Bridging machine learning and compartment models to predict an epidemic in Canada

Minh Triet Chau
s6michau@uni-bonn.de

Abstract

This work proposes a Physics-informed Machine learning method to model and emulate the progression of COVID-19. Besides the high accuracy, lower data need, and interpretability, the method also estimates hidden parameters from data, which are useful for policymakers to flatten the curve and better understand public healthcare system.

1 Introduction

It is beneficial to forecast the course of the pandemic to plan effective control strategies. There are two approaches to this prediction problem, the first is purely data-driven techniques without epidemiology prior knowledge [Alfred and Obit, 2021]. On the other hand, epidemiologists develop compartment models that embed their prior knowledge [William and Anderson, 1927, Aron and Schwartz, 1984]. As the COVID-19 epidemic poses more questions than these models can answer, scientists develop more sophisticated models with more parameters, but it is getting more complicated to tune them manually. We propose the use of machine learning (ML) to optimize the parameters fitting a compartment model [Giordano et al., 2020] and compare it with data-driven methods. The source code is available at `https://github.com/minhtriet/covid_ode`.

This work uses the Canada data from [Dong et al., 2020]. It contains the number of infected, recovered, and death cases. Due to quality control difficulties, they discontinue tracking the number of recovered data after 4th August 2021[1]. To better demonstrate the selected compartment model, this work only uses the time frame when the numbers of infected, recovered, and deaths are available, which contains 555 days. We split the data into train, validation, and test sets whose lengths are 277, 55, and 223 days, respectively. A visualization is in Figure 1.

2 Methodology

We experiment with ARIMA, Temporal Fusion Transformers (TFT) [Lim et al., 2021], and our proposed method. ARIMA is an earlier autoregressive model integrated with the moving averages. On the other hand, TFT is a more modern DL method that uses attention mechanism [Vaswani et al., 2017]. Attention mechanism achieved impressive result in Natural language processing, as is widely adapted to time series forecast. ARIMA and TFT implementation is from [Pedregosa et al., 2011] and [Herzen et al., 2022], respectively. A shared property of all three architectures is their internal parameters that we have to fit with the training data.

Designed specifically to model COVID-19, [Giordano et al., 2020] partition the population into eight classes (Table 1a) with transition probability parameters between them (Table 1b). Unlike the other compartment models, it discriminates between detected and undetected cases of infection and between different severity of illness. This separation helps policymakers estimate the number of unknown patients hidden in the community.

[1] `https://github.com/CSSEGISandData/COVID-19/issues/4465`

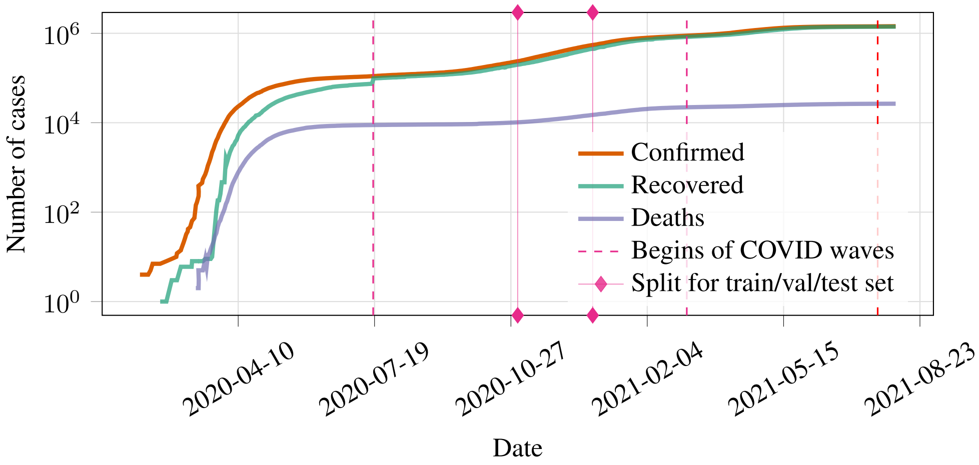


Figure 1: COVID in Canada through figures. The number of cases is on the log scale. The days of the beginning of the waves are from [Public Health Agency of Canada, 2022]. There is one infection wave (the dashed line) in the training set, zero waves in the validation set, and two waves in the test set.

Meaning	Infected		Detected
	Asymptomatic	Symptomatic	
Susceptible	No	No	-
Infected	Yes	No	No
Diagnosed	Yes	No	Yes
Ailing	No	Yes	No
Recognized	No	Yes	Yes
Threatened	No	Yes	Yes
Healed	-	-	-
Extinct	-	-	-

(a)

	I	D	A	R	T	H	E
S	$\alpha, \beta, \gamma, \delta$	-	-	-	-	-	-
I	-	ϵ	ζ	-	-	λ	-
D	-	-	-	η	-	ρ	-
A	-	-	-	θ	μ	κ	-
R	-	-	-	-	ν	ξ	-
T	-	-	-	-	-	σ	τ

(b)

Table 1: Terms and their explanation for the SIDARTHE model. In Table 1a, **T** means that the patient has life threatening symptoms while **R** does not. In Table 1b, the parameters map terms from the row to the column. While the rest of the parameters are self-explanatory, the meaning of α, β, γ, and δ are more specific. Concretely, α is the transmission rate of an S from I, β from D, γ from A, and δ from R.

When feeding the input data to this model, it is unclear whether to map the number of infected people to R or D, as the number reported in the data does not differentiate between asymptomatic and symptomatic infected. We settled on mapping it to D and using the parameter η as the transition from D to R. Besides that, H and E map directly to the number of recovered and death present in input data. All of the other parameters are unknown and learnable.

The training procedures follow [Wang et al., 2021]. We first use the Runge–Kutta 4 implementation of [Chen, 2018] to discretize and approximate from the current model's parameters. After that, we compute the loss function based on D, H, E and backpropagate the error to tune every parameter until convergence.

3 Experiment

The goal is to minimize a loss function L between the predicted and actual results. In the case of root mean squared error loss (RMSE), $L = \sqrt{\frac{(\sum_{n=1}^{N} \hat{y}_n - y_n)^2}{N}}$, with N the number of data points, $\hat{y}_n, y_n$

Parameter	Value	Discussion
Transmission rate		
α	0.4497	$\alpha > \gamma$ means people avoid individuals with symptoms. $\gamma > \beta, \delta$ (assuming that subjects who have been diagnosed are properly isolated). These parameters correlate social-distancing or face masks mandate
β	0.3382	
γ	0.4148	
δ	0.1402	
Detection rate		
ϵ	0.1708	$\theta > \epsilon$ means a symptomatic subject is more likely to be tested than an asymptomatic.
θ	0.4218	
Develop relevant symptoms rate		
ζ	0.0863	These parameters correlate with the community adaptive immunity [Boyton and Altmann, 2021]
η	0.0100	
Develop life-threatening symptoms rate		
μ	0.1974	$\mu > \nu$ means undetected cases are more likely to turn critical than detected cases.
ν	0.0100	
Mortality rate		
τ	0.1013	This is higher than reported by [Government of Canada, 2022] (approximately 0.0007% around the test time). However, note that it is the death rate from T.
Recovery rate		
$\lambda, \kappa, \xi,$ ρ, σ	0.0100	The higher these parameters are, the better the treatments and community immunity

Table 2: The values of hidden parameters and discussion. Table 1b contains more information about the above parameters.

the prediction and actual data. To be closer to reality, in the test phase, we use its past data of the test phase to continuously retrain all of the models.

3.1 Baseline: ARIMA and TFT

Since DL methods depend on the hyper-parameters they are initialized with, in the case of TFT, we perform a grid search to get the best values for the number of encoder and decoder layers, dropout rate, number of input days, and the number of attention heads. TFT would take I, E, and R as a a multivariate time series, while ARIMA only supports univariate time series. We also use the number of input dates used for TFT for ARIMA.

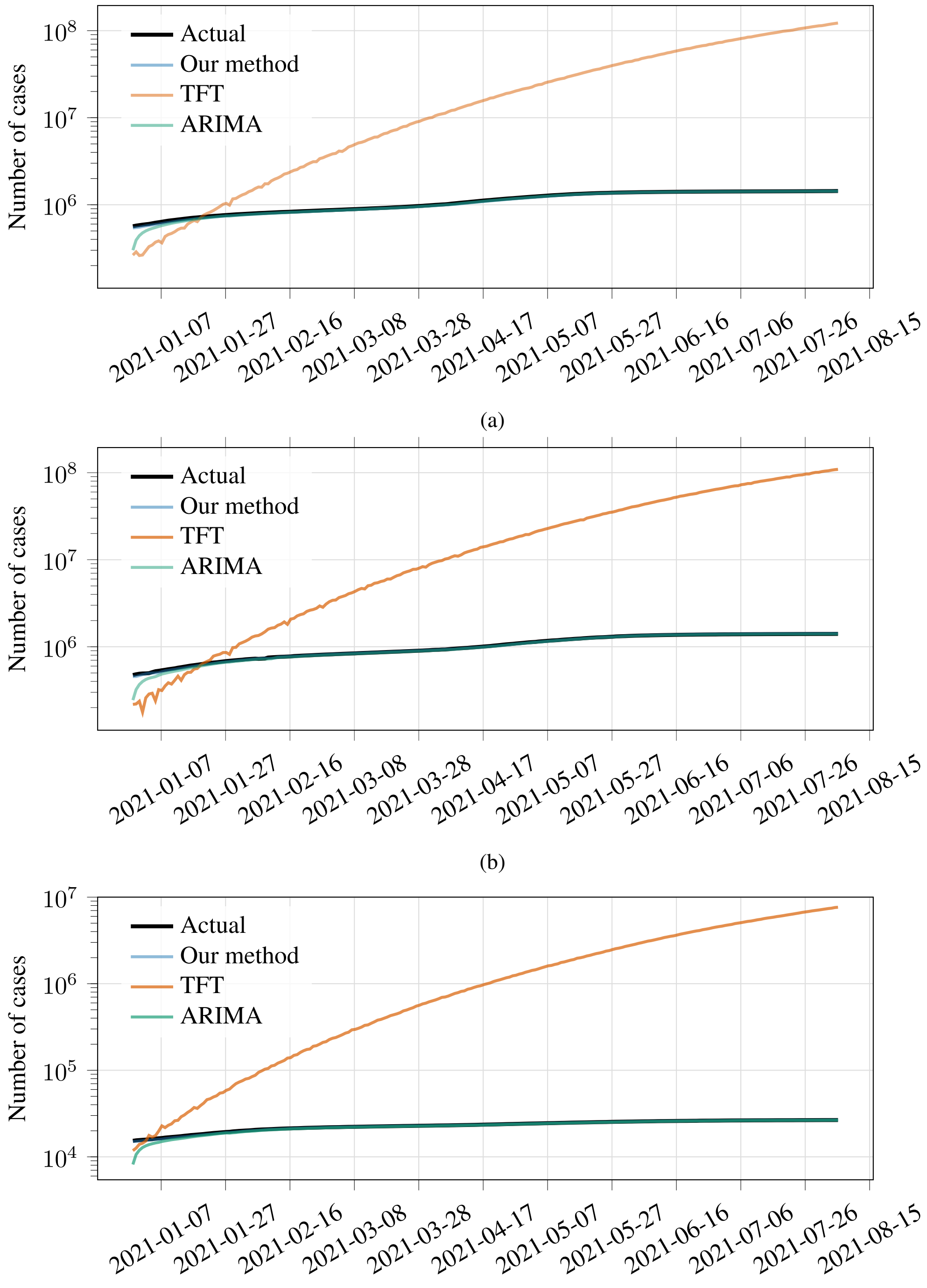
Even though TFT reports a low training error, it fails to extrapolate. It is not unexpected, as TFT has only hundreds of training data points to tune a large number of parameters (around 21000). Therefore it is prone to overfitting. Retraining during the test phase does not help improve the performance of TFT. Regardless of their performance, however, the values of parameters of both TFT and ARIMA do not help in terms of epidemiology and control. We discuss a model that addresses this shortcoming in the next session.

3.2 ML to train the compartment model

The compartment model takes the shortest time to train and has the best performance compared to other experimented methods (See Table 3). Therefore, we assume that the values of the hidden parameters after training fit the data well. As each parameter either helps gauge the healthcare system or corresponds to a different policy, we discuss more of them in Table 2. The RMSEs are in 3 and the visualization of the prediction is in Figure 2.

4 Conclusion

We pose COVID-19 as a non-linear dynamical system and propose the use of ML and a compartment model for forecasting. DL suffers from the low number of data samples and the distribution shift in the epidemic dynamics. On the contrary, through different infection waves, the ML model was



(a)

(b)

(c)

Figure 2: The prediction performance for each quantity of each method. Figure 2a, 2b, and 2c show the prediction of each method for the number of infected, recovered, and deaths, respectively. TFT consistently overestimates the number of cases, while our method and ARIMA have much smaller error margins.

Input	ARIMA	TFT	Our method
Confirmed	32189.884	46690140.864	**16410.667**
Recovered	28208.252	41675153.819	**10267.877**
Death	809.910	2958386.149	**168.567**

Table 3: The RMSE of ARIMA, TFT, and our approach.

able to learn sensible values to the parameters and produced an accurate prediction. Its parameters for unobservable data are important to understand the course of the epidemic and keep track of the healthcare system.

References

Rayner Alfred and Joe Henry Obit. The roles of machine learning methods in limiting the spread of deadly diseases: A systematic review. *Heliyon*, 7(6):e07371, June 2021. doi: 10.1016/j.heliyon.2021.e07371. URL https://doi.org/10.1016/j.heliyon.2021.e07371.

Joan L. Aron and Ira B. Schwartz. Seasonality and period-doubling bifurcations in an epidemic model. *Journal of Theoretical Biology*, 110(4):665–679, 1984. ISSN 0022-5193. doi: https://doi.org/10.1016/S0022-5193(84)80150-2. URL https://www.sciencedirect.com/science/article/pii/S0022519384801502.

Rosemary J. Boyton and Daniel M. Altmann. The immunology of asymptomatic SARS-CoV-2 infection: what are the key questions? *Nature Reviews Immunology*, 21(12):762–768, October 2021. doi: 10.1038/s41577-021-00631-x. URL https://doi.org/10.1038/s41577-021-00631-x.

Ricky T. Q. Chen. torchdiffeq, 2018. URL https://github.com/rtqichen/torchdiffeq.

Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, May 2020. doi: 10.1016/s1473-3099(20)30120-1. URL https://doi.org/10.1016/s1473-3099(20)30120-1.

Giulia Giordano, Franco Blanchini, Raffaele Bruno, Patrizio Colaneri, Alessandro Di Filippo, Angela Di Matteo, and Marta Colaneri. Modelling the COVID-19 epidemic and implementation of population-wide interventions in italy. *Nature Medicine*, 26(6):855–860, April 2020. doi: 10.1038/s41591-020-0883-7. URL https://doi.org/10.1038/s41591-020-0883-7.

Government of Canada. Public Health Infobase | Public Health Agency of Canada, October 2022. URL https://health-infobase.canada.ca/src/data/covidLive/Epidemiological-summary-of-COVID-19-cases-in-Canada-Canada.ca.pdf.

Julien Herzen, Francesco Lässig, Samuele Giuliano Piazzetta, Thomas Neuer, Léo Tafti, Guillaume Raille, Tomas Van Pottelbergh, Marek Pasieka, Andrzej Skrodzki, Nicolas Huguenin, Maxime Dumonal, Jan Kościsz, Dennis Bader, Frédérick Gusset, Mounir Benheddi, Camila Williamson, Michal Kosinski, Matej Petrik, and Gaël Grosch. Darts: User-friendly modern machine learning for time series. *Journal of Machine Learning Research*, 23(124):1–6, 2022. URL http://jmlr.org/papers/v23/21-1177.html.

Bryan Lim, Nicolas Loeff, Sercan Arik, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. 2021.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Public Health Agency of Canada. Covid-19 epidemiology update: Detailed data, maps, charts, April 2022. URL https://health-infobase.canada.ca/covid-19/#a4.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

Rui Wang, Danielle Maddix, Christos Faloutsos, Yuyang Wang, and Rose Yu. Bridging physics-based and data-driven modeling for learning dynamical systems. In Ali Jadbabaie, John Lygeros, George J. Pappas, Pablo A. Parrilo, Benjamin Recht, Claire J. Tomlin, and Melanie N. Zeilinger, editors, *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, volume 144 of *Proceedings of Machine Learning Research*, pages 385–398. PMLR, 07 – 08 June 2021. URL https://proceedings.mlr.press/v144/wang21a.html.

Ogilvy Kermack William and Gray McKendrick Anderson. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721, August 1927. doi: 10.1098/rspa.1927.0118. URL `https://doi.org/10.1098/rspa.1927.0118`.