Fitting models to realistically simulated time-series data and evaluating their performance using cross validation

Aaryan Bhatia, supervised by August Hao and Jennifer Flegg

2023/2024 Mathematics and Statistics Vacation Scholarship Program, The University of Melbourne

Introduction

Time-series data are often of interest in real-life applications –for example, include the annual amount of carbon dioxide produced by Australia or the month-on-month value of local residential buildings. In applications like these, there is interest in using statistical models for inference or prediction.

However, working with real time-series data is often nuanced and complex:

- Influencing variables may change gradually or suddenly
- Imperfections in calculating quantities of interest arise from method changes, biases, or missing data.
- Fitting models to time-series data is complex, requiring flexibility in how the response is modelled over time. However, flexible statistical and machine learning models are prone to overfitting to complex time-series patterns.

The above challenges highlight two key tasks in the development of modelling methods suitable for time-series data:

1. Simulation of time-series data that flexibly exhibit behaviours like step changes, sampling bias, and inconsistent sampling.
2. Development of a cross-validation approach for evaluating interpolation performance and understanding model limitations.

The work presented in this poster details the construction of frameworks used to address these goals.

Simulation Methodology

We simulate daily counts of close contacts in a hypothetical scenario, often used in infectious disease management.

- 1,066 days of contacts data are simulated within the hypothetical context, representing every day from January 1st, 2020, to December 31st, 2022.
- The expected number of contacts for a given day is calculated as the product of several mechanistic factors with varying levels of significance.
- Each day, 10 samples are taken from a Poisson distribution with an expected value given by the above product. This is summarised in equation 1.

Factor	Minimum value	Maximum value
Random uniform base	10.0	12.0
Public Holiday	1.00	1.15
School Holiday	1.00	1.15
Week date	0.90	1.15
Month	0.99	1.01
Lockdown	0.30	1.00
Week following lockdown	1.00	1.20
Average Temperature	0.90	1.10
Possible Variance	**2.41**	**24.33**

Table 1: The factors affecting daily contact expectations. Temperature follows a sinusoidal pattern, peaking in January. Lockdown and holiday data are based on Melbourne's history, as discrete variables that take only minimum or maximum values

$$X_{ij} \sim \text{Pois}\left(\prod_k f_k(i)\right)$$

Equation 1: The sampling distribution of X_{ij}, the jth sample of contact observations on the ith day. f_k is the kth factor of the above table.

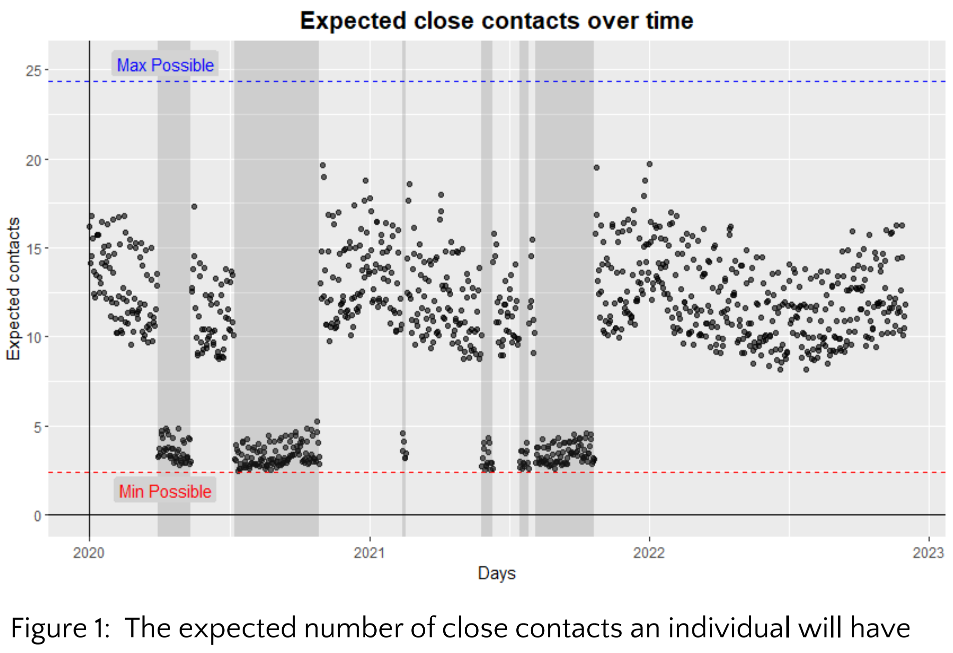

Figure 1: The expected number of close contacts an individual will have on a given day. The drop in number of contacts in the dark regions are due to lockdown restrictions.

Missing and Erroneous Data

The count data generated in the previous section depict a 'perfect' observation scenario. However, real-life time-series are imperfectly observed. To simulate an 'imperfect' process, the following modifications were made to the original data:

1. Maximum of 5 samples taken in the first 100 days:

This simulates low response rates of survey-based data, which, in context, is likely in the early stages of the survey effort with lower public interest and awareness

2. Samples with more than 20 close contacts have an 50% chance of being undercounted:

In context, this represents individuals with many close contacts who might misremember the exact number of contacts. Undercounting happens by subtracting an integer from 1 to 5.

3. 10% fewer samples in days that are not weekends or public holidays:

Represents a baseline imperfect response rate (i.e. the survey is unable to fill its response quota). 10% of all samples are removed from data at random.

4. 60% fewer samples in weekends and public holidays:

Represents potential effect of non-working days on survey response rate.

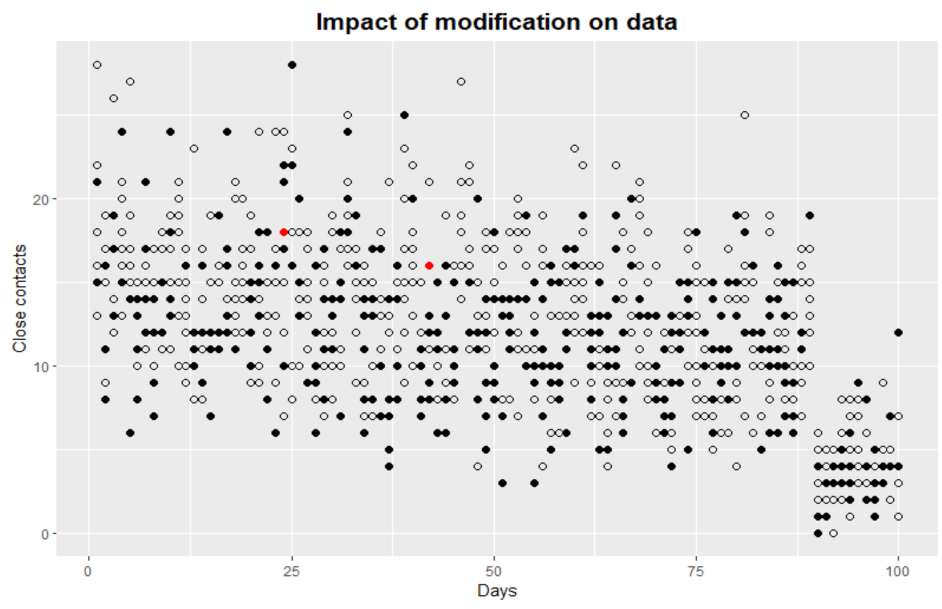

Figure 2: The impact of the removal of data and creation of bias in the first 100 days of data – full black circles are true observations recorded, empty ones are missed. Full red circles represent biased (undercounted) observations.

Cross-validation Methodology

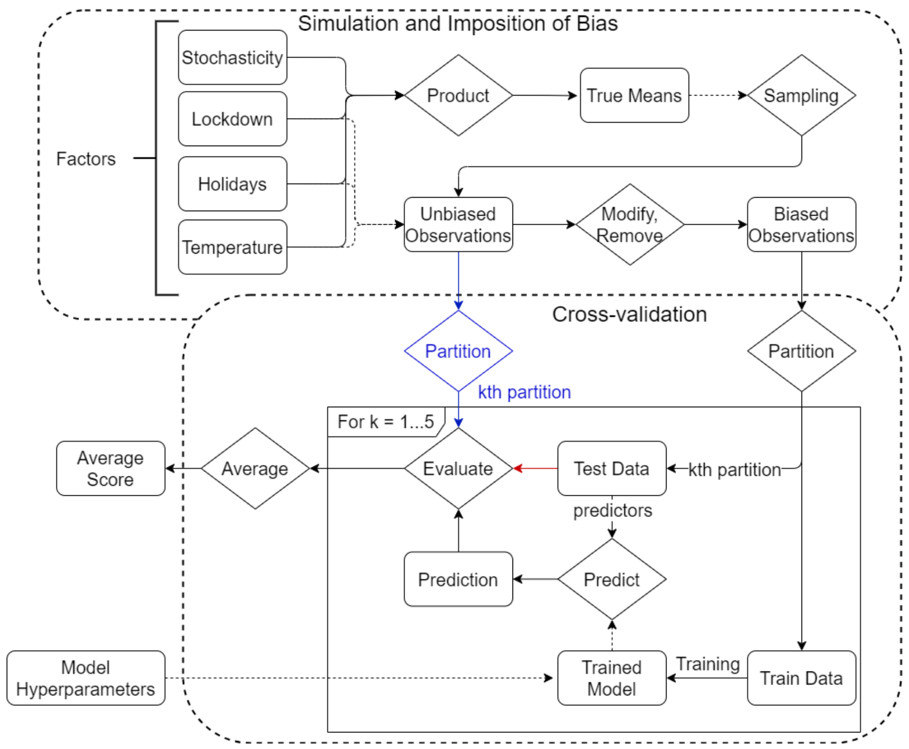

Figure 3: A schematic that shows the steps involved with data simulation and cross-validation. Rounded rectangles represent pieces of data while diamonds represent the operations on data. The red and blue arrows represent alternative paths of evaluation.

k-fold cross validation is a technique used in machine learning to provide a robust evaluation of model performance on unseen data[3].

- Training data is divided into k parts for iterative model evaluation. Each time, one part is kept aside for testing, and the rest is used for training.
- To account for the time structure of our data, partitioning was done in continuous time intervals[1].
- Two evaluation methods were considered: comparing model predictions to withheld training data or to the unbiased data generated for the project.
  - The latter helps understand how models trained on biased data would perform if tested with "correct" data later.

Cross-validation Example

While generalized additive models (GAMs) are demonstrated here for testing the cross-validation framework, the framework can be applied to any statistical or machine learning model for time-series data. That said, I found invaluable guidance on GAMs from Simpson's recorded webinar[2].

The following models were tested:

1. A GAM that fits contacts based on lockdown presence and time. The model is highly non-parametric, with an upper limit of 300 degrees of freedom to make up for lack of other information.
2. Another GAM that considers additional factors like holiday, week date, and month, without prior knowledge of their impact on expected number of contacts. It has 100 degrees of freedom.
3. A benchmark model that returns the time-constant mean number of contacts in the training data.

During each cross-validation iteration, the models were evaluated using the root mean square error (RMSE) of prediction values compared to unbiased and biased observations.

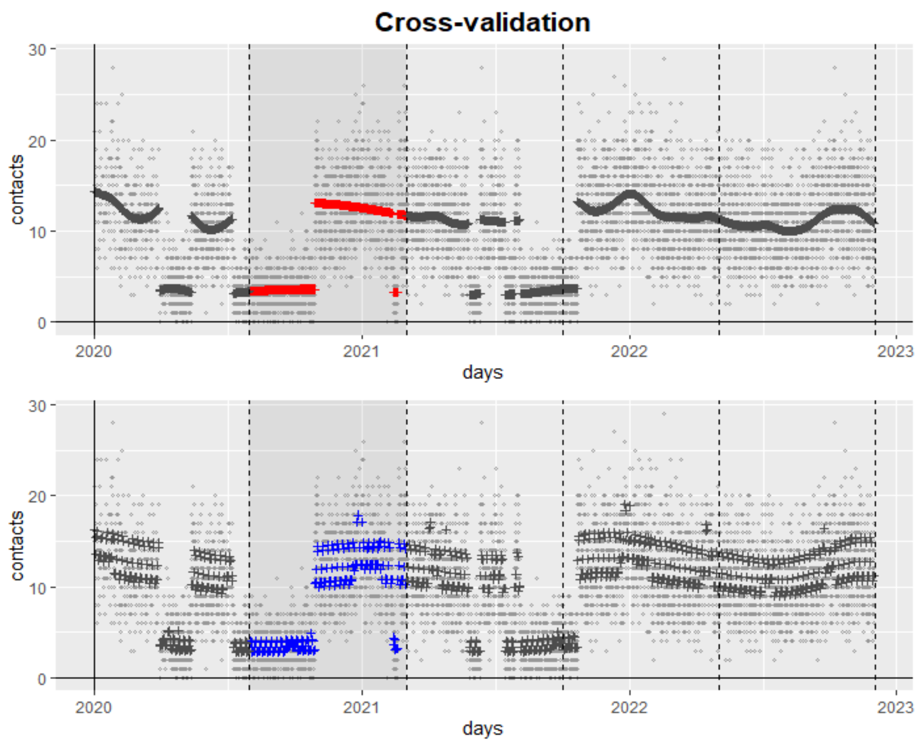

Figure 4: An illustration of $k = 5$ cross-validation. The second partition (darker shade) has been withheld for testing. The small grey circles depict biased observations, while the red/blue crosses show the predictions of the first and second GAMs in the second partition, respectively.

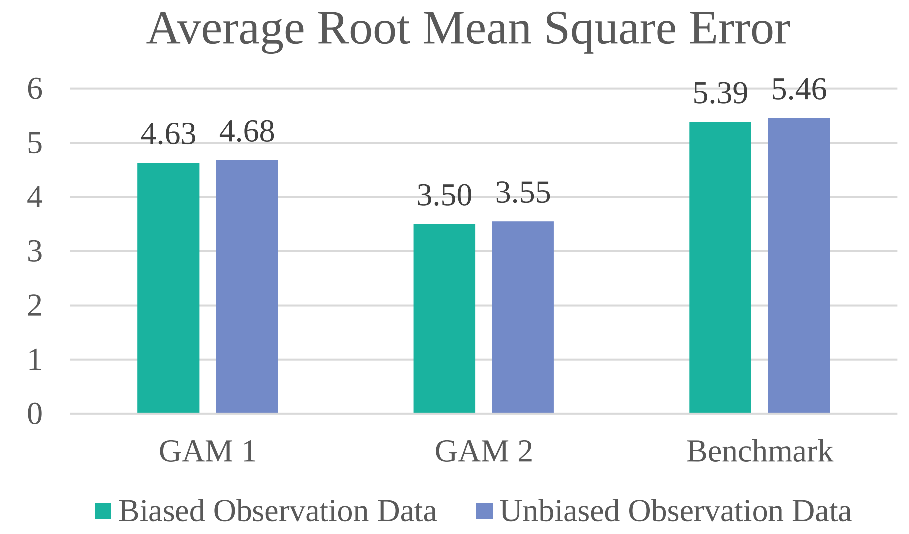

Figure 5: The average RMSE of the above models.

The second GAM model had the lowest error, indicating that the model fitted with more information was less susceptible to overfitting. Testing on biased/unbiased data yield similar results.

Extensions

The frameworks and codes in this model are designed to be highly adaptable, making this poster's work an excellent tool for model evaluation. Extensions encompass:

- Employing an alternative sampling distribution; for instance, a negative binomial distribution accommodates larger variances than feasible with a Poisson distribution.
- Exploring various factor types using real-world data to justify their influence on distribution parameters.
- Incorporating different forms of bias or missing data to better reflect real-world scenarios.
- Cross-validating GAMs with diverse hyperparameters or entirely different statistical and machine learning models.
- Comparing the cross-validation performance of models on simulated data to real-world samples

References

[1] Roberts, DR, Bahn, V & Ciuti, S 2016, 'Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure', in M Araújo (ed.), Ecography, vol. 40, no. 8.

[2] Simpson, G 2020, Introduction to Generalized Additive Models with R and mgcv, www.youtube.com, viewed 6 November 2023, <https://www.youtube.com/watch?v=sgw4cu8hrZM&t=1332s>.

[3] Wikipedia Contributors 2019, Cross-validation (statistics), Wikipedia, Wikimedia Foundation, viewed 6 February 2024, <https://en.wikipedia.org/wiki/Cross-validation_(statistics)>.