Time Series Analysis and Forecast Predictions for COVID-19 Vaccine Administrations

Sam Stern, Yash Kamat
School of Information
University of Michigan
sternsam@umich.edu, ykamat@umich.edu

Abstract

Our project aims to investigate the similarities and differences in COVID-19 vaccine roll-outs across 32 countries in the European Union and North America through time-series decomposition and time series distance metrics. Additionally, our project analyzes pairwise dynamic time warping costs within the EU and within North America, as well as modeling vaccination forecasts using the FBProphet open source model for the United States and Canada.

1 Introduction

Vaccines provide us with our best chance to fight the COVID-19 pandemic that has affected our world for nearly the past two years. The advent of successful vaccines gave governments around the world an opportunity to shield their population from the SARS-CoV2 virus and slow down rates of community spread. After clinical trials showed promising results, governments were eager to work on a fast and efficient vaccine roll-out. Now, 12 months after the first inoculations began across North America and Europe, we aim to study the trends in vaccine roll-outs in 32 different European and North American countries, using time series methods.

We believe that the findings of our project might be of interest to healthcare organizations and governments who would be interested in understanding not just the different components of the vaccine roll-out in these countries, but also the factors that affected them. It could also help

officials of countries with a low and delayed supply of COVID-19 vaccines, who could use the analysis of the roll-outs in these countries to optimize their own vaccine roll-out process.

In this project, we aim to compare and predict vaccination levels in 30 countries in the European Union (EU) through two subproblems and one time series forecast:

I. We aim to compare vaccination rates for countries in the EU on the basis of countries, age groups, and vaccine manufacturers. This will help us understand how different countries in Europe rolled out different vaccines in different stages and allow us to get an insight into how quickly different vaccines from manufacturers were adopted.

II. We aim to compare our EU data with two countries in North America - Canada and the United States. The motivation behind this intercontinental comparison is to see if similarities exist across countries outside of a region. While we are cognizant that there are factual differences between the two regions (such as population, nation-wide healthcare systems, etc.) and subtle ones as well (like cultural differences) we hope to understand the similarities/differences between these regions.

III. Thirdly, we aim to compare the North American data from part two and make predictions for vaccine administrations for the United States and Canada over the next year. This is important as the rate in which people are getting vaccinated has slowed down, despite countries being below the threshold of herd immunity. At the same time, COVID-19 booster doses have recently seen government approvals in many countries. So, while the rate at which new 1st or 2nd dose vaccine administration has slowed substantially, we will begin to see an uptick in vaccine doses as more people become eligible for their 1 dose booster shot. We will attempt to create a model to predict this expected increase in vaccine administrations over the next year using FBProphet.

2 Related Work

For our study, we are choosing to conduct an analysis looking at total COVID-19 vaccine doses administered. A vaccine dose can be a first dose, second dose, booster dose, or unknown dose. In a paper conducted by Diesel, Jill et al. [1], researchers decided to investigate weekly initiation rates (beginning a vaccine), but also second dose completion percentages for those who started a qualifying 2-dose vaccine. At the time of the paper, they found that 89.3 percent of adults who initiated a 2-dose vaccine series in the United States had received their second dose [1].

In a research paper written by N. Kumar and S. Susan [2], researchers attempted to forecast COVID-19 case data among the top 10 most affected countries at the time of the paper (July 2020). One model they utilized was a FBProphet time series forecasting model and evaluated it based on mean absolute error, root mean square error, root relative squared error, and mean absolute percentage error. In our case, we plan to utilize FBProphet to predict vaccination data, rather than COVID-19 case data, in an effort to make a forecast on our countries of interest. [2]

When it comes to research in COVID-19 vaccination predictions, a paper by P. Cihan [3] has attempted to create a prediction model using an ARIMA model to predict vaccinations across the world's major continents. Using ARIMA, this paper attempted to predict the amount of fully vaccinated people in each continent through May 2021 and beyond (paper was written in January 2021). At the time of this article, vaccine rates among adults were still increasing. However, now more countries in the EU and North America have had trouble convincing a section of their populations to receive a COVID-19 vaccine. As a result, the previous vaccination trends are now vastly different, most likely prolonging a herd immunity situation [3].

3 Data

We are using two sources of data for our project.

First, we have a dataset of vaccination rates for 30 European countries. The dataset includes weekly statistics of the number of vaccines administered for different manufacturers and age groups for each of the aforementioned countries. This data has been sourced from the European Centre for Disease Prevention and Control (ECDC) and accurate up to Week 48 of 2021, or the week that started on November 14th, 2021. The dataset consists of 182,842 rows and 10 columns.

Table 1: *Data on COVID-19 vaccination in the EU/EEA (Relevant Variables) [4]*

Variable	**Definition**
YearWeekISO	Date when the vaccine was received/administered
ReportingCountry	Two letter code for each EU country
FirstDose	Number of first dose vaccine administered to individuals during the reporting week.
SecondDose	Number of second dose vaccine administered to individuals during the reporting week.

DoseAdditional1	Number of additional doses administered to individuals during the reporting week
UnknownDose	Number of doses administered during the reporting week where the type of dose was not specified
TargetGroup	Age group ranges
Vaccine	Name of vaccine manufacturer

Our second dataset contains similar vaccination data for the United States and Canada. The data is sourced from Our World in Data, an organization run jointly by the University of Oxford and Global Change Data Lab. While this data is updated daily, we will be looking at data collected until the week beginning in November 14th, 2021 (similar to the EU Dataset). Once filtered to United States and Canada data, this dataset contains 1,352 rows and 68 columns.

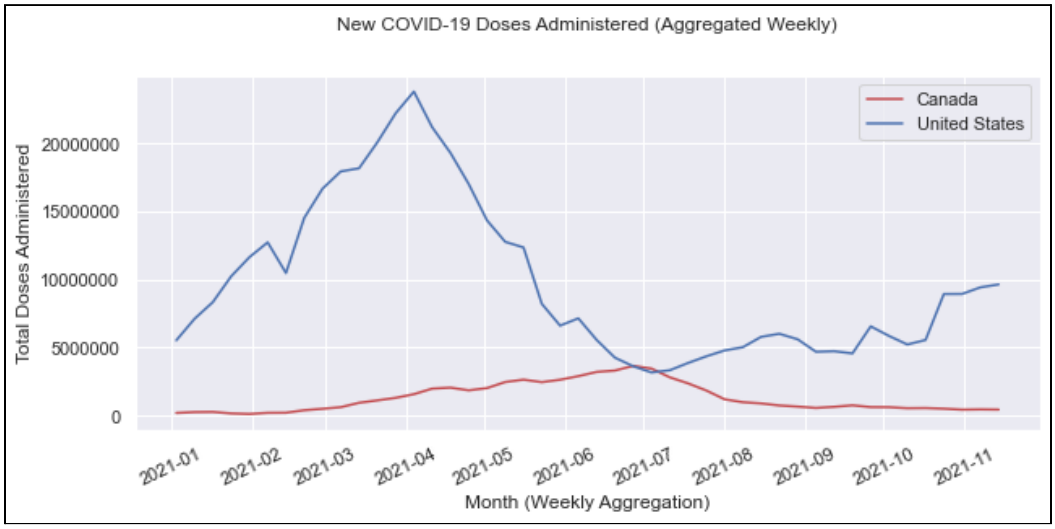

Fig 1: Weekly Aggregated New Vaccine Doses for the United States and Canada

Our datasets vary in the specific nature of the information that they contain. The North American dataset, for example, contains data for each day while the EU dataset contains weekly-aggregated data. Both datasets also contain columns of data that are extraneous to the scope of our project. This means both datasets required a fair amount of cleaning for us to compare and analyze our data in a fair manner.

Cleaning for our EU data frame included creating a custom column, adding up all vaccine dose columns to find ‘total doses’ per week, and eventually pivoting to a time series based dataframe. Pivoting had to be completed to create separate dataframes for a country by country time series, a vaccine manufacturer time series, and an age group time series.

Our North American dataset contained daily vaccine data and already had a total vaccine doses column. The majority of cleaning for this source of data was aggregating the data from daily to weekly to align with the EU data source. Once both datasets had similar week formatting - '2021-W01' to indicate the first week of 2021 and beyond - each week needed to be formatted to the first date of that week in order to function properly with our time series Python packages.

Finally, to analyze an aggregation of all EU countries v. North American countries (United States and Canada), both the EU country time series and the North American country time series needed to be summed to create a merged dataset. This merged dataset resulted in a time series with one column of all EU countries in our analysis and one column with the United States and Canada combined.

First_Day_of_Week	All NA	All EU
2021-10-17	6095496.0	7914985.0
2021-10-24	9425655.0	8515826.0
2021-10-31	9360575.0	8759840.0
2021-11-07	9862748.0	11818307.0
2021-11-14	10064449.0	14595338.0

Table 2: Example of EU v. North American Aggregated Dataset

4 Methodology

Our project focused on 4 different categories of time series analysis.

4.1 Seasonal Decomposition

Our first method was to capture the seasonal decomposition for our 3 European Union groupings: country, age groupings, and vaccine manufacturers, as well as our EU v. NA aggregation. We utilized a custom function which used the *seasonal_decompose* function from the *statsmodels.tsa.seasonal* python package to transfer our time series data frames into seasonal decompositions, focusing on 1-month (4 week) periods as a comparison.

4.2 Time Series Distance Metrics

Our second method was to analyze time series distance metrics: euclidean distance and cosine similarity. We analyzed not only our 3 EU data frames, but also took a look at these distance metrics for the seasonal decomposition data frames created from 4.1. We utilized a custom

function to find pairwise metrics for all of our countries, age groupings, and manufacturers. Therefore, rather than our output dataframe having the same *row x column* shape as our input dataframes, our output data frames had the shape of *N x N* (N = the number of variables per dataframe). For example, as we were analyzing 30 EU countries in our dataset, our distance metric output data frames that were to be analyzed were of shape *30 x 30.* Afterwards, we conducted the same analysis on our aggregated data frame that compares all of our EU countries (N = 30) and all of our North American countries (N = 2).

$$d(x,\ y)\ =\ \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

Fig 2: Euclidean Distance Formula

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Fig 3: Cosine Similarity Formula

4.3 Pairwise Dynamic Time Warping

Our third method was to analyze pairwise dynamic time warping for our EU data frames and EU v. NA aggregation, analyzing our original time series dataframes, as well as our seasonal decomposition dataframes. We utilized a custom function (incorporating the *dtw* function from the *dtaidistance* python package) to find pairwise DTW cost for all of our countries, age groupings, and manufacturers. Therefore, rather than our output dataframe having the same *row x column* shape as our input dataframes, our output data frames had the shape of *N x N* (N = the number of variables per dataframe). Afterwards, we conducted the same analysis on our aggregated data frame that compares all of our EU countries (N = 30) and all of our North American countries (N = 2).

4.4 Time Series Forecast: FBProphet

Our fourth method was to utilize FBProphet, an open source time series model developed by Facebook, to predict daily new vaccine doses administered for our two North American countries: United States and Canada. According to the github.io page for FBProphet, “Prophet is

a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality". [5]

Firstly, for each country we created an in-sample model for each country. This consisted of creating a 75%/25% split of our available data - fitting our model on the first 75% of dates and measuring the model on the remaining 25% of testing data. We then evaluated these in-sample forecasts looking at MAE, MSE, and R^2 values. FBProphet allows users to create custom seasonality cycles in their model, depending on the given problem that is being worked on. In our case, we had to incorporate future booster doses being administered into our model's seasonality. For this, we took into account two variables:

1) The timeline of each national government's vaccine booster guidance, availability, and access
2) All approved booster doses being 1 shot administrations

Secondly, for each country we created an out-sample forecast to predict the next 365 days after our last date in our model (November 24th, 2021). An out-sample forecast, in this instance, is a FBProphet model that we fit on the entirety of our available time series data (January 1st - November 24th 2021), rather than on 75% of our data. This allows us to create a forecast with as much seasonality to train the model as possible. The above 2 variables were taken into account with our custom FBProphet seasonality parameters.

5 Evaluation

5.1 Distance Metrics and Pairwise DTW Cost

Evaluation of our distance metrics and DTW cost focused primarily on the maximum and minimum pairings (and subsequent values) of each metric: euclidean distance, cosine similarity, and pairwise dtw cost. There were some interesting patterns that came to light when analyzing these results.

For example, age groups 0-4 & 5-9 had the smallest euclidean distance values for their original time series data, as well as the seasonal decomposition data (SD). As COVID-19 vaccines were approved for older age groups first and later approved for lower age groups, these age groups having the shortest euclidean distance is not particularly shocking. Additionally, JANNS (Johnson & Johnson) was responsible for the lowest cosine similarity with SPU (Sputnik V) for non-SD data and MOD (Moderna) for SD data. This was not shocking either, as SPU and MOD are both 2 shot vaccine series, while JANNS is only a 1 shot series (not accounting for boosters). You can see the tables for both maximum and minimum values for each metric below:

Table 3: Max Value Pairings for Euclidean Distance, Cosine Similarity, and DTW

Metric	Country Pair	Country Value	Age Group Pair	Age Group Value	Manufacturer Pair	Manufacturer Value
Euclidean	FR, IL	3.290725e07	0-4, 25-49	4.313324e07	COM, SPU	9.847537e07
Euclidean (SD)	PL, FR	830,400.15	15-17, 25-49	765429.78	BECNBG, COM	1.364758e06
Cosine	FR, IT	0.988743	10-14, 15-17	0.954548	MOD, UNK	0.989050
Cosine (SD)	IS, LT	0.997906	50-59, 60-69	0.961854	SPU, UNK	0.968399
DTW	FR, LI	1.082887e15	0-4, 25-49	1.860476e15	COM, SPU	9.697399e15
DTW (SD)	FR, LI	3.886232e11	0-4, 25-49	4.986828e11	COM, SPU	1.346141e12

Table 4: Minimum Value Pairings for Euclidean Distance, Cosine Similarity, and DTW

Metric	Country Pair	Country Value	Age Group Pair	Age Group Value	Manufacturer Pair	Manufacturer Value
Euclidean	LU, IS	73720.27	0-4, 5-9	4.032515e03	BECNBG, SPU	5.403386e05
Euclidean (SD)	MT, LI	4552.56	0-4, 5-9	84.52	JANSS, SPU	100156.08
Cosine	NO, RO	0.565944	5-9, 50-59	0.150395	JANSS, SPU	0.245590
Cosine (SD)	IT, PL	-0.995021	15-17, 60-69	-0.950831	JANSS, MOD	-0.907893
DTW	MT, LU	7.399429e08	0-4, 5-9	1.585219e07	BECNBG, SPU	1.356879e11
DTW (SD)	MT, LI	1.490197e07	0-4, 5-9	6.148452e03	JANSS, MOD	5.718813e09

Table 5: Distance Metrics for EU and North America Aggregations

Metric	EU and North America
Euclidean	1.002093 e08
Euclidean (SD)	2.158441 e06
Cosine	0.782895
Cosine (SD)	-0.27935
DTW	1.536610 e15
DTW (SD)	2.803306 e12

5.2 Time Series Forecasting

5.2.1 In-Sample Forecasting

Using FBProphet to build our in-sample models for Canada and the United States, we evaluated the models using MAE, MSE, and R^2 regression metrics. Both of these models' regression predictions are visualized below. The regression evaluation metrics are in a table below, which make it clear that the United States model was a much better fit, compared to Canada.

Our Canadian model came out with a negative R^2 value, which means that the model created is a worse fit than a horizontal line continuing from our training data. One of the reasons for this may have to do with the custom seasonality that we added to all of our models, accounting for booster shots. As Canada has been behind the United States in COVID-19 vaccine booster approvals, perhaps the training data was not up to date enough to process this seasonality well. In future models, we would want to experiment with the custom seasonality of Canada more. However, our out-sample forecasts are more relevant for future forecasting, which would be of more interest to citizens and government officials.

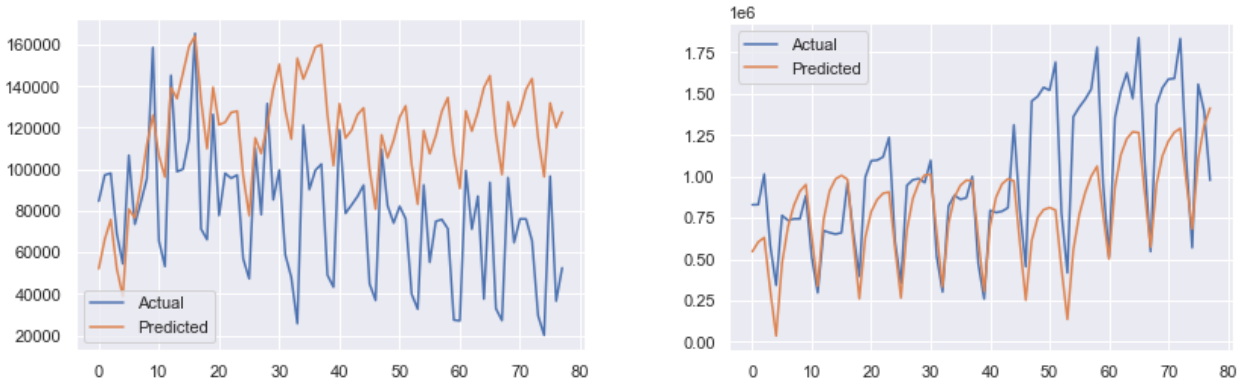

Fig 4: *FBProphet Canada (left) & United States Prediction (right) v. Actual*

Table 6: Evaluation Metrics for In-Sample Forecasting

Metric	Canada	United States
MAE	44485.898	270458.646
MSE	255,31,63,075	12,23,83,61,8854
R^2	-1.685	0.290

5.2.2 Out-Sample Forecasting

For training our out-sample forecasting to predict the next 365 days of vaccine data, we used all the data available to us at the time of this report. This allows us to capture a "full" season of vaccine administrations, as well as the beginning of booster dose administrations in both countries. Below you will see our forecast for total vaccine dose administration over the next year, where the solid blue line indicates our yhat (dose prediction), and the lighter blue shaded area represents the 95% confidence interval of our forecast. The reasoning behind the uptick shown in 2022 is due to recent government guidance for all adults to receive an additional vaccine dose ("booster" shots), with children and teenagers likely to follow in the near future. As all booster doses to date are administered as 1 shot and most original vaccine series being a 2

dose administration, we decided to cut our predictions and confidence interval outputs from the model in half.

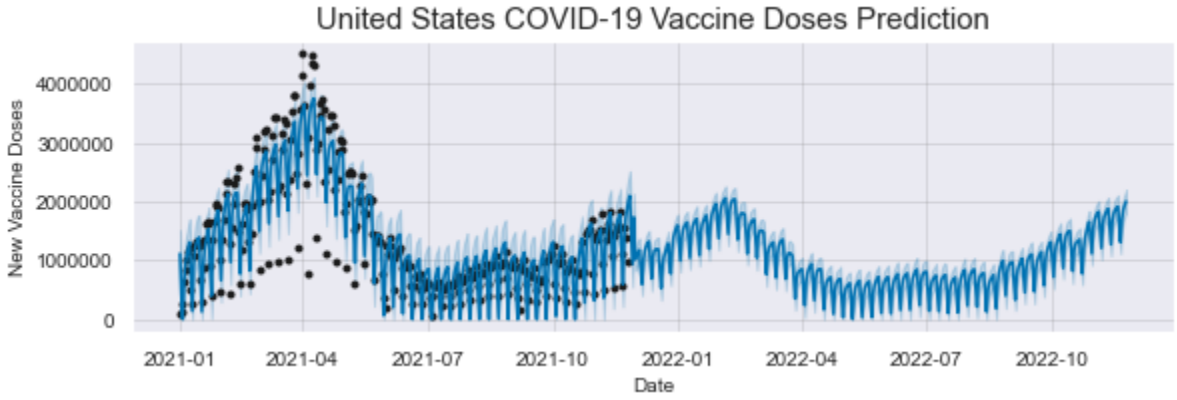

Fig 5: FBProphet United States Predictions

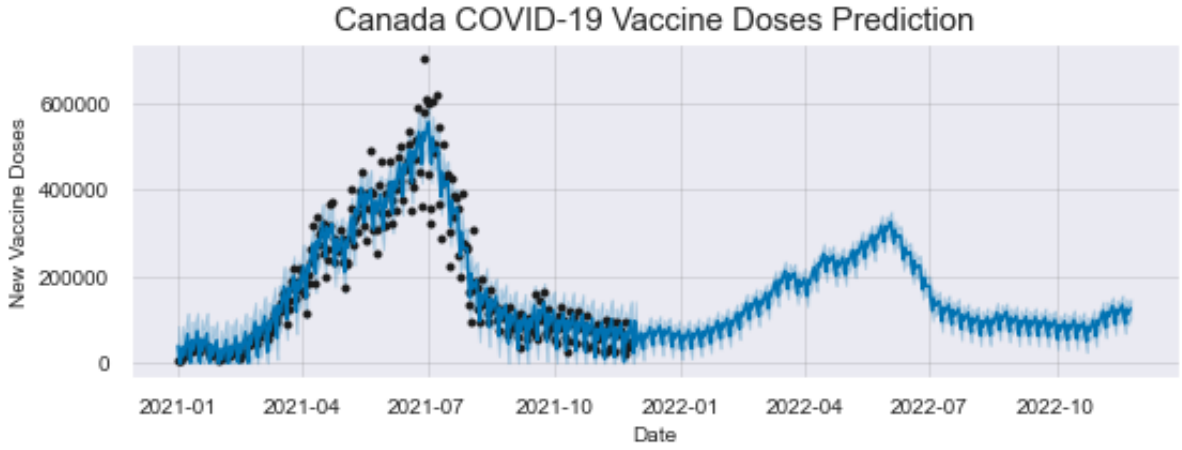

Fig 6: FBProphet Canada Predictions

6 Conclusion

This project attempted to analyze time series trends that we have seen from COVID-19 vaccine data from January 2021 - November 2021, specifically focusing on EU countries and the North American countries of Canada and the United States. On top of the insights that our distance metrics were able to provide, we utilized FBProphet to create forecasts for how vaccine administrations may look throughout 2022 with booster doses recently being approved. Further studies will be able to tweak their time series models to have more of an accurate trend of how booster trends look. As of this report, booster shots have been approved for all adults in the

United States for around a month and only a few weeks in Canada. Having just a few more months of vaccine booster data could provide additional trend information that would be very beneficial to future forecasting. Additionally, our code has been made to be very easily repeatable with other geographical based vaccine dose time series. Repeating similar methods for states within the United States could provide many beneficial insights into the United States' efforts in approaching herd immunity levels (estimated to be roughly 70% - 80%), as well as project booster administration as new COVID variants inevitably emerge**.**

References

[1] N. Kumar and S. Susan, "COVID-19 Pandemic Prediction using Time Series Forecasting Models," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-7, doi: 10.1109/ICCCNT49239.2020.9225319.

[2] Diesel J, Sterrett N, Dasgupta S, et al. COVID-19 Vaccination Coverage Among Adults - United States, December 14, 2020-May 22, 2021. MMWR Morb Mortal Wkly Rep. 2021;70(25):922-927. Published 2021 Jun 25. doi:10.15585/mmwr.mm7025e1

[3] Cihan, P. (2021). Forecasting fully vaccinated people against COVID-19 and examining future vaccination rate for herd immunity in the US, Asia, Europe, Africa, South America, and the World. Applied Soft Computing, 111, 107708. doi:10.1016/j.asoc.2021.107708

[4] Data on COVID-19 vaccination in the EU/EEA. European Centre for Disease Prevention and Control. (2021, December 10). Retrieved December 12, 2021, from https://www.ecdc.europa.eu/en/publications-data/data-covid-19-vaccination-eu-eea

[5] Forecasting at scale. Prophet. (n.d.). Retrieved December 12, 2021, from https://facebook.github.io/prophet/.