For more information:

<http://www.shc.psu.edu/students/thesis/>

For thesis examples:

<https://honors.libraries.psu.edu/search/>

\*\*\*\*\***Do not type into the table of contents, list of figures, or list of tables.** Update/complete your work throughout the thesis and then return to the list of tables, figures, or TOC and right click to update the fields\*\*\*\*\*

All yellow highlighted text, including this page, should be replaced or deleted before uploading your thesis.

THE PENNSYLVANIA STATE UNIVERSITY

SCHREYER HONORS COLLEGE

DEPARTMENT OF STATISTICS

PREDICTING THE OUTCOME OF THE FINAL SIXTEEN TEAMS IN COLLEGE BASKETBALL USING TIME SERIES ANALYSIS AND MARKOV CHAINS

CHRISTOPER KENNETH MILLER

SPRING 2023

A thesis

submitted in partial fulfillment

of the requirements

for baccalaureate degrees

in Statistics and Mathematics

with honors in Statistics

Reviewed and approved\* by the following:

Faculty Name

Faculty Title (Professor of ?)

Thesis Supervisor

Faculty Name

Faculty Title (Professor of ?)

Honors Adviser

\* Electronic approvals are on file.

ABSTRACT

Start typing or insert file here ...(double space)

TABLE OF CONTENTS

LIST OF FIGURES [Edit these first 3 lines manually as needed, delete this text] iii

LIST OF TABLES iv

ACKNOWLEDGEMENTS v

[The rest of the TOC is a "field" which can be updated: right-click on it, choose "update field," choose "update entire table." Delete these two lines of text.]

[Chapter 1 Introduction 6](#_Toc126519112)

[Motivation and Overview 8](#_Toc126519113)

[Chapter 2 Efficiency 9](#_Toc126519114)

[Chapter 3 Data Considerations and Setup 11](#_Toc126519115)

[Chapter 4 Time-Series Model 13](#_Toc126519116)

[Chapter 5 Markov Chain to Create Ranks 14](#_Toc126519117)

[Chapter 6 Analysis of Results 15](#_Toc126519118)

[Chapter 7 Limitations and Conclusions 15](#_Toc126519119)

[Appendix A Replace with Appendix Title 16](#_Toc126519120)

LIST OF FIGURES

[The List of Figures is a "field" which can be updated: right-click on it, choose "update field," choose "update entire table." Delete these two lines of text.]

[Figure 1. Atherton Hall, SHC Est. 1997 1](#_Toc29892751)

LIST OF TABLES

[The List of Tables is a "field" which can be updated: right-click on it, choose "update field," choose "update entire table." Delete these two lines of text.]

[Table 1. Contact List for Experiment 2](#_Toc29892752)

ACKNOWLEDGEMENTS

# Introduction

National Collegiate Athletic Association (NCAA) Division I men’s basketball is a widely publicized sport in the United States, and like many other sports, the use of statistics to make some informed decision has become relevant for the league, analysts, coaches, and even the casual fan. The validity and usefulness of certain statistically backed reasoning can be questioned and explored but it is undeniable that number-backed decisions provide concreteness to any given conclusion. In the eyes of analysis, this sport provides an extra level of difficulty due to the collegiate aspect. In a traditional professional league, you can note some small material differences between teams, for example, payroll size, organization location, owner investment strategies, but overall, you can do analysis with the assumption that the professional teams are all on the same level of fairness. When we look at the collegiate level there are two main levels of unfairness that we must consider when looking to do any sort of statistical analysis, access to funds and recruiting level.

It is known that the NCAA basketball tournament, also known as “March Madness”, the final tournament of the NCAA season is regarded as the pinnacle of sports. It is a single elimination tournament that decides the overall champion of college basketball each year. The tournament is set up such that out of the 32 Division I conferences, the champion of each is guaranteed a spot in the tournament, then 36 other teams that “impressed” the NCAA committee(1). This guarantee’s representation of every conference, and rewards teams that play “tougher” conferences, which goes back to the unfairness factor in this sport. When the field of 68 teams is set, the NCAA committee then decides seeding, such that the best teams would play the worst teams on a path to the championship, this rewards the teams that did the best in the regular season. This seeding decision by the NCAA is at least in part, statistically based, and by creating an order of teams, the NCAA is essentially making their own “prediction” of what teams they think are better than others(1). If the NCAA’s ranking was completely true, then the lower seed would always win with the top ranked number one seed winning the whole tournament. We know this is not true, for example, since 1984 when the tournament expanded to 68 teams the seed 5 teams only have a 63% win rate in the initial matchup against the seed 12 teams(2). Much of this randomness in predicting relative team performance has to do with the complexity of valuing how much the aforementioned “unfairness” contributed to the team’s performance.

This complexity has led to many other analyst’s having different ideas than the NCAA, other ranking systems focus on different metrics and methods like efficiency and relative performance. Some years these methods perform better than others, but no method is “correct”, but rather other methods provide a reasonable prediction performance and more importantly a unique way to explore the relative strength of the teams in the tournament. This brings about the objective of this research paper, to provide a new unique way to look at March Madness teams relative strength while also having a reasonable prediction performance.

## Motivation and Overview

One thing you’ll commonly hear when it comes to March Madness predictions is, “what teams are hot?” This prompt usually explores what teams are trending up recently and what teams are trending down and how that all plays into overall predictions. If we look at popular prediction methods, we can see how they take this prompt into consideration.

The industry standard prediction, KenPom, is an adjusted efficiency margin rating that uses some adjusted offense and defense efficiency (this will be discussed in more detail later), and uses an additive model to create efficiencies for each game throughout the season then averages the two metrics to rank the teams(3). Many of the other popular rankings will have the same general idea, calculate a rating for a team's performance in any given game weighted by SOS, venue advantage, etc., then take the average rating to rank the teams. Most of these methodologies tackle “recency” in a similar way, KenPom states “The adjusted game efficiencies are then averaged (with more weighting to recent games)”(4), and RoundTable ratings states, “More recent games are weighted heavier than games from the early season” (5). Sagarin comes the closest to focusing purely on teams’ trends by stating, “The RECENT, is score-based and weights RECENT play more heavily than earlier games. Its effect will become

more pronounced the longer a season goes if a given team happens to have an upward or downward trend.”(6) It is unclear how Sagarin decides what a trend is and how much “more pronounced” effect is. It is also important to note that Sagarin uses this factor as one of three factors in his overall rankings.

The methodology used in this paper will completely focus on a time-based approach to the prediction of overall teams’ ratings at the time of the start of the final tournament. An adjusted offensive and defensive efficiency will be calculated for each team and game and those metrics will be combined for an overall rating for each team and game. These game efficiencies will be used in a simple moving average model and an automatically detected time-series model to predict ranking values for the field of 68 teams in the tournament. Then a Markovian approach will be taken to calculate the probabilities for the team's success in the first two rounds in the March Madness tournament.

# Efficiency

The first component to understanding the calculations behind the performance rating we will give to a team for each game to be used in the time-series analysis is adjusted efficiency. For efficiency calculations we will not stray away from the industry standard. This is because exploring and testing new efficiency calculations is not in the scope of this paper and is unreasonable given most efficiency calculations don’t stray too far from each other. For this data we will use data from Bart Torvik and his efficiency calculations. For each game the winner and loser are assigned an offensive and defensive efficiency. These efficiencies are calculated using points scored/allowed per 100 possessions. Possessions are defined as . (https://kenpom.com/blog/the-possession/) The first component is a shot by the offense that is discounted by continued possessions, which is denoted as offensive rebounds. The second component is an offensive turnover and the last is a random number from 0 to 1 times free throws attempted because it is not known how much a free throw should count as a possession. After calculating the PPP for the offense and defense for a team (the opposing teams PPP for offense will be the other teams PPP for defense and vice versa), we can adjust these PPP’s to get our adjusted offensive and defensive efficiencies.

The PPP’s are then adjusted by opposing team strength and venue location. The formula for this adjustment goes as follows: (http://adamcwisports.blogspot.com/p/every-possession-counts.html)

The venue factor is 1.4% so if the team is home the factor is equal to 98.6% and is 101.4% on the road. The calculation boils down to a performance metric to see how well a team scores/defends comparative to the relative strength of their opponent. It is important to note that a higher game adjusted OE is related to a strong offensive performance and a lower game adjusted DE is related to a strong defensive performance.

At this step, the method used in this paper will diverge from most predictive methods. Here, most predictive methods will use some algorithm combined with outside factors to create predictive ranks. For example, KenPom popularized using the Bill James’ pythagorean expectation method for calculating actual rank. (<https://arxiv.org/abs/math/0509698>) I will calculate a game performance metric for any given team as:

The predictiveness in my method will be taking this game performance metric and applying it to time-series forecasting methods. The reasoning for not using a more complex method to calculate the game performance metric is because I did not want to have focus on the background data going into the time-series model. If a more complex method was used like Bill James’ pythagorean expectation method, then the interpretability of the effectiveness of the time-series model could be brought into question as the pythagorean expectation method is predictive in its own nature.

# Data Considerations and Setup

It is important to note that this method, like the other popular methods, is predictive and not retrodictive. When carrying out this methodology my rankings are supposed to predict how well the teams will do going forward, it is not a model to see what factors best explain results that have already happened, a better explanation of this can be found at (7). That said, the data range used for the time series analysis is the regular season games for any given team. Only regular season games that are in the year of the predicted tournament are considered. This is because year over year college programs can experience a lot of turnover with transfers, graduates, and NBA draft declarations. It should be noted that in a sense that the previous season does have a small implication on a couple of the first games because there needs to be a base ranking set to calculate the first games efficiency ranking, and the base ranking is in part based on preseason rankings. When saying that just regular season games are used that also means that conference games are not included in the data. This is because it can be argued that teams don’t have similar amount of games when some teams play more conference tournament games than others, also any one-bid conference (a conference in which only the champion makes the tournament) will automatically have higher predictions than intended just because they succeeded in the tournament format.

The regular season data is collected from Bart Torvik’s T-Rank site which allows me to go back into previous seasons and collect season data at a specific date. The tournament is collected from (kaggle), combining these two datasets to cover the seasons from 2011-2019. This allows for predictions for 9 tournaments, and it is important to note that the format over this time period is consistent year over year, in 2011 the NCAA committee added the First Four games but we will not be predicting those games as most brackets do not count those games. From the regular season data we will use the offensive efficiency and defensive efficiency statistics that were discussed in the previous section. We can see that the overall data counts 47460 regular season games, 603 tourney games,  and 352 unique teams over the 9 seasons. (Figure)

To provide some context to efficiencies we can see how the efficiencies compare to some more popular household statistics. (eda, graphics and explain, overall efficiency spilt by winners and losers, correlation plot)

# Time-Series Model

Given a set of adjusted offensive efficiencies and a set of adjusted defensive efficiencies and the combination of both we can now apply a time-series analysis on the season data for the teams. The main problem is that we are dealing with many different teams in multiple years and the main goal is to fit a time-series that is right for all each one individually in order to forecast future values. This can be done using the auto.arima function in R in the forecast package.

The function uses an algorithm called the Hyndman-Khandakar algorithm which determines the best ARIMA (Autoregressive Integrated Moving Average) model for a given team each year. We will consider three parameters for this model, p, which is the degree of the Autoregressive term, d which is the amount of times the time-series data has to be differenced in order to create stationary data, and q, which is the degree of the Moving Average term. For the sake of our data there will be an assumption made that there is no seasonal component to our data because we are only modeling within a singular year. The formula for this model can turn out to be:

Where B is a backshift operator (https://otexts.com/fpp2/backshift.html) and the phi and theta functions of B correspond to a degree polynomial of the order p and q, respectively. The is a white noise process with mean zero and variance . If then there is an implied order of differencing of d.

This algorithm works by first finding the order of differencing needed in the time-series model. Approaches before have suggested minimizing AIC from different combinations of models, but this can lead to over-differencing, instead the algorithm uses successive KPSS unit-root tests to determine the differencing coefficient. Then p and q are found using a step-wise selection minimizes the AIC. This is important in relation to working with team data as a time series. Many team’s data could already have a stabilized mean throughout the season, for example a team that has never seemed to go on a stretch of good or bad games, compared to a team that is more streaky and would have not have a stabilized mean. The algorithm will take both of those scenarios into account and determine which is appropriate according to the team’s trends. The last thing to note is that in order for this algorithm to work as intended there is assumption of homoscedasticity that must be made about the time-series data beforehand. With team game data we will find that there is not enough games in a season to identify a trend in variance that would violate the homoscedasticity assumption. This is helpful in the fact that the ARIMA model can be used with this data, but it also begins to uncover a limitation with the time-series model and this team data in the fact that ARIMA models tend to work stronger with more data points. Here is an example of Ohio State in 2011 and how they were fit to an ARIMA model.

# Markov Chain to Create Ranks

For every year we have a set of numbers defining the teams to participate in the final NCAA tournament. As stated in the previous section it was stated how we are using a time-series analysis to populate these set of numbers, but we can see that the following markov chain method to get expected probabilities of teams making it to the second and third rounds can be applied to any set of numbers that are defining the teams as long as those set of numbers are positive. This stochastic method takes the set of any pair of four connected teams in the march madness bracket, for example, seeds (1,16,8,9) or (2,15,7,10). (link to bracket) We will use seed numbers as a baseline example for creating these probabilities. For any given seed we have the model:

# Analysis of Results

With the time-series model set and Markov-chain creating winning probabilities. The R code now runs a loop for every tournament team in the nine years of study to simulate predictions for the first two rounds of each of those years tournaments. First, we can refer back to the seed method as a baseline to show results of these predictions. The seed method performed as follows: (Figure)

(Analysis seed method prediction)

(Time series one and two performed as follows)

(Analysis)

(Deeper analysis into why time series performed as such with figures of performance against upsets vs. favorites)

# Limitations and Conclusions

# Appendix A Replace with Appendix Title

Start here

**BIBLIOGRAPHY**

**ACADEMIC VITA**

Insert your resume here. **Please OMIT your GPA, address and phone number.**