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

TITLE LINE (Do not use ALL CAPS, use Title Case)

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 1](#_Toc125580560)

[Motivation and Overview 3](#_Toc125580561)

[Chapter 2 Data Considerations and Setup 5](#_Toc125580562)

[Chapter 3 Efficiency 5](#_Toc125580563)

[Discussion 5](#_Toc125580564)

[Example in Data 6](#_Toc125580565)

[Chapter 4 Time-Series Model 7](#_Toc125580566)

[Chapter 5 Markov Chain to Create Ranks 8](#_Toc125580567)

[Chapter 6 Analysis of Results 9](#_Toc125580568)

[Chapter 7 Limitations and Conclusions 10](#_Toc125580569)

[Appendix A Replace with Appendix Title 11](#_Toc125580570)

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.



Figure 1. Atherton Hall, SHC Est. 1997

To insert pictures using Word 2007 & 2013 (Windows):

1. Click on Insert > Pictures
2. Right click on the picture and choose Insert Caption
3. Go to the List of Figures page and choose Update Field (Update Entire Field)

Note: If you delete the List of Figures we have in the template, you need to add it back in using References Tab > Insert Table of Figures.

To insert pictures using Word 2011 (Mac)

1. Click on Insert > Photo
2. Right click on the picture and choose Insert Caption…
3. Go to the List of Figures page and choose Update Field

If that doesn’t work, on a Mac, here’s how to add the Table of Figures:

1. Click on Insert > Index and Tables
2. Choose Table of Figures tab, and choose Figure under Caption label, and click OK

# 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

# Efficiency

## Discussion

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 to stray too far from each other. For this data we will use data from Bart Torvik and his efficiency calculations. The general formul

## Example in Data

# 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 data was tested using two different time-series methods. One method was a model using a simple moving average (MA) model of 5 data points. The second method uses the auto.arima function in the forecast package in R.

# Markov Chain to Create Ranks

Start here

# Analysis of Results

# 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.**