Modeling Growth Trajectories of Applied Statistics Fields: Public Interest, Degree Offerings, and Popular Sentiment

Nana Amma B. Asamoah and Samantha Robinson
Department of Mathematical Sciences, University of Arkansas

Background

Understanding that there is a rising demand for analytical skills due to the vast amount of data available, this study sought to explore whether there is a corresponding echo in the interest of applied statistics fields.

Objectives

- Examine time trends in google searches for Biostatistics and Data Science.
- Forecast the growth of degrees offered in these fields.
- Explore sentiment analysis that examines public attitudes about the fields.

Methods

Exponential Smoothing Method: Two different methods were considered in this study.

❶ Holt's Linear Method: This method allows forecasting of data with a trend.
- Forecast equation: $y_{t+h|t} = l_t + hb_t$
- Level equation: $l_t = \alpha y_t + (1 - \alpha)(l_{t-1} + b_{t-1})$
- Trend equation: $b_t = \beta^*(l_t - l_{t-1}) + (1 - \beta^* b_{t-1})$

❷ Holt's-Winters' Seasonal Method: This extends the above method to capture seasonality.
- Forecast equation: $y_{t+h|t} = (l_t + hb_t)s_{t+h-m(k+1)}$
- Level equation: $l_t = \alpha \frac{y_t}{s_{t-m}} + (1 - \alpha)(l_{t-1} + b_{t-1})$
- Trend equation: $b_t = \beta^*(l_t - l_{t-1}) + (1 - \beta^* b_{t-1})$
- Seasonal equation: $s_t = \gamma \frac{y_t}{l_{t-1} - b_{t-1}} + (1 - \gamma)s_{t-m}$

Sentiment Analysis Method: The following steps were used in exploring public attitudes and sentiments about Biostatistics and Data Science

- After Twitter authentication with R, tweets including search terms #biostatistics and #datascience were extracted from Twitter.
- The text of 5000 tweets per search term were then cleaned to remove URLS, usernames, spaces, extra whitespace, common stop words and nonsense words.
- Word frequency was inspected and wordclouds for each term were then constructed in R with word size allocated in accordance to the observed word frequency.
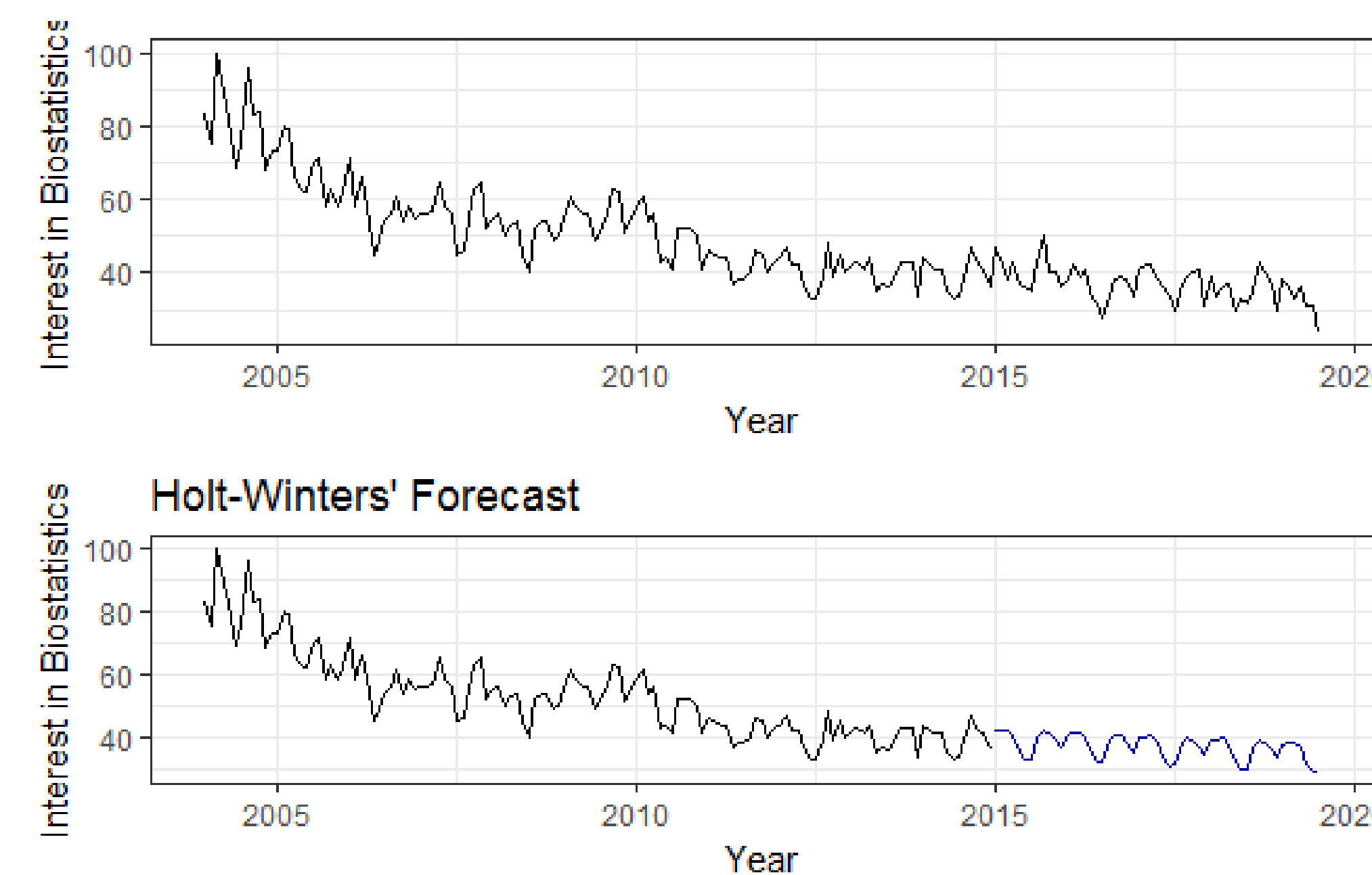
Results - Biostatistics

Figure 1: Public Interest in Biostatistics
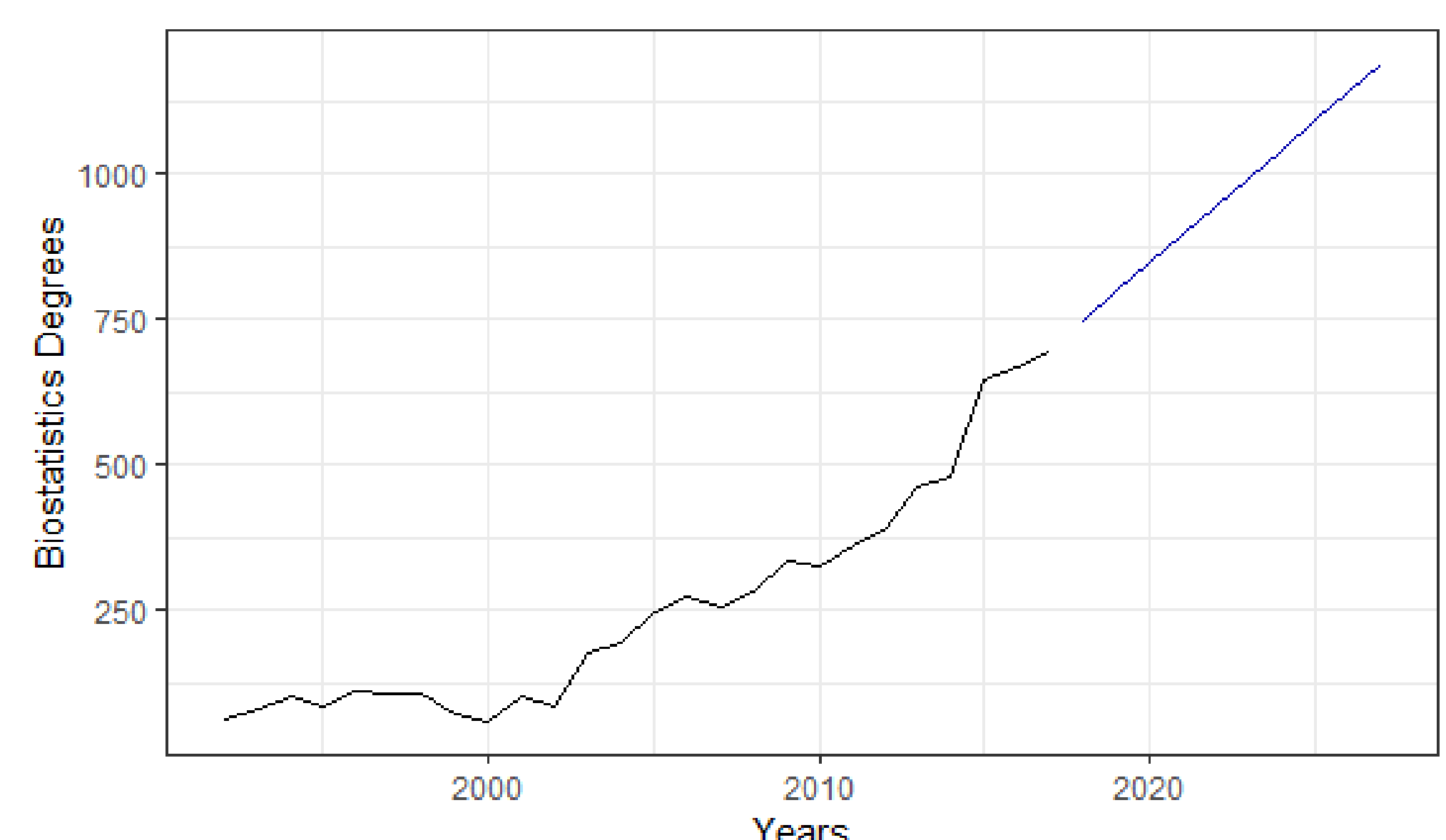

Figure 2: Biostatistics Degrees Awarded

Figure 3: Biostatistics Wordcloud
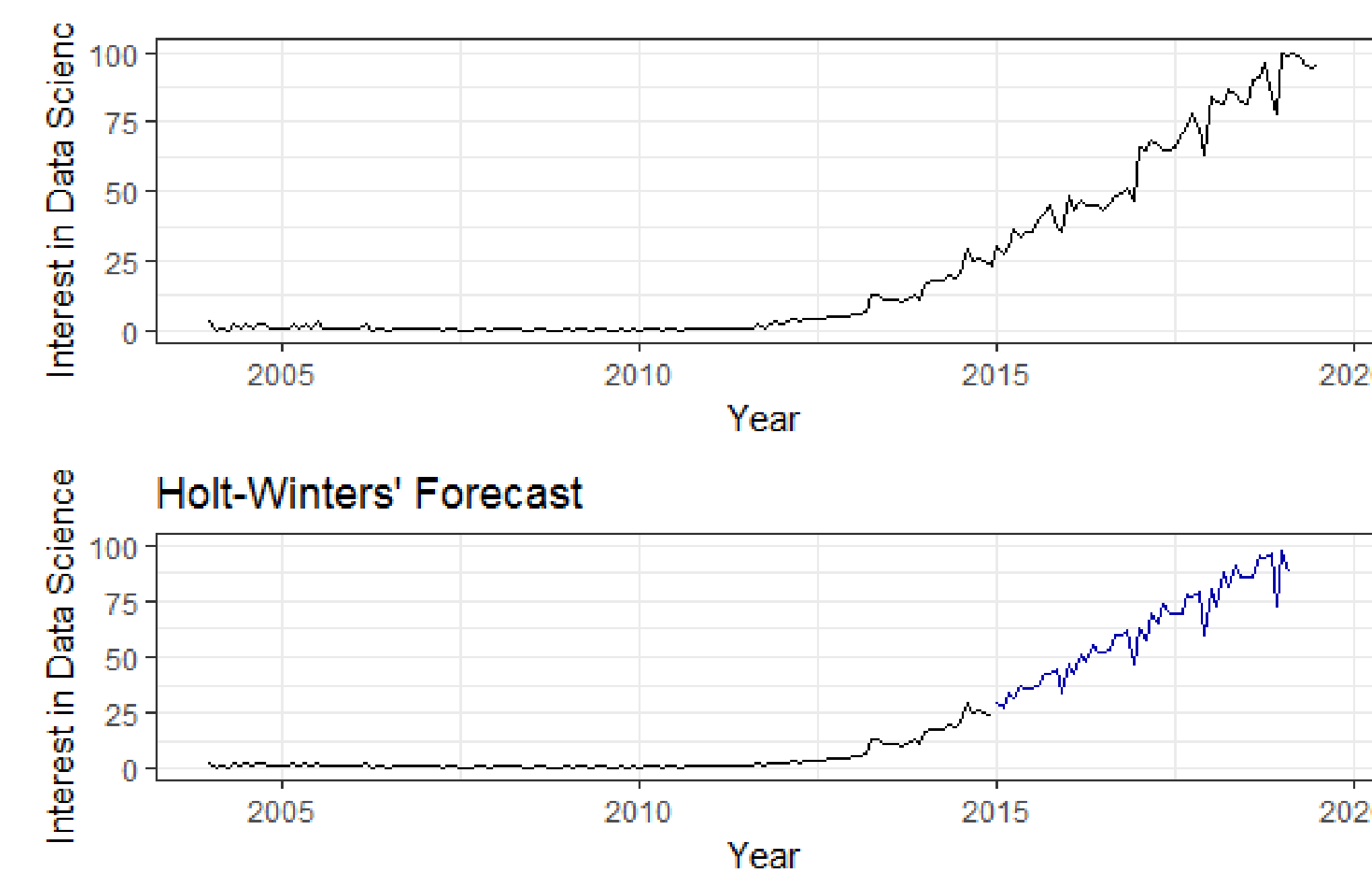
Results - Data Science
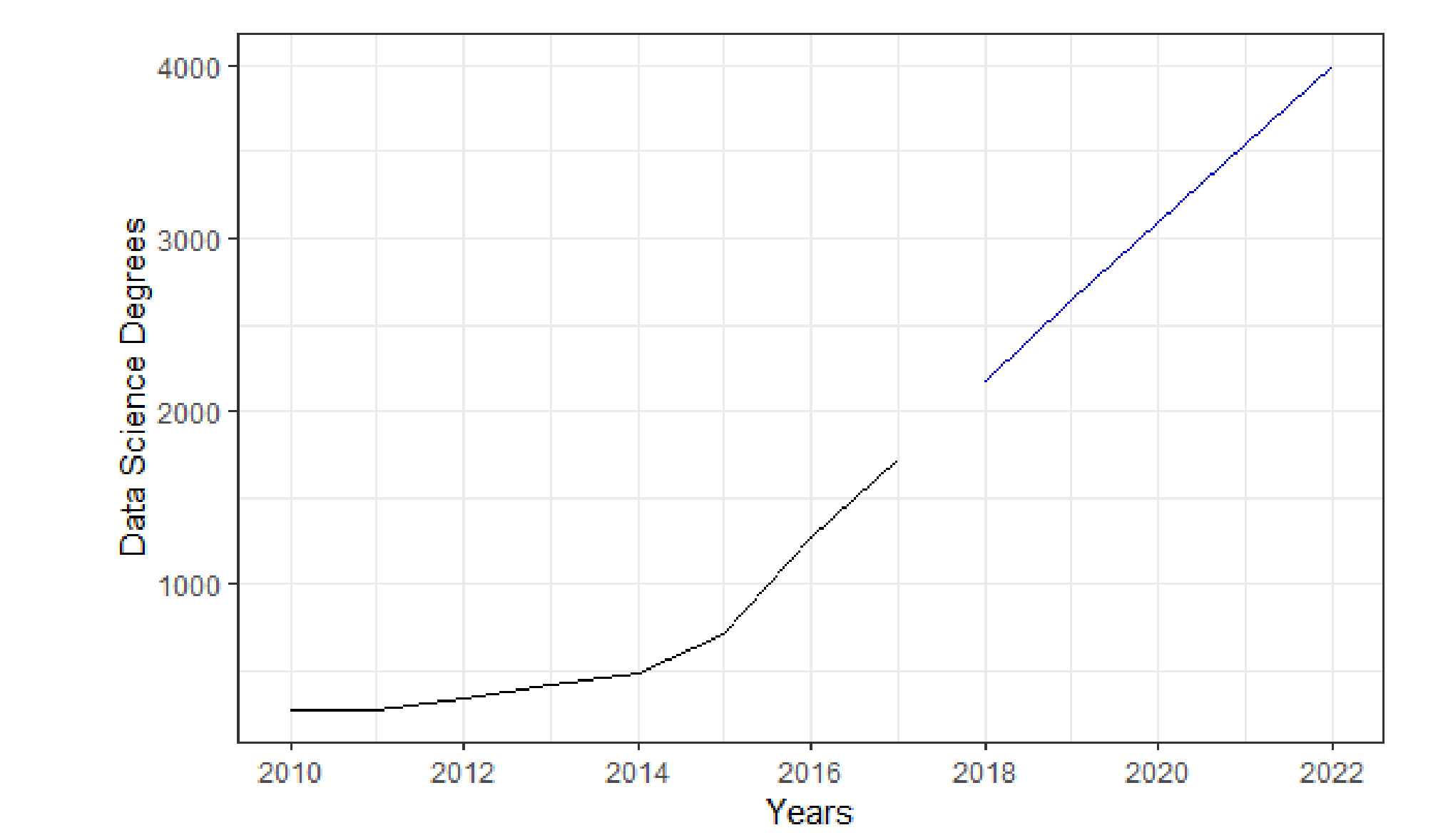

Figure 4: Public Interest in Data Science

Figure 5: Data Science Degrees Awarded

Figure 6: Data Science Wordcloud

Conclusion

❶ **Biostatistics**
- Since data was collected on google trends, Biostatistics has had a relatively higher interest in that regards. However, this has been decreasing over the years.
- Using data currently available, we predict that degree offerings in Biostatistics will increase over the next ten years.
- The output from the sentiment analysis run for Biostatistics show that some of the words that come up often when people tweet about Biostatistics are 'online courses', 'university', 'Hopkins', 'research', 'mph' and 'learning'. The nature of tweets related to Biostatistics are then academia oriented.

❷ **Data Science**
- At the start of 2004, we observe that searches in Data Science were not as significant, however it starts to pick up around 2014 and has been increasing since then. The Holt-Winters' forecast as a predictor is sufficiently adequate and matches the true values approximately well enough
- Forecasting five years ahead, we predict that the number of data science degrees awarded will continue to increase.
- Some of the terms that appear more frequently in tweets for Data Science are 'machine learning','AI', 'analytics', 'big data', 'iot', and 'python'. Tweets including data science make the more programming aspect of the field evident.
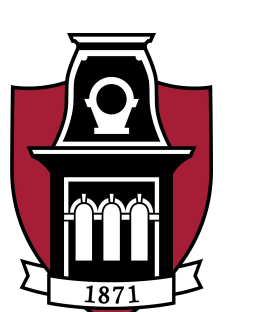
Data Sources

❶ Google Trends
❷ American Statistical Association
❸ Twitter

References

Hyndman, R.J., Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.

Contact Information

- Email: nbasamoa@uark.edu
- LinkedIn: linkedin.com/in/nab-asamoah/

UNIVERSITY OF ARKANSAS