Modeling Infectious Disease Trend using Sobolev Polynomials

Rolly Czar Joseph Castillo[1], **Victoria May Mendoza**[1,2], **Jose Ernie Lope**[1], **Renier Mendoza**[1,2,*]

[1]Institute of Mathematics, University of the Philippines Diliman, Quezon City 1101, Philippines
[2]Natural Sciences Research Institute, University of the Philippines Diliman, Quezon City 1101, Philippines

[*]Email: rmendoza@math.upd.edu.ph

Abstract

Trend analysis plays an important role in infectious disease control. An analysis of the underlying trend in the number of cases or the mortality of a particular disease allows one to characterize its growth. Trend analysis may also be used to evaluate the effectiveness of an intervention to control the spread of an infectious disease. However, trends are often not readily observable because of noise in data that is commonly caused by random factors, short-term repeated patterns, or measurement error. In this paper, a smoothing technique that generalizes the Whittaker-Henderson method to infinite dimension and whose solution is represented by a polynomial is applied to extract the underlying trend in infectious disease data. The solution is obtained by projecting the problem to a finite-dimensional space using an orthonormal Sobolev polynomial basis obtained from Gram-Schmidt orthogonalization procedure and a smoothing parameter computed using the Philippine Eagle Optimization Algorithm, which is more efficient and consistent than a hybrid model used in earlier work. Because the trend is represented by the polynomial solution, extreme points, concavity, and periods when infectious disease cases are increasing or decreasing can be easily determined. Moreover, one can easily generate forecast of cases using the polynomial solution. This approach is applied in the analysis of trends, and in forecasting cases of different infectious diseases.

Keywords: covid-19, data smoothing, Mpox, schistosomiasis, Sobolev polynomials, Whittaker-Henderson method

MSC2020 classification number: 65D10, 65D15, 92-10

1. INTRODUCTION

Trends of infectious diseases have drawn a lot of attention especially in the past three years because of the urgency of controlling COVID-19 and more recently, mpox disease. Trend analysis plays an important role in infectious disease control because of the need to constantly monitor how diseases spread across individuals. An analysis of the underlying trend in the number of cases or the mortality of a particular disease allows one to characterize how serious an epidemic is and to determine interventions required to halt or slow its spread. Examining the trend over an interval of time in the past may help in conducting assessments of the impact on disease transmission of a certain event or an intervention that occurred in the period of interest. Meanwhile, if the purpose of trend analysis is to generate short-term forecast of the extent of disease transmission, one may focus on trend in the most recent period.

However, trends are often not readily observable in infectious disease data because of the presence of noise. This has been observed especially in time series drawn over short intervals of time such as daily COVID-19 cases. Fluctuations in data are commonly caused by random factors, short-term repeated patterns, or measurement error. A typical approach in dealing with noisy data is to apply smoothing techniques to obtain *graduated* data that shows more regular pattern, which is easier to analyze.

Several data smoothing techniques have been developed and these methods are extensively discussed and compared in [4] and [9]. Most common techniques include the moving average and its variations, kernel smoothing, locally-weighted regression techniques, the Whittaker-Henderson method and its variant, the Hodrick-Prescott filter, and smoothing splines. Most of these techniques are discrete in nature, that is, they find smoothed data points corresponding to the crude data points, which can be then used in further analysis.

[*]Corresponding author
Received April 18^{th}, 2023, Revised June 1^{st}, 2023, Accepted for publication August 31^{st}, 2023. e-ISSN: 2549-2896, DOI:10.5614/cbms.2023.6.2.2

The Whittaker-Henderson method and smoothing splines are extended further in [1] and [11], both of which solve for a continuous function that interpolates the smoothed data instead of discrete points. This paper mainly uses the smoothing approach in [1]. The next section discusses this method.

We apply the smoothing technique proposed by [1] to extract the underlying trend in infectious diseases. This approach generalizes the Whittaker-Henderson method to infinite dimension such that the resulting solution is represented by a polynomial. The solution is obtained by projecting the problem to a finite-dimensional space using an orthonormal Sobolev polynomial basis obtained from Gram-Schmidt orthogonalization procedure and a smoothing parameter computed using evolutionary algorithms. Because the trend is represented by the polynomial solution, extreme points, concavity, and periods when infectious disease cases are increasing or decreasing can be easily determined. Moreover, the solution can also be used in generating short-term forecasts of cases.

In this work, the algorithm in [1] is modified to incorporate a recent global optimization algorithm, called the Philippine Eagle Optimization Algorithm (PEOA), to calculate the smoothing parameter. Moreover, we also show using an example, that the method can be used for short-time forecasting. Another improvement of this work from [1] is that we explore through numerical simulations how the degree and order of Sobolev polynomial can be adjusted using the L^2-norm of the difference between the data and the smooth polynomial.

The paper is organized as follows: Section 2 discusses the smoothing method, Section 3 presents numerical results on COVID-19, mpox and schistosomiasis, Section 4 concludes the paper and outlines some directions for future work.

2. METHODOLOGY

2.1. Infinite-dimensional Whittaker-Henderson method

The Whittaker-Henderson method involves finding $u \in \mathbb{R}^n$ that minimizes the penalized least squares problem given by

$$Q(u) = \sum_{i=1}^{n} \omega_i (u_i - f_i)^2 + \lambda \sum_{i=1}^{n} (\Delta u_i)^2, \tag{1}$$

where f is a vector of crude data points to be smoothed, $\omega \in \mathbb{R}^n$ is a vector of positive weights, $\lambda > 0$ is the smoothing parameter and Δ is the difference operator. A generalization of the method is presented in [1] and [11] where Equation (1) is treated as a minimization problem in the function space $L^2(\Omega)$, $\Omega \subseteq \mathbb{R}^n$. In this approach, one seeks a polynomial function u that minimizes the objective function

$$J(u) = \frac{1}{2} \int_\Omega \omega (u - f)^2 \, dx + \frac{\lambda}{2} \int_\Omega |\mathbf{D}^{(m)} u|^2 \, dx, \tag{2}$$

where f is an interpolation of crude data, ω is an interpolation of weights, $\mathbf{D}$ is the derivative operator and m is a positive integer. In [11], m is set to 1. Similar to [1], in this paper, m can be any positive integer greater than 1.

The solution u of Equation (2), which is a polynomial, is a smooth representation of data that is possibly noisy. The main difference between the result of the infinite-dimensional Whittaker-Henderson method and other smoothing methods is in the smoothness of the solution. In the infinite-dimensional case, the trend is guaranteed to be continuous and smooth because it is represented by a polynomial. All that is left is to ensure that the solution exhibits good fit to data. In Equation (2), the first term represents the fidelity of the solution to the data. However, to prevent over-fitting, which may cause poor out-of-sample predictions, the second term is added. The parameter λ is set to balance the fidelity and smoothness of the solution. As λ approaches 0, fidelity to the data is favored. Meanwhile, larger values of λ forces the solution to favor smoothness.

2.2. Smoothing using Sobolev polynomials

In a nutshell, the Sobolev space $H^m(\Omega)$ is the space of functions that are in $L^2(\Omega)$ and whose derivatives are also square-integrable. An inner product in $H^m(\Omega)$ is defined by

$$\langle u, v \rangle_{H^m(\Omega)} = \int_\Omega \omega (uv) \, dx + \lambda \int_\Omega \mathbf{D}^{(m)} u \cdot \mathbf{D}^{(m)} v \, dx. \tag{3}$$

The Sobolev space $H^m(\Omega)$ is of special interest because the minimizer u of Eq. (2) must be m-times differentiable.

Rather than directly minimizing Equation (2), one can instead solve an equivalent variational formulation of the problem. On this note, [1] showed that for an interpolation of weights ω that is bounded above, u is the minimizer of Equation (2) if and only if

$$\int_\Omega \omega(uv)\,dx + \lambda \int_\Omega \mathbf{D}^{(m)}u \cdot \mathbf{D}^{(m)}v\,dx = \int_\Omega \omega f v\,dx \quad \forall v \in H^m(\Omega), \tag{4}$$

where $\omega > 0$ is a continuous function, $\lambda > 0$, and $m \in \mathbb{N}$. Notice that the left- and right-hand sides of Equation (4) are inner products in $H^m(\Omega)$ and $L^2(\Omega)$, respectively. Thus, this variational problem can be written as

$$\langle u, v\rangle_{H^m(\Omega)} = \langle \omega f, v\rangle_{L^2(\Omega)}, \quad \forall\, v \in H^m(\Omega).$$

Using the Lax-Milgram's lemma, [1] showed that this variational problem has a unique solution in the Sobolev space $H^m(\Omega)$.

By posing the variational problem in Equation (4) in a finite subspace S in $H^m(\Omega)$ of dimension l, one can solve for u by projecting f on S, that is,

$$u = \sum_{i=1}^{l} p_i \langle \omega f, p_i\rangle_{L^2(\Omega)},$$

where $\{p_1, p_2, \ldots, p_l\}$ is an orthonormal Sobolev polynomial basis. Based on this formula, the data smoothing approach proposed by [1] only requires interpolations of data f and corresponding weights ω, and an orthonormal Sobolev polynomial basis. The construction of orthonormal Sobolev polynomials using the Gram-Schmidt orthogonalization procedure uses the inner product in $H^m(\Omega)$ given by Equation (3). To initialize the Gram-Schmidt procedure, an initial polynomial basis is required. While the standard monomial basis can be used, this paper uses Chebyshev polynomials over the interval $[a, b]$ as proposed by [1], where a and b are endpoints of the period of interest based on the given data.

2.3. Parameter values

The values of three parameters—the dimension l of the orthonormal Sobolev polynomial basis, the order of derivative m, and the smoothing parameter λ—need to be set. These parameters can be set according to the purpose of smoothing. For instance, the dimension of the orthonormal basis, which also determines the degree of the smoothing polynomial, can be chosen according to a fitness measure. One may use the L^2-norm error given by

$$L^2\text{-norm error} = \left(\int_\Omega (u - f)^2\,dx\right)^{\frac{1}{2}}$$

to gauge the goodness-of-fit of the solution to the entire data set especially if the purpose of smoothing is just to filter noise. The same approach can be applied in choosing the order of the derivative operator.

For the smoothing parameter λ, [3], [7], [10], [17] and [18] proposed minimizing the generalized cross validation score (GCV) given by

$$GCV(\lambda) = \frac{n\sum_{i=1}^{n}(\hat{f}_i - f_i)^2}{\left(n - \sum_{i=1}^{n}(1 + \lambda\gamma_i^2)^{-1}\right)^2}, \tag{5}$$

where

$$\hat{f} = \left(I_n + \lambda D^T D\right)^{-1} f,$$

n is the number of data points, I_n is the $n \times n$ identity matrix, D is a tridiagonal matrix with entries

$$D_{i,i-1} = \frac{2}{h_{i-1}(h_{i-1} + h_i)}, \quad D_{i,i} = \frac{-2}{h_{i-1}h_i}, \quad D_{i-1,i} = \frac{2}{h_i(h_{i-1} + h_i)},$$

with h_i representing the step between $\hat{f}_i$ and $\hat{f}_{i+1}$, and the γ_is are the eigenvalues of $D^T D$.

Consider the GCV score of a COVID-19 dataset as a function of λ as shown in Figure 1. In [1], the minimizer of GCV score is computed using a hybrid of MATLAB's genetic algorithm (`ga`) and constrained nonlinear multivariable optimization problem solver (`fmincon`). When we ran `ga-fmincon` 100 times on a COVID-19 dataset, 88 successfully obtained the global minimizer (red dot) and 12 calculated the local minimizer (magenta dot). The inconsistency and inefficiency of solutions of the hybrid method is remedied by [1] by running genetic algorithm 10 times and choosing the solution that produces the lowest GCV score, before using this solution as a starting point for `fmincon`. However, this can be computationally expensive and may still result in a local minimizer. To resolve this, the minimizer of the GCV score is obtained numerically using an evolutionary algorithm called *Philippine Eagle Optimization Algorithm* [8], which is a nature-inspired, meta-heuristic optimization technique that utilizes the hunting behavior of the Philippine Eagle. It employs three distinct global operators for its exploration strategy, and it also features an intensive local search during each iteration, resulting in a strong ability to obtain accurate minimizers. The approach's ability to balance exploration and exploitation makes it a reliable algorithm in solving intricate optimization problems. PEOA was applied 100 times to find the minimizer of GCV in Figure 1. The inputs of PEOA are the dimension, the objective function, the lower bound, the upper bound, and the maximum number of function evaluations, which we set to 1, the function in Equation (5), 0.01, 1, and 1000, respectively. All the other hyperparameters of PEOA were set to their default values. PEOA was able to estimate the global minimizer 100% of the time.

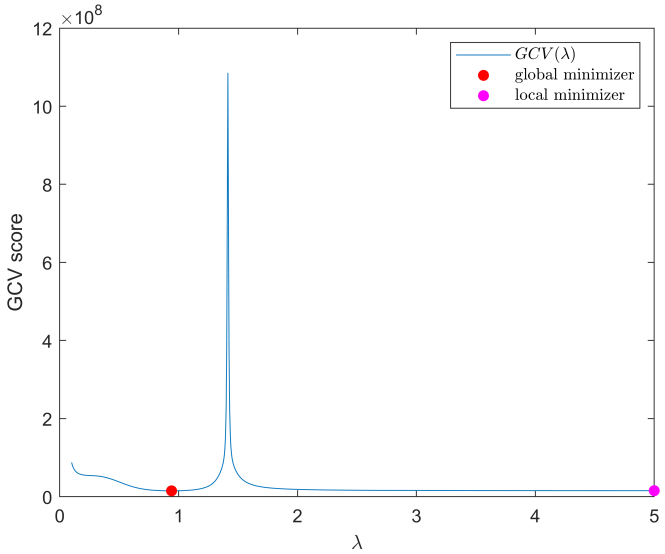

Figure 1: GCV score as a function of the smoothing parameter λ. The two minimizers (one local and one global) are shown.

2.4. Forecasting using the smoothing polynomial

Forecasting using the smoothing polynomial is straightforward because one only needs to evaluate the polynomial for every point outside of the sample. As in other methods, short forecast horizons are preferred when forecasting using the smoothing polynomial. A common practice in forecasting time series is to include a confidence interval for the forecast. This paper adopts the bootstrap method carried out in [1] based on the method proposed by [2] and [6].

2.5. Smoothing algorithm

The smoothing method discussed above is summarized in Algorithm 1.

Algorithm 1 Data Graduation Using Sobolev Polynomials

1: *Input:* The crude data $\{f_i\}_{i=1}^n$ and corresponding weights $\{\omega_i\}_{i=1}^n$, if any.
2: Set the value for l (dimension of the polynomial space) and m (order of differentiation).
3: Determine the polynomial interpolations, ω and f, of the weights and the data, respectively.
4: Choose the value of the smoothing parameter λ or estimate the minimizer of the function GCV in Equation (5) using the Philippine Eagle Optimization Algorithm.
5: Generate an initial polynomial basis (monomials, or Chebyshev polynomials, etc.).
6: Using the smoothing parameter λ and the initial basis polynomials, obtain a set of orthonormal Sobolev polynomial functions $\{p_1, p_2, \ldots, p_l\}$ using Gram-Schmidt orthogonalization procedure and the inner product in Equation (3).
7: *Output:* The smooth (polynomial) approximation of the data given by $u = \sum_{i=1}^{l} p_i \langle \omega f, p_i \rangle_{L^2(\Omega)}$.

3. Results

The smoothing method is applied to analyze the trends in mpox cases in Germany, COVID-19 cases in Canada, and schistosomiasis cases in the Philippines over specific time periods. Data on mpox and COVID-19 used in analyses were obtained from Our World in Data [13], [12], while schistosomiasis data were obtained from [15]. The data points in all the examples have uniform weights of 1.

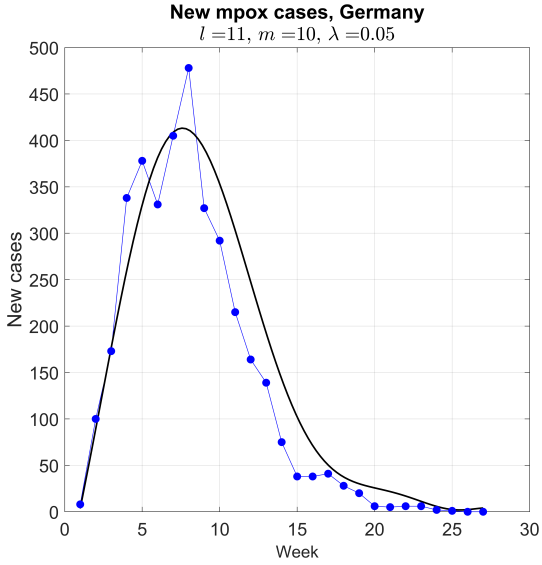

Figure 2: Smoothing of weekly new mpox data, Germany. The black line represents the smoothed data.

Figure 2 presents a scatter plot of weekly new mpox cases in Germany over 30 weeks since 3 June 2022. Weekly data is obtained by taking the sum of reported daily new cases of mpox. This adjustment is done to address the effects of non-reporting especially during weekends. The black line represents the smoothed data using the proposed method. The smoothing polynomial is obtained using an orthonormal Sobolev polynomial basis with dimension 11, order of derivative 10, and a smoothing parameter λ calculated as 0.05. As exhibited by the smoothed data, daily mpox cases in Germany peaked at the 8th week and persistently died down afterward. The change in daily cases slowed after the 12th week.

The smoothing of daily new COVID-19 cases in Canada is presented in Figure 3. The raw data represented by blue dots cover 180 days from 9 August 2020. For this example, the dimension of the orthonormal Sobolev polynomial basis is set to 9 and the order of derivative is set to 7 [16].

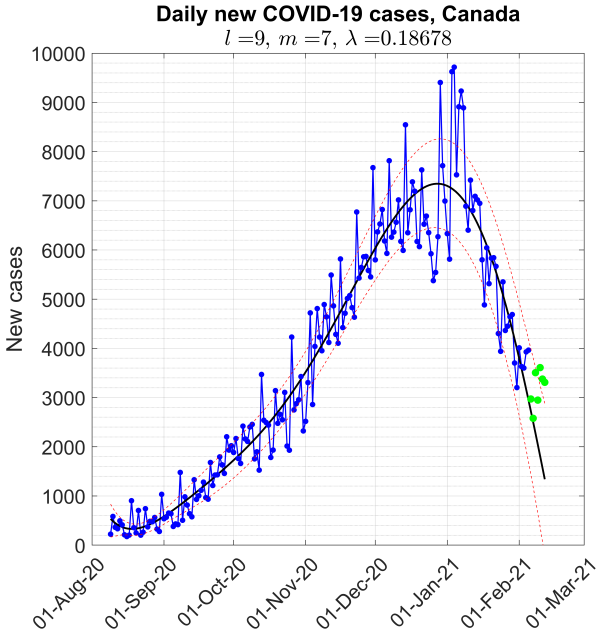

Figure 3: Smoothing and forecasting of daily COVID-19 cases, Canada. The black line represents the smoothed data while the blue dots represent the crude data used to estimate the smoothing polynomial. The green dots represent out-of-sample data. The segment of the smoothed data without blue dots represents a 1-week forecast of new COVID-19 cases. The red broken lines represent the bounds of a 95% confidence interval for the smoothed data.

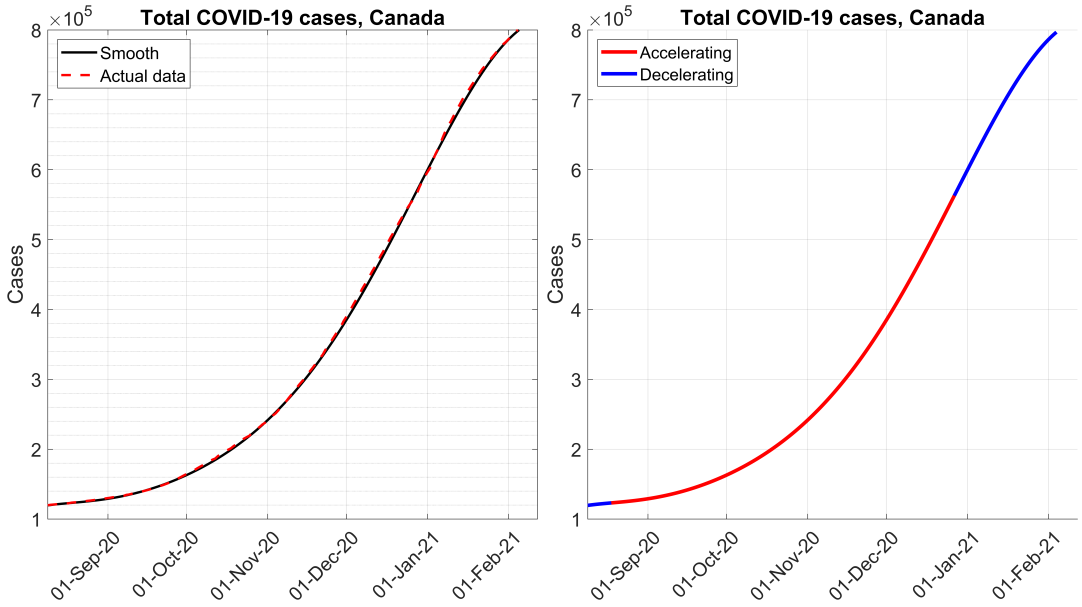

Figure 4: Structural analysis of cumulative COVID-19 cases, Canada. The left panel compares the plot of cumulative COVID-19 cases obtained from the smoothing polynomial and the actual data. The right panel characterizes the rate at which COVID-19 spreads over the time horizon of interest.

We now illustrate how our method can be used to forecast trends based solely on the crude data. The length of forecast horizon does not have an exact definition in the literature, and in health forecasting, this may be set depending on the situation. For public health situations characterized by more frequent *shocks*, short forecast horizons may be more useful especially in policy-making [14], [5]. This is so because apart from the need to constantly update forecasts as soon as new data becomes available, long term forecasts in this case may suffer from larger errors due to more unexpected movements that may change the trend in actual data [16]. In Figure 3, the smoothing polynomial is evaluated beyond the given data to provide a one-week forecast of new COVID-19 cases. The forecast horizon is set to one week as suggested in [14] for COVID-19 cases. Note that the user can modify the forecast window since extrapolation can easily be done by evaluating the obtained polynomial over an extended time period.

A 95% confidence interval for the smoothed data, shown in Figure 3, is also constructed by bootstrapping to provide a range of possible values especially for out-of-sample forecast. Integrating the smoothing polynomial over the period of interest yields an approximation for the actual cumulative COVID-19 cases as shown in the left panel of Figure 4. Meanwhile, the derivative of the smoothing polynomial provides information about periods when total COVID-19 cases are rising at a faster or slower rate. The right panel of Figure 4 shows that total COVID-19 cases rapidly grew from September 2020 to December 2020 and slowed down starting on January 2021.

Figure 5 presents the application of the smoothing method in the analysis of trends in schistosomiasis in the Philippines over the years 2000 to 2019. Unfortunately, this data set is infested by missing data points particularly for the years 2009-2011 and 2013. Missing data is a common problem encountered in data analysis and is commonly addressed by imputation. Discrete smoothing techniques, such as the Whittaker Henderson Method and the moving average, may fail in datasets with missing data points mainly because both techniques require data points to be evenly spaced. In contrast, the method used in this paper handles this problem well because data imputation is implicit in the interpolation. Interpolation of data points creates a continuous function over the entire domain, even in intervals where there are missing data. Since the interpolation of data is all that is required by the algorithm, having missing data points is easily remedied.

In the smoothing of schistosomiasis cases, piecewise linear interpolation is applied on the crude data set and this is presented as the red broken line in Figure 5. As shown, the trend in schistosomiasis cases becomes clearer after the smoothing despite the time series having breaks in certain periods because of missing data.

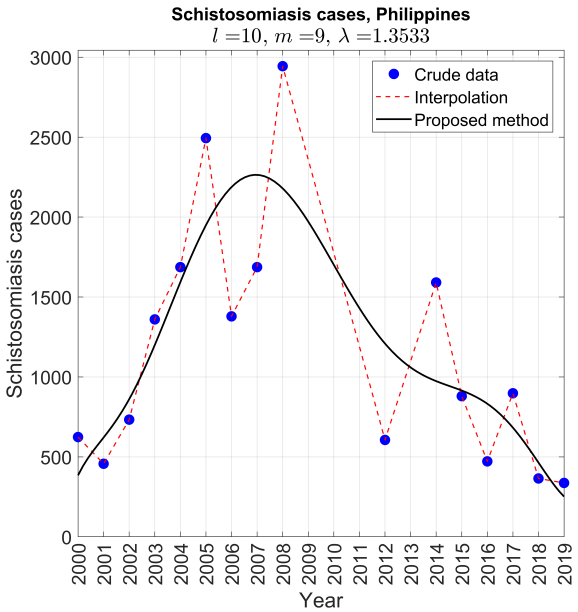

Figure 5: Smoothing of schistosomiasis cases with missing data, Philippines. The blue points represent the crude data. The red broken like represents the piecewise linear interpolation of data. The black line represents the smoothed data.

4. CONCLUSION

Trends analysis provides insights that can be used in controlling the spread of infectious diseases. In most disease data, trends may not be readily observable because of noise, thus the need for data smoothing. This paper applied a data smoothing technique that uses Sobolev polynomials. The underlying trend in data generated by this technique is represented by a polynomial. Using this technique, this paper generated the underlying trends in mpox, COVID-19, and schistosomiasis in selected countries. Forecasting using this smoothing technique is also demonstrated.

Future work on data smoothing of infectious disease data may focus on optimizing the dimension of the basis polynomial l, and the order of derivative m. Along with the smoothing parameter λ, both l and m can be considered as hyperparameters that can be tuned to improve goodness-of-fit of the smoothing polynomial to crude data, and to produce more accurate out-of-sample forecast. Another topic that can be explored is the extension of the smoothing method on disease data that have dimensions other than time such as geographic area. One can also explore using other basis functions in generating the smooth approximation. Lastly, one can also explore how the Sobolev polynomial smoothing technique relates to parameter estimation and solutions of dynamical systems for infectious diseases.

ACKNOWLEDGEMENT

RCJ Castillo, JE Lope, and R Mendoza were supported by a grant from the Computational Research Laboratory of the Institute of Mathematics, University of the Philippines Diliman. VM Mendoza and R Mendoza are funded by the Natural Sciences Research Institute, University of the Philippines Diliman under the research grant MAT-21-1-04.

REFERENCES

[1] Castillo, R. and Mendoza, R., On smoothing of data using Sobolev polynomials, AIMS Mathematics, 7(10), pp.19202-19220, 2022.

[2] Chowell, G., Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts, Infectious Disease Modelling, 2(3), pp. 379-398, 2017.

[3] Craven, P. and Wahba, G., Smoothing noisy data with spline function: estimating the correct degree of smoothing by the method of Generalized Cross-Validaton, Numerische Mathematik, 31(4), pp. 377-403, 1978.

[4] Debon, A., Montes, F. and Sala, R., A comparison of nonparametric methods in the graduation of mortality: Application to data from the Valencia region (Spain), International Statistical Review, 74(2), pp. 215-233, 2006.

[5] Regional Committee for the Eastern Mediterranean, Forecasting in communicable diseases. World Health Organization, 1999.

[6] Efron, B. and Tibshirani, V., An introduction to the Bootstrap, CRC Press, 1992.

[7] Eilers, P., A perfect smoother, Analytical Chemistry, 75(14), pp. 3631-3636, 2003.

[8] Enriquez, E., Mendoza, R. and Velasco, A., Philippine eagle optimization algorithm, IEEE Access, 10, pp. 29089-29119, 2022.

[9] Friedman, J., Hastie, T. and Tibshirani, R., The elements of statistical learning: Data mining, inference and prediction, Springer, 2009.

[10] Garcia, D., Robust smoothing of gridded data in one and higher dimensions with missing values, Computational Statistics And Data Analysis, 54(4), pp. 1167-1178, 2010.

[11] Manejero, J. and Mendoza, R., Variational approach to data graduation, Philippine Journal of Science, 149(2), pp. 431-449, 2020.

[12] Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hasell, J., Macdonald, B., Beltekian, D. and Roser, M., Coronavirus Pandemic (COVID-19), Our World In Data, 2020. https://ourworldindata.org/coronavirus

[13] Mathieu, E., Spooner, F., Dattani, S., Ritchie, H. and Roser, M. Monkeypox., Our World In Data, 2022. https://ourworldindata.org/monkeypox

[14] Oliveira, T. and Moral, R., Global short-term forecasting of COVID-19 cases, Scientific Reports, 11(1), ID 7555, 2021.

[15] Riñon, J., Mendoza, R., Reyes V, A., Belizario Jr., V. and Mendoza, V., Management and control of schistosomiasis in Agusan del Sur, Philippines: A modeling study, Research Square, 2020, Preprint.

[16] Soyiri, I.N. and Reidpath, D.D., An overview of health forecasting, Environmental Health and Preventive Medicine, 18(1), pp. 1–9, 2013.

[17] Wahba, G., Spline models for observational data, Society for Industrial and Applied Mathematics, 1990.

[18] Weinert, H., Efficient computation for Whittaker–Henderson smoothing, Computational Statistics and Data Analysis, 52(2), pp. 959-974, 2007.