Predicting Disease Incidence with Natural Cubic Splines

Yew-Meng Koh[1], Noah Kochanski[1]
[1]Dept of Mathematics, Hope College, 27 Graves Place, Holland MI 49423

Abstract
High-degree polynomials provide great flexibility and potentially perfect fit of historical time series data. Such flexibility, however, often leads to overfitting and results in models with poor predictive performance. Splines are a low-degree polynomial smoothing method which reduces these overfitting effects. We use a cross validation method for time series in order to compare the performance of various models which utilize smoothing splines with regard to their forecast accuracy of Singaporean dengue fever counts.

Key Words: Dengue fever, time series, smoothing splines, cross validation

1. Dengue fever background, Time Series, Predictor Variables

Dengue fever (DF) is a disease, borne by *Aedes* mosquitoes, and is endemic in tropical areas of the world. Typical symptoms are high fever, rash, joint and muscle pain. DF, if not treated, can lead to a potentially fatal, more serious disease, known as dengue hemorrhagic fever. Knowledge of the disease enables us to use information from environmental predictor variables, assisting us in forecasting DF counts more accurately.

1.1 Dengue fever count dataset

The DF count data used in this paper were obtained from the Ministry of Health, Singapore [1]. Figure 1 shows the DF counts used in this paper's analysis.

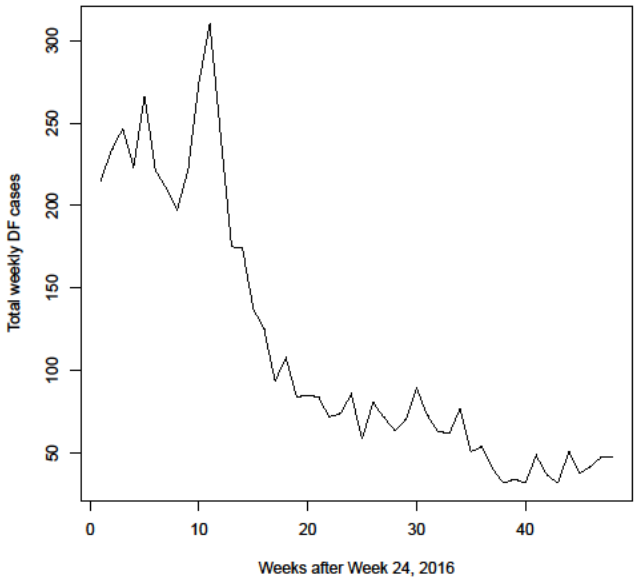

Figure 1: Weekly, reported DF counts across Singapore from Week 24, 2016.

2. Some Possible Predictor Variables

As mentioned in the previous section, various environmental predictor variables were considered when predicting DF counts. Among these were the weekly maximum temperature and average daily humidity. These variables were found to have weak correlation with DF counts, and did not offer much improvement in forecast accuracy of DF Counts. The same phenomenon was observed when using Zika disease counts as a predictor (Zika is also carried by the same *Aedes* mosquito). Zika counts had a surprisingly weak correlation with DF counts. This challenge was compounded by the fact that the Singaporean Ministry of Health data for Zika counts was only available starting from 2016.

Total weekly rainfall, when lagged appropriately, was found to be quite strongly negatively correlated with DF counts. Figure 2 shows total weekly rainfall in Singapore for the time period considered in this paper. Rainfall data was obtained from [3].

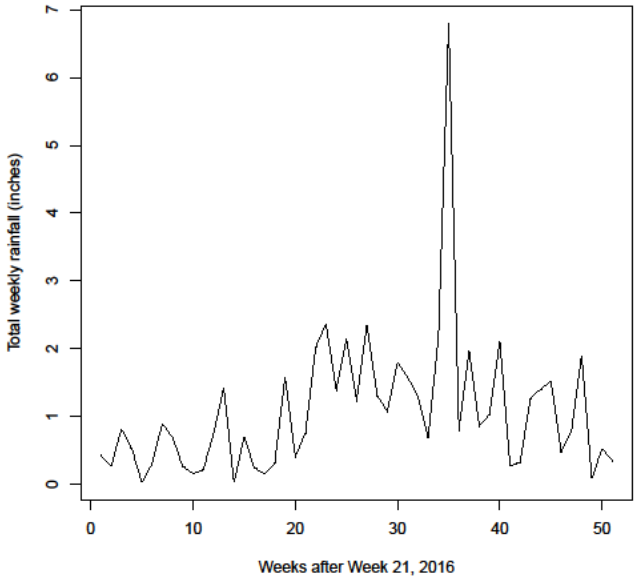

Figure 2: Total weekly rainfall in Singapore, from Week 21, 2016.

Three weeks after a week with heavy total rainfall, a drop in DF counts was observed. The reverse was also observed to be the case (see [2] for more details). The biological explanation for this was the destruction of *Aedes* mosquito larvae due to heavy rain, and the two to three week incubation period before a DF infected person begins to show symptoms of the disease. Figure 3 plots the weekly DF counts and the total weekly rainfall (at a lag of 3 weeks) together and shows this phenomenon.

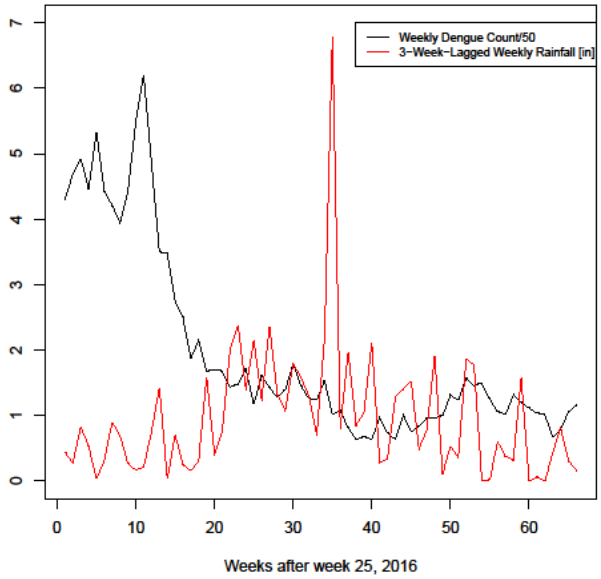

Figure 3: Weekly, reported DF counts across Singapore from Week 24, 2016 and Total weekly rainfall in Singapore from Week 21, 2016.

3. Natural Cubic Splines and Cross-validation in Time Series

3.1 Natural cubic splines

Given a time series $\{x_{t_0}, x_{t_1}, \dots, x_{t_n}\}$, a natural cubic spline is a sequence of cubic polynomials $S_0(t), S_1(t), \dots, S_{n-1}(t)$, where $S_{i-1}(t) = a_{i-1}t^3 + b_{i-1}t^2 + c_{i-1}t + d_{i-1}$, defined on the interval $t_{i-1} \le t \le t_i$, for $i = 1, \dots, n$. The **knots** associated with a cubic spline fit are the points $t_0, t_1, \dots, t_n$. In our paper, the **degrees of freedom** associated with a cubic spline fit is a linear function of the number of knots considered.

These cubic polynomials $S_0(t), S_1(t), \dots, S_{n-1}(t)$ are required to satisfy the following requirements:

- $S_{i-1}(t_i) = S_i(t_i)$ for $i = 1, \dots, n-1$
- $S'_{i-1}(t_i) = S'_i(t_i)$ for $i = 1, \dots, n-1$
- $S''_{i-1}(t_i) = S''_i(t_i)$ for $i = 1, \dots, n-1$
- $S''_0(t_0) = S''_{n-1}(t_n) = 0$
- $S_{i-1}(t_{i-1}) = x_{t_{i-1}}$ for $i = 1, \dots, n$
- $S_{n-1}(t_n) = x_{t_n}$.

In our paper, the time series $\{x_{t_0}, x_{t_1}, \dots, x_{t_n}\}$ would be the 3-week lagged total rainfall values. Thus, a natural cubic spline is fit to the historical total rainfall data.

3.2 Cross-Validation in Time Series Data

Time series data pose a unique challenge to the formation of training sets. Training data sets can only be subsetted in subsequential time frames (not randomized subsets). This is due to the sequential nature of the data, and the need to avoid creating missing values at random time points.

In this paper, we have data at 65 data points, of which the first 45 are always used for training. As an example, we train the model on the first 1, 2,..., 45 time points. We then predict for time points 46, 47, 48,..., 63, 64, and 65. This process is repeated, each time increasing the training set by 1 time point. Finally, we train on 1, 2,..., 61, 62 time points and predict for time points 63, 64, and 65. Figures 4 and 5 show two examples of the Time Series Training Sets used in this paper.

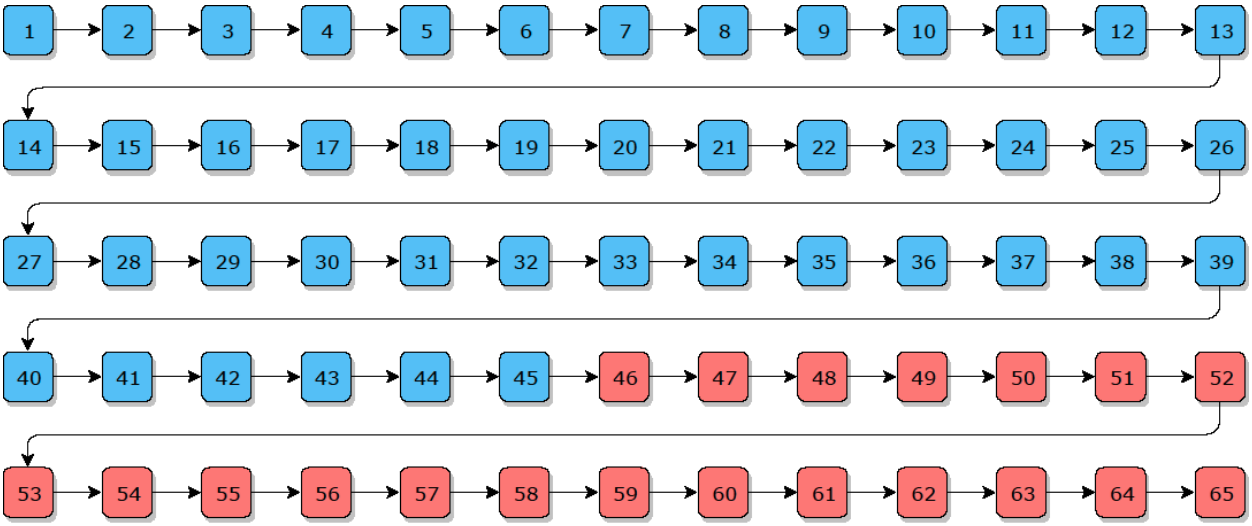

Figure 4: Using data from time points 1 through 45 as training data for a model, and using the model to predict DF counts at time points 46 through 65.

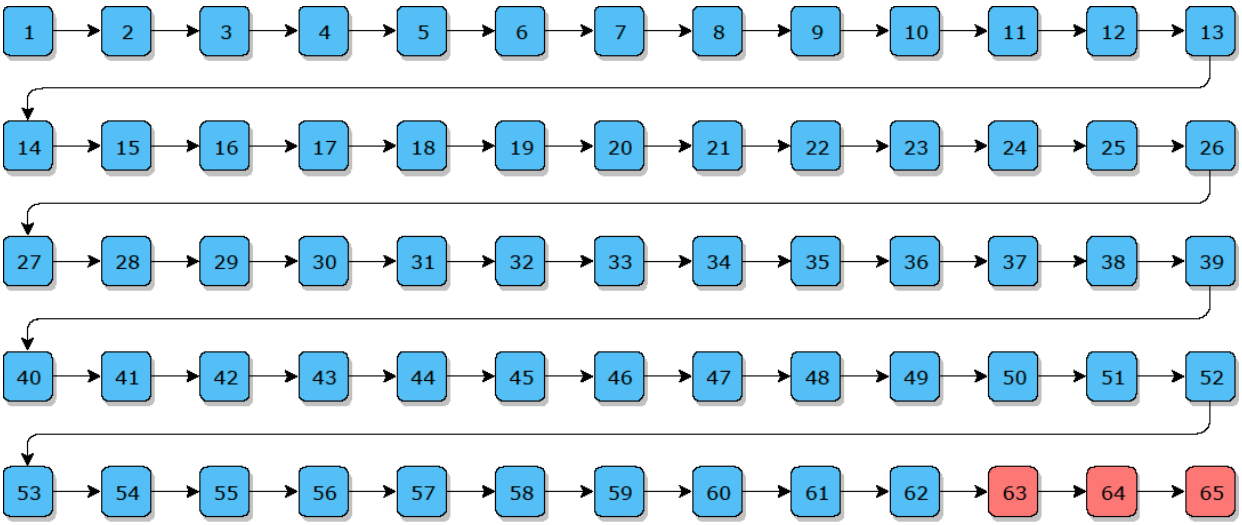

Figure 5: Using data from time points 1 through 62 as training data for a model, and using the model to predict DF counts at time points 63 through 65.

4. Natural Cubic Spline Model 1

The first natural cubic spline model we consider (Model 1) has the form

$$y_t = ns(x_{t-3}, m) + \varepsilon_t,$$

where y_t is the Dengue Fever count for week t, x_{t-3} is the total rainfall for week $t-3$, $ns(.)$ is the natural cubic spline function fit to the $\{x_{t-3}\}$ time series data with degrees of freedom m, and the ϵ_t are identically and independently distributed (iid) $N(0, \sigma^2)$ random variables.

Figure 6 shows an example of the actual and predicted Dengue Fever counts when using Model 1 with 10 degrees of freedom. The ns function in the splines R package was used for model fitting (see [4]).

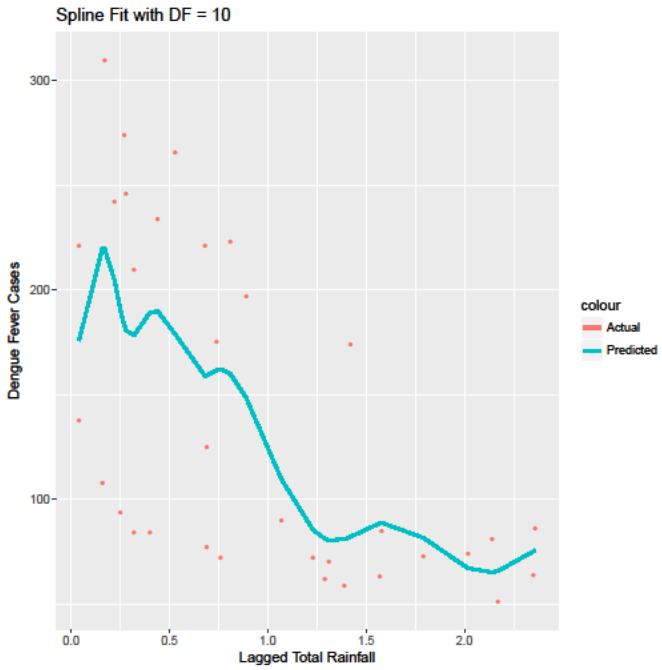

Figure 6: Actual and predicted Dengue Fever counts when using Model 1 with 10 degrees of freedom.

5. A Metric for Prediction Accuracy and choosing the optimal Degrees of Freedom for Model 1

The following metric for prediction accuracy across the different training sets was defined:

$$MSE_{T,P} = \frac{1}{P}\Sigma_{i=1}^{P}(y_{T+i} - \hat{y}_{T+i})^2,$$

for $T = 46, 47, \ldots, 62$ and P is the number of required predictions (i.e. at time points $T+1, \ldots, T+P$).

For each training set, the optimal degrees of freedom in the natural cubic spline model (in the sense of minimizing $MSE_{T,P}$) was determined. Table 1 shows these optimal degrees of freedom, indexing the results by T (the time point at which prediction begins for a

particular training set) and prednum, *P* (the number of timepoints after the final time point in a training set) for which predictions are made by Model 1.

We note that the maximum degrees of freedom we considered was 20, as a spline fit with a high number of knots increases the likelihood of an overfit model. Also, in an actual prediction scenario, predicting for a higher number of timepoints than we have rainfall data for would be facilitated by using historical data from the same time the previous year as surrogate values.

Table 1: Optimal degrees of freedom for Model 1 (natural cubic spline fit on 3-week lagged rainfall) based on training set (indexed by *T*) and number of predictions, *P*. (Please note that prednum = *P*)

Prednum	T=46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65
1	4	11	1	11	12	7	8	1	9	6	12	19	6	20	12	20	13	2	17	1
2	4	1	1	11	20	7	1	1	7	7	19	19	20	16	12	19	19	1	1	
3	20	1	1	8	15	1	1	1	9	12	19	20	16	8	13	19	17	1		
4	1	20	1	8	9	1	1	1	9	12	19	16	8	20	11	19	17			
5	20	20	1	8	1	1	1	1	9	9	16	8	20	20	12	19				
6	20	20	1	1	1	1	1	1	9	12	19	20	20	20	20					

Some conclusions we come to from Table 1:

- As the training set gets bigger (more time points), the optimal degrees of freedom increases (in many cases to the maximum we considered, which was 20). This indicates the need for an increasingly flexible spline fit for increased forecast accuracy as the number of time points used for training the model increases.
- For any value of *T*, there is more agreement on the optimal degrees of freedom as the value of *P* increases. This is unsurprising as the predictions at one time point and the next time point are correlated.

6. Natural Cubic Spline Model 2

The second natural cubic spline model we consider (Model 2) has the form

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + ns(x_{t-3}, m) + \varepsilon_t,$$

where y_t is the Dengue Fever count for week *t*, y_{t-1} is the Dengue Fever count for week $t-1$, x_{t-3} is the total rainfall for week $t-3$, α_0 and α_1 are fixed constants, *ns*(.) is the natural cubic spline function fit to the $\{x_{t-3}\}$ time series data with degrees of freedom *m*, and the ϵ_t are identically and independently distributed (iid) $N(0, \sigma^2)$ random variables.

Figure 7 shows a plot of the Dengue Fever count vs a 1-week lagged Dengue Fever count. The positive correlation between these variables justified the inclusion of the 1-week lagged Dengue Fever count in Model 2.

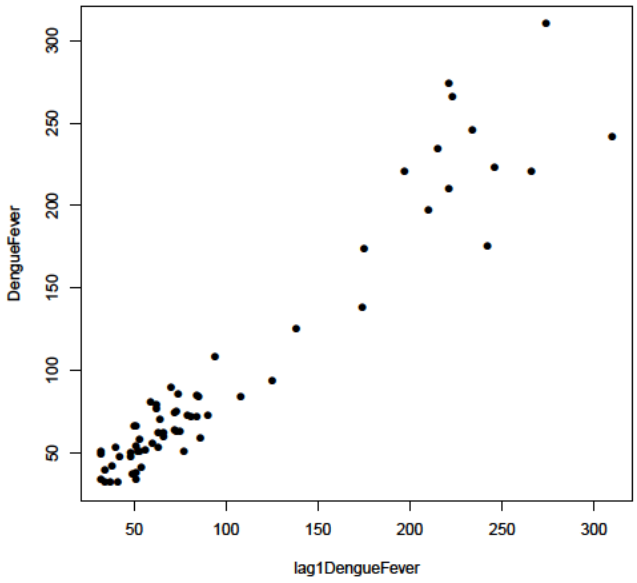

Figure 7: Dengue Fever count vs a 1-week lagged Dengue Fever count.

7. Choosing the Degrees of Freedom for Model 2

For each training set, the optimal degrees of freedom in the natural cubic spline model (in the sense of minimizing $MSE_{T,P}$) was determined. Table 2 shows these optimal degrees of freedom, indexing the results by T (the time point at which prediction begins for a particular training set) and prednum, P (the number of timepoints after the final time point in a training set) for which predictions are made by Model 2.

Table 2: Optimal degrees of freedom for Model 2 (which included a 1-lagged Dengue Fever count and a natural cubic spline fit on 3-week lagged rainfall) based on training set (indexed by T) and number of predictions, P. (Please note that prednum = P)

Prednum	T=46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65
1	1	19	1	20	17	17	12	1	1	6	14	6	6	20	1	18	13	17	20	1
2	1	1	1	20	16	18	1	1	1	13	7	6	9	1	1	10	13	15	4	
3	1	1	1	20	17	1	1	1	1	7	8	10	1	1	11	10	13	2		
4	1	1	1	17	1	1	1	1	1	7	9	1	1	11	11	10	13			
5	1	18	1	1	1	1	1	1	1	8	1	1	11	11	11	19				
6	20	1	1	1	1	1	1	1	1	7	1	11	11	11	1					

Some conclusions we come to from Table 2:

- As the training set gets bigger (more time points), the optimal degrees of freedom increases. We note a difference in Model 2 here, where the optimal number of degrees is around 10 (the optimal degrees of freedom in the large training set scenario for Model 1 was around the maximum we considered, which was 20). This indicates the need for an increasingly flexible spline fit for increased forecast accuracy as the number of time points used for training the model is lessened by the inclusion of the 1-week lagged Dengue Fever count in Model 2.

- As was the case in Model 1, for any value of *T*, there is more agreement on the optimal degrees of freedom as the value of *P* increases.

8. Summary

- Cubic spline fits provide the availability of a flexible polynomial fit to time series data, without the need for high orders.
- Historical rainfall data, when lagged appropriately, can be helpful as a predictor variable for Dengue Fever counts.
- Cross-validation carried out this way is a useful method for identifying the optimal degrees of freedom for increased prediction accuracy. This cross-validation method takes into account how much data is available and how many time points you want to predict ahead.
- The higher the number of time points at which predictions are required, there is more agreement on the optimal degrees of freedom for the spline fit.
- There is a shift in the optimal degrees of freedom for the spline fit depending on the time point at which prediction begins.

Acknowledgements

The authors would like to thank NASA and the Michigan Space Grant Consortium (MSGC) for their support.

References

[1] https://www.moh.gov.sg/resources-statistics?type=statistics (accessed 23 September 2018)
[2] Koh Y-M, Spindler R, Sandgren M, Jiang J. A model comparison algorithm for increased forecast accuracy of dengue fever incidence in Singapore and the auxiliary role of total precipitation information,
Int J Environ Health Res. 2018 Oct; 28(5):535-552
[3] https://weather.com/weather/monthly/l/SNXX0006:1:SN (accessed 23 September 2018)
[4] Splines R package
https://www.rdocumentation.org/packages/splines/versions/3.5.1 (accessed 23 September 2018)