Analytics for Decision-Making in Amateur Sport Organizations

by

Islay Wright

A thesis submitted in conformity with the requirements for the degree of Master of Applied Science
Graduate Department of Mechanical and Industrial Engineering
University of Toronto

© Copyright by Islay Wright, 2017
Analytics for Decision-Making in Amateur Sport Organizations

Islay Wright

Master of Applied Science
Graduate Department of Mechanical and Industrial Engineering
University of Toronto
2017

Abstract

In this thesis, we demonstrate the value that analytics can have in guiding decisions at amateur sport organizations. We first use machine learning methods to estimate how the demand for tennis is distributed across Canada on a geographical level; geographical prediction of demand for sport has not been seen in the sport management literature. We develop novel validation techniques to show that our estimates outperform the naive assumption that the same proportion of tennis players lives in each region in Canada. To present the estimated demand to decision-makers, we build interactive maps that display the supply and demand for tennis across Canada. Through a case study in facility location, we show that the predicted demand can be used to determine the optimal locations of new tennis programs. Lastly, we present interactive tools that we built for decision-makers to analyze their membership and make targeted decisions about recruitment and retention strategies.
Dedicated to my grandparents.
Acknowledgements

I would first like to thank my supervisor, Timothy Chan, for your gracious support and mentorship. I have trouble expressing how much I learned from you, both academically and personally. I will always hold you in the highest regard, and I aspire to approach my work with the same level of rigour and compassion that you put into yours.

I would also like to extend thanks to my committee members, Michael Carter and Daniel Frances, for taking the time to sit on my committee and provide me with valuable feedback. Thank you to Tulla Bateman, Rob Cianfarani, and Richard Crowell from Tennis Canada, and James Anderson from the CFSA for providing me with data and insight into Canadian NSFs. I would also like to thank my undergraduate thesis student, Amina Sherif, for helping to design and build the maps in this thesis.

Thank you to all the members - past and present - of the Applied Optimization Lab, and the “Greater AOL.” I would especially like to thank the following labmates for their support and friendship. Thank you to Aaron and Neal for your closeness while we went through these masters degrees together: you have both inspired me over the past two years: Aaron, with your positive attitude and hard work ethic, and Neal, with your immense wisdom and humility. Thank you Chris, for your kindness and for being the first friend I made in the AOL three years ago. Thank you Justin, for being a mentor of mine over the past three years and for teaching me a lot of what I know about research. Thank you Philip, for your enthusiasm (particularly about math, puns, and Star Wars), and Rafid for always knowing an answer to all my questions. Thank you to Minha for your pragmatism and keeping us all healthy (i.e., supplying us with Vitamin C), Ian for your indefatigable drive, and Ben for pushing us with your endless curiosity. You have all taught me millions of things, and I’m proud to call you all close friends. On that note, Tim, I would like to apologize for perhaps being a little too social in the lab.

I also would like to thank all of my family for your support over the years. In particular, I’d like to thank my brothers, Noah, Asher, and Sam, for your love and encouragement, and my aunt Lisa, for welcoming me to Toronto as if I was one of your own kids. Thank you, as well, to Dan and Fiona for your generosity.

Penultimately, I would like to thank my parents for your unwavering support for as long as I can remember. Thank you Rob for teaching all of your children the value of curiosity and necessity of the scientific method. Thank you Liz for your guidance and example that I can do anything I want to do.

Lastly I would like to thank Ben again; your companionship has made Toronto, finally, home.
Contents

1 Introduction 1
  1.1 Summary of Contributions ............................................. 2
  1.2 Outline of Thesis ..................................................... 3

2 Geographical Demand Prediction 5
  2.1 Literature Review ..................................................... 6
    2.1.1 Methods .......................................................... 6
    2.1.2 Variables ......................................................... 7
    2.1.3 Validation ........................................................ 8
    2.1.4 Outcomes ........................................................ 8
  2.2 Data ................................................................. 9
    2.2.1 Survey Data ........................................................ 10
    2.2.2 Canadian Census Data .......................................... 11
    2.2.3 Geographical Data ............................................... 12
  2.3 Demand Prediction ................................................... 14
    2.3.1 Overview of Model Fitting Procedure .......................... 15
    2.3.2 Clustering ....................................................... 18
    2.3.3 Individual Approach ............................................. 19
    2.3.4 Aggregate Approach ............................................. 22
    2.3.5 Inputting Census Data to Models .............................. 24
    2.3.6 Validation Procedure ........................................... 24
    2.3.7 Summary and Results of Validation Procedure ............... 24
  2.4 Presentation to Decision-Makers .................................... 28
Chapter 1

Introduction

Analytics is a rapidly growing approach to problem solving which seeks to gain insights from data for the purpose of guiding better decisions. In particular, sports analytics has grown substantially over the past two decades. For instance, the MIT Sloan Sports Analytics Conference has grown from 175 attendees to 3,500 attendees since its fruition in 2007 [19], showing the increasing evidence of the potential that analytics can have in solving sport-related problems. In addition, every major sports team employs an analytics expert or has an analytics department [46]. However, the majority of this growth in analytics in sport is concerning (1) sport performance rather than sport management and (2) professional sports rather than amateur sports. To our knowledge, analytics methods have rarely been applied to business problems in amateur sport. Our research focuses on illustrating the immense use that descriptive, predictive, and prescriptive analytics can have in the amateur sport context.

Our research was first motivated by the National Sport Federation Enhancement Initiative (NSFEI), which is a four-year project led by the Canadian Olympic Committee (COC) with the aim of improving the organizational capacity of amateur sport organizations. One of the main objectives of this initiative is to help amateur sport organizations effectively recruit new athletes into their sports [33]. In Canada, these organizations are often constrained by scarce resources. David Patterson, the leader of the NSFEI, said that by using analytics to better allocate their resources, amateur sport organizations will benefit “on the field of play and away from competition, where we work hard to attract and retain more Canadians to a lifestyle in sport” [33]. With this clear need for analytics to help make decisions in amateur sport organizations, this thesis is concerned with how analytics can be used to help these organizations with
the recruitment of new participants and retention of current participants.

While increasing participation in sport is important for amateur sport organizations, it is also important for the Canadian population in general. Sport participation has been shown to be exceptionally beneficial to individuals [23] and communities [50, 35]. However, few of the efforts of amateur sport organizations to increase sport participation are founded in analytics. In addition, much of the quantitative research on sport participation has been focused on participation at the individual level. However, with the importance of sport participation in communities, we can use analytics to understand sport participation at a population level. By using analytics to understand populations of Canadians, amateur sport organizations will be able to better allocate their scarce resources and develop more effective methods to engage more Canadians in sport.

Each of the three types of analytics, descriptive (e.g., basic statistics), predictive (e.g., machine learning), and prescriptive (e.g., optimization), can help amateur sport organizations make decisions about how to better manage their resources. In our research, we demonstrate how each of these types of analytics can be used to assist sport managers in making decisions. We show that descriptive analytics can help sport managers understand characteristics of current members, allowing them to better target efforts to keep existing members or recruit new members. Predictive analytics can be used to help amateur sport organizations understand the demand for their sport, assisting them in making decisions about where to put new facilities or programs. Prescriptive analytics can help them determine the optimal locations to place these new facilities or programs if the organization has a specific goal for the facility or program. While in the past there have been many studies using qualitative (e.g., [50]) and quantitative (e.g., [52]) methods to understand participation in sport, few elaborate on how the information gained can help sport managers make decisions. We hope to bridge this gap by using analytics methods with the aim of guiding decision making in amateur sport.

1.1 Summary of Contributions

In this thesis we make the following three contributions:

1. First, we design novel methods to predict demand for tennis on a geographical level, and validate these methods using Canadian census data. We then use the resulting predictions, as well as locations of tennis courts, to build an interactive map that highlights the supply and demand for
tennis across Canada. The information gained from this map is useful for decision-makers looking to build new courts or develop new tennis programs with the aim of recruiting new members.

2. Second, to demonstrate the relevance and novelty of the demand prediction, we present a case study that uses the predicted demand as input into an optimization model. This optimization model selects the best indoor turf facilities to hold winter tennis programs to teach children the fundamentals of the sport. We show that by using the demand prediction as input into this optimization model, we can achieve different results from those if we assume demand is uniformly distributed across the country. We illustrate that by having a specific objective, the demand prediction can be used to make directed decisions.

3. Finally, we review practical, descriptive tools that we built for the analytics sessions that we ran as part of the NSFEI. These tools focus on how National Sport Federations (NSFs) can use analytics to understand their membership, and create targeted recruitment or retention strategies. We show that the tools were effective in helping the Canadian Freestyle Ski Association (CFSA) increase their analytics capabilities. By increasing their analytics capabilities, the CFSA has been able to compare their membership data to that of other snow sports, which has in turn has helped to inform their sport development and high performance development decisions.

1.2 Outline of Thesis

The remainder of this thesis is organized as follows. In Chapter 2, we review the quantitative sport management literature related to sport participation. We then explain the framework that we developed to predict demand for tennis on a geographical level across Canada. We demonstrate the validity of this framework through our novel validation technique. In addition, we present how we combined the predicted demand with the supply of tennis courts in interactive maps that decision-makers in amateur sport can use to understand supply and demand for tennis.

Chapter 3 illustrates, through a case study, that the demand prediction from the previous chapter can be used as input into an optimization model to make effective decisions in amateur sport. This shows the value that prescriptive analytics can have in amateur sport organizations if the organizations have specific business objectives.

In Chapter 4, we demonstrate the power that descriptive analytics can have in amateur sport orga-
nizations by presenting descriptive tools we built to aid decision making regarding membership.

Lastly, Chapter 5 concludes the previous chapters and offers recommendations for further research in the amateur sport analytics space.
Chapter 2

Geographical Demand Prediction

One of the primary functions of Canadian National Sport Federations (NSFs) is “implementing national initiatives to develop and promote their sport” [39]. Part of the role of developing and promoting their sport on a national level requires sport managers at the NSFs to be concerned with creating new programs or new facilities. In order to make directed decisions regarding locating new programs or new facilities, it is vital for these sport managers to know how demand for their sport is distributed geographically. Despite the growth of analytics in sport and the need of decision-makers in sport management to understand demand for sport on a geographical rather than individual level, models to estimate the demand geographically have not been seen in the sport management literature.

In this chapter we show that we can estimate the demand for sport on a geographical level by using demographic information of individuals and populations. We build a prediction model based on survey data and census data that estimates the proportion of people who play tennis in a given region. For the rest of this thesis we will refer to this estimate as the demand prediction. Our hope in creating this demand prediction is that it will help sport managers make directed decisions regarding the placement of new programs or facilities.

This chapter first reviews the relevant literature on sport participation in Section 2.1. Section 2.2 provides a detailed overview of the data used for our research and how it was appropriately processed. The following section, 2.3, reviews the methods to predict demand for tennis on a geographical level and the methods to validate these predictions. Lastly, in Section 2.4 we show how these results are presented in a way that makes them easy to use and interpret by decision-makers in sport management.
2.1 Literature Review

We consider literature along four dimensions: methods, variables, validation, and outcomes. We first review the quantitative methods used in the sport management literature studying participation and explain how the difference between the questions the literature answers and the question we are trying to answer leads to a difference in methods. We then look at which variables examined in the literature influence participation in sport, and discuss which of these should be considered in our research. Next, we review the methods used to validate the models in the literature and how our approach will differ. Finally, we will review the outcomes and policy recommendations produced from the sport management literature and explain how these differ from the outcomes and we hope to achieve in our research.

2.1.1 Methods

In the sport management literature there has been significant effort to understand the determinants of sport participation. Many studies have examined how socio-economic, demographic, psychological, health, and location factors impact frequency of sport participation or physical activity by an individual [16, 20, 43, 49, 44, 48, 34, 27, 36, 21, 32, 52, 51, 38]. All of these studies ask what factors influence an individual to participate in sport / physical activity or what factors influence the frequency of an individual’s participation in sport / physical activity. Alternatively, we are asking how to can determine the proportion of a population who participates in sport based on demographic factors of individuals and populations. Knowing which demographic features influence participation, and the methods used to determine this, is relevant to our research.

The methods in the previous literature use statistical and econometric models to understand the factors that influence an individual to play sports. Logistic regression models are commonly used to find the demographic determinants of participation in sport or physical activity [16, 20, 43, 49, 44, 48, 34, 27]. Other statistical models, such as ordered probit models [36, 21] have also been used to examine the relationship between individual demographic features and participation in sport. In addition, econometric models, such as double hurdle models, have been used to look at the impact of income, age, gender, and family structure on both the choice to participate in sport and the amount of time spent participating in sport [32]. All of these models are fit using demographic features of individuals as independent variables and some metric representing the individuals’ participation in sport or physical activity as the dependent variable. A main distinction between the previous research question and our research question is that
their question is focused on participation by an individual and our question is focused on participation of a population. Hence, the models built in the previous research do not have the capacity to predict demand for sport on a population level.

2.1.2 Variables

Many of the papers listed in Subsection 2.1.1 are focused on finding the determinants of physical activity or sport in general. The findings in these papers are fairly consistent despite using datasets from different countries across North America, Europe, Asia, and Australia, and analysis for different combinations of sports. In general, being older and female are negatively correlated with participation in sports [16, 20, 43, 49, 44, 48, 34, 27]. Individuals with higher income are generally more likely to play sport [32, 20, 48], however Cheah and Poh [16] find that individuals with higher income are less likely to participate in physical activity. Although generally higher education is positively correlated with participation in sport [20, 43, 48, 34, 27], Cheah and Poh [16] find that higher education is negatively correlated with being physically active. People with professional occupations are more likely to participate in sport than people with any other occupations [20, 48, 34]. Ethnicity and language can also be factors that impact the frequency of sport participation. Downward showed that in the UK, white, British individuals are more likely to participate in sport than the rest of the population [20]. Stratton et al showed that individuals in Australia who do not speak English are less likely to participate in sport [48]. All of these studies show that socio-economic and demographic factors are important variables to consider with respect to frequency of sport participation, hence we consider these variables in our prediction of demand for sport.

Other sport participation studies focus on finding the determinants of playing racket sports or tennis in particular. Through a random probit model, Farrell et al find that age is negatively correlated with playing racket sports, men are more likely to play racket sports than women, and education is positively correlated with playing racket sports [26]. In a study on the relationship between cultural and economic capital (education and income respectively), and sport participation in adults, Stempel finds that tennis is the sport that is most exclusive to individuals with both high cultural and high economic capital in the United States. The individuals with highest income and highest education had an estimated odds ratio of playing tennis that was 53.48 times higher than that for the individuals with lowest income and lowest education. [47]. A more recent study by Eime et al looks at the the effects of socio-economic
status and geographical remoteness on participation in 95 types of sport and physical activity, including tennis. Using logistic regression models, the researchers find that there is a positive correlation between living in a rural region and tennis participation, which holds for 14 of the 95 types of physical activity considered [22].

These results on the determinants of participation in tennis and racket sports reflect the majority of the results on the determinants of participation in sport in general. Based on the variables that significantly impacted participation in sport in the previous studies, we choose to consider age, gender, income, education, occupation, ethnicity, language, and province of residence as independent variables in our models. With an extensive list of independent variables to consider based on the results of the previous studies, we are able to use these variables to build a model that focuses on prediction accuracy rather than information gain.

2.1.3 Validation

Because the models in the previous research are focused on learning about the factors that influence participation in sport, the studies validate their models with goodness-of-fit metrics. These metrics include likelihood ratio tests [16], Wald tests [32, 47] $R^2$ [52], or Pseudo $R^2$ [20] to quantify the fit of their models. These methods are sufficient for the statistical purposes of the previous papers. Since our model is intended to be used to accurately predict on an unknown population of a geographical region, it is vital for us to validate that it performs well out of sample. Validation of the models on an out-of-sample population has not been seen in the previous literature.

2.1.4 Outcomes

Many studies in the literature reviewed in Subsections 2.1.1 and 2.1.2 seek to discover relationships between demographic features of an individual and participation in sport. Hence the outcomes of these studies describe which factors influence an individual to participate in sport. Because the results are specific to the individual, and are not differentiated by location, the outcomes of these studies do not examine actionable insights into how the results can be used to increase sport participation.

Other studies conclude with general policy recommendations, based on the odds ratios of their regression models, about how the government or local sporting organizations can help increase participation in sport. For instance, Cheah and Poh [16] suggest that office-based exercise programs should be introduced
to encourage well-educated individuals to stay active. This suggestion is derived from the odds ratio of 0.797 associated with high education, implying that the highly educated people are less physically active than the general population. With limited resources, it can be very difficult for a policy maker to make directed decisions based on a general suggestion.

In another study, Lera-Lopez and Rapun-Garate suggest that in order to encourage more people to participate in sport, policy-makers should focus on young age groups, women, and workers, rather than low-income groups. [36]. While these recommendations are validated by their findings, they are too general to inform directed policy decisions. For instance, it would be useful for policy-makers to know how much to shift their resources from low-income groups to youth, women, and workers, and in which locations they may have the largest increase in participation among these groups. By understanding demand for sport geographically we are able to better target populations that can benefit from more sport programming.

These examples highlight the main difference between the insights gained from the previous research and the intended insights to be gained from our research. The previous research focuses on understanding the determinants of individual participation in sport, so the gained insights from the research are also at the individual level. Our research will focus on understanding how participation in sport is geographically distributed on a population level.

By understanding the distribution, our analysis can be used to make more targeted policy recommendations and our quantitative results can be used as input for further models, as evidenced in the case study in Chapter 3. The case study illustrates how the geographical distribution of participation discovered in our model can be combined with optimization modelling to make targeted and insightful policy decisions about where to place new tennis programs.

\section*{2.2 Data}

We use many different sources of data throughout this study. The following subsections include detailed explanations of how the data were collected, which data we use from each source, and some summary statistics.
2.2.1 Survey Data

The Canadian Tennis Brand and Health Study is a study completed for Tennis Canada with the aim of better understanding tennis participation across Canada [14, 15]. The study is based on survey data that has been collected annually since 2008. Charlton Strategic Research Inc. has administered this 25 minute online survey in which respondents aged 12-17 answered for themselves and respondents older than 18 answered for their households. In this project, we use data obtained from the 2015 and 2016 surveys.

In 2015 there were 3,005 respondents from across the ten provinces in Canada; since many answered for their households, we have data from a total of 7,372 Canadians. In 2016 there were 3,075 respondents from across the provinces, and we have data for 7,233 Canadians.

The data we use from the survey includes demographic information about the respondents and frequency of tennis play. While respondents did answer for their households, there is not enough data on the children to build models based on demographics, hence, we only focus on data from people who directly responded to the survey. In addition, the children ages 12-17 who directly responded to the survey have similar responses for fields such as occupation and education, allowing for little diversity in the data; we hence only focus on respondents age 18 and older. If a respondent did not answer a relevant question or responded with “prefer not to say” we remove the respondent. This led to the removal of 1,540 respondents. In addition, the 2015 and 2016 surveys were sent to the same survey panel, so possible duplicate respondents were removed. Since respondents have been anonymized, if respondents in the 2015 and 2016 dataset had the same postal code, gender, and ethnicity, and the 2016 respondent was either 0, 1, or 2 years older than the 2015 respondent, the 2016 respondent was removed. The choice of removing the 2016 data was arbitrary since only age differs between the 2015 and 2016 data for these respondents. This led to the removal of 400 additional respondents. See Figure 2.1 for a summary of the process of removing respondents from the data set.

After removing these respondents we are left with data from 4,140 respondents. We use data for fields listed in Table 2.1 as well as frequency of tennis play in the respondent’s most active 8 consecutive weeks of play in the past year. Age and income are ordinal variables, while every other variable is converted into dummy binary categorical variables.
# Respondents
6080
with age < 18
with relevant questions unanswered
Remove respondents:
4981
4540
possible duplicates between 2015 & 2016
4140

Figure 2.1: Detailed process outlining the removal of survey respondents.

<table>
<thead>
<tr>
<th>Survey Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Province of residence &amp; postal code</td>
</tr>
<tr>
<td>Ethnicity</td>
</tr>
<tr>
<td>Place of birth (Canada or elsewhere)</td>
</tr>
<tr>
<td>Language spoken most at home (English, French, or other)</td>
</tr>
<tr>
<td>Highest level of education</td>
</tr>
<tr>
<td>Occupation</td>
</tr>
<tr>
<td>Household income</td>
</tr>
</tbody>
</table>

Table 2.1: Relevant Survey Data

## 2.2.2 Canadian Census Data

We use two sources of data from the 2011 Canadian census: the National Household Survey [7] and the Census Profile [5]. From the Census Profile, we use population, median age, gender, and language spoken most often at home (English, French, other). From the National Household Survey we use place of birth (born in Canada or elsewhere), ethnicity, occupation, education, and median household income. Data are provided at the Forward Sortation Area (FSA) level. An FSA is the area represented by the first 3 characters of a postal code, and there are 1,598 FSAs across the 10 provinces in Canada. The relevant census data used in this study are summarized in Table 2.2.

### Comparison Between Survey Data and Census Data

For modelling purposes, we require the census data to have the same categories as the survey data. For the occupation category, we develop three buckets of occupations (white collar, blue collar, and not in
<table>
<thead>
<tr>
<th><strong>Census Profile (2011)</strong></th>
<th><strong>National Household Survey (2011)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>Place of birth (Canada or elsewhere)</td>
</tr>
<tr>
<td>Median age</td>
<td>Ethnicity</td>
</tr>
<tr>
<td>Gender</td>
<td>Occupation</td>
</tr>
<tr>
<td>Language spoken most at home (English, French, or other)</td>
<td>Education</td>
</tr>
<tr>
<td></td>
<td>Median household income</td>
</tr>
</tbody>
</table>

Table 2.2: Relevant Census Data

<table>
<thead>
<tr>
<th><strong>Survey Data Variables</strong></th>
<th><strong>Survey Variable Type</strong></th>
<th><strong>Census Data Variables</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender, place of birth (Canada or elsewhere)</td>
<td>Binary</td>
<td>Proportion of population female, Proportion of population Canadian born</td>
</tr>
<tr>
<td>Province, ethnicity, language, education, occupation</td>
<td>Categorical</td>
<td>Proportion of population in each category</td>
</tr>
<tr>
<td>Age, income</td>
<td>Ordinal</td>
<td>Median age, median income</td>
</tr>
</tbody>
</table>

Table 2.3: Mapping of survey data variables to census data variables.

workforce), and each option for occupation in both the survey data and census data is sorted into one of these buckets (see Figure A.1 in Appendix A).

The census data is then formatted to have the same layout as the survey data. Variables in the survey data that are binary or categorical are formatted in the census data to be proportion of the FSA. Variables that are ordinal in the survey data are listed as median values over the population of the FSA in the census data. See Table 2.3 for the details of the mapping of survey data to census data.

It should be noted that in the process of collecting the survey data, age and gender quotas were set to be reflective of the Canadian population. However, a minimum target was set for the number of respondents from each province to allow Charlton Strategic Research Inc. to do analysis across all age groups in each province. This results in a substantially different distribution of survey respondents by province than the true Canadian population (Figure 2.2). The other variables in the survey and census data are compared in Appendix A in Figures A.2 to A.9.

### 2.2.3 Geographical Data

Because this study is focused on understanding the geographical distribution of demand for sport, there are many geographical data sets we use. To collect data on the locations of tennis courts in the ten provinces of Canada, we scraped data from the Global Tennis Network [30] and had data provided by Tennis Canada. We then searched this database for major Canadian cities and towns, and if these were not included in the database we searched manually on Google for tennis courts in these regions. For
instance, there were initially no tennis courts in our database on Prince Edward Island, so we searched for tennis courts in this province and manually added the results to our database.

The tennis court location data include addresses of the tennis courts, number of courts at each location, and the type of court (categorized into indoor, bubble, or permanent). Once our database was as extensive as possible through this process, we geocoded the addresses of the courts using the Google API to get the latitude and longitude coordinates and then calculated the Universal Transverse Mercator (UTM) coordinates. See Figure 2.3 for the complete process of geocoding the court locations. At the end of this process, we have data for 6,882 tennis courts in 1,930 locations. While it is unlikely that this list is exhaustive, we suspect that most of the private or membership-based courts in Canada are in our database, due to many of them having a relationship with Tennis Canada or simply having websites and being easily searchable. However, we are likely missing some public courts where the people who play at these courts do not add them to the Global Tennis Network and are less easily searchable on the internet.

The last set of geographical data we use contains latitude and longitude coordinates of centroids of each FSA and postal code in the ten provinces in Canada. This is crowd-sourced data that has since been removed from the internet. We converted the centroid coordinates into UTM coordinates. Having all the geographical data in UTM coordinates is useful because the distance in meters between the two points can be calculated simply using the Euclidean distance.
2.3 Demand Prediction

We use survey data from the 2015 and 2016 Canadian Tennis Brand Heath Studies to build prediction models. The aim of the prediction models is to be able to accurately predict the proportion of tennis players in each FSA in the ten provinces of Canada. These models use the demographic features from Table 2.1 to predict the proportion of a group of Canadians that are tennis players. We then input census data into these models to predict the proportion of people in a given FSA who play tennis.

It is important to consider how frequency of playing tennis translates to being a tennis player. Addressing our objective, which is to accurately predict the potential demand for tennis geographically, we make the assumption that if a person has played tennis at least once in the past season, he/she should be included in the people demanding tennis. For the remainder of this thesis we will refer to a person as a tennis player if he/she played at least once in the season prior to the survey being administered.

For clarity, let $X^S$ be the matrix of demographic variables of the survey respondents and let $X^C$ be the matrix of demographic variables of the FSAs. The dependent variable corresponding to $X^S$ is called $Y^S$ and element $Y^S_i$ is the binary indicator of whether or not a survey respondent $i$ plays tennis. Let the row of $X^S$ that represents survey respondent $i$ be $X_i^S$. Note that from $Y^S$ it is possible to calculate the proportion of survey respondents in each FSA: call this $Y^C$. Then $Y^C_j$ is the proportion of tennis players in FSA $j$ and corresponds with the row of $X^C$ that represents FSA $j$, $X_j^C$. The models are
trained using $X^S$ and $Y^S$, and then we validate how well they perform, in comparison to $Y^C$, when $X^C$ is inputted.

### 2.3.1 Overview of Model Fitting Procedure

In this section we use supervised machine learning models to predict the proportion of people in each FSA who play tennis. Before we explain the specific details behind building and validating the models, there are some important higher-level concepts to note. Typical supervised machine learning problems are trained on features (independent variables) and targets (dependent variables). Before a model is trained, a subset of the features and corresponding targets is set aside to be a validation set, which is not used in any of the training. The models are then built using the data that was not removed. The remaining data is often partitioned further into training and test sets, where the features and targets of the training set are used to fit a model. The features of the test set are then inputted into this model to get resulting predictions, which are compared to the true values in the test set. After seeing how well the model performs on the test set, it may be adjusted to try to achieve better predictions, after which a final model is decided upon based on the model that achieves the best predictions. Note that in this entire process, the validation set is not used at all. Once a final model has been decided upon, the features from the validation set can be inputted into the model and the resulting predictions can be compared to targets in the validation set. A norm difference between predictions from the validation set and validation set targets is often used as a metric of how well the model performs [3].

In our methods, we use two separate approaches, each of which trains many models on the entire set of survey data $\{X^S, Y^S\}$. Rather than removing a validation set before training the two approaches, we use the census data, $X^C$ (features) and the corresponding true proportions of tennis players in each FSA, $Y^C$ (targets) as the validation set. It is fair to use the census data as the validation set since it is independent from the survey data. In addition, using the census data as the validation set will help us answer the question about whether we can accurately predict demand for tennis on a geographical level; if one of the two approaches we use trains models that perform well on the validation set then we can confidently say that we can accurately predict demand for tennis. To determine whether the approaches train models that perform well on the census data, we compare both of them to a naive prediction that supposes the same proportion of tennis players live in each FSA.

In each of the two approaches, we partition the survey data $\{X^S, Y^S\}$, into training and test sets,
and train various machine learning models on the training set. After comparing the resulting predictions to the test set, we choose the machine learning model that performs best, and train this model on all the survey data. We refer to this best performing model as the “final model.” The census data (validation set) is then inputted into the final model and the validation process explained above is continued. Hence note that there are two validation procedures: we validate that the model we choose performs well on a test set of the survey data, and then we validate that the census data inputted into this model performs well compared to a naive prediction.

The difference between the two main approaches we build is in the way that the data are processed prior to training and the way the data are compared to the true number of tennis players in the test set. The first main approach is called the individual approach. The individual approach builds a model that inputs individual-level data and predicts the probability that a given person plays tennis. Recall that the elements of $Y_S$ are binary, indicating whether or not a person plays tennis, and we are predicting a probability. Hence, to check the accuracy of these predictions, we create groups of survey respondents and compare the mean predicted probability of playing tennis over a group to the true proportion of tennis players in the group.

The second main approach we build is called the aggregate approach. Rather than training on individual respondents like the individual approach, the aggregate approach trains on data where each data point represents a group of respondents. We choose to try this approach since each data point in the census data is technically aggregate information of groups of Canadians (FSAs), and this is the data that we for which want to be able to accurately predict the percentage of tennis players. Hence, the output of the final model in the aggregate approach is the percentage of the respondents in each group who are tennis players, rather than the probability that a given survey respondent is a tennis player. Note that for both the individual approach and the aggregate approach, respondents must be placed in groups.

Figure 2.4 outlines the entire model fitting procedure. The final output of this procedure is a measure of error concerning how well the individual model, aggregate model, and naive prediction, perform on the census data. This output can be seen in the bottom-right corner of the figure. The entire validation procedure is 5 steps, and outputs from each of the first four steps are used in a later step, either as inputs into an algorithm or as a model.

The following subsections explain the procedure from Figure 2.4 in detail. Subsection 2.3.2 explains
the first step of the procedure, presenting the clustering heuristic used to group respondents for the individual and aggregate approaches. In Subsection 2.3.3, these groups, or clusters, are then inputted into Algorithm 2 from the second step of the procedure. This is the individual approach, and the output is a final individual model (FIM). Likewise, Subsection 2.3.4 explains the aggregate approach in Algorithm 3, which outputs a final aggregate model (FAM). In step 4 of the procedure, census data is inputted into the outputted models from steps 2 and 3 to achieve predictions of the proportion of tennis players in each FSA, with one prediction from the individual model and another prediction from the aggregate model. This is explained in Subsection 2.3.5. Lastly, step 5 of the procedure is explained in Subsection 2.3.6, and uses the predicted proportion of tennis players in each FSA from the individual model and from the aggregate model, as well as the FSA clusters, to determine the error of the predictions.

We summarize this entire procedure again, in pseudo-code, in Subsection 2.3.7. In this subsection, we repeat the procedure outlined in Figure 2.4 500 times to get a range of 500 errors for each model.
The results are analyzed and the model with the lowest median error is selected as the model that most accurately predicts demand for tennis geographically.

### 2.3.2 Clustering

Since we are working under the hypothesis that the proportion of tennis players is different according to the region, a natural division of groups is by region in which the respondents live. To ensure granularity of at most 5% in the aggregate data set, there must be at least 20 respondents in each group. Examining the FSA in which survey respondents live, we note that only 29 of the 1589 FSAs have at least 20 survey respondents, and 829 of the 4140 respondents live in FSAs with at least 20 respondents. It is thus unreasonable to group respondents by FSA. Figure 2.5 shows that most FSAs have few - if any - survey respondents.

![Figure 2.5: Histogram of the distribution of the number of respondents in all FSAs.](image)

Instead, FSAs are grouped by similarity until there are at least 20 survey respondents in the group of FSAs. This process is outlined in Algorithm 1. Two FSAs are similar if the 2-norm difference between the demographic features of the FSAs is small. Recall that in $\mathbf{X^C}$, every variable is on the interval $[0, 1]$ except for age and income. In order to avoid age and income heavily weighting the 2-norm difference between FSAs, we normalize these variables such that they are between 0 and 1. The normalized data set is called $N(\mathbf{X^C})$. We write $N(\mathbf{X^C}_j)$ as the vector of demographic features from $N(\mathbf{X^C})$ for FSA $j$.

In this algorithm, after there are at least 20 survey respondents in a group, it is considered full and a
new group is started. This is to create as large a training set as possible, since each group translates to one data point in the aggregate approach. Training on more data will make a stronger model. In Step 3 of Algorithm 1, an FSA is selected at random, hence the resulting clusters may differ each time the algorithm is run.

**Algorithm 1** Clustering Heuristic.

**Input:** Normalized census data, \(N(X^C)\); Postal codes of each respondent in survey data;

**Output:** Cluster assignment for each FSA, \(C\);

1: Label every FSA unassigned;
2: If an FSA has at least 20 survey respondents, assign it to its own group and label it assigned;
3: Randomly select an unassigned FSA, \(j\). Find the unassigned FSA, \(k \neq j\), such that \(\|N(X^C_j) - N(X^C_k)\|_2\) is minimized, i.e., \(k\) is the closest FSA to \(j\). Assign both of these FSAs to the same new group. If there are no unassigned FSAs, stop;
4: If this group has at least 20 respondents, go back to step 3. If this group does not have at least 20 respondents and there are still unassigned FSAs, move on to step 5. If there are no more unassigned FSAs, stop;
5: Find the next closest unassigned FSA to the original one and add this to the group. Return to step 4;
6: return Cluster assignments, \(C\);

### 2.3.3 Individual Approach

The individual approach trains machine learning models on the survey data, with \(X^S\) as the features and \(Y^S\) as the targets. The machine learning models considered are L1 regularized logistic regression [29], random forest [4], and gradient boosting [29]. Each model is trained using 10-fold cross validation, which divides the survey data, \(\{X^S, Y^S\}\), into 10 disjoint subsets, trains a model on 9 of them, and then tests the model on the subset that was left out. This is repeated with each of the subsets as the test set; each repetition is called a fold. For the rest of the section, suppose the training set contains \(r\) survey respondents and is denoted \(\{X^R, Y^R\}\). Let the test set be denoted \(\{X^T, Y^T\}\). Let \(\hat{Y}^T\) be the vector of predicted probabilities of each of the survey respondents in the test set playing tennis.

The logistic regression predicts each element of \(\hat{Y}^T\) based on the logistic function:

\[
\hat{Y}^T_i = \frac{1}{1 + e^{-(\beta_0 + X^T_i \beta)}}
\]  

(2.1)

To fit the model, constant \(\beta_0\) and row vector \(\beta\) are parameters that are tuned in order to minimize some
error function. The error function for the L1 regularized logistic regression is:

\[
\min_{\beta_0, \beta} \frac{1}{2} \sum_{i=1}^{r} (Y_i - \beta_0 - X^R_i \beta)^2 + \lambda |\beta| \tag{2.2}
\]

The scalar \( \lambda \) is a regularization parameter that is chosen prior to fitting the model [29]. When we run the individual approach we fit five different L1 regularized logistic regression models, one for each value of \( \lambda \in \{0.0001, 0.001, 0.01, 0.1, 1\} \).

Both the random forest and gradient boosting model are ensemble methods, which combine many weak prediction models to create a single model with high prediction accuracy.

The random forest trains \( n \) decision trees on various samples of size \( r \), with replacement, from \( \{X^R, Y^R\} \). The resulting predictions for each member of the test set, \( \hat{Y}^T \), is the mean prediction over all \( n \) trees [4]. We run five different random forest models, one for each value of \( n \in \{100, 250, 500, 750, 1000\} \).

The gradient boosting model is also an ensemble method of decision trees. We use \( w \) decision trees as the weak learners in our gradient boosting model. The gradient boosting model is iterative; at each stage, \( m \), of the model, the decision tree, \( T_m \), is updated according to some loss function. In our case, use a squared error loss function, \( L(\hat{Y}^R) = \frac{1}{2}(Y^R - \hat{Y}^R)^2 \), where \( \hat{Y}^R \) are predictions of \( Y^R \) produced in training the model.

To begin the process of building the gradient boosting model, we fit two decision trees \( T_1 \) and \( T_1' \). The first decision tree, \( T_1 \) is fit to predict \( Y^R \), resulting in predictions \( T_1(X^R) \). The second decision tree, \( T_1' \), is fit to predict the “residual” which is the negative of the derivative of the loss function:

\[
- \frac{\partial L}{\partial T_1(X^R)} = (Y^R - T_1(X^R)). \tag{2.3}
\]

The resulting predictions from \( T_1' \) are \( T_1'(X^R) \). We then iterate on both \( T_1 \) and \( T_1' \) with a shrinkage parameter, \( v = 0.01 \), with the following two step process:

1. \( T_{m+1} \) is fit to predict \( T_m(X^R) + v T_m'(X^R) \).
2. \( T_{m+1}' \) is fit to predict \( Y^R - T_{m+1}(X^R) \)

This iterative process continues until we have the final decision tree, \( T_w \). The shrinkage parameter controls the learning rate by shrinking the size of the gradient step. As with the logistic regression and random forest models, we train five different gradient boosting models, each using a different number of
decision trees, \( w \in \{100, 250, 500, 750, 1000\} \).

The machine learning models are compared to a naive prediction model, which sets the percentage of tennis players in the test set to be the percentage of tennis players in every cluster in the training set. See Algorithm 2 for the complete procedure of the individual approach.

**Algorithm 2 Individual Approach.**

**Input:** Cluster of each survey respondent; Survey data, \( X^S \) & \( Y^S \);

**Output:** Final individual model;

1. If there is a cluster with fewer than 20 people, remove the respondents in that cluster from \( X^S \);
2. Partition the remaining clusters into 10 folds;
3. \textbf{for} \( f := 1 \) to 10 \textbf{do}
4. Create test set with survey data belonging to clusters in fold \( f \);
5. Create training set with survey data belonging to clusters in remaining folds;
6. Train naive model on training set & predict on the test set;
7. \textbf{for} \( \lambda \in \{0.0001, 0.001, 0.01, 0.1, 1\} \) \textbf{do}
8. On training set, train L1 regularized logistic regression with regularization parameter \( \lambda \);
9. \textbf{end for}
10. \textbf{for} \( n \) in \( \{100,250,500,750,1000\} \) \textbf{do}
11. On training set, train random forest model with number of trees = \( n \);
12. \textbf{end for}
13. \textbf{for} \( w \) in \( \{100,250,500,750,1000\} \) \textbf{do}
14. On training set, train gradient boosting model with number of weak learners = \( w \) and learning
rate = 0.01;
15. \textbf{end for}
16. For each cluster, find the mean prediction over all individuals in the test set for every trained
model;
17. Calculate MSE between mean predictions & true proportion of tennis players in each cluster in
the test set for each model and hyperparameter;
18. \textbf{end for}
19. Record the model and hyperparameter that lead to the lowest median MSE over all folds;
20. Train a final model on all the data using this model and hyperparameter;
21. \textbf{return} Final individual model;

In each fold of the individual approach, the mean squared error (MSE) between the predicted percentage of tennis players in each cluster of the test set is compared to the true percentage of tennis players in each cluster in the test set (Step 17). To calculate MSE, suppose that for a given fold, \( \hat{y} \) is a vector of predictions of the proportion of tennis players in each cluster, and \( y \) is a vector of the true proportion of tennis players in each cluster. If there are \( n \) clusters in this fold, then the MSE is calculated simply by \( \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \).

Whichever model and hyperparameters leads to the lowest median MSE over the 10 folds of Algorithm 2 is chosen as the final individual model in Step 20. The resulting MSEs from a sample run of Algorithm 3 are displayed in Figure 2.7. Note that over the 10 folds, the error is essentially the same for many different models. However, because any choice between these models with the same error is arbitrary,
we select the model that has the lowest median error as the final model. If there is a draw then the model closest to the left of Figure 2.7 is selected.

![Figure 2.6: MSE over 10 folds of every model in a sample run of Algorithm 2. The numbers next to the model choices are choices of hyperparameters.](image)

### 2.3.4 Aggregate Approach

The same machine learning models and hyperparameters are considered for the aggregate approach as are used in the individual approach. In addition to these models we train a weighted naive model. Since each observation in the the aggregate approach is comprised of aggregate data from various numbers of respondents, the weighted naive model uses the number of respondents comprising each observation to calculate a weighted mean, representing the proportion of tennis players in the training set. This is then set to be the predicted proportion of tennis players in every cluster of the test set, in the same fashion as the naive model. The data points in the three machine learning models are also weighted according to the number of respondents that constitute each point. Algorithm 3 details the aggregate approach.

Note that the aggregate measure used for age and income is median, but for the other variables in $X^S$, mean is used. This reflects the aggregation of the census data; recall that age and income from the census data, in Subsection 2.2.2, are median values, while the other variables are proportions of the population.

The resulting MSEs from Step 17 of a sample run of Algorithm 3 are displayed in Figure 2.7. The
Algorithm 3 Aggregate Approach.

Input: Cluster of each survey respondent; Survey data, \( X^S \) & \( Y^S \);

Output: Final aggregate model;

1. If there is a cluster with fewer than 20 people, remove the respondents in that cluster from \( X^S \);
2. Calculate median age and income of survey data in each remaining cluster;
3. Calculate mean of every other variable in survey data in each cluster;
4. Partition this aggregate dataset into 10 folds;
5. for \( f := 1 \) to 10 do
6. Create test set with fold \( f \) and training set with remaining folds;
7. Train naive model and weighted naive model on training set & predict on the test set;
8. for \( \lambda \) in \( \{0.0001, 0.001, 0.01, 0.1, 1\} \) do
9. On training set, train L1 regularized logistic regression with regularization parameter \( \lambda \);
10. end for
11. for \( n \) in \( \{100, 250, 500, 750, 1000\} \) do
12. On training set, train random forest model with number of trees = \( n \);
13. end for
14. for \( w \) in \( \{100, 250, 500, 750, 1000\} \) do
15. On training set, train gradient boosting model with number of weak learners = \( w \) and learning rate = 0.01;
16. end for
17. For each trained model, calculate MSE between predictions on the test set and true values in the test set;
18. end for
19. Record the model and hyperparameter that lead to the lowest median MSE over all folds;
20. Train a final model on all the data using this model and hyperparameter;
21. return Final aggregate model;

same technique to choose the final model is used in this case as in Algorithm 2.

Figure 2.7: MSE over 10 folds of every model in a sample run of Algorithm 3. The numbers next to the model choices are choices of hyperparameters.
2.3.5 Inputting Census Data to Models

Recall that the census data was formatted to match the survey data in Section 2.2. Hence, census data, \( \mathbf{X}^C \), can be inputted into both the final individual and aggregate models resulting from Algorithms 2 and 3, which were trained on survey data \( \{ \mathbf{X}^S, \mathbf{Y}^S \} \). The resulting outputs will be predictions of the proportion of tennis players in every FSA. These predictions will be validated in Subsection 2.3.6.

2.3.6 Validation Procedure

Most FSAs have fewer than 5 respondents; from this information there is no way to know the true proportion of tennis players in an FSA. However, because the clusters built in Algorithm 1 actually cluster FSAs rather than survey respondents, we can compare the predicted proportion of tennis players in all the FSAs in each cluster to the true proportion of tennis players in each cluster. This validation technique is outlined in Algorithm 4.

**Algorithm 4 Validation Procedure.**

**Input:** Predictions of number of tennis players in every FSA from final aggregate model; Predictions of number of tennis players in every FSA from final individual model; Population of each FSA; Cluster each FSA belongs to; Survey data targets, \( \mathbf{Y}^S \);

**Output:** Mean Squared Error between predictions and true value for individual model (\( \text{MSE}_i \)), aggregate model (\( \text{MSE}_a \)), and naive model (\( \text{MSE}_n \));

1: Calculate the true percentage of tennis players in each cluster;
2: With predictions from aggregate model, calculate predicted proportion of tennis players in each cluster, weighted by population of FSAs;
3: With predictions from individual model, calculate predicted proportion of tennis players in each cluster, weighted by population of FSAs;
4: Let naive prediction for each cluster be the proportion of tennis players in \( \mathbf{Y}^S \);
5: Calculate MSE between all predictions and true percentage of tennis players over all clusters;
6: return \( \text{MSE}_i, \text{MSE}_a, \text{MSE}_n \);

2.3.7 Summary and Results of Validation Procedure

We use Algorithm 5 to run the entire procedure from Figure 2.4, with all the algorithms explained above. We complete 500 repetitions of this procedure since each time the clustering algorithm from Algorithm 1 is run, the resulting clusters and partition of ten folds may be different.

The final models chosen in Step 19 of Algorithm 2 and Step 21 of Algorithm 3 may differ on every repetition of Algorithm 5. Figures 2.8 and 2.9 illustrate the models chosen as the final model for all 500 repetitions. It is evident from Figure 2.8 that no one model in the individual approach is dominant, and the naive model is never chosen. This justifies that it is important to combine many models, since the
**Algorithm 5** Entire Validation Procedure.

**Input:** Census data, $X^C$; Survey data, $X^S$ & $Y^S$.

**Output:** Mean Squared Error between predictions and true value for each rep for individual model ($MSE_i$), aggregate model ($MSE_a$), and naive model ($MSE_n$);

1. **for** $rep := 1$ to $500$ **do**
2. Create clusters of FSAs using clustering heuristic from Algorithm 1;
3. Run Algorithm 2 to output the final individual model;
4. Run Algorithm 3 to output the final aggregate model;
5. Input census data, $X^C$ into final aggregate and individual models & obtain predictions for proportion of tennis players in each FSA from each model;
6. Run validation procedure from Algorithm 4 and record $MSE_i$, $MSE_a$, and $MSE_n$;
7. **end for**
8. **return** $MSE_i$, $MSE_a$, and $MSE_n$ for each rep;

best model differs depending on the assignment of clusters and partition of folds. While in the aggregate approach, it is clear that L1 logistic regression with $\lambda = 0.01$ is the favoured model, the others still play an important role, comprising over half of the choices. In the aggregate approach, the naive is chosen only twice, and the weighted naive is never chosen.

![Individual model breakdown over 500 reps](image)

Figure 2.8: Final model choices from 500 repetitions of the individual model. N = naive, LR = logistic regression, RF = random forest, GB = gradient boosting.

It is evident from Figure 2.10 that the median MSE from the individual model and aggregate model is less than the median MSE from the naive model. We perform three sign tests to see if there is significant difference between the MSE from 500 repetitions between the individual and naive models, aggregate and naive models, and individual and aggregate models. The results show that the aggregate model has MSE significantly less than both the individual model (p value 0.002) and the naive model.
Aggregate model breakdown over 500 reps

Figure 2.9: Final model choices from 500 repetitions of the aggregate model. N = naive, wN = weighted naive, LR = logistic regression, RF = random forest, GB = gradient boosting.

(p value $2.61 \times 10^{-54}$). The aggregate model has significantly less error than the individual model (p value $2.29 \times 10^{-28}$).

MSE over 500 reps

Figure 2.10: Box plot of mean squared error for each of 500 repetitions from Algorithm 5 with outliers removed. The same plot containing outliers can be seen in Appendix A in Figure A.10

While it is evident that the aggregate model does not perfectly predict the proportion of tennis players in each FSA, it does a significantly better job than the assumption that the same proportion of tennis players live in every FSA. Considering the high cost associated with starting a new program...
or building a new facility, the small difference in accuracy between the aggregate model and the naive model can result in large savings.

In order to calculate a final prediction for the proportion of tennis players in each FSA, we take the median prediction, from the final model chosen, over the 500 repetitions of the aggregate approach in Algorithm 5. The distribution of resulting predictions over all FSAs is displayed in Figure 2.11. These predictions are then multiplied by the population in each FSA to get the estimated number of tennis players in each FSA. This information can be used to help inform decision-makers at Tennis Canada and in the amateur tennis community about the distribution of demand for tennis geographically.

![Distribution of Final Aggregate Model Predictions](image)

**Figure 2.11:** Distribution of median predictions from aggregate model over 500 repetitions.

The methods and validation used in this section are clearly distinct from those in the previous sport management literature reviewed in Section 2.1. The benefit of using these novel methods is that we are able to predict demand on a geographical level. This ties back to our objective of helping decision-makers make more targeted decisions regarding creating new programming or building new facilities.

To show that our predictions are useful, we illustrate in Section 2.4 the maps and dashboards intended to be used by decision-makers at Tennis Canada and in the tennis community to understand supply and demand. If the decision-makers have more directed objectives, they can input the predictions into optimization models as in the case study in Chapter 3 to find locations for new programming or new facilities.
2.4 Presentation to Decision-Makers

Our objective to predict demand geographically is motivated by the need of sport managers to make more targeted policy decisions concerning sport participation. To present our findings to decision-makers, we display the predicted demand from Section 2.3 on a map overlaid with locations of tennis courts.

The novelty of using this geographical demand prediction to inform policy-makers is that policy-makers are able to understand supply and demand at a much more granular level. For instance, a study by Wicker et al [52] suggests that there may be an oversupply of tennis courts in Germany, which may discourage a policy-maker from building any more tennis facilities in Germany. A more granular geographical understanding of supply and demand can help the policy maker understand whether the entire country is in fact oversupplied with courts. By combining the geographical prediction of demand with the location of current facilities in a map, the policy maker can see the precise locations where there is an over supply and target the locations where the supply is low.

2.4.1 Map

The final predictions from the aggregate model can be used by decision-makers to make targeted decisions. Figure 2.12 shows the map of demand in each FSA overlaid with locations of tennis courts. This map was built in Tableau 10.2 [45]. A benefit of using Tableau is that it is a user friendly software; when the court data is updated in the future, inputting this new information will be seamless for the user. In addition, this map is interactive; the user can zoom in to areas of interest and it is possible to view only one type of court if desired.

2.4.2 Dashboard

While the map is useful on its own, it has some obvious drawbacks. If the user has a place of interest, for instance a location where he/she would like to build a court, it is easy to look at that area and see how many courts there are currently, what types of courts they are, and what the current demand is. However, it is difficult to decipher from the map which locations are best to consider if the user does not have a specific location in mind. In addition, if the user would like to find similar FSAs or compare two FSAs, there is no clear way to do this with the map. The user also cannot see the proportion of tennis players in a region, so it can be difficult to discern if there are many tennis players or just a large population. The dashboard shown in Figure 2.13 addresses these problems.
Figure 2.12: Map of predicted demand for tennis (in number of people) by FSA and locations of courts. Darker FSAs have more tennis players than lighter FSAs. Tennis courts are sorted into 3 categories: outdoor, bubble (an outdoor court with a bubble that goes over it in the winter), and permanent (indoor courts). The size of the dot illustrates how many courts are at each location; the larger the dot, the more courts.

The dashboard is broken into two related figures. The figure on the right is the same map as in Figure 2.12. The figure on the left is a scatter plot of FSAs with population on the y-axis and the predicted proportion on tennis players in the FSA on the x-axis.

To help users focus on size and density of the region, FSAs are classified to be in urban or rural regions. Knowing that playing tennis is positively correlated with living in rural regions, from the study by Eime et al [22], a decision-maker may only want to focus on these regions. With this in mind, we can see that decision making in urban and rural regions may need to be approached from different angles, and so the user should be able to choose which type of region he/she wants to see. In addition, our metrics may have different meanings depending on whether the FSA is urban or rural. This is because the centroid of a small FSA is a much better approximation of the location in which the population of the FSA live than the centroid of a large FSA. We can see from the map that larger FSAs tend to be rural and smaller FSAs tend to be urban. According to Canada Post, postal codes with zero as the second digit are rural, while all other postal codes are urban [42]. We add this information to the dashboard.

It should be noted that the Canada Post definition of urban and rural is not entirely consistent with the Statistics Canada definition of urban and rural. Statistics Canada states that rural populations are those living outside centres with both populations less than 1,000 and density less than 400 people per square km [8]. However, New Brunswick, for instance, does not have any rural postal codes, but 48% of the population lived in rural regions in 2011 according to Statistics Canada [9]. Since we have no
way to calculate whether an FSA is urban or rural based on the Statistics Canada definition, we use the Canada Post definition.

Understanding Supply and Demand from Dashboard

The dashboard is built with the goal of being easy for a sport manager to use to analyse supply and demand for tennis. To assist the user’s understanding of supply, the size of the dot in the scatter plot is proportional to the distance from the centroid of the FSA to the closest tennis court. This feature is particularly useful when the user wants to find undersupplied areas with high demand.

The upper right quadrant of the scatter plot contains FSAs with high demand since both the proportion of tennis players and population in each FSA are above average. Hence for users to find areas with high demand and low supply, or more specifically areas to build tennis courts, they can look in the
upper right quadrant for FSAs with large dots.

An interesting anecdote is that this procedure can also be used to find missing courts in the database. Figure 2.14 shows four FSAs selected with high predicted percentage of tennis players but large distances to courts. After following up on these FSAs to confirm that they do not have nearby courts, we found that they each, in fact, have at least one court that was missing from our database. While this shows that the dashboard is limited without comprehensive data, it also shows that the tool works; these clearly would be target locations if they did not have courts.

Figure 2.14: Example of targeted FSAs that actually have courts. From top to bottom: T3M (Auburn Bay, AB), K8H (Petawawa, ON), V8J (Prince Rupert, BC), P8T (Sioux Lookout, ON).

Like the map, this dashboard is also interactive. An FSA can be selected on the map and it will be displayed on the scatter plot or vice versa. When an FSA is selected, additional information is given to the user such as population, predicted proportion of tennis players, and minimum distance to a tennis court (Figure 2.15). A benefit of using the dashboard as opposed to solely the map is that both the predicted proportion of tennis players and predicted number of tennis players in each FSA can easily be seen.

Example: Fredericton

In a discussion with Tennis Canada we were asked whether the tool we built could indicate whether Fredericton had a predisposition for tennis. Our connections at Tennis Canada told us that historically there has been very little tennis in Fredericton, New Brunswick, with very few clubs and low membership.
A few years ago, some local advocates raised money to build a six court indoor facility in the city ([13]), but it took a long time for tennis to become popular in the city.

To check whether Fredericton has low demand for tennis, we selected all the FSAs in the city and the surrounding areas, which can be seen in Figure 2.16. We found that all seven of the FSAs in and surrounding Fredericton have lower than average predicted proportion of tennis players, indicating that the demand in Fredericton is low.

The map and dashboard presented in this chapter can be effectively used to display our demand prediction. With a general managerial inquiry about where to put new programming or courts, this dashboard can supplement the decision-maker’s knowledge of supply and demand for tennis and help them make informed decisions.

If the decision-maker has a specific objective, such as wanting to locate courts at five of ten possible
Figure 2.16: Dashboard with the seven FSAs in Fredericton and surrounding areas selected. Note that each of these FSAs has below average predicted proportion of tennis players (seen in scatter plot).

locations, then the best approach is to use an optimization model with the demand prediction as input rather than to use these descriptive maps. Chapter 3 presents a detailed case study on how sport managers can use the geographical understanding of demand we developed in this chapter to make targeted business decisions through optimization modelling.
Chapter 3

Facility Location Case Study

This case study illustrates how the results from the demand prediction can be used by NSFs to make directed decisions that aim to increase participation in sport. The ability of our demand prediction to be used for this purpose extends beyond the previous literature, using the quantitative results from our statistical analysis to make optimal decisions.

In addition to showing how the demand prediction can be used in an optimization problem, we demonstrate (1) that the case study is motivated by a real sport management problem and (2) that the resulting solutions from using the demand prediction at the FSA level are different from the resulting solutions using strictly population of FSAs. This will illustrate that the results from Chapter 2 can be used in a meaningful way to guide managerial decisions in amateur sport.

This chapter will first review the motivation for the case study, inspired by a real business problem of Tennis Canada, in Section 3.1. Section 3.2 will review literature on the relationship between sport participation and location of facilities, as well as literature on the relevant optimization models we will use. The following section (3.3) presents the optimization model used in this case study. Then Section 3.4 shows that the optimization model produces different results using the demand prediction versus the naive demand (i.e., population alone). Lastly, Section 3.5 demonstrates how this model can be easily adapted to meet different objectives of Tennis Canada.
3.1 Background

A long-term goal of Tennis Canada is to increase the participation of Canadian children in tennis. Two barriers to achieving this goal include (1) lack of access to tennis facilities, particularly in the winter and (2) perceptions that tennis is an expensive sport. To understand to what extent these factors are barriers to children’s involvement in tennis, Tennis Canada asked questions about what prevents parents from enrolling their children in tennis in the Canadian Tennis Brand Health Studies [14]. Of parents whose kids do not play tennis, 10.9% list lack of access to facilities or programs as the primary reason, while 6.7% of them list the high price-tag as the main deterrent. The other common responses were that their children prefer other sports, are not interested in sports, do not have time to play tennis, or that the parents themselves do not play tennis. While 17.6% of parents may seem like a small group to target for new programming, it is evident that parents seeking affordable, accessible programming are the largest group interested in getting their children involved in the sport. This is evident from a further question in the survey, for which 17.6% of parents whose kids are not enrolled in tennis say that if there were programs for their child’s age or skill level they would consider enrolling them, and 15.5% said that if there were more affordable classes they would consider enrolling them [14]. Overcoming these two barriers is extremely important for Tennis Canada to expand participation in the sport among children, hence they wanted to design a program that would help remove these barriers for many families.

The large shortage of affordable tennis programming in Canada happens during the winter months, since the weather in Canada prohibits the use of outdoor courts. It is important that children continue playing tennis throughout the winter season so they do not lose the skills they have learned or lose interest in the sport. This is a problem that Tennis Canada struggles with because tennis at indoor facilities can be very expensive, due to many of them being privately owned and not readily located. However, tennis does not necessarily have to be played in indoor tennis facilities, particularly when children are learning the fundamentals of the sport.

Kids Tennis is a widely used tennis development program intended for introducing young children to the sport, and is fully endorsed by Tennis Canada [12]. The program introduces children to tennis by using smaller courts and racquets, and large, soft balls that do not bounce as high as regular tennis balls. The children are taught the fundamentals of the sport and their skills are developed much faster playing Kids Tennis, which allows for an easy transition to a full court when they are ready [12]. Another benefit of Kids Tennis is that it does not have to be run on a tennis court; programs can be held in school gyms,
baseball pitches, or even on grass or turf. By holding Kids Tennis programs at indoor turf facilities, Tennis Canada will be able to significantly reduce the cost of programming due to the large number of children who can play tennis in the space at a given time and the lower value of the space. The large capacity of indoor turf facilities increases the accessibility of tennis in the winter to many children living in surrounding areas. Hence, holding Kids Tennis programs at indoor turf facilities will be beneficial for Tennis Canada to expand participation among kids. For the remainder of this chapter we will refer to these particular programs as “Serve and Turf” programs.

If Tennis Canada were to hold the Serve and Turf programs, they would need to know which turf facilities are the best locations to do this. In this chapter, we show that the demand prediction we built in Chapter 2 can be seamlessly used to help guide this decision. Further, we show that in certain circumstances, this demand prediction performs better than locating facilities where there is highest population.

3.1.1 Data

In the case study, we use the data from Section 2.2 as well as data about the indoor turf facilities in the eastern provinces of Canada. To collect data on the turf facility locations we searched on Google for any databases or lists of indoor artificial turf facilities in Canada, but no such lists exist. Without an easily accessible list of turf facility locations in Canada, we decided to limit this case study to the eastern provinces of Canada: Ontario, Quebec, New Brunswick, Prince Edward Island, Nova Scotia, and Newfoundland and Labrador.

To find the turf facility locations in these provinces, we ran searches on Google maps in these provinces such as “indoor soccer” and “indoor turf” and collected the resulting turf names and locations. We then manually checked to see if these locations were in fact indoor turfs or outdoor turfs that have a bubble over them in the winter. In addition to this, we searched Google for soccer clubs in these regions and checked their websites manually to see where they play indoor soccer. We geocoded the addresses of the turfs to get the latitude and longitude coordinates and then calculated the UTM coordinates.

Note that in the following analysis we assume that all turf facilities have the same capacity, since many are the size of a single soccer field. Turf facilities with more than one soccer field, and thus larger capacity, could be easily worked into the models, however for simplicity we assume they all have a single field.
3.2 Literature Review

The literature review in Chapter 2 highlighted that previous research has not looked at understanding participation in sport at a population level. Because this has not been done in the past, there has not been any research about applying population level information to inform decisions on where to place new facilities. However, there has been relevant past research looking at the relationship between the location of sporting facilities and participation in sport.

Previous research [52, 51, 38] shows that close proximity to various sporting facilities has a positive impact on sport participation. A study by Hallmann et al showed that different sized communities have different requirements for the types of sport facilities that will positively impact participation the most [31]. A policy recommendation resulting from this study was to place swimming pools in large metropolitan areas or sports fields in medium-sized municipalities. For policy-makers, this recommendation will be useful when deciding which facilities to build. Rather than looking at which types of facilities to build, as the previous research does, our case study looks at where to locate new programs in pre-existing facilities. To do this, we use the demand prediction from Chapter 2 combined with optimization methods.

The optimization model we use in our case study is based on the Maximal Covering Location Problem (MCLP) introduced by Church and ReVelle in 1974 [17]. The objective of the MCLP is to “maximize coverage (population covered) within a desired service distance \( S \) by locating a fixed number of facilities” [17]. In our case study, we aim to maximize the number of potential tennis players we can reach (based on the demand prediction) within a certain distance by choosing a fixed number of turf facilities to hold the Serve and Turf programs.

Covering Location Problems similar to the MCLP have been used in an assortment of applications [25, 24]. Some applications include: locating emergency medical services [37], locating distribution centres for humanitarian relief [2], and locating distribution centres in e-commerce [24]. To our knowledge, this case study is the first application of a covering location problem in the sports industry.
3.3 Maximizing Demand Covered

3.3.1 Creating Demand Nodes

The MCLP requires knowledge of the location and number of people demanding a service at specific points called demand nodes. From our prediction, we know the demand in each FSA. Demand nodes could be chosen to be the centroids of each FSA, but there is the possibility that a facility could cover a large portion of an FSA but not the centroid, and hence would not count any of the demand from that FSA. In fact, seven turf facilities are not within 5 km from any FSA centroids - including the FSA in which they are located. Clearly demand points must be more granular.

Because every FSA is composed of postal codes, we choose to use postal code centroids as demand nodes in the MCLP. Note that the problem explained above could also occur here, as some postal codes are also large in size and demand could be missed if the postal code centroid is not covered. However, Figure 3.1 shows that most turf facilities have at least 1,000 postal code centroids within 5km, and only four facilities have fewer than 100 surrounding postal code centroids. This shows that using postal code centroids as demand points should be sufficient.

![Figure 3.1: Number of postal code centroids within 5km of each indoor turf facility in eastern Canada.](image)

To calculate the demand at each postal code we first must estimate the population at each postal code. To estimate the population at postal code $i$ we divide the population of the FSA in which $i$ belongs by the total number of postal codes that make up that FSA. This is the best estimate we can make.
for population at the postal code level because the census data does not include population of postal codes. We then multiply this estimated population of postal code $i$ by the predicted proportion of tennis players in the FSA in which $i$ belongs, calculated in Chapter 2, to get the number of people demanding tennis at the centroid of postal code $i$.

### 3.3.2 General Model

Let $i$ index the $m$ demand nodes in the set of postal codes, $P$. Let $j$ index the $n$ potential facilities in the set of potential facilities, $F$. Then $h_i$ is the demand at demand node $i$. Let $c$ be the number of distinct Serve and Turf programs that Tennis Canada plans to hold. We set $R_i$ to be the maximum distance from home that parents living in demand point $i$ are willing to send their children to an extra-curricular program.

The location of each demand node and facility is in UTM coordinates. Recall from Subsection 2.2.3 that the Euclidean distance between UTM coordinates is the direct distance in meters. We use this distance as an approximation; arguably Manhattan distance is more accurate since this research is concerned with driving distance. Let $d(i, j)$ be the Euclidean distance between the coordinates of demand node $i$ and facility $j$. Matrix $a \in \{0, 1\}^{m \times n}$ has binary elements, $a_{ij}$, defined in Equation 3.1.

$$a_{ij} = \begin{cases} 
1 & \text{if } d(i, j) \leq R_i \\
0 & \text{if } d(i, j) > R_i 
\end{cases} \quad (3.1)$$

The decision variables are the vectors $z \in \{0, 1\}^m$ and $x \in \{0, 1\}^n$ with elements defined by Equations 3.2a and 3.2b.

$$z_i = \begin{cases} 
1 & \text{if demand node } i \text{ is covered} \\
0 & \text{else} 
\end{cases} \quad (3.2a)$$

$$x_j = \begin{cases} 
1 & \text{if facility } j \text{ is chosen} \\
0 & \text{else} 
\end{cases} \quad (3.2b)$$

Equation 3.3 outlines the optimization model.
Chapter 3. Facility Location Case Study

\[ \text{max} \sum_{i=1}^{m} h_i z_i \quad (3.3a) \]

s.t. \[ z_i \leq \sum_{j=1}^{n} a_{ij} x_j, \forall i \in P \quad (3.3b) \]
\[ \sum_{j=1}^{n} x_j \leq c \quad (3.3c) \]
\[ z_i \in \{0, 1\}, \forall i \in P \quad (3.3d) \]
\[ x_j \in \{0, 1\}, \forall j \in F \quad (3.3e) \]

The objective function 3.3a, maximizes the total demand met by the turf facilities that are chosen to host Serve and Turf programs. Constraint 3.3b allows a demand node to be covered only if it is within a certain distance from a chosen facility. Constraint 3.3c limits the number of chosen facilities to \( n \), and 3.3d and 3.3e are integrality constraints.

3.3.3 Parameters

For our application, the two parameters that we must specify to solve optimization model 3.3 are the number of Serve and Turf programs to host, \( c \), and the distance that a child from demand point \( i \) is willing to travel to a program, \( R_i \). The Serve and Turf program is a pilot project, so Tennis Canada would only host a handful of new programs. We analyze results from a range of 1 to 12 new programs.

To determine reasonable values for each \( R_i \), we asked Tennis Canada to include a question about the distance parents are willing to travel in the 2016 survey. The survey asked parents how many minutes they are willing to travel to their child’s extra-curricular activities. The distribution of parents’ responses is plotted in Figure 3.2. We looked at the distribution of responses for each province separately as well, and the median response for every province was 30 minutes, except for Quebec which was 20 minutes.

For simplicity, we will assume that parents are willing to travel 30 minutes, based on the median response. We then convert this time to a distance, to approximate \( R_i \), by multiplying it by an estimate for the parents’ average driving speed. The mean driving speed in Toronto fluctuates daily between 24 km/h and 32 km/h [28], so we estimate the mean driving speed in the eastern provinces to be 30 km/h. This is a conservative estimate, reflecting the off-peak driving speeds in Toronto, since Toronto will have lower mean driving speeds than other regions based on its high density. Hence, in the following examples,
assuming that $R = R_i \ \forall i \in P$, we choose $R = 15\text{km}$ to be the maximum distance from home that parents will send their children to a Serve and Turf program. Note that this estimate for distance that parents are willing to travel is extremely simplified; it is unlikely that the maximum distance parents are willing to travel is the same everywhere, and better data on this could help determine $R_i$ more accurately.

Figure 3.2: Time parents from the 2016 survey are willing to travel for their children’s extra-curricular activities.

### 3.4 Validation Method & Results

In order to see how using our demand prediction is useful in the facility location context, we must compare it to facility location with the naive prediction of demand. We define two facility location models, Facility Location Demand Prediction (FLDP) and Facility Location Naive (FLN). FLDP uses the predicted demand at postal code $i$ (calculated in 3.3.1) as the cost $h_i$, while FLN uses the percentage of tennis players in the eastern provinces, 13.7%, multiplied by the estimated population at $i$ as $h_i$. Aside from this change in the cost of the objective function, the models are identical. Both FLDP and FLN can be written as functions of $c$ and $R$; for instance, FLDP($c$,$R$) is the facility location model with $h_i$ as the predicted demand at demand node $i$, $c$ as the number of programs to locate, and $R$ as the maximum distance parents will send their kids to a program.

To validate that FLDP covers more demand than FLN, we run the facility location validation in
Algorithm 6. In this validation method, we calculate the number of tennis players in the survey data covered by FLDP ($t_{dp}$), the number covered by FLN ($t_n$), and the percentage of covered respondents who are tennis players for FLDP and FLN ($s_{dp}$ and $s_n$ respectively). Figures 3.3 and 3.4 show the results of this method.

**Algorithm 6** Facility Location Validation.

**Input:** Final predictions of number of tennis players in every FSA from aggregate model; Naive prediction of number of tennis players in every FSA; Survey data; Location of postal code centroids;

**Output:** Number of survey respondents covered by FLDP for each number of programs, $s_{dp}$, and number of tennis players covered, $t_{dp}$; Number of survey respondents covered by FLN for each number of programs, $s_n$, and number of tennis players covered, $t_n$;

1: Distribute the demand uniformly for each FSA among its postal codes;
2: for $c := 1$ to 30 do
3:  Run FLDP($c$, 5km) and record $p_{dp}$, the demand points covered;
4:  Record $s_{dp}$, the number of survey respondents who live in demand points $p_{dp}$ and the number of these survey respondents who are tennis players, $t_{dp}$;
5:  Run FLN($c$, 5km) and record $p_n$, the demand points covered;
6:  Record $s_n$, the number of survey respondents who live in demand points $p_n$ and the number of these survey respondents who are tennis players, $t_n$;
7: end for
8: return $s_{dp}, t_{dp}, s_n, t_n$;

Figure 3.3: Number of tennis players covered by facilities selected by FLDP and FLN for various numbers of Serve and Turf programs.

Examining the turf facilities that are selected with the FLDP and the FLN (Figure 3.5), we see that five of the twelve possible number of programs to locate select different turf locations between the FLDP and the FLN. It follows, in Figures 3.3 and 3.4, that only these five numbers of programs to locate obtain
Figure 3.4: Percentage of survey respondents covered by facilities selected by FLDP and FLN who are tennis players, for various numbers of Serve and Turf programs.

The results of the facility location validation show that by locating fewer than 12 Serve and Turf
programs, the FLDP does at least as well, if not better, than the FLN when measured on the survey data. While the FLDP only reaches at most four more people than the FLN, if the survey data is representative of the population, then the number of additional people reached by it could be on the order of many thousands of people (i.e., number of Canadians per survey respondent).

However, the problem with using this validation method is that the locations of survey respondents are not necessarily representative of the Canadian population. Recall from Subsection 2.2.1 that minimum target quotas were set for each province, resulting in the geography of respondents not being representative of the Canadian population. For instance, 7.9% of Canada’s population lives in FSAs starting with “M” (Toronto region) but only 2.9% of the survey respondents are from FSAs starting with “M”. On the other hand 8.9% of the survey respondents are from Newfoundland while only 1.5% of Canada’s population lives in Newfoundland [14, 15, 7]. This discrepancy between the survey data and the census data must be highlighted since it may make the results of the facility location validation less convincing.

Despite this possible survey bias issue, we take an optimistic view and assume that our prediction for demand is, in fact, the true demand for tennis. Then by definition, FLDP covers at least as many tennis players in the Canadian population as FLN covers. To see how much better the FLDP performs in comparison to the FLN, we evaluate the solutions of both the FLDP and the FLN with the FLDP objective function. From Figure 3.6, which shows the percentage of the FLDP optimal value that the FLN solution covers, it is evident that the FLN covers over 99% of the tennis players that the FLDP covers for each number of programs. However, when we examine the number of tennis players that the FLDP covers that the FLN does not (Figure 3.7), it is substantial. This is not surprising, based on the results from the facility location validation.

Note that the number of tennis players from Figure 3.7 does not directly reflect the number of additional tennis players who would join the programs, since it is reflective of the entire population (not

<table>
<thead>
<tr>
<th># programs</th>
<th>(t_{dp})</th>
<th>(t_n)</th>
<th>(s_{dp})</th>
<th>(s_n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>104</td>
<td>100</td>
<td>27.3%</td>
<td>25.4%</td>
</tr>
<tr>
<td>6</td>
<td>108</td>
<td>104</td>
<td>25.6%</td>
<td>23.9%</td>
</tr>
<tr>
<td>7</td>
<td>111</td>
<td>107</td>
<td>24.7%</td>
<td>23.7%</td>
</tr>
<tr>
<td>11</td>
<td>120</td>
<td>119</td>
<td>24.1%</td>
<td>23.4%</td>
</tr>
<tr>
<td>12</td>
<td>122</td>
<td>130</td>
<td>23.7%</td>
<td>24.4%</td>
</tr>
</tbody>
</table>

Table 3.1: Comparison between metrics \(t_{dp}\) & \(t_n\) and \(s_{dp}\) & \(s_n\) for number of programs for which \(t_{dp}\) differs from \(t_n\) and \(s_{dp}\) differs from \(s_n\). Highlighted entries cover more tennis players or higher percentage of tennis players than white entries.
just 6-11 year olds). However, by locating programs at the facilities that the FLDP chooses rather than the FLN, it is clear that more demand may be covered.

![Graph showing the percentage of the FLDP optimal value met by the FLN optimal solution for various numbers of Serve and Turf programs.](image)

Figure 3.6: Percentage of the FLDP optimal value met by the FLN optimal solution evaluated with the FLDP objective function for various numbers of Serve and Turf programs.

![Graph showing the absolute number of tennis players covered by FLDP but not by FLN for various numbers of Serve and Turf programs.](image)

Figure 3.7: Absolute number of tennis players that the FLDP covers that the FLN does not, for various numbers of Serve and Turf programs.

As an example, we choose to place six Serve and Turf programs. Figure 3.8 illustrates the six program locations that the FLDP chooses versus the FLN.
Chapter 3. Facility Location Case Study

3.5 Extensions of Case Study

Up until now we have been focused on maximizing the number of people covered by the facilities chosen. However, in reality, Tennis Canada may have a different objective. Through conversations with them, we identified some reasonable objectives. This section will address some instances of how this model can be easily adapted to meet these alternative objectives.

3.5.1 Equality Constraint

From Section 3.4 we see that all of the Serve and Turf programs are placed in Ontario and Quebec when 6 programs are to be placed. Along with the goal of increasing participation in tennis, Tennis Canada wanted to spread these programs across the country, so these results may not be desirable. To create more fair selections, we add a constraint that will ensure at least half of the programs are hosted in the Atlantic provinces. Suppose that $A$ is the set of turf facilities in the Atlantic Provinces, then the optimization problem is exactly the same as 3.3 but with following constraint added:
\[ \sum_{j \in A} x_j \geq \frac{c}{2} \]  

For the case of \( c = 6 \), the results are identical for both the FLDP and FLN. By ensuring that 3 Serve and Turf programs are in the Atlantic provinces, a turf is chosen to host a program in each of Moncton, Halifax, and St. John’s, as illustrated in Figure 3.9. Despite the fact that Ontario has two programs, and Prince Edward Island has no programs, this solution likely more accurately represents the objective of Tennis Canada.

![Figure 3.9: Red markers indicate locations of the 6 turf facilities selected by the FLDP with the equality constraint. Blue markers are all other turf facility locations in the eastern provinces.](image)

### 3.5.2 Other Extensions

**Low Proportions of Tennis Players**

Another idea that Tennis Canada mentioned to us was wanting to extend the sport to populations who do not typically play tennis. An extension of the case study would be to locate Serve and Turf programs in areas with low predicted proportions of tennis players. Previously, demand, \( h_i \), at postal code \( i \) was defined as the population, \( p_i \), times the predicted proportion of tennis players, \( \alpha_i \) at that postal code. In this new case, we simply define new demand points \( h_i' \) such that \( h_i' = \alpha_i \). The new objective function is:

\[ \min_x \sum_{i=1}^{m} h_i' z_i. \]  

All the constraints from the MCLP in Equation 3.3 from Subsection 3.3.2 remain the same. This will only locate programs in locations where there are low proportions of predicted tennis players.
Nearby Outdoor Courts

Another reasonable objective may be for Tennis Canada to locate programs at a turf facility only if there is an outdoor court nearby. This may help to increase the retention rates of the children involved with Kids Tennis because they will be able to continue playing in the summer once the Serve and Turf program has ended.

To solve this problem, we create a new matrix, \( a' \in \{0,1\}^{m \times n} \) to indicate whether turf facility \( j \) and any outdoor tennis court are within distance \( R_i \) from postal code \( i \). To explicitly define the elements of \( a' \), we define a new vector \( b \in \{0,1\}^m \) that indicates whether there is an outdoor court within \( R_i \) distance from postal code \( i \). Let \( k \) be an element in the set \( K \) of all tennis courts. We define the elements of \( b \) in Equation 3.6.

\[
b_i = \begin{cases} 
1 & \text{if } \min_{k \in K} d(i,k) \leq R_i \\
0 & \text{if } \min_{k \in K} d(i,j) > R_i 
\end{cases} \tag{3.6}
\]

With vector \( b \) defined, and \( a \) defined identically to Equation 3.1 in Subsection 3.3.2, we define the elements of \( a' \) in Equation 3.7.

\[
a'_{ij} = \begin{cases} 
1 & \text{if } a_{ij}b_i = 1 \\
0 & \text{if } a_{ij}b_i = 0 
\end{cases} \tag{3.7}
\]

The MCLP from Equation 3.3 from Subsection 3.3.2 remains the same except constraint 3.3b is altered to be:

\[
z_i \leq \sum_{j=1}^{n} a'_{ij}x_j, \forall i \in P. \tag{3.8}
\]

Budget Oriented Planning

Turf facility costs vary across Canada to rent. For instance, to rent the turf field at the Varsity Centre in Toronto in the winter, the cost is $650.45 per hour [41]. On the other hand, to rent the field at the Kenora Indoor Sportsplex, it costs $160.00 per hour [40]. It is plausible that Tennis Canada would have some set budget allotted to the Serve and Turf programs rather than a specific number to locate. Suppose that starting a new program at facility \( j \) has a fixed cost, \( f \), that is common to all new programs, and a
cost $c_j$ that varies according to the location of $j$. Then if Tennis Canada has a budget of $B$ allotted to the Serve and Turf programs in a given year, constraint 3.3c from the MCLP presented in Equation 3.3 can be changed to Equation 3.9. This will ensure the budget is not exceeded by the cost of the programs selected.

$$\sum_{j=1}^{n} (f + c_j) x_j \leq B$$

These examples illustrate that if Tennis Canada has a specific objective, such as those above, it is easy to use our demand prediction as input into an optimization model that will help them make decisions to achieve this objective. This is novel, since the previous literature on sport participation did not suggest specific locations to hold programming that will help to reach business objectives. In addition, the results of the FLDP differ from the results of the FLN, indicating that the demand prediction from Chapter 2 may be needed to determine the best locations for the programs. Of course, there are many more factors that could be incorporated into the MCLP given the appropriate data, such as availability of tennis instructors, or locations of current programming. However, we hope that this case study has illustrated that once these factors are known, they will be easy to incorporate into the model to make an optimal decision.
Chapter 4

Tools for National Sport Federations

The Canadian Olympic Committee (COC) is a not-for-profit organization that is “responsible for all aspects of Canada’s involvement in the Olympic Movement” [18]. This responsibility includes working closely with the National Sport Federations (NSFs) [18]. NSFs are the governing bodies for each sport in Canada, and they are responsible for overseeing all aspects of their sport in Canada including “managing their high performance programs,” but also developing and promoting their sport more generally [39].

The COC is holding an NSF Enhancement Initiative, which is a four-year project with the objective of helping the NSFs grow in their business operations, leadership and governance. In addition it focuses on helping NSFs with recruitment of new athletes. As part of this initiative we were asked to run two day-long sessions teaching NSFs about how they can use analytics to drive decision making. The audience included executives and staff from various NSFs.

For the sessions, we developed two tools to showcase a hands-on example of how analytics can be used to drive decisions. The objective of demonstrating these tools was to inspire the NSFs to think of specific ways they can solve their own business or sport-related problems with analytics.

Since the audience of the analytics sessions was diverse in both the sport they represented and their roles within their NSF, we were presented with the challenge of building tools that would be relevant to everyone in the audience. As a result of many sports being significantly different in nature (for instance single player sports versus team sports, or sports that depend on a geography such as skiing versus sports that can be played anywhere such as soccer), the membership in NSFs varies significantly, as do the membership objectives. The Canadian Freestyle Ski Association, for instance, is more focused on
retention of current members rather than growing membership [J. Anderson, pers. comm.], while other
NSFs, such as Swimming Canada, are focused on growing their membership base [11].

One common factor among all NSFs, however, is their interest in understanding who their members
are and where they live. Naturally, analytics is a tool that can help with this. We built two interactive
tools in Excel that decision-makers in NSFs can use to understand their membership. The following
section outlines the details of the tools we built and shows some figures of screenshots of the tools.
Section 4.2 explains a follow-up tool we build for one NSF, and reviews the feedback we received about
it and the analytics workshops.

4.1 Tools

The tools we built reflect the priority of many NSFs to understand their membership. Whether the NSF
is focused on retaining members or growing their membership, we make the assumption that it will be
useful for them to see where people live who are similar to their current members. By knowing where
people live who are similar to their current members, NSFs can effectively target recruitment of new
members. If their focus is on retention, they can search for locations with people who are similar to
their long-term members, while if their focus is on growing membership they can search for locations
with people who are similar to all their current members.

Based on the literature reviewed in Section 2.1, we know that demographic characteristics are pre-
dictive of participation in sport, so we define similarity between members and the population based
on demographic characteristics. While the two tools are similar in this sense, the inputs for each are
distinct, based on what information the NSF currently knows about their members. If the NSF knows
typical demographic characteristics of their members, they can search for locations with populations
that have high proportions of these characteristics with the Characteristic Search Tool (CST), explained
in Subsection 4.1.1. On the other hand, if the NSF knows locations in Canada where many of their
members live, they can use the Location Search Tool (LST), explained in Subsection 4.1.2, to search
for similar locations. Subsection 4.1.3 reviews the dashboard that supplements both tools, which shows
how the demographic characteristics of the population in areas that the NSFs members live compare to
the characteristics of Canada’s population.
Variables | # Characteristics | Characteristics
---|---|---
Gender | 2 | female, male
Age | 18 | 5 year ranges
Marital status | 6 | married, common law, single, separated, divorced, widowed
Official languages spoken | 4 | French, English, both, neither
Immigration status | 7 | non-immigrants, non permanent residents, immigrants (5 ranges for date of immigration)
Ethnic origins | 8 | each continent, Caribbean, North American, Aboriginal
Religion | 9 | Buddhist, Christian, Hindu, Jewish, Muslim, Sikh, Traditional Spirituality, Other, None
Education | 7 | no certificate, high school, trades diploma, college, university below bachelor’s, bachelor’s university degree above bachelor’s
Occupation | 10 | (see NOC 2011 [6])
Industry | 20 | (see NAICS 2012 [10])
Household income | 13 | no income or 12 ranges of income

Table 4.1: Variables that the user is able to consider in the CST.

### 4.1.1 Characteristic Search Tool

The Characteristic Search Tool (CST) finds FSAs in Canada whose population is similar to a user inputted target person. The users input demographic characteristics (from Table 4.1) representing a target individual or population, similar to their members. The output of the tool is a rank ordered list of FSAs in Canada based on a measure of similarity to the target. Figure 4.1 shows a screenshot of the CST where the user is looking for FSAs with high proportions of males ages 20 to 29. Note that the user can select any combination of characteristics to consider and does not have to choose each characteristic specific to a given variable.

To calculate similarity between target characteristics and every FSA, we define variables as follows.

Let \( c \in \{0, 1\}^n \) be a vector indicating the user’s selection of characteristics, where \( n = 104 \) is the number of possible demographic characteristics that can be selected. Each element \( c_i = 1 \) if characteristic \( i \) is of interest and \( c_i = 0 \) if \( i \) is not of interest. For each of \( m = 1607 \) FSAs, indexed by \( j \), let \( p_{ij} \in [0, 1] \) be the proportion of people in FSA \( j \) with characteristic \( i \). To calculate the similarity of FSA \( j \) to the user’s selections, we simply use the 2-norm, resulting in similarity score \( s_j = \sqrt{\sum_{i=1}^{n} (c_i p_{ij})^2} \). After doing this for each FSA, we have a vector \( s \in \mathbb{R}^m \) of similarities. These similarities are then normalized to achieve the similarity vector \( s' \), where \( s'_j = 100 \frac{s_j - \min(s)}{\max(s) - \min(s)} \). The benefit of normalizing \( s \) to be between 0 and 100 is that a score from 0 to 100 is interpretable for the user. We present \( s' \) to the user both in alphabetical order by FSA and sorted by score (“Output” portion of Figure 4.1).
Recall from the literature review in Section 2.1 that there has been a substantial amount of research on how demographic features influence participation in sport. An NSF can use the results from one of these studies or a similar study to determine the FSAs where people are the most (or least) likely to participate. If a user knows demographic characteristics of their members, but does not know where they should focus their efforts to gain new members, they can use this tool to see where there are high proportions of these people.

### 4.1.2 Location Search Tool

In contrast to the CST, the Location Search Tool (LST) finds FSAs with populations that have similar demographic characteristics to the people in a particular FSA. This is useful when the NSF does not know what the typical demographic characteristics are of someone who plays its sport, but it knows where many of their members live. The high level differences between the CST and the LST are summarized in Table 4.2.

In the LST (Figure 4.2), the user must select a target FSA in which he/she is interested. As in the similarity calculation for the CST, for FSA $j$, let $p_{ij} \in [0, 1]$ be the proportion of people in FSA $j$ with characteristic $i$. The similarity of FSA $j$ to the user’s selected FSA, $k$, is simply $t_{jk} = \sqrt{\sum_{i=1}^{n} (p_{ij} - p_{ik})^2}$.
Chapter 4. Tools for National Sport Federations

<table>
<thead>
<tr>
<th>CST</th>
<th>LST</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Objective</strong></td>
<td>Find areas that have people similar to a target person</td>
</tr>
<tr>
<td><strong>Input</strong></td>
<td>Target demographic characteristics</td>
</tr>
<tr>
<td><strong>Output</strong></td>
<td>Prioritized list of similar FSAs</td>
</tr>
</tbody>
</table>

Table 4.2: Comparison of the CST and the LST

The vector $t$ is normalized to become $t'$ in the same way that $s$ was normalized to become $s'$ in Subsection 4.1.1. Likewise, $t'$ presented in the same way that $s'$ was presented in the CST.

Note that in the calculation of $t$ above, there is no way for the user to select which characteristics to consider. We add the option for the user to select demographic characteristics. However, for simplicity, the users must select an entire variable (e.g., gender) rather than choose a specific characteristic (e.g., male) within a variable as they could in the CST. The variables the users can choose include those used in the CST, listed in Table 4.4, and more, which are listed in Table 4.3. In total there are 27 variables that the user can consider that have a total of 453 characteristics.

Suppose the user selects demographic variables $c \in \{0, 1\}^{27}$. Let $l_i$ be the number of characteristics...
Additional Variables | # Characteristics
---|---
Family size | 32
Household characteristics | 45
Languages spoken at home | 111
Citizenship | 2
Place of birth | 65
Generation status | 3
Visible minority | 14
Aboriginal identity | 20
Mobility status | 8
Field of study | 13
Location of study | 6
Labour force status | 3
Class of worker | 3
Work activity | 11
Place of work | 4
Transport to work | 9

Table 4.3: Additional variables included in the LST that are not included in the CST.

corresponding with variable \( i \). Then we create a new vector \( c' \in \{0, 1\}^r \), \( r = \sum_{i=1}^{27} l_i = 453 \), such that:

\[
c' = \begin{bmatrix}
c_1 1^{l_1} \\
\vdots \\
c_{27} 1^{l_{27}}
\end{bmatrix}
\]  \hspace{1cm} (4.1)

where \( 1^x \) is a vector of all ones of length \( x \). Then we can define the similarity between FSAs \( j \) and \( k \) as

\[
t_{jk} = \sqrt{\sum_{i=1}^{n} c'_i (p_{ij} - p_{ik})^2}.
\]

4.1.3 Dashboard

Included in both tools is a dashboard (Figure 4.3) with which the user can compare the demographic characteristics of a given FSA with the average characteristics of the country. This is useful when the user is interested in an FSA that was highlighted when using the tool and would like to know more about the population of that FSA.

4.2 CFSA Tool

Following the workshops we received a request to make a similar, but more specific tool for the Canadian Freestyle Ski Association (CFSA). The objective of this tool was for the CFSA to be able to understand which FSAs have high proportions of members with target membership characteristics. The CFSA
Figure 4.3: Screenshot of the dashboard included with the CST and the LST. The user inputs an FSA of interest at the top and the dashboard shows the similarity score that the FSA was given in the first plot. It also displays a plot for each variable in 4.1 (not all shown above) with demographic comparisons between the FSA and the rest of the country.

asked us to include their membership data from approximately 3,800 members directly in the tool, with information from the fields in Table 4.4 for each member. The general membership type can be one of: athlete, coach, associate, judge or major official. In addition, each member must have a license to participate in a CFSA sanctioned club or tournament. The license type can be one of Can Free 1& 2 (entry-level athlete), Can Free 3 (athlete competing in provincial or national competitions), Can Free 4 - FIS Canada (athlete competing in International Ski Federation competitions in Canada), Can Free 4 - FIS International (athlete competing in International Ski Federation competitions internationally), Try Freestyle (temporary athlete license) or Unknown.

The tool works in a similar way to the Characteristic Search Tool. Users can select any of the characteristics from Table 4.4 to be target characteristics. The tool finds FSAs with high proportions of members who have the selected characteristics, calculating the similarity metric in the same way as the CST.

This tool is also equipped with a dashboard (Figure 4.4), similar to the one included in Tools A & B. The CFSA inputs an FSA of interest and they can see how their members in that FSA compare to
<table>
<thead>
<tr>
<th>Gender</th>
<th>Province</th>
<th>Postal code</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Provincial / Territorial Freestyle Ski Organization (P/TSO)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>General membership type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Athlete license type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Table 4.4: Features of CFSA members.

their general membership. In the case of age, the user can also see how their members compare to the general Canadian population.

4.2.1 Feedback

Fourteen months after we built the CFSA Tool, we followed up with our contact at the CFSA to see how it had been used. The following information is from a personal communication with J. Anderson on July 13, 2017. The CFSA had hoped to use this tool to inform their marketing team about locations to target long-term high performance members. Because their focus is on high performance rather than expanding membership, they found it more useful to compare their niche membership to memberships of other snow sports rather than the general Canadian population, as we had built the tool. They were able to easily adapt the tool by editing the back-end to replace the census data with various relevant data sources,
such as membership data from the Canadian Snowsports Association [1] which represents the eight snow
sport NSFs internationally. The information that the CFSA gained from this adapted tool shows how
the membership of the CFSA compares to membership of the other snow sports. This information was
used to build a report which was shared at the Freestyle Summit, a conference that engaged various
stakeholders from CFSA, the Provincial Sport Organizations (PSOs), and clubs, helping develop CFSA’s
strategic plan for the next Olympic cycle. More detailed findings, specific to each province, were shared
with the PSOs, some of which used the information for provincial grant applications. Membership
data from the previous seasons has also helped the CFSA to establish a baseline upon which future
membership data can be compared to.

Perhaps more important than their use of the tool, the CFSA has increased their use of analytics
substantially. While before the workshop they admitted to having very little analytics experience aside
from creating tables in Excel, they are now manipulating the tool we built for them and using Tableau
to visualize the results of their own analyses. Our contact told us that since our workshop, they have
been using Tableau a lot to analyze their membership data and inform their sport development or high
performance development decisions.

A natural next step of this research would be to build a tool to learn about the locations and features
that are predictive of retention and attrition of members. This is an important topic of research; the
CFSA has started analyzing longitudinal membership data from the past three years to determine annual
retention and attrition rates. They were able to determine who was consistently registered, who dropped
out, the average age of people dropping out, the freestyle program they were in when they dropped out,
and how many years they were in their program and freestyle skiing. They then compared this to general
data they acquired about participation in sports such as the general attrition rate for certain age groups
in Canada. They were able to use these findings to develop strategies to increase program retention
rates for athletes 6-12yrs and increase participation for girls in particular. For example, their findings
show that our attrition rate were 6% higher than the national sport average. These descriptive statistics
have been extremely useful for the CFSA; they have been able to use them to create a rating that helps
identify centres for excellence. Understanding the factors that lead to retention and attrition would help
the CFSA better target efforts to retain current members.

By adapting the CST to fit the needs of the CFSA, we were able to help them expand their analytics
capabilities. On this topic, in a testimonial, James Anderson said: “introducing me to Tableau was game-
changer for our organization. It has been a great tool. The workshop helped with the conceptualization of organizing our membership data, and possible implications for a NSO." Further, through demonstration of our tools in the analytics workshops, we showed the CFSA the ease and benefit of using analytics to help guide business decisions. The CFSA has now embraced using analytics to guide decisions, all of which stemmed from the original tools we showed them at the workshop [J. Anderson, pers. comm.].
Chapter 5

Conclusion

In this thesis, we used predictive, prescriptive, and descriptive, analytics to help Canadian amateur sport organizations make decisions with the objective of maximizing participation in their sport. Through novel prediction methods, we accurately predicted the distribution of demand for tennis on a geographical level. By validating these models with census data, we showed that our methods are more accurate than naively assuming that the same proportion of tennis players live in each region of Canada. In order to use these predictions to help Tennis Canada make decisions, we illustrated the predicted demand and the supply of tennis courts on an interactive map. In particular, this map is useful for a sport manager to determine which locations may need new facilities or programming, and which locations may be oversupplied.

We then showed, in a case study, how the predicted demand can be used as input for an optimization problem to solve a real business problem. The case study locates new winter tennis programs called “Serve and Turf” programs based on where there is high demand for tennis. In this analysis, we demonstrate that if an amateur sport organization has specific objectives regarding locating a new program, prescriptive analytics can be exceptionally useful to find the optimal locations. In addition, it validates that locating facilities based on the predicted demand produces different results from locating facilities based on population. This implies that predicting demand on a geographical level can be used to inform decision-makers of specific decisions they may not be able to make without analytics.

In the final chapter, we reviewed two tools that we built for the National Sport Federation Enhancement Initiative (NSFEI) analytics sessions, called the Characteristic Search Tool (CST) and the Location Search Tool (LST). These tools helped us effectively explain to many National Sport Federations (NSFs)
the value of using analytics in amateur sport to make business decisions. In addition, we adapted the CST such that it could input the membership data from the Canadian Freestyle Ski Association (CFSA). Both the NSFEI session and this adapted tool helped the CFSA increase their use of analytics and create more targeted strategies concerning retaining membership.

There are many future directions for using analytics in amateur sport organizations. To expand upon our geographical prediction of demand, more variables could be considered other than demographical and socio-economic variable. Health and psychological variables have been shown to impact participation in sport, and so if we could infer some of these variables from the survey data, they would be useful to consider in our models.

Additionally, it would be advantageous to approach the problems from Chapters 2 and 3 with the objective of increasing retention in sport rather than increasing participation in sport. Most of our research was focused on increasing participation in sport, but many NSFs are concerned with retention of members in addition to increasing membership. Regarding future work, it would be interesting to examine how predictive and prescriptive models can be used to increase retention rates in sport. Using longitudinal membership data, rather than survey data, we could use similar modelling techniques to predict whether a current member will still play a sport in the future. Based on this prediction, we could solve similar facility location problem but rather than locating facilities to maximize demand covered, it could locate facilities to maximize the likelihood of retention of covered athletes.

The tools we built in Chapter 4 can adapt in many ways to help NSFs with their current membership objectives. Along the idea of incorporating retention into the demand prediction and facility location, many NSFs may find use in tools that have metrics of retention incorporated. After analyzing the the demographic factors that lead to attrition, these features can also be incorporated into the tools, helping NSFs better address their attrition concerns. Future work will help amateur sport organizations make more directed membership and participation related decisions driven by analytics.
Bibliography


[41] University of Toronto Faculty of Kinesiology and Physical Education. Rental information. [https://kpe.utoronto.ca/facilities-memberships/rental-information](https://kpe.utoronto.ca/facilities-memberships/rental-information), 2017.


Appendix A

Supplemental Figures

Figure A.1: Illustration of the grouping of occupation categories from the National Household Survey and the tennis survey into 3 buckets.
Figure A.2: Distribution of median age in FSAs in the census data. Vertical red line is median age of survey respondents.

Figure A.3: Distribution of median age in FSAs in the census data. Vertical red line is median age of survey respondents.
Figure A.4: Comparison between census data and survey data by gender.

Figure A.5: Comparison between census data and survey data by ethnicity.
Appendix A. Supplemental Figures

Census and Survey Data by Place of Birth

Figure A.6: Comparison between census data and survey data by place of birth.

Census and Survey Data by Language

Figure A.7: Comparison between census data and survey data by language spoken most often at home.
Figure A.8: Comparison between census data and survey data by education.

Figure A.9: Comparison between census data and survey data by occupation.
Figure A.10: Box plot of mean squared error for each of 500 repetitions from Algorithm 5 with outliers included.