Forecasting COVID-19 Case numbers over the next 10 days in Canada: A comparision of ARMA, SIR and SVR models

Mahnoor Babar (261032538)

Abstract

This research project sets out to predict COVID-19 case numbers over the next 10 days. This shall be done using three different types of models: the time series model class Autoregressive Moving Average (ARMA); the Susceptible-Infected-Recovered (SIR) model commonly used in the epidemiology literature; and the Machine Learning model Support Vector Regression (SVR). Each of these models have their advantages and disadvantages which are reflected in the forecasts produced and are discussed after each section. Accurate case predictions are necessary from a policy perspective as they can help guide government decisions regarding lockdowns, making adequate hospitalization arrangements and lately even easing of restrictions. The resulting prediction show no single model has an edge over the rest however, it is evident that sensible predictions can only be made in the short run which is why a 10-day forecasting window is chosen. Moreover, complex models do not necessarily produce better forecasts as is the case. This is especially because COVID-19 has been fast mutating, resulting in newer variants that keep causing spikes in case numbers (when a traditional epidemiology model would predict none). These can be hard for models to capture. Yet, the ARMA models seemed to have best captured the variability in daily case numbers and even it terms of root mean square error (RMSE). This is corroborated by the literature in this topic.

Section 1: Background

COVID-19 emerged in December 2019 in the Chinese city of Wuhan and was declared a pandemic by March 2020 as it spread around the world. There is scientific evidence that the virus spreads via close human contact, the air, and by touching surfaces etcetera. The incubation period of the virus is at least 14 days and it is during this time that it can spread between people. Due to the high transmission of the virus, nearly the entire world has been in lockdown for the better part of two years now which has hampered economic activity and burdened health systems. So far COVID-19 has infected 447 million individuals across the globe, causing 6 million deaths in the wake (Worldometer, 2022).

Given the facts around the disease it is essential to produce good enough forecasts of "confirmed" and "recovered" cases as this will allow policy makers to plan utilization of healthcare resources and even the degree to which lockdowns can be eased. Different types of forecasting models have been used for this purpose.

This paper will run three broad types of models used to predict COVID-19 infections. The Epidemiology specific SIR (Susceptible, Infected, Recovered), the time series models ARMA models, as well as SVR Machine Learning models will be employed to forecast case numbers and their results will be compared based on how well each is able to predict out of sample when compared to the actual data (once it has been collected and incorporated).

Following Section 1: Background, Section 2 will look at brief review of seminal works in the literature using each of these three classes of models. Section 3 will give an overview of the methodology, including data . Section 4 produces the main results of this research. Finally, Section 5 concludes this analysis.

Section 2: Literature Review

The flexible SIR model class has shown decent success in predicting number of infections as shown in the seminal works (Wong et al. 2020; Tang et al. 2020; Khajanchi and Sarkar, 2020; Fenelli and Piazza, 2020; Wu et al. 2020) over the short run however their long-term predictions are questionable.

Maleki et al. (2020), Yue et al. (2020), Papastefanopoulos et al. (2020), Alzahrani et al. (2020) and Dehesh et al. (2020) have produced some of the instrumental works using the ARMA class of models. While Chimmulas and Zhang (2020), Riberio et al. (2020) and Kirbas et al. (2020) use Machine learning and Neural networks techniques for forecasting.

A number of authors also use hybrid models. Malavika et al. (2020), Wang et al. (2020) use the Logistic Growth model together with the SIR and FbProphet models in their seminal work. Syage et al. (2020) use the asymmetric Gaussian with SEIR. Chakrabortry and Ghosh (2020) use combination models WBF-ARIMA to adequately account for periodic residuals. Combining models allows to make up for the disadvantages of any particular model.

Section 2.1: Autoregressive Moving Average (ARMA) models

Kufel (2020, pp.182-187) tests the efficacy of the simple ARIMA model to forecast COVID cases at different stages of the virus' evolution. The ARIMA (1,2,0) was found to have the least AIC and therefore the closest to observed numbers.

Dehesh et al (2020, pp.1-12) also use ARIMA for forecasts. Different models (p,d,q) were achieved for different China, Italy, South Korea and Iran. 17-day forecasts were made (using 41 days of data).

Alzahrani et al (2020, pp. 914-919) used all variants in this class of models to determine the best model namely: AR, MA, ARMA and ARIMA. These were evaluated using RMSE, MAPE, RMSRE and R2 values. ARIMA gives the best results followed by ARMA, then AR and MA models. Their results show cases in Saudi Arabia are expected to increase exponentially over the next four weeks if strict measures are not put in place. This was important because the annual Muslim pilgrimage would have seriously exacerbated the situation if allowed to take place as usual.

Maleki et al.'s (Nov 2020, pp. 1-7) instrumental contribution states the symmetric error distribution assumption in time series but they may not be feasible in the current case so they consider the two-piece scale mixture model (TP-SMN) distributions. The best fitting time series ARMA model is first fit to the data, then then models are used for predictions. Their results show their model performs better than ordinary time series. Aggregate number of confirmed and recovered cases depend highly on the numbers in the previous days. This makes the Autoregressive time series model to be a useful tool to analyze and forecast these numbers. While SIR modeling seems more optimal at a local level, AR can be useful to look at general patterns (Maleki et al. 2020, p.2). As a matter of fact, AR models look at the behavior of the current value based on the past values where error terms are assumed to uncorrelated and identically distributed random variables.

However, in attempting to model this real-world scenario, classical modeling assuming symmetric distributions do not do too well, the methodology by Maleki et al. uses asymmetric or heavy tailed TP-SMN (Two-piece scale mixture normal) distributions. The models then include symmetric Gaussian (regular AR) and asymmetric non-Gaussian autoregressive model. First, the AR models are fitted to historic numbers and then, the selected models are used to predict values. The difference between these two reflects the performance of the models. Here, the TP-SMN distributions improve the simple autoregressive model which is better able to predict the number of cases and recovered. Moreover, credence has been provided to this model by the Akaike and Bayesian Information Criterion when outcomes from this model are compared to the standard Autoregressive model.

Yue et al. (2020) analyze the question in point by using the trend comparison method to forecast the inflection point (IP) and key point (KP) of COVID-19 and then compares these to those of the 2003 SARS virus in China in their seminal work. The two are found to be very different. They use ARMA, ARIMA, Seasonal ARIMA with exogenous regressors as well as exponential smoothing to predict infections. After carrying out the usual procedure outlined above, they use the ARMA model for forecasting. They concluded the pandemic would be curtailed in February 2020.

Seminal work by Papastefanopoulos et al. compares a variety of time series methods to forecast active cases. The use the simple ARIMA method but they also explore more complex methods such as the Holt-Winters additive model (HWAAS), TBAT (characterized by trigonometric seasonal formulation, ARMA errors and a trend component), Facebook's Prophet Model as well as the DeepAR (Autoregressive Recurrent Neural Networks) model for ten countries that had the highest number of cases during the first wave of the virus. The results are compared and evaluated using the RMSE. Their results showed that while there was no single model that performed best in every situation, the simple ARMA model had the second-best performance as compared to the other methods (TBAT performed the best). Even the Deep Learning Neural Network models (mentioned above) were unable to perform as well but this is likely due to limited data availability as opposed to the loads of data they prefer to feed on.

Petropoulous et al. (2020, pp.1-7) use a simple time series model to predict cases and

deaths over the next ten day as well as their confidence interval. They use exponential smoothing models which are capable of capturing and extrapolating point estimates, trends and seasonal patterns. The exponential smoothing model with non-seasonal multiplicative error term ETS(MMN) fits this criterion however, these do provide wider prediction intervals compared to a model with additive error. However, this univariate model does not take into account the primary drivers of the two variables (infections and deaths), such as lockdown policies. The model only extrapolates established patterns in the data, assuming that these patterns are true and will continue to hold in the future.

Section 2.2: Susceptible-Infected-Recovered (SIR)

The seminal work of Khajanchi and Sarker (2020, p.2-17) propose an SIR with six different components susceptible (S), asymptomatic (A), reported symptomatic (I), unreported symptomatic (U), quarantine (Q), and recovered (R) called SAIUQR. Actual COVID data is used for four different states in India.

They attempt to introduce COVID specific situations such as the distinction between symptomatic and asymptomatic as the later are more likely to infect others. In their model they also factor in population dynamics including migration. The transmission rates are a function of the safety protocols followed. Moreover, they decide to distinguish between those who were tested negative but were in their incubation period and those who were actually uninfected. However, at the time there wasn't a lot of evidence for reinfection (and they predict over a short period) so that is not accounted for in the model. The reproduction number R0 is calculated to be greater than 1 for all states which explains the explosion in cases around the time. In order to smooth out the data, three-day moving averages have been used. The parameters are estimated using root mean square error (RMSE). The model works well for three states, with higher RMSE for the fourth. Between the two disease-free and endemic equilibria, the model concludes a disease-free equilibrium but only due to vaccines. Their model is able to close capture the growing trend in cases in each of the provinces.

Wong et al. (2020, pp.2-8) use the Kermack-McKendrik age-of-infection model which is similar to the SIR but allows for delays in transition between moving from one state to the next to be randomly drawn from probability density functions. Non-Markovian models are used so the delay can be represented as an arbitrary time scale as opposed to a single exponential rate of the SIR models.

In this seminal work, they take the susceptible and infected population to be unobserved due to biased testing and so infer the dynamics using lagged and indirect indicators including total number of hospitalized, critical patients and total deaths. Their model is calibrated using the Markov Chain Monte Carlo (MCMC) approach so that the global posterior probability density is produced. This allows explore correlations between parameter, put an interval on the uncertainty and to directly marginalize some modeling uncertainties. The Bayesian framework allows similar treatment of time series data and the model parameters. Prior means are calculated using averaged clinical data, weighted by sample size. The unweighted RMSE is used see the deviation between model and observed values.

Their model curves follow the data trend well. It seems capable to capturing the situation and fitting empirical data. Exponential sensitivity means predictability for epidemic models is over a short term only as when Rt ¿ 1 which means the uncertainty of the environment makes it difficult to make accurate forecasts.

Seminal medical research by Tang et al. (2020, pp.3-9) use the SEIR model by catering

for exposed, pre-symptomatic, hospitalized, and quarantined. They use the next generation matrix to drive the control reproduction number Rc for when social distancing restrictions are in place. They also used the likelihood-based estimations of Rc. These two time series estimations are close enough. Their results show how different social distancing controls can lead to very different transmission and therefore case rates. Case numbers are also close to actual. They also look at the effect of complete travel restrictions. Although the reproduction number R0 they find is higher than that calculated by other researchers at the time, they concluded cases would peak in February 2020.

Wu et al (2020, pp.689-697) also use SEIR to simulate the pandemic across major Chinese cities in their seminal work. The R0 was estimated using MCMC (with Gibbs sampling and non-informative flat prior) and presented results using posterior mean and credible interval. They find an R0 of 2.68, with nearly 76000 infected in Wuhan by late January 2020. Time taken for the epidemic to double was 6.4 days. Assuming same transmissibility across provinces, they infer exponential growth of the pandemic with a lag time of 1-2 weeks behind Wuhan.

For Wuhan, the size of the outbreak is estimated by the number of cases "exported" to other cities as well as beyond China. They also use travel information and model movement of the population to simulate the spread of the disease across China. They find the epidemic would peak in April 2020 if no measures were taken to reduce transmission. They also imply a complete quarantine of Wuhan city would be ineffective as considerable movement into other areas has already taken pace.

Cooper et al. (2020, pp.1-6) modify the SIR such that it takes the susceptible population as a variable (and so does not consider the total population) to account for new infected individuals spread through communities. This allows the model for surges in the number of susceptible and infected populations. They carry out the analysis for China, South Korea, India, Australia, USA and Italy and predict a huge increase in case numbers around the world, especially if movement of individuals continues across borders.

The seminal work of Fanelli and Piazza (2020, p.1-3) uses the simple SIR model, followed by SIRD model to include deaths. Their results show some universality in the virus (for China, Italy and France) as seem by time lag plots of confirmed infected segments in the three countries. According to the SIRD recovery rates were similar for China and Italy but the difference in death and infection rates can be ascribed to personal hygiene practices, social restraint measures, average age and health conditions. Using kinetic parameters fitted to Italy's data, they conclude the effectiveness of a lockdown to be at its peak if measures taken are both quick and drastic.

A case study by Moein et al., (2020, pp.2-5) in Iran explores the predictive power of the SIR. They simulate the model using four different R0 values so as to reflect the different practices (i.e social distancing) at that point in time. Their results reveal that their model worked in the short run and they were correctly able to predict numbers however they were unable to predict new peaks in cases and variants which renders the model useless for long term forecasting.

Section 2.3: Machine Learning (ML) models

Riberio et al. (2020, p.1) evaluate several different models for their accuracy in forecasting short-term COVID case numbers in their seminal work in Brazil. The CUBIST regression, RF, RIDGE and SVR models are adopted as base-learners (models that perform slightly better than random guesses) and the Gaussian process is used as the meta-learners (these

improve performance of a learning algorithm by adjusting it based on experience). The effectiveness of the model is then gauged using improvement index, mean absolute error and symmetric mean absolute percentage error. They find that a majority of the time the support vector regression (SVR) and stacking-ensemble learning achieve improved performance than comparison models. The errors in one, three- and six-days forecasts were 0.87%–3.51%, 1.02%–5.63%, and 0.95%–6.90% respectively. From best to worst accuracy, the models ranked as follows: SVR, stacking-ensemble learning, ARIMA, CUBIST, RIDGE, and RF models. Therefore, stacking models may provide an additional edge over simple models. ARIMA is effective for the very short run but deteriorates in performance over longer horizons. SVR's efficiency is because of its ability to deal with a small dataset whereas the stacking-ensemble learning combines the advantages of multiple models into one by learning the pattern of data and providing very similar forecasts.

Another seminal work by Kirbas et al., (2020, pp.1-8) compare ARIMA, (Non-Linear Autoregression Neural Network) NARNN and LSTM models for COVID case numbers in Denmark, Belgium, Germany, France, United Kingdom, Finland, Switzerland and Turkey. The model performance was measured using Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Root-Mean-Square Error (RMSE), Normalized Root-Mean-Square Error (NRMSE), Mean Absolute Percentage Error (MAPE) and Symmetric Mean Absolute Percentage Error (SMAPE).

In the first stage LSTM was found most accurate. Chimmula and Zhang (2020, pp.1-9) is one of the early studies (and a seminal paper) that uses LSTM as well as Artificial Intelligence and Deep Learning models for an infectious disease. Linear methods used previously for COVID-19 ignored temporal component. Those who have used LSTM models before could not represent spatio-temporal components simultaneously. These issues are addressed by modifying LSTM and established alternate connections between input and output cells to help the networks maintain represent spatio-temporal components as well as transfer past information between them.

To maintain time-frequency components and prevent random noise, wavelet transformation is applied to the data. The results show the LSTM model to have a RMSE of 34.8 with an accuracy of 93.4 for short term predictions in Canada. For the testing dataset, the RMSE is 45.7 with an accuracy of 92.67.

Recurrent LSTM networks are capable of addressing the limitations of time series forecasting techniques by adapting nonlinearities of given COVID-19 dataset. RNNs with LSTM blocks are the efficient algorithms to build a time series sequential model. The reason behind the sigmoidal activation function is because we need to pass only positive values to the next gates for getting a clear output. The 3 gates of LSTM network equations are given as follows:

In the first equation, input gate gives the information that needs to be stored in the cell state. The second churns out information based on the forget gate activation output. The third output gate equation combines the information from the cell state and the output of forget gate at time t to generate the output.

Self-loops are initiated to create a path so that the weights can be shared for long durations. This helps while modelling deep networks where vanishing gradient is a common problem. By adjusting weights as self-looped gates, the time scale to detect the dynamically changing parameters can be adjusted.

Similarly, Arora et al. (2020) also use deep learning models to predict cases in India. They use Recurrent Neural Networks (RNN) based on LSTM variants including Deep LSTM, Convolutional LSTM and Bi-directional LSTM. The results show Bi-directional LSTM give

the best results with daily predictions errors of just 3% while weakly measures have errors of 8%.

Section 3: Methodology

Section 3.1: Data

For the purpose of this project, data was retrieved from 'Our World in Data' which is online scientific publication founded by the Economist Max Roser. It focuses on global issues such as poverty, disease, inequality etcetera and also collects large amounts of data on the subject which is open access and open source which makes it extremely accessible.

During the COVID-19 pandemic, Our World in Data became one of the primary sources of data on the subject. It uses a web scrapper to gather data from key sources such the the World Health Organisation (WHO), John Hopkins University, Financial Times and The New York Times for instance.

This database records case numbers for COVID-19 since 23rd January 2020 upto the present for all countries. The metrics recorded include total cases, new cases, total deaths, new deaths, ICU patients, testing, vaccination as well as certain demographic variables such as population, GDP per capita etcetera.

For the purpose of this study, new Covid-19 daily case numbers were predicted for Canada over the next 10 days. This means, I had 792 observations from the onset of the pandemic till 24th March which is when the data was first retrieved; and 802 observation till 3rd April which is when additional data was retrieved for evaluation of the results. Because of the nature of this predictive exercise the only variables used were date/time, new cases (for Canada) and the population numbers. Our World in Data also calculates real-time Reproduction values (R0) for COVID-19 which was helpful in the SIR modeling.

The idea behind this project is to find the model class that is most suitable to the prediction of COVID-19 daily case numbers. Based on the literature reviewed, three different model classes were chosen: the time series Autoregressive Moving Average (ARMA) models, epidemiology SIR models and Machine Learning Support Vector Regressions (SVR). These models are evaluated using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

Section 3.2: Time Series Models

Time series data consists of trend, seasonality and error. A trend is the repetition of a pattern in regular intervals of time. The nature of time series allows us to see which methods can be used in a particular case. Two broad classifications of the series are stationary and non-stationary. The former indicates a series does not depend on time-factors like seasonality or trend so the mean and variance are constant with respect to time; the latter is when the series has a trend or seasonality that changes with time. The Augmented Dickey Fuller (ADF) test is used to check stationarity of data. It is the unit root test to see the influence of trends on data. Its results are interpreted using the p-value. For $p < 0.05$, the data has a unit root and is therefore stationary (Chuimmula and Zhang 2020).

In general, time series forecasting analyses past observations of a random variable to settle on a model that in able to best capture the fundamental relationship and any patterns. The best model is the one that fits the training set best, and so is used to predict future value. (Papastefanopoulos et al., 2020, p.3) The biggest advantage of using time-series

methods can be reaped when (i) there is no explanatory model that can relate the variable to be predicted with potential explanatory variables and (ii) when there is little information about the fundamental process.

In this study, three different ARMA models were estimated: the AR, MA and the ARMA. Each of these was used to predict COVID-19 numbers over the next 10 days. The results were evaluated using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). This class of models was built using STATA.

Section 3.21: Autoregressive Moving Average (ARMA)

ARMA is a broad class of models commonly used in time series forecasting. Firstly, these models primarily require the history of a particular variable. This information can then be used to make forecast models even when the researcher does not know what the underlying model is. Secondly, because of their easy computation, these models can form baseline models for comparison with other models. Thirdly, these models have strong theoretical underpinnings such as the Wold theorem which proves all covariance stationary processing can be denoted a moving average form (Elliot et al., 2007, p.136). Their simplicity makes them great models for broad range of applications. ARMA models have several variations that make them especially convenient. A drawback of these models in their inability to deal with non-linear relationship which depending on the application may not be ideal for certain real-life problems (Papastefanopoulos et al., 2020, p. 3).

ARMA is appropriate when a system is a function of a series of unobserved shocks (the MA or moving average part) as well as its own behavior. Given a time series of data , the ARMA model is a tool for understanding and predicting future values in this series. The AR part involves regressing the variable on its own lagged values. The MA part involves modeling the error term as a linear combination of error terms occurring contemporaneously and at various times in the past. The model is usually referred to as the ARMA(p,q) model where p is the order of the AR part and q is the order of the MA part. An equation depicting this is as follows.

$y_t = \phi_1 y_{t-1} + ...+ \phi y_{t-1} + \epsilon_t + \theta_1 \epsilon_1 + ...+ \theta_q \epsilon_{t-q}$

According to the Wold representation theorem (Elliot, Timmerman. p.40) the infinite order MA processes can be used to represent the first two moments of any covariance stationary process. In practice, we can use finite order ARMA models. These can be viewed as approximations to such MA(∞) processes. ARMA models are based on linear projections which provide reasonable forecasts of linear processes under MSE loss.

The Autocorrelation function (ACF) and Partial Autocorrelation function (PACF) can be to determine lags (p) and size of the moving average (q). In case there is an obvious correlation beyond k lags in only ACF, the MA should be used; if the former is true just for PACF, AR should be used; if neither shows correlation, ARMA should be our model of choice (Yue et al., 2020, p.8).

In the same vein, an AR model drops all the error term values in the equation above, whereas the MA would only account for error terms and the remaining part of the equation will be dropped. Another model is ARIMA(p,d,q). This model is only employed when the data at hand is non-stationary and so can be differenced (parameter d in ARIMA) to obtain a stationary process that can then be forecasted as before.

In general model selection tries to come up with the best model that usually is based on the researcher's loss function, the data generating procedure and the researcher's information and horizon (Elliot et al., 2007, p.89). Common in-sample model selection methods include

the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) as well as the Lasso with some variations. AIC attempts to minimize distance between density between density of a model under consideration and the actual model (reality). BIC chooses the model with the highest updated (posterior) probability. Both information criteria penalize larger number of parameters. Further, the lasso will allow a parameter in the model if its p-value is below a predetermined criterion (Elliot et al., 2007, p.107).

Section 3.22: Strengths

Further advantages of this model class are as follows: One, interpretation of the results is relatively straight forward as based on the assumptions of the model, any relationship between the dependent and independent variables can be easily understood. This allows for understanding of the existing state of a given variable with its lags and also exogenous variables. Two, the model selection for this class of models is automated by the Akaike and Bayesian Information Criteria (AIC and BIC respectively). This maximizes the accuracy of forecasts.

Section 3.33: Limitations

However, these models are not without any limitations. These fairly simple models have shown reasonable accuracy in prediction of short-term numbers. Their fairly easy implementation and accessibility has meant numerous papers have used these models either on their own, in combination, or in comparison to others. Certain factors that can influence ARMA results is the inadequate data. At present, even more than two years into the pandemic, we have around 800 data points to work with. The high variation in testing capacity as well as different strategies across countries means that data will paint a completely different picture for two distinct points in time (say early days of the pandemic compared to the present time). ARMA models cannot account for the viral characteristics of the spread of COVID as they simply use the trend between past variables and make forecasts based on that. They can also not account for the arrival of multiple variants of the virus and therefore re-surges in case numbers (however this drawback is common to other models as well). They can only model a linear relationship but non-linear relationships cannot be accounted for.

Section 3.3: Epidemiology Models

The study of epidemics, pandemics, and even health conditions that are not caused by disease is called epidemiology. An infection can progress within populations both because of the behavior of the infectious agent and the population itself. Epidemiological models are based on assumptions and statistics which are used to establish a set of parameters that inform how effective intervention will be. This can be used to predict which interventions to implement or avoid, including preparing for mass hospitalisations as well as future growth and spread patterns etcetera.

The complexity of epidemiologic models can vary. They can be simple deterministic models or complex spatially explicit stochastic simulations. The chosen approach depends on several variables including how much is known about the disease's epidemiology, the purpose of the study, and the amount of data available, and its quality.

In this study, two versions of the SIR models are used. For the SI model estimations were made using two different parameter values. For the SIS estimations were made using three different parameter values. The parameters were different activity rate and infection

probabilities resulting in 9 models, 5 of the results have been included in this paper. These models are evaluated using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The were programmed using the epimodels package in R.

Section 3.31: Susceptible-Infected-Recovered (SIR)

This is a very popular model class in this category and has been widely used for COVID-19 predictions. SIR stands for the three states of the model: Susceptible, Infected, and Recovered. At a given time t, each individual in the population is in one of the three states. The proportion in each is given by S(t), I(t) and R(t) where for a population of unit mass:

$$S(t) + I(t) + R(t) = 1$$

An individual can move from one state to another in two ways: (i) people who are infected at present may move to the recovered compartment; (ii) a susceptible individual can become infected by coming into contact with an infected person. It is possible that those who have recovered may still be sick (or dead) but they are not infectious. In addition, they are assumed to be not likely to contract the infection again. The transition rates between these states are given by the parameters a and b. These can be interpreted as the transmission rate and recovery rate respectively. One of the ways in which this model can be simulated is by assuming uniform random distribution of agents in continuous time. This means each individual meets on average a others per unit time and any susceptible coming in contact with infected is assumed infected. So, rate of change for the susceptible population is given by -aSI; and the rate of change for the recovered population is given by bI. The resulting derivatives for the change in the three components per unit time are given below:

$$\frac{dS}{dt} = -\alpha SI$$
$$\frac{dI}{dt} = -\alpha SI - \beta I$$
$$\frac{dR}{dt} = -\beta I$$

Another important parameter is the R0 which is the expected number of individuals an infected person will infect when the entire population is susceptible. This can be a crucial parameter to understanding the kind of a situation we are dealing with. If R0 ¿ 1, the number of infections is greater than the number of infections at this stage and so infections will continue to grow as the peak has not yet been reached. If R0 ¿ 0, the opposite is true and we are past the peak at that point in time (Christopher et al., 2020, pp.80-83). Based on the daily R0 values calculated by 'Our World in Data', R0 for Canada at present is about 0.7.

The basic SIR model can be tweaked in many different ways. For instance, The SEIR model accounts for those who have been infected but are asymptomatic. This takes into account the incubation period of the virus as an individual may not be showing signs of the infection yet or it can also include the scenario where an individual does not show any symptoms at all (even after the incubation period). The differential equations now account for this change. This model assumes exposed individuals do not infect other however if that is not the case it will have to be factored-in. Similarly, the SIR with vital dynamics incorporates birth rates, death rates and even migration. In the same manner, vaccines and other factors can be modeled in. Another interesting variant is the SI model. Depending on the type of infection, there might not be a recovered population. Contracting the virus does not mean one will not be infected again. This means susceptible can become infected and become susceptible again. Viral mutations lead to this kind of a situation.

In this particular study, the SI model has been used. I chose to work on the simple SI model because it seems more plausible in the case of COVID-19 than it's counterparts. This

means the 'recovered' category in the SIR is set to zero. This makes a lot of sense in the current scenario as we know nobody really recovers from COVID in the manner prescribed by the SIR model: their is high re-infection despite the high use of double-dose vaccines and even booster doses more recently. I model this by decreasing the transmission and recovery rates while keeping the recovered at zero. To explain this, recovery rates would increase as infected people would be sick for much shorter period of time compared to pre-vaccine days, mortality rates have reduced drastically (not accounted for separately), the extent or magnitude of the sickness has also decreased substantially as a result. Because of all these changes, transmission rates also lowered since individuals are less sick, less often and as a result transmission of the infection also decreases.

Section 3.31: Strengths

SIR is a simple class of models that has had great success predicting numbers for a range of infectious diseases including the 2003 SARS virus, measles, ebola etcetera. The fact that it was made for epidemiological predictions and is widely used in the field lends it great credence. They allow for considerable flexibility and the various situations can be easily incorporated within the model.

Section 3.32: Limitations

While the SIR is a great model to predict patterns for infectious diseases such as SARS and measles to name a few. However, certain assumptions behind the model do not hold for the case of COVID-19. For one, the infected and recovered individuals are considered 'not susceptible' or immune to the virus in SIR. However, after numerous viral mutations and new, ever rising peaks in COVID case numbers as well as other evidence, we are certain of COVID reinfections. This is greatly affected by the social distancing policies in place. Moreover, the presence of asymptomatic individuals also complicates matters for both prevention and prediction purposes (as more individuals are prone to infection than believed to be). (Moein et al., 2020, pp.2-5).

Moreover, the way the SIR is modeled is based on an exponential increase in the trajectory of the virus (starting from Day 1), reaching a peak value based on the input parameter and then gradually falls back to zero. In general, this makes great sense for most viral trajectories. However, COVID-19 has seen way too many variants in these in the past two years which leads to a re-surge in case numbers instead on the virus dying out. This phenomenon cannot be captured by the SIR. This makes the model useful for short term forecasts only as the long term situation would be pretty difficult to model.

Section 3.4: Machine Learning (ML) Models

Machine Learning (ML) models are a broad class of models that learn patterns or trends from a given dataset, also called training set and uses it to make predictions over the so called testing set. These have wide applications and are particularly popular at present. Their popularity is probably because of their 'hands-off' application as they do a lot of the work for the researcher.

There are three types of ML models: binary classifications, multiclass classifications and regressions. In this analysis one of the regression methods is used since we are trying to forecast COVID case numbers and so it is best suited to our needs. Here, the Support Vector Regression (SVR) is used to make predictions. The optimal model is found by tuning 1100

models and then choosing the best model based of Root Mean Square Error (RMSE). A k-fold cross validation is done to provide evaluate the optimal model from these 1100, such that the RMSE is the lowest. The program is modeled in R using the e1071 package.

Section 3.41: Support Vector Regression

Support Vector Regression (SVR) seeks to determine support vectors points close to a hyperplane that maximizes the margin between two-point classes (obtained from the difference between the target value and a threshold). It takes into account kernel functions to deal with non-linearity. The advantage of the use of SVR lies in its capacity to capture the predictor non-linearity and then use it to improve the forecasting cases (James et al., 2013, pp.344-350). A support vector regression is an adaptation of the support vector machine when the dependant variable is numeric (rather than categorical), as is the case for covid case numbers.

Section 3.42: Strengths

One advantage of the SVR is that it is a non-parametric technique so its result does not depend on the distribution of the dependant and independent variables. In fact, the SVR relies on kernel functions. The second benefit of the SVR is that it allows the generation of a non-linear model which can be greatly useful in understanding the model. By the 'principle of maximal margin', the SVR is to make the error less than a certain value. In this manner, it can be seen as a convex optimisation problem. In order to prevent overfitting, the regression is penalised by a cost parameter. Therefore, SVR provides high flexibility in terms of distribution of underlying variables, relationship between independent and dependent variables and the control on the penalty term. In fact, the SVR is capable of modeling a non-linear relationship. Moreover, SVR can avoid overfitting (Rahimi et al. 2020, p.8)

Section 3.43: Limitations

The SVR can be easily influenced by the choice of kernel together with the speed and size of training data can easily influence it (Rahimi et al. 2020, p.8). In addition, it does not work well on large datasets (this works in our favour). It's performance is deteriorated if the data has more noise. Lastly, SVM doesn't directly provide probability estimates, these are calculated using an expensive k-fold cross-validation.

The complexity of ML models is not always worth the cost. Papers that compare various models to ARMA models do not always find a significant improvement in prediction. In fact, most of the time ARMA models perform better to at least as well as ML models including Ridge, RF, SVR at least in the short run. For long run predictions ARMA models break down. Certain ML models such as SVR have some advantage in predictions over three to six days. However, other authors have made reasonable predictions for up to 17-day forecasts with ARMA models.

Section 5: Results

Section 5.1: ARMA

The first step to fitting ARMA models to our data is to check for stationarity. One way to tell is by plotting our variable of interest, new cases against time. A stationarity process has is one whose unconditional joint probability distribution does not change when shifted in time. Graphically, its mean and variance do not change over time.

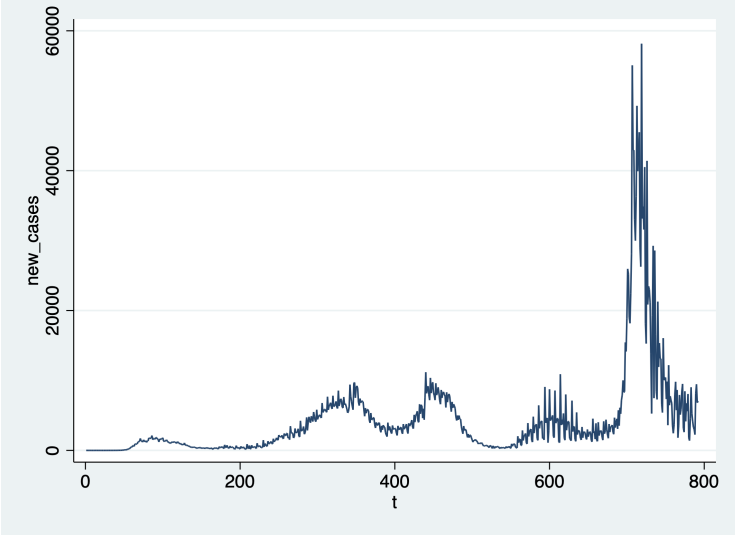

Fig. 1: time trend for New Cases in Canada

Figure 1 shows new cases plotted over time. We see that case numbers roughly oscillate around a mean value and shows somewhat constant variance. There is a start peak around the 700th day which also quickly drops back to the mean of the series. One possible reason for this peak due to major change in testing capacity and the availability of self testing or rapid testing facilities. However, for this present analysis I will assume testing capacity and any other factors that can drastically change our numbers remain consistent and the data available is representative of the actual situation at that point in time. Overall, the plot alone reveals our data is possibly stationary.

The more formal method to test stationarity is by conducting a statistical test. To do so we conduct the Augmented Dickey-Fuller test to check for the unit root. I four types of stationarity separately as shown below:

When checking for difference stationarity, our 10% critical value is -2.57. Our test statistic is -7.066 hence it surpasses the 10% critical value. The p-value is 0 so we reject the null of a unit root at 1% therefore our sequence is stationary in terms of difference stationarity. The implied value for our a1 (1-0.12) is 0.88 so we reject the null that a1=1.

Checking for drift stationarity changes the critical values but our conclusion remains unchanged.

Covariance non-stationarity test does change the value of the test statistic and the p-value. The results for the lagged variables are negative but highly significant. In fact,

lags are highly significant for upto 21 days. So our sequence is stationary in difference or covariance.

Trend non-stationarity shows the same result with the trend coefficient being positive and significant at 1%.

Hence, our data process is stationary. We do not have to worry about using differencing before using ARMA models, which would in fact result in an ARIMA. So we proceed with using the AR, MA and ARMA models.

Next, I check the autocorrelation (ACF) and partial autocorrelation (PACF) plots for the variable new cases. This is done to get a sense of the p and q for our AR(p), MA(q) and ARMA(p,q) models. From the graph below it can be seen that the ACF is geometrically decaying. The first 23 lags are significant. The PACF is significant from lag 1 to 4 and then for lags 6 to 10 (though in different directions) as well as lag 13, 15 and 21. While there is definitely a strong AR component to our series, there is probably an MA component as well resulting in an ARMA process.

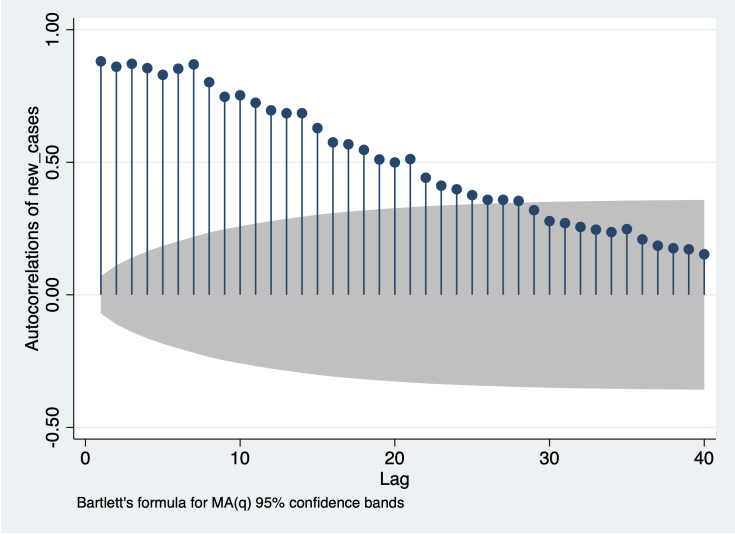

Fig. 2: Autocorrelation plot for New Cases

By analysing the resulting ACF (Figure 2) and PACF (Figure 3) plots, several potential models can be identified: the AR(22), MA(9) and ARMA(22,9). But we can do better.

In order to evaluate our models and their respective parameter, I check the Akaike (AIC)and Bayesian Information Criterion (BIC). The purpose of these is to estimate prediction error and hence the relative quality of statistical models for a given set of data. Given a collection of models for the data, these information criteria estimate the quality

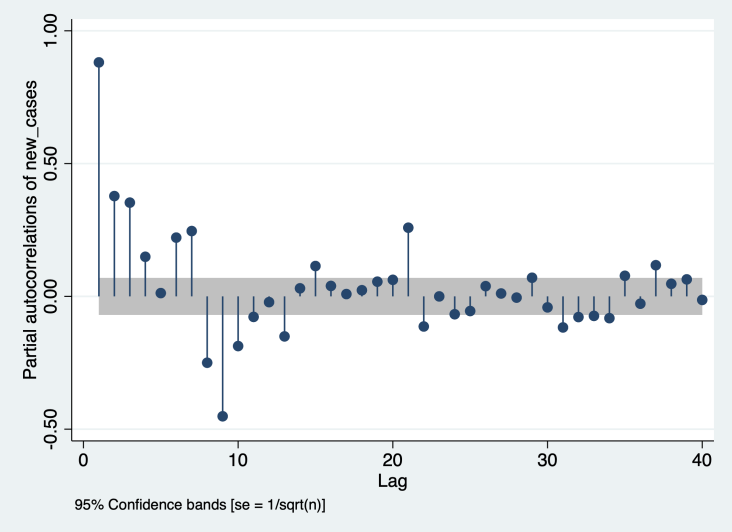

Fig. 3: Partial Autocorrelation plot for New Cases

of each model, relative to each of the other models. Thereby providing a means for model selection.

Therefore I check all possible models from AR(1) to AR(24); MA(1) to MA(22); and ARMA(1,1) to ARMA(24,22). By comparing the information criteria for each model respectively. AR(22) had the lowest AIC as well as the lowest BIC. For MA, the the appropriate model is MA(20) with the lowest AIC and BIC. For ARMA, ARMA(21,7) had the lowest BIC and ARMA(21,14) had the lowest AIC. Therefore, the prediction exercise were carried out with AR(22), MA(20) and ARMA(21,7).

Although briefly mentioned before, I used data from the onset of the pandemic till 24th March 2022 and made predictions over the next 10 days. This gives me 792 data points. I then used actual data from 24th March 2022 till 3rd April 2022 to evaluate the model predictions over that time period. The results for each of the ARMA models is plotted below.

Figure 4 and Figure 5 show our prediction results for the AR(22). New daily cases have been plotted against time. The actual numbers are depicted by the blue line whereas the prediction is depicted by the red. The vertical red line separates the pseudo out-of-sample predicted values from the rest. As expected, the model shows great in-sample predictive power. Although not perfect, the highs and lows in the numbers have been captured quite well.

However, when we look at the actual out of sample predictions, they do not seem as neat. Here the vertical line separates actual out-of-sample prediction from the remaining. The model is able to capture the trend till around the 800th day. After that the prediction seems to be flattening out and so our predictions are not as reliable.

Next, we can see Figure 6 and Figure 7 for the MA(20) model which follow that same

Fig. 4: AR(22) In sample prediction

Fig. 5: AR(22) Out of sample prediction

format as mentioned previously. Here too, pseudo out-of-sample predictions seem decent as they are able to capture the peaks and troughs in daily numbers. Yet, the true out of sample predictions start to lose their strength at around the 800th day, making the predictions less and less reliable after that.

Lastly, we can see Figure 8 and Figure 9 for ARMA(21,7). The results depict a very similar story to that of AR and MA above. The exact numbers predicted by each of the models are provided in Table 1.

Fig. 6: MA(20) In sample prediction

Fig. 7: MA(20) Out of sample prediction

Table 1

Date	Actual numbers	Predicted (ARMA)	Predicted (AR)	Predicted (MA)
18-3-2022	3340	6337	6196	4947
19-3-2022	3007	1365	2022	975
20-3-2022	2253	576	861	633
21-3-2022	8153	6289	6011	6179
22-3-2022	9466	1365	6401	5009
23-3-2022	6824	576	4904	4420
24-3-2022	7053	1365	3847	3252
25-3-2022	8944	5462	5757	5320
26-3-2022	4374	5692	4682	6381
27-3-2022	3654	5406	5857	7674
28-3-2022	6150	7376	9008	9103
29-3-2022	10561	9704	8627	9119
30-3-2022	10976	7203	8158	6307
31-3-2022	8791	6827	6845	6597
01-4-2022	13015	8451	7683	8691
02-4-2022	3357	6644	6421	7801
03-4-2022	2581	6290	7064	8272

Fig. 8: ARMA(21,7)

Fig. 9: ARMA(21,7) prediction

Section 5.11: Evaluation

Because we cannot visually distinguish our results anymore or compare them to one another effectively, we can use loss functions in order to evaluate our forecasts. A central point in the construction of loss functions is that the loss function should reflect the actual trade-offs between different forecast errors. In this sense the loss function is a primitive to the forecasting problem.

I use the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to evaluate the models both in and out of sample. Table 2 summarizes the results in terms of loss functions. As obvious, all three models perform better insample than out-of-sample. In terms of both RMSE and MAE, ARMA seems to have performed the best, followed by AR and then the MA. Note: The values for RMSE and MAE look huge at first however, given I am forecasting COVID-19 case numbers which are ranging in the 1000s everyday, this kind

of error is to be expected. Further, a lot of the seminal literature on the subject is from 2020 when the pandemic had just started and case numbers had barely reached two-digits at the time, RMSE and MAE observed there was not as large simply because case numbers were not that large to begin with. Therefore, to the best of my understanding, these number make sense.

Table 2

Model type	RMSE (in sample)	RMSE (out of sample)	MAE(in sample)	MAE (out of sample)
ARMA	2026	3316	16	270
AR	2092	3687	19	546
MA	2290	4183	18	722

On the whole, this analysis suffers from a small sample size. While the question at hand is pressing and requires immediate answers, we only have around 800 data points per country starting from January 2020 till present. While this number has greatly improved from the data points available to researcher who set out to answer this question in 2020, by general econometric standards the current number of data points is still relatively small. Further, earlier studies involving the ARMA class of models reached optimal models similar to the order ARMA(1,0,2), this is remarkably different from what I find to be optimal models for this study which include ARMA(21,0,7). While this may seem wildly different at first, the sheer difference in the number of data points available for instance 3 months vs. 2+ years is likely responsible for this. The presence of a richer number of data points allows for more complex models that better fit the data in and out of sample with a much larger number of significant lags and moving average components.

Section 5.2: SIR

Figure 10 illustrates my decision not to use the standard SIR model. When programming, the model takes as input parameters the the activity rate which is the extent of interactions between the population (α based on the equations used in the last section) as well as the infection probability which is the probability of an individual coming into contact with an infected person actually catching the disease (β). Further, the model requires numbers for the susceptible group which in this case is the entire population, as well as the infected and recovered numbers on a given day. The recovered population is set to zero initially due to the very high chance of reinfections. However, because the model inherently assumes anybody infected will directly enter the recovered group, we barely see the infection numbers show any useful movement before the virus effectively dies down (we wish).

Therefore, I first use the SI model. On top of no recovered class, this model assumes that each person in the susceptible population is equally likely to be transmitted the disease through contact with an infected individual. Once a person is infected, they cannot recover; they remain in the Infected class forever. This may make sense if we consider the long-term effects of COVID such as weakness and other symptoms that may persist long after the infection although this may differ from person to person. Although COVID has led to significant deaths, they will not be included in the model as death rates have drastically decreased over time.

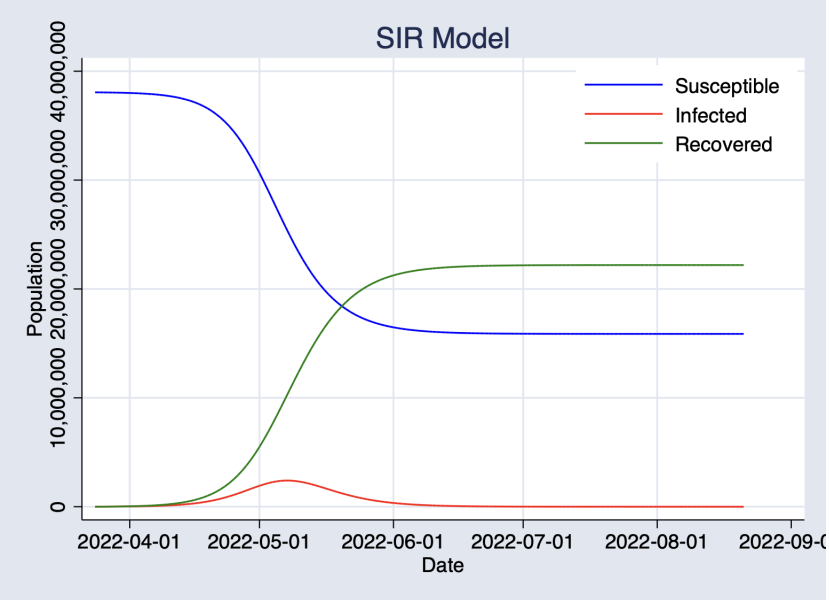

Fig. 10: Preliminary SIR

Further, the population of Canada is kept constant. Dynamics such as birth, death and migration are not considered, especially because we are forecasting over only 10 days and these are unlikely to change significantly in such a short time. Since vaccinations have not been completely immunising individuals, they are not modeled for explicitly but the activity rate and infection probability (α and β) have been decreased as a result.

In order to initiate this model, we first need to estimate α and β values. In our case, we are trying to model for 10 days ahead, starting from 25th March to 3rd April, 2022. As seen in the literature on this topic, it is common to use estimated R0 values from other researchers or data collecting organisations without calculating it anew. According to our data source, the R0 value on 24th March was 0.73 therefore, the α and β values were chosen such that the R0 equals 0.73. I run several of these and compare them to actual numbers.

The model is fed with the following parameters. As of 24th March (Day 0 for feeding our model) Population=38067913

S=S(t) is the number of susceptible individuals =38067913-7053 (total-new case)=38060860

I=I(t) is the number of infected individuals =7053

R=R(t) is the number of recovered individuals =0

(Makes a lot of sense to me)

s(t) = S(t)/N, the susceptible fraction of the population, = 38060860/38067913=0.999

i(t) = I(t)/N, the infected fraction of the population, =7053/38067913=0.0001

r(t) = R(t)/N, the recovered fraction of the population.=0

Figure 11 shows that SI model parametrised with (α, β = 0.5, 0.6) and (α, β = 0.3, 0.6). The activity level is changed but the probability of infections is kept constant. Both these values can be impacted by the level of social distancing and other measures to curb the spread of the virus. The red line shows infections and the blue line shows the susceptible population. The population is plotted o y-axis and number of days on x-axis. As evident,

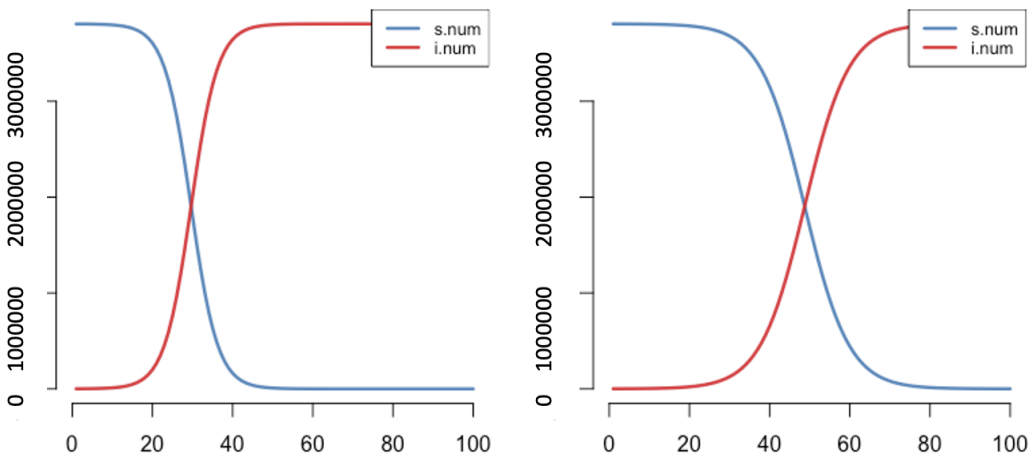

Fig. 11: SI model (DCM)

with higher activity levels α, the peak in case is observed a lot faster.

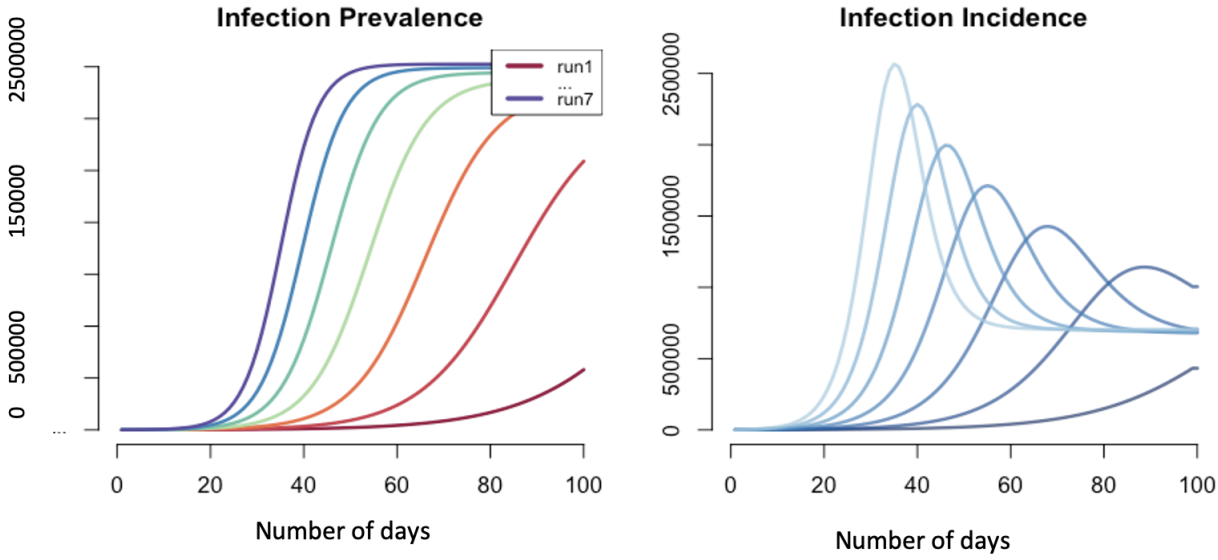

Fig. 12: SIS Sensitivity Analysis

Next, I use the SIS model. This model assumes infected individuals return to the susceptible class on recovery because the disease confers no immunity against reinfection. This makes a whole lot of sense in our scenario. The model is parameterized as usual but this time the activity levels is kept constant at 0.3 while the infection probability is varied from 0.3 to 0.9 resulting in seven different models only three of which are shown as the results of

some were pretty close to the SI models above.

Figure 12 shows infection prevalence in the population for the seven different models. This is the proportion of infected individuals in the population. It can be seen that higher infection probability leads to faster peaks (the bluer lines represent higher infection probability). Simultaneously, the infection incidence also changes drastically. This represents the probability of infection in a population (the lighter lines show higher infection probability). This analysis goes in favour of lockdowns and social distancing depending on the stage of the pandemic.

Section 5.21: Evaluation

Table 3

Date	SI (0.5 0.6)	SI (0.3 0.6)	SIS (0.3 0.7)	SIS (0.3 0.8)	SIS (0.3 0.9)
25-3-2022	3329	1664	1971	2358	2776
26-3-2022	4493	1992	2384	2938	3564
27-3-2022	6063	2385	2882	3660	4576
28-3-2022	8181	2855	3485	4560	5874
29-3-2022	11037	3418	4213	5681	7540
30-3-2022	14889	4091	5094	7077	9678
31-3-2022	20079	4897	6158	8815	12420
01-4-2022	27070	5861	7444	10978	15936
02-4-2022	36479	7015	8998	11367	20443
03-4-2022	49130	8395	10876	17023	26218

Table 3 gives the case number predictions for each of the variants of our models. In both the tables the values in the brackets are the α and β parameter values. The actual numbers have not been added to this table due to space constraints, however the RMSE and MAE calculated have been shown in the table below.

Table 4

Model type	RMSE (out of sample)	MAE (out of sample)
SI (0.5 0.6)	19083	10816
SI (0.3 0.6)	5334	3001
SIS (0.3 0.7)	5241	1907
SIS (0.3 0.8)	6032	187
SIS (0.3 0.9)	9599	3644

From the results in Table 4, the sweet spot / best model seems to be SIS with infection probability rate of 0.3 and activity rate of 0.7. This makes sense given the government's decision to open up activity and move away from the lockdown model implemented so far.

Note: All these are out of sample evaluations as the SIR does not take into account case numbers on each day and instead build the model based on single figures for the susceptible and infected population at a given time as well as estimates for activity rate and infection probability.

Further, an issue with this class of models is the logistic trajectory that is built into it. I can see this being a feasible option for a lot of our diseases but covid cases have been showing a great deal of variance. Even if one is to take information on the general trend out of these models it may not be very helpful. Especially, at this stage, where we have dealt with the virus for over two years and have a fair bit of information on how the virus has

been progressing and it is not as simple. What these numbers and trajectories do tell us however, is to expect another peak within the next 50(odd) days which could be very useful from a policy perspective.

However, the deterministic nature of the model makes it infeasible for the modeling of the present scenario. Further, by simply mapping out trajectories, this class of models fails to account for the day-to-day variability in case numbers which depending on what we require from our models could be exactly what we are looking for. For instance, the variability in these daily case numbers gives policy makers an idea of what to expect over the next few days, months from a planning perspective. If cases start to increase drastically, it can point to the onset of a new variant for which hospitalisation will have to be planned and social distancing measures will have to be put in place. While it is okay for the model to predict a steady rate whereas the real numbers fluctuate about that mean, however, it is the rise of new peaks in case numbers (due to viral mutations and formation of new variants) that these epidemiological models fail to predict, making them incapable of providing near accurate predictions.

Section 5.3: Support Vector Regression (SVR)

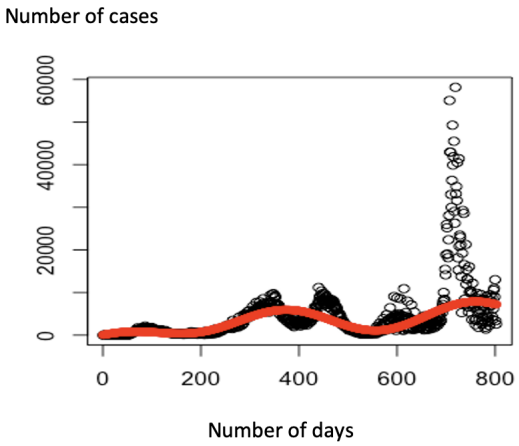

Fig. 13: SVM prediction

Figure 13 shows an SVR model fitted to actual case numbers. The white dots (with black outlines) represent actual values and the red dots represent values predicted by the SVR. The strength of this model is its ability to move smoothly with predicted numbers as well as its 'persistence' (daily case numbers hovering around previous numbers rather than showing exponential changes like those shown by the SIR model) quality which makes it closer to reality.

For this particular model, the automated kernel selection is used (which is provided by R). In this model we use Radial Basis Function kernal (RBF) due to non-linearity of the relationship. The kernel function transforms data from non-linear to linear space. The kernel allows SVR to find a model fit where the data then is mapped to the original space. This SVR gives a RMSE of 5870.

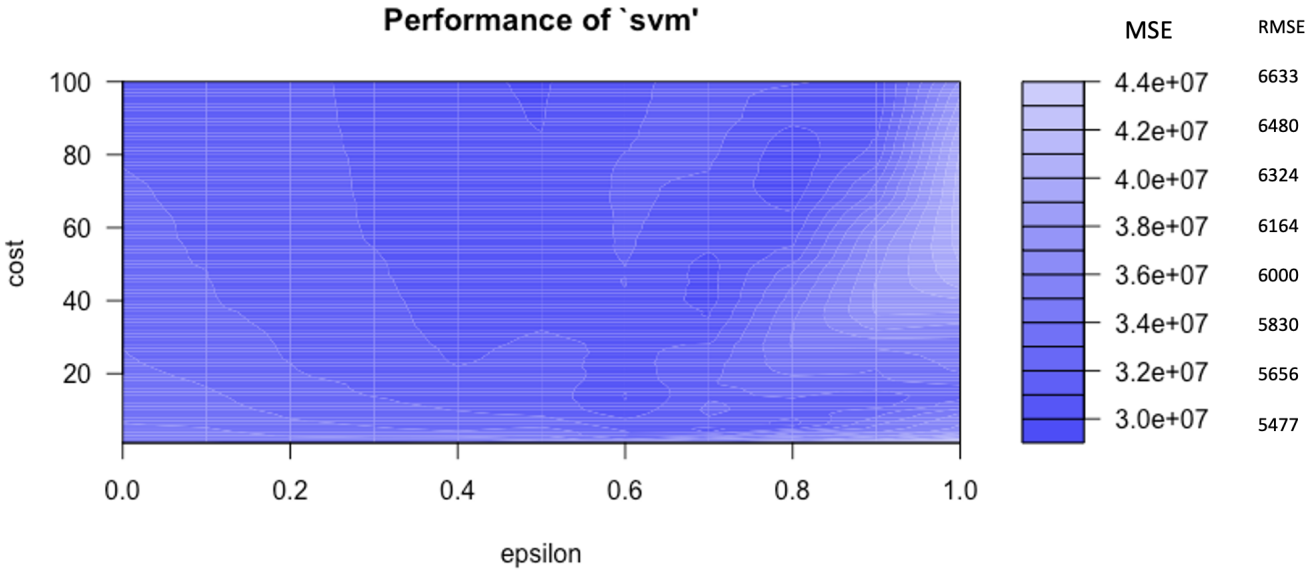

Fig. 14: Sensitivity Analysis for SVM

Next, I tune the SVR by changing the maximum error (epsilon of the error function) and cost parameters. Error is varied from 0 to 1, with intervals of 0.1 whereas the cost parameters are varied from 1 to 100. This results in 1100 models being evaluated, with its Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) reported to the right. The optimal model has the lowest RMSE. As shown in Figure 14, the darker the region, the lower the RMSE. When asked to pick the best model, R picks the model with an epsilon of 0.8 and cost of 73. It has a MSE of 29250560 which translates to an RMSE of 5310. This sampling is done using the 10-fold cross validation.

Finally, I plot the SVR as well as the tuned model in the same plot in Figure 15. The SVR is given by the blue line and the tuned model is given by the red line (with the black dots showing our actual data as before). As we can see, the tuned model fits the data much better than the SVR since the RMSE decreases from 5870 to 5310.

Section 7: Conclusion

The three models compared have all been very popular in the literature on CIVID prediction, including inter model comparisons. By comparing the different types of models among each other, ARMA(21,7) outperforms all variants of the SIR and SVR based on both the RMSE and MAE. This means the simplicity of the ARMA trumps the complexity of both the SIR and the SVR.

For ARMA interpretation of the results is relatively easy and we don't need to know alot about a process in order to make forecasts. These fairly simple models have shown reasonable accuracy in prediction over the short-term. Factors such as inadequate data and

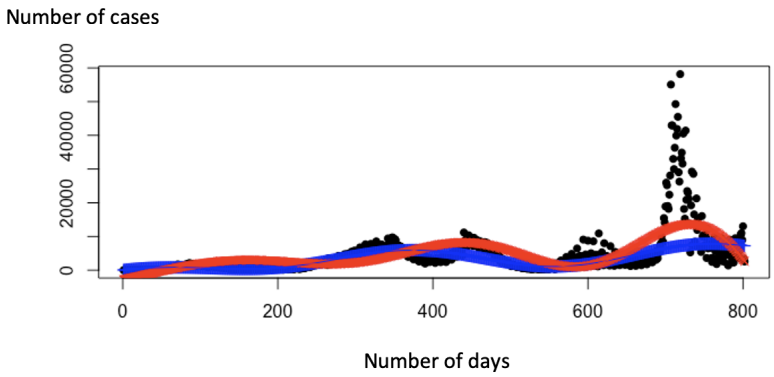

Fig. 15: Overlapped SVM models

high variation in testing strategy can affect ARMA results. Further, these models cannot account for the viral characteristics of the spread of COVID as they simply use the trend between past variables and make forecasts based on that. They can only model a linear relationship.

SIR has demonstrated remarkable success in predicting diseases. They allow for considerable flexibility and so can be adjusted based on the situation however, some of its assumptions break down for COVID. Moreover, the way the SIR is modeled is based on an exponential increase in the trajectory of the virus. Given all the COVID variants in these in the past two years which leads to a re-surge in case numbers instead on the virus dying out. This phenomenon cannot be captured by the SIR. This makes the model useful for short term forecasts only as the long term situation would be pretty difficult to model.

The SVR allow non-linear model fitting. It avoids overfitting and provides high flexibility. However, the complexity of ML models is not always worth the cost. As evident from the results in this research as well as prior literature simple ARMA models seem to outperform both SIR and SVR in short run forecasts.

Importantly, none of the models can account for mutations and multiple variants of the virus and therefore re-surges in case numbers.

Section 8: References

1. Alzahrani, Saleh I. Aljamaan, Ibrahim A and Al-Fakih, Ebrahim A. "Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA predicition model under current public health interventions". Elsevier. Volume 13. Issue 7. (2020): 914-919. https://doi.org/10.1016/j.jiph.2020.06.001

2. Arora, Parul. Kumar, Himanshu and Panigrahi, Bijaya Ketan. "Prediction and analysis of COVID-19 positive cases using deep-learning models: A descriptive case study of India". Elsevier. Volume 139 (2020):1-9. https://doi.org/10.1016/j.chaos.2020.110017

3. Avery, Christopher. Bossert, William. Clark, Adam. Ellison, Glenn and Ellison, Sara Fisher. "An Economist's Guide to Epidemiology Models of Infectious Diseases". The Journal of Economic Perspectives. Vol. 34, No. 4 (Fall 2020), pp. 79-104. https://www.jstor.org/stable/26940891

4. Chakraborty, Tanujit and Ghosh, Indrajit. "Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis". Elsevier. Volume 135 (2020): 1-11. https://doi.org/10.1016/j.chaos.2020.109850

5. Chimmula, Vinay Kumar Reddy and Zhang, Lei. "Time series forecasting of COVID-19 transmission in Canada using LSTM networks". Elsevier. Volume 135 (2020): 1-9. https://doi.org/10.1016/j.chaos.2020.109864

6. Cooper, Ian. Mondal, Argha and Antonopoulos. "A SIR model assumption for the spread of COVID-19 in different communities". Elsevier. Volume 139 (2020): 1-8. https://doi.org/10.1016/j.chaos.2020.110057

7. "COVID-19 Coronavirus Pandemic". Worldometer. Accessed 7th March, 2022. https://www.worldometers.info/coronavirus/

"COVID-19 data".Our World in Data. Accessed 3rd April, 2022. https://github.com/owid/covid-19-data/tree/master/public/data

8. Dehesh, Tania. Mardani-Fard, H.A. Dehesh, Paria. "Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA Models". Medrxiv (2020): 1-12. https://doi.org/10.1101/2020.03.13.20035345

9. Fenelli, Duccio and Piazza, Francesco. "Analysis and forecast of COVID-19 spreading in China, Italy and France". Elsevier. Volume 134 (2020): 1-12. https://doi.org/10.1016/j.chaos.2020.109761

10. "Forecasting at scale". Github. Assessed 1st March, 2022. https://facebook.github.io/prophet/

11. Gareth, James. Witten, Daniela. Hastie Trevor and Tibshirani, Robert. (2013). An Introduction to Statistical Learning with Applications in R. Springer.

12. Graham Elliot and Allan Timmerman. (2007). Economic Forecasting. Princeton University Press.

13. Hastie, Trevor. Tibshirani, Robert and Friedman, Jermome. (2008) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

14. "How COVID-19 Spreads". Centers for Disease Control and Prevention. Accessed 28th February, 2022. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html

15. Khajanchi, Subhas and Sarker, Kankkan. "Forecasting the daily and cumulative number of cases for the COVID-19 pandemic in India". Chaos 30, 071101 (2020): https://doi.org/10.1063/5.0016240

16. Kirbas, Ismail. Sozen, Adnan. Tuncer, Azim Dogus and Kazancioglu, Fikret Sinasi. "Comparitive analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches". Elsevier. Volume 138 (2020): 1-8. https://doi.org/10.1016/j.chaos.2020.110015

17. Kufel, T. “ARIMA-based forecasting of the dynamics of confirmed Covid-19 cases for selected European countries”. Equilibrium. Quarterly Journal of Economics and Eco-nomic Policy,15(2), (2020): 181–204. doi: 10.24136/eq.2020.009

18. Malavika, B. Marimuthu, S. Joy, Melvin. Najaraj, Ambily. Asirvatham, Edwin Sam. Jeyaseelan, L. “Forecasting COVID-19 epidemic in India and high incidence states using SIR and logistic growth models”. Clinical Epidemiology and Global Health. Volume 9 (2021): 26-33. https://doi.org/10.1016/j.cegh.2020.06.006

19. Maleki, Mohsen. Mahmoudi, Reza M. Heydari, M. Hossein, Pho, Kin-Hung. “Model-ing and Forecasting the spread and death rate of coronavirus (COVID-19) in the world using time series models”. Elsevier. Volume 140. (November 2020): 1-7. https://doi.org/10.1016/j.chaos.2020.110151

20. Maleki, Mohsen. Mahmoudi, M. Reza. Wraith, Darren. and Pho, Kim-Hung. “Time series modelling to forecast the confirmed and recovered cases of COVID-19”. Travel Medicine and Infectious Disease. Volume 37. October 2020. https://doi.org/10.1016/j.tmaid.2020.101742

21. Moein, Siva. Nickaeen, Niloofer. Roointan, Amir. Borhani, Niloofer. Heidary, Zarifah. Javanmard, Shaghayegh. Haghjooy, and Ghaisari, Jafar. “Inefficiency of SIR models in forecasting COVID-19 epidemic: a case study of Isfahan”. Sci Rep 11, 4725 (2021). https://www.nature.com/articles/s41598-021-84055-6

22. Nawaz, Saqib Ali. Li, Jingbing. Aslam Bhatti, Uzair. Bazai, Sibghat Ullah. Zafar, Asmat. Bhatti, Mughair Aslam. Mehmood, Anum. Ain, Qurat ul. Shoukat, Muhammad Usman. “A hybrid approach to forecast the COVID-19 epidemic trend”. PLOS ONE (2021): 1-13. https://doi.org/10.1371/journal.pone.0256971

23. Papastefanopoulos, Vasilis. Linardatos, Pantelis. and Kotsiantis. Sotiris, “COVID-19: A Comparison of Time Series Methods to Forecast Percentage of Active Cases per Population”. Applied Sciences (2020): 1-15. https://doi.org/10.3390/app10113880

24. Petropoulos, Fotios. Makridakis, Spyros. and Stylianou, Neophytos. “COVID-19: Forecasting confirmed cases and deaths with a simple time series model”, International Journal of Forecasting (2020). https://doi.org/10.1016/j.ijforecast.2020.11.010

25. Quinlan, J.R. (1992) Learning with Continuous Classes. Proceedings of Australian Joint Conference on Artificial Intelligence, Hobart 16-18, 343-348.

26. Rahimi, Iman. Chen, Fang and Gandomi, Amir H. “A Review on COVID-19 fore-casting models”. Neural Computing and Applications (2020): 1-9. https://doi.org/10.1007/s00521-020-05626-8

27. Ribeiro, Matheus Dal Molin. Gomes da Silva, Ramon. Mariani, Viviana Cocco. Coelho, Leandro dos Santos. “Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil”. Chaos Solitons Fractals. Volume 135. (2020): 1-14. 10.1016/j.chaos.2020.109853

28. Severe Acute Respiratory Syndrome (SARS). World Health Organization. Accessed 28th February 2022.

29. Siyan. Baker, Julien S. Liu, Liting. and Dong, Kechen. “Risk Prediction and Assessment: Duration, Infections, and Death Toll of the COVID-19 and Its Impact on China’s Economy”, Journal of Risk and Financial Management. Volume 13, Issue 4 (2020): 1-26. https://doi.org/10.3390/jrfm13040066

30. Syage, Jack A. “A Statistical and Dynamic Model for Forecasting COVID-19 Deaths based on a Hybrid Assymetric Gaussian and SEIR construct”. Medrxiv (2020): 1-8. doi: https://doi.org/10.1101/2020.06.21.20136937

31. Tang, Biao. Wang, Xia. Bragazzi, Nicola Luigi. Tang, Sanyi. Xiao, Yanni and Wu, Jianhong. “Estimation of the Transmission Risk of the 2019-nCoV and Its Implication for Public Health Interventions”. J Clin Med, Volume 9, Issue 2 (2020): 462. doi:

10.3390/jcm9020462

32. Wang, Peipei. Zheng, Xinqi. Li, Jiayang and Zhu, Bangren. "Prediction of the epidemic trends in COVID-19 with logistic model and machine learning techniques". Elsevier. Volume 139 (2020): 1-6. https://doi.org/10.1016/j.chaos.2020.110058

33. "WHO Director-General's opening remarks at the media briefing on COVID-19". World Health Organization. 11 March 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020

34. Wong, George N. Weiner, Zachary J. Tkachenko, Alexei V. Elbanna, Ahmed. Maslov, Sergei. Goldenfeld, Nigel. "Modeling COVID-19 dynamics in Illinois under non-pharmaceutical interventions". Medrix. (2020): 1-21. https://doi.org/10.1101/2020.06.03.20120691

35. Wu, Joseph T. Leung, Kathy and Leung, Gabriel M. "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCov out-break originating in Wuhan, China: a modelling study". The Lancet. Volume 395. Issue 10225 (2020): pp. 689-697. https://doi.org/10.1016/S0140-6736(20)30260-9

36. Yue, Xiao-Guang. Shao, Xue-Feng. Li, Rita Yi Man. Crabbe, M. James C. Mi, Lili. Hu, "Risk Prediction and Assessment: duration, infections, and death toll of COVID-19 and its impact on China's economy". Journal of Risk and Management. Volume 13. Issue 4 (2020): 1-13. https://doi.org/10.3390/jrfm13040066