**Predicting Covid-19 Spread using Kalman Filter**

Team: Prateek Kakkar, Swathi Kolar Ravikumar, Aishwarya Saibewar

Introduction

We all have been affected by the current COVID-19 pandemic. However, the impact of the pandemic and its consequences are felt differently depending on our status as individuals and as members of society. Our different social identities and the social groups we belong to determine our inclusion within society and, by extension, our vulnerability to epidemics.

Problem Statement

Coronavirus (COVID-19) has recently caused a paramount worldwide concern. As the number of coronavirus cases reportedly increases, the spread of COVID-19 is a pressing threat to global health. As of January 30, 2021, more than 74.2 million people across every state in the United States and its four territories had tested positive for COVID-19.

In this project, we'll try to predict the spread of coronavirus for each of the infected regions by fitting a time series analysis and statistical algorithm to produce the best short term and long-term predictions. An adaptive Kalman filter would be a good approach for one-day prediction for each region as it doesn’t require historical values like a batch estimator and hence it is a recursive estimator.

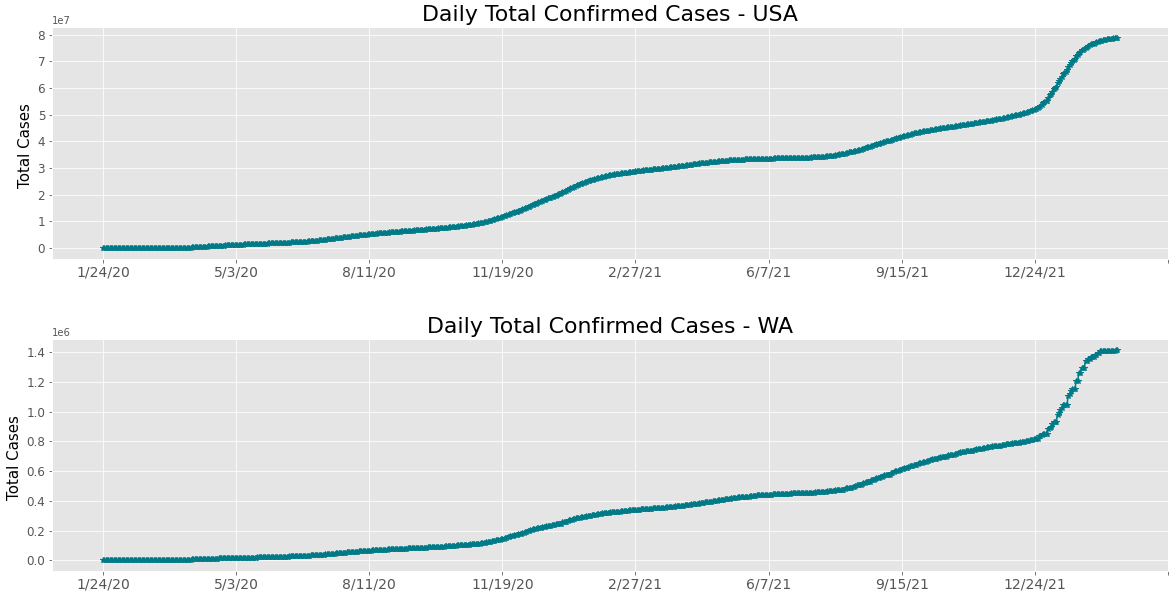
The Kalman filter is an algorithm for efficiently performing exact inference in a linear dynamic system where the state space of the latent variables is discrete and all latent and observed variables have a Gaussian distribution. With the assumption that the time series is linear and Gaussian, we will be applying the Kalman Filter to predict the covid 19 spread across Washington State.

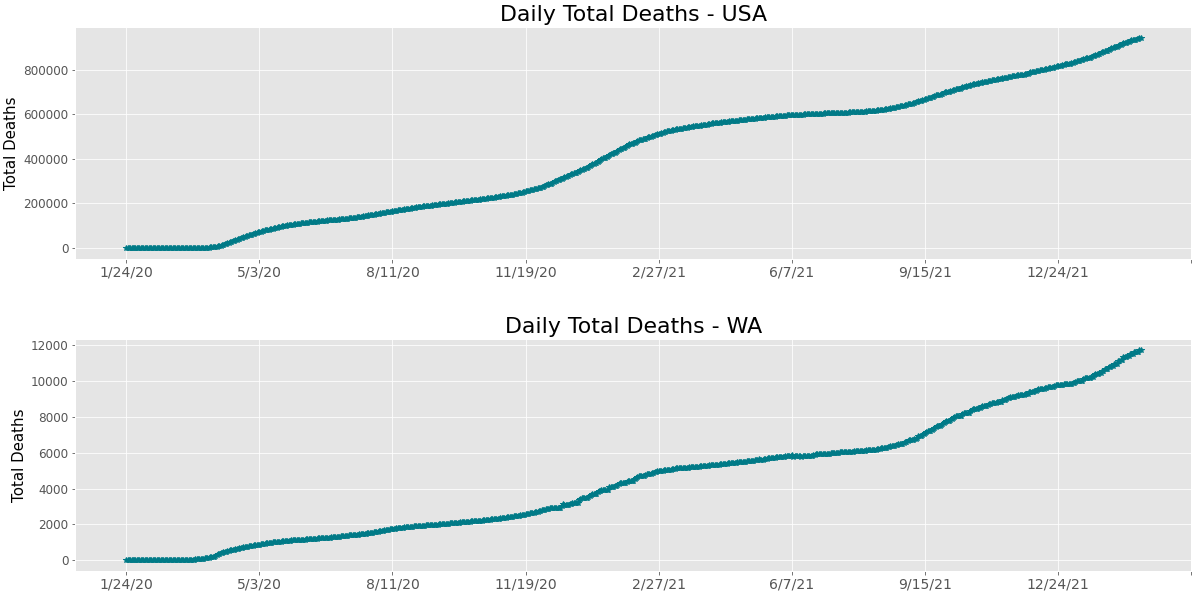
Data source

We’ll use the data set from the COVID-19 [GitHub Data Repository](https://github.com/CSSEGISandData/COVID-19) by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. This repository is updated daily and consists of daily case reports – confirmed, deaths and recovery from each country/region in the US. We’ll also be using the county-level [estimated population data](https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/) from Economic Research Service USDA.

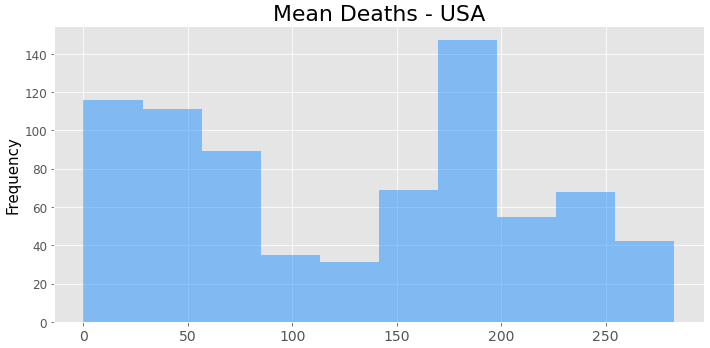
Exploratory Analysis

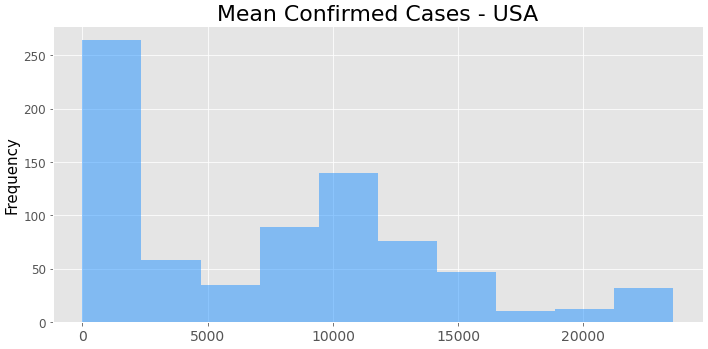
We have access to our data source and were able to perform some exploratory data analysis. To begin with, we started with loading the csv files from the GitHub repo directly to our Colab Notebook. We have two different data sets - Deaths and Confirmed COVID cases for each county in the United States from January 2020 to Current Date as the data is updated every day. Both datasets have been merged with the population variable for each county. Both these datasets consist of 776 (increases by one each day) attributes and 3342 records. The below time series figures the total count of confirmed cases and deaths each day in the US and Washington state.





The below distribution presents the mean Deaths and the Confirmed Covid cases in the USA.

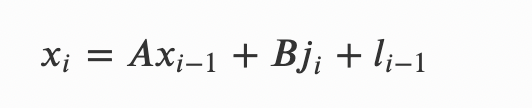
****

****

Model:

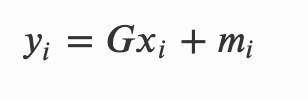
The spread of COVID is due to demographic and environmental factors.  In our model we are including the state variables - population and weather (in consideration), input variables - confirmed cases and deaths and the observation variable as predicted number of COVID cases in a day(s).

Through our research (Singh, 2021), we found out that the Kalman filter tries to predict the state X ∈ Rn of a discrete-time dependent process that is controlled by linear stochastic difference equation:



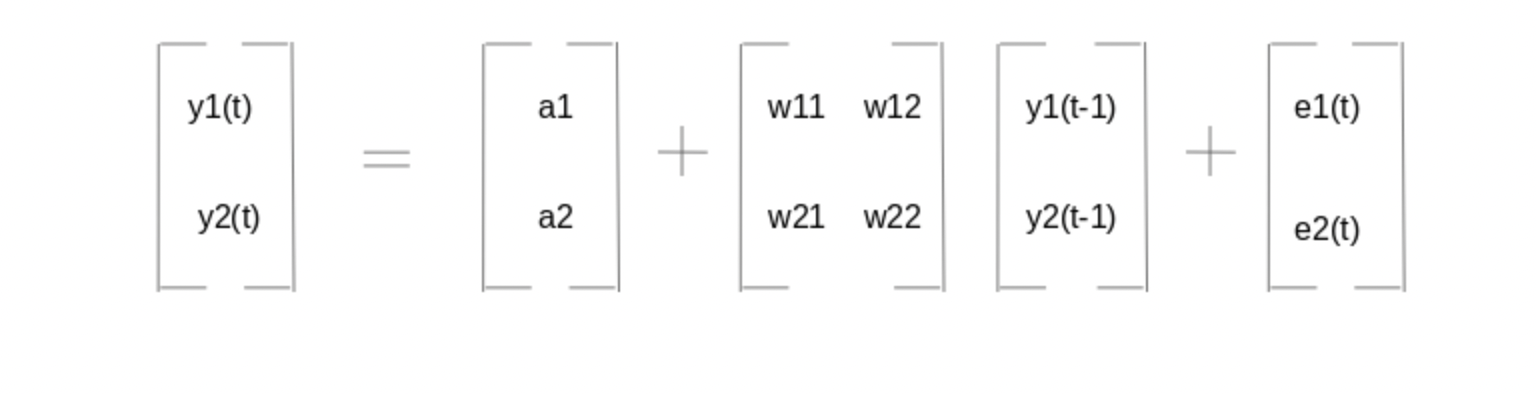
Here, A is a matrix that denotes the relation between states of previous and current time stamp and B is also a matrix which denotes the relationship between input and state function.

With a measurement Y ∈ Rm which is



The random variable li represents the process noise and mi represents the measurement noise. Both these variables are assumed to be independent of each other with normal probability distributions.

The equation for our model may look as below,



Where y1, y2 are the input variable i.e., weather, population, deaths, and total confirmed cases and

w11, w12, w21, w22 are the coefficients of the variables. While a1, a2 are the constant terms

and e1, e2 represents the noise

Plan for Completion

Time-Series Data is taken as input and some data pre-processing was done on the collected data.  A time-series data frame is generated which is ready to be used as input in the Kalman Algorithm, with the state variable population for prediction. By next mid-week we’ll reach to decision whether to have weather forecast included as a state variable in covid19 spread prediction. This decision depends on the factor that the data is easily merged with the existing data and have the daily reading by county and state.

The team has decided to devote 3-4 hours a week per person on this project to take it to the finish line. We are currently at the stage of acquainting ourselves about Kalman Filter and its importance in the prediction models. Once it is done, we’d be ready to fit the model on the data. The team would require meeting up with the professor the following weeks on the approaches decided to be able to reach successful result by the presentation day.

Risk and Challenges:

* *Subjectivity of the topic:* Team requires more understanding and clarity on the novel concept - Kalman Filter.
* *Weather Data:*We are considering the weather as our second state variable where finding data for the past years which matches with our existing data is a challenge for us.
* *Forming the equation:* Would require the guidance on the formation of the equation to describe the true relationship between the elements of weather data and reported covid/deaths cases.
* *Different parameters:* Another challenge would be analyzing the weather data which will affect covid19 spread based on different parameters involving Growth in 1 day, Growth rate in 3 days, Growth in 5 days, Growth in 7 days, Max-Min Temperature in C Humidity.

Group Dynamics

*General Communication:*

We are communicating via WhatsApp group chat for informal project related information, Outlook to communicate with the professor and Canvas portal to share important links and documents for collaboration.

*Sharing data and code:*

We’ll be usingthe Google Colab Notebook to code in collaboration. We’ll be uploading the necessary data files to individual Google Drives so that they can be mounted on the Colab file system.

*Periodic Meeting:*

We are meeting on Zoom on a regular interval to discuss what has been done, what we’ll be doing next and whether there are any roadblocks.