COVID-19 Analysis

Gabriel Lapointe

March 29, 2020

Contents

1 Introduction

1.1 Context Summary

The COVID-19 is infecting many hundred of thousands people in the world and is very virulant. Among these people, many recovered from the infection while some of them died. These statistics increase day after day and many scientists are working to understand the virus and find out a vaccin to stop the propagation. While the COVID-19 is propagating around the world, safety measures have been enforced in many countries in order to reduce the propagation between infected people and non infected people. However, they discovered that elder people and people with chronic diseases are more at risk to die.

Here are some safety measures applied in Canada:

- Social distancing rule where people have to be at a distance of at least 2 meters between each other. People that are not respecting that rule could get a fine (from 1500$ to 6000$). It also means that gatherings of people are strongly forbidden and is punishable by fine.

- Non-essential services and stores are closed. Services like hospitals, police, gaz stations, drug stores, grocery stores are considered essential services and stores.
- There is a limitation of people that can enter the stores considered as essential. Also, people have to wash their hands with purell when entering the store. In grocery stores, baskets are all desinfected once a client finished his grocery and leave the store.

1.2 Problem and Questions

The COVID-19 is not fully understood and many thousands of people are getting infected and die day after day. Some safety measures are in place in many countries, but other problems of psychologic nature may arise from these measures. Are those safety measures really as efficient as we thought? We would say yes because it seems to be the common sense for many people to reduce the propagation. However, other factors might be important to consider and might have more impacts on the propagation than we may think.

Since the virus is not fully understood, many questions have to be answered. Here are some of these questions:

1. In which countries the propagation of the virus slowed down the most quickly?
2. Which countries have the greater ratio of deaths over the population and the total infected people?
3. Which countries have the greater ratio of recovery over the population and the total infected people?
4. What is the age category that is more susceptible to die from the COVID-19 after being infected?
5. What is the age category that got mostly infected.
6. Is there a correlation between the sex of a person and the infection rate, death rate and recovery rate?
7. Which chronic diseases are the most vulnerable against the COVID-19?
8. Does the weather have an impact on the COVID-19 propagation?
9. Do the pollution rates have an impact on the COVID-19 propagation?
10. Does the hospitals capacity have an impact on the number of deaths caused by the COVID-19? Which countries are mostly impacted?
11. Is there a correlation between the density of the population and the propagation velocity of the COVID-19?

1.3 Objective

The objective of this analysis is to understand the propagation of the COVID-19 in countries and more precisely in Canada and in the province of Québec. It means to identify factors that appear to impact the propagation velocity of the COVID-19. Understanding these factors will help to understand the propagation of the virus and know how to slow it down quicker.

2 Data Preparation

The objective is to gather necessary data in order to answer our questions. The following datasets are used in our analysis:

- Total population by country;
- Worldwide Covid-19 cases prepared by the John Hopkins University Center for Systems Science and Engineering.

The dataset shared by the John Hopkins University Center for Systems Science and Engineering provides the information on the:

- Country or region
- Province or state for some countries
- Latitude
- Longitude
- Date

- Cummulative number of people confirmed with the COVID-19
- Cummulative number of people that died from the COVID-19
- Cummulative number of people that recovered from the COVID-19

```
Number of countries or regions:  188
```

```
Start date:  2020-01-22
```

```
End date:  2020-06-04
```

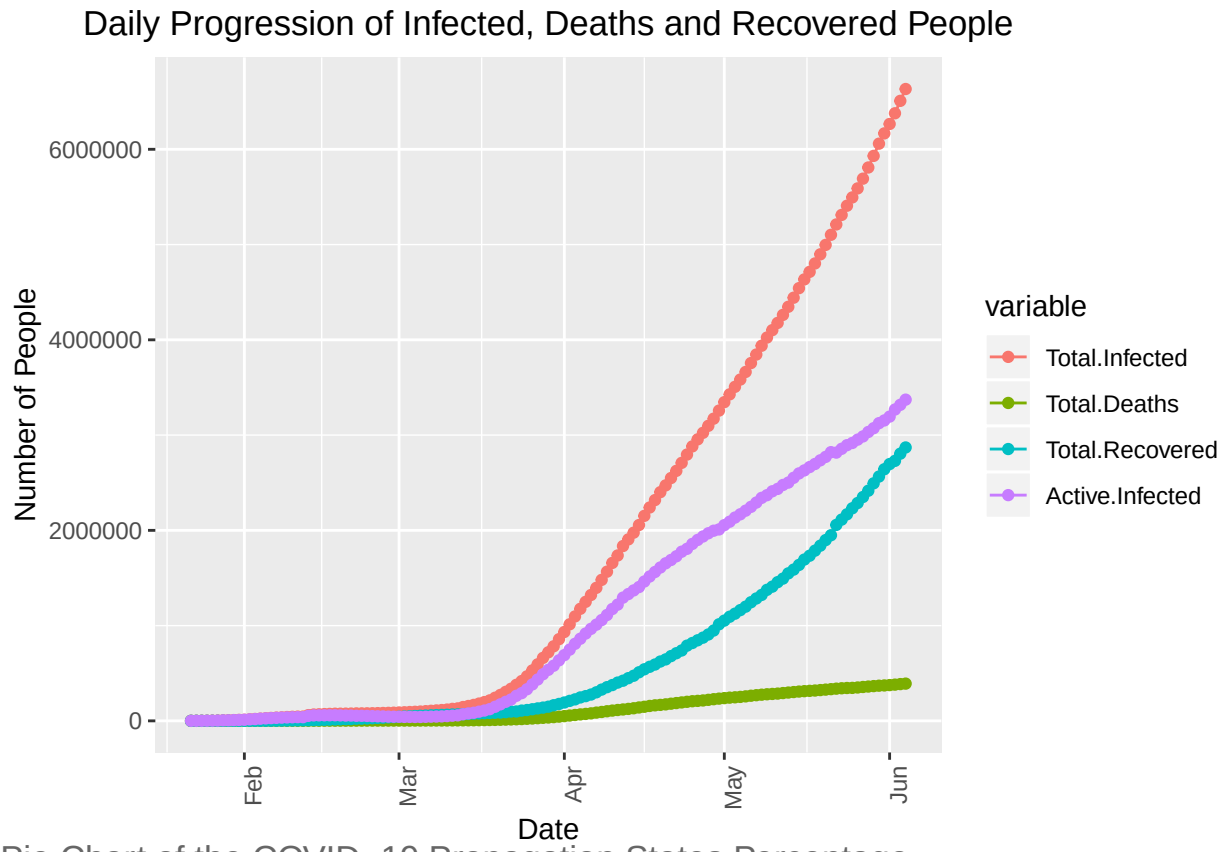
3 Dataset Exploration

3.1 Worldwide Propagation

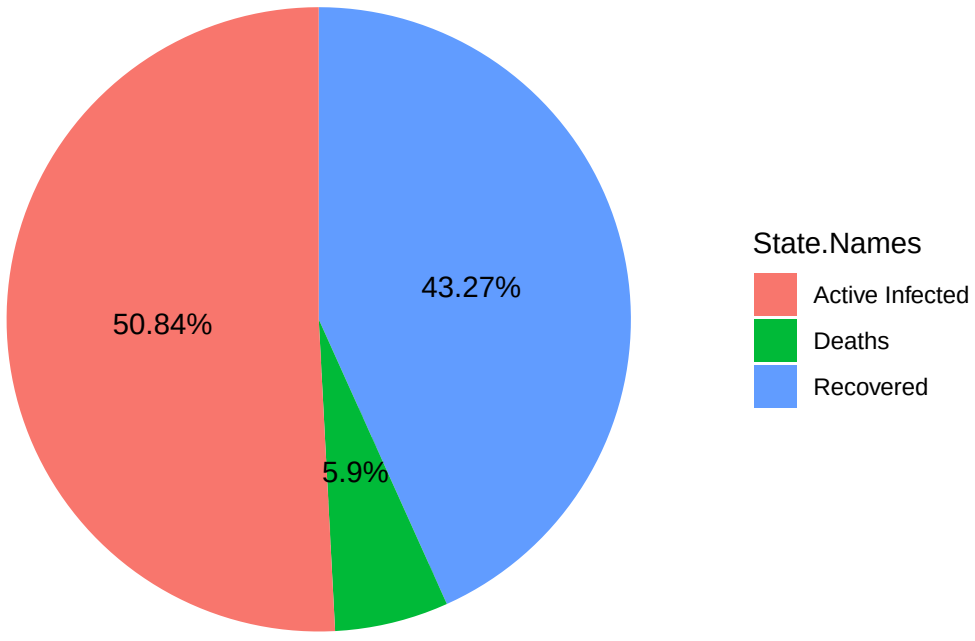
3.1.1 Daily Progression

The objective is to know the distribution of the number of people infected, dead and that recovered from the COVID-19 over days in the world.

Date	Total.Infected	Total.Deaths	Total.Recovered	Active.Infected
2020-05-22	5211156	338233	2056984	2815939
2020-05-23	5311020	342213	2112862	2855945
2020-05-24	5407613	345058	2168563	2893992
2020-05-25	5495061	346231	2231738	2917092
2020-05-26	5589626	350452	2286956	2952218
2020-05-27	5691790	355628	2350088	2986074
2020-05-28	5808946	360308	2415960	3032678
2020-05-29	5930781	364998	2493535	3072248
2020-05-30	6059017	369126	2564693	3125198
2020-05-31	6166946	372035	2641329	3153582
2020-06-01	6265852	375543	2696009	3194300
2020-06-02	6378237	380249	2729527	3268461
2020-06-03	6508635	385947	2804982	3317706
2020-06-04	6632985	391136	2869963	3371886



Pie Chart of the COVID−19 Propagation States Percentage

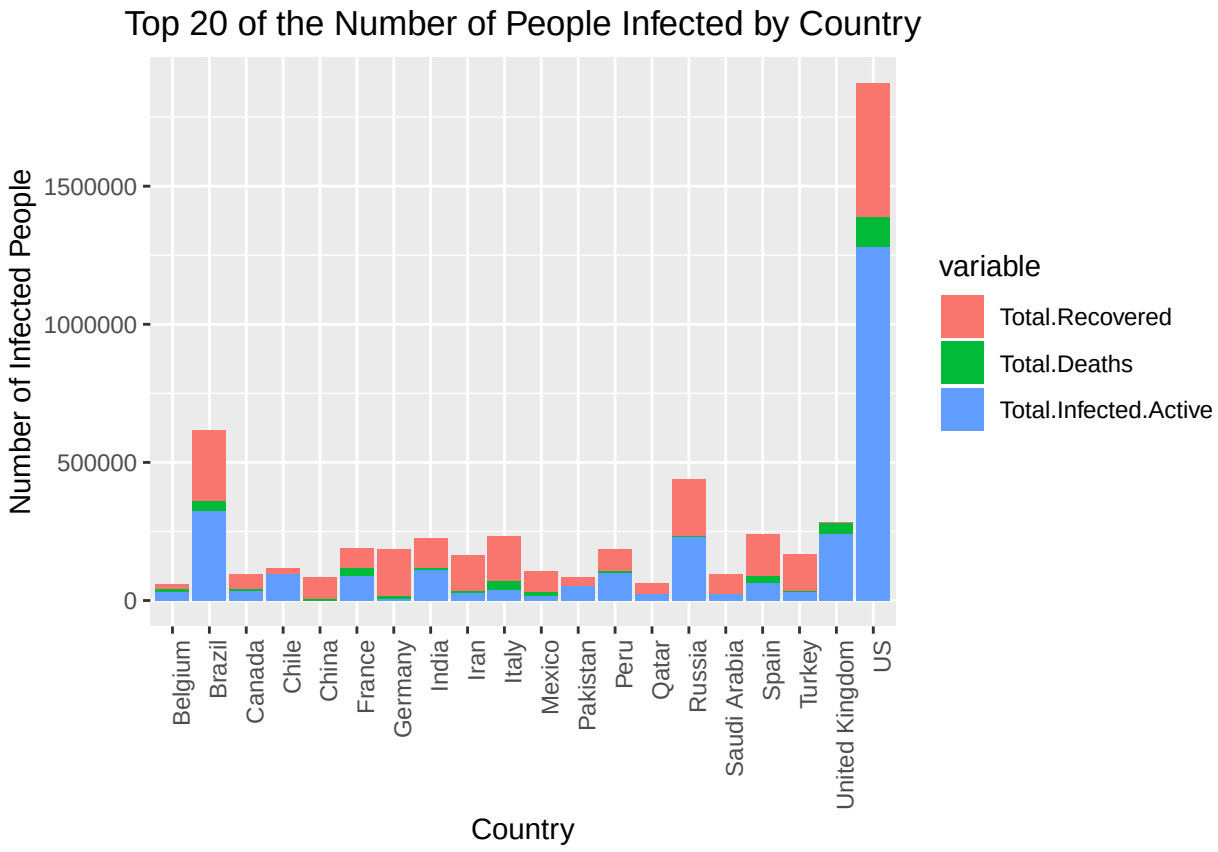


The distribution of the number of people infected, dead or recovered seems to be exponential. This would make sense because if we assume that all people will be infected one day or another, we expect that the curve will describe a sigmoid curve. The reasons behind this is explained in the **Propagation Model section**.

3.1.2 Countries With Highest Number of People Infected

The objective is to show in which countries there are the most infected people until today. Since there are many countries and because the list may be huge enough, we only display the 20 countries with the greatest number of infected people.

Country.Region	Total.Infected	Total.Deaths	Total.Recovered	Total.Infected.Active
US	1872660	108211	485002	1279447
Brazil	614941	34021	254963	325957
Russia	440538	5376	204197	230965
United Kingdom	283079	39987	1219	241873
Spain	240660	27133	150376	63151
Italy	234013	33689	161895	38429
India	226713	6363	108450	111900
France	189569	29068	70094	90407
Germany	184472	8635	167909	7928
Peru	183198	5031	76228	101939
Turkey	167410	4630	131778	31002
Iran	164270	8071	127485	28714
Chile	118292	1356	21305	95631
Mexico	105680	12545	74758	18377
Canada	95269	7717	52184	35368
Saudi Arabia	93157	611	68965	23581
Pakistan	85264	1770	30128	53366
China	84171	4638	79415	118
Qatar	63741	45	39468	24228
Belgium	58767	9548	16048	33171

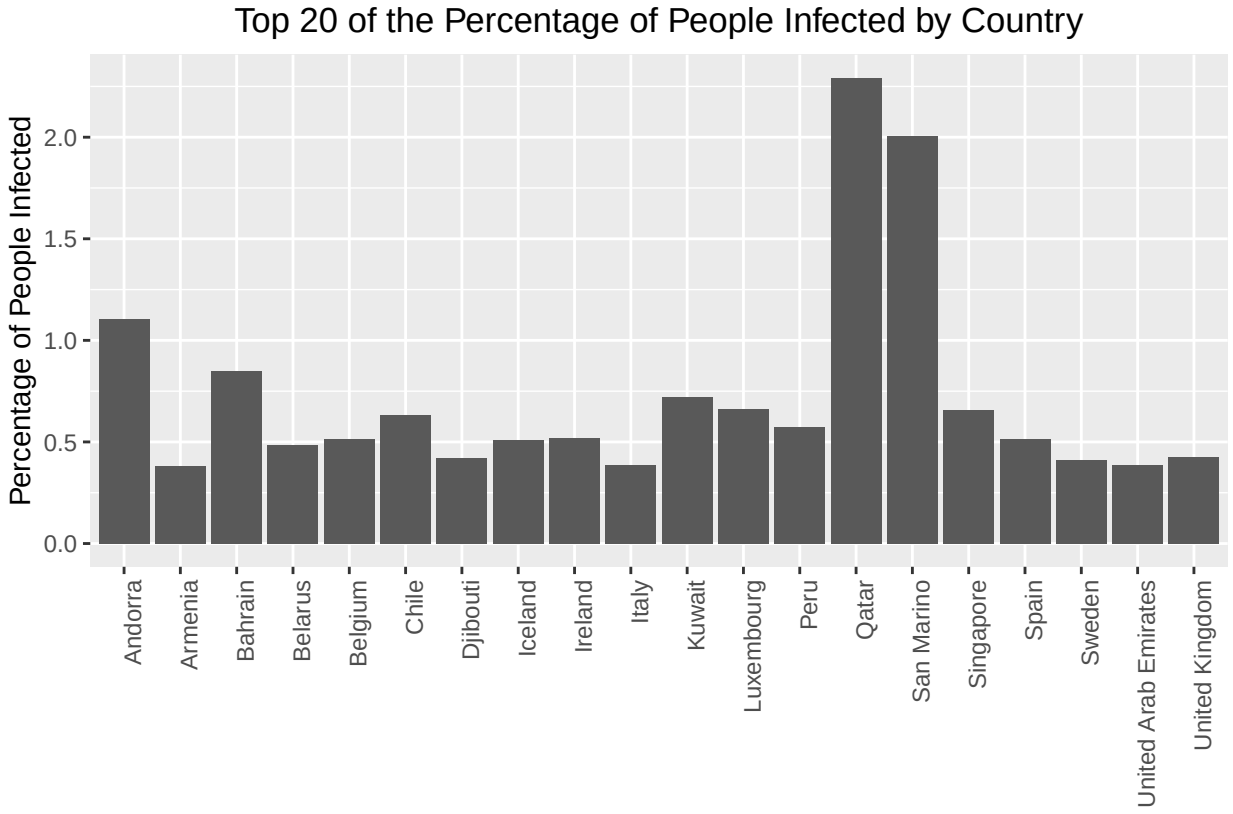


3.1.3 Countries Worst Ratio of Infected and Death People

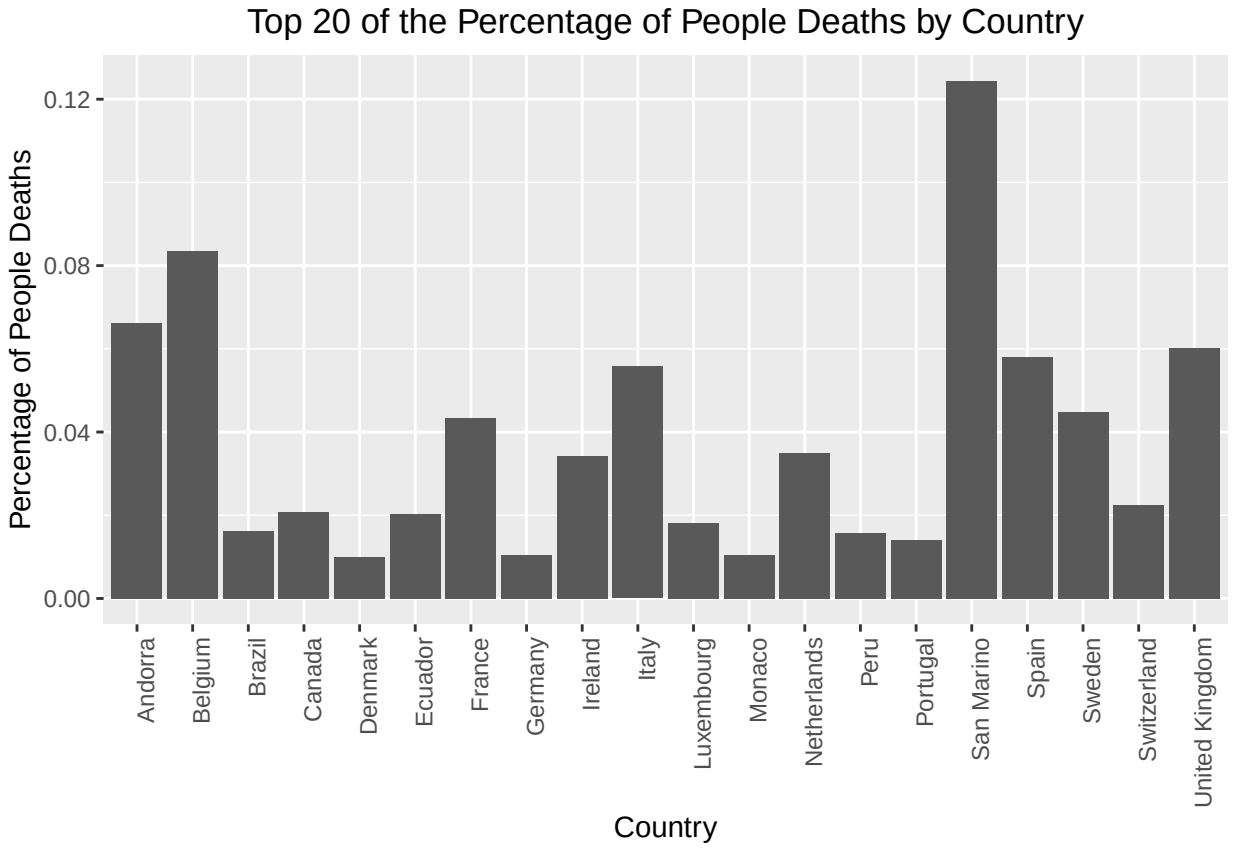
The objective is to know which countries have the worst ratio of deaths and infected people over their population.

Country.Region	Population	Percent.Infected
Qatar	2781677	2.2914594
San Marino	33785	2.0068078
Andorra	77006	1.1064073
Bahrain	1569439	0.8471817
Kuwait	4137309	0.7231995
Luxembourg	607728	0.6626320
Singapore	5638676	0.6547991
Chile	18729160	0.6315927
Peru	31989256	0.5726860
Ireland	4853506	0.5180173
Spain	46723749	0.5150700
Belgium	11422068	0.5145040
Iceland	353574	0.5107842
Belarus	9485386	0.4847562
United Kingdom	66488991	0.4257532
Djibouti	958920	0.4227673
Sweden	10183175	0.4112961
Italy	60431283	0.3872382
United Arab Emirates	9630959	0.3843646
Armenia	2951776	0.3801440

Country.Region	Population	Percent.Infected



Country.Region	Population	Percent.Deaths
San Marino	33785	0.1243155
Belgium	11422068	0.0835926
Andorra	77006	0.0662286
United Kingdom	66488991	0.0601408
Spain	46723749	0.0580711
Italy	60431283	0.0557476
Sweden	10183175	0.0447994
France	66987244	0.0433933
Netherlands	17231017	0.0348732
Ireland	4853506	0.0342845
Switzerland	8516543	0.0225561
Canada	37058856	0.0208236
Ecuador	17084357	0.0204046
Luxembourg	607728	0.0181002
Brazil	209469333	0.0162415
Peru	31989256	0.0157272
Portugal	10281762	0.0141513
Germany	82927922	0.0104127
Monaco	38682	0.0103407
Denmark	5797446	0.0100389

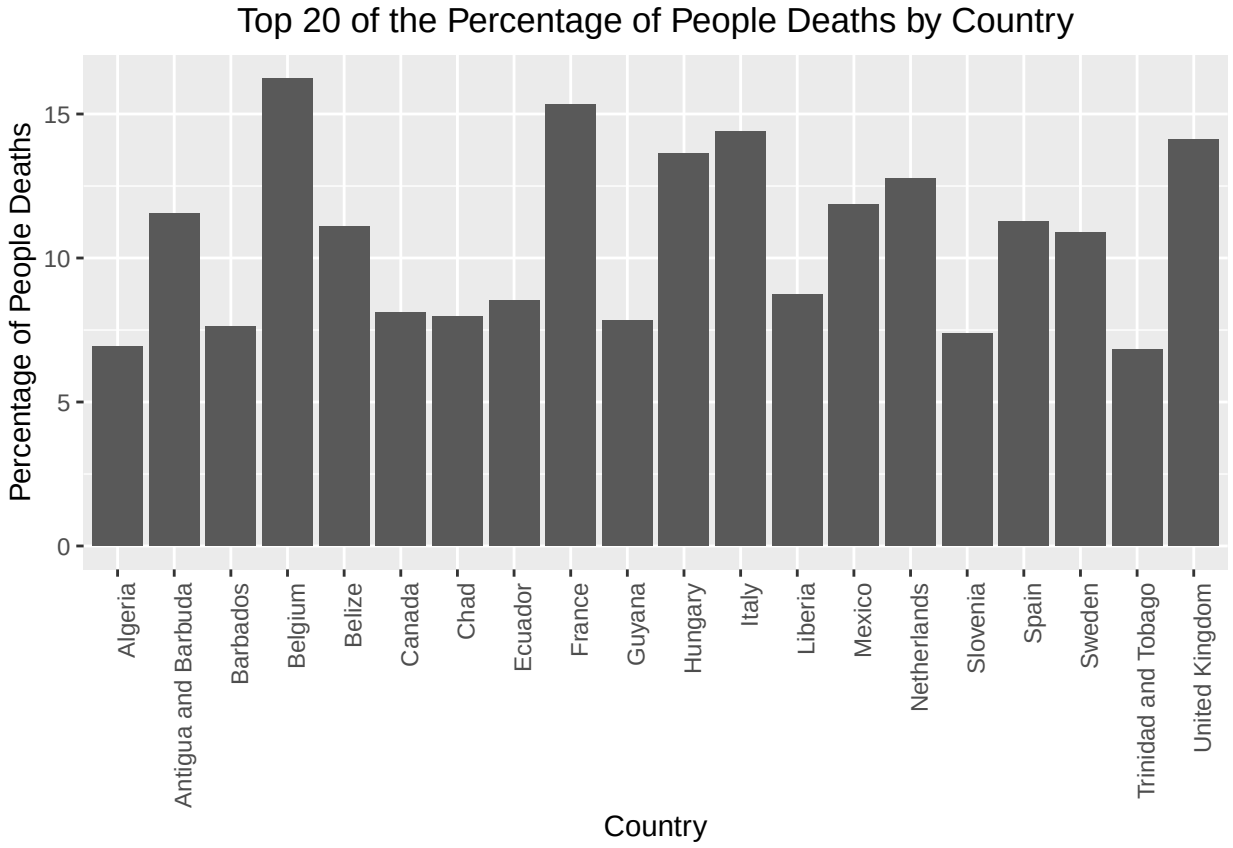


3.1.4 Countries Worst Ratio of Deaths Over Cumulative Infected People

The objective is to know which countries (top 20) have the worst ratio of dead people over the cumulative infected people.

Country.Region	Total.Infected	Total.Deaths	Percent.Deaths
Belgium	58767	9548	16.247214
France	189569	29068	15.333731
Italy	234013	33689	14.396209
United Kingdom	283079	39987	14.125739
Hungary	3954	539	13.631765
Netherlands	47148	6009	12.744973
Mexico	105680	12545	11.870742
Antigua and Barbuda	26	3	11.538462
Spain	240660	27133	11.274412
Belize	18	2	11.111111
Sweden	41883	4562	10.892247
Liberia	321	28	8.722741
Ecuador	40966	3486	8.509496
Canada	95269	7717	8.100221
Chad	828	66	7.971014
Guyana	153	12	7.843137
Barbados	92	7	7.608696
Slovenia	1477	109	7.379824
Algeria	9831	681	6.927067
Trinidad and Tobago	117	8	6.837607

Country.Region	Total.Infected	Total.Deaths	Percent.Deaths

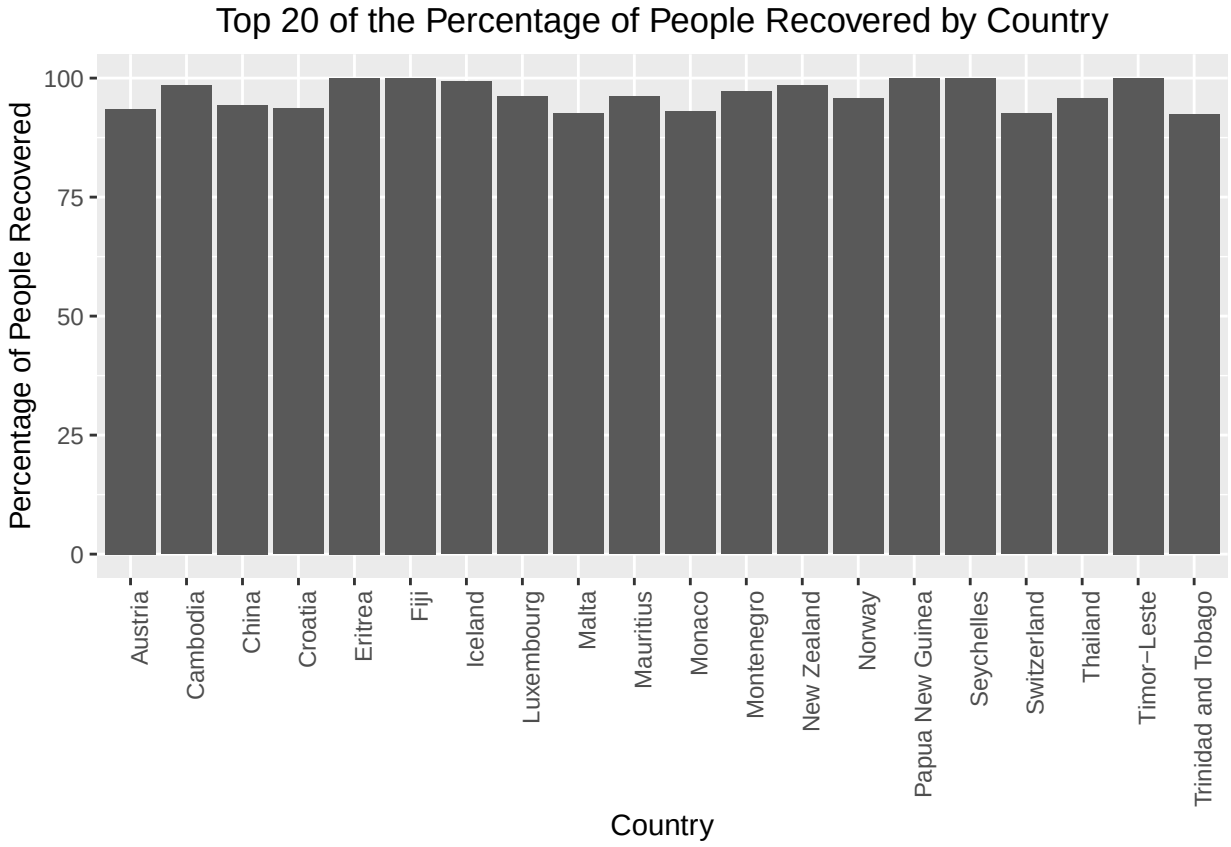


3.1.5 Countries Best Ratio of Recovers Over Cumulative Infected People

The objective is to know which countries (top 20) have the best ratio of recovered people over the cumulative infected people.

Country.Region	Total.Infected	Total.Recovered	Percent.Recovered
Eritrea	39	39	100.00000
Timor-Leste	24	24	100.00000
Fiji	18	18	100.00000
Seychelles	11	11	100.00000
Papua New Guinea	8	8	100.00000
Iceland	1806	1794	99.33555
New Zealand	1504	1481	98.47074
Cambodia	125	123	98.40000
Montenegro	324	315	97.22222
Luxembourg	4027	3874	96.20065
Mauritius	335	322	96.11940
Thailand	3101	2968	95.71106
Norway	8504	8138	95.69614
China	84171	79415	94.34960
Croatia	2247	2105	93.68046
Austria	16805	15717	93.52574

Country.Region	Total.Infected	Total.Recovered	Percent.Recovered
Monaco	99	92	92.92929
Malta	622	576	92.60450
Switzerland	30913	28600	92.51771
Trinidad and Tobago	117	108	92.30769



3.2 Countries With Stable or Decreasing Propagation

The objective is to identify all countries whose propagation is stable or decreasing. In order to find these countries, we have to define what precisely is the meaning of *stable* or *decreasing* propagation.

Let $I(t)$ be the cumulative number of infected people at day $t \in \mathbb{N}$. If we take the difference between $I(t)$ at day t and $t+1$, we get the **propagation velocity**. In mathematical terms, it is expressed as

$$\Delta I(t) = \frac{I(t+1) - I(t)}{(t+1) - t} = I(t+1) - I(t)$$

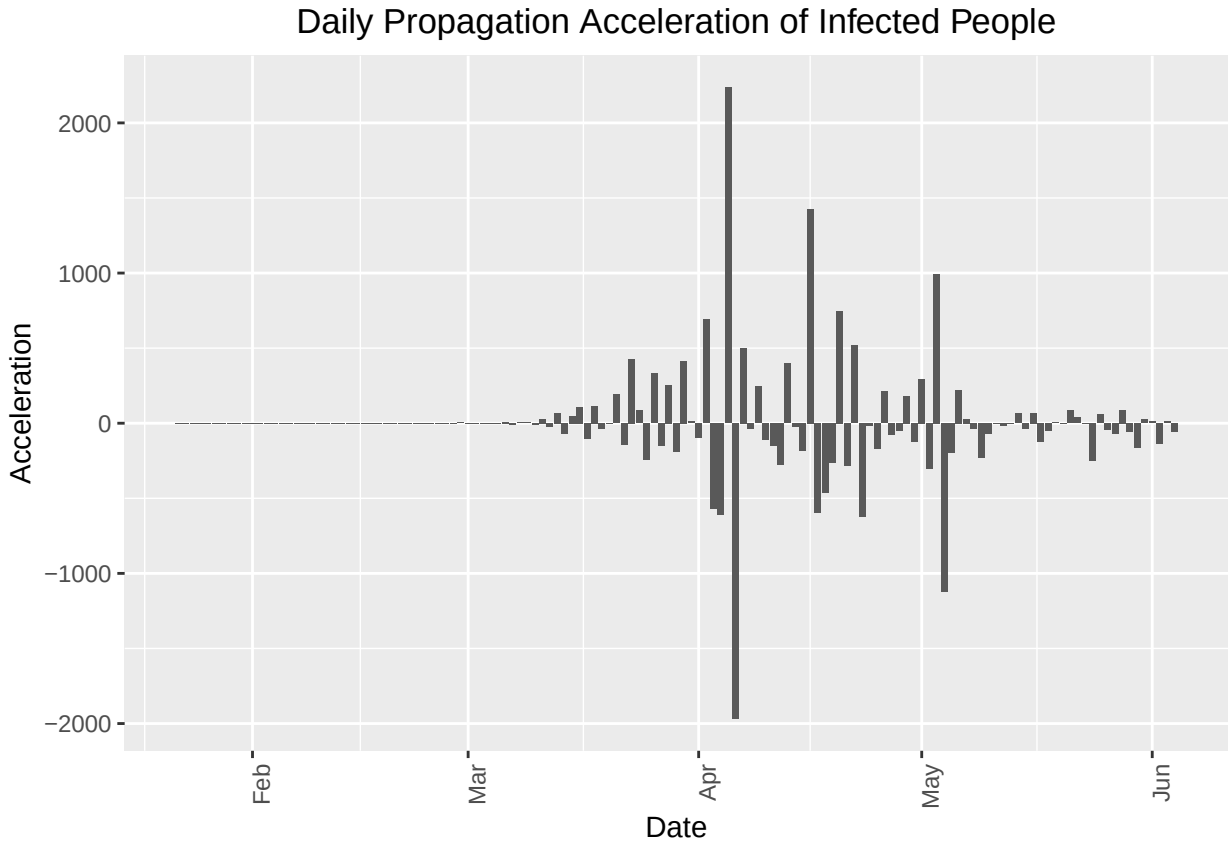
which can be seen as the derivative of $I(t)$ but in a discrete case. The same method is used to define the **propagation acceleration** where we look at the difference between the propagation velocity at day t and $t+1$ expressed as

$$\Delta^2 I(t) = \frac{\Delta I(t+1) - \Delta I(t)}{(t+1) - t} = \Delta I(t+1) - \Delta I(t) = I(t+2) - 2I(t+1) + I(t).$$

The propagation is said **stable** if and only if the acceleration or deceleration is negligible which means that $\Delta^2 I(t) \approx 0$. Therefore, we need to find countries whose propagation acceleration is $\Delta^2 I(t) \lesssim 0$. However,

some of the values of $I(t)$ contained in the dataset may be aberrant values. The first step is to determine and remove those aberrant values. The second step is to determine when the propagation is accelerating, stabilized and decelering.

Date	Total.Infected	infected.delta	infected.acceleration
2020-05-22	83947	1205	38
2020-05-23	85151	1204	-1
2020-05-24	86106	955	-249
2020-05-25	87119	1013	58
2020-05-26	88090	971	-42
2020-05-27	88989	899	-72
2020-05-28	89976	987	88
2020-05-29	90909	933	-54
2020-05-30	91681	772	-161
2020-05-31	92479	798	26
2020-06-01	93288	809	11
2020-06-02	93959	671	-138
2020-06-03	94641	682	11
2020-06-04	95269	628	-54



It comes now the following questions:

1. What are the conditions to determine that a propagation velocity $\Delta I(t)$ is categorized as an aberrant value?
2. What is the range of propagation accelerations considerated as approximative to 0 in the expression $\Delta^2 I(t) \approx 0$ and how to determine it?

3.2.1 Aberrant Propagation Acceleration Detection Model

The objective is to define what is an aberrant value based on our context and remove them in order to get a better estimation on the propagation acceleration. An aberrant value is a value that seems to be out of the "normality" according to our context. For example, if $\Delta^2(t)$ oscillates normally between -250 and 300 and at a day k, $\Delta^2(k) = 1500$, then $\Delta^2(k)$ could be considered as an aberrant value.

Since we want to find when the propagation speed is stable or is decreasing over days, the *value* here is the propagation acceleration. Assuming that the propagation accelerations are independent and identically distributed (i.i.d.) between days, the idea is to assume that the propagation accelaration is normally distributed (equivalently $\Delta^2 I(t) \sim N(\mu, \sigma^2)$). The mean μ is expressed as

$$\mu = \frac{1}{n-2} \sum_{t=1}^{n-2} \Delta^2 I(t)$$

where n is the number of observations in the dataset. Because there are n observations in the dataset, it follows that there are $n-2$ propagation accelerations. The variance is expressed as

$$\sigma^2 = \frac{1}{n-2} \sum_{t=1}^{n-2} (\Delta^2 I(t) - \mu)^2.$$

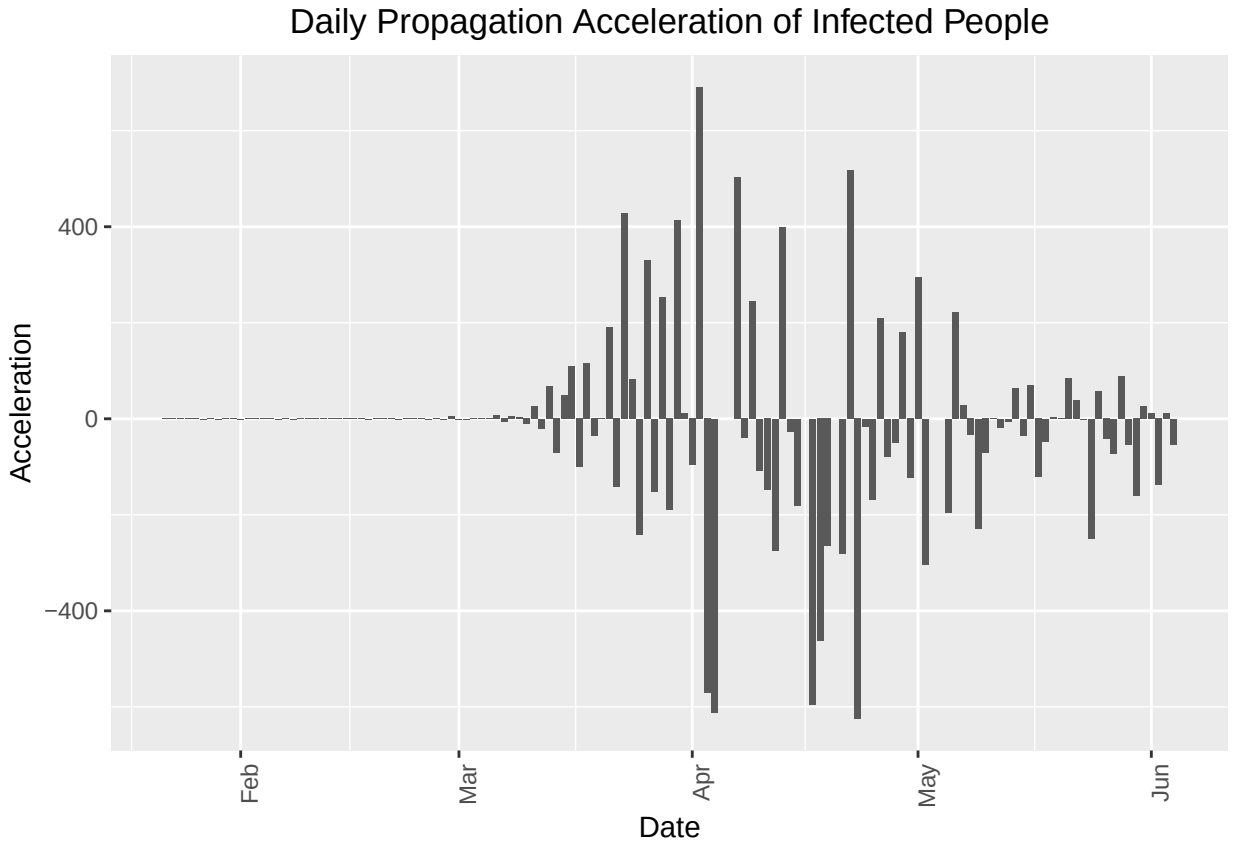
Let's take the data of the Canada as an example. The mean of the propagation accelerations is $\mu = 4.6518519$ and the variance is $\sigma^2 = 136494.8555003$. Therefore, we have $\Delta^2 I(t) \sim N(4.6518519, 136494.8555003)$ where the positive standard deviation is $\sigma = 369.4521018$.

Therefore, the propagation acceleration is **aberrant** if and only if

$$\Delta^2 I(t) \in]-\infty, \mu - k\sigma] \cup [\mu + k\sigma, \infty[.$$

where $k \in \mathbb{R}$ is the parameter to provide. For example with $k = 2$, non-aberrant propagation accelerations in the Canada have to be exclusively between 2 standard deviations -734.2523517 and 743.5560554.

Actually, there are 4.4444444 % of the propagation accelerations that are aberrant.



3.2.2 Propagation Phases Model

We know that a propagation can accelerate, become stable and decelerate on its curve. The objective is to find a model that determines if the propagation is still accelerating, stabilized or decelering based on a set of points chronologically ordered. The idea is to find the global maximum among the non-aberrant propagation accelerations and define what is considered as an acceleration, a stabilization and a deceleration on the curve.

We define a **propagation cycle** when the 3 propagation phases occur in this order: Acceleration, Stabilization and Deceleration. We saw that a propagation is stable if $\Delta^2(t) \approx 0$. This is equivalent to say that a propagation is **stable** if and only if

$$|\Delta^2(t)| < \epsilon$$

where $\epsilon \in \mathbb{N}_*$. The stabilization phase may take many days and for this reason, we have to extend this definition to $|\Delta^2(t+k)| < \epsilon$ where $t+k \leq n$. However, this is not enough because it may happen that, between days t and $t+k$, the propagation is not stable. But overall, the bar chart will show that the propagation is stable. Therefore, we have to consider that a propagation is stable if the average of accelerations within a range of k days is near 0. In other terms, the **propagation is stable** between days t and $t+k$ if and only if

$$|\mu(\Delta^2(t), k)| = \frac{1}{k}\left|\sum_{i=1}^{k} \Delta^2(t+i)\right| < \epsilon.$$

We define the **propagation acceleration** as an acceleration of the number of infected people between day t and day $t+k$. The same method as the stabilization is used:

$$\mu(\Delta^2(t), k) = \frac{1}{k}\sum_{i=1}^{k} \Delta^2(t+i) \geq \epsilon.$$

We define the **propagation deceleration** as a deceleration of the number of infected people between day t and day $t + k$. The same method as the acceleration is used:

$$\mu(\Delta^2(t), k) = \frac{1}{k} \sum_{i=1}^{k} \Delta^2(t + i) \leq -\epsilon.$$

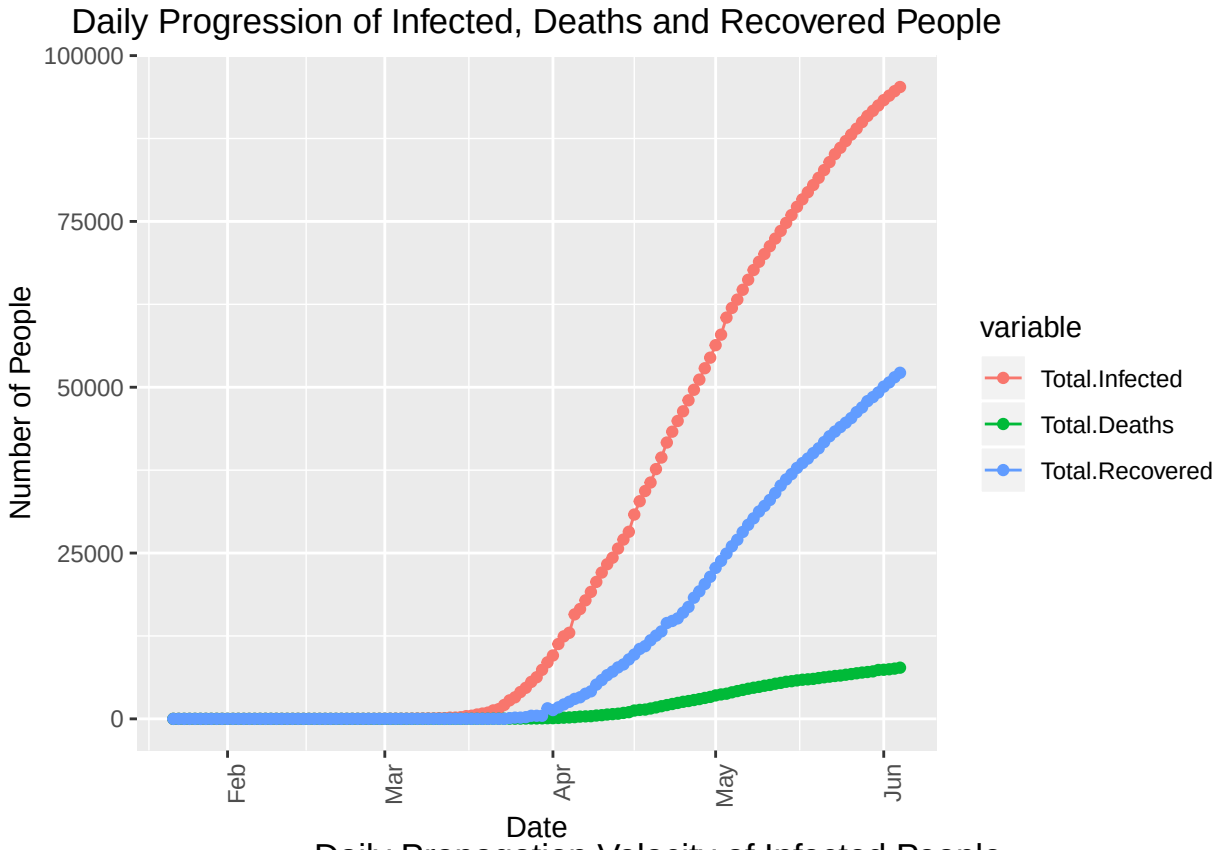
It follows that we have to provide the following parameters:

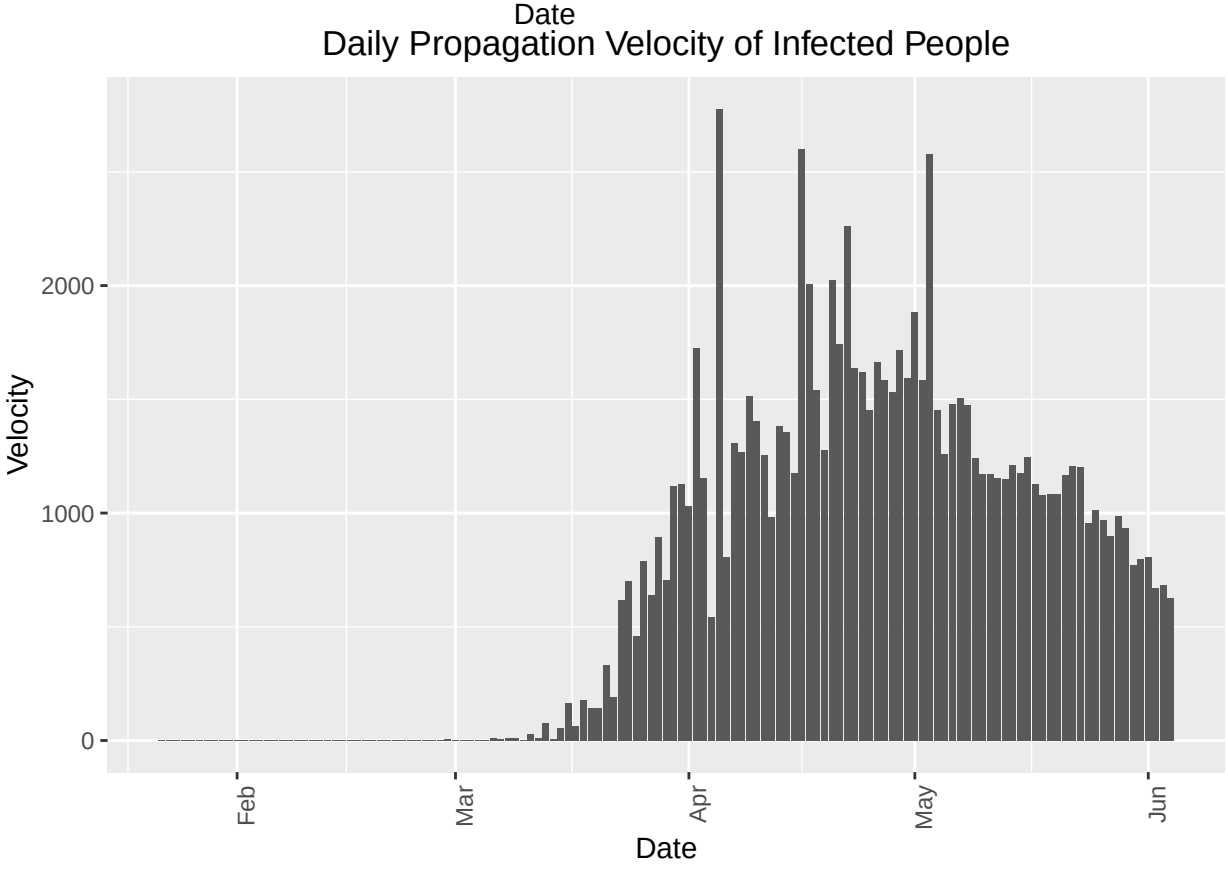
- the stabilization upper bound ϵ;
- the number of days k on which the mean $\mu(\Delta^2(t), k)$ will be calculated.
- The step of days s between the calculation of means.

3.3 Canada Propagation Overview

The objective is to know the overall propagation in Canada.

Date	Total.Infected	Total.Deaths	Total.Recovered	infected.delta	deaths.delta	recovered.delta
2020-05-22	83947	6360	42608	1205	93	877
2020-05-23	85151	6466	43318	1204	106	710
2020-05-24	86106	6534	43998	955	68	680
2020-05-25	87119	6655	44651	1013	121	653
2020-05-26	88090	6753	45352	971	98	701
2020-05-27	88989	6876	46248	899	123	896
2020-05-28	89976	6982	46961	987	106	713
2020-05-29	90909	7063	47905	933	81	944
2020-05-30	91681	7159	48517	772	96	612
2020-05-31	92479	7374	49213	798	215	696
2020-06-01	93288	7404	50091	809	30	878
2020-06-02	93959	7476	50725	671	72	634
2020-06-03	94641	7579	51506	682	103	781
2020-06-04	95269	7717	52184	628	138	678


Daily Progression of Infected, Deaths and Recovered People
100000
75000
50000
25000
0
Number of People
Feb
Mar
Apr
May
Jun
Date
variable
Total.Infected
Total.Deaths
Total.Recovered


Daily Propagation Velocity of Infected People
2000
1000
0
Velocity
Feb
Mar
Apr
May
Jun
Date

4 Propagation Model

The COVID-19 is currently a worldwide pandemic virus which is considered very virulant. It means that it propagate from infected people to non-infected people by direct or indirect contacts. For exemple, if an infected person touches an object, a non-infected person touching the same object after a short time is mostly at risk to be infected.

The objective is to define a model that represents the propagation of the COVID-19 based on assumptions. However, we should take a look on variables that could have an impact on the propagation of the virus. Here is a list of some of those variables:

1. *Population density*: Countries with high population density should be more at risk because the contact between people is much easier hence more at risk to propagate the virus.
2. *Age of people*: Elder people are mostly to die after being infected by the virus because they are more fragile than younger people.
3. *People with chronic diseases*: People with chronic diseases like heart disease, lung disease, kidney disease, cancer, Alzheimer, diabetes, asthma and many others are more at risk to die after being infected.
4. *Births and deaths*: Since the propagation of the virus is a long time period, during this peiod, some people will die from any other causes than the COVID-19 which will decrease the population. In the other case, some women will give birth which will increase the population.
5. *Safety measures*: During the pandemic, many countries adopted safety measures in order to help reducing the contamination between people.
6. *Number of COVID-19 tests*: Since these tests are expensive, they are limited. Coutries that are part of the third world coutries will have less tests than the other countries. Therefore, the number of tests should at least be function of the country.
7. *Number of infected people not tested*: Some people that want to be tested because they might be infected by the COVID-19 are not tested because they are not considered as *essential*. By essential, we mean that the probability they infect others is much greater than other people that do not interact with people in their work (examples of essential people: police officers, nurses, docters). Other people could be infected and prefer to stay at home without asking to be tested.
8. *Infected error factor*: It may happen that a person has been tested positive to the COVID-19 but is not infected at all. Thus, errors when testing could happen.
9. *Recovery error factor*: Errors could happen when people are considered to have recovered but in fact, they did not recover yet. They identified them as recovered too soon.
10. *Death error factor*: It may happen that a person has not died from the COVID-19 but is counted as being dead because of the COVID-19.

We did not consider the immunity against the COVID-19 because we are incertain if there are people that will never be infected by the COVID-19 because they are immune. The same incertainty holds for people recovering from the COVID-19. We do not know if they are immune for the rest of their life or it is like the Influenza; they can be infected after a period of time (virus mutation for example).

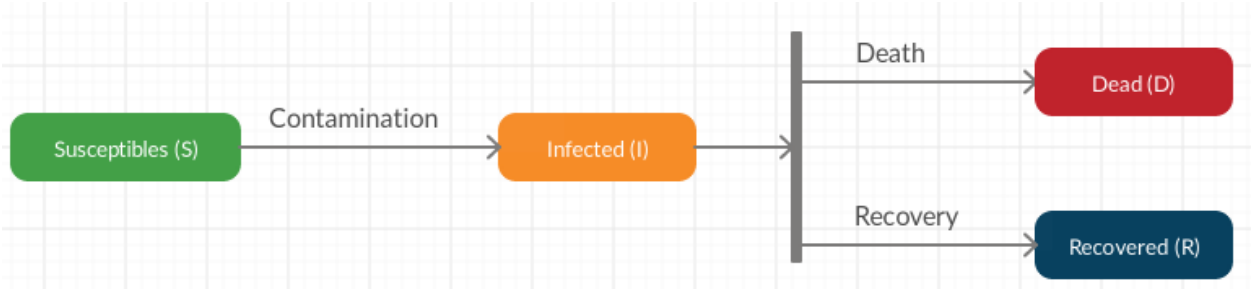
4.1 States and Transitions

A person can be in one of the following states during the propagation:

- Susceptible (noted $S(t)$): People susceptible to be infected by the COVID-19 at day t.
- Infected (noted $I_a(t)$): People tested positive to the COVID-19 at day t but did neither recovered nor died yet.
- Dead (noted $D(t)$): People died from the COVID-19 at day t.
- Recovered (noted $R(t)$): People recovered from the COVID-19 at day t.

Initially, all people in the population are in the *Susceptible* state. Then, it needs at least one person to start the propagation of the virus. This starts at day 1 (2020-01-22) given in our dataset and ends actually on 2020-06-04.

Here is a state diagram describing the interaction between the states:



The arrows between the states represent the transitions between 2 states.

4.2 Assumptions

The following assumptions are made to simplify the model:

1. A person taken randomly in the population has the same probabilities to be infected than any other person taken randomly in the population (it follows a uniform distribution). Therefore, any people have the same probabilities (homogenous population) to be infected without considering their age or if they have a chronical diseases.
2. An infected person could stay infected, recover or die the next day. There is no error made when a person is categorized as infected by the COVID-19. It means that it is not possible for a person to be in the *Infected* state and then transits back to the *Susceptible* state because an error has been made.
3. Every person that recovers from the COVID-19 are immune against it. It means that once a person recovered, that person cannot be infected anymore. Therefore, there is no transition between the *Recovered* state and the *Susceptible* state or *Infected* state.
4. During the propagation of the virus, there are neither births nor deaths (demography is ignored). The initial population is fixed to a constant N.
5. We know that there are more people infected than what the dataset is providing. Many circumstances make that these people have not been tested yet against the COVID-19. For simplicity, we assume that the dataset provides the right values of infected people. It means that we will not add an additional estimator to estimate the number of susceptible people that are in fact infected but not tested yet. However, we know that this assumption is not representative of the reality because the number of tests is limited. Indeed, some people that want to be tested because they have the COVID-19 symptomes are not tested because they are not considered as *essential*. By essential people, we mean that the probability that they infect others is greater than other people that do not interact with people in their work (e.g. police officers, nurses, docters). These tests are also expensive which also explain why they are limited.
6. There are no safety measures taken during the pandemy. It means that there is no quarantine and social distancing between people, and any other safety measures.
7. All the population will have been infected one day. It does **not** take for account that some of the susceptible people could be immune against the virus, could never be infected or have not been tested but got infected by the COVID-19 and recovered.
8. The density of the population is independent of the propagation of the COVID-19.

4.3 SIR Epidemic Model

According to the assumptions, there are 3 transition phases between states on which our model is based:

- From *Susceptible* to *Infected* between days t and $t+1$
- From *Infected* to *Recovered* between days t and $t+1$
- From *Infected* to *Dead* between days t and $t+1$

For exemple, if at day $t=1$ there are 2 infected people and at day $t=2$, there are 5 infected people and 1 dead, then between days $t=1$ and $t=2$, there are 4 people that transited from the *Susceptible* state to the *Infected* state and 1 person transited from the *Infected* state to the *Dead* state.

Per assumption 4, let the initial fixed population noted N be

$$N = S(t) + I_a(t) + R(t) + D(t)$$

where $S(t)$ will decrease while $I_a(t)+R(t)+D(t)$ will increase over days. On the first day of the propagation, there has to have at least one person infected in order to prapagate the virus to susceptible people. Generally, at this initial state, there are neither recovered nor dead people because they have to be infected before. However, it depends on the initial values given in the dataset. It may happen that the data have been gathered later like it is for our dataset.

We introduce $I_c(t)$ the **cumulative number of infected people** at day t because our dataset provides this feature. It means that $N = S(t) + I_c(t)$.

Per assumption 1, each infected person can be in contact with susceptible people and has the probability β to infect each of them. Therefore, each infected person generates $\beta S(t)$ infected people every day. This is true for all infected people ($I_c(t)$), therefore the total number of infected people generated is $\beta S(t) I_c(t)$. The population will then decrease at this rate.

The transition between the suceptible state and the infected state is represented by the equation

$$\frac{\partial S(t)}{\partial t} = -\beta S(t) I_c(t).$$

Per assumption 2, there is a probability γ that infected people will recover (transition from *Infected* to *Recovered* state) or a probability of α that an infected person will die (transition from *Infected* to *Dead* state) fromt day $t-1$ to t. The transitions between the *Infected* state and the *Recovered* state or *Dead* state are given by

$$\frac{\partial R(t)}{\partial t} = \gamma I_c(t)$$
$$\frac{\partial D(t)}{\partial t} = \alpha I_c(t).$$

We know that the number of susceptible people decreses when they become infected. It follows that the number of infected people increases by the same value. Therefore, we have that

$$\frac{\partial I_c(t)}{\partial t} = -\frac{\partial S(t)}{\partial t} = \beta S(t) I_c(t).$$

We did not remove the deaths and recovered people from the infected ones because in our case, the number of infected people is cumulative (I_c). However, to fit with our state diagram, we have to consider the **active** infected people ($I(t)$). This means that these people are infected by the COVID-19 but did neither recovered or died yet. Therefore, we have to subtract the deaths and recovered from the number of infected people:

$$\frac{\partial I_a(t)}{\partial t} = -\frac{\partial S(t)}{\partial t} - \frac{\partial R(t)}{\partial t} - \frac{\partial D(t)}{\partial t} = \beta S(t) I_c(t) - (\gamma + \alpha) I_c(t).$$

We have the following equations that represent our state transition model:

$$\begin{aligned}
\frac{\partial S(t)}{\partial t} &= -\beta S(t) I_c(t) \\
\frac{\partial I_c(t)}{\partial t} &= \beta S(t) I_c(t) \\
\frac{\partial I_a(t)}{\partial t} &= \beta S(t) I_c(t) - (\gamma + \alpha) I_c(t) \\
\frac{\partial R(t)}{\partial t} &= \gamma I_c(t) \\
\frac{\partial D(t)}{\partial t} &= \alpha I_c(t)
\end{aligned}$$

Let $R_0 = \frac{\beta}{\gamma+\alpha}$ be the number of infected people over the revovered and dead ones where $0 < \gamma + \alpha \leq 1$. We expect that $R_0 > 1$ will increase during the rising part of the propagation (contamination phase). Then, we expect that R_0 will decrease over the days and be nearer to 0 because the contamination phase will slow down while the recovery and death phases will increase faster. Finally, all phases will stabilize slowly to $R_0 = 1$.

4.4 Example

The example is based on the data we have for the Canada in this dataset. The first infected person appears to be on 2020-01-26.

Thus, let $I_c(0) = 1$, $R(0) = 0$, $D(0) = 0$ and fix the population to $N = 37500000$ people. It follows that $S(0) = 37499999$. Lets also fix the model parameters to $\alpha = 0.0065$, $\beta = 0.0000000045$ and $\gamma = 0.045$.

Lets see the results of the first iteration:

$$\begin{aligned}
\frac{\partial S(t)}{\partial t} &= -0.0000000045 \times 37499999 \times 1 = -0.1687499955 \\
\frac{\partial I_a(t)}{\partial t} &= 0.0000000045 \times 37499999 \times 1 - 0.045 \times 1 - 0.0065 \times 1 = 0.1172499955 \\
\frac{\partial R(t)}{\partial t} &= 0.045 \times 1 = 0.045 \\
\frac{\partial D(t)}{\partial t} &= 0.0065 \times 1 = 0.0065
\end{aligned}$$

Therefore, we obtain $I(1) = 0.8312500045$, $S(1) = 37499998.83125$, $R(1) = 0.045$ and $D(1) = 0.0065$.

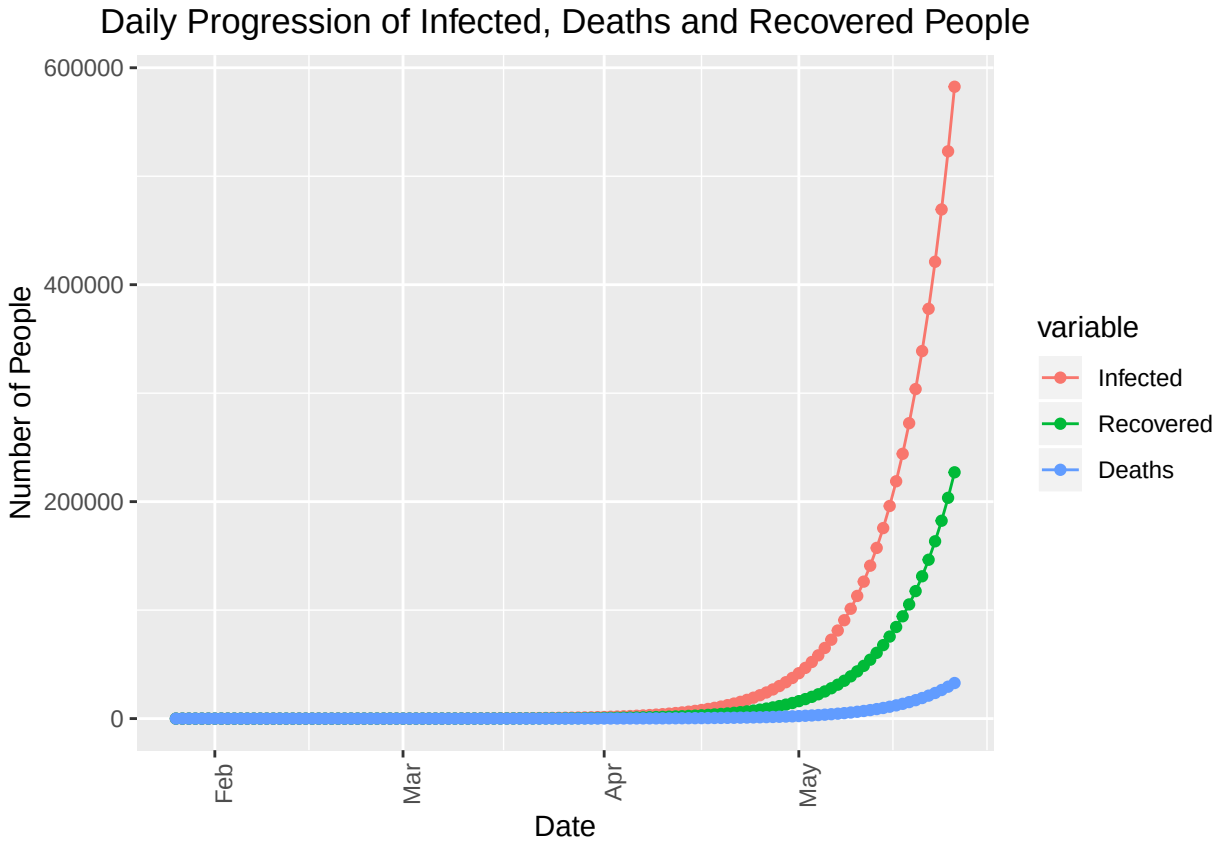
Lets simulate our model for 100 days to see the propression of each state over the days.

Date	Susceptibles	Infected	Recovered	Deaths
2020-01-26	37499999	1	0	0
2020-01-27	37499998	1	0	0
2020-01-28	37499998	1	0	0
2020-01-29	37499998	1	0	0
2020-01-30	37499998	1	0	0
2020-01-31	37499997	1	0	0
2020-02-01	37499997	1	0	0
2020-02-02	37499997	2	0	0
2020-02-03	37499996	2	0	0
2020-02-04	37499996	2	0	0
2020-02-05	37499996	3	0	0
2020-02-06	37499995	3	0	0

Date	Susceptibles	Infected	Recovered	Deaths
2020-02-07	37499994	3	1	0
2020-02-08	37499994	4	1	0
2020-02-09	37499993	4	1	0
2020-02-10	37499992	5	1	0
2020-02-11	37499991	5	1	0
2020-02-12	37499990	6	2	0
2020-02-13	37499989	7	2	0
2020-02-14	37499988	8	2	0
2020-02-15	37499987	9	3	0
2020-02-16	37499985	10	3	0
2020-02-17	37499983	11	4	0
2020-02-18	37499982	12	4	0
2020-02-19	37499979	14	5	0
2020-02-20	37499977	15	5	0
2020-02-21	37499974	17	6	0
2020-02-22	37499971	19	7	1
2020-02-23	37499968	22	8	1
2020-02-24	37499964	24	9	1
2020-02-25	37499960	27	10	1
2020-02-26	37499955	31	11	1
2020-02-27	37499950	34	12	1
2020-02-28	37499944	38	14	2
2020-02-29	37499938	43	16	2
2020-03-01	37499930	48	18	2
2020-03-02	37499922	54	20	2
2020-03-03	37499913	60	22	3
2020-03-04	37499903	67	25	3
2020-03-05	37499891	75	28	4
2020-03-06	37499879	84	31	4
2020-03-07	37499864	94	35	5
2020-03-08	37499848	105	40	5
2020-03-09	37499831	117	44	6
2020-03-10	37499811	131	50	7
2020-03-11	37499789	146	55	8
2020-03-12	37499764	164	62	9
2020-03-13	37499736	183	69	10
2020-03-14	37499705	204	78	11
2020-03-15	37499671	228	87	12
2020-03-16	37499632	255	97	14
2020-03-17	37499589	285	109	15
2020-03-18	37499541	319	122	17
2020-03-19	37499487	356	136	19
2020-03-20	37499427	398	152	22
2020-03-21	37499360	444	170	24
2020-03-22	37499285	497	190	27
2020-03-23	37499201	555	212	30
2020-03-24	37499107	620	237	34
2020-03-25	37499002	693	265	38
2020-03-26	37498885	774	296	42
2020-03-27	37498755	865	331	47
2020-03-28	37498609	966	370	53
2020-03-29	37498445	1080	414	59

Date	Susceptibles	Infected	Recovered	Deaths
2020-03-30	37498263	1206	462	66
2020-03-31	37498060	1348	517	74
2020-04-01	37497832	1506	577	83
2020-04-02	37497578	1682	645	93
2020-04-03	37497294	1880	721	104
2020-04-04	37496977	2100	805	116
2020-04-05	37496622	2346	900	130
2020-04-06	37496226	2621	1005	145
2020-04-07	37495784	2929	1123	162
2020-04-08	37495290	3272	1255	181
2020-04-09	37494737	3656	1403	202
2020-04-10	37494121	4084	1567	226
2020-04-11	37493431	4563	1751	252
2020-04-12	37492661	5098	1956	282
2020-04-13	37491801	5696	2186	315
2020-04-14	37490840	6364	2442	352
2020-04-15	37489766	7110	2728	394
2020-04-16	37488567	7943	3048	440
2020-04-17	37487227	8874	3406	492
2020-04-18	37485730	9914	3805	549
2020-04-19	37484057	11076	4251	614
2020-04-20	37482189	12374	4750	686
2020-04-21	37480102	13824	5307	766
2020-04-22	37477770	15443	5929	856
2020-04-23	37475166	17252	6624	956
2020-04-24	37472256	19273	7400	1068
2020-04-25	37469006	21531	8267	1194
2020-04-26	37465376	24052	9236	1334
2020-04-27	37461321	26869	10319	1490
2020-04-28	37456791	30015	11528	1665
2020-04-29	37451732	33528	12878	1860
2020-04-30	37446081	37452	14387	2078
2020-05-01	37439770	41834	16073	2321
2020-05-02	37432722	46728	17955	2593
2020-05-03	37424851	52193	20058	2897
2020-05-04	37416061	58295	22407	3236
2020-05-05	37406245	65108	25030	3615
2020-05-06	37395286	72714	27960	4038
2020-05-07	37383050	81206	31232	4511
2020-05-08	37369389	90684	34886	5039
2020-05-09	37354139	101264	38967	5628
2020-05-10	37337117	113071	43524	6286
2020-05-11	37318119	126245	48612	7021
2020-05-12	37296919	140944	54293	7842
2020-05-13	37273263	157341	60636	8758
2020-05-14	37246872	175629	67716	9781
2020-05-15	37217435	196022	75619	10922
2020-05-16	37184605	218756	84440	12197
2020-05-17	37148000	244095	94284	13618
2020-05-18	37107196	272328	105269	15205
2020-05-19	37061722	303777	117524	16975
2020-05-20	37011058	338796	131194	18950

Date	Susceptibles	Infected	Recovered	Deaths
2020-05-21	36954632	377775	146439	21152
2020-05-22	36891809	421142	163439	23607
2020-05-23	36821894	469368	182391	26345
2020-05-24	36744121	522969	203512	29396
2020-05-25	36657648	582509	227046	32795



4.5 Model Representation With Markov Chain

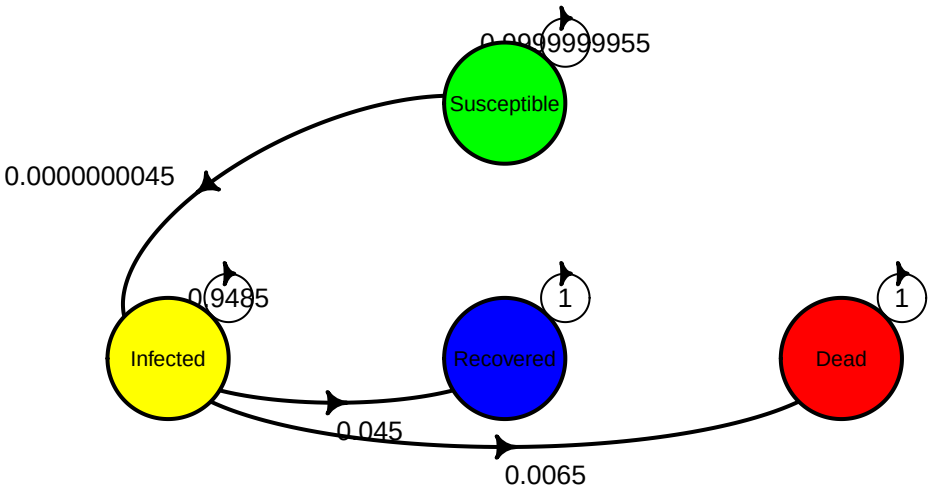
Let $X = (X_t)_{t\geq 0}$ be a sequence of random variables where each random variable is representing a person X at day $t \geq 0$. Let $\mathbb{E} = \{S, I, R, D\}$ the COVID-19 state space. At any day t, a person X_t has to be in a state of $\mathbb{E}$. If $i, j \in \mathbb{E}$, then $\mathbb{P}(X_t \in j | X_{t-1} \in i) = p_{i,j}$. A person at day $t = 0$ as to start in a state of $\mathbb{E}$.

The probability that a susceptible person becomes infected at day t is defined as $\mathbb{P}(X_t \in I | X_{t-1} \in S) = \beta$. The susceptible people that do not transit to the infected state is expressed as $\mathbb{P}(X_t \in S | X_{t-1} \in S) = 1 - \beta$. Per assumption 5, it is impossible to transit from the *Infected* state to the *Susceptible* state. Therefore, we have $\mathbb{P}(X_t \in S | X_{t-1} \in I) = 0$.

The assumption 2 states that an infexcted person can transit to the death state or to the recovery state. It means that $\mathbb{P}(X_t \in R | X_{t-1} \in I) = \gamma$ and $\mathbb{P}(X_t \in D | X_{t-1} \in I) = \alpha$. Since the sets of recovered R and deaths D are mutually disjoint, we have that

$$\mathbb{P}(X_t \in R \cup D | X_{t-1} \in I) = \mathbb{P}(X_t \in R | X_{t-1} \in I) + \mathbb{P}(X_t \in D | X_{t-1} \in I) = \gamma + \alpha$$

We deduce that the remaining infected people that will stay in the *Infected* state (will not transit to another state on the next day) from day $t-1$ to day t is $\mathbb{P}(X_t \in I | X_{t-1} \in I) = 1 - \alpha - \gamma$.



The transition matrix (noted P) is the following considering the order in $\mathbb{E}$ and the transition between day $t-1$ and day t:

$$P = \begin{bmatrix} p_{S,S} & p_{S,I} & p_{S,R} & p_{S,D} \\ p_{I,S} & p_{I,I} & p_{I,R} & p_{I,D} \\ p_{R,S} & p_{R,I} & p_{R,R} & p_{R,D} \\ p_{D,S} & p_{D,I} & p_{D,R} & p_{D,D} \end{bmatrix} = \begin{bmatrix} 1-\beta & \beta & 0 & 0 \\ 0 & 1-\alpha-\gamma & \gamma & \alpha \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Let's assume that there is an infected person, no deaths and no recovery on the first day. Including that there are $N-1$ susceptible people, this means that if $\mathbf{x}^{(0)}$ is the initial vector, we have

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)}P = \begin{bmatrix} N-1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1-\beta & \beta & 0 & 0 \\ 0 & 1-\alpha-\gamma & \gamma & \alpha \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} (N-1)(1-\beta) & (N-1)\beta + (1-\alpha-\gamma) & \gamma & \alpha \end{bmatrix}$$

On the second day, we have

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)}P = \begin{bmatrix} (N-1)(1-\beta)^2 \\ \beta(N-1)((1-\beta)+(1-\alpha-\gamma)) + (1-\alpha-\gamma)^2 \\ \beta(N-1)\gamma + \gamma((1-\alpha-\gamma)+1) \\ \beta(N-1)\alpha + \alpha((1-\alpha-\gamma)+1) \end{bmatrix}^T$$

For n days, we have $\mathbf{x}^{(n)} = \mathbf{x}^{(0)}P^n$. By induction on $n \geq 1$, one shows that

$$\mathbf{x}^{(n)} = \mathbf{x}^{(0)}P^n = \begin{bmatrix} (N-1)(1-\beta)^n \\ \beta(N-1)\sum_{i=0}^{n-1}(1-\beta)^{n-1-i}(1-\alpha-\gamma)^i + (1-\alpha-\gamma)^n \\ \beta(N-1)\gamma\sum_{i=0}^{n-2}(1-\beta)^i(1-\alpha-\gamma)^{n-2-i} + \gamma\sum_{i=0}^{n-1}(1-\alpha-\gamma)^i \\ \beta(N-1)\alpha\sum_{i=0}^{n-2}(1-\beta)^i(1-\alpha-\gamma)^{n-2-i} + \alpha\sum_{i=0}^{n-1}(1-\alpha-\gamma)^i \end{bmatrix}^T$$

Let's see what is the probability law $\mathbf{x}$. We have to find $\mathbf{x} = \lim_{n\to\infty} \mathbf{x}^{(n)}$. We calculate this limit term by term in $\mathbf{x}^{(n)}$. The first term:

$$\lim_{n\to\infty} (N-1)(1-\beta)^n = 0$$

because $0 < \beta \leq 1$ then $0 \leq (1-\beta) < 1$. For the second term, we have that

$$\lim_{n\to\infty} \beta(N-1)\sum_{i=0}^{n-1}(1-\beta)^{n-1-i}(1-\alpha-\gamma)^i+(1-\alpha-\gamma)^n = \beta(N-1)\lim_{n\to\infty}\sum_{i=0}^{n-1}(1-\beta)^{n-1-i}(1-\alpha-\gamma)^i+\lim_{n\to\infty}(1-\alpha-\gamma)^n = 0.$$

Indeed, we get $\lim_{n\to\infty}(1-\alpha-\gamma)^n = 0$ because $0 \leq (1-\alpha-\gamma) < 1$. We also have that $\lim_{n\to\infty}\sum_{i=0}^{n-1}(1-\beta)^{n-1-i}(1-\alpha-\gamma)^i = 0$ because $\lim_{n\to\infty}(1-\beta)^{n-1-i} = 0$ since $0 \leq \beta < 1$. For the third term, we have that

$$\beta(N-1)\gamma\lim_{n\to\infty}\sum_{i=0}^{n-2}(1-\beta)^i(1-\alpha-\gamma)^{n-2-i}+\gamma\lim_{n\to\infty}\sum_{i=0}^{n-1}(1-\alpha-\gamma)^i = \gamma\lim_{n\to\infty}\sum_{i=0}^{n-1}(1-\alpha-\gamma)^i = \gamma\sum_{i=0}^{\infty}(1-\alpha-\gamma)^i$$

because using the same properties as for the second term, the first limit is evaluated to 0. For the second limit, we have a geometric series of ratio $(1-\alpha-\gamma) < 1$. Note that it is impossible to have $(1-\alpha-\gamma) = 1$ because it would mean that $\alpha = \gamma = 0$ which contradicts our assumption stating that an infected person at day t has to transit to either the *Recovered* state or the *Dead* state at day $t+k$ where $k \geq 1$.

Therefore, we have that

$$\gamma\sum_{i=0}^{\infty}(1-\alpha-\gamma)^i = \frac{\gamma}{1-(1-\alpha-\gamma)} = \frac{\gamma}{\alpha+\gamma}.$$

For the fourth term, we use the same logic as the third term

$$\beta(N-1)\alpha\lim_{n\to\infty}\sum_{i=0}^{n-2}(1-\beta)^i(1-\alpha-\gamma)^{n-2-i}+\alpha\lim_{n\to\infty}\sum_{i=0}^{n-1}(1-\alpha-\gamma)^i = \alpha\lim_{n\to\infty}\sum_{i=0}^{n-1}(1-\alpha-\gamma)^i = \frac{\alpha}{\alpha+\gamma}.$$

Therefore, the final result is

$$\mathbf{x} = \begin{bmatrix} 0 & 0 & \frac{\gamma}{\alpha+\gamma} & \frac{\alpha}{\alpha+\gamma} \end{bmatrix}.$$

We expect that with the SIR model based on our assumptions, all susceptible people will be infected and then, all infected people will either recover or die. In other terms, we expect that at the end of the COVID-19 pandemic, a percentage of the population will recover while the rest of the population will die. Therefore, our vector $\mathbf{x}$ makes sense because $\frac{\gamma}{\alpha+\gamma} + \frac{\alpha}{\alpha+\gamma} = 1$.

If we take back our example where $\gamma = 0.045$ and $\alpha = 0.0065$, we obtain

$$\frac{\gamma}{\alpha+\gamma} = \frac{0.045}{0.0515} = 0.8737864078$$

This means that 87.3786407767 % of the population will recover and 12.6213592233 % will die once the pandemic will end.

4.6 SIR Model Solutions

The objective is to solve the first-order differential equations defined for the 3 transitions in order to find $S(t)$, $I_a(t)$ and $R(t)$ in function of t. However, it is useful to mention that the propagation is described basically by the 3 following phases:

1. **Initialization** The propagation will start and slowly progress; only a small number of people will be infected at the beginning of the pandemic.
2. **Acceleration** The propagation will increase quickly because every person infected will infect other people that were not infected making the propression increasing exponentially.
3. **Resolution** The worst case is when the majority of the population is infected. Indeed, there are less non-infected people remaining and the progression will have no choice but to slow down until all people of the population are infected.

These phases justify why the sigmoid function is a good choice. This sigmoid function $I_c(t)$ starts at $t = 0$ and ends until all the population is infected. Therefore, we deduce that the bounds of t and $I_c(t)$ are $t \in [0, \infty[$ and $I_c(t) \in [0, N]$ where $I_c(t)$ will increase over days (cumulative function). It also follows that $\lim_{t\to\infty} I_c(t) = N$.

4.6.1 Cumulative Infected and Susceptible People Models

We know that when susceptible people are getting infected, the same number of people decreases from susceptible to increase in the infected state from day $t-1$ to day t. It means that $I_c(t) = N - S(t)$. Therefore, we have

$$\frac{\partial I_c(t)}{\partial t} = \beta I_c(t)(N - I_c(t)).$$

We have $\frac{\partial I_c(t)}{\partial t} = N\beta I_c(t) - \beta I_c^2(t)$ which is a Bernouilli's differential equation. Dividing the equation by $I_c^2(t)$ gives

$$\frac{\frac{\partial I_c(t)}{\partial t}}{I_c^2(t)} - \frac{N\beta}{I_c(t)} = -\beta.$$

Let $y(t) = -\frac{1}{I_c(t)}$. Then, we have $\frac{\partial y(t)}{\partial t} = \frac{1}{I_c^2(t)}\frac{\partial I_c(t)}{\partial t}$. Replacing in the equation above, we get

$$\frac{\partial y(t)}{\partial t} = -\beta(1 + Ny(t)).$$

Dividing by $1 + Ny(t)$ on both sides gives

$$\frac{\frac{\partial y(t)}{\partial t}}{1 + Ny(t)} = -\beta.$$

Since $\int \frac{1}{1+Ny(t)}dt = \frac{1}{N}\ln(1 + Ny(t))$, we have

$$\frac{1}{N}\frac{\partial(\ln(1 + Ny(t)))}{\partial t} = -\beta.$$

Integrating and multiplying by N on both sides gives

$$\ln(1 + Ny(t)) = -N(\beta t + I_0)$$

where $I_0 \in \mathbb{R}$ is the integration constant. Using the exponential on both sides, it follows that

$$1 + Ny(t) = e^{-N(\beta t + I_0)}$$

which is equivalent to

$$y(t) = \frac{e^{-N(\beta t + I_0)} - 1}{N}.$$

Since $y(t) = -\frac{1}{I_c(t)}$, we have

$$I_c(t) = \frac{N}{1 - e^{-N(\beta t + I_0)}}.$$

Using the same method to solve $\frac{\partial S(t)}{\partial t} = -\beta S(t)(N - S(t))$, we obtain

$$S(t) = \frac{N}{1 + e^{N(\beta t + S_0)}}.$$

4.6.2 Number of Recovered and Deaths Models

We know that $\frac{\partial R(t)}{\partial t} = \gamma I_c(t)$. Replacing $I_c(t)$ by the equation we found for $I_c(t)$ and integrating on both sides, we have to solve

$$R(t) = \int \frac{\gamma N}{1 - e^{-N(\beta t + I_0)}} dt.$$

Equivalently, by multiplying by $1 = \frac{e^{N(\beta t + I_0)}}{e^{N(\beta t + I_0)}}$, we get

$$R(t) = \gamma N \int \frac{e^{N(\beta t + I_0)}}{e^{N(\beta t + I_0)} - 1} dt.$$

Let $u = e^{N(\beta t + I_0)}$. Then, $du = N\beta e^{N(\beta t + I_0)}$ or equivalently $\frac{du}{N\beta} = u$. Replacing in $R(t)$ and solving give

$$\begin{aligned} R(t) &= \gamma N \int \frac{du}{N\beta(u-1)} \\ &= \frac{\gamma}{\beta} \int \frac{du}{u-1} \\ &= \frac{\gamma}{\beta} \ln|u-1| + R_0 \end{aligned}$$

where $R_0 \in \mathbb{R}$ is the integration constant. Replacing u in the equation gives

$$R(t) = \frac{\gamma}{\beta} \ln(e^{N(\beta t + I_0)} - 1) + R_0.$$

The number of deaths at day t is found the same way but γ is replaced by α

$$D(t) = \frac{\alpha}{\beta} \ln(e^{N(\beta t + I_0)} - 1) + D_0.$$

Finally, we know that the active cases of infection exclude the recovered and death people: $I_a(t) = I_c(t) - R(t) - D(t)$.