STAT 6550 Fall 2021

Final Project Proposal

Matthew Lister Jake Rhodes

Assignment due: Nov 17

1. **Dataset**

   Our dataset was produced by an AI using DPIV (cite) to count the level of bee activity outside a hive marked R_4_5. These csv files are part of the output of Dr. Sarbajit Mukherjee during his dissertation at Utah State University. We have received permission of Dr. Kulyukin to use these files for time series analysis. The counts correspond to a 28 second period recorded approximately every 15 minutes throughout the day. The monitors shutdown at night and went off line several times throughout the season leaving significant coverage gaps.

2. **Proposed Analysis**

   We have a non uniform time series with multiple values to impute. We believe the data can be shown to have a daily auto-regressive pattern with some degree of seasonality. Furthermore, weather variables can be obtained from USU weather station and can be combined with the raw count data to produce time series models with environmental predictors.

3. **In what way would R functions, organized in an R package, aid in the analysis of the selected dataset?**

   There is no single source R package that combines multiple time series techniques into one convenient package. Current packages tend to focus on a single technique and seldom make any mention of other analysis pipelines.

4. **What challenges do you foresee in completing this analysis? How do you intend to address these challenges?**

   The ideal would be approach the predictive accuracy of facebook's prophet [cite], which, like other modern AR techniques, makes little mention or use of the other methods in the field. While it is unlikely we can reach that ideal, progress can be made by proceeding from the simplest models into more complex and recent additions to the field and incorporating the knowledge and experience gained into later experiments.

5. **In what way would your analysis be of interest to the other students in the class?**

   We feel there is a barrier among our peers to attempting time series analysis with real world data. That barrier is cause by missing data points, uneven time segments, and interpretability of the output. We hope, by following our vignette that the students will be able to remove some of that difficulty.