Generalized additive time series parametrization and modelling with smoothing term controls and mechanistic predictors

Luoman (Mary) Huang, supervised by Dr. August Hao and Prof. Jennifer Flegg

2023/2024 Mathematics and Statistics Vacation Scholarships Program

Introduction

Generalized Additive Models (GAMs) are a well-established tool for time series analysis, known for their flexibility in capturing intricate and sporadic temporal patterns. This work focuses on the appropriate application of GAMs for time-series analysis. Our goal is to rigorously test how GAMs are best calibrated for complex time-series count data, examining the impact of different mechanistic predictors and levels of rigidity in the smoothing of response over time. In navigating these complexities, we aim to gain insight to important decisions and strategies in using GAMs for time-series analysis.

Method & model

Context of simulated data: daily close contacts

We explore GAMs with a hypothetical dataset of close contact numbers surveyd by an infectious disease surveillance team. The response variable is "Number of Contacts," with varies with "Days," but is also affected by "Lockdown," "Public Holiday," "School Holiday," "Weekend," "Month," and "Year."

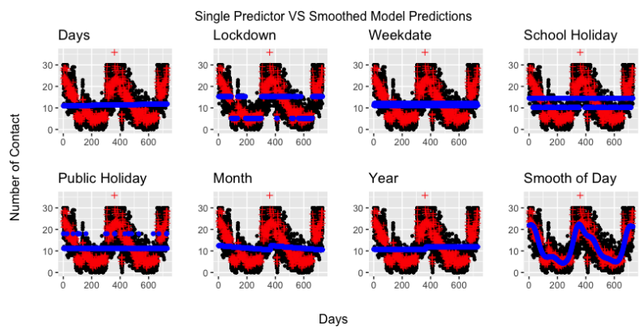

Fig. 1: Single predictor model predictions (without smooth terms VS with simple smooth of day)

Controlling smoothing terms for flexibility & model robustness:

GAMs introduce smooth terms, a key feature allowing flexible response shapes with respect to a predictor. Figure 1 visually contrasts models with and without smooth terms over the "Days" predictor. Strategic adjustment of the number of knots for smooth terms crucial for model accuracy and resilience. Figure 3 vividly illustrates the impact of varying knots on model performance.

Leveraging categorical predictors:

Our methodology utilizes categorical variables beyond step changes; we harness them as valuable mechanistic predictors. Events like lockdowns and school holidays offer crucial insights into underlying data patterns. This approach allows for a nuanced strategy — identifying these events enables the use of more rigid smoothing terms for the time-since-start predictor without compromising overall model fit. This efficient use of mechanistic predictors enhances interpretability and predictive accuracy (Figure 4) and contrasts those without categorical mechanisms' flexible but lower predictive accuracy (Figure 2), unraveling intricate patterns within the time series data. (Figure 3)

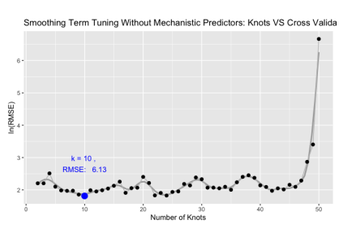

Fig. 2: Smooth day only model, without any mechanistic predictors: Knots tuning and resulted RMSE trend

These are some methods that modellers can use to model their GAMs. However, ultimately we use cross-validation to choose the best parameter inputs.

Results

Our exploration into model parameterization, encompassing various predictors and smoothing settings in GAMs, extends beyond knots, providing insights into diverse parameterization strategies via cross-validation.

Parameterization strategies

- **Predictors Variation:** Inclusion of diverse predictors adds complexity, capturing underlying patterns.
- **Smoothing Settings:** Altering settings, notably knot count, is critical for refining model, to avoid underfitting or overfitting.

Performance metrics

To gauge efficacy, we used both RMSE and % Deviance Explained.

- **RMSE:** Indicator of predictive accuracy.
- **% Deviance Explained:** Measures the proportion of variability in the response variable that is accounted for by the model.

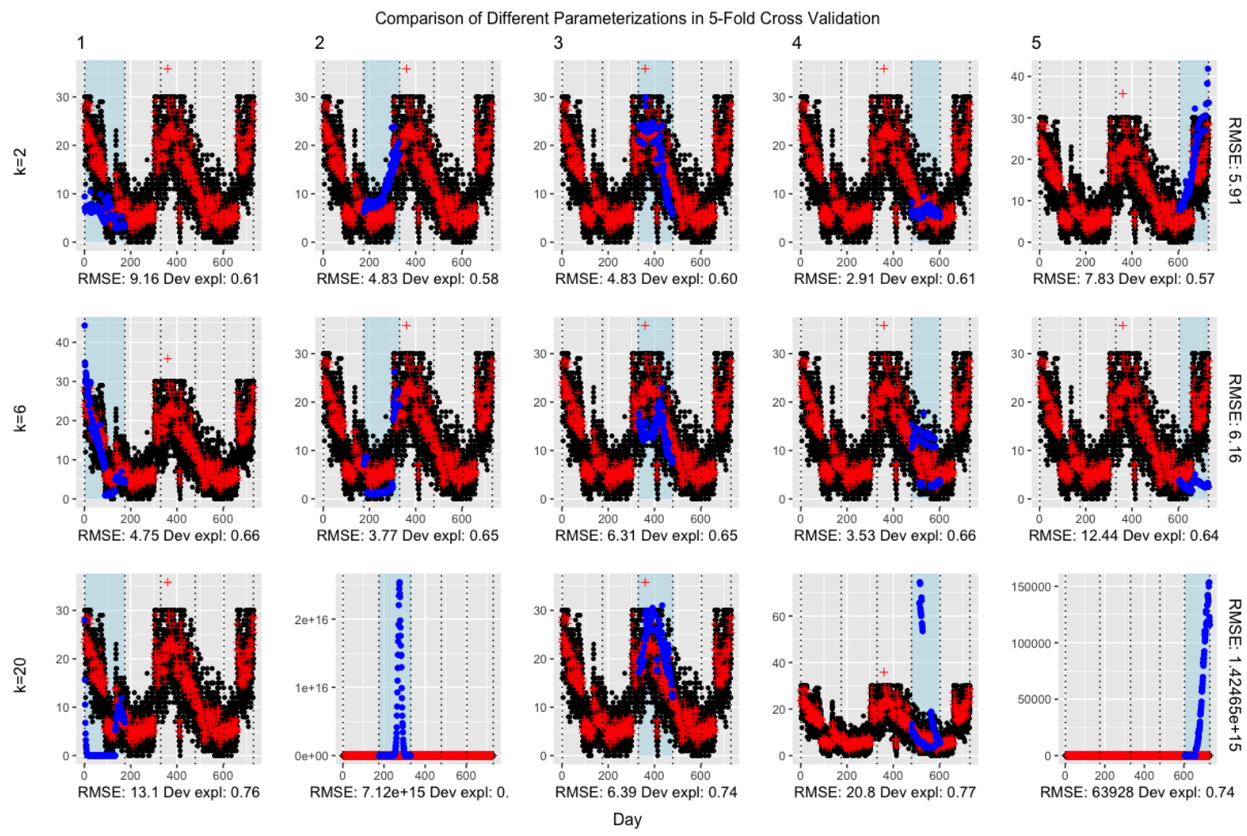

Fig. 3: Comparison of different parameterizations in 5-fold cross validation, where the dataset is split into 5 even-bin subsets for iterative training (4 bins), testing (1 bin), and evaluation. Data points are represented as follows: Black = Simulated data collected; Red = True mean; Blue = Model prediction. In the bottom row, models with smoothing terms of 20 knots exhibit a common pattern of spurious high values in the testing folds, suggesting potential overfitting to the training data due to excessive model complexity leading to poor generalization.

Balancing trade-offs through cross validation

Our approach involves systematically testing different parameterizations using cross-validation. This not only identifies the optimal model complexity but also ensures stability and robustness by assessing the model across various parameter configurations. For instance, through systematic cross validation of the GAM, k=6 visibly demonstrates the best predictive accuracy as well as a good robustness. (Figure 4)

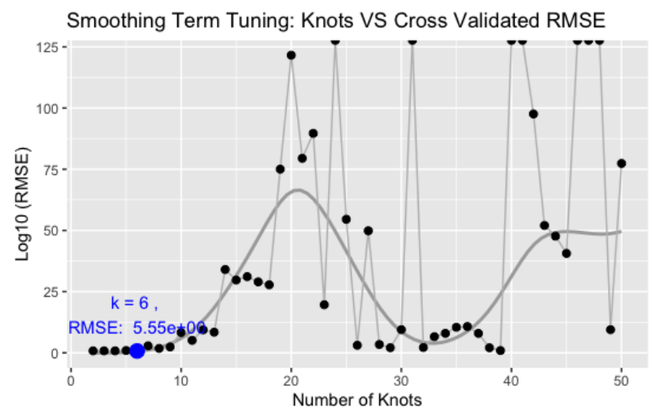

Fig. 4: Tuned model with mechanistic predictors: knots tuning and resulting RMSE trend. The current model, incorporating optimized mechanistic features, outperforms the simpler model lacking such predictors. However, its optimal performance is attained with fewer knots for the smoothing term, as evidenced by the low RMSE at k=6 (optimal performance) and other viable candidates with k<6. Conversely, more knots generally yields extremely poor performance in the testing fold (Figure 3). In contrast, the simpler model accommodates a higher number of knots before encountering overfitting issues, typically after approximately knots >= 40, resulting in artificially high RMSE values (Figure 2).

Remarks

This study highlights the nuanced nature of fitting GAMs to time series data, considering factors like mechanistic inputs and smoothing term flexibility. While we present an optimization framework, it's important to acknowledge inherent methodological limitations. Simulating data with known mechanisms simplifies tasks but differs from real-world challenges. Identifying suitable mechanistic predictors requires domain expertise. Our model's success depends on context and may not universally apply. Recognizing these challenges underscores the need for refinement in diverse scenarios.

Acknowledgements

Thanks to supervisors August Hao and Jennifer Fleggs for invaluable guidance, to colleague Aaryan Bhatia for contributions to data simulation, cross-validation.

References

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. CRC Press, Boca Raton, FL, USA.

Wright, D. B., & London, K. (2009). *Modern Regression Techniques Using R: A Practical Guide*. SAGE Publications, Limited.