applied. Although, even if no intervention at all is done, an exponential trend is destined to convert to a logistic curve once a large number of the population gets infected by the disease (in fact, the probability that an exposure can lead to an infect automatically decreases if the majority of the population and the people we meet are already infected). Applying any type of restriction would then help us in making it feasible to reach our inflection point between these two trends as soon as possible.

Exponential growths can be easier to inspect when plotting on a logarithmic scale. Using this type of graph, an exponential curve would then look like approximately a straight line. As we can see from the graph on the left of Figure 3.1, all the different considered countries follow at first the same exponential pattern which then seems to start converting into a logistic curve. The graph on the right of Figure 3.1, was designed in order to try to amplify this change [23]. While going through an exponential growth, it can be difficult to understand how long it will last (if the growth is going to still keep being exponential or is going to start decaying). One possible way to approach this problem is to focus our attention on the rate of change in new cases from one week to another. Plotting this with both axis on a logarithmic scale, we would then clearly see that all the different countries have the same linear growth in cases. Although, using some form of containment, some of these countries are successfully able to escape from this linear growth in cases. Using this type of approach, we can successfully emphasize the deviation in the growth of an exponential curve.

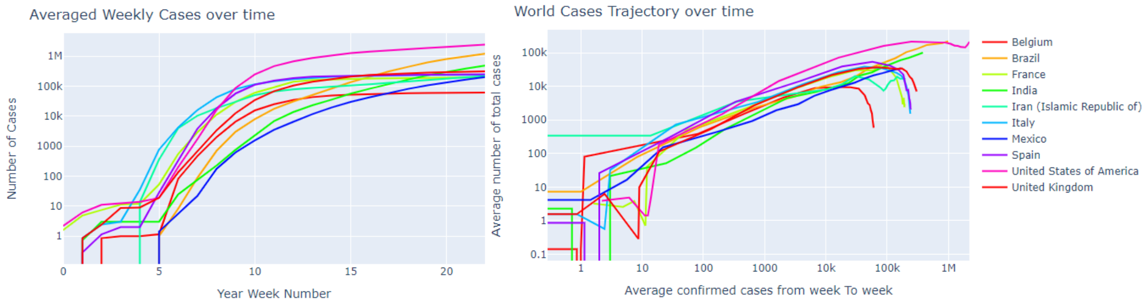


Figure 3.1: Exponential Growth Evolution (COVID-19, 28th June 2020)

Using the logarithmic linear graphs, we could then perform a linear regression to find the line of best fit and find out how many days it takes for cases to increase by a fixed constant. Finally, using metrics such as the R^2 score, we could then quantitatively measure how far are our curves from an exponential curve.

Another way to understand if we are reaching the end of an exponential curve is by examining the slope (Growth Factor, Equation 3.2).

$$Growth\ Factor = \frac{\Delta N_d}{\Delta N_{d-1}} \tag{3.2}$$

A growth factor of more than one will show us that we are still going through an exponential growth, while a growth factor equal to one can tell us we might now be approaching our inflection point.

A worked out example of logistic/exponential curve fitting on real world Coronavirus data, is Available in Appendix C.

3.2.3 Quantifying the spread of a disease

One of the main units used to measure how easily a disease is able to diffuse in a community is the "Effective Reproductive Number" (R), which is measured as the average number of people infected by each individual carrying the disease. In a fully susceptible population, R is also referred as R_0 ("Basic reproductive Number"). The Basic reproductive Number for COVID-19 is currently estimated to be around 2.5.

As shown in Equation 3.3, R_0 can be calculated as the number of people someone positive to the disease can infect each day (β) multiplied by the number of days each person remains positive to the disease (D). Equivalently, R_0 can also be estimated as the number of people someone positive to the disease can infect each day multiplied by the proportion of individuals infected recovering each day ($\gamma = 1/D$). Furthermore, as shown in Equation 3.4, β can be also calculated to be equal to the probability that an exposure might lead to an infection (p) multiplied by the average number of people we are exposed to every day (E) [24].

$$R_0 = \beta \times D = \frac{\beta}{\gamma} \tag{3.3}$$

$$\beta = E \times p \tag{3.4}$$

If R is greater than one, then the disease is still in the exponentially growing phase (we have an epidemic), if R is instead equal to one we are then in an endemic (therefore the number of cases stays approximately constant) and if R is less than one, we are finally in the eradication phase and the disease could disappear entirely soon from our population.

One of the most common approaches to eradicate a disease is through Herd Immunity. The percentage proportion of the individuals in a population necessary to reach Herd Immunity can be estimated using Equation 3.5. In the case of COVID-19, this is estimated to be around 60%. Herd Immunity can potentially be achieved either through vaccination or natural selection [25].

$$Herd\,Immunity\;(H.I.) = 1 - \frac{1}{R_0} \tag{3.5}$$

Different typology's of Epidemics Modelling have been developed in the past few years, such as:

- Compartmental Models
- Agent Based Models
- Network Models
- Meta-populations Models

In this chapter, we will explore the first two approaches.

3.3 Compartmental Models

In Compartmental Models it is assumed that each individual in a population is assigned to a compartment. During the course of the simulations, individuals can then be free to move from one compartment to another depending on the dynamics of the model. Some examples of common departments in Epidemic Modelling are: Susceptible, Exposed, Infectious, Recovered, Dead, Vaccinated, etc...

These models can be designed using either ordinary differential equations or stochastic elements as well. Diagrams representations of this type of models can be of great help in order to understand how the model equations works and what are the possible movements between different states. In Appendix E, are additionally available the Causal Diagrams equivalent representation of the implemented Compartmental Models.

3.3.1 SIR (Susceptible-Infected-Recovered)

Causal Question: How does the number of people I am in contact with on average in a day affect the evolution of the pandemic?

Experimental Results: The SIR model is one of the most widely used epidemiology model and is composed by just three different compartments: Susceptible (S), Infected (I) and Recovered (R). This model can be described by the following three formulas, where N is the total number of elements in the population, β represents the average amount of people an infected element can be able to infect in a day and γ the percentage of how many individuals recover from the disease each day.

$$\frac{\partial S}{\partial t} = -\beta \times I \times \frac{S}{N} \tag{3.6}$$

$$\frac{\partial I}{\partial t} = \beta \times I \times \frac{S}{N} - \gamma \times I \tag{3.7}$$

$$\frac{\partial R}{\partial t} = \gamma \times I \tag{3.8}$$

In order to better visualise the situation, this set of equations can then be converted into a block diagram representation using the following structure (Figure 3.2).

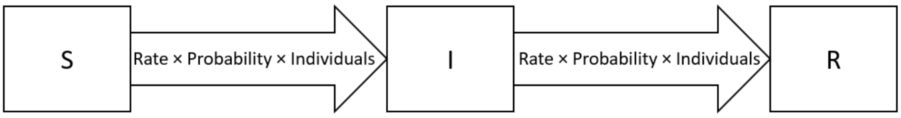

S
Rate × Probability × Individuals
I
Rate × Probability × Individuals
R

Figure 3.2: Compartmental Models Diagram Representation

Therefore, in order to move from one compartment to another, we need to take into consideration how long would this transaction take (Rate), the probability that it will actually happen for each individual (Probability) and the portion of individuals for which this transition takes place (Individuals). Converting our SIR set of equations into this representation, we can then obtain the diagram in Figure 3.3. When converting between these two representations, we can then see how a minus sign in the Ordinary Differential Equation (ODE) corresponds to an arrow leaving that compartment, while a plus sign corresponds to an arrow pointing towards that compartment. This same procedure can then be used in order to design any other type of compartmental model.

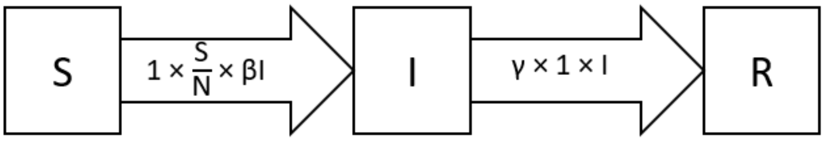

S
$1 \times \frac{S}{N} \times \beta I$
I
$\gamma \times 1 \times I$
R

Figure 3.3: SIR Diagram Representation

Using the parameters in Table 3.1, it can then be possible to obtain the results in Figure 3.4.

Parameter Type	Value
Population Size	100
Number of Days	100
Number of individuals originally infected	3
Number of individuals at close contact in a day	5
Probability of infection if in contact with an infected	0.1%
Number of days a the disease can last	7

Table 3.1: SIR Model Parameters

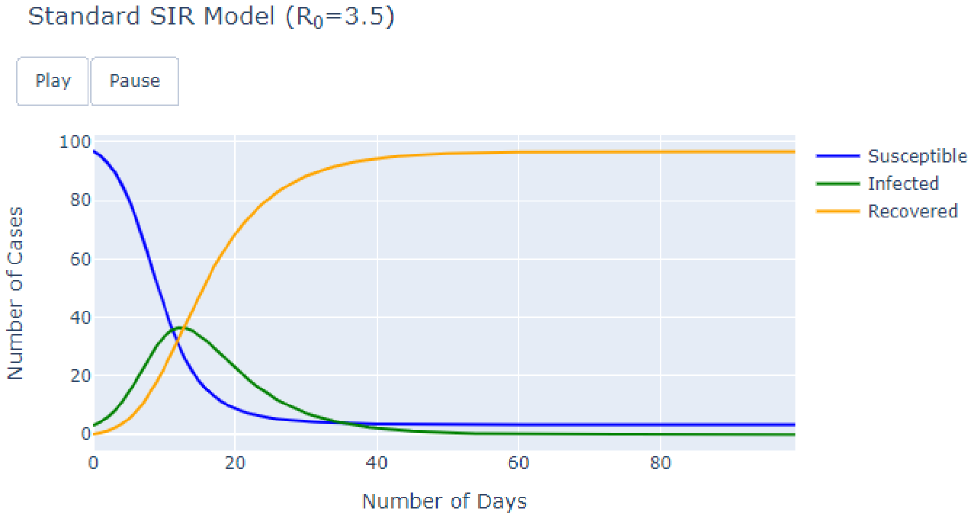


Figure 3.4: SIR Model

Experimental results demonstrated that even slight increases in the number of people individuals are in contact with on average in a day can cause major changes in the possible evolution of the pandemic, leading to higher peaks in the curve of infected and possibly overwhelming hospitalization systems.

3.3.2 SEIR (Susceptible-Exposed-Infected-Recovered)

In order to make our model more realistic, we can then add an additional state E representing all the population elements which are still in the incubation stage before becoming infected. To apply these modifications, we just need to update $\frac{\partial I}{\partial t}$ and add this extra stage just before it. The only variable which needs to be added compared to the SIR model is the proportion of how many individuals move from the incubation period to being infected ($\delta = \frac{1}{Days\ of\ incubation}$).

$$\frac{\partial E}{\partial t} = \beta \times I \times \frac{S}{N} - \delta \times E \tag{3.9}$$

$$\frac{\partial I}{\partial t} = \delta \times E \times -\gamma \times I \tag{3.10}$$

Following the same procedure as for the SIR model, we can then convert our SEIR model in block diagram form (Figure 3.5).

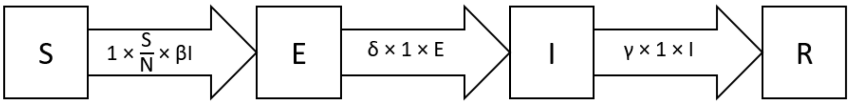

S
$1 \times \frac{S}{N} \times \beta I$
E
$\delta \times 1 \times E$
I
$\gamma \times 1 \times I$
R

Figure 3.5: SEIR Diagram Representation

Using the same parameters as in Table 3.1, and choosing one day as the number of incubation days for the disease, can then be produced the results in Figure 3.6.

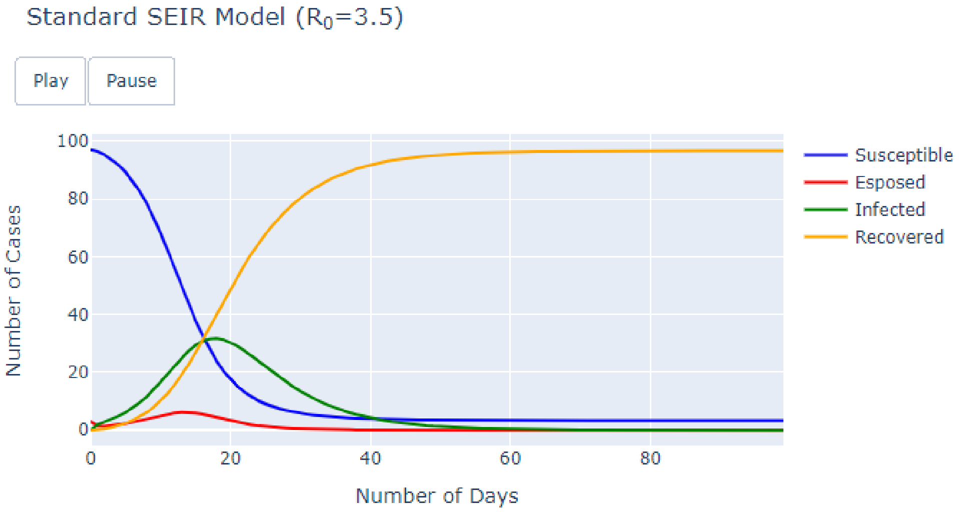


Figure 3.6: SEIR Modelling

3.3.3 Advanced SEIR Modelling

Causal Question: What are the effects of varying social distancing measures over time using different approaches (landscapes)? Given some user defined population age distribution, how would that effect the overall number of deaths?

Experimental Results: Starting from the developed SEIR Model, it can now be possible to add more compartments and make different elements time-dependent so to better capture real-world dynamics. The main additions engineered in this model are:

- Take into account the portion of infected individuals whom die instead of recover. This can be done by adding a new compartment after the Infected Stage and introducing two new variables ρ (the rate, in terms of days, which takes on average to die) and α (the disease death rate) which determines the probability an individual infected might either die or recover.

- R_0 is not anymore static but dynamically changes over time. In this example, two functions have been used in order to simulate R_0 behaviour over time: a Sigmoid or Sinusoidal. In this way, we are now able to model how a government might react in order to control the spread of the disease by exercising social distancing measures. A Sigmoid in it's minimum point can in fact represent a lock-down and the smoothness by which it reaches its minimum can easily represent how gradually the restrictions have been applied. An additional parameter is provided in order to decide from what day onward to start applying the restrictions (so that to observe what could be the consequences of a late or early intervention). The Sinusoidal landscape, can instead be used in order to model possible sequential waves a disease can lead to. Because, β is dependent on R_0, this will be indicated to be our time dependent variable in our set of ODEs.

- Also the death rate has been designed to be time and age dependent. To each different age group is assigned a different base death rate (the older, the greater), which increases linearly with the increase in the number of infected at each time-step. Therefore, higher peaks of individuals infected all at the same time increases the likelihood of each individual to die (mimicking strained healthcare system which doesn't have enough resources to cure everyone at the same time).

The main equations summarising the flow between the different compartments can be summarised as:

$$\frac{\partial S}{\partial t} = -\beta(t) \times I \times \frac{S}{N} \tag{3.11}$$

$$\frac{\partial E}{\partial t} = \beta(t) \times I \times \frac{S}{N} - \delta \times E \tag{3.12}$$

$$\frac{\partial I}{\partial t} = \delta \times E - (1 - \alpha(t)) \times \gamma \times I - \alpha(t) \times \rho \times I \tag{3.13}$$

$$\frac{\partial R}{\partial t} = (1 - \alpha(t)) \times \gamma \times I \tag{3.14}$$

$$\frac{\partial D}{\partial t} = \alpha(t) \times \rho \times I \tag{3.15}$$

Which results in the following diagram flow architecture:

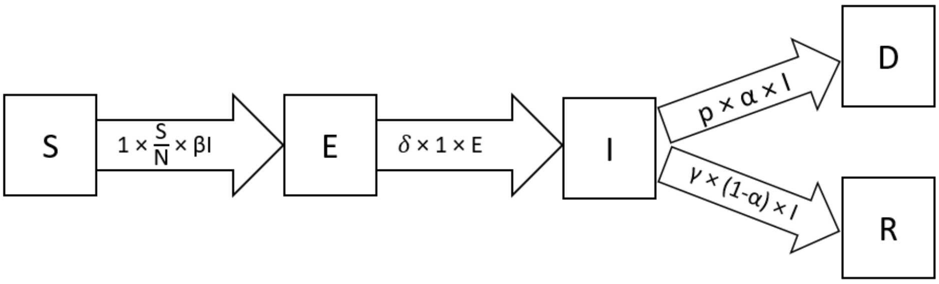


Figure 3.7: Advanced SEIR Diagram Representation

More specifically, the adaptive R_0 for the Sigmoid case has been designed to be equal to Equation 3.16.

$$R_0(t) = \frac{R_{start} - R_{end}}{1 + \exp^{-k(-x+x_0)}} - R_{end} \tag{3.16}$$

where: R_{start}, R_{end} Initial and desired final value for R_0 in the sigmoid

x_0 Sigmoid inflection point date

k Defines how drastically rapidly the restrictions have been applied

Following from our $R_0(t)$ calculation, $\beta(t)$ can then be estimated to be equal to:

$$\beta(t) = R_0(t) \times \frac{1}{\alpha \times \frac{1}{\rho} + (1-\alpha) \times \frac{1}{\gamma}} \tag{3.17}$$

In the case of a Sinusoidal landscape, R_0 has instead designed to be dependent on the following expression:

$$R_0(t) = Oscillatory\ Scale \times \sin\left(\frac{x}{\frac{N\ Simulation\ Days}{10}}\right) + Oscillatory\ Scale + 1 \tag{3.18}$$

In this way, R_0 is designed to always go through two peaks and one low for any type of possible simulation. Using an $Oscillatory\,Scale$ of 1, the peaks will be equal to an R value of 3, while the low would be equal to 1. Increasing the value of the $Oscillatory\ Scale$, will then lead to higher values for our landscape peaks.

Finally, in order to make our death rate time variant and dependent on the population age and number of infected cases at the same time, the following expression has been used:

$$\alpha(t) = s \times \frac{I(t)}{N} + \alpha_{opt} \tag{3.19}$$

where: s Regulates in what proportion the death rate rises when the number of infected increases.

α_{opt} The death rate calculated depending on the population age. Overall older populations will therefore have an higher base death rate than younger populations.

This advanced SEIR version has been inspired by Henri Froese's [24] work.

Starting with the same parameters already specified for the SIR and standard SEIR model we can then specify the number of days the disease can take to become lethal to be equal to 3 and the percentage weight of age on the death rate to 30%.

The proportions of individuals within different age ranges could then be specified as shown in 3.2.

Demographic	Age			
	0-20	20-50	50-70	70-110
Proportion Percentage (%)	0.15	0.25	0.4	0.2

Table 3.2: Population Demographics

In the case of the Sigmoid Landscape, using the parameters in Table 3.3, it has been possible to record the results in Figure 3.8.

Parameter Type	Value
Social Distancing start date	20
Percentage weight of how rapidly the restrictions are applied	30%
Maximum possible R value	5
Minimum possible R value	1

Table 3.3: Sigmoid Landscape Parameters

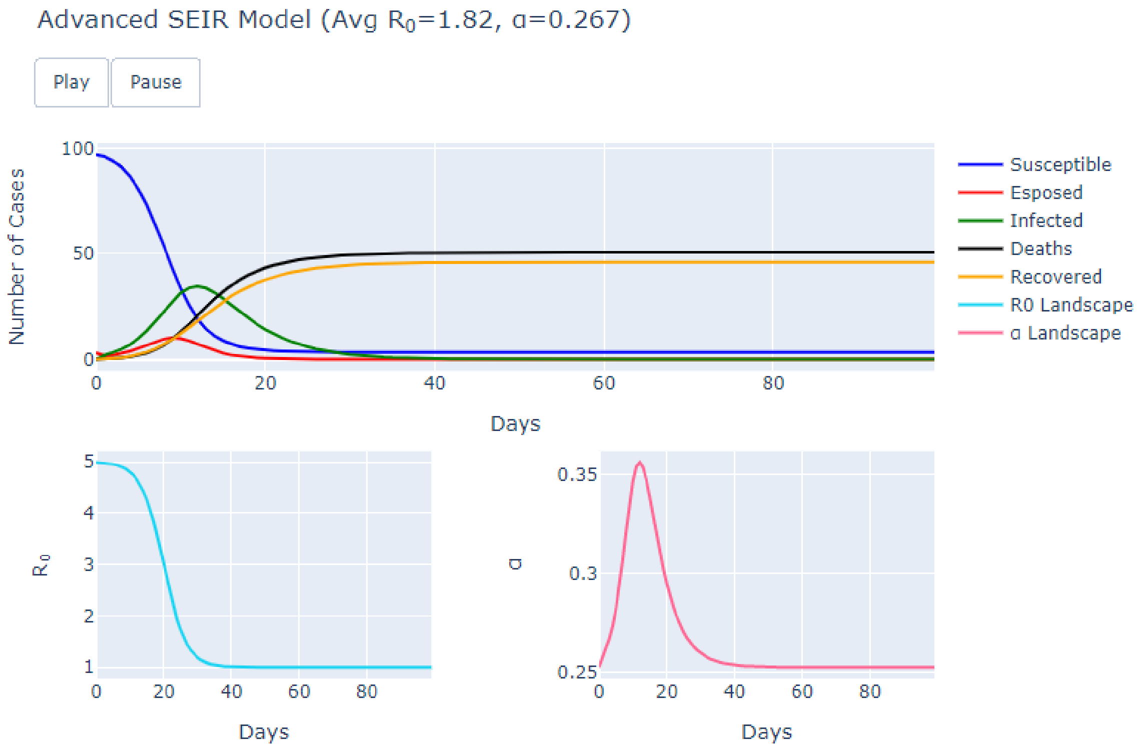


Figure 3.8: Advanced SEIR (Sigmoid) Modelling

Using instead the Sinusoidal landscape and a scaling factor of 2, have been registered the results in Figure 3.9.

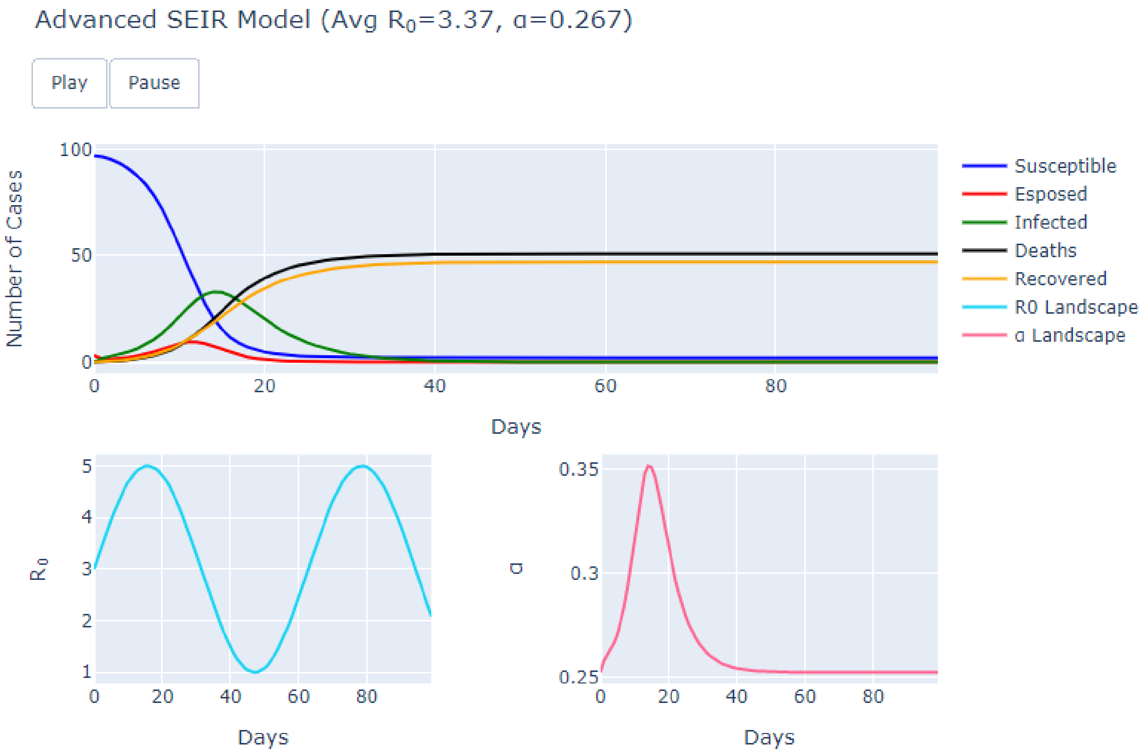


Figure 3.9: Advanced SEIR (Sinusoidal) Modelling

3.3.4 Time Limited Immunity and Vaccination Modelling

3.3.4.1 Time Limited Immunity

Causal Question: What if recovered patients could be infected again? Could it be possible to eradicate the disease?

Experimental Results: Updating our SIR model (Section 3.3.1), we can then be able to take into account the possibility that individuals might not gain lifetime immunity from a disease when recovering from it, but that they might instead be re-infected again in the future after some time. The amount of time an individual might be immune from a disease can be represented by just adding a new variable to our model (v) [26].

$$\frac{\partial S}{\partial t} = -\beta \times I \times \frac{S}{N} + v \times R \tag{3.20}$$

$$\frac{\partial I}{\partial t} = \beta \times I \times \frac{S}{N} - \gamma \times I \tag{3.21}$$

$$\frac{\partial R}{\partial t} = \gamma \times I - v \times R \tag{3.22}$$

Converting our set of equations into a Block Diagram representation, will then result in Figure 3.10.

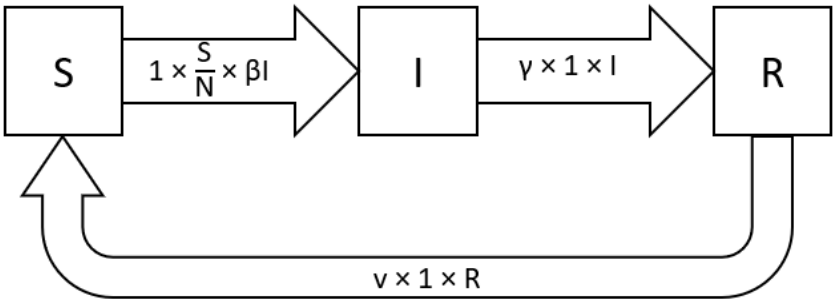

S
$1 \times \frac{S}{N} \times \beta I$
I
$\gamma \times 1 \times I$
R
$v \times 1 \times R$

Figure 3.10: SIR Time Limited Immunity Diagram Representation

Using the same SIR model parameters as in Table 3.1, setting to 25 the maximum number of days the immunity to the disease can last, we can then produce the following results in Figure 3.11.

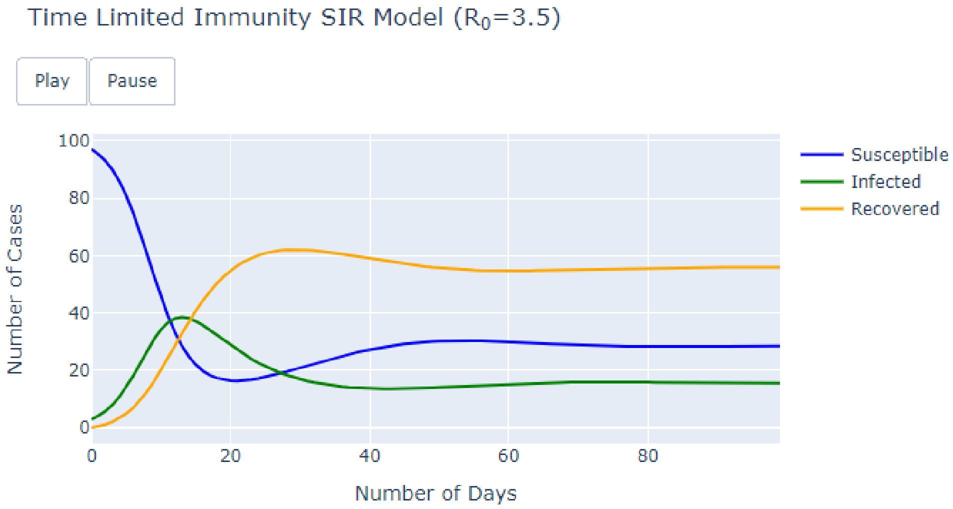


Figure 3.11: Time Limited Immunity Modelling

As can be seen from our experimental results, having a low time limited immunity would make it almost impossible to eradicate the disease from the population.

3.3.4.2 Vaccination

Causal Question: Adding vaccination to this model, would that make possible to eradicate the disease even with low time-limited immunity?

Experimental Results: Extending our set of equations (adding an extra stage $\frac{\partial V}{\partial t}$), we can be able to take into account how an epidemic will evolve once a vaccine is available. In order to apply these modifications, we just need to update $\frac{\partial S}{\partial t}$ and add the vaccination stage just after it. To make the simulation more realistic, we can then also specify from when in time a vaccine could start being distributed and how fast it can produced and shipped (p). Finally, a stage used to record the possible amount of deaths is included (using the same notation for the Advanced SEIR models).

$$\frac{\partial S}{\partial t} = -\beta \times I \times \frac{S}{N} + \upsilon \times R - p \times S \tag{3.23}$$

$$\frac{\partial V}{\partial t} = p \times S \tag{3.24}$$

$$\frac{\partial I}{\partial t} = \beta \times I \times \frac{S}{N} - (1 - \alpha) \times \gamma \times I - \alpha \times \rho \times I \tag{3.25}$$

$$\frac{\partial R}{\partial t} = (1 - \alpha) \times \gamma \times I - \upsilon \times R \tag{3.26}$$

$$\frac{\partial D}{\partial t} = \alpha \times \rho \times I \tag{3.27}$$

Converting our set of equations into a Block Diagram representation, we can then obtain Figure 3.12.

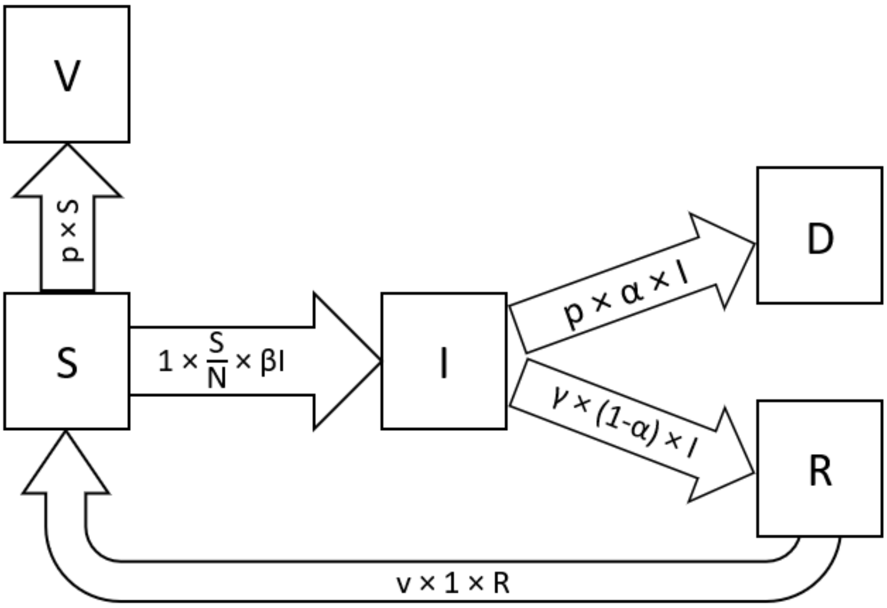


Figure 3.12: SIR Time Limited Immunity and Vaccination Diagram Representation

Extending our set of parameters used for the Time Limited Immunity model, using the values from Table 3.4, it has been possible to obtain the results in Figure 3.13.

Parameter Type	Value
Number of days the disease can take to become lethal	5
Death rate	0.2%
Number of Days to start Vaccine Distribution	30
Vaccine Distribution rate	0.1%

Table 3.4: SIR Vaccination Model Parameters

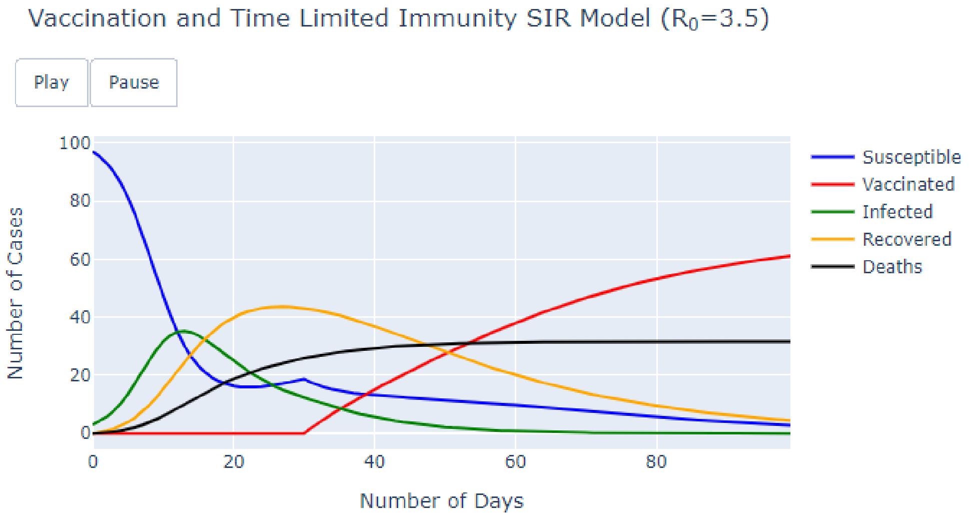


Figure 3.13: Vaccination Modelling

Making use of the provided Time Limited Immunity and Vaccination models, in conjunction with Equation 3.5, it could then be possible to get estimates of what proportion of the population would be necessary in order to gain Herd Immunity.

Adding vaccination in our model (especially in an early stage and distributing it at a high rate), would therefore make it possible to completely eradicate a disease even in the case of low time-limited immunity [27].

3.3.5 Coronavirus Modelling

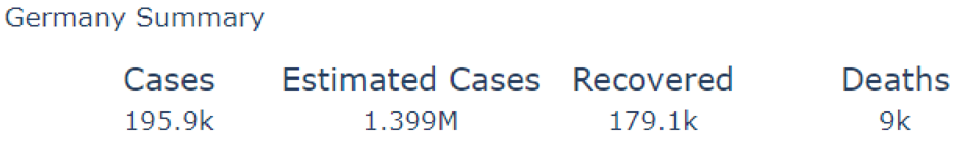
In order to test on real world data one of the created models, the architecture of the Advanced SEIR model outlined in Section 3.3.3 has been tested. The different parameters of the model, have then been personalised depending on the examined country specifics.

In order to do so, information about a large number of countries population size and demographics has been processed making use of the United Nations World Population Prospects 2019 dataset [28]. Additionally, information about the number of hospital beds available for 1000 people for each country has been used from the The World Bank Data Hospital Beds dataset [29], in order to calculate estimates of when hospitals in different countries might get overwhelmed and how many avoidable deaths would that cause. Finally, an estimation of the number of actual cases for each country has been calculated using the approach outlined in ”Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)” [30].

3.3.5.1 Germany Case Study

Causal Question: How can we prevent a healthcare system becoming overwhelmed? How many lives could be saved?

Experimental Results: As of the second of July 2020, Germany Coronavirus record can be summarised as in Figure 3.14. Germany has a total population of 84 millions and the total number of estimated cases has been calculated taking in consideration that about 86% of the total Coronavirus cases have been estimated to have been undocumented [30] in China.



Germany Summary

Cases	Estimated Cases	Recovered	Deaths
195.9k	1.399M	179.1k	9k

Figure 3.14: Germany Statistics

Another point, which could be necessary to take into account in order to understand the true number of cases in a country, is the possibility of false positives and false negatives during testing. More accurate estimates can be calculated using Bayes Rule and the Law of Total Probability (Equation 3.28). For instance, if a patient does a Coronavirus test and results positive, what are its chances that he/she actually has Coronavirus?

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)} = \frac{P(A|B) \times P(B)}{P(A|B) \times P(B) + P(A|B) \times P(B)} \tag{3.28}$$

In this example, we are going to take into consideration the COVID-19: Roche Antibody Test [31]. This test has an estimated sensitivity of 100% (ability to correctly identify patients with the disease) and specificity of 99.8% (ability to correctly identify patients without the disease). Assuming a prevalence of the disease in the population of about 5%, the following estimates can be calculated:

$$P(B|A) = \frac{1 \times 0.05}{1 \times 0.05 + 0.002 \times 0.95} = 96.33\% \tag{3.29}$$

If the test is positive, there would be a 96.33% probability the patient actually has Coronavirus. As can be seen from Equation 3.28, the rate of the Coronavirus cases in a population, plays an important factor.

An example of a simulation run using Germany as case study is available in Figure 3.15. As can be noticed, the key difference between 3.15 and 3.8 is the fact that we are taking into account this time also when the healthcare system of the selected country runs out of beds for critical patients. This can be of vital help in order to understand to what extent the healthcare system can be able to provide support the all the patients in need at any point in the pandemic and how many deaths could be avoided.

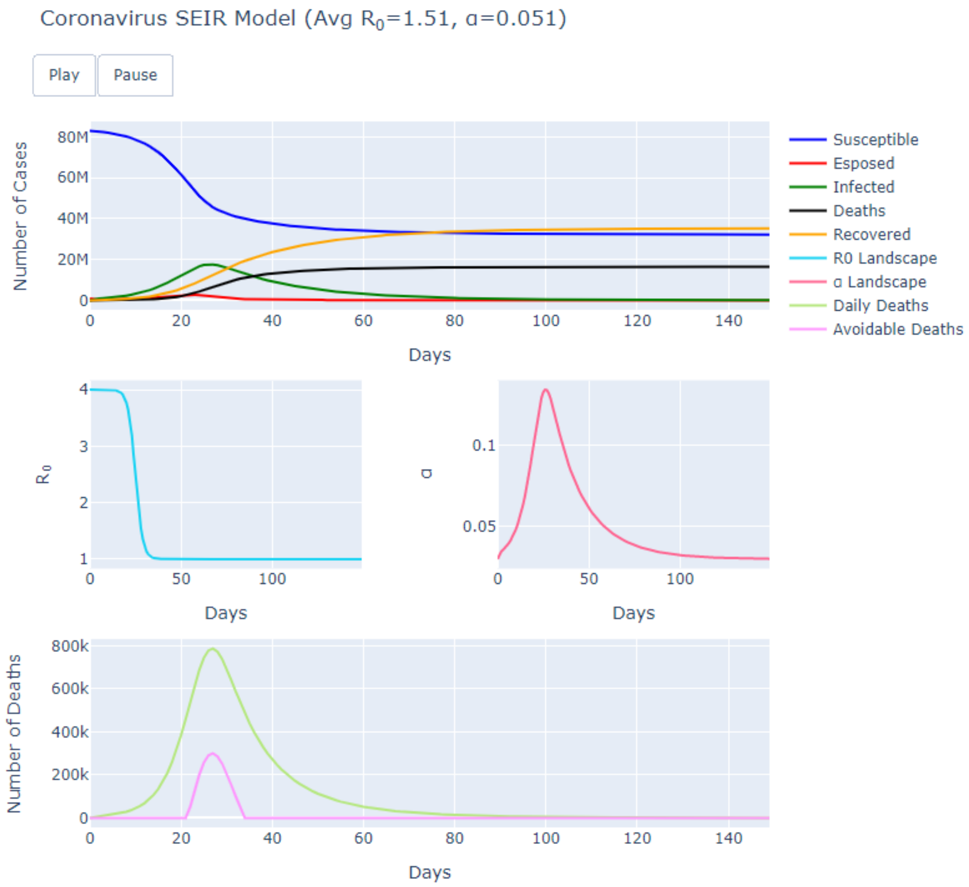


Figure 3.15: Germany Simulation

Additionally, an estimate of the number of beds needed to go through the peak of the simulated pandemic while providing support for all the critically ill patients is provided in Figure 3.16 alongside with the current number of beds available.

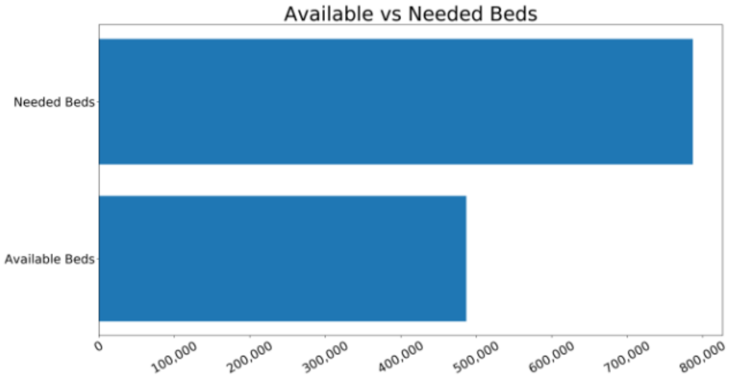
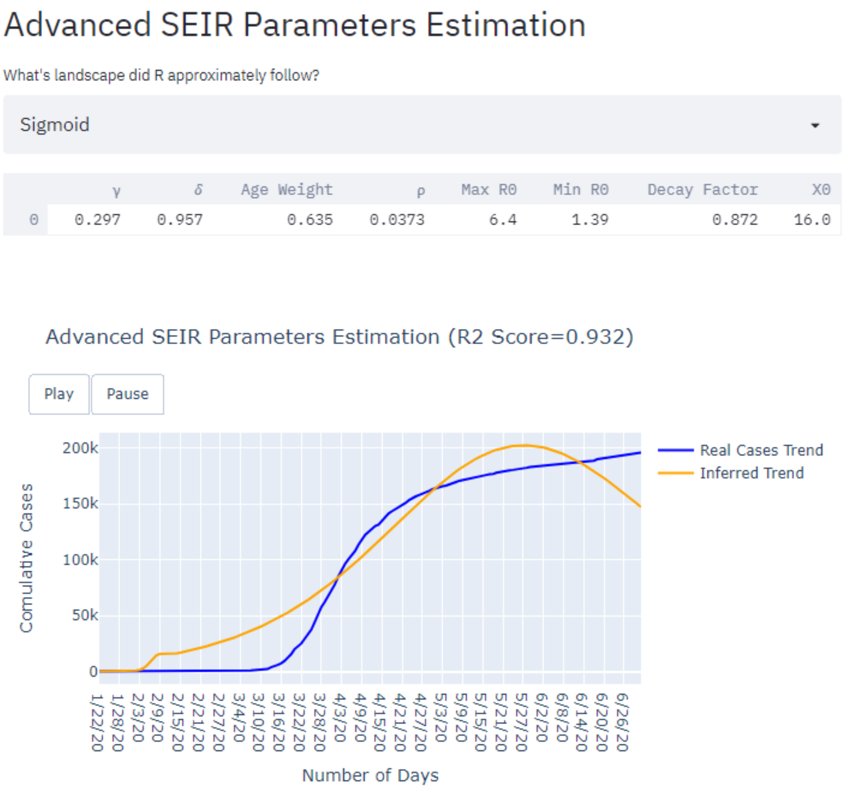


Figure 3.16: Germany Hospital Beds Analysis

Finally, it is calculated an automatic estimation of the parameters of the Advanced SEIR model, given the provided data (Figure 3.17). This estimation is computed using non-linear least squares and the final R^2 score is provided as metric for the convergence of the fitted curve compared to the original one. According to these estimates, thanks to public health restrictions, Germany's R value, varied from a maximum of 6.37 to a minimum of 1.45 during the last 5 months.



Advanced SEIR Parameters Estimation

What's landscape did R approximately follow?

Sigmoid

	γ	δ	Age Weight	ρ	Max R0	Min R0	Decay Factor	X0
0	0.297	0.957	0.635	0.0373	6.4	1.39	0.872	16.0

Figure 3.17: Advanced SEIR Parameters Estimation

3.4 Agent Based Models

An alternative approach which can be used in order to simulate compartmental-like models is by creating an Agent Based simulation. In this case, each single individual in the population is created following an Object Oriented Programming approach (e.g. using a programming class) and it's behaviour in an environment while in contact with the rest of the population is simulated. Using this type of approach can therefore enable us to keep track of the position of the different individuals in a population and attribute them different characteristics such as an age or daily income. Similar results could be obtained using standard compartmental models by converting our Ordinary Differential Equations into Partial Differential Equations and making them dependent on both time and space [32].

The proposed Agent Based Model, is directly inspired from The Epidemiological Triad paradigm (Figure 3.18).

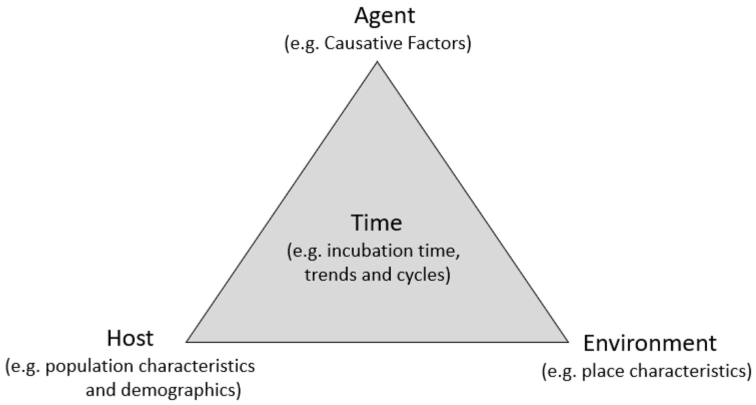

Agent
(e.g. Causative Factors)
Time
(e.g. incubation time, trends and cycles)
Host
(e.g. population characteristics and demographics)
Environment
(e.g. place characteristics)

Figure 3.18: The Epidemiological Triad

3.4.1 Population Modelling

Causal Question: How does population density and community distribution affect the spread of a disease?

Experimental Results: As described in Section 3.2.2, the number of new cases can vary according to Equation 3.1. Therefore, the only way we can be able to decrease the number of cases, is by decreasing the values of E and p.

- E can decrease if travelling and meetings of people are reduced as much as possible.
- p can be reduced instead for example by making it less likely to catch the disease by taking precautions such as washing hands, wearing masks, avoid touching our faces, etc...

This trend can be observed in the following proposed model (Appendix F) by the **Contact Radius** (E) and **Probability of how unlikely it is to spread the virus if within the contact radius** (complementary of p) variables. In this way, causal effects of social distancing and improved hygiene can be easily inspected. Furthermore, the role of dividing individuals in different communities is additionally studied. Having different communities with a central shared point and random infected initialization can in fact resemble how contagious disease hot-spots can be created.

An example output of this type of simulation, is available in Figure 3.19.

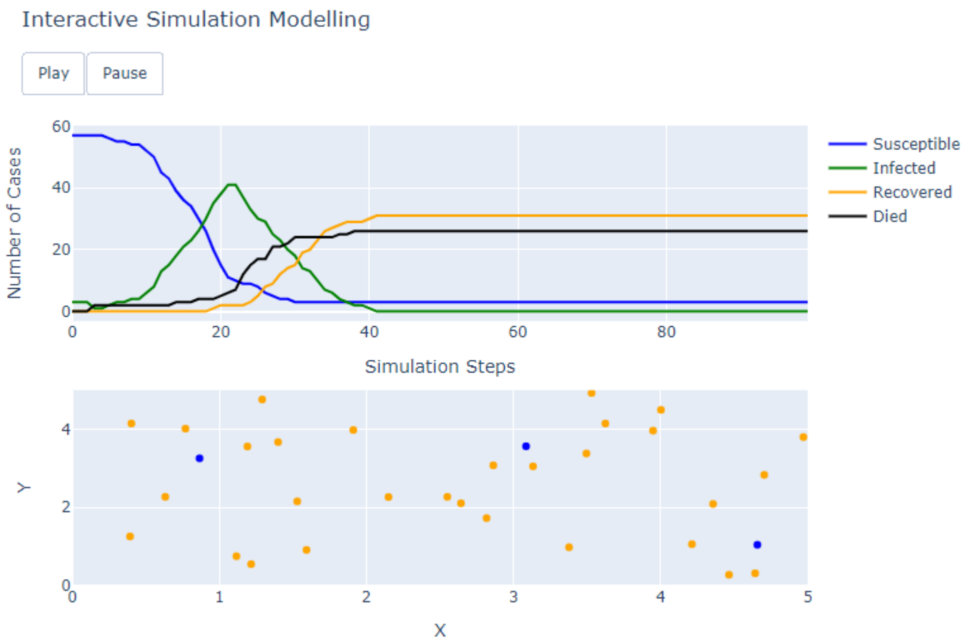


Figure 3.19: Population Modelling

In this type of model, each individual is born with unique combination of characteristics such as: their situation (susceptible to the virus, positive, recovered or dead), position in X and Y coordinates on a grid which represents their world, the speed with which they move and the X and Y directions at which they point to (e.g. mimicking individuals commuting from one place to another every time), a counter used to keep track of how many days of rehabilitation an individual went through with the virus and the individual age (the greater the age, the more likely to die).

An original death probability is assigned to each individual, indicating the probability someone would have to die for any general reason (this base probability is then increased depending on age if being affected by the virus). In this case, the death probability is defined as:

$$Updated\ Death\ Probability = Death\ Probability \times (\frac{Age}{100} + 1) \tag{3.30}$$

For example, if a 50 years old individual has a base probability to die of 1%, this will be increased because of the virus to 1.5%. Additionally, a Boolean value (**Static**) can be used in order to make the population static (therefore mimicking the effect of being permanently in a same location in order to avoid spreading the virus).

During each iteration, using the Euclidean Distance in Equation 3.31, it is measured how close each individual is with the others in the space and if it is close to some people affected by the virus (within the defined Contact Radius). The number of occurrences are then counted and the more they are and the more likely it will be that the individual will catch the virus (how likely it is that someone can catch the virus is additionally dependent on the specified probability of how unlikely it is to catch the virus - the lower it is and the more likely we will become infected).

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{3.31}$$

Additionally, with each iteration, it is registered how many individuals are affected by the virus. With each iteration, depending on the mortality probability, different individuals might die and if an individual survives for 14 iterations without dying, it is then considered as a survivor (recovered).

Successively, depending on the speed and the travelling direction, the position of all the different individuals is updated. Additional conditions are imposed if an individual reaches an edge of the grid in order to make it bounce back (therefore inverting the direction of movement). All the different metrics are then stored in order to produce the plots.

To simulate different separated communities, the population has been divided into groups which all have different grid limits within which they can move. Creating subpopulations with overlapping grid limits, individuals can then be allowed to move from one subset to another. The overall workflow is summarised in Appendix F.

3.4.2 Track and Tracing

Causal Question: What if instead of using social distancing we would try to increase our test capacity and successfully track and isolate all the infected people and the ones they have been in contact with? How would this system change if a portion of individuals would remain un-tracked?

Experimental Results: Track and Tracing can be considered to be the most effective approach in order to take under control a pandemic. Although, one of the main limitations of this approach, is that in less lethal diseases it might be difficult to correctly identify in time all the individuals infected (some might be asymptomatic). Developing contact tracing apps using cryptography could therefore enable us to keep our privacy intact whilst reducing the risk of spreading the disease.

In the following model, it is presented how an epidemic might evolve if all the infected individuals are successfully identified and then make their way to a quarantine location designed for all the individuals affected by the disease. Individuals are represented with different associated velocities in order to simulate the fact that some might be tracked before others and might interact with susceptible individuals along the way.

As we can see from Figure 3.20, using this type of technique, led to a sharp decrease in the overall number of deaths (17) and infected than our general model. In this case, has been created a scenario with four different communities and a fraction of individuals allowed to move between different hubs.

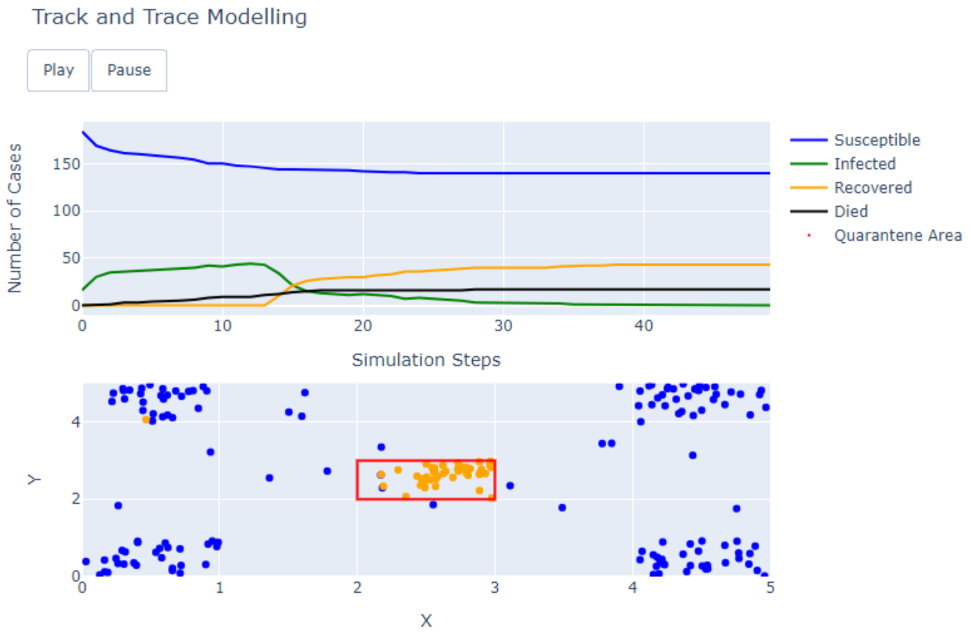


Figure 3.20: Perfect Track and Tracing

The following model extends instead the first model by adding a probability value that some individuals might not get traced at all and might therefore end up spreading the disease (e.g. Coronavirus asymptomatic or limited available testing capacity). An example simulation result using a 50% probability to not be traced is shown in Figure 3.21. As the results demonstrated, allowing even a small portion of un-tracked infected individuals, can potentially lead to far worse scenarios compared to Figure 3.20. Analogous results have been registered also in [33].

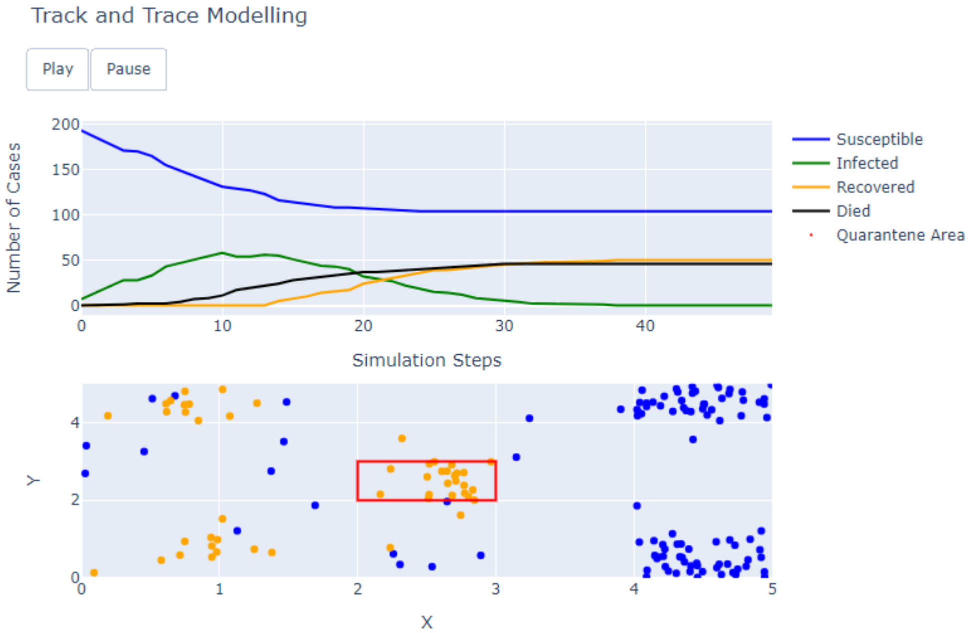


Figure 3.21: Imperfect Track and Tracing

3.4.3 Central Hubs

Causal Question: How does having a central hub visited by different communities members affect the spread of the disease?

Experimental Results: Imposing travel restrictions can greatly help in lowering the rate at which a disease can spread. Individuals although still have at times to visit centrals hubs such as supermarkets during lock-downs. In this simulation, we can easily observe how having even just a single central hub can lead to a fast spreading of the disease across different communities.

As the results show in Figure 3.22, allowing about 30% of the population to visit a central hub can potentially be quite dangerous because if anyone in the hub is effected, it can then easily allow to spread the disease to individuals belonging to different communities. Possible outcomes in this type of situation, can greatly vary from different runs, depending on causalities and random initialization.

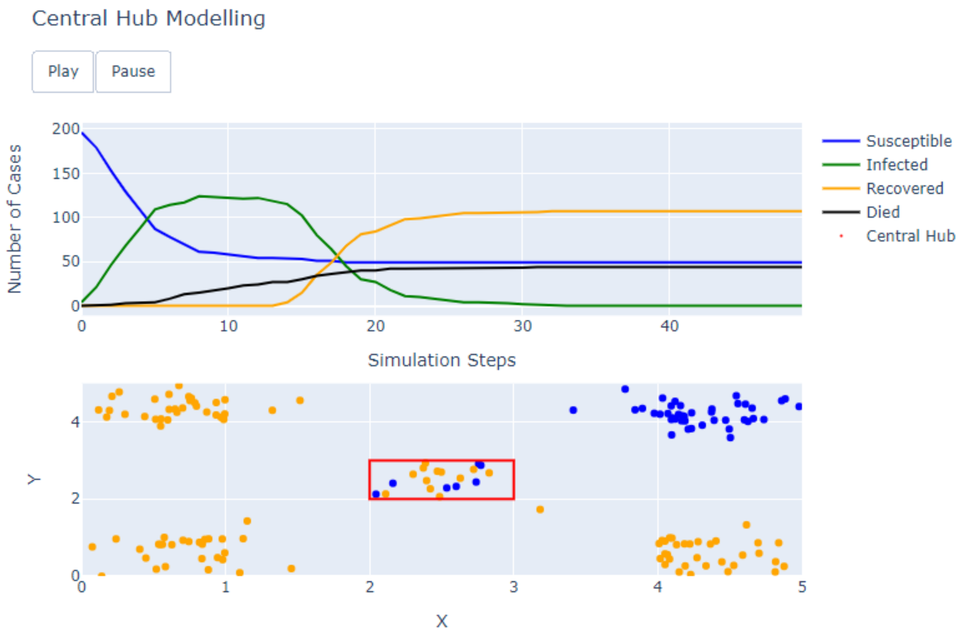


Figure 3.22: Single Central Hub

One of the most efficient strategies in order to make central hubs safer, would be to make sure that all the individuals that travel to the hub are virus-free, therefore making it impossible to spread the disease across different communities through the central hub.

3.4.4 Finance Simulation

Causal Question: How can social distancing techniques affect a community economy? Which social classes would suffer the greatest impact?

Experimental Results: Applying different types of social distancing and limited movement restrictions, could potentially lead to a good containment of the spreading of a disease, but also to a major shrink of the whole economy. In the following simulation, two main types of responses are simulated: no containment at all or imposing a hard lock-down. In order to keep track of the economic consequences of these two different approaches, the created population has been divided into 3 different classes: Working Class, Middle Class and Upper Class. Which have assigned different types of incomes and expenses they have to pay on daily basis depending on their income. The government offers the opportunity to give financial support on daily basis in case any of the citizens are struggling to pay their expenses. In a fully functioning society, most of the citizen are able to pay their expenses without having to use their savings or ask for help. As restrictions are imposed and freedom of movement is limited, citizens can only continue earning and being self-sufficient if they are able to work from home. Otherwise, they will have to make use of their savings and of the government support provided. The financial support covers just the daily expenses and can be asked every day. Because of the nature of their work, middle and higher class workers are more likely be able to work remotely.

In standard conditions, the updated income of a citizen on daily basis is equal to:

$$Updated\,Income = (Income - Expenses) \times \|Current\ Position - Previous\ Position\| \tag{3.32}$$

In this way, the ability of moving freely is represented as a way to possibly increase their income (eg. citizen can travel to work and during working activities).

Running an example simulation for 50 days, without applying any restriction would then lead to the results shown in Figure 3.23, no financial support requests and overall £4464.241 of average savings accumulated by the population.

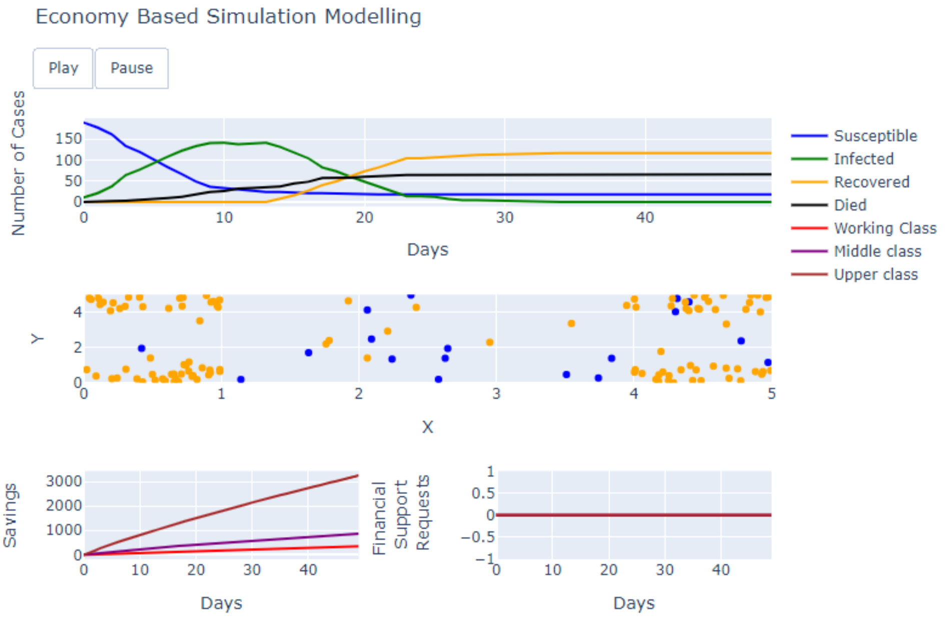


Figure 3.23: Financial Modelling (No restrictions applied)

Applying instead a lock-down would then result to just £3402.764 accumulated savings (mainly from higher social classes) and 241 financial support requests received just from the working class. Applying restrictions, led, on the other side, to a sharp decrease in the overall number of infected and deaths in the community.

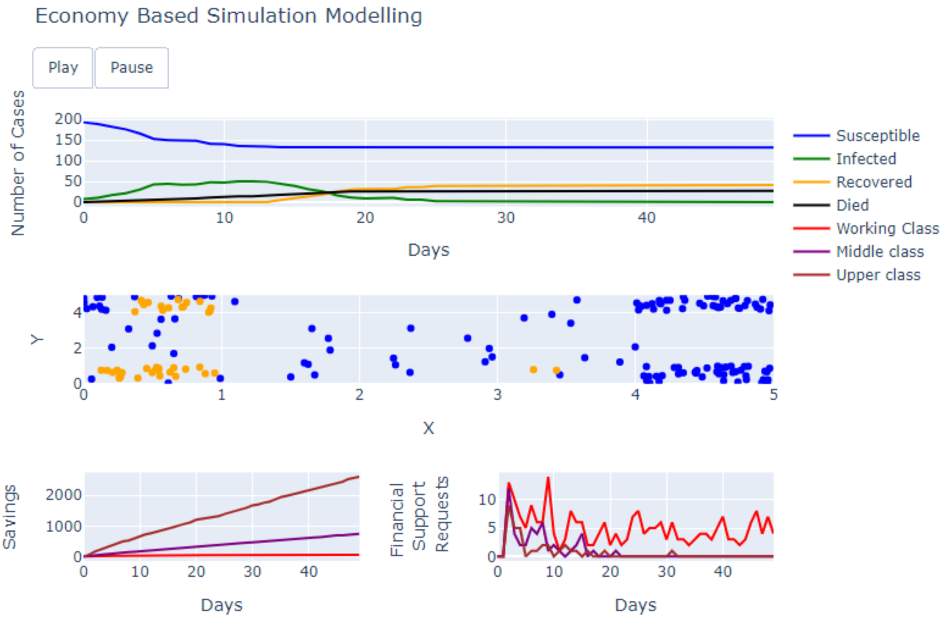


Figure 3.24: Financial Modelling (Lock-down)

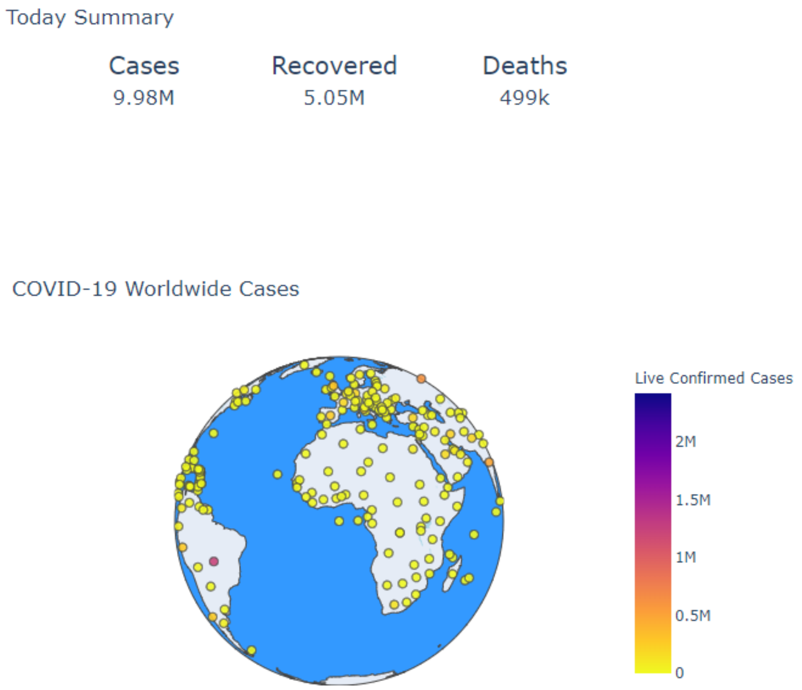
3.5 Extras

In order to make the web application, a complete tool which could be used by any type of users, a professional and user-friendly design was created using Streamlit and the app has been made available to the web by creating an Amazon Web Services (AWS) EC2 Linux Instance. This enabled to easily fetch real time data from various sources and to implement animated and interactive plots.

Some additional features which have been added to the web application are a live report of how Coronavirus has been spreading around the world and live news updates about this topic.

3.5.1 Live World Statistics and Records

In this section of the web application, summary statistics of the number of cases/recovered/deaths due to Coronavirus, have been provided (Figure 3.25).



Today Summary

Cases	Recovered	Deaths
9.98M	5.05M	499k

Figure 3.25: World Statistics Example

These included:

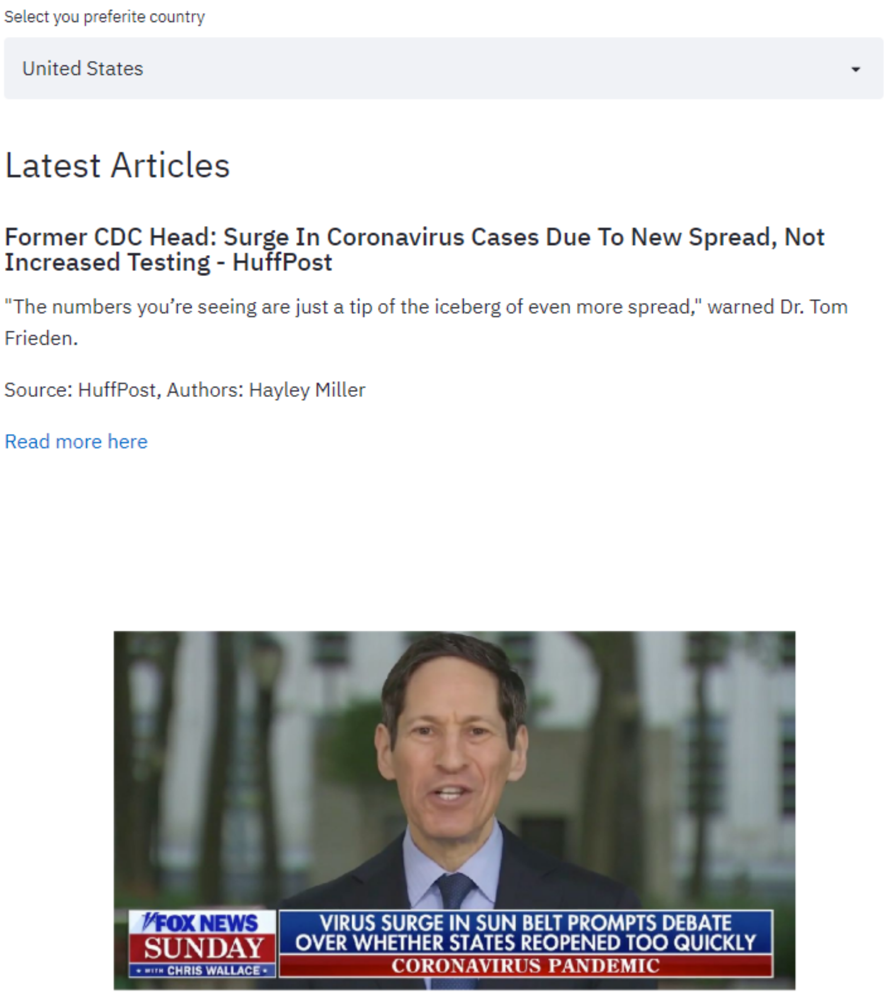
- Record of the number of cases/recovered/deaths in the world up to date.
- Interactive world view of how the number of cases are spread around the word.

- Charts displaying what are today's top 10 countries for number of cases/deaths and how their numbers changed in the last 24 hours.
- Interactive animations displaying how Coronavirus spread across different countries over time.

The data used in order to create these charts was provided by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [34] and automatically updated every 24 hours.

3.5.2 Live World News and Sentiment Analysis

This section was possible thanks to the use of the Python News API [35]. Making use of this API, live news are fetched every 2 hours from a large number of countries around the world about Coronavirus.



Select you preferite country

United States

Latest Articles

Former CDC Head: Surge In Coronavirus Cases Due To New Spread, Not Increased Testing - HuffPost

"The numbers you're seeing are just a tip of the iceberg of even more spread," warned Dr. Tom Frieden.

Source: HuffPost, Authors: Hayley Miller

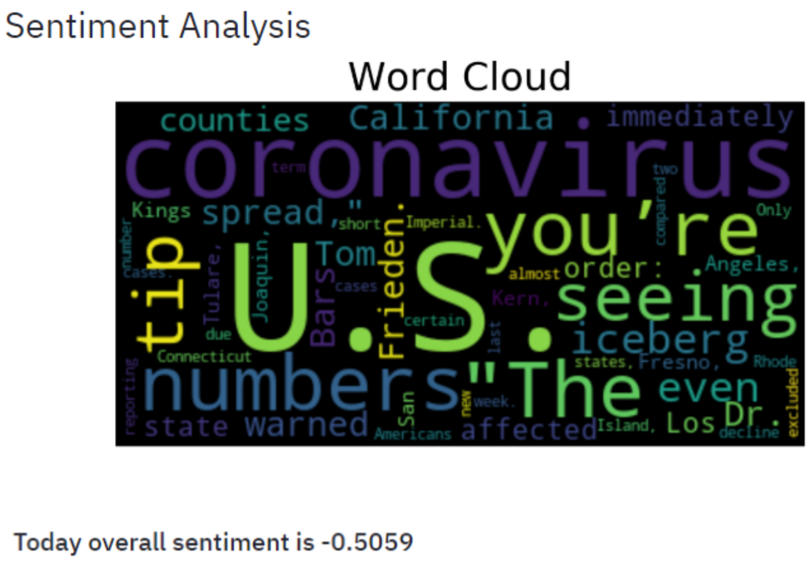
Read more here

FOX NEWS
SUNDAY
WITH CHRIS WALLACE
VIRUS SURGE IN SUN BELT PROMPTS DEBATE OVER WHETHER STATES REOPENED TOO QUICKLY
CORONAVIRUS PANDEMIC

Figure 3.26: World News Example

Using the fetched news articles, it was then possible to apply sentiment analysis in each of the different countries in order to understand what are the key words of the day and what's the overall sentiment about the outbreak. Sentiment Analysis is one of the most popular Natural Language Pre-processing (NLP) techniques used in order to infer sentiments from text data and expressing them as a numeric score. If applied to large databases (e.g. Social Networks, News Websites), this technique can be able to provide valuable information about how the public is reacting to macro changes in fields such as politics and economics.

Standard NPL techniques (eg. Tokenizzation, Stop Words Removal, Stemming) have been applied in this case study using the Python NLTK (Natural Language Toolkit) library and the sentiment was calculated using the VADER (Valence Aware Dictionary and sEntiment Reasoner) model. This type of model would then return a sentiment score between -1 (negative) and 1 (positive).



Sentiment Analysis

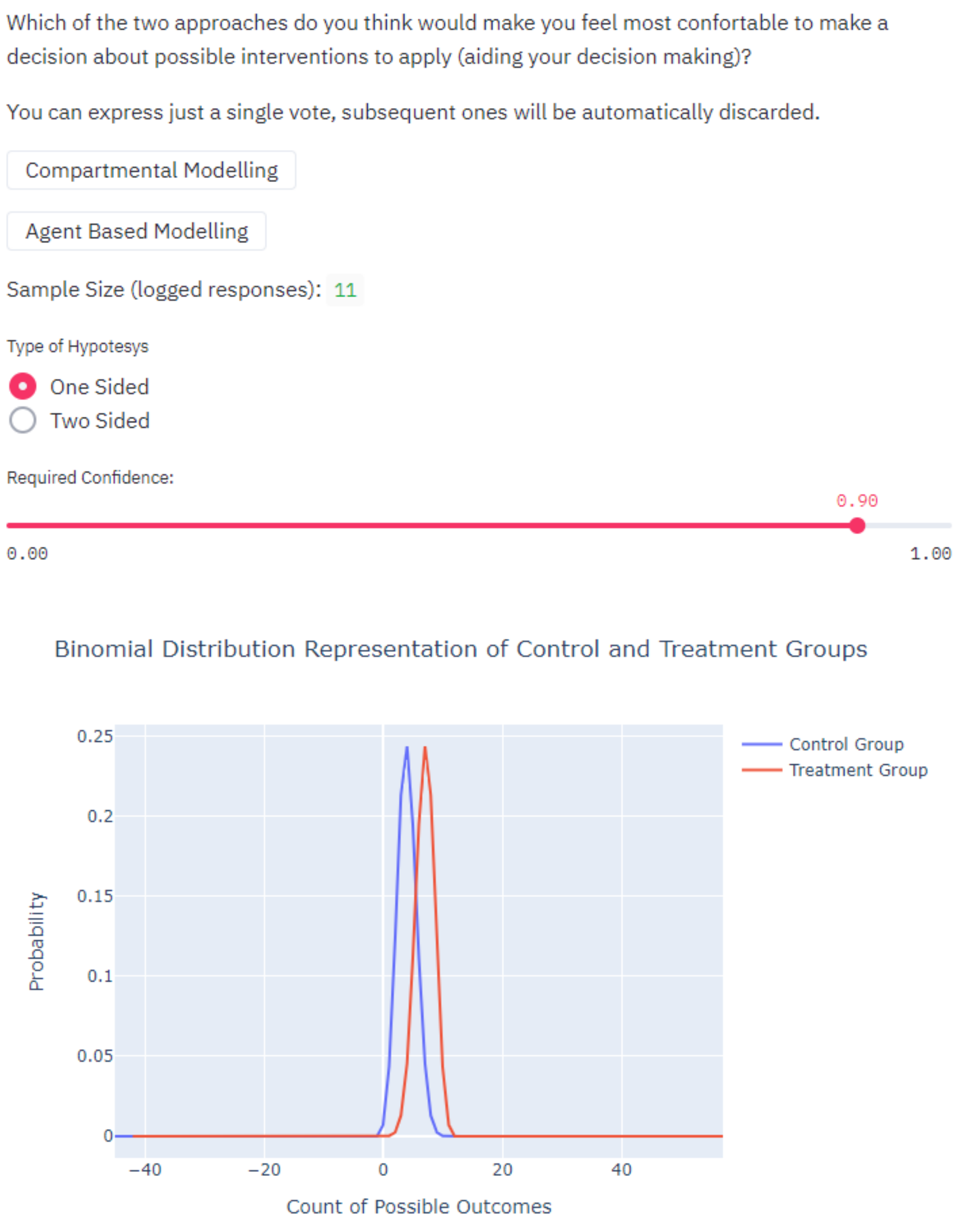
Word Cloud

Today overall sentiment is -0.5059

Figure 3.27: World News Sentiment Analysis

3.5.3 Live Feedback A/B Testing

Finally, there has been designed a section to get some feedback from the web application users and automatically generate an A/B Testing report to asses if either Agent Based Modelling or Compartmental Modelling are able to provide a statistically significant improvement for policy makers to derive conclusions (Figure 3.28).



Which of the two approaches do you think would make you feel most confortable to make a decision about possible interventions to apply (aiding your decision making)?

You can express just a single vote, subsequent ones will be automatically discarded.

Sample Size (logged responses): 11

Type of Hypotesys

● One Sided

○ Two Sided

Required Confidence: 0.90

0.00
1.00
Binomial Distribution Representation of Control and Treatment Groups
Control Group
Treatment Group
Probability
0.25
0.2
0.15
0.1
0.05
0
−40
−20
0
20
40
Count of Possible Outcomes

Figure 3.28: Real Time A/B Testing

Once logged the responses from the users, a series of summary statistics are automatically generated:

- Binomial distributions of the two examined groups.
- Conversion Rates (Equation 3.33).
- Relative Uplift (Equation 3.34).
- Z Score.
- P Value.
- Making use of the calculated metrics, is then generated a Boolean value representing if the experiment results can be considered statistically relevant or not.

$$CR_C = \frac{Conversion\,C}{Visitors} \times 100\% \qquad CR_T = \frac{Conversion\,T}{Visitors} \times 100\% \tag{3.33}$$

$$Relative\,Uplift = \frac{CR_C - CR_T}{CR_C} \times 100\% \tag{3.34}$$

Additionally, it has been made also possible for users to vary the requirements of the statistical test by choosing a confidence percentage and if to consider our type of hypothesis to be One Sided or Two Sided.

In order to validate this A/B testing tool, it has been tested against different online tools such as AB Testguide [36] for different input values.

Additional background information about A/B Testing and it's application in COVID-19 clinical trials, can be found in Appendix G.

3.6 Remarks

The different Compartmental and Agent Based Models proposed in this chapter aimed to provide an interactive tool in order to simulate how viruses can spread in a community and what can be done in order to try to control the rate of transmission (using COVID-19 as a practical case study). Although, these models can only be considered to be an approximation of the real world dynamics and might therefore not be able to capture all the key aspects related to a disease. As Scott Ambler [37] said: "The primary purpose of modelling is to provide an opportunity to think before you act".

Some examples of limitations in the proposed models are [38]:

- An inaccurate estimation of the growth rate can potentially lead to a poor prediction in the long term of the estimates of the number of infected and deaths (for example due to the stochasticity of the Agent Based Models).
- Small changes in how/when different intervention techniques are applied can potentially lead to completely different simulation results (exponential growth vs exponential decay).

- Simple models might at times provide a more accurate bigger picture of the simulation compared to complex models. In fact, simpler models might not have some constrains embedded instead in complex models which have been designed to study more specific situations.

- Using a continuous probability fraction to express how likely it is that the virus might spread from an individual to another might result in an excessively pessimistic view of how the epidemic might evolve (e.g. if a lock-down is lifted before all the cases are eradicated then the number of infected will most likely start to increase exponentially again). This problem has been partially addressed by adding a ”minimum contact radius” in the Agent Based Simulations.

- Concerning the examined Compartmental Models, systems of differential equations are highly sensitive to initial conditions and changes in them can lead to completely different outcomes.

Finally, the COVID-19 data used to fit the models might not be accurate since many countries around the world have not been able in the past few months to achieve an adequate testing capacity.

Chapter 4

COVID-19 Analysis

During Chapter 3, we had the opportunity to understand how to potentially model a disease spreading in a community. In this chapter, we are instead going to expand this topic in order to take into account how the spreading of the disease can be forecasted (using real world data) and how comorbodities can be taken into account in the case of COVID-19.

4.1 SIR Time Series Estimation

Building on from the SIR model constructed in Section 3.3.1, we can make use of it in order to approximate real world data and predict future trends [39].

This process can be summarised in the following two steps:

1. Estimating β and γ, given the data about the number of cases and recovered in a country. In order to optimise iteratively these two parameters, a modified form of the Root Mean Squared Error (RMSE) equation has been used to take into account of both the number of cases series and the number of recovered individual (Equation 4.1). In Equation 4.1, there has been additionally added an hyper-parameter (α) to decide if to give more weight either to optimising the number of cases or recovered curve fit. This exercise has therefore been designed to be an optimization minimization problem, in which the parameters estimation is improved iteratively by minimising the overall loss. This optimization process is then carried out making use of the Limited-memory BFGS (L-BFGS) algorithm. The L-BFGS algorithm is an approximation[i] of the traditional BFGS (Broyden–Fletcher–Goldfarb–Shanno) algorithm which works by estimating the Inverse of the Hessian Matrix in order to move through the search space. The calculated β and γ are then going to be used as our pa-

[i] Limited-memory BFGS, approximating the traditional BFGS algorithm manages to in fact to keep a linear memory consumption.

rameters for the SIR model.

$$Loss = \alpha \times \sqrt{\frac{1}{n}\sum_{t=1}^{n}(I_t - ID_t)^2} + (1-\alpha) \times \sqrt{\frac{1}{n}\sum_{t=1}^{n}(R_t - RD_t)^2} \tag{4.1}$$

where:	I_t, R_t	SIR Infected and Recovered timestep
	ID_t, RD_t	Infected and Recovered from data timestep
	n	Number of timesteps

2. Solve the SIR model equations by numerically integrating them and providing some initial condition values for the number of susceptible, infected and recovered in the population. The values of the initial conditions can then be calculated by taking into account the population size of the country we are examining and the number of cases registered so far. The integration method used instead to solve the system of differential equations (4^{th} order for 3 dimensions), was the Explicit Runge-Kutta method [40].

In Figure 4.1 are available the prediction results of Italy and Germany as of the 9^{th} of July 2020 in order to predict the following 30 days trends. As can been seen from the results, both countries have been fairly well approximated and the number of infected predicted in the simulation have been slightly overestimated. This mismatch although can still look quite realistic in reality because of the limited amount of tests available and presence of asymptomatic patients.

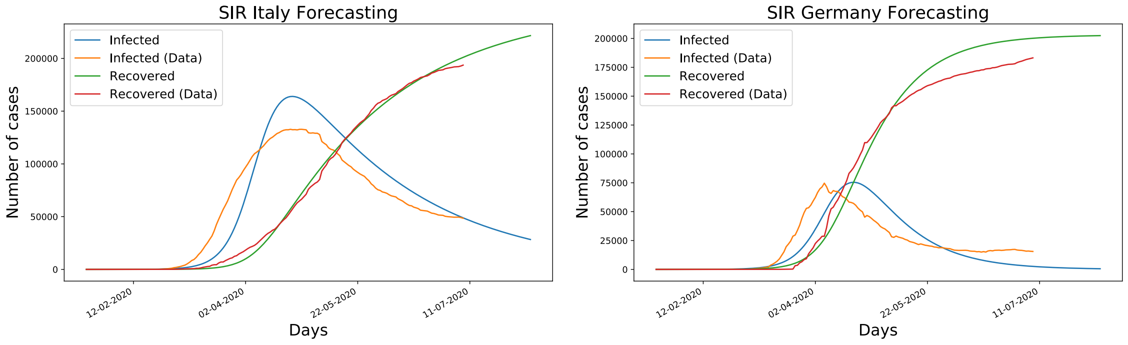


Figure 4.1: Italy and Germany SIR Forecasting

4.1.1 ML Forecasting

Another possible approach which can be taken in order to forecast time series is to use standard Machine Learning and Deep Learning techniques. In this case, the number of infected cases in Germany over time is going to be taken as an example [ii] to forecast the number of cases in 30 days time. In order to accomplish this task, the Python Darts library [41] has been used and the following models have been taken into consideration:

1. **Auto ARIMA (Auto Regressive Integrated Moving Average)**: is a time series method which can be used in order to make predictions as a linear weighted sum of past input data. In the Auto version of ARIMA, the model parameters are automatically inferred through differencing tests and optimised by recording Information Criterion (e.g. Akaike Information Criterion (AIC)) metrics.

2. **Exponential Smoothing**: this technique follows the same approach of standard ARIMA models, but the model assigns exponentially decreasing weights for past observations.

3. **LSTM (Long-Short-Term-Memory)**: The LSTM is a type of Recurrent Neural Network (RNN) ideated in order to add a memory mechanism suitable to analyse time series (the information is kept in a loop and data is fed in sequentially).

4. **T-CNN (Temporal Convolutional Neural Network)**: The T-CNN is a type of Convoluational Neural Network used for time series forecasting. This type of model is composed by a one dimensional convoluational network combined with causal convolutions. In causal convolutions, outputs at a specific timestep are convolved just with elements of the same timestep and of previous layers (in order to add time dependencies).

The results obtained from this analysis are available in Figure 4.2.

[ii] Using data up to the 9th of July 2020.

In this case, all the models have been used in order to predict a portion of the current series (so that to estimate a fit loss) and predict the next 30 days in the future of the number of cases in Germany.

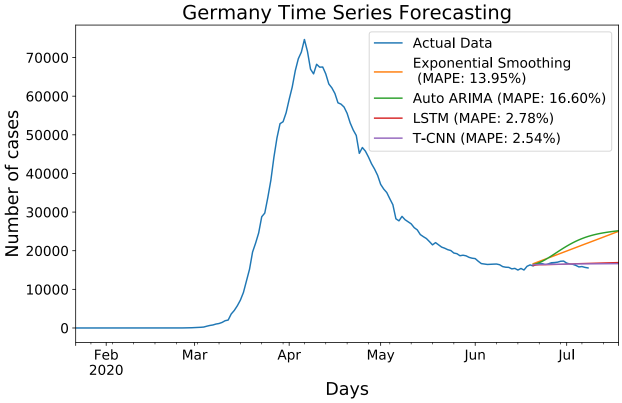


Figure 4.2: Germany ML Forecasting

In this case, MAPE (Mean Absolute Percentage Error) has been used as our loss function. MAPE is in fact one of the most commonly used loss function for regression tasks (Equation 4.2). It's main advantages are interpratibility (we work using percentage terms) and scale-independency. One of the main disadvantages of MAPE is that it can be undefined for actual values close to zero.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \times 100 \tag{4.2}$$

where: A_t, F_t Actual and forecasted time-step

n Number of datapoints

Overall, the LSTM and T-CNN managed to best fit the original data, while both Exponential Smoothing and Auto ARIMA predicted an increase in the number of cases over the following month.

Complex Deep Learning models are currently able to offer us good performances on a wide variety of tasks but they heavily rely on past data and they are not able to give us any insight about how the system might work behind the scenes. On the other hand, using for example an SIR model in order to make predictions, can allow us to not only to make estimates (like just done using ML based techniques), but also to infer underlying epidemiology parameters such as β and γ which can in turn give us more information about how the disease is spreading in a community.

4.2 Coronavirus comorbidities

Two of the greatest factors which seem to have the greatest impact over the mortality of Coronavirus for different patients, are age and possible pre-existing conditions. In different models implemented in Chapter 3, the age factor has been taken into account by varying the mortality likelihood depending on age. In this section[iii], we are instead going to have a look at which pre-existing conditions seem to have the greatest impact on COVID-19 mortality (Figure 4.3).

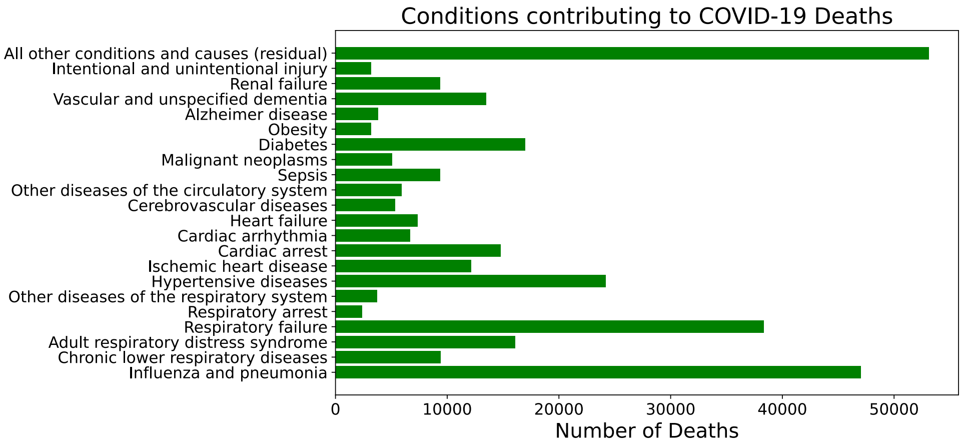


Figure 4.3: Conditions contributing to COVID-19 Deaths

As can be seen from Figure 4.3, influenza and pneumonia seem to be one of the main causes of deaths related to COVID-19. Taking a greater look at the patients ages whom died having these pre-conditions, we can then see how having a greater age can increase the overall likelihood of dying because of COVID-19 (Figure 4.4).

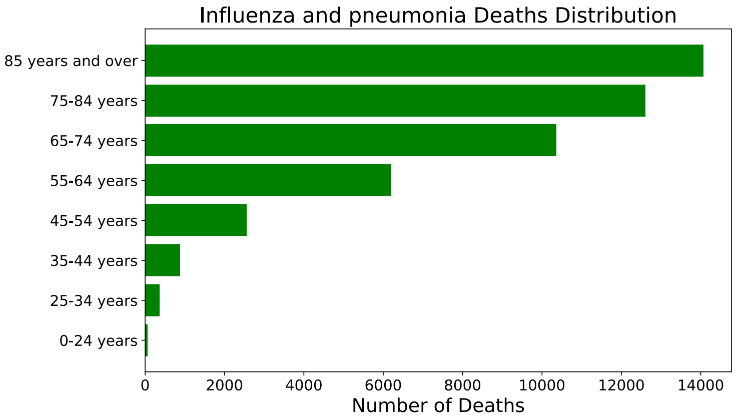


Figure 4.4: COVID-19 Deaths Distribution

[iii] Making us of the data provided by the National Center for Health Statistics [42].

4.2.1 Survival Analysis

Survival Analysis is one of the most common mathematical modelling approaches which can be used in order to estimate the time it can take for a particular event to take place. This type of approach can therefore be used in order to answer causal questions such as: What is the likelihood a COVID-19 patient would die within a given time frame? How would that change given the individual specifics (e.g. pre-existing conditions) and provided cures?

The time for an event to happen (e.g. a patient death) can be characterised in this case by a continuous non-negative random variable. This random variable can then be summarised in terms of its probability density function ($g(t)$) and cumulative density function ($G(t)$). The cumulative density function at a specific time would then give us the probability that someone might have died by then (Equation 4.3).

$$G(t) = \int_t^0 g(t)dx \tag{4.3}$$

The probability that a death might have not occurred by a specific point (Survival Function), could then be specified as shown in Equation 4.4 as $S(t) = 1 - G(t)$.

$$S(t) = \int_\infty^t g(t)dx \tag{4.4}$$

Finally, making use of the Survival Function, we can then estimate the rate at which the different patients die in our population [iv] (Hazard Function). The Hazard Function, can then be considered to be as our measure of risk to die in the provided time interval (Equation 4.5).

$$H(t) = \frac{\frac{S(t) - S(t - dt)}{dt}}{S(t)} = \frac{g(t)}{S(t)} \tag{4.5}$$

Making use of these basis, we can then implement a non-parametric model such as the **Kaplan-Meier Estimate** in order to create a probabilistic survival curve of the patients survival against time [43].

[iv] Given the provided individual characteristics.

The Kaplan-Meier Estimate can be calculated using the expression shown below. In Equation 4.6, n_i represents the patients at risk at time t_i, while d_i the number of deaths occurred so far in time.

$$\widehat{S(t)} = \prod_{i=t_i}^{t} \frac{n_i - d_i}{n_i} \tag{4.6}$$

As a practical example, let us consider we are trying to test the efficacy of an antidote for COVID-19 in order to reduce the mortality rate of patients which suffer at the same time of hypertensive diseases. We run an experiment with 400 patients and divide them into a Control (No Antidote) group and a Treatment (Using Antidote) group of 200 patients each. Using the Kaplan-Meier Estimate, we can then be able to estimate if the antidote can have a positive impact or not. For example, in Figure 4.5 there is shown a possible outcome in case the antidote has a positive effect compared to no intervention.

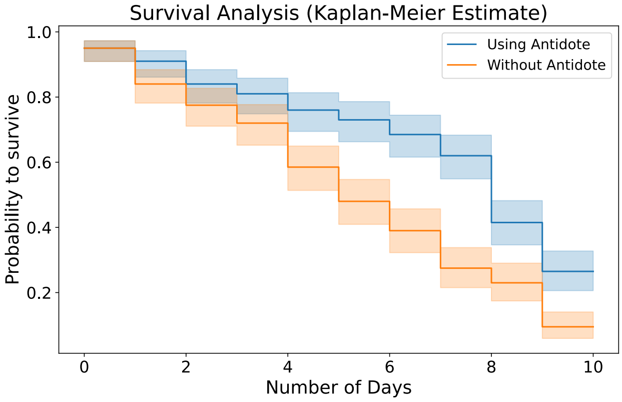


Figure 4.5: Survival Curve

As shown in Figure 4.5, all patients start with a probability to survive equal to one (before being infected) and then using the antidote can manage to limit the negative impact the disease can have on their survival probability. In Figure 4.5, there have additionally been included confidence intervals in order to take into account uncertainties due to the reduced sample size and possibility of bias in the created data.

Another possible approach which can be used to visualise how using an antidote can be of help or not to prolong the life span of the considered patients is to plot the number of days a patient survived on a time line (Figure 4.6). In order to make the data visualization easier, just a random sample of 20 patients has been considered in Figure 4.6. Patients which have survived more than 9 days have automatically been considered as recovered in this example.

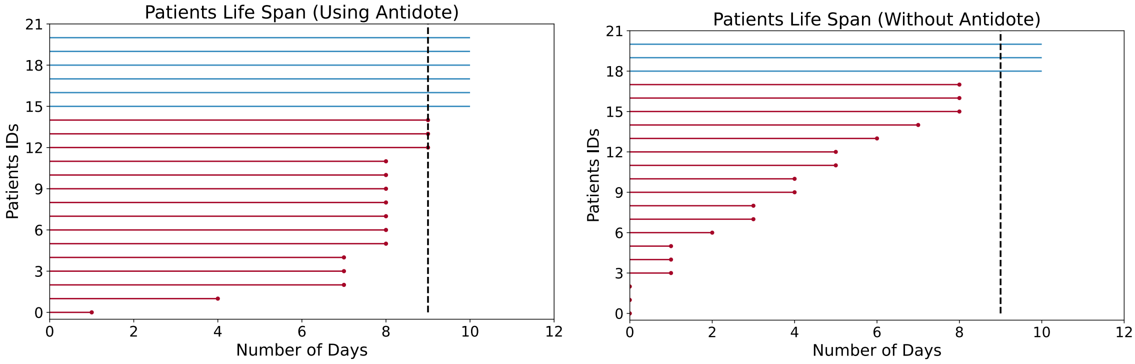


Figure 4.6: Patients survival span throughout experiment

In order to make our experiment more accurate and personalised for each individual patient, we could then try to make use of additional information about the patients (e.g. age, sex, economic status). This could be easily done, making use of more advanced models such as the Cox Proportional Hazard Model. This approach has in fact been taken in different research publications such as ”Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan” [44] and ”Survival Analysis of COVID-19 on Democracy with Cox Proportional Hazards Model” [45], so that to test new approaches to reduce the mortality rate of COVID-19.

Chapter 5

Project Management

5.1 Time Management

Throughout this project I made use of backup repositories such as:

- **Dropbox**: to store files related to this project, and to be able to access them from any type of station at any moment.
- **GitHub**: to retain a version-control of all the written code, Latex files and create a website page to interactively share code and simulation animations[i]
- **Amazon AWS EC2 Instance**: to store the Epidemic Modelling Dashboard Application and create the live version.
- **Trello Board**: all the different planned tasks have been recorded and divided on an online Trello board in order to easily plan and keep track of what tasks are left to do and any possible related reference list.

In Appendix A, there are additionally available a series of project management techniques which have been used in order to best organise and plan this project.

1. A Gantt Chart representing the planned project-schedule.
2. A Gantt Chart summarising the actual project-schedule.
3. A Risk Assessment Matrix summing up all the possible risks related to this study.
4. A Work Breakdown Structure displaying in a tree-like format the main project milestones.

Finally, in Appendix J there is available the Design Archive Guide, while in Appendix K is registered the total Word Count for the project report.

[i] Additional information available in Appendix H.

5.2 Data Management

In order to complete this project, use has been made of different freely available data sources in order to create the Epidemic Modelling Dashboard. All the different data sources used, have been acknowledged and referenced as part of this research study.

5.3 Project Challenges

One of the greatest challenges faced as part of this project, was creating the Agent Based Models outlined in Section 3.4. These models have in fact been created entirely from scratch (not having to follow strictly any mathematical model). Although, taking this approach made possible to build an easily scalable framework able to incorporate different society aspects such as spatial division in communities and economy simulations.

The different components composing these models have in fact been constructed in order to closely resemble SIR based models, while making it possible to easily incorporate much more complicated dynamics. All the different formulas used in order to design these systems had then to be constructed by hand and through experimentation (e.g. Death probability, Income Update).

Chapter 6

Conclusion

Overall, this project had a positive outcome and all the objectives established in the Project Brief (Appendix I) have been accomplished.

6.1 Summary

As part of this study, different approaches to find causal relationships in epidemiology studies have been examined by using: Compartmental Models, Agent Based Models, A/B Testing and Survival Analysis.

Using Compartmental Models, it was possible to create computationally effective simulations in order to deterministically keep track of a simulation dynamics given a set of initial conditions. One drawback of this approach, was the difficultly to keep track of individuals behaviours and spatial movements (this could be partially resolved by using sets of Partial Differential Equations).

Agent Based Modelling made it instead much easier to uniquely define the characteristics of each individual in a population creating custom behaviours both on an individual and sub-group level. Two of the main drawbacks about this type of approach is the overall high level of stochastic behaviour and the increased time complexity.

A/B testing can instead be used in situations in which we are able to run some form of controlled experiment. Controlled experiments are in fact commonly referred as the **gold standard** for causal analysis. Although, A/B tests can be quite difficult to run in situations in which it is unethical to apply some form of intervention on a group of people or when experimentation can be quite costly.

Finally, Survival Analysis is another common approach which can be used in order to asses the statistical significance of an experiment. Using simple non-parametric models such as the Kaplan-Meier Estimate it can be relatively easy to gain insights on a population level. While using instead more complex parametric models such as the Cox Proportional Hazard Model, designed intervention results can be obtained for each participant in the experiment.

6.2 Further Advancements

In order to take this project further, different aspects could be taken into consideration such as:

- The Live Feedback A/B Testing web application page could be further improved by calculating the observed power and making adjustments depending on the provided sample size.

- Using NPL, it could be possible to find patterns in research publications about research advancements to find a cure/vaccine against Coronavirus. In this way, underlying patterns could be analysed and used in order to make progresses (e.g. by combining different approaches).

- As more data about Clinical Trials would become available in the following months, it could be possible to make use of it in order to find causal relationships in how different treatments might effect different patients.

- Introduce other causality related techniques such as Knowledge Graphs and Explainable AI.

- Make use of open source causality libraries such as Microsoft DoWhy [46], Uber CausalML [47] and QuantumBlack CausalNex [48].

Finally, the principles of the proposed Agent Based and Compartmental models could possibly be applied not only to simulate the spreading of a disease but also other types of viral content such as news and ideas through a network of individual.

Bibliography

[1] To Build Truly Intelligent Machines, Teach Them Cause and Effect *Quanta Magazine. Accessed: https://www.quantamagazine.org/to-buildtruly-intelligent-machines-teach-them-cause-and-effect-20180515* July 2020.

[2] Simpson's Paradox: How to Prove Opposite Arguments with the Same Data *Will Koehrsen, Towards Data Science. Accessed: https://shorturl.at/jkyT9* August 2020.

[3] "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Marco Tulio Ribeiro, University of Washington et. al. Accessed: https://arxiv.org/pdf/1602.04938.pdf* July 2020.

[4] Black-box vs. white-box models. *Lars Hulstaert, Towards Data Science. Accessed: https://towardsdatascience.com/machine-learning-interpretability-techniques-662c723454f3* July 2020.

[5] Truly Explainable AI: Putting the "Cause" in "Because". *CausaLens. Accessed: https://media-exp1.licdn.com/dms/document/C4E1FAQFj_hm_0X1QlQ/feedshare-document-pdf-analyzed/0?e=1595084400&v=beta&t=zYhikB7P63ZSkJi2tTwGqxtM-SSlipOCF_6LZpgXUTU* July 2020.

[6] Introduction to modelling and simulation *Anu Maria, State University of New York at Binghamton. Accessed: http://acqnotes.com/Attachments/White%20Paper%20Introduction%20to%20Modeling%20and%20Simulation%20by%20Anu%20Maria.pdf* August 2020.

[7] Defence - Advanced multi-domain synthetic environments *Improbable.io. Accessed: https://improbable.io/defence* August 2020.

[8] Threat Modeling Security Fundamentals *Microsoft Learn. Accessed: https://docs.microsoft.com/en-us/learn/paths/tm-threat-modeling-fundamentals/* August 2020.

[9] The Rules of Contagion: Why Things Spread - and Why They Stop. *Adam Kucharski. Accessed: https://www.amazon.co.uk/*

Rules-Contagion-Outbreaks-Infectious-Diseases-ebook/dp/B07JLSHT7M August 2020.

[10] COVID-19 reports *Imperial College of London. Accessed: https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/covid-19-reports/* August 2020.

[11] The Seven Tools of Causal Inference with Reflections on Machine Learning. *JUDEA PEARL, UCLA Computer Science Department, USA. Accessed: https://ftp.cs.ucla.edu/pub/stat_ser/r481.pdf* July 2020.

[12] The Book of Why: The New Science of Cause and Effect. *Pearl, Judea and Mackenzie, Dana. Accessed: https://dl.acm.org/doi/book/10.5555/3238230* July 2020.

[13] Systems Thinking Systems Modelling. *A Course for Understanding Systems and Creating Systems Models. The Sustainability Laboratory's. Accessed: https://systemsinnovation.io/system-dynamics-book/* July 2020.

[14] Introduction to Causal Inference *Peter Spirtes, Department of Philosophy, Carnegie Mellon University. Accessed: http://www.jmlr.org/papers/volume11/spirtes10a/spirtes10a.pdf* July 2020.

[15] Causal Bayesian Networks: A flexible tool to enable fairer machine learning *Silvia Chiappa and William Isaac, DeepMind. Accessed: https://deepmind.com/blog/article/Causal_Bayesian_Networks* July 2020.

[16] Path-Specific Counterfactual Fairness *Silvia Chiappa and Thomas P. S. Gillam, DeepMind. Accessed: https://arxiv.org/pdf/1802.08139.pdf* July 2020.

[17] Are you guilty of using the word "experiment" incorrectly?. *Cassie Kozyrkov - Towards Data Science. Accessed: https://towardsdatascience.com/are-you-guilty-of-using-the-word-experiment-incorrectly-9068baeab7a4* July 2020.

[18] Causal Inference that's not A/B Testing: Theory Practical Guide. *Eva Gong - Towards Data Science. Accessed: https://towardsdatascience.com/causal-inference-thats-not-a-b-testing-theory-practical-guide-f3c824ac9ed2* July 2020.

[19] Elements of Causal Inference: Foundations and Learning Algorithms. *Jonas Peters, Dominik Janzing et. al. The MIT Press. Accessed: https://mitpress.mit.edu/books/elements-causal-inference* July 2020.

[20] CAUSALITY FOR MACHINE LEARNING. *Bernhard Schölkopf, Max Planck Institute for Intelligent Systems. Accessed: https://arxiv.org/pdf/1911.10500.pdf* July 2020.

[21] 2 Minute Classroom *Endemic vs Epidemic vs Pandemic — How Epidemiologists Classify Disease Prevalence. Accessed: https://www.youtube.com/watch?v=nclAnJXdgqs* June 2020.

[22] Exponential growth and epidemics *3Blue1Brown. Accessed: https://www.youtube.com/watch?v=Kas0tIxDvrg* June 2020.

[23] How To Tell If We're Beating COVID-19 *minutephysics. Accessed: https://www.youtube.com/watch?v=54XLXg4fYsc* June 2020.

[24] Infectious Disease Modelling: Understanding the models that are used to model Coronavirus *Henri Froese, Towards Data Science. Accessed: https://www.youtube.com/watch?v=nclAnJXdgqs* June 2020.

[25] Epidemic, Endemic, and Eradication Simulations *Primer. Accessed: https://www.youtube.com/watch?v=7OLpKqTriio* June 2020.

[26] Modeling the Spreading of Diseases *Center for Biomedical Computing, Simula Research Laboratory, University of Oslo. Accessed: http://hplgit.github.io/disease-modeling/doc/pub/disease_modeling-beamer-red_shadow.pdf* June 2020.

[27] What Happens Next? COVID-19 Futures, Explained With Playable Simulations *It's Nicky!, Marcel Salathé Nicky Case. Accessed: https://ncase.me/covid-19/* June 2020.

[28] United Nations *Department of Economic and Social Affairs, World Population Prospects 2019. Accessed: https://population.un.org/wpp/DataQuery/* June 2020.

[29] The World Bank Data *Hospital beds (per 1,000 people). Accessed: https://data.worldbank.org/indicator/SH.MED.BEDS.ZS* June 2020.

[30] ScienceMag.org *Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2), Ruiyun Li et. al. Accessed: https://science.sciencemag.org/content/368/6490/489* June 2020.

[31] The Centre for Evidence-Based Medicine develops, promotes and disseminates better evidence for healthcare. *COVID-19: Roche Antibody Test – 14th May.Susannah Fleming et. al. Accessed: https://www.cebm.net/covid-19/covid-19-roche-antibody-test-14th-may/* June 2020.

[32] Oxford Mathematician explains SIR Travelling Wave Disease Model for COVID-19 (Coronavirus) *Tom Rocks Maths. Accessed: https://www.youtube.com/watch?v=uSLFudKBnBI&t=1s* June 2020.

[33] Simulating an epidemic *3Blue1Brown. Accessed: https://www.youtube.com/watch?v=gxAaO2rsdIs* June 2020.

[34] COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University *CSSEGISandData, Accessed: https://github.com/CSSEGISandData/COVID-19* June 2020.

[35] Python News API *Accessed: https://newsapi.org/docs/get-started* June 2020.

[36] AB Testguide, Online Dialogue *Accessed: https://abtestguide.com/calc/* August 2020.

[37] Scott Ambler, Wikipedia *Accessed: https://en.wikipedia.org/wiki/Scott_Ambler* July 2020.

[38] What models can and cannot tell us about COVID-19. *Alexander F. Siegenfeld et. al. Accessed: https://www.pnas.org/content/pnas/early/2020/06/23/2011542117.full.pdf* July 2020.

[39] COVID-19 dynamics with SIR model *The First Cry of Atom. Accessed: https://www.lewuathe.com/covid-19-dynamics-with-sir-model.html* July 2020.

[40] Programming for Computations – Python *Linge, H.P. Langtangen. Accessed: https://core.ac.uk/download/pdf/81856903.pdf* July 2020.

[41] Time Series Made Easy in Python *Darts. Accessed: https://unit8co.github.io/darts/* July 2020.

[42] Conditions contributing to deaths involving coronavirus disease 2019 (COVID-19), by age group, United States. *National Center for Health Statistics. Accessed: https://data.cdc.gov/NCHS/Conditions-contributing-to-deaths-involving-corona/hk9y-quqm* July 2020.

[43] Survival Analysis: Intuition Implementation in Python. *Anurag Pandey, Towards Data Science. Accessed: https://towardsdatascience.com/survival-analysis-intuition-implementation-in-python-504fde4fcf8e* July 2020.

[44] Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. *Xiaochen Li, MD et. al. Accessed: https://www.sciencedirect.com/science/article/pii/S0091674920304954* July 2020.

[45] Survival Analysis of COVID-19 on Democracy with Cox Proportional Hazards Model. *Yue Zhao and Deepika Dilip. Accessed: https://www.researchgate.net/publication/342228658_Survival_Analysis_of_COVID-19_on_Democracy_with_Cox_Proportional_Hazards_Model* July 2020.

[46] DoWhy — Making causal inference easy. *Amit Sharma, Emre Kiciman - Microsoft. Accessed: https://microsoft.github.io/dowhy/* July 2020.

[47] Causal ML: A Python Package for Uplift Modeling and Causal Inference with ML. *UBER. Accessed: https://github.com/uber/causalml* July 2020.

[48] CausalNex: A toolkit for causal reasoning with Bayesian Networks. *QuantumBlack. Accessed: https://github.com/quantumblacklabs/causalnex* July 2020.

[49] RIP correlation. Introducing the Predictive Power Score *Florian Wetschoreck, Towards Data Science. Accessed: https://towardsdatascience.com/rip-correlation-introducing-the-predictive-power-score-3d90808b9598* July 2020.

[50] Correlation and dependence *Wikipedia, the free encyclopedia. Accessed: https://en.wikipedia.org/wiki/Correlation_and_dependence* July 2020.

[51] COVID-19 Clinical Trials dataset *Parul Pandey, Kaggle. Accessed: https://www.kaggle.com/parulpandey/covid19-clinical-trials-dataset* July 2020.

[52] Understanding Power Analysis in AB Testing *Paulynn Yu, Towards Data Science. Accessed: https://towardsdatascience.com/*

understanding-power-analysis-in-ab-testing-14808e8a1554 July 2020.

[53] Doc Word Counter. *PDF - Word Counter, Counts the real number of words in any document format. Accessed: https://docwordcounter.com/en/PDF-word-counter* July 2020.

Appendices

A Project Management

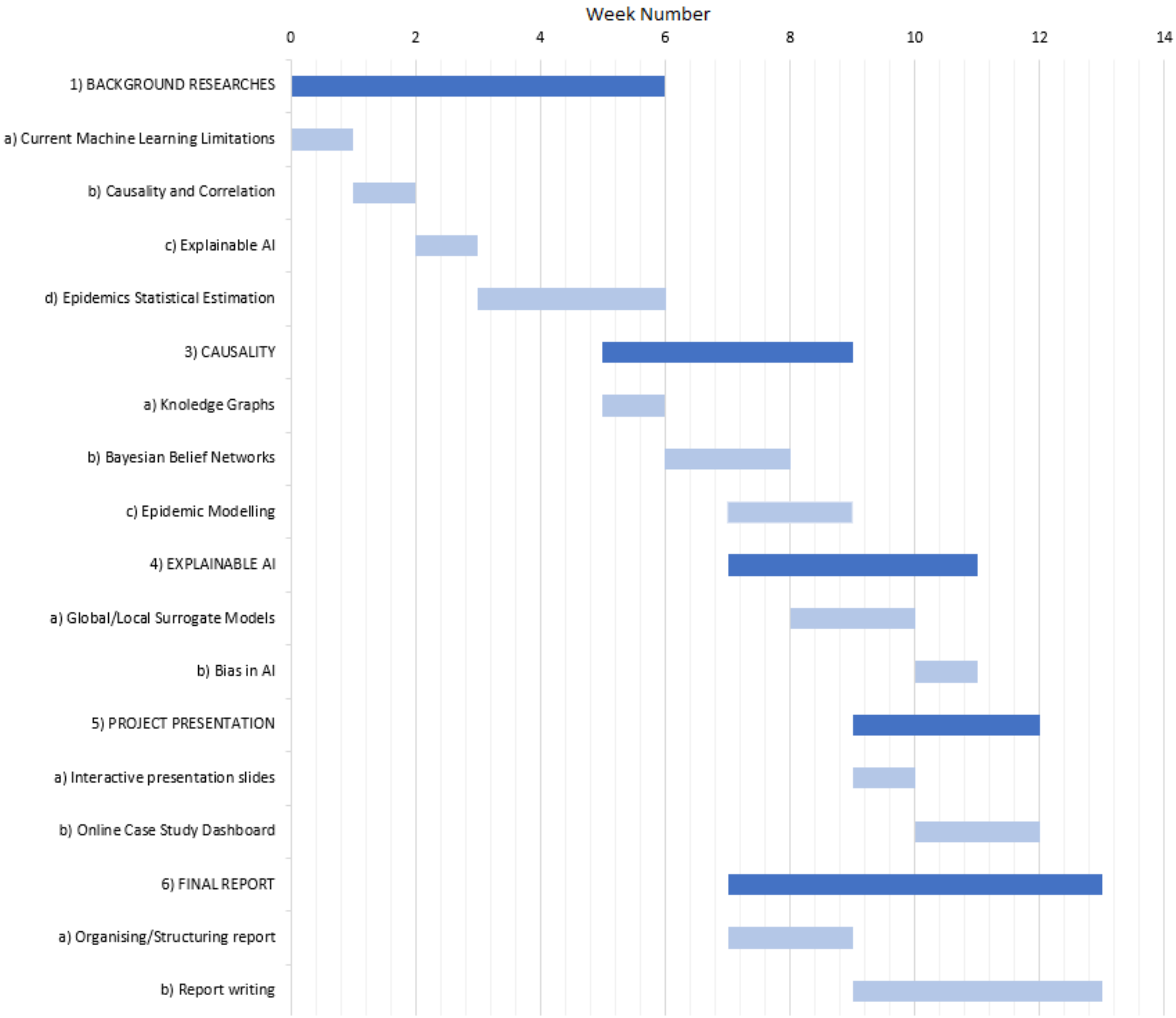


Figure A.1: Planned Gantt Chart

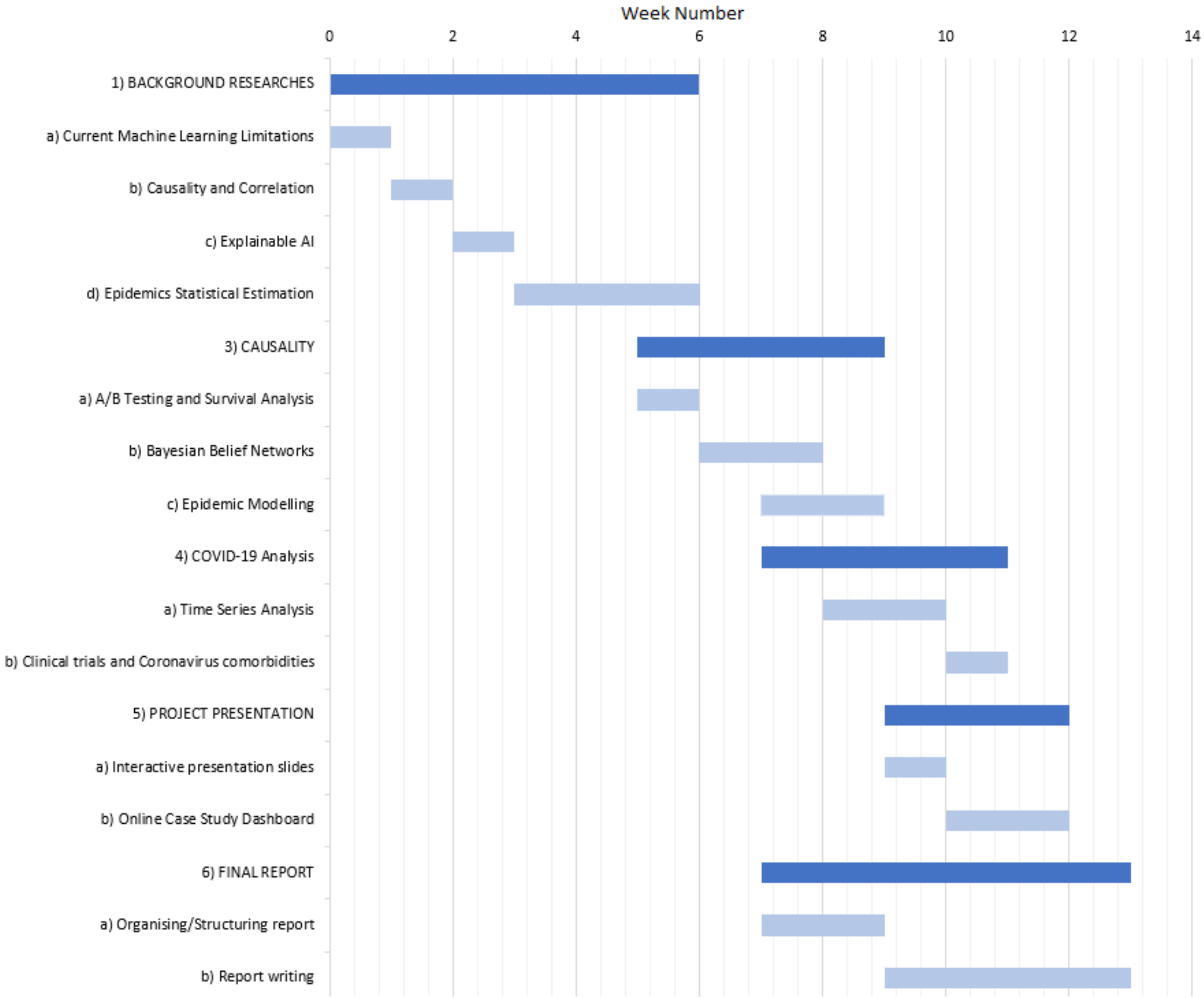
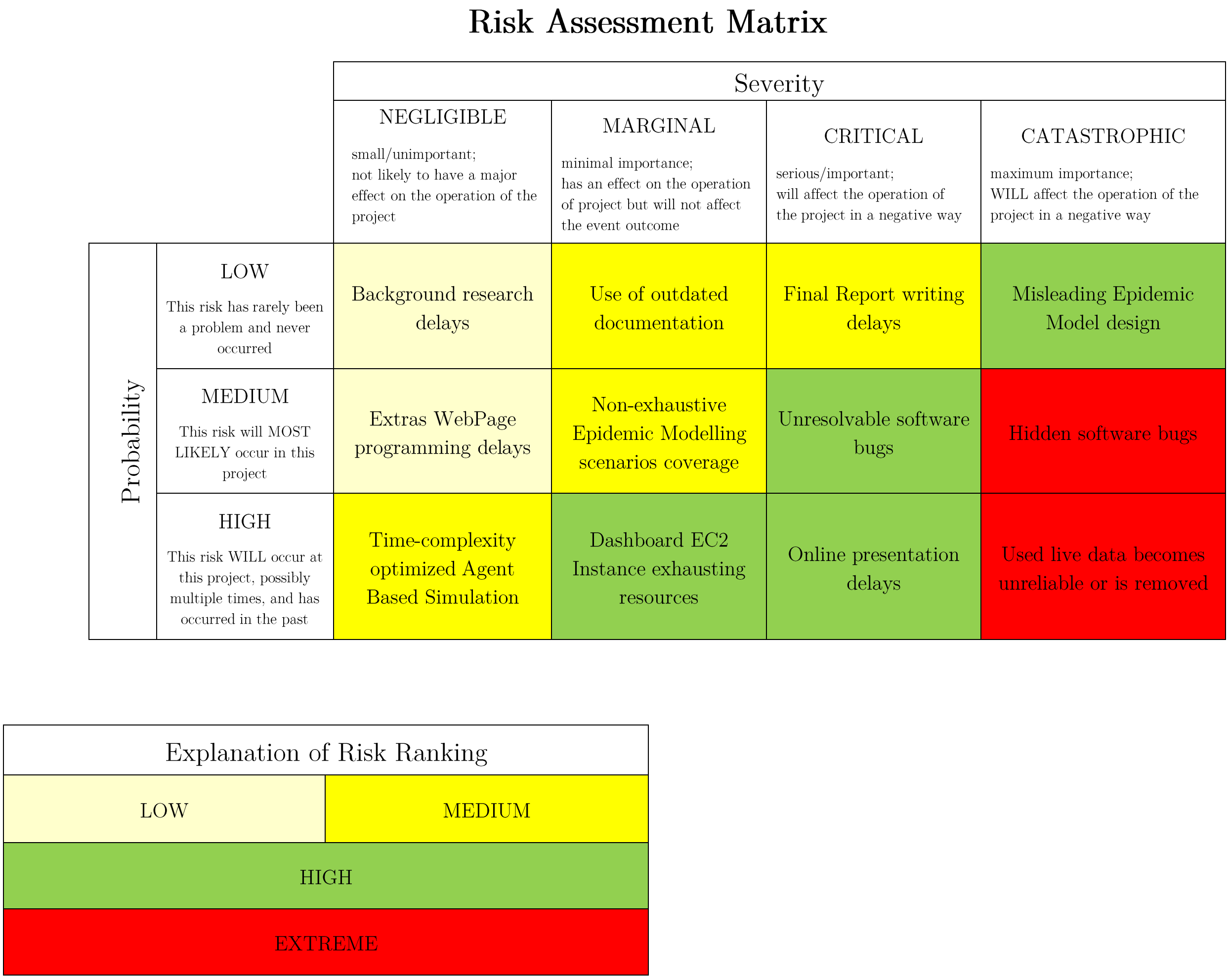


Figure A.2: Actual Gantt Chart



Risk Assessment Matrix

		Severity			
		NEGLIGIBLE small/unimportant; not likely to have a major effect on the operation of the project	MARGINAL minimal importance; has an effect on the operation of project but will not affect the event outcome	CRITICAL serious/important; will affect the operation of the project in a negative way	CATASTROPHIC maximum importance; WILL affect the operation of the project in a negative way
Probability	LOW This risk has rarely been a problem and never occurred	Background research delays	Use of outdated documentation	Final Report writing delays	Misleading Epidemic Model design
	MEDIUM This risk will MOST LIKELY occur in this project	Extras WebPage programming delays	Non-exhaustive Epidemic Modelling scenarios coverage	Unresolvable software bugs	Hidden software bugs
	HIGH This risk WILL occur at this project, possibly multiple times, and has occurred in the past	Time-complexity optimized Agent Based Simulation	Dashboard EC2 Instance exhausting resources	Online presentation delays	Used live data becomes unreliable or is removed

Explanation of Risk Ranking	
LOW	MEDIUM
HIGH	
EXTREME	

Figure A.3: Risk Assesment Matrix

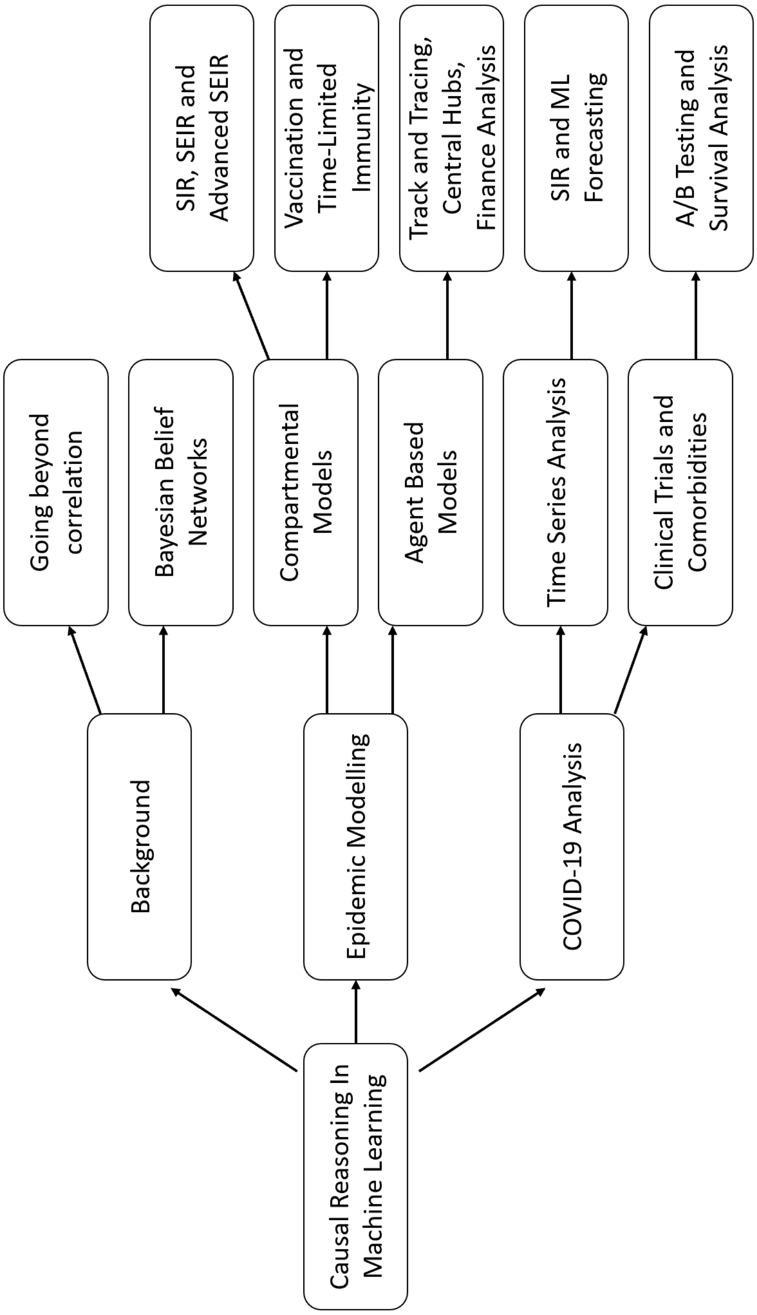

Causal Reasoning In Machine Learning
Background
Epidemic Modelling
COVID-19 Analysis
Going beyond correlation
Bayesian Belief Networks
Compartmental Models
Agent Based Models
Time Series Analysis
Clinical Trials and Comorbidities
SIR, SEIR and Advanced SEIR
Vaccination and Time-Limited Immunity
Track and Tracing, Central Hubs, Finance Analysis
SIR and ML Forecasting
A/B Testing and Survival Analysis

Figure A.4: Work Breakdown Structure

B Power Predictive Score (PPS)

During the last few years, different approaches have been taken in order to try to overcome correlation limitations. This research focus led then to the development of the "Causality Revolution" and alternative metrics to correlation such as the **Predictive Power Score (PPS)** [49].

Typical correlation analysis is able to tell us if there is a linear relationship between different variables by returning a score between -1 (e.g. if one variable increase in values, the other decreases) and 1 (e.g. if one variable increases in value, the other one will follow a similar trend). Although, correlation is not able to identify any non-linear relationship and is not able to handle non-numeric data. In the case of categorical data, this could potentially be converted into numerical data by using for example One Hot Encoding or Word Embedding techniques but would most likely lead to an increase of the dataset dimensionality in order to achieve good results. Finally, correlation might not be able to understand if relationships between the different columns are symmetric or asymmetric.

In Figure B.5, is available a summary of some examples of correlation trends and limitations.

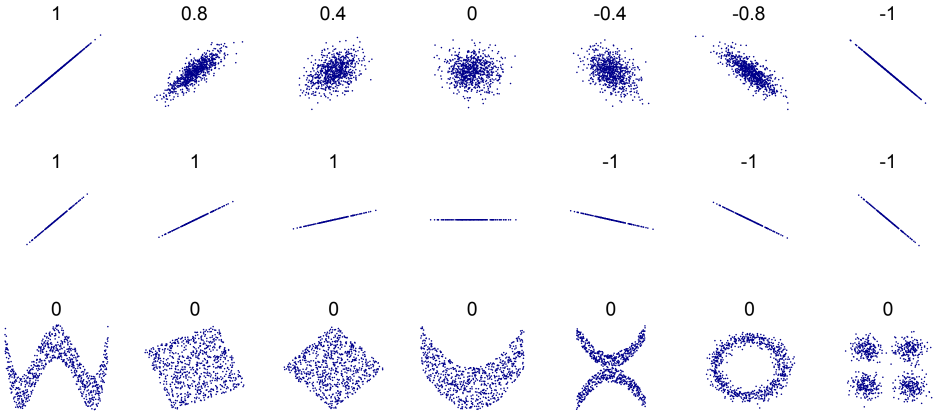


Figure B.5: Correlation Trends (Image reproduced from [50])

The predictive power score has been ideated in order to try to overcome the presented correlation limitations. One possible way to calculate the PPS is to train a Cross-Validated Decision Tree model on one feature and consider the other feature we want to consider as our label. We can then evaluate the model using an appropriate evaluation metric and then normalise the score by comparing it with the score obtained by a naive predictor. A possible set-up for a PPS calculation is shown in Table 1.

	Numeric Data	Categorical Data
ML Model	Decision Tree Regressor	Decision Tree Classifier
Evaluation Metric	Mean Absolute Error	Weighted F1 Score
Naive Predictor	Predict median value	Predict most common class

Table 1: Predictive Power Score Set-up

Where the Mean Absolute Error (MAE) and F1 score can be defined as:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad MAE = \frac{1}{n}\sum_{i=1}^{n} |x_i - x| \tag{1}$$

The result from the evaluation metric can then be normalised by comparing it with the results from the naive predictor. In the case of the F1 Score, one is going to be considered our upper limit and the naive predictor score as our lower limit.

$$PPS = \frac{Decision\,Tree\;(F1) - Naive\;Predictor\;(F1)}{1 - Naive\;Predictor\;(F1)} \tag{2}$$

A similar formula could then be calculated for the MAE case, but zero should be considered as our lower limit (in this case, lower scores are considered as better).

$$PPS = 1 - \frac{Decision\,Tree\;(MAE)}{Naive\;Predictor\;(MAE)} \tag{3}$$

Following this procedure, we would then have a PPS score between 0 (no relationship) and 1 (perfect relationship) able to capture either linear/non-linear relationships and to work with either numerical or categorical data.

As a simple demonstration of the PPS score, there is shown in Figure B.6 a noisy cosine function. In this example, the X axis has been realised by creating a uniform range between 0 and 1000, while the Y axis has been created by passing the respective X value in a cosine function and adding some small noise on the result. In this way, it is designed a clear non-linear dependence between X and Y.

Calculating the correlation between the two features would then lead to a result equal to zero (from either points of view). Using instead the PPS would then lead as expected to a score of 0.737 of X respect to Y and of 0 for Y respect to X.

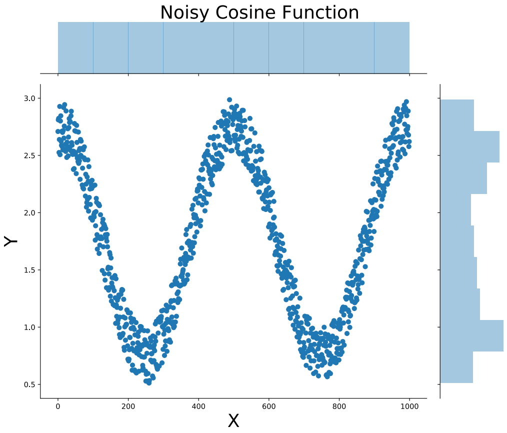


Figure B.6: PPS Score Example

The Predictive Power Score is just one of the different approaches which can be taken in order to go beyond correlation traditional limitations, other examples are: Causality, relative entropy and Granger techniques.

C Logistic/Exponential Curve Fitting

For a logistic curve at the turning point:

$$Slope = Growth\ Factor/2 \Rightarrow \quad Doubling\ Time\ (DT) = \frac{ln(2)}{Growth\ Factor/2} \tag{4}$$

Instead, for an exponential curve:

$$Slope = Growth\ Factor \Rightarrow \quad Doubling\ Time\ (DT) = \frac{ln(2)}{Growth\ Factor} \tag{5}$$

A worked out example with the results from the top three countries with the most number of Coronavirus Cases as of the end of June 2020, is available below. From this example, we can easily see how well our data resembles a logistic/exponential curve (using the R^2 score to quantify the mismatch) and what's the predicted time for the number of cases to double given the current trends.

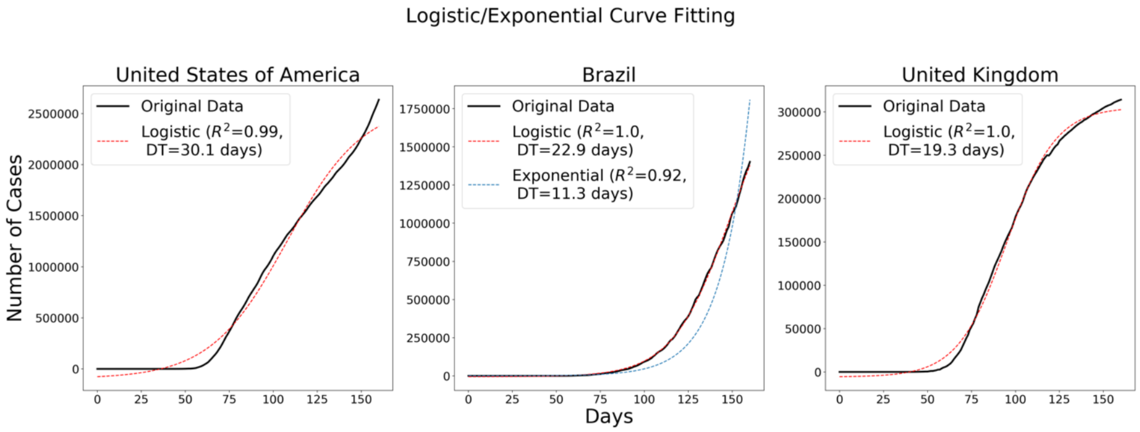
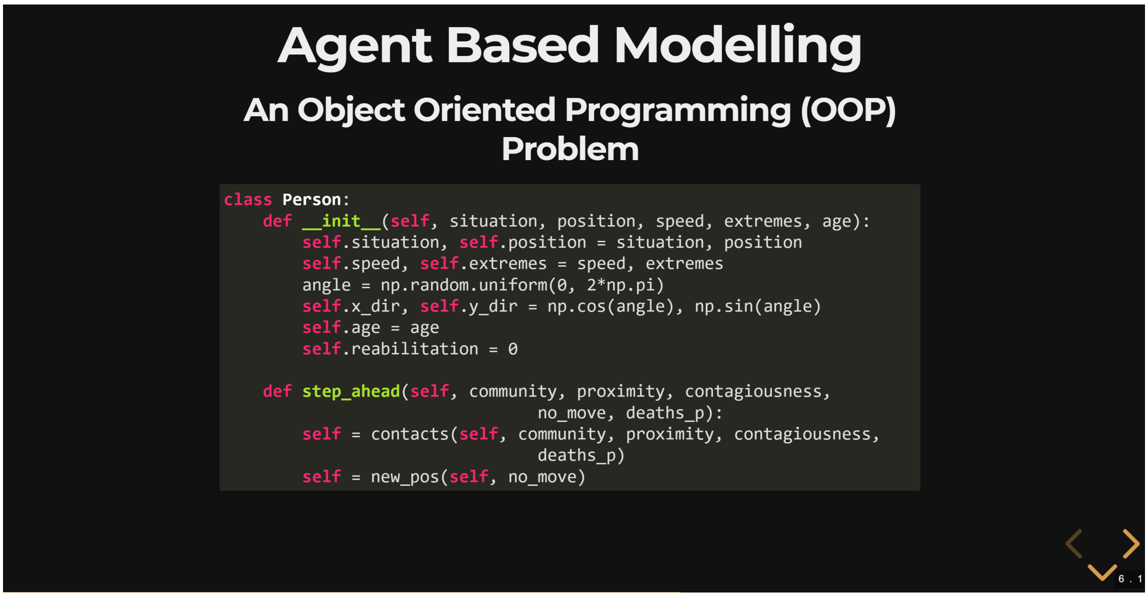


Figure C.7: Curve Fitting

D Project Demonstration

Two of the main functionalities of the created secondary GitHub pages website are a Reveal.js online presentation of the whole project and a D3.js scroller page created for interactively presenting and explaining different concepts of this research project. In Figure D.8, there is available an example of part of the created Reveal.js presentation.



Agent Based Modelling

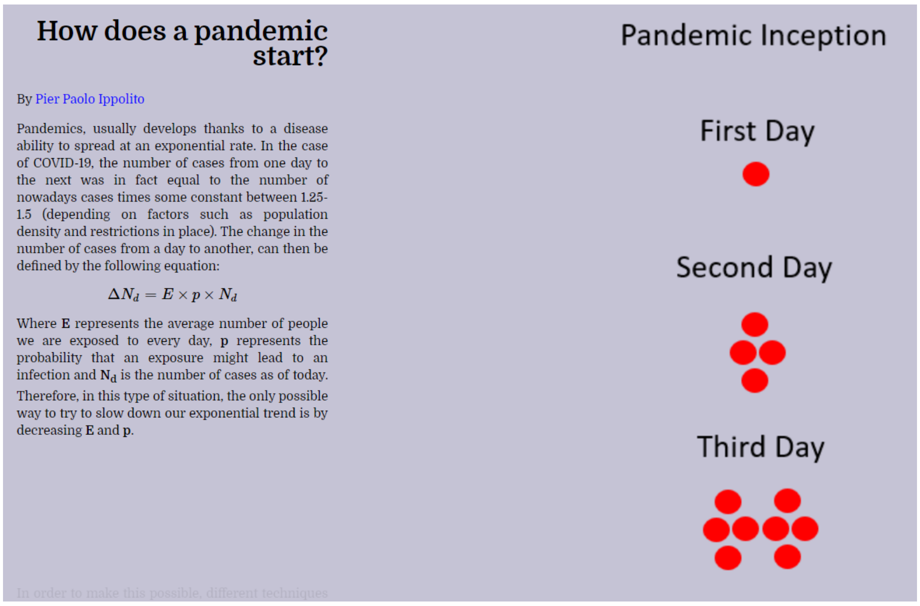
An Object Oriented Programming (OOP) Problem

```
class Person:
    def __init__(self, situation, position, speed, extremes, age):
        self.situation, self.position = situation, position
        self.speed, self.extremes = speed, extremes
        angle = np.random.uniform(0, 2*np.pi)
        self.x_dir, self.y_dir = np.cos(angle), np.sin(angle)
        self.age = age
        self.reabilitation = 0

    def step_ahead(self, community, proximity, contagiousness,
                            no_move, deaths_p):
        self = contacts(self, community, proximity, contagiousness,
                            deaths_p)
        self = new_pos(self, no_move)
```

Figure D.8: Reveal.js Presentation

In Figure D.9, there is instead shown the first section of the created D3.js story-telling narrative.



How does a pandemic start?

By Pier Paolo Ippolito

Pandemics, usually develops thanks to a disease ability to spread at an exponential rate. In the case of COVID-19, the number of cases from one day to the next was in fact equal to the number of nowadays cases times some constant between 1.25-1.5 (depending on factors such as population density and restrictions in place). The change in the number of cases from a day to another, can then be defined by the following equation:

$$\Delta N_d = E \times p \times N_d$$

Where **E** represents the average number of people we are exposed to every day, **p** represents the probability that an exposure might lead to an infection and $\mathbf{N_d}$ is the number of cases as of today.

Therefore, in this type of situation, the only possible way to try to slow down our exponential trend is by decreasing **E** and **p**.

In order to make this possible, different techniques

Figure D.9: D3.js Scroller

E Compartmental Models Causal Diagrams

In this Appendix there are available the Causal Diagrams of the epidemic compartmental models introduced in Section 3.3.1.

The SIR and SEIR models can be described as shown in Figure E.10 using a causal chain.

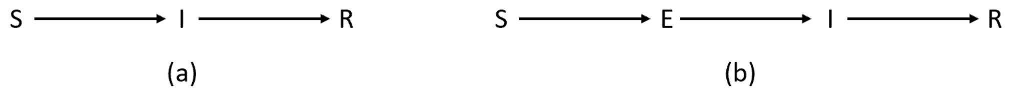


Figure E.10: SIR and SEIR

The SEIR model including also a deaths compartment can instead be described by a chain followed by a fork junction. Finally, models including time-limited immunity can be created by introducing a possibility for cycles in the graph.

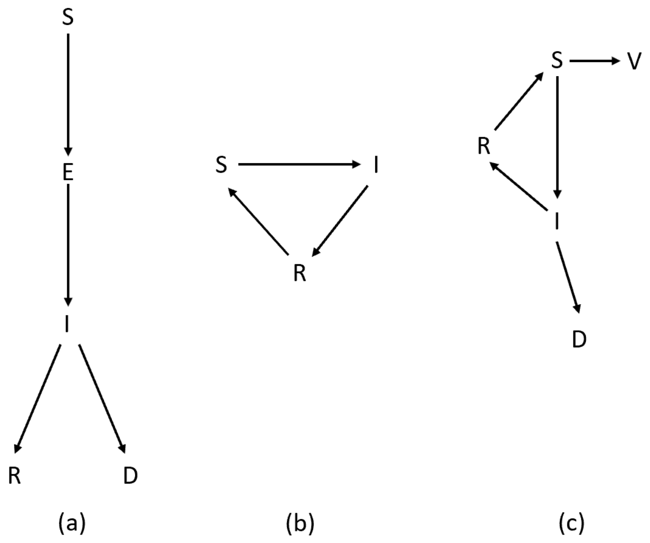


Figure E.11: Advanced Models

F Population Modelling Pseudo-code

Algorithm 1 Population Modelling Pseudo-code Outline

```
$E \Leftarrow Contact\ Radius$
$\overline{p} \Leftarrow Unlikeliness\ of\ Spread$
$p\_died \Leftarrow Death\ Probability\ dependent\ on\ age$
for $day\ in\ simulation\ days$ do
    for $individual\ in\ population$ do
        $Record\ individual\ status$
        if $Infected$ then
            if $Draw\ with\ probability\ (p\_died \times age == 1)$ then
                $individual\ status \Leftarrow Dead$
            else
                if $(Rehabilitation\ days == 14)$ then
                    $individual\ status \Leftarrow Recovered$
                end if
                $Rehabilitation\ days += 1$
            end if
        else if $Susceptible$ then
            $close\_people = 0$
            for $friend\ in\ community$ do
                if $(Friend == Infected)\ and\ (Euclid\ Dist < E)$ then
                    $close\_people += 1$
                end if
            end for
            if $(Draw\ with\ probability\ dependent\ on\ close\_people\ and\ \overline{p} == 1)$ then
                $individual\ status \Leftarrow Infected$
            end if
        end if
        if $(Static == False)$ then
            $individual\ X\ and\ Y\ position\ update$
            if $individual\ X\ or\ Y\ position\ out\ of\ boundaries$ then
                $Adjust\ position\ and\ reverse\ movement\ direction$
            end if
        end if
    end for
end for
```

G Clinical trials

Making use of the data from the United States ClinicalTrials.gov website, it has been possible to gain insights about current clinical trials as of July 2020 against COVID-19 [51]. As can be seen from Figure G.12, Hydroxychloroquine has been so far the most common treatment tried.

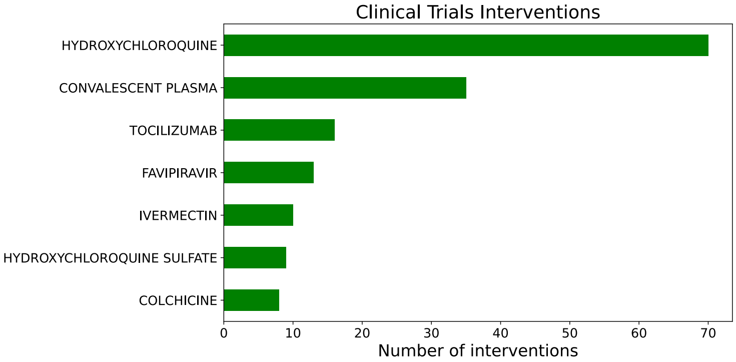


Figure G.12: COVID-19 Clinical Trials

As outlined in Section 2.4.1, clinical trials are one of the main examples of A/B testing application. Gathering results from an A/B test, we can then be able to infer causal relationships about what can be the potential effects of a treatment [52].

In this setting, the Causal question we are asking ourselves is: Does a certain treatment decrease COVID-19 mortality rate? This question can then be formulated in statistical terms as a **null and alternative hypothesis**. In the null hypothesis, applying the treatment would not lead to any major change in the mortality rate and therefore both the treatment and control groups will be quite similar. Instead, in the alternative hypothesis, the treatment would cause a statistically significant change between the two groups.

In our case, patients affected by COVID-19 would be considered as our population and our intervention (providing experimental medical treatment), would then be compared to no intervention. Variation in mortality rate could then be used as our metrics to asses the results. Patients should then be randomly assigned to either groups so that to avoid introduction of any form of bias (e.g. patients age, comorbidities, geographical location). If bias is unconsciously introduced, then this could lead to some form of **confounding bias**, which would then make it really difficult to disentangle what are effects due to the intervention and which ones are instead caused by a flaw in the randomization process.

Following on with our example, the number of times an intervention leads to a substantial difference compared to the control group can then be summarised over a number of trials as a Binomial Distribution. In this distribution, the X axis will represent the count of possible outcomes, while the Y axis will represent the probability associated with an outcome. Although, according to the Central Limit Theorem, as we would increase our sample size, we would then end up with a Gaussian Distribution for each group in the experiment.

In Figure G.13, there is shown a possible outcome for our example. As can be seen from the diagram, four different areas are present: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).

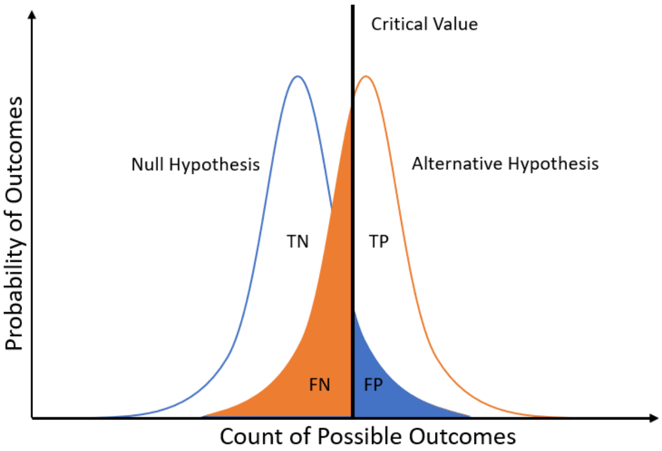


Figure G.13: A/B Test Distributions

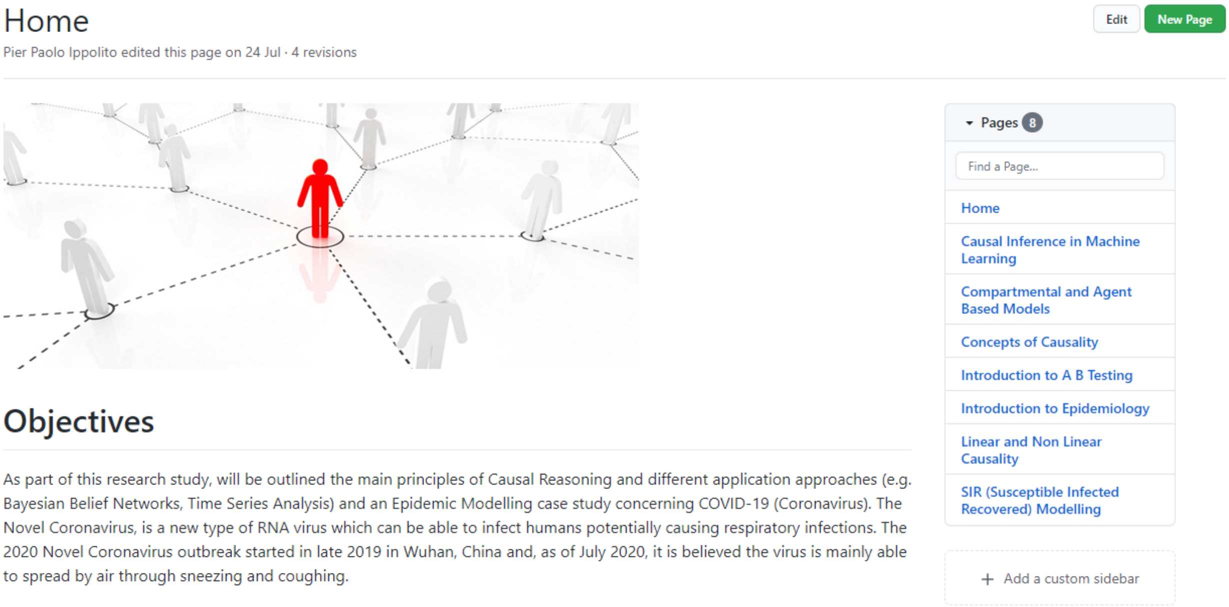
In the TP case, we can affirm our treatment is beneficial since it managed to pass our test for the hypothesis (e.g. decreasing the mortality rate). In the case of the TN area, we can instead be confident that our treatment is not beneficial. While in the FP area, we might be deceived to believe our intervention was beneficial while it wasn't (the opposite holds true instead in the FN area). In these last two cases, it is then vital to look for any possible form of bias interference.

H Project Repository

Throughout this project, a private GitHub repository has been used as a key version control tool. The overall project has been structured into 4 different branches:

- **Master:** In this branch, all the code and support files necessary in order to create the web application outlined in Chapter 3 have been stored. The web application has been mainly developed on a local machine and then deployed on an Amazon Web Services EC2 Instance through Git control.
- **Extras:** In this section there has instead been developed different parts of this project such as the code used to explain the Simpson Paradox (Section 1.2), Coronavirus comorbidities (Section 4.2), Survival Analysis (Section 4.2.1), Power Predictive Score (Appendix B), Clinical trials (Appendix G). Additionally, for the SIR Time Series Estimation model introduced in Section 4.1, the has also been created an additional set-up in order to make possible to run this model easily through the command line.
- **Gh-pages:** This branch has instead been used to create and deploy through GitHub Pages a support website. This website has then been used to share two online presentations about the project (for the second examiner demonstration) and to create two online interactive shareable code notebook in order to explain how different aspects of the project have been programmed (Appendix D).
- **Thesis:** This project report has been entirely created using LaTeX and synchronised with this branch of the repository in order to keep a backup and version control of the writing itself.

Additionally, a Wiki page has been included as part of the repository documentation in order to provide adequate background reading and context in case this project is going to be developed any further in the future (Figure H.14).


Edit
New Page
Pages 8
Find a Page...
Home
Causal Inference in Machine Learning
Compartmental and Agent Based Models
Concepts of Causality
Introduction to A B Testing
Introduction to Epidemiology
Linear and Non Linear Causality
SIR (Susceptible Infected Recovered) Modelling
+ Add a custom sidebar

Home

Pier Paolo Ippolito edited this page on 24 Jul · 4 revisions

Objectives

As part of this research study, will be outlined the main principles of Causal Reasoning and different application approaches (e.g. Bayesian Belief Networks, Time Series Analysis) and an Epidemic Modelling case study concerning COVID-19 (Coronavirus). The Novel Coronavirus, is a new type of RNA virus which can be able to infect humans potentially causing respiratory infections. The 2020 Novel Coronavirus outbreak started in late 2019 in Wuhan, China and, as of July 2020, it is believed the virus is mainly able to spread by air through sneezing and coughing.

Figure H.14: Repository Wikipedia

Finally, Github Actions have been included in order to add some form of Continuous Integration (CI) support in line with common DevOps (Development-Operations) principles.

I Project Brief

MSc Project Brief
Causal Reasoning in Machine Learning

Student:
Ippolito, Pier Paolo
ppi1u16@soton.ac.uk

Supervisor:
Dasmahapatra, Srinandan
sd@ecs.soton.ac.uk

- Problem

Thanks to recent advancements in Machine Learning and Deep Learning, has been possible for Artificial Intelligence (AI) models to achieve superhuman performance in specific applications. Although, these type of models are currently not able to generalise to a good extent for different (but similar) types of applications. Additionally, good performance of Deep Learning models is highly dependent on providing large amount of data. Due to these limitations, it could then be almost impossible to manage to create any form of Strong AI architecture. One possible approach which can be used in order to overcome these type of limitations, is to design models able to capture Causal Relationships between different variables in a dataset (e.g. Supervised/Unsupervised Learning) or elements in an environment (e.g. Reinforcement Learning).

- Goals

This project aims to:

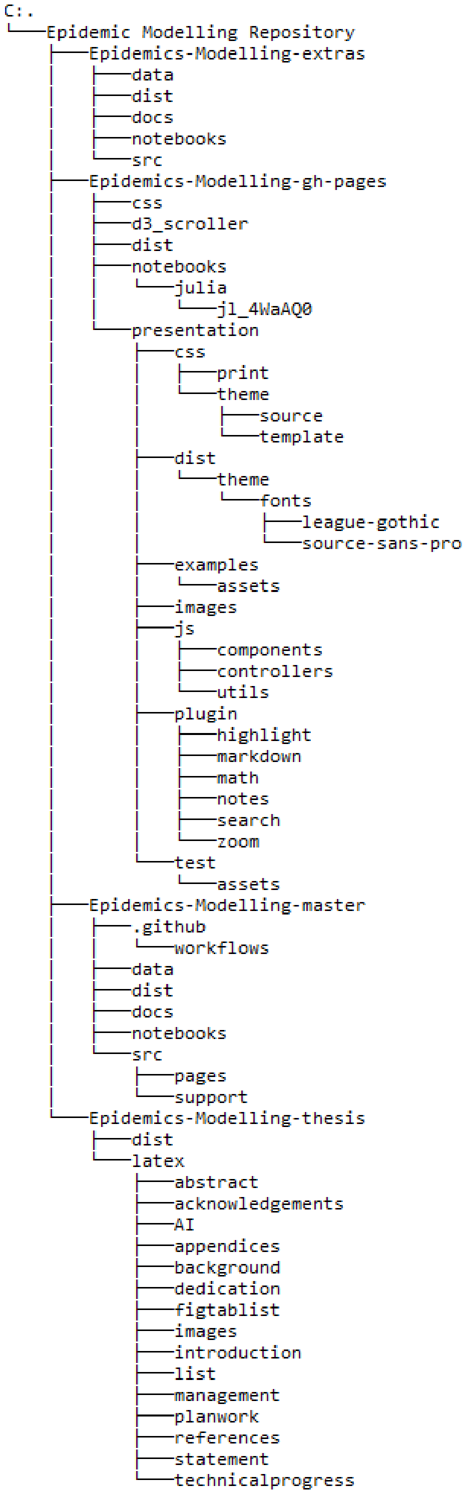
1. Outline today's main Machine Learning limitations and propose possible alternatives.
2. Create an Epidemic Modelling online dashboard which can be used in order to design different possible scenarios and answer causality based questions.
3. Research different ways to overcome Machine Learning limitations and provide example applications using Graphical Methods and Explainable AI.

- Scopes

Successful application of Causality in Machine Learning, could potentially lead to major breakthroughs in the field of Artificial Intelligence and have a huge commercial impact. Causal Reasoning, could in fact allow us to create more explainable models which could then be applied in different sensitive fields such as Medicine or Law. Additionally, it could make possible to apply AI not just in automation based tasks but also in more creative applications (e.g. text/audio automatic generation). Finally, increasing the transparency of the decision making process of the model, would ultimately also make end users more confident/comfortable in using causal based models.

J Design Archive Guide

A tree representation of the Project Design Archive is represented in the figure below. This image has been created through the windows command prompt using the tree command in the designed directory.



```
C:.
└───Epidemic Modelling Repository
    ├───Epidemics-Modelling-extras
    │   ├───data
    │   ├───dist
    │   ├───docs
    │   ├───notebooks
    │   └───src
    ├───Epidemics-Modelling-gh-pages
    │   ├───css
    │   ├───d3_scroller
    │   ├───dist
    │   ├───notebooks
    │   │   └───julia
    │   │       └───jl_4WaAQ0
    │   └───presentation
    │       ├───css
    │       │   ├───print
    │       │   └───theme
    │       │       ├───source
    │       │       └───template
    │       ├───dist
    │       │   └───theme
    │       │       └───fonts
    │       │           ├───league-gothic
    │       │           └───source-sans-pro
    │       ├───examples
    │       │   └───assets
    │       ├───images
    │       ├───js
    │       │   ├───components
    │       │   ├───controllers
    │       │   └───utils
    │       ├───plugin
    │       │   ├───highlight
    │       │   ├───markdown
    │       │   ├───math
    │       │   ├───notes
    │       │   ├───search
    │       │   └───zoom
    │       └───test
    │           └───assets
    ├───Epidemics-Modelling-master
    │   ├───.github
    │   │   └───workflows
    │   ├───data
    │   ├───dist
    │   ├───docs
    │   ├───notebooks
    │   └───src
    │       ├───pages
    │       └───support
    └───Epidemics-Modelling-thesis
        ├───dist
        └───latex
            ├───abstract
            ├───acknowledgements
            ├───AI
            ├───appendices
            ├───background
            ├───dedication
            ├───figtablist
            ├───images
            ├───introduction
            ├───list
            ├───management
            ├───planwork
            ├───references
            ├───statement
            └───technicalprogress
```

Figure J.15: Design Archive

K Word Count

The registered word count for this report, starting from Chapter 1 to Chapter 6, was equal to 14,999 words. This has been measured using Doc Word Counter [53].