Modeling the 2020 COVID-19 outbreak in Blaine County, Idaho

By H Melvin Dyck[1]

Hailey ID, 11 May 2020

[1]Retired research astrophysicist. Formerly employed by the National Optical Astronomy Observatories, Tucson AZ; the Institute for Astronomy, University of Hawaii, Honolulu HI; the Department of Physics and Astronomy, University of Wyoming, Laramie WY; and the U.S. Naval Observatory, Flagstaff AZ. Visiting scientist at the Max-Planck-Institut-für-Astronomie, Heidelberg, Germany.

I. Introduction

The 2020 world pandemic of COVID-19 resulting from the coronavirus SARS CoV-2 (Mayo Clinic, 2020) hit the United States particularly hard and, for a while, Blaine County, Idaho had one of the highest per-capita rates of positive cases in the country (Ames, 2020). Once the outbreak began, local and state officials reacted quickly and issued orders for citizens to shelter in place, not to travel to neighboring communities and to practice social distancing. Unnecessary businesses were ordered closed and travelers from neighboring communities were discouraged from entering the County. The result sequestered the County which has few travel routes through the area. That means that Blaine County, in this state of lockdown, was an almost isolated population.

Under these circumstances, I was curious to see how accurately a simple empirical model could predict the extent and severity of the outbreak. One such model which has been used for isolated communities is the logistic growth model (Verhulst, 1845; Wikipedia, 2020a) which predicts the growth of populations that may be constrained by environmental factors such as physical boundaries, limited food, external hostile populations and the like. A 'generalized' version of the logistic growth model and two other model variations have recently been used to analyze the progression of the COVID-19 outbreak in China (Roosa, *et al.*, 2020) in two provinces. The results presented in that paper show good agreement between the model and the data and provide short-term forecasts (5 and 10 days into the future) of the growth of cases. The populations in these provinces were restricted by the Chinese government to prevent the spread of the disease and were, therefore, also isolated.

In the Blaine County case the 'growth population' is the coronavirus and the limits to its growth are the finite number of citizens in the County and the concomitant restrictions on those citizens. Clearly, the virus case-count cannot grow past the total human population and will be limited by the imposed restrictions which, in turn, will determine the number of re-infections that a single individual will be responsible for. I was not so much interested in short-term forecasts, as in the Roosa, *et al.* study but, rather, in the overall impact of the infection in Blaine County. Schwartz (2020), in another study of the Chinese data, performed an almost identical analysis to mine, using the simple form of the logistic growth model. He was mostly interested in trying to predict the maximum number of cases as the pandemic was in progress, recording updates as new data became available. It is interesting to see in that analysis when the prediction of the maximum number of cases became stable. He also explores other aspects of the pandemic.

For this report, I used data from the Idaho South Central Public Health District (2020a) and the Idaho Novel Coronavirus (2020) websites, both primary sources of information, and the ***Idaho Statesman*** (2020), a secondary source that maintained an archive of case reports. Since the peak of the infection wave (judged by reported positive tests) occurred in early April, I have arbitrarily stopped the analysis on May 1. The peak of the onset of symptoms actually occurred earlier, in mid-March (Idaho South Central Public Health District, 2020b). In Figure 1, I show two sets of data: the blue graphic shows a replot of the District report of the onset of symptoms and the orange graphic shows the daily confirmed case report which I have used for the analysis in this report. It is clear that there is about a two-week lag between the two.

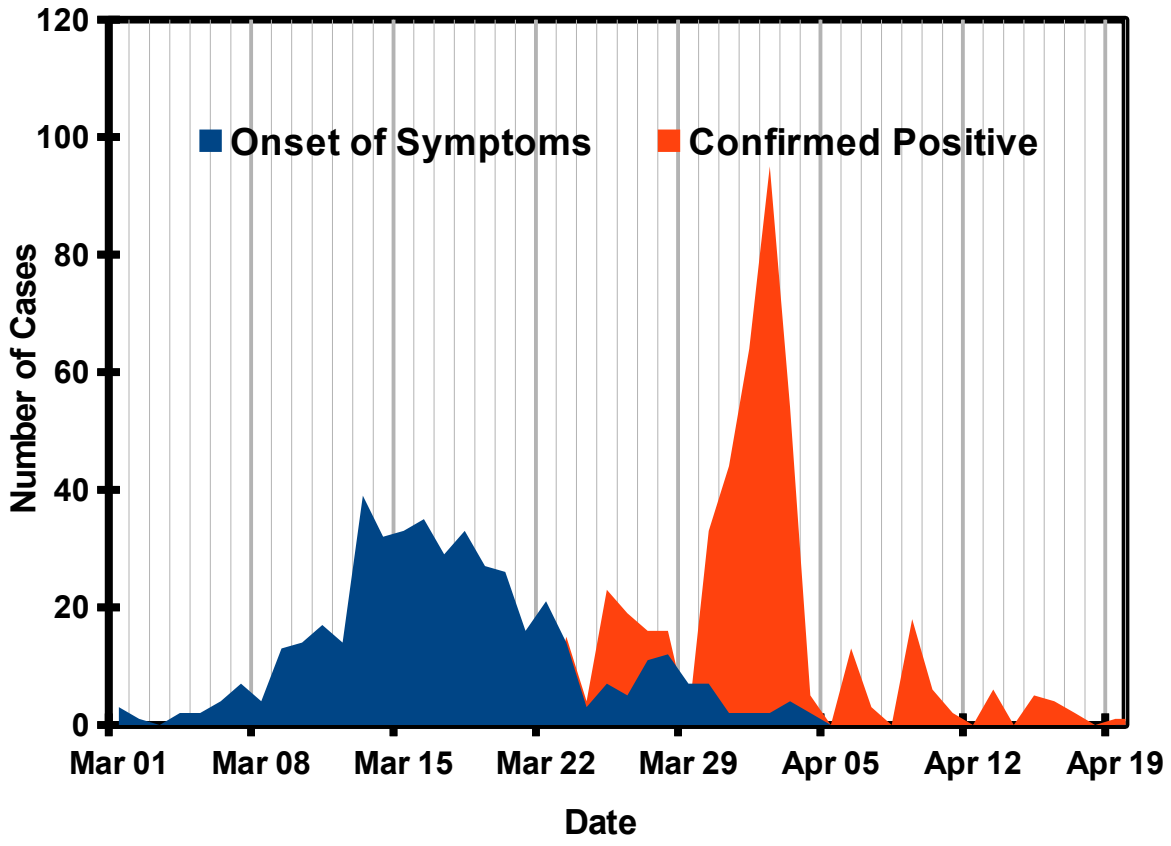

Figure 1: Date of onset of symptoms versus confirmed positive test results.

II. The Growth Rate of Positive Test Results

The logistic growth model is a mathematical function that predicts the total (cumulative) number of cases versus date. The cumulative distribution function (*CDF*) is given by

$$CDF = \frac{C_{max}}{1 + \beta\, e^{-\gamma(T - T_o)}}$$

where C_{max} is the maximum number of reported cases (positive tests not probable cases), T_o is the time of the mid-point of the *CDF* and γ is the growth rate for the outbreak (related to the re-infection rate); β is a scaling factor. I determined these constants by using a non-linear minimization of Pearson's χ^2-statistic (Pearson, 1900). In this analysis, I did some data selection: The South Central Public Health District did not routinely report cases on weekends, so I collected only data that were reported Monday through Friday. That eliminates an artificial source of noise in the data and results in blocks of 5 data points. No data are omitted, just collated in 5-day groups. I then graphed and fit the logistic growth *CDF* weekly, after the Friday data posting, keeping track of three interesting parameters: C_{max}, the predicted maximum number of cases, and the dates at which the *CDF* was predicted to reach 99% of C_{max} (T_{99}) and 90% of C_{max} (T_{90}). In Figure 2, I show the data set obtained and fitted on 24 April 2020.

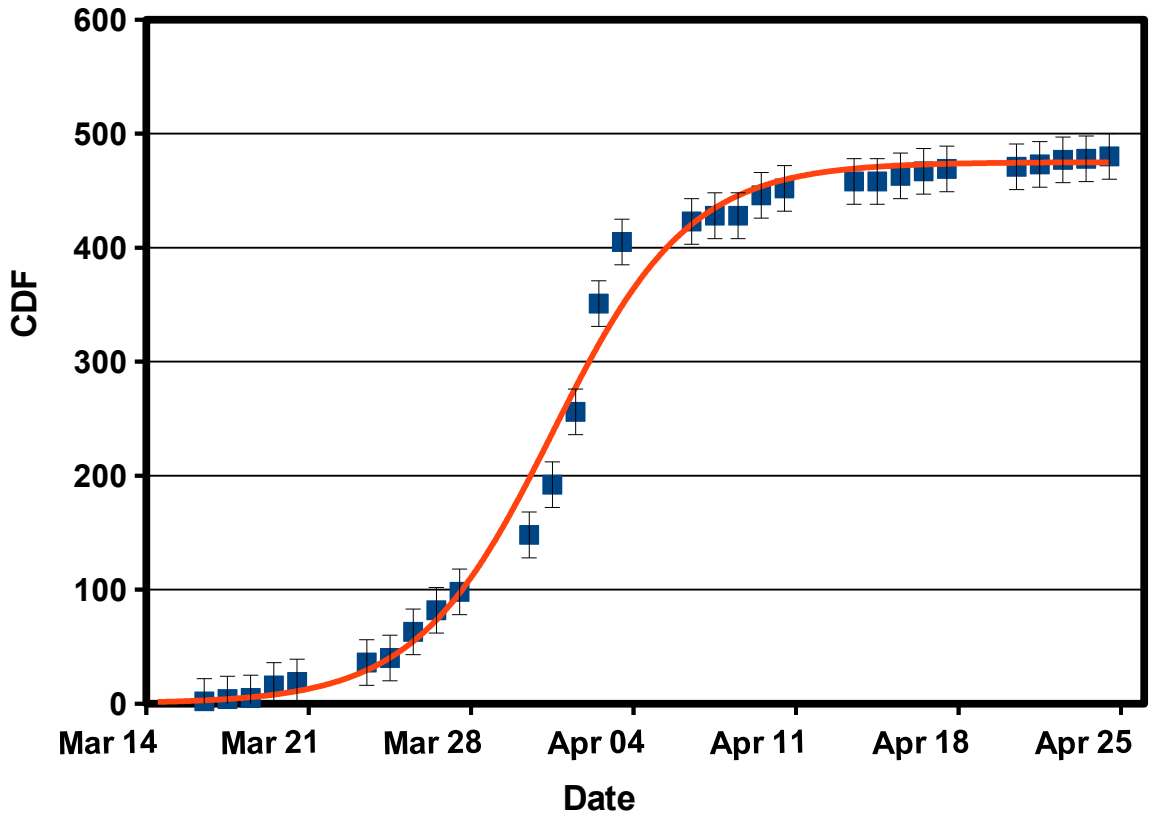

Figure 2: Cumulative data through 24 April.

The 1-σ error bars shown were determined from differences between the model fit and the reported case data and represent ±20 cases. There are 30 data points shown and only 4 lie more than 1-σ away from the model. Simple arguments would say that as many as 10 could lie that far away from the model (i.e., 33% of the data), so this is a good quality fit. However, there are two sections of the model that are systematically incorrect. In the exponential growth section the observed data grow at a steeper rate than the model predicts. In the part of the curve past the exponential growth section, the model flattens faster than the actual data. That

difference is most strongly visible in the final data block for 1 May (not shown) but only slightly apparent in the above figure.

The point of my analysis is to look at the derived parameters (T_{99}, T_{90} *and* C_{max}) for all the 5-day data blocks, beginning on 27 March and ending on 1 May and to compare those predictions to the reported data on 1 May. This starts past the point where the data first begin to rise exponentially. Data earlier than the 27 March block haven't risen sufficiently to generate a very meaningful model although they could be used for very crude estimates. For the remaining data blocks, I have summarized the results of the model fits in the table, individually, and computed their mean values as well. The errors listed are the standard deviations of those mean values tabulated. I have also tabulated the reduced χ^2 value (Bevington, 1969) resulting from the model fits. As may be seen from the table, the simple logistic growth model yields parameters that are pretty self-consistent, once the curve starts to grow exponentially, through the end of the sampling cycle. Those parameters are to be compared to the observed data listed in the last line of the table.

Derived parameters from the *CDF* model fits:				
Date	**T_{99}**	**T_{90}**	**C_{max}**	**Reduced χ^2**
03/27/20	04/17/20	04/08/20	465	1.6
04/03/20	04/16/20	04/08/20	525	4.2
04/10/20	04/14/20	04/07/20	485	3.5
04/17/20	04/14/20	04/07/20	470	2.5
04/24/20	04/14/20	04/07/20	475	2.1
05/01/20	04/14/20	04/07/20	480	1.8
Averages	**04/14/20**	**04/07/20**	**483**	**N/A**
σ_{mean}	**±1**	**±1**	**±9**	**N/A**
Reported data:				
05/01/20	04/25/20	04/09/20	487	N/A

The average model prediction for when the growth curve would reach 99% of maximum occurs about 11 days earlier than the observed date to a very high level of statistical significance (11-σ). There is very nearly zero probability that this difference is simply the result of accidental errors. For example, a 3-σ difference would have 3 chances in 1000 of being an accidental error (Bevington, 1969; Wikipedia, 2020b); a 5-σ difference would only have 6 chances in 10,000,000 of being an accident. The systematic differences between the model and the data, mentioned above, are the likely cause of this discrepancy. The difference is almost certainly attributable to inadequacies of the model. This notion is also corroborated by the reduced χ^2 value which should be near 1 for a perfect model but, on the average, is closer to 2. There is a very clear spike in the reduced χ^2 value in the second data block when the observed

data rise much more steeply than the model predicts. This gradually decays to lower values as the model and data begin to converge.

On the other hand, the predicted dates for reaching 90% of the maximum is within about 2 days of the observed value. This is very nearly within the estimated 1-σ error. So, using the model with a looser constraint on the case count results in much better agreement with the observed data.

The average predicted maximum case count is within about 1% of the observed value and lies well within the estimated error.

III. The Distribution Function

Since Blaine County was a nearly isolated community, I expected that the distribution of daily reported cases would be nearly Gaussian (*i.e.*, a normal distribution). Several independent factors determined a positive case report: contact with an infected person, acquiring the virus,

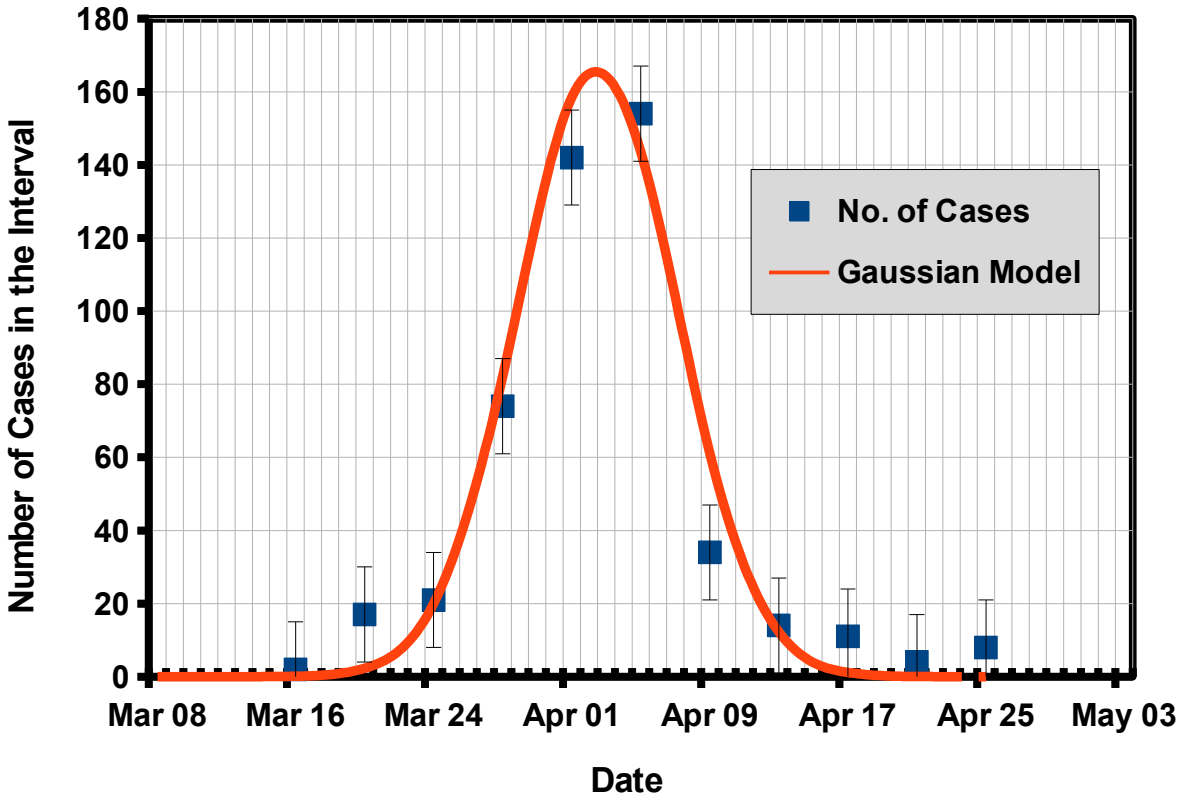

Figure 3: Case data, binned into 4-day intervals, versus date compared to a Gaussian model.

having symptoms that would be recognized as valid, obtaining a test and getting the results back and the paperwork filed. Since these are not likely to be related, the central limit theorem (Wikipedia, 2020c) predicts that the interaction of these random variables will result in a Gaussian distribution. In Figure 3, I have plotted the reported case data (shown as the blue points), binned into 4-day intervals, and fitted a Gaussian distribution (shown as the orange line) to those data. The function was forced to equal the maximum number of cases.

The error bars shown were determined from the difference between the model and the observed data and correspond to ±13 cases. The fit looks pretty good with 2-3 points (out of 11) lying more than 1-σ away from the model, indicating that the Gaussian is probably the correct distribution function. This raises the question of the validity of the logistic growth model since the *CDF* for a Gaussian is not the same. Possibly this could account for the systematic differences between the logistic growth prediction and the actual observations.

There are two features of the model that are interesting: first the peak occurs on 3 April, compared to the actual peak of 2 April. The difference results from the data binning. Second, the resulting characteristic width (σ) for the model is 4.6 days, so (roughly) 99% of the cases will lie within ±3σ, or about 28 days. From the plot, one can see that 99% of the cases occured between 20 March and 17 April. This still places the end of the infection wave earlier than the value inferred above from the observed data but is consistent with the logistic growth model *CDF* used earlier to model the growth of cases.

IV. Conclusion

In summary, the logistic growth model for Blaine County is a quick and simple approach that does not require particularly sophisticated computing power. It should allow public officials to rapidly estimate roughly how long a containment order could be in place, to within 10 days or so, and to estimate most of the infections and what the case load on health care facilities might be. Adopting a looser constraint on the predictions yields a date-of-maximum model estimate that is closer to the observed data. These are not perfect predictions but can be useful as a planning tool. Even as early as one month before the case count reached maximum, the model gave a crude estimate of both parameters. Better accuracy forecasting the date of the infection peak will require some *ad hoc* changes to the model. That is beyond the scope of this investigation.

References

Ames, M (2020). “Why an Idaho Ski Destination Has One of the Highest COVID-19 Infection Rates in the Nation,” ***The New Yorker***, 3 April 2020, www.newyorker.com/news/news-desk/why-an-idaho-ski-destination-has-one-of-the-highest-covid-19-rates-in-the-nation, Retrieved 27 April 2020.

Bevington, P R (1969). ***Data Reduction and Error Analysis for the Physical Sciences***, McGraw-Hill Book Company, New York, 336 pp.

Idaho South Central Public Health District (2020a). www.phd5.idaho.gov/CoronaVirus, Retrieved 28 April 2020.

Idaho South Central Public Health District (2020b). www.dropbox.com/sh/fig4hbbqozzwm4x/AADkFk6kkD0hIGJOtNGjjLtza/2020/4.%20April%202020?dl=0&preview=6.0++FACH+Board+Report.pdf&subfolder_nav_tracking=1, Retrieved 6 May 2020.

Idaho Novel Coronavirus (2020). coronavirus.idaho.gov, Retrieved 28 April 2020.

Idaho Statesman (2020). www.idahostatesman.com, Retrieved 28 April 2020.

Mayo Clinic (2020). "Coronavirus disease 2019 (COVID-19)," www.mayoclinic.org/diseases-conditions/coronavirus/symptoms-causes/syc-20479963, Retrieved 27 April 2020.

Pearson, K (1900). "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," ***Philosophical Magazine***, Series 5, Volume 50, Issue 302, pp. 157-175. doi.org/10.1080%2F14786440009463897, Retrieved 28 April 2020.

Roosa, K, Lee, Y, Luo, R, Kirpich, A, Rothenberg, R, Hyman, J M, Yan, P and Chowell, G (2020). "Short-term Forecasts of the COVID-19 Epidemic in Guangdong and Zhejiang, China: February 13-23, 2020," ***Journal of Clinical Medicine*****,** Volume 9, Number 2, p. 596, doi.org/10.3390/jcm9020596, Retrieved 29 April 2020.

Schwartz, S (2020). "The Mathematics Behind the Coronavirus Spread," mastermathmentor.com/mmm-archive/CoronaVirus.pdf, Retrieved 9 May 2020.

Verhulst, P-F (1845). "Recherches mathématiques sur la loi d'accroissement de la population [Mathematical Researches into the Law of Population Growth Increase]," ***Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles***, Volume 18, pp. 1-42. gdz.sub.uni-goettingen.de/id/PPN129323640_0018?tify={%22view%22:%22info%22}, Retrieved 28 April 2020.

Wikipedia (2020a). "Logistic function," en.wikipedia.org/wiki/Logistic_function, Retrieved 27 April 2020.

Wikipedia (2020b). "Standard deviation," en.wikipedia.org/wiki/Standard_deviation, Retrieved 2 May 2020.

Wikipedia (2020c). "Central limit theorem," en.wikipedia.org/wiki/Central_limit_theorem, Retrieved 10 May 2020.