Mathematics in Applied Sciences and Engineering
Volume **2**, Number **1**, March **2021**, pp.**1-9**

https://ojs.lib.uwo.ca/mase
https://doi.org/10.5206/mase/11101

AN EMPIRICAL FORECASTING METHOD FOR EPIDEMIC OUTBREAKS WITH APPLICATION TO COVID-19

BO DENG

Abstract. In this paper we describe an empirical forecasting method for epidemic outbreaks. It is an iterative process to find possible parameter values for epidemic models to best fit real data. As a demonstration of principle, we used the logistic model, the simplest model in epidemiology, for an experiment of live forecasting. Short-term forecasts can last for five or more days with relative errors consistently kept below 5%. The method should improve with more realistic models.

1. Introduction

It would be beneficial for planning if we were able to forecast Covid-19 outbreak like we do for weather forecasting. Unfortunately such practice is non-existent. Surely, infectious diseases modeling has been and still is very active among theorists. But there are no known reliable ways for short-term forecasting, loosely defined to be no more than one week in time, and by extension long-term forecasting, one month or more into the future, remains even more elusive. The main problem lies in a reality that for reported cases of infection over a period of time there can be infinitely many parameter combinations in any epidemic model that can just fit the data similarly. Compounding a hard problem still, all possible future trajectories are driven apart by the inherent exponential growth in viral transmission dynamics. Long-term forecasting is more of a guessing game than a science, even with reasonable epidemic models. The aim of this paper is to explore ways for short-term forecasting with the hope that it may shed some light on medium-term forecasting and long-term forecasting. More accurately, the rest of the paper was a contemporaneous account on a live forecasting experiment. As a result, the presentation below is kept in the present tense when it was written.

Every epidemic will come to an end, by which time there will be a total tally of infected for a given region which may or may not be chosen for some arbitrary reasons. For lack of a word, let us call the infected total *epidemity*, and denote it by its initial, E. E is a function of many variables. Among them includes the virus or diseases intrinsic ability to transmit among the host population, the intrinsic infectivity parameter, call it r throughout. It is also determined by how vigilant we are about the threat and what we do to keep it at bay. e.g., the practice of social distancing, which is mostly hidden but affects E greatly. Epidemity is real, and there will be an E. It can be as large as the whole population, or as small as zero. Most likely it is somewhere in between simply because a portion of population will avoid infection once the population has reached herd immunity.

Received by the editors 23 October 2019; accepted 14 December 2020; published online 22 December 2020.
2010 *Mathematics Subject Classification.* 92D30, 92-05, 92-08, 92-10.
Key words and phrases. Covid-19, logistic model, forecasting, gradient search.

To know E is to require the condition that everyone infected is tested (ignoring the efficacy of a test) and counted. In reality, we only get an approximation of E, called E^*. If there is no limitation on testing, then $E^* = E$. For most cases, $E^* < E$, or vastly under estimated, $E^* \ll E$. In the model below, we will use E throughout, but with the caveat that we are forecasting E^* in most cases.

Similarly, let $i(t)$ denote the cumulative total of infected at time t, measured in days. Then what we are actually estimating is $i^*(t)$ instead, with the same test caveat as for E^*. Also, there is a time delay between the actual infection and the reported infection, denoted by $c(t)$ for confirmed total cases at time t. But, assuming most infections undergo a mean time delay between infection and confirmation of infection, both variables slide on the time line with a mean translation in length. Therefore, if we can forecast $c(t)$ into the future, for all practical purposes, we are predicting i^* or i if testing is not an issue.

To forecast $c(t)$ is to approximate it by a variable, $x(t)$, of an epidemiological model. Let $t_1, t_2, ..., t_\ell$ be the days on which $c(t)$ is used to fit the parameters of the model. Here, the time sequence $\{t_k\}$ does not need to be every day, nor consecutive days, nor all reported days. It should be chosen by the forecaster who deems the data $c(t)$ reliable on those $\{t_k\}$ days. For example, pretty much all South Koreas data for the first 30 reported days can be used for model fitting because they are not limited by testing according news report. For another example, on February 12, 2020, there was a huge jump in c for the Wuhan, China, outbreak. For forecasters, that was a piece of good news, because it could be an accurate count for a long stretch in days, and it should be included to calibrate their forecasts. On the other hand, if we used most of the daily-case numbers prior to February 12, 2020, we would end up forecasting i^* and E^*, grossly distorting the reality. A similar problem took place for the New York State outbreak, a Wuhan lesson not learned and repeated. To reliably forecast a disease, just like for weather, a few accurate readings is a necessity. For our method to work, a few data points of $c(t_k)$ are required to be usable because of their close approximations to $i(t_k)$.

Equivalently, one can forecast the daily case numbers, which are simply the difference of the cumulative total, $\Delta c(t) = c(t) - c(t-1)$. More effectively though, we can best-fit both the total c and the daily Δc to models in combination to find the best forecasting trajectories.

2. Model

The model we will use is the most basic one for infectious diseases,

$$\frac{dx}{dt} = rx\left(1 - \frac{x}{E}\right), \tag{2.1}$$

where E is the epidemity and r is the intrinsic infectivity rate. This is the same logistic equation named by Pierre Franois Verhulst, a Belgian mathematician who introduced it in the 19th century for population studies ([7, 4, 1]). This model can be justified for infectious diseases in a few ways as follows.

Biologically, we assume that virus has evolved to sweep through a host population in a rapid and simple way. It aims to spread exponentially, $\dot{x} = rx$, the exponential growth model with a constant per spore growth rate $\dot{x}/x = r$, with r the intrinsic rate for infectivity. But, it can't grow exponentially indefinitely. It must run into the epidemity buffer E. Therefore, its per spore infection rate is modulated, say, to be $\dot{x}/x = r(1 - x/E)$. That is, when it is near E, the spread must slow down to a halt. The

factor is not a square of $1-x/E$ or a square root of $1-x/E$, which may or may not take some delicate evolutionary trickeries to do so, but for simplicity, the logistic buffering is a reasonable choice.

A second justification goes as follows. Treat E as a constant for the moment. Then $p(t) = x(t)/E$ can be treated as the fraction of infected, and $(1-p(t))$ the fraction of uninfected. The rate of change for the infected depends on social contact between the infected and the uninfected, in product with an intrinsic rate r. The intrinsic rate r depends on the virus' ability to transmit and the host population's social defensive measures against the viral transmission, which only applies to the human population.

The third justification is similar to the second, except that we assume $x(t)$ directly to be the product of the epidemity E and the infection probability $p(t)$. In this formulation, we have to treat the time t as a random variable. Unlike a usual random variable, we have to experience the time t only in the forward direction, and every instance of it. In this view, the probability density function in the time random variable is assumed to take the simplest bionomial distribution

$$\frac{dp}{dt} = rp(1-p).$$

This probabilistic formulation conforms to our view that the viral transmission is fundamentally a random and probabilistic process.

All justifications above are based on a common hypothesis that all infected, x, acquire immunity and are removed from the susceptible pool, $E-x$, and a proportion of which are infectious and the proportionality is absorbed into the intrinsic infectivity rate r. The last justification is to explain why the epidemity E can be treated as a constant rather than a time-evolving variable, because as we have been gradually increasing our social distance, E must be decreasing. The explanation lies in the time scales in which the two variables, x and E, change. The transmission in x takes place at a much faster time scale, measured in perhaps minutes, or hours. But the epidemity E changes at a slow time scale, at least in days, and delayed. Because of this fast-slow time asymmetry, for all modeling and computational purposes E can be treated as a constant relative to $x(t)$. Thus, each new forecast on $x(t)$ can forecast a slowly evolving E by the model.

As as result, the forecasting curve is the logistic solution

$$x(t) = \tfrac{x_0 E}{x_0 + (E-x_0)\exp(-rt)} \tag{2.2}$$

where x_0 is the initial condition $x(0) = x_0$, with $t = 0$ setting at any day the forecaster chooses to be the start of an outbreak. A forecaster's task is to determine the parameter values in r, E, and the initial value x_0 so that the daily prediction $x(t)$ can be made. The most basic requirement is to have $x(t_k)$ as close to $c(t_k)$ as possible. This requires to minimize the error between the reported data c and the predicted function x. For definitiveness, we will use a common daily relative error function denoted as

$$H(x_0, r, E) = \tfrac{1}{n}\textstyle\sum_{k=1}^{\ell} \tfrac{(x(t_k)-c(t_k))^2}{c(t_k)^2}. \tag{2.3}$$

It is a per data relative error. We also add one consistency constraint that

$$x(t_k) \geq c(t_k), \text{ for } k = 1, 2, \ldots, \ell \tag{2.4}$$

because the fitted counts must be no less than the data reported. The problem therefore becomes to minimize the error function H in the parameter space $D = \{q : x_0 > 0, r > 0, E > 0\}$ subject to the constraint above.

At face value, this looks like a straightforward nonlinear optimization problem to solve. This is very similar to energy minimization and is easy to solve by undergraduate students. This basic minimization approach is insufficient and will fail every time for a number of reasons. First, H does not have a unique minimum. It is an empirical fact that there seems to be no limit on the number of local minimums for H. For all practical purposes, one may just think there are infinitely many. Second, there are vastly different values in the epidemity E in combination with reasonable r that make H to be similarly small. We can call such a minimizer a critical point. A forecaster can't just pick any critical point to broadcast. This is why all attempts of forecasting the spread of diseases fall short here. It is only when an outbreak ends that everyone can find the same or similar critical values of E and r, fitting the full outbreak trajectory $c(t)$ posterior, [3].

3. Method

We now describe an algorithm in pseudo code to find the forecasting curve.

(1) Start a reasonable initial guess $q = (x_0, r, E)$.

(2) Use a nonlinear regression algorithm to find its corresponding critical point or minimizer, denoted by $\tilde{q} = (\tilde{x}_0, \tilde{r}, \tilde{E})$. Such an algorithm is usually based on Newton's gradient search idea ([5, 2]).

(3) Choose an integer, $m \geq 2$, call it a multiplier, and another number $s > 1$, call it a range scale. Pick $m-1$ many parameter points at random. Specifically, for parameter E, randomly pick $m-1$ many values from the interval $[\tilde{E}/s, s\tilde{E}]$, using uniformly distributed random numbers for definitiveness. Do the same for $\tilde{x}_0$ and respectively for $\tilde{r}$. This creates $m-1$ many new initial guesses. Run the nonlinear minimization program to create $m-1$ many new critical points, which are offsprings of the mother minimizer $\tilde{q}$. Denote by M the set of all minimizers. Notice that the choice of the lower end of the interval $[\tilde{E}/s, s\tilde{E}]$, for example, is just a simple way to avoid the needless choice 0 and we use only one parameter to fix two ends of the searching interval for simplicity.

(4) Choose a number $0 < b \leq 100$, call it a breeder percentage parameter. Let B be a set of minimizers that comprises $b\%$ of all minimizers from M. A minimizer is selected to be in B if its H error is inside the better b-percentile of M. It is for the purpose of selecting minimizers with smaller errors which are hard to find because of their small basins of attraction for Newton's fastest-descend searching algorithm. This step may be referred as selection.

(5) Choose an integer n. Repeat Step 2 and Step 3 for the breeder minimizer set B to obtain a new generation of M and B. In each iteration, the parameter $m,\ s,\ b$ may stay the same or vary. By the nth iteration, denoted by M_n or just M the set of all minimizers for simpler notation.

(6) Choose a small error tolerance, $\epsilon > 0$. Denote by M_ϵ the family of all minimizers whose errors are no greater than ϵ, and the corresponding parameter families by $M_{\epsilon,x_0},\ M_{\epsilon,r},\ M_{\epsilon,E}$. Denote the forecasting parameter values with respect to this error tolerance by $f_{\epsilon,x_0},\ f_{\epsilon,r},\ f_{\epsilon,E}$. Then each value is the *median* of its respective family:

$$f_{\epsilon,x_0} = \text{median}(M_{\epsilon,x_0}), \quad f_{\epsilon,r} = \text{median}(M_{\epsilon,r}), \quad f_{\epsilon,E} = \text{median}(M_{\epsilon,E}).$$

(7) For a sequence of $\epsilon = \epsilon_1 < \epsilon_2 < \cdots < \epsilon_k$, if f_{ϵ_k} stabilizes, the one with the smallest fit error H is used as the forecasting parameters, $f^* = f_{\epsilon_i}$ for some $1 \leq i \leq k$.

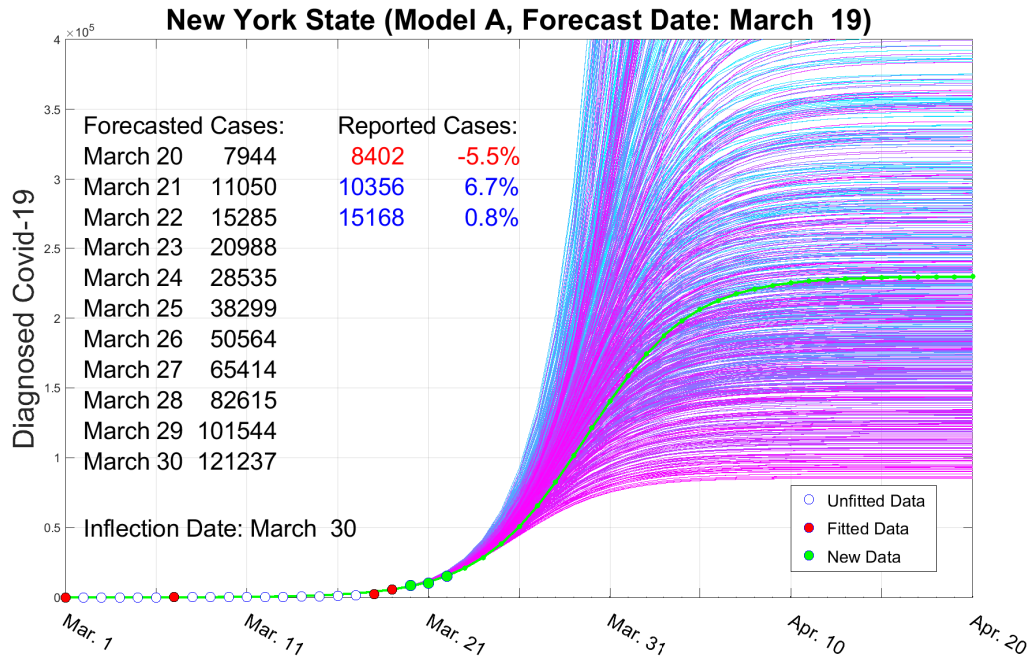
We refer to this iterative method *the median-path method*. Notice that, if ϵ is chosen to be the smallest error H for the last set M, then the median path method simply yields the least error of the minimizers. This variation is referred to as the *the least-error method*, which has been used in combination with the median-path protocol for comparison purposes.

The main justification for the least-error method is as follows. It was used to test to see if it can keep the global minimizer if the global minimizer is known. Specifically, we first generated a data sequence by the outbreak logistic function for a given parameter value q. We used this q as an initial guess for the least error method. For whatever searching parameter values in $m,\ s,\ n,\ b$, the method expectedly returned the known global minimizer q with least error $H = 0$. We also tested the case that initial guesses were chosen away from the known global minimizer q. All simulations had resulted in least-error minimizers near the known global minimizer q. The median-path method had also resulted in forecasting values close to q for small enough ϵ. Because of the selective nature of median values, some of the simulation runs landed on the true global minimizer q. As a result, one used both the median-path method and the least-error method to approximate the true global minimizer solution.

4. Experiment

I started an open experiment, first in Linkedin when the Wuhan outbreak started, and then on Twitter (@BoDeng17567961) when the US outbreak started. It was intended to simulate live forecasting. I first did it for the epicenter, Wuhan, and then the Hubei province of which Wuhan is the provincial capital, and the whole mainland of China. I did one or two forecasts for the outbreaks of South Korea, Iran, and Germany. The forecasts for China gave reasonably good predictions on the E numbers, the inflection date, from which the daily cases start to decrease, except before the big spike on c on February 12, 2020. These exercises demonstrated two points. One, when Covid-19 testing was limited for the population, our forecast could only predict i^* and E^*. Two, one can bypass problematic data and use only a few reliable c to do forecasting for $i(t)$ and E. The reason is because the logistic curve has only three free parameters, and if $c(t)$ is the ideal logistic curve, one only needs three data points in theory to determine the parameters. The reported data c for South Korea was almost a perfect logistic curve when I stopped the forecasting exercise. I only needed one repeating forecasting run to capture the outbreak curve well enough. It can be used as a textbook data set in the future for the logistic model for the early phase of disease outbreaks. I did two forecasts for Italy for a short period in the early phase of its outbreak. The short term match to real data was very good. I started the exercise for United States on March 9, 2020. All short term matches went well until March 18, 2020. That was the time when I realized there were at least two major outbreaks taking place in US, with vastly different onset date, inflection date, and epidemity. One should be for the New York State and the other for the rest. Ideally, each state should have its own forecasting when an outbreak occurs. All reported case data c were taken from Wikipedia for its transparency and checkability by crowdsourcing. From March 20, 2020, I focused my experiment exclusively on the outbreak of New York State.

All experiments were time-stamped by the contemporaneous online posts. Figure 1 shows two sample online posts, the first post and the last post for the NYS experiment. For the algorithm parameters, after tried a few values such as $s = 2$, I used $s = 5$ consistently for the posted predictions. I used typical values $m = 10,\ 50,\ 100$, and $b = 10,\ 20,\ 50,\ 100$, for variable m and b, respectively. The size of M was usually greater than 5,000. After sorting, I usually kept the first 4,000 of M to see if the median



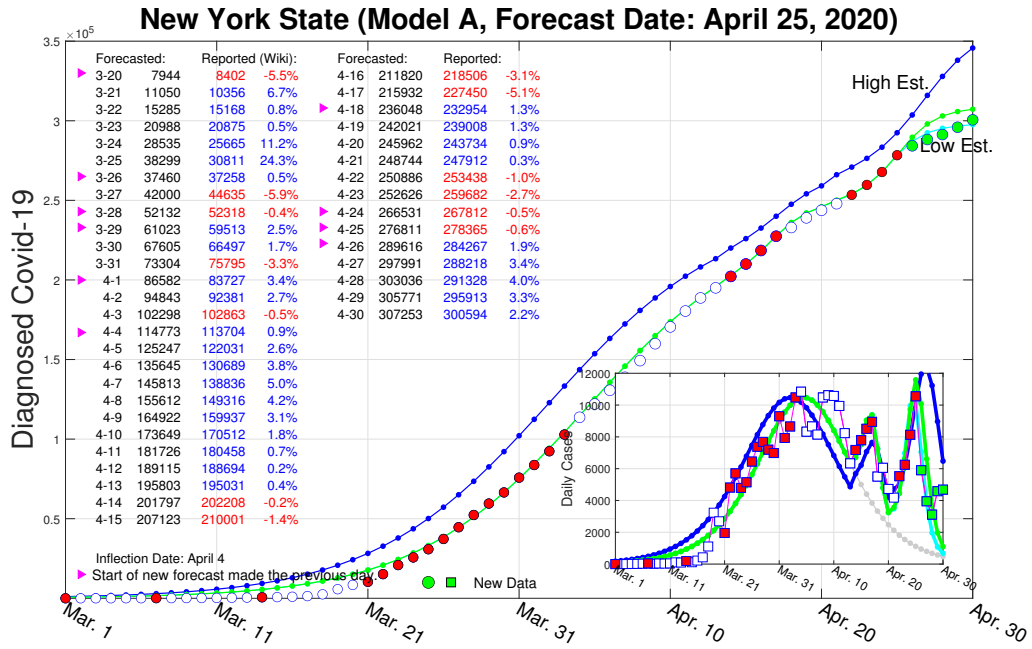


FIGURE 1. All forecasts were made when the algorithm seems to converge to a fixed $f_{\epsilon,q}$. Under-predicted are shown in red and over-predicated are shown in blue. Relative percentage errors are also shown. Usually, a new forecast was made when an under-predication occurred, or a grossly over-predication occurred such as the 3-26 forecast.

values stabilized for the first 1,000, 2,000, 3,000, and 4,000 of M if the median-path protocol was used, each of which corresponds to an M_ϵ for some increasing value ϵ. If it did, a convergence was called and a forecasting was made. For practice, I sorted the set M according to the H error from low to high. I would drop a percentage of the high end of M because they were usually not good fits to the data c visually. As an example, the top plot in Fig.1 includes the forecasting curve in green and searching curves of the first 2000 parameter values from M. For the last forecasting plot from the bottom graph of Fig.1, only the forecasting curve in green was shown, and all other searching curves were omitted.

The exercise ended on April 30, 2020, when it became apparent that the logistic model had exhausted its usefulness. More specifically, notice that after the inflection point (around April 4), the daily data exhibited an oscillation of about every seven to ten days. This secondary oscillation, mini-outbreaks on top of the underlining outbreak which the logistic model is aimed to model, was present in the data even before the logistic curve reaches the inflection point, but became increasingly pronounced afterwards. For forecasts made on April 18 and thereafter, the method was modified as follows. We used the difference between the real data and the forecasted data for the secondary mini-outbreak. We used the same method to fit the logistic model to this data difference. We then added the forecasted data difference back to the underlining logistic curve for the new forecast. This is why the forecasted daily curve exhibited the saw-tooth feature after the highest inflection point, or the inflection point for the underlining logistic curve. That is, after that point, the forecasting curve is the superposition of two or more logistic curves, the first for the underlining outbreak, the second for the first mini-outbreak after the highest inflection point, and the third for the second mini-outbreak superimposed onto the superposition of the first two, and so on.

From April 4 on, I also included two new curves to the forecasting. One was label "High Est." and the other "Low Est.", for an upper estimate and a lower estimate for c. The former was obtained by using only local maximal values of c, and the latter was a curve from the search curves having the lowest E value from the forecasting pool, usually the first 4000 searches ranked by their search errors H from the lowest to the highest.

From March 31 on, I also used the method to best-fit the daily case numbers. It became apparent that after the inflection date around April 4, there have been a persistent oscillation in daily case number of a period between 7 and 12 days, exposing the major shortcoming of the logistic model.

5. Discussion

Our method is similar to most artificial intelligence algorithms because both are based on Newton's gradient search method together with large training sets with the goal to recognize an input and to output accordingly. Our method has to generate its own training sets, which are generated from a mathematical model from an iterative process rather than existing patterns. In other words, the model carries infinitely many training sets and we have to judiciously select relevant ones. For example, if I included some obviously irrelevant logistic curves, the top plot of Fig.1 would look very different, and unorganized. The purpose of our search algorithm is to match a live data set c to some curves from the training set M_ϵ. Obviously the outcome varies with the size of M_ϵ and how uniformly the set M_ϵ is distributed. Alternative ways were tried to generate training sets, but no convergence was observed, and therefore no forecasting could be made. Averaging M_ϵ was also tried without success.

The importance of making realistic and reasonable forecasting on epidemic outbreaks is obvious, by which we can estimate other epidemiological variables and parameters, such ICU required, mortality rate, etc. That is, a better planning can be made. Perhaps more importantly, it may help prevent mass hysteria because when people can see what is to expect they tend to act more rationally.

There are many rooms for improvement and standardization. The obvious one is on the model. The simple logistic model used for the experiment is good enough for the outbreak phase of an epidemic. Once it passes the inflection, the symmetric nature of the logistic curve becomes inadequate for human population because unlike animal populations we will actively adopt counter measures against the viral spread. More sophisticated models should be used. However, the forecasting method used in this paper should be extended in principle to such models. Although more sophisticated model, such as the basic SIR model or modifications with more compartments ([6]), can do better in general, removing the symmetry artificiality of the logistic curves, I have not seen in literature that such a model has exhibited the saw-tooth like feature along the outbreak data *c*. Our *ad hoc* treatment, using superimposed logistic curves, may be used as a patchy work, but the model becomes cumbersome and forecasting becomes passive, reactive, and laborious. A worthy project for an immediate future would be to find a variant of SIR model capable of secondary-outbursts riding on a primary trajectory. With or without such a more realistic model, our method presented here should also be tested and refined for other epidemic models. Given the urgent nature of outbreaks across the globe, the method should be put to a wider test, and hopefully it can be added to a toolset to fight the Covid-19 pandemic.

The method present here is imperfect but it is demonstrated in principle that it can work for short-term forecasting. However, because the epidemity E is changing, though at a slower time scale, and it varies in a wide range, apparent further into future times, (c.f. Fig.1), long-term forecasting on E becomes unattainable. It is doubtful this situation can be improved significantly with higher dimensional models.

To summarize, a short-term forecasting method was used on real data that can find good approximations of known parameter values of a model if the data are generated by the model. The forecasting seems to work well on the outbreak phase of the epidemic. However, it is difficult to assign a theoretical degree of certainty to the empirical method because we don't know at the more basic level how the model or any model is measured against a process, natural or human-influenced or both, other than the posterior measure on the error between the real and forecasted data. Nevertheless, we believe the method can do better with more sophisticated epidemic models, empirically.

Acknowledgement: The author thanks Prof. J.-Y. Cai of University of Wisconsin-Madison, O.X. Deng of University of Nebraska-Lincoln, and the anonymous reviewers for their insightful comments on the manuscript.

References

[1] F. Brauer, C. Castillo-Chavez, *Mathematical Models in Population Biology and Epidemiology*, Springer Science and Business Media, 2013.

[2] B. Deng, *An Inverse Problem: Trappers Drove Hares to Eat Lynx*, Acta Biotheor, **66**(2018), 213-242.

[3] E. Massad, M.N. Burattini, L.F. Lopez and F.A. Coutinho, *Forecasting versus projection models in epidemiology: the case of the SARS epidemics*, Medical Hypotheses, **65**(2005), 17-22.

[4] J. R. Miner and Pierre-Franois Verhulst, *The Discoverer of the logistic curve*, Human Biology, **5**(1933), 673-689.

[5] A.P. Ruszczynski, *Nonlinear Optimization*, Princeton University Press, Princeton, 2006.

[6] P. van den Driessche and J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Mathematical Biosciences, **180**(2002), 29-48.

[7] P.-F. Verhulst, Recherches mathématiques sur la loi d'accroissement de la population [Mathematical Researches into the Law of Population Growth Increase], Nouv. mém. de l'Academie Royale des Sci. et Belles-Lettres de Bruxelles, **18**(1845), 1-41.

Supporting Materials

- All Linkedin posts (https://www.linkedin.com/in/bo-deng-180b96).
- On-line supporting materials: The median-path/least-error searching algorithm in Matlab m-files. Set M of 10,000 minimizers for March-19 forecast on New York State outbreak:
  (http://www.math.unl.edu/ bdeng1/Research/Open Experiment Covid-19 Supporting Materials/)
- New York State data from Wikipedia. Data for other countries and regions were also taken from Wikipedia:
  (https://en.wikipedia.org/wiki/COVID-19_pandemic_in_New_York_(state))

Department of Mathematics, University of Nebraska-Lincoln, Lincoln, NE 68588
E-mail address: bdeng@math.unl.edu