Universidad Autónoma de Madrid
Escuela Politécnica Superior

Trabajo de Fin de Grado

COMBINING CLUSTERING AND TIME SERIES FOR BASEBALL FORECASTING

Grado en Ingeniería Informática

Miguel Vázquez Fernández de Lezeta
Mayo 2014

COMBINING CLUSTERING AND TIME SERIES FOR BASEBALL FORECASTING

AUTOR: Miguel Vázquez Fernández de Lezeta
TUTOR: Héctor Menéndez Benito
PONENTE: David Camacho Fernández

Dpto. de Ingeniera Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Mayo 2014

Resumen

Resumen

El mundo de las apuestas deportivas ha crecido como uno de los mayores negocios en los últimos años. Las compañías de apuestas generan enormes cantidades de dinero con la ayuda de las nuevas tecnologías, que acercan los deportes profesionales a gente de todo el mundo, interesados tanto en esos deportes como en las apuestas. Los sistemas de predicción son una de las principales herramientas requeridas por dichas compañías de apuestas para establecer las cuotas. El análisis estadístico de los eventos deportivos es necesario para conseguir esta meta. Uno de los deportes más influenciado es el Baseball. El análisis deportivo a través de estadísticas ha revolucionado el Baseball desde que los Oakland Athletics comenzaron a fichar jugadores por eficiencia estadística a bajo coste, proceso denominado como "*Moneyball*". Esta estrategia supuso un gran éxito, así que muchos equipos se adaptaron a ella durante las siguientes temporadas. Ya que las estadísticas de Baseball se hicieron tan importantes, algunas compañías como Retrosheet han creado datasets recopilando información de partidos de Baseball, jugadores y equipos durante muchas temporadas. Este trabajo intenta generar un modelo de predicción de resultados de Baseball usando información estadística previa. El modelo combina series temporales y algoritmos de clustering para general un modelo que aprende de la evolución de equipos y partidos e intenta predecir resultados finales.

Palabras Clave

Data Mining, Series temporales, Clustering, Baseball, Predicción, Estadísticas, Métricas

Abstract

Sports betting has grown as one of the greatest bussiness over the past years. Gambling companies generate enourmous amounts of money due to the help of new technologies at meeting professional sports to people all around the globe, interested in those sports as well as betting. Prediction systems are one of the main tools required by those gambling companies to state the odds. Statistical analysis of the sport events is necesary in order to achieve this goal. One of the most influenced sports is Baseball. Sports analysis through statistics has revolutionized Baseball since the Oakland Athletics started singing players by statistical efficiency at low costs, this process was named as "*Moneyball*". This strategy provided a great success, so many teams adapted to it over the following seasons. As Baseball statistics became so important, some companies such as Retrosheet have created datasets collecting information from Baseball games, players and teams through many seasons. This work intends to generate a forecasting model which predicts Baseball results using previous statistical information. The methodology combines time-series and clustering algorithms to geneate a model which learns about the teams and matches evolution and tries to predict final results.

Key words

Data Mining, Time-series, Clustering, Baseball, Prediction, Statistics, Metrics

Agradecimientos

Quisiera dedicar este trabajo a todos aquellos que me han apoyado, aunque sea con el más mínimo gesto de ánimo. Quiero dedicárselo fuertemente a mi familia, especialmente a mis padres, Miguel y Milagros, que han estado ahí todo este tiempo, fueran las cosas bien o mal. También quiero hacer mención de la gente del departamento que han hecho que disfrute de este trabajo con un ambiente muy agradable y gran ayuda, en especial a Héctor por todo su esfuerzo y dedicación. Por último, quiero dedicarlo a aquellos que han estado ahí todos estos años: Alfredo, por meterme de carambola en esta carrera, Alejandro por transmitirme la vocación por las ingenierías, Carlos por compartir su sufrimiento en este fin de recta, Isaac por amenizar las largas noches de trabajo con su humor y Guillermo y Carolina por ser capaces de arrancarme una sonrisa cuando la cabeza necesita despejarse. A todos: Gracias.

Contents

List of Figures

List of Tables

1 Introduction

Forecasting has been one of the most challenging areas of knowledge for a long time. From weather forecast to financial predictions, knowing what is happening in the future has always provided advantage for those who had the knowledge of what is about to happen. However predicting the future is not an easy task as there are many factors to consider at any kind of predictions. In the end, the best prediction is the most accurate one according to the available information. Gambling is one of the most influenced areas. Gambling bussiness require of solid predictions in order to entablish consistent odds that may be attractive to the costumers as well as secure benefits. Besides, knowing with a high chance the result of an sport event previous to its celebration may result in a high benefit at betting.

Prediction systems are some of the most complex models within the data analysis. Many of these systems have been implemented through differente Data Mining techniques such as Time-Series [1] or Clustering [2], reaching notorious results in fields like investment funds or even sports results. In relation to this last area of interest, several betting systems have been updated in order to being able to apply this kind of techniques, thus they can provide a higher rentability to their clients.

There are also several studies that have used Data Mining at sports analysis before [3]. It is important to mention how some of those studies have affected the sports they focus at, turning into players improvement by addapting training habits or increasing teams performance by more sofisticated strategies. However, they focus mostly at teams or players analysis and understanding in order to find patterns or strategies, which means that using Data Mining technics for prediction purposes is still a field to be worked at.

These Data Mining and automatic learning techniques may be applied to sports analysis for several reasons. Maybe the two most interesting ones are team management and results predictions. In the first case, we can find an example in 2002 when the Oaklands Athletics highly improved their team performance by singing up players by statistical performance, this process was named *Moneyball*. Related to results prediction, these techiniques are used by the gambling bussiness to set their odds as well as betting experts to get advantage at their investments.

Data Mining processes must lead with some potential issues. On the one hand, working with big amounts of data may generate some issues in terms of resources and time. As the amount of information to be analyzed increases, better hardware is required as well as better algorithms

and software techniques must be used. Otherwise, the Data Mining systems take too long to achieve their goals, making their work impracticable at some points. On the other hand, the lack of usefull information may lead to a situation where the analysis can not be made.

This project intends to develop a prediction system, oriented to sport results forecast, based on some classic Data Mining models and generating a methodology that allows, through different tools, to compare the results from these models facilitating the information management. The project focuses specially in forecasting Baseball results by combining both, time-series and clustering, as well as adding graph theory to generate a prediction model.

The rest of the project is structured as follows: Section 2 explains the related work as well as some of relevant Data Mining techinques and tools. Section 3 describes the prediction model proposed by this project and Section 4 shows all the experimentation related to the model described in Section 3. At last, Section 5 adds conclusions and proposes some ideas for future work.

1.1 Motivation

Sports have a huge impact in society and generate large amounts of money through all the globe. Some of the most important sport events are followed by millions of people all around the world, such as the Super Bowl in the NFL in USA or the final game of the UEFA Champions League in Europe. Gambling is one of the bussiness areas that are highly benefit from the sports relevance all around the world. Gambling companies have a lot of costumers who can bet their money on different sports, which generates millions of euros per year for some of the greatest companies such as *Bwin.Party Digital Entertainment*.

The automatization process of sports analysis is an interesting task, as well as statistical research and predictions. This last feature is really important at sports betting and also one of the most difficult tasks. This problem motivates this work, which tries to deal with it combining unsupervised learning techniques and time-series.

1.2 Objetives

The main goal of this project is to generate a forecasting system which can deal with baseball data. In order to achieve this goal, this project has been divided in the following objectives:

- **Studying Data Mining techniques** and how different supervised and unsupervided models can be applied to extract patterns from data.
- **Studying Time-series techniques** and how they are used for forecasting problems.
- **Extract data from baseball** and prepare the data to be analyzed using Data Mining preprocessing methods.
- **Generate a forecasting model** which combines unsupervised Data Mining tecniques and time series.
- **Apply the model** to the baseball dataset in order to generate the prediction system.
- **Test the model** and evaluate its results.

2
Related Work

This chapter presents a general overview around the sport industry, its history and the current approaches around this market. It also introduces some Data Mining methodologies applied to Sports and some tools which have been studied, but not necessarily used.

2.1 Sports and Society

Sports are currently one of the most profitable business around the world. The Olympics [4] or the FIFA World Cup [5] are good examples of how this industry has become really important over the last few decades.
There are several fields inside the sport industry. Some of the most important are:

1. **Marketing**: The influence of sports in market is really remarkable not only currently but during all its history. The different brands have found a complete market to publicise their products using sport players or teams [6].

2. **Medicine**: Several physical advances have been provided through the high inversions of medical or health resources in sport in order to keep the player health [7].

3. **Social**: Sports also have influence in countries motivation and social psicology, specially when an important event, such as the Olympics, is organized [4].

4. **Economical**: Different sports can affect to the economy of a whole country or the economy of a small regions which design products related to an specific sport [8].

5. **Gambling**: National Lotteries and Gambling business also find profits from sport competitions having influence on the decisions and events that the leagues takes [9].

Besides, sport Forecast is one of the most challenging problems. The problem consist on predict match results based on a dataset extracted from different teams, players and matches of a concrete sport. One of the most difficult steps for this process is to find the most appropriate dataset. Usually, some sport datasets contains general season information while other contains more detailed information about the play-by-play, game logs, players aligment, etc. Baseball

is one of the sports which contains more detailed data about the different teams and players. There exists a dataset called Retrosheet [1] which contains lots of information about player, teams and matches. They accumulate the information using game logs of the different events during a match.

There are several studies which are focused on the importance of sports in society. According to the different aspects mentioned above, some works have been based on the effect of the most relevant events. For example, Lee and Taylor [5] analyse the influence of megaevents specially focused on the FIFA World Cup of 2002. They focused their research on how these events create sport tourism, improving the economy of the hosting country (in this case South Korea). They calculate that the World Cup generated an economical impact of 1.35$ billions of sales, 307$ millions of incomes and 713$ million added for the country. These results show how this kind of tourism generate more profit than foreign leisure tourists (1.8 times more). From a different perspective, Waitt [4] analyses the social impact of Sydney Olimpics. Author studies how this event has affected from its organization in 1998 to its celebration in 2000. He shows that the social euphoria intensifies during this period. He also studies the implication of this event in different social groups.

Besides that, other works are focused on the economical impact for the industry. For example, Pinch and Henry [8] studies how small firms of motor industry can affect to national economy in United Kingdom. From a more general perspective, Pitts et al. [10] apply industry segmentation theory to sport industry in order to create a model for this kind of industry. They use information about sport manufacture to discriminate three main categories: sport performance, sport production and sport promotion. They also focus their study on how categorize the different clients into these groups. In this context, it is also important to study how the different business affect to sports. A good example, studied by Forrest and Simmons [9], is the relationship between sports and gambling. They focus their reseach on how gambling can affect to sport organization and operation. Also, they study the danger of corruption from wagering and how these methods have affected to different sports in history (such as cricket) and the influence of different National Lotteries on them.

Finally, as every business, sport requires quality measure systems in order to improve their services. In this context, Ko and Pastore [11] propose a service quality model for sport. They focus their research on four main features: program quality, interaction quality with the clients, outcome or social repercusion and environment quality.

2.2 Teams and Players in Sports

This section explores the influence of teams and players from different perspectives. It is focused on different analysis carried out to teams, and latter to players, in order to remark the importance of these factors.

2.2.1 The importance of Team Analysis

Teams can be analysed from different perspectives. In sport business, it is important to consider how teams affect to the brands. For example, Gladden and Funk [12] focused their work on understanding brand management in sports creating a model to identifies the different dimensions of brand associations. They identify three main categories related to team: attributes, benefits and attitude. Their models are tested on sport costumers. From a similar perspective, Bauer et al. [13] study the importance of brand image for fan loyalty in team sports. They show

[1] `http://retrosheet.org/`

how there are relationships between these factors. According to the three categories described above, brand attitude is the most influencial for fan loyalty, which is confirmed using structural equation modeling.

In order to improve the team, there are different studies about the relevant aspects in sport teams. One of the most important is the team cohesion. Carron et al. [14] try to study the cohesion-performance relationship in sports. They also examine the influence of moderators in this process. They discover that the highest cohesion is performed in female teams. They also remak the importance of cohesion during team building, and the team targets. Also, the team motivation is relevant in order to improve the cohesion. From this perspective, Roberts and Ommundsen [15] studies how the motivation influence in the teams performance. They divide the study in two main goal: ego and task. They study how inner competition affect in the team and how different factors can influence in the cohesion of the team according these two goals. They also try to determinane the satisfaction factors of the team.

Another important factor is the coach. It is important to understand how the coach influence in the team. For example, Gilbert and Trudel [16] studies the coach role in youth team sports. They try to understand the main influence of the coach in the partitioners, taking into account enviromental conditions and personal views. They also study how to examine these features. From a similar perspective, it is also important to understant the consecuence of replacing a coach. A good example is [17], where Mc. Teer et al. analyse the effect on performance of the mid-season replacement of coaches. The study was oriented to four sports: basketball, baseball, hoskey and football. They study the effect on leadership in two different ways: how a new leadership is introduce and how its news ideas affect to the team and also how the performance is influenced by the changes. They conclude that there are minimal effects for these kinds of changes.

Finally, it is important to understand how to train the team. Baker et al. [18] study the influence of different expert opinions in three main sports: hockey, netball, and basketball. They took information from experts and non-experts athletes about the practice activities they have had during their carrers. They conclude that experts usually spend 4000 hours of sport-specific training before reaching international standard. They also conclude that there were a negative correlation between the number of additional sports and the hours of specific training.

2.2.2 The importance of Player Analysis

Teams are formed by players. It is also important to study not just the global perspective of teams, but also the individual perspective of a player.

One of the most relevant fields for players study is health. Sports demands some physical aptitudes that usually produce injuries in players. A good example of the influence of these demands can be found in [7], where Abdelkrim studies the physical demands produced by the changes introduced in basketball rules in May 2000, which reduce attack time 6 seconds and create 4 quarters instead of the 2 halves. They analyse heart rate effect and blood for four different positions. Generally, players spend more time in high specific movements. Centers have higher intensity than guards and forwards. The changes slightly increased the cardiact effects involved during competition, and the intensity differ according to the player positition. It is also important to study how players recover from their injuries. For example, Verral et al. [19] study the decrease in playing performance of athletes returning to sport after recovery from hamstring muscle strain injury. Atheles has a significanly lower performance immediatly upon return. The study concludes than some athletes may return to sport prior to complete resolution of their injuries. Also, It is needed to prevent these problems. A good example in this field is found in [20], where Surve et al. evaluate the effect of semi rigid ankle orthosis on

the incidence of ankle sprains in soccer during one season. The study concludes that a semirigid orthosis reduced the incidence of recurrent ankle sprains in soccer players with previous history of ankle sprains.

From a social perspective it is also important to study the influence of players in team brands. A good example can be found in [21], where Wilson et al. studies how player transgressions in sport bring negative repercusions for stakeholders as a result of their association with the team athlete or sport. They focused their study on the effects that these incidents can have on relationships with sponsors from a sport organization perspective. They also discuss several factors around the impact of transgressions.

2.3 Data Mining in Sports

Data mining techniques are based on knowledge extraction or pattern identification inside an information source. They usually are focused on classification or clustering techniques. Classification techniques use the information of a class attribute, while clustering techniques identify patterns that can group similar data points into categories.

Classification [22] techniques have been designed to consider the class information of the data. These classes form part of the data instances and are considered by the algorithm in order to generate a general predictive model that describe the data. Different classification models have been designed in Machine Learning. The most common models are based on [22]: decision trees (C4.5), classification rules (RIPPER), artificial neural networks, random decision forest, support vector machines, naïve Bayes and k-nearest neighbours, amongst others.

Clustering [22] is a blind methodology to group similar data minimizing a cost function. These techniques are usually divided in three types of clustering [23]: partitional (each instances belongs to a single cluster), overlapping (each instance belongs to one or more clusters) and hierarchical (partitional solutions are nested to generate a tree of clusters). Well-known clustering techniques are K-means [24] and EM [25]. Both K-means and EM are partitional clustering algorithms.

This work is focused on hierarchical clustering. Hierarchical clustering is usually divided in two main categories [26]:

- **Agglomerative:** this is based on a bottom-up approach, pairs of clusters are joined in each move up within the hierarchy.
- **Divisive:** this is based on a top-down approach, the clusters are splited in each move down within the hierarchy.

Hierarchical clustering has been highly used in time-series clustering [1]. This kind of analysis is based on time-series grouping, where the algorithm is designed to find similarities between different time-series in order to use these series to forecast the behaviour of others.

Several techniques such as statistics, data mining and machine learning have been used to analyse the performance of teams and players in games like soccer [27, 28], football [29], basket [30, 31, 32], etc. These approaches, usually named human or robot behaviour modelling, has been applied in different domains like Robosoccer simulations, but, in these examples, all the information is totally controlled and simulated. Moreover, Raines et al. [33] create a multi-agent framework to analyse team behaviour. They generate an automatic agent for offline team analysis. This tool have multiple types of models for team behaviour in order to analyze different events such as actions, interactions and performance. All these methodologies implies

machine learning models focus on the generation of a human interpretation. Their test domain is the Robosoccer.

Other similar analysis applied to human team games can be found in the NBA league. Vaz de Melo et al. [32] analyse the evolution of this league during its whole history creating a complex network model and studying its evolution. In the same context, Bhandari et al. [30] describes a data mining application, named Advanced Acout, used by the National Basketball Association. They focus their explanation on how this application deal with the different steps of Data Mining: data extraction, data pre-processing, pattern discovery and model application. They also generate a visualization model based on patterns and video tapes. From a dirrent perspective, Ivankovic et al. [34] apply Data Mining techniques to basketball data. The main goal of their work is to discriminate the most relevant features of the data to predict match results.

To the football or soccer analysis problem, Onody and Castro [28] propose a model also based on complex networks but only applied to analyse Brazilian players, and Bitter et al. [31] generates a statistical model, modifying classical probability distribution such as Bernouilli and Gaussian distribution to create a score model for different leagues. Besides, Dawson et al. [29] studies the Australian Football League movements and activites using video analysis. They extract information about movements patterns and game activities studing statistics of different positions. They also propose improvements in specific training practices for different positions using the information extracted.

Finally, Shumaker et al. [3] presents a roadmap about different aspects of Sport Data Mining. They focus their work on different methodologies providing several data sources for sports, research details of different sports, tools for the analysis, prediction models, methods to analyse multimedia content, methodologies to extract data from web pages and some case studies using classifiers.

2.3.1 Baseball Analysis

From the baseball point of view, there are also several works that are focused on baseball analysis. In [35] they propose a visualization framework for baseball to extract information about different teams and matches in order to query different aspects of baseball. Hakes and Sauer [36] study the Moneyballl effect from an economic perspective. Their goal was to prove that there was an inefficiency players evaluation for baseball market over a prolonged period of time. Exploitation of this inefficiency by the Oakland Athletics team suppose an outstanding progress for the baseball strategies. Other research is focused on different analytical perspectives, for example, Marchi and Albert [37] introduces several techniques to analyse the different parts of a baseball macth, team, player, etc, using R. They provide several analytical methods extracted from mathematics. They also introduce some machine learning methodology but only focused on classification and regression.

2.4 Data Mining and Baseball Analysis Tools

There are several Data Mining tools which are helpful for baseball analysis. Here, we enumerate some of them:

- **R**[2]: R is a language and a toolset for statistical computing and graphics. It is under GPL license and possesses a huge community providing different packages for statistical

[2] `http://www.r-project.org/`

analysis, data and big data handling, pattern recognition and visualization. It is based on a previous language called S and much code written for S runs unaltered under R.

- **Matlab**[3]: MATLAB is a high-level language and also an IDE for numerical computation, visualization, and programming. It allows to analyze data, develop algorithms, and create models and applications. It has several optimizations focused on computation. It can be used for a range of applications, including signal processing and communications, image and video processing, control systems, test and measurement, computational finance, and computational biology.

- **Octave**[4]: Octave is a high-level interpreted language similar to Matlab, primarily intended for numerical computations. Is is under GPL license. It has been focused on numerical solution of linear and nonlinear problems, and numerical experiments. It also provides extensive graphics capabilities for data visualization and manipulation. The Octave language is quite similar to Matlab so that most programs are easily portable.

- **Weka**[5]: Weka is a collection of machine learning algorithms for data mining tasks. It has tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It also allows analyzers to develop their own machine learning schemes.

- **Improvise**[6]: Improvise is a Java software architecture and user interface that enables users to build visualizations interactively. It has been designed to synchronize different visualizations and queries in order to combine information from different sources. Users can navigate and select the appearance of data across multiple views, using a number of variations on well-known coordination patterns such as synchronized scrolling, brushing, drill-down, and semantic zoom.

- **GameDay**[7]: GameDay is a software program that allows sports fans to track games with live stats of baseball. For Major League Baseball, it was introduced in 2002, a year after all team sites were migrated to MLB.com. The software provides several features which can help in game analysis.

- **Retrosheet**[8]: Retrosheet is one of the biggest baseball datasets. It also provides different tools to analyse the data it contains. Retrosheet has a huge collection of games. They try to make play by play information publicly available for all interested researchers.

[3] http://www.mathworks.co.uk/products/matlab/
[4] http://www.gnu.org/software/octave/
[5] http://www.cs.waikato.ac.nz/ml/weka/
[6] http://www.cs.ou.edu/~weaver/improvise/index.html
[7] http://mlb.mlb.com/stats
[8] http://retrosheet.org/

3

Prediction Methodology

This project focuses on the development of a new prediction methodology for baseball games in order to predict the result of the matches. To achieve this goal, clustering, graph theory and time-series techniques are combined. Predictions will be done based on collected data about past games so the model tries to find similarities between teams and games to find possible trends and forecast the outcome of the game according to these trends. In this chapter a detailed explanation of the model is given as well as all the processes the project has come through.

3.1 Data Description

Retrosheet[1], as mentioned in Section 2, is an organization created to recollect data about Major League Baseball games whose database is continuosly updated and contains a large amount of data. Games from 1871 to these days may be found recorded within this database and also detailed information about these games like play-by-play data and other interesting features related to the greatest American baseball leagues. The main reason why this database has been selected for this project is the time-based data it contains, we can find games whose statistics can be orderer chronologically as well as play-by-play data within every game data. Therefore, we can create time series with the collected data and apply clustering to those as we show in the following sections.

3.1.1 Retrosheet Data

Data are downloadable from Retrosheet website, where every file name is formed by the Season Year and a sufix determining the kind of data it contains (*eve* for Regular Season games, *seve* for Regular Season games played during a 10 years period or *post* for Post Season games), e.g. *2003eve.zip* contains data about Regular Season games during 2003 while *2000seve.zip* covers data about all Regular Season games played from 2000 to 2009. Here, we give a brief summary of the information they contain, while further information can be found in the Appendix A or the Retrosheet Webpage[2]. Compressed within the *.zip* files three kinds of files can be found:

[1] "The information used here was obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at 20 Sunset Rd., Newark, DE 19711."

[2] http://www.retrosheet.org/

- **Event Files**: *.eva*, *.evn* and *.eve* contain data about games played in American League, National League or play-offs respectively. For *.eva* and *.evn* files the filename consists of the season year followed by a team code and contains information about all games played as home team by the team corresponding to that code, e.g., *2003OAK.eva*. For *.eve* files the team code in the filename is substituted by an abreviation of the round name of the play-offs the games are played at. The game information provided by these files is about both the game characteristics, such as teams, time, weather conditions or starting rosters, and play-by-play information such as batting, running or player substitutions.
- **Teams Files**: These files have no extension but are named by the season year concatenated to the string *TEAM* and have information about every team playing during the corresponding season or post-season, e.g., *2003TEAM*.
- **Roster Files**: Each *.ros* file provides the information about a team roster during a year and some detailed information about the players that are part of that roster. These files name is formed by the corresponding team code and the year, e.g., *OAK2003.ros*.

3.1.2 Data Transformation

Although Retrosheet provides tons of interesting data, data must be transformed to a format we can work with, as well as we must select only data that may be meaningful for us, which is data we can transform into Time Series. However, we were dealing with large amounts of data, so even though the event files contain their data as record entries with different fields per record, parsing and analyzing every file record by record is not an easy task and extracting meaningfull data, as well as translating the information, may need a significant amount of resources and time, so automation of this process was needed in order to speed it up and save computation and unnecesary efforts. Some software tools were needed to complete the translation and cleaning of the extracted data, in adition to some human interaction to specific kinds of data such as missing values or outlayers unmanageable by those tools. In this section we will analyze the main tools we have used for this automation process.

Retrosheet Software Tools

Retrosheet offers some software tools to transform data collected in their database into a more usable format. In this case we have selected *BGAME.EXE* and *BEVENT.EXE*, which are programs that can be executed in command prompt to transform the data within the downloable files from Retrosheet to *CSV* files that could be later loaded into a local database. The main advantage of having these tools is the data standarization they provide, in the sense of all the data obtained as output is in the same format and in an easier way to read by programs, avoiding future needs of parsing the whole event documents in order to access to specific data. The usage of these programs as well as the output format is explained in the Appendix A.

MySQL Local Database

Althought we will not explain the technical details about MySQL technology it is important to understand the ideas behind the choice of using a local database for this project and why MySQL was apropiated for this task. As mentioned before, it is important to collect meaningful data and discard whatever is not needed and so is having fast access to the data. Using a local database provides help in this way through creation of tables where the data may be loaded in and the utilities provided for it, such as management or different functions to work with the data. Having *CSV* outputs from the Retrosheet tools, data only have to be loaded and tables

created following the fields within the *CSV*. The main reasons why MySQL was selected for this purpose is the access provided by many tools and programming languages through libraries and plug-ins to this database format due to its extended usage.

3.2 Architecture Topology

The model for this project is focused on forecasting through statistics over time comparison. This means that teams which act in similar ways will perform similarly in concrete situatons with a high probability. In the same way, games should develop in a similar manner if the playing teams are similar.

Once the data have been collected, the model has to be generated. Having data from the games in play-by-play form allows to generate player statistics for each game as a whole or statistics per play. The reason why player statistics were chosen is to generate team statistics, considering a team as the sum of all players that play for that team in a concrete game. This reasoning allows us to generate team statistics that are the base to our time-series generation. The other part of the model is also developed in this way as games may be measured by the playing teams performance.

3.2.1 Teams Graph

Due to similarity measure for teams is needed we decided to create a graphs topology connecting teams according to their similarity. Using graphs also provides information about the search space, which helps during the analysis process. The first step is to create the time-series that measure the performance of teams over time. In team analysis, games were used as the chronologycal variable for the time series, so the team-related statistics are calculated by game for each team, which provided us time-series like the example shown in Figure 3.1. This figure shows how a time-serie is generated per metric and team. Later, we apply clustering to group the time-series per metric.

Those time series are the first step to create similarity between teams, as their performance may be compared in a mathematical way. Next we need to group the time series per metric using clustering. In Figure 3.2 we can see and example of how different teams time-series are assigned to the different clusters per metric. In this first approach, our model used hierarchical clustering[2] based on the correlation dissimilarities between team time series about the same statistic. After all series were grouped in clusters per metric and the matrix related to Figure 3.2is generated, similarity between teams were calculated using the following formula:

$$sim(t_i, t_j) = \frac{\sum_{C_q} \delta^i_{C_q} \cdot \delta^j_{C_q}}{M} \tag{3.1}$$

Where M is the number of chosen metrics, t_i, t_j are the teams to be compared, C_q represents the posisble clusters per metric, and $\delta^i_{C_q}$ defines whether the metric is included in the clusters or not by the following definition:

$$\delta^i_{C_q} = \begin{cases} 1 & \text{if } t_i \in C_q \\ 0 & \text{Otherwise} \end{cases}$$

For example, according to Figure 3.2 *Team 1* and *Team 2* are similar at 3 metrics (C_1^{hr}, C_1^{er} and C_1^{wh}) so if we calculate their similariti we would get:

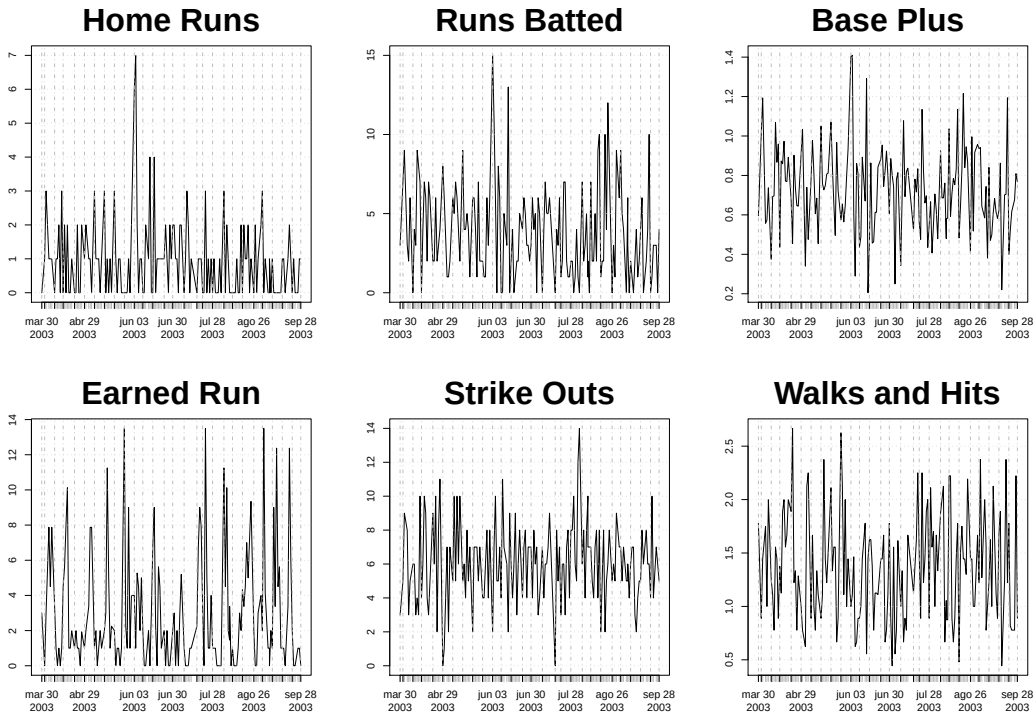

Figure 3.1: Example of Team Time Series (Anaheim Angels during the 2003 season).

Metric	Team 1	Team 2	Team 3	...	Team N
Home Runs	C_1^{hr}	C_1^{hr}	C_2^{hr}	...	C_5^{hr}
Runs Batted	C_1^{rb}	C_2^{rb}	C_2^{rb}	...	C_1^{rb}
Base Plus	C_1^{bp}	C_3^{bp}	C_2^{bp}	...	C_5^{bp}
Earned Run	C_1^{er}	C_1^{er}	C_3^{er}	...	C_2^{er}
Strike Outs	C_1^{so}	C_4^{so}	C_2^{so}	...	C_2^{so}
Walks and Hits	C_1^{wh}	C_1^{wh}	C_4^{wh}	...	C_3^{wh}

Figure 3.2: Clusters matrix of the 6 chosen metrics for all teams.

$$sim(Team1, Team2) = \frac{1+0+0+1+0+1}{6} = 0,5$$

Using these formulas, a similarity matrix for teams was created in order to have all similarities between each pair of teams within the training set. Even though this would not be used for similarities to new teams, it is necessary for similarities between games as we will see in the following section. This similarity matrix also works as a similarity graph, where teams are nodes and their similarity states how close those nodes are, taking the place of the edges of the graph. An example may be found in the right side of Figure 3.3.

3.2.2 Games Graph

Following the same process used for teams, time series for games were created in order to find similar games. However, there are differences between games and teams analysis. First, teams time-series used games as their time variable, but games time-series need a different one. As baseball statistics may be recorded per play and plays group within innings, innings where chosen as the time variable for games time-series. In the other hand, two teams play at each

game, so statistics from both, local and visitor teams, would work as metrics for the game time series. In first place, similarity between games is calculated according to their time-series clustering as did for the teams graph:

$$sim_c(g_i, g_j) = \frac{\sum_{C_q} \delta^i_{C_q} \cdot \delta^j_{C_q}}{M}$$

Where, similarly to Formula 3.1[3], M is the number of chosen metrics, g_i, g_j are the games to be compared, C_q represents the posisble clusters per metric, and $\delta^i_{C_q}$ defines whether the metric is included in the clusters or not by the following definition:

$$\delta^i_{C_q} = \begin{cases} 1 & \text{if } g_i \in C_q \\ 0 & \text{Otherwise} \end{cases}$$

As teams playing a specific game are also relevant, the game similarity formula must also take that relevance into consideration, so the second part of the similarity calculation is team-related. This model uses the similarity between teams playing the games in order to add information to games similarity. However, as each team plays whether as local team or visiting team during a game, the highest similarity between teams playing a pair games to be compared must be chosen. Based on this process, games similarity per teams is calculated using the following formula:

$$sim_t(g_i, g_j) = max\left(\frac{sim(t^1_i, t^1_j) + sim(t^2_i, t^2_j)}{2}, \frac{sim(t^1_i, t^2_j) + sim(t^2_i, t^1_j)}{2}\right)$$

Where t^1_k and t^2_k are respectively the local and visitor teams playing at game g_k and $sim(t^n_i, t^{n'}_j)$ is the similarity between teams according to the team similarity matrix generated before.

In the end, games similarity is calculated using both, clustering similarity and team-based similarity. In this approach, both methods are mixed taking a 50% weight each as shown in the following formula:

$$sim(g_i, g_j) = \frac{1}{2} \cdot sim_c(g_i, g_j) + \frac{1}{2} \cdot sim_t(g_i, g_j)$$

Using this formula, the games similarity matrix is generated as well as the games graph, as shown in the right side of Figure 3.3. This graph also shows how games are connected to teams.

3.2.3 New Game Prediction

Once the graphs are created, any game can be introduced to the model in order to be predicted. It is important to mention that, as this model uses past information to predict future behaviours, teams and games graphs must contain a large amount of samples of past games and teams, so their data can be statistically meaningful. As a new game is to be predicted its known previous information of the game is analyzed because it can be useful. In this first approach, only information about the teams playing the new game is used.

Due to team information is recorded within the graphs, new teams playing a game must be compared within the teams included in the graphs. When a new team is brought into the model, its past information is needed in order to generate time-series through statistics. This allows teams comparison between the new team time-series and the teams included in the team graphs. As recent information provides a more accurate information about team behaviour, last

[3]Note that C_q now correspond to the game metrics and may differ to the team statistics used for the teams formula as well as M may be a different number of metrics.

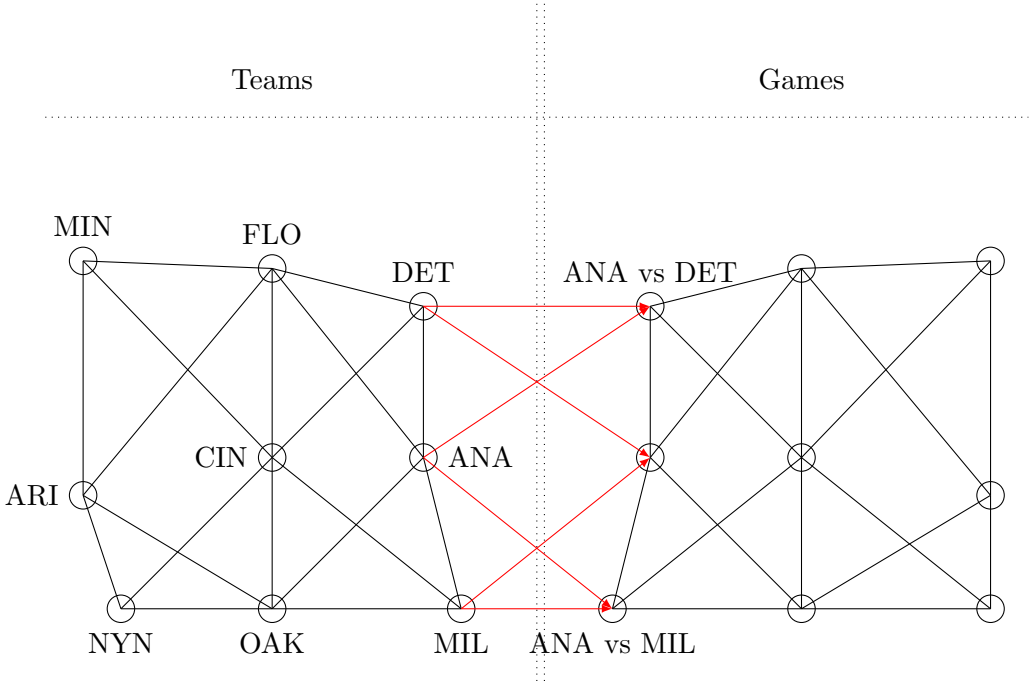

Figure 3.3: Representation of the similarity graphs amongst teams (left) and games (right) and the connections between these two graphs related to those teams which have played a match.

N games played by the teams will be used to extract team statistics and generate the time-series to be compared. For this approach, whether there is not enough past information from a team the prediction can not be made, but some ideas are explained in following sections in order to solve this feature in future work.

Using the similarities between the teams playing the new game and the teams within the graphs, the τ most similar teams are selected from the graph. Let T_1 be the τ most similar teams to t_1 and T_2 be the τ most similar teams to t_2, where t_1 and t_2 are the local and visitor teams of the new game respectively. A number of games G is then selected and the last G games played by each pair of teams (T_1^i, T_2^j) are extracted from the games graph[4]. Then, a similarity degree is calculated for each one of the G games in relation with the game to be predicted using the following formula:

$$sdeg(n, g_i) = \frac{1}{2} sim(t_1, T_1^i) + \frac{1}{2} sim(t_2, T_2^i)$$

Where n is the new game, g_i is one of the games to compare n with, t_k are teams playing at n and T_k^i are teams playing at g_i.

In the last step the game prediction is made by chosing the team with the most victory probability according to the following formula:

$$V(n) = max\left(\frac{\sum_{i=1}^{G} v(T_1^i) \cdot sdeg(n, g_i)}{G}, \frac{\sum_{i=1}^{G} v(T_2^i) \cdot sdeg(n, g_i)}{G}\right)$$

Where G is the number of previous games to compare the new one with and $v(T_k^i)$ is a function defined by:

$$v(T_k^i) = \begin{cases} 1 & \text{if } T_k^i \text{ won } g_i \\ 0 & \text{if } T_k^i \text{ lost } g_i \end{cases}$$

[4]In this approach only games with T_1^i as local team and T_2^j as visitor are chosen.

Following all these steps all the model is created and games may be forecasted. In the next section the model will be tested for a performance measurement by using some sample data from the Retrosheet Database to create the graphs and compare the predictions to the reality. Despite of being a first approach, the model combines a lot of features, such as clustering algorithms or similarity measures, that may be modificated in future works in order to adapt the model to future studies and possible improvements, as we will see in Section 5.

4

Experimental Results

This chapter shows the application of the architecture to the data in order to analyze the model performance. For this first approach, we have selected the games played at the National League and American League during the Regular Seasons and Playoffs (also called Post Season games) between 2003 and 2005. The reasons why we have chosen those three seasons are:

- **Training & Test Data**: In order to test our model, we selected the related data from 2003 and 2004 as Training Data and 2005 as Test Data. On the one hand, Training Data was used to create time series that will work as references to the games that would be predicted later. In the other hand, Test Data is meant to help at checking whether the prediction results are good or not by comparing the predicted games to their real outcome.
- **Moneyball Impact**: We have selected the three seasons that follow the 2002 season, where the Oakland Athletics started "playing Moneyball", creating a trend in this way so other teams addapted to this new way of fielding teams through signing players up by statistics.

4.1 Experimental Setup

This section explains the different parameters chosen for the model experimentation. In order to fill the model requirements, data selection must be done to have a sample to work with, as explained before, but also metrics must be chosen to be used for the prediction. Metrics, parameters and other features will be described in detail as well as the reasons why those have been chosen.

4.1.1 Time-series Parameters

For this first approach we chose six representative statistics that would act as variables for our time series, chosen from the statistics set the MLB organization records about the players that participate in the Major American Leagues[1]:

[1] All statistics recorded by the MLB organization may be found at http://mlb.mlb.com/stats/sortable.jsp

- (B)**Home Runs**: The number of times a batter hits the ball and reaches home plate scoring a run either by hitting the ball out of play in fair territory or without aid of an error or fielder's choice.
- (B) **Runs Batted In**: The number of runs scored safely due to a batter hitting a ball or drawing a base on balls.
- (B) **On Base Plus Slugging**: The number of times each batter reaches base by hit, walk or hit by pitch, divided by plate appearances including at-bats, walks, hit by pitch and sacrifice flies.
- (P) **Earned Run Average**: The average number of earned runs allowed by a pitcher, total number of earned runs allowed multiplied by 9 divided by the number of innings pitched.
- (P) **Strike Outs**: The number of strikeouts by a pitcher.
- (P) **Walks and Hits per inning**: The average number of walks and hits by a pitcher, hits plus walks allowed divided by innings pitched.

Measures preceded by B (Batting) are offensive statistics, while the ones preceded by P (Pitching) are defensive statistics.

These six metrics provide information about team performance and are quite related to the endgame score, which make them trully representative facing a result prediction for games. It is important to mention that there are 6 time-series per team, while there are 12 per game (6 corresponding to the local team and 6 to the visiting team).

As explained earlier, this model uses time-series clustering in order to group similar teams within the training set so hierarchical clustering analysis has been chosen, applied on time-series dissimilarities for this first experimentation. The reasons why hierarchical clustering has been chosen over other clustering techniques are, on the one hand, because hierarchical clustering is deterministic, and, on the other hand, it is typically used for time-series clustering. Time-series dissimilarity has been calculated using the correlation-based dissimilarity metric (wich computes the estimated Pearson correlation of two given time series), defined by:

$$d = \sqrt{\left(\frac{1-\rho}{1+\rho}\right)^{\beta}}$$

The selected value of β is 2 in order to obtain a positive value from the metric. The chosen number of clusters in this case of study after some testing has been 6, as the time-series distribution through clusters seemed good enough with a relatively small amount of clusters.

When comparing time-series, they must have the same length or the comparison could not be made. Following this rule we have to choose a length for our time-series. As our team time-series use games as the time variable, we decided to use the last 161 games as their length. The reason why this value was chosen is because a regular season for any team is 162 games long plus games played at play-offs, however, some games are sometimes suspended, and teams may have no time to play their continuation at the end of the season, so they may play less games. As this happens few often, teams play, at least, 161 games per season, hence we consider this value as the number of games needed for our time-series. In the other side of the graph, games use innings as their time variable for the time-series. Every game is played, at least, 9 innings long, so this is the chosen value as the length for games time-series. Games can last any number of innings while teams are tied, but extra innings do not provide representative

Parameter	value or value range	sequences (ranges)
S (t.s. metrics)	6	-
C (t.s. clusters)	6	-
P (past games)	161	-
N (new games)	100	-
I (iterations)	7	-
T (sim. teams)	[3-9]	2
G (sim. games)	[3-9], all	2

Table 4.1: Parameter selection for the model.

information about game outcome as teams play until the tie is broken, so their first 9 innings are collected into the time-series representing really close games with a virtual tie in the end. In the end, our training set provided us information about 60 teams (30 per season) and 4930 games in form of time series and clusters.

4.1.2 Prediction Parameters

Once the graphs are created using the parameters and algorithms described in the previous section, we tested the model using different parameters during the prediction phase. In this first approach, the only modified parameters are the number of similar teams for each team playing the game to be predicted and the number of games between similar teams[2]. These values set the amount of information from our training set to be compared in order to determine the new game result. Each pair of teams-games was tested 7 times in order to have a good sampling for the experiment. In each experimental iteration, 100 games where randomly selected from our test dataset, their result was predicted and later compared to the real result, so we could measure the performance of the model by the number of right predictions made over the total, as shown in Table 4.1.

4.2 Experimental Results

The experimental results have been analyzed in three different ways. The first analysis focuses the results by model, while the second and third analysis group the models by parameters in order to find the behaviour of the models by those different parameters as well as their best values.

Results by model

Table 4.2 shows the results of all the models which have been generated in the experimental phase. Each experimentation model consist on a pair of values of the parameters *"Similar Teams"* and *"Similar Games"*. The maximum, minimum, mean and standard deviation values of correct predictions for each model have been represented in this table, considering that each model has been tested 7 times on 100 random games per iteration, which means a total of 700 predicted games per model.

Checking the results shown in Table 4.2 we can have a clear overview of how parameters affect to the prediction rates. Small values for both, similar teams and games selected, show the

[2]An odd number of similar games has been selected in order to avoid unpredictable games due to both teams having same chances of victory according to the model formulas.

Teams-Games	Max	Min	Mean	SD
3-3	57.95%	42.53%	46.51%	± 0.0551
3-5	52.81%	47.19%	48.84%	± 0.0200
3-7	59.14%	41.76%	54.12%	± 0.0639
3-9	55.91%	42.70%	54.35%	± 0.0575
3-all	59.76%	42.17%	52.17%	± 0.0661
5-3	62.24%	43.16%	52.75%	± 0.0661
5-5	54.26%	40.00%	46.81%	± 0.0492
5-7	54.95%	47.87%	52.17%	± 0.0273
5-9	60.44%	47.87%	51.11%	± 0.0399
5-all	57.14%	42.22%	50.54%	± 0.0589
7-3	56.84%	43.33%	48.45%	± 0.0462
7-5	59.18%	43.96%	52.27%	± 0.0506
7-7	60.00%	***51.06%***	53.41%	± 0.0364
7-9	60.44%	42.22%	50.52%	± 0.0575
7-all	54.74%	46.32%	51.16%	± 0.0272
9-3	**63.44%**	45.45%	50.54%	± 0.0665
9-5	***62.77%***	**51.58%**	**56.12%**	± 0.0433
9-7	62.64%	38.30%	***56.04%***	± 0.0785
9-9	55.43%	42.55%	53.76%	± 0.0524
9-all	56.52%	39.78%	50.55%	± 0.0568

Table 4.2: Experiment results according to all the models which have been designed. Maximum, minimum, mean and standard deviation of correct predictions are shown.

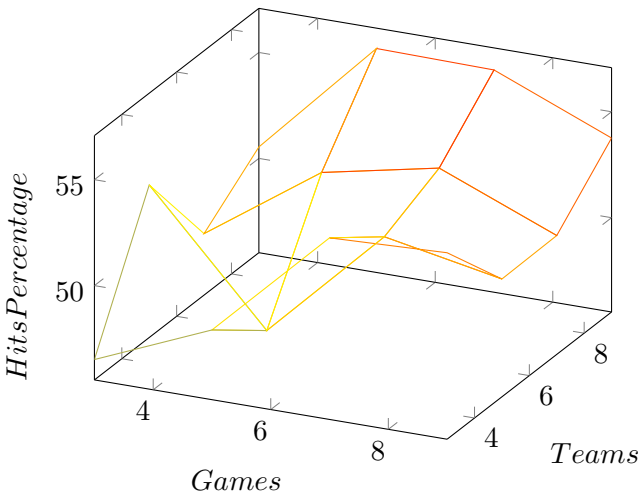

Figure 4.1: Mean percentage of correct predictions per model.

lowest mean results (46.51% for 3-3 and 46.81% for 5-5), while the highest values are found at a high number of teams and a medium number of similar games (56.12% for 9-5 and 56.04% for 9-7). Figure 4.1 shows graphically the evolution of the model performance through the different paremeter values according to the mean results. The highest maximum results are also found at a high number of similar teams but few similar games (63.44 for 9-3 and 62.77 for 9-5) while the lowest maximum is found at 3-5 (52.81%), again few teams few games. The highest minimums appear at a high number of similar games and medium number of similar games (51.58% for 9-5 and 51.06% for 7-7), but this time the lowest minimums are found at those models with the highest number of similar teams and a high number of similar games (38.30% for 9-7 and 39.78% for 9-all). The last column of the table shows the highest and lowest standard deviation

values at 3-5 (±0.0200) and 9-7 (± 0.0785) respectively. However, there is not a pattern that states how the standard deviation, hence the models regularity, behaves according to this pair of parameters.

Results by similar teams

Table 4.3 shows the results grouped by number of similar teams chosen for each team of the two playing a game to be predicted. As we have done at model analysis, the maximum, minimum, mean and standard deviation values of correct predictions are shown in this table. In this case, we can analyze how the model behaves by the modification of a single parameter in order to understand its performance, as well as finding the best values for this parameter.

Teams	Max	Min	Mean	SD
3	59.76%	***41.76*** %	50.00%	± 0.0529
5	***62.24%***	40.00%	51.09%	± 0.0517
7	60.44%	**42.22%**	***52.27%***	± 0.0460
9	**63.44%**	38.30%	**53.68%**	± 0.0611

Table 4.3: Experiment results grouped by chosen number of similar teams for all the different number of similar games posibilities.

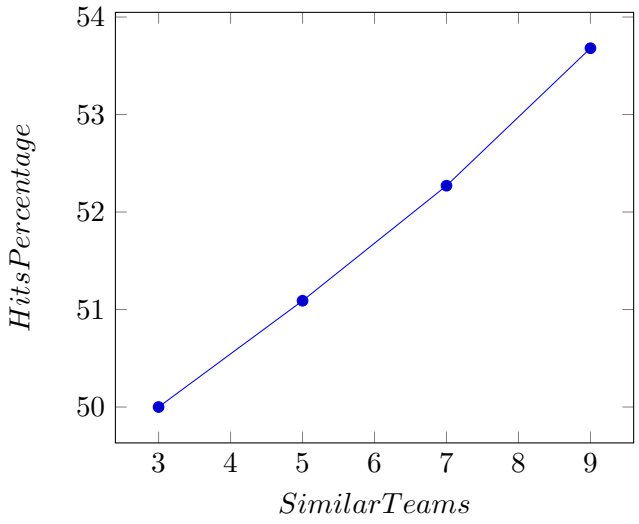

Figure 4.2: Mean percentage of correct predictions per similar teams.

Based on this parameter, the experiment results show how, increasing the number of similar teams, the mean percentage of good predictions increases too (as shown in Figure 4.2), being 9 the value for this parameter with a 53.68% of good predictions. However, logic states that, the more teams chosen, the less similar they will be and the less accurate the prediction should be as happens with similar games. Further experimentation should show the value that makes the model reach its highest performance, according to mean results, based on the number of similar teams selected. In this experiment, both, the maximum and the minimum values fluctuate over the parameter values without providing any clear information about their behaviour. Finally, and according to the results table, the parameter with the highest stability is 7 (± 0.0460 of standard deviation) while the lowest is 9 (± 0.0611), but, as in the model analysis, the standard deviation varies without a meaningful pattern.

Results by similar games

Table 4.4 shows the results by number of similar games per pair of similar teams. It is important to mention that the total number of similar games at each iteration does not grow linearly, as it depends on the number of similar teams, nor it is a consist number of games, as each pair of similar teams may have not played the same number of games. So the total number of similar games is:

$$G \leq N^2 \times M$$

Where N is the value of the *"Similar Teams"* parameter and M is the value of the *"Similar Games"* parameter.

Games	Max	Min	Mean	SD
3	**63.44%**	**42.53%**	48.95%	± 0.0571
5	***62.77%***	40.00%	50.83%	± 0.0525
7	62.64%	38.30%	**53.76%**	± 0.0545
9	60.44%	***42.22%***	***51.87%***	± 0.0499
all	59.76%	39.78%	50.86%	± 0.0521

Table 4.4: Experiment results grouped by chosen number of similar games for all the different number of similar teams posibilities.

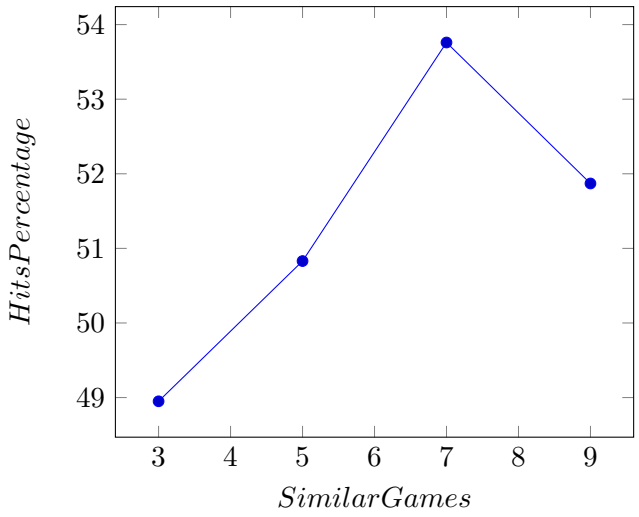

Figure 4.3: Mean percentage of correct predictions per similar games.

Results based only on number of similar games show an interesting performance. On the one hand, the number of similar games chosen for the prediction model seems to have its highest hit ratio at a value of 7 (53.76%). Higher or lower values do not make the model perform as well, but the closer they are to 7 the better they are, drawing a curve, as shown in Figure 4.3. On the other hand, the highest maximum values are shown with small number of games (63.44% for 3 similar games) and decreases as the number of similar games increases (59.76% for all possible similar games). In this case the highest minimum occurs again at 3 similar games (42.53%), but there is no consistency at minimum values behaviour through this parameter. In this table all values show a standard deviation around ± 0.05 without any growing pattern which means this parameter is not a relevant factor for the model consistency. The "all similar matches" choice does not achieve remarkable results in any case, which suggests that it is not relevant to include

all the information, but it is enough to use relevant and more reduced information sources. This makes sense as the more similar games are chosen, the less grouped they are chronologically speaking, so they provide less trustable information.

4.3 Discussion

Results have shown how different metrics may cause significative differences at the model performance. This model still allows much more analysis by modifying a small set of metrics and parameters as well as the possibility of being tested with different datasets, but the first experiments have been satisfactory facing the novelty of the project and knowing all the improvement potential behind the future work related to it.

In order to improve the model, a confidence factor could be added. This feature would reinforce the realiability of the predictions, specially for those scenarios where the number of well predicted matches is lower or inconsistent. Having a games graph, clustering can be applied to group the different games by similarity. Once the games within the graph were grouped and having a new game to be predicted, the distribution of those similar games through the clusters could be used to determine a confidence degree to state whether the prediction of the new game is consistent or not. For example, the highest number of games that are in the same cluster divided by the total number of similar games could work as confidence degree:

$$confidence(g_i) = \frac{max_q C_q}{\sum_{q=1}^{N} C_q} \tag{4.1}$$

This confedence degree would not improve the predictions made by the model, but would help to measure their performance by finding which predictions are trustworthy and which are not. This task is not part of this study and would require a lot of effort, time and resources, so we just mention the possibility of adding this kind of feature as it could be an interesting task for future approaches.

5

Conclusions and Future Work

5.1 Conclusions

This project has proposed a prediction model based on similarities through time-series clustering in order to forecast baseball games from the Major League Baseball. In order to achieve this goal, ingame data must me analyzed and managed, being the source of the time-series that will lately be clusterized and compared. Compared time-series result in graphs that connect teams and games, being the tool that will be used in the prediction phase. Once a new game is to be predicted, its known data is analyzed, playing teams' information is extracted in order to generate time-series, so information of the teams' status at the moment of the game is obtained and may be compared to the graphs information. Similar teams are suposed to play similar games, so past similar games provide the information needed to know the chances each team has to win the game.

Model experimentation has provided us some interesting results to be analyzed, despite of the early stage of the model:

- The model has shown an acceptable performance at predicting game results, facing the complexity of this task.
- Different parameters show different performances, allowing us to find metrics that work better for the model.
- Not only the prediction performance but the consistency of the parameters could be studied by the statistical measures provided by mean and standard deviation.
- There is still a lot of potential improvement for the model and future studies about this topic, as explained in the following section.

It is important to mention that large amounts of data may create bottlenecks in the data analysis. Specifically we found some issues at data transformation and time-series comparison, in both graph generation and new games prediction, as we were dealing with thousands of games to be compared one by one.

5.2 Future work

As this is new prediction methodology, there is still room for improvement and a long path to walk:

- The first and most obvious improvement would be the application of this model to other sports. However, it is difficult to find dataset that provide detailed, meaningful and manageable information as Retrosheet does.
- Different parameters or metrics may show different results for this model and a larger experimentation could show the best ones. Also testing the already used parameters with different seasons could provide us more accurate results, as we are working with random samples.
- Back to baseball specifics, there are factors that could also be relevant for the prediction, and may be added to the model in order to provide information, such as weather, field conditions or umpires.
- While this approach of the model based on teams, the model could also be oriented to a player analysis. While teams keep being the same through time, players come and go and even the starting roster may vary a lot from game to game. Injuries are a good example of player-related circumstances that may hit team's performance, even if the team keeps using the same strategies and tactics.
- Adding a confidence degree, as explained in Section 4.3, could provide reliability to the predictions in order to state whether it is consistent or not.
- This model finds some situations where predictions can not be made. E. g., lack of information from teams which are new to the league could be fixed by using information about the performance of games that were new in past seasons.

Glossary

- **TS**: Time-serie
- **BB**: Baseball
- **MVP**: Most Valuable Player
- **HR**: Home Runs
- **RBI**: Runs Batted In
- **OBPS**: On Base Plus Slugging
- **ERA**: Earned Runs Average
- **SO**: Strikes Out
- **WHPI**: Walks and Hits Per Inning
- **Sim**: Similarity

Bibliography

[1] T. Warren Liao. Clustering of time series data—a survey. *Pattern Recognition*, 38(11):1857 – 1874, 2005.

[2] Rui Xu and Don Wunsch. *Clustering (IEEE Press Series on Computational Intelligence)*. Wiley-IEEE Press, illustrated edition edition, October 2008.

[3] RobertP. Schumaker, OsamaK. Solieman, and Hsinchun Chen. *Sports Data Mining*. Integrated Series in Information Systems. Springer US, 2010.

[4] Gordon Waitt. Social impacts of the sydney olympics. *Annals of Tourism Research*, 30(1):194 – 215, 2003.

[5] Choong-Ki Lee and Tracy Taylor. Critical reflections on the economic impact assessment of a mega-event: the case of 2002 {FIFA} world cup. *Tourism Management*, 26(4):595 – 603, 2005.

[6] Brenda G Pitts and David Kent Stotlar. *Fundamentals of sport marketing*. Fitness information technology Morgantown, WV, 2002.

[7] Nidhal Ben Abdelkrim, Saloua El Fazaa, and Jalila El Ati. Time–motion analysis and physiological data of elite under-19-year-old basketball players during competition. *British Journal of Sports Medicine*, 41(2):69–75, 2007.

[8] Steven Pinch and Nick Henry. Paul krugman's geographical economics, industrial clustering and the british motor sport industry. *Regional Studies*, 33(9):815–827, 1999.

[9] David Forrest and Robert Simmons. Sport and gambling. *Oxford Review of Economic Policy*, 19(4):598–611, 2003.

[10] Brenda G Pitts, Lawrence W Fielding, and Lori K Miller. Industry segmentation theory and the sport industry: Developing a sport industry segment model. *Sport Marketing Quarterly*, 3(1):15–24, 1994.

[11] Yong Jae Ko and Donna L Pastore. A hierarchial model of service quality for the recreational sport industry. *Sport Marketing Quarterly*, 14(2), 2005.

[12] James M Gladden and Daniel C Funk. Developing an understanding of brand associations in team sport: Empirical evidence from consumers of professional sport. *Journal of Sport management*, 16(1), 2002.

[13] Hans H Bauer, Nicola E Stokburger-Sauer, and Stefanie Exler. Brand image and fan loyalty in professional team sport: A refined model and empirical assessment. *Journal of sport Management*, 22(2), 2008.

[14] Albert V Carron, Michelle M Colman, Jennifer Wheeler, and Diane Stevens. Cohesion and performance in sport: A meta analysis. *Journal of Sport & Exercise Psychology*, 24(2), 2002.

[15] GC Roberts and Y Ommundsen. Effect of goal orientation on achievement beliefs, cognition and strategies in team sport. *Scandinavian journal of medicine & science in sports*, 6(1):46–56, 1996.

[16] Wade D Gilbert and Pierre Trudel. Role of the coach: How model youth team sport coaches frame their roles. *Sport Psychologist*, 18(1), 2004.

[17] William McTeer, Phillip G White, Sheldon Persad, et al. Manager/coach mid-season replacement and team performance in professional team sport. *Journal of Sport Behavior*, 18(1):58–68, 1995.

[18] Joseph Baker, Jeane Cote, and Bruce Abernethy. Sport-specific practice and the development of expert decision-making in team ball sports. *Journal of applied sport psychology*, 15(1):12–25, 2003.

[19] G.M. Verrall, Y. Kalairajah, J.P. Slavotinek, and A.J. Spriggins. Assessment of player performance following return to sport after hamstring muscle strain injury. *Journal of Science and Medicine in Sport*, 9(1–2):87 – 90, 2006.

[20] Iqbal Surve, Martin P. Schwellnus, Tim Noakes, and Carl Lombard. A fivefold reduction in the incidence of recurrent ankle sprains in soccer players using the sport-stirrup orthosis. *The American Journal of Sports Medicine*, 22(5):601–606, 1994.

[21] Bradley Wilson, Constantino Stavros, and Kate Westberg. Player transgressions and the management of the sport sponsor relationship. *Public Relations Review*, 34(2):99 – 107, 2008. Special Issue: Public Relations and Sport.

[22] Daniel T. Larose. *Discovering Knowledge in Data*. John Wiley & Sons, 2005.

[23] E.R. Hruschka, R.J.G.B. Campello, A.A. Freitas, and A.C.P.L.F. de Carvalho. A survey of evolutionary algorithms for clustering. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 39(2):133 –155, march 2009.

[24] J. B. Macqueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[25] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[26] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[27] Hector Menendez, Gema Bello-Orgaz, and David Camacho. Extracting behavioural models from 2010 fifa world cup. *Journal of Systems Science and Complexity*, 26:43–61, 2013.

[28] Roberto N. Onody and Paulo A. de Castro. Complex network study of brazilian soccer players. *Phys. Rev. E*, 70:037103, Sep 2004.

[29] B Dawson, R Hopkinson, B Appleby, G Stewart, and C Roberts. Player movement patterns and game activities in the australian football league. *Journal of Science and Medicine in Sport*, 7(3):278 – 291, 2004.

[30] Inderpal Bhandari, Edward Colet, Jennifer Parker, Zachary Pines, Rajiv Pratap, and Krishnakumar Ramanujam. Advanced scout: Data mining and knowledge discovery in nba data. *Data Mining and Knowledge Discovery*, 1(1):121–125, 1997.

[31] E. Bittner, A. NuBbaumer, W. Janke, and M. Weigel. Self-affirmation model for football goal distributions. *EPL (Europhysics Letters)*, 78(5):58002, 2007.

[32] Pedro O.S. Vaz de Melo, Virgilio A.F. Almeida, and Antonio A.F. Loureiro. Can complex network metrics predict the behavior of nba teams? In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 695–703, New York, NY, USA, 2008. ACM.

[33] Taylor Raines, Milind Tambe, and Stacy Marsella. Automated assistants to aid humans in understanding team behaviors. In Manuela Veloso, Enrico Pagello, and Hiroaki Kitano, editors, *RoboCup-99: Robot Soccer World Cup III*, volume 1856 of *Lecture Notes in Computer Science*, pages 85–102. Springer Berlin Heidelberg, 2000.

[34] Z. Ivankovic, M. Rackovic, B. Markoski, D. Radosav, and M. Ivkovic. Analysis of basketball games using neural networks. In *Computational Intelligence and Informatics (CINTI), 2010 11th International Symposium on*, pages 251–256, Nov 2010.

[35] Andy Cox and John Stasko. Sportsvis: Discovering meaning in sports statistics through information visualization. In *Compendium of Symposium on Information Visualization*, pages 114–115. Citeseer, 2006.

[36] Jahn K Hakes and Raymond D Sauer. An economic evaluation of the moneyball hypothesis. *The Journal of Economic Perspectives*, 20(3):173–185, 2006.

[37] Max Marchi and Jim Albert. *Analyzing Baseball Data with R.* CRC Press, Taylor and Francis Group, 2013.

A

Retrosheet Software Output

This appendix contains information about the output provided by the Retrosheet Software Tools used for this project. As *BGAME.EXE* and *BEVENT.EXE* had the function of parsing the event files provided by Retrosheet to collect the play-by-play information and translate it into a more suitable format for automatization processes, it is important to analyze the output of these programs in order to understand the information they provide. The output is a group of entries compound by different fields. Each field has been analyzed and explained in the following tables. The tables have been divided in four columns:

- **FIELD**: This column shows the name of the field according to the program's output
- **FORMAT**: This column states whether the value is in a categorical or numerical format. It is important to mention that some numerical fields such as *wind direction*, *field condition* or *precipitaion* use numerical value as a classification for different situations (e.g., a *precipitaion* value of 4 means *Rain* and 5 means *Snow*). In a similar way, categorical fields like *pictch sequence* or *event text* are codes that describe the play in a compact manner but knowing how Retrosheet records the plays is necesary to understand what those codes mean[1].
- **CHRONOLOGY**: (BGAME only) States the moment the information within the field is known (previous to the start of the game, or as result of the game being played).
- **INFERRED FROM**: (BEVENT only) This column says where the field information can be found aside from the field itself (e.g., most of play details may be found in the *event text field* but many fields such as *passed ball flag* contain just specific information the play that help to the information access).
- **MEANING**: This column explains what kind of information the field contains if the field name is not clear enough.

[1]Detailed information about these fields may be found at http://www.retrosheet.org/eventfile.htm

FIELD	FORMAT	CHRONOLOGY	MEANING
game id	Categorical	Pregame	
date	Categorical	Pregame	
game number	Numerical	Pregame	
day of week	Categorical	Pregame	
start time	Categorical	Pregame	
DH used flag	Categorical	Pregame	Designated hitter flag
day night flag	Categorical	Pregame	Day/night flag
visiting team	Categorical	Pregame	
home team	Categorical	Pregame	
game site	Categorical	Pregame	
vis starting pitcher	Categorical	Pregame	
home starting pitcher	Categorical	Pregame	
home plate umpire	Categorical	Pregame	
first base umpire	Categorical	Pregame	
second base umpire	Categorical	Pregame	
third base umpire	Categorical	Pregame	
left field umpire	Categorical	Pregame	
right field umpire	Categorical	Pregame	
attendance	Numerical	Pregame	
temperature	Numerical	Pregame	
wind direction	Numerical	Pregame	
wind speed	Numerical	Pregame	
field condition	Numerical	Pregame	
precipitation	Numerical	Pregame	
sky	Numerical	Pregame	
time of game	Numerical	Postgame	
number of innings	Numerical	Postgame	
visitor final score	Numerical	Postgame	
home final score	Numerical	Postgame	
visitor hits	Numerical	Postgame	
home hits	Numerical	Postgame	
visitor errors	Numerical	Postgame	
home errors	Numerical	Postgame	
visitor left on base	Numerical	Postgame	
home left on base	Numerical	Postgame	
winning pitcher	Categorical	Postgame	
losing pitcher	Categorical	Postgame	
save for	Categorical	Postgame	Pitcher credited with Save
GW RBI	Categorical	Postgame	Batter credited with the game-winning run batted in
visitor batter 1	Categorical		

Table A.1: Bgame output fields

FIELD	FORMAT	CHRONOLOGY	DESCRIPTION
visitor position 1	Numerical	Pregame	
visitor batter 2	Categorical	Pregame	
visitor position 2	Numerical	Pregame	
visitor batter 3	Categorical	Pregame	
visitor position 3	Numerical	Pregame	
visitor batter 4	Categorical	Pregame	
visitor position 4	Numerical	Pregame	
visitor batter 5	Categorical	Pregame	
visitor position 5	Numerical	Pregame	
visitor batter 6	Categorical	Pregame	
visitor position 6	Numerical	Pregame	
visitor batter 7	Categorical	Pregame	
visitor position 7	Numerical	Pregame	
visitor batter 8	Categorical	Pregame	
visitor position 8	Numerical	Pregame	
visitor batter 9	Categorical	Pregame	
visitor position 9	Numerical	Pregame	
home batter 1	Categorical	Pregame	
home position 1	Numerical	Pregame	
home batter 2	Categorical	Pregame	
home position 2	Numerical	Pregame	
home batter 3	Categorical	Pregame	
home position 3	Numerical	Pregame	
home batter 4	Categorical	Pregame	
home position 4	Numerical	Pregame	
home batter 5	Categorical	Pregame	
home position 5	Numerical	Pregame	
home batter 6	Categorical	Pregame	
home position 6	Numerical	Pregame	
home batter 7	Categorical	Pregame	
home position 7	Numerical	Pregame	
home batter 8	Categorical	Pregame	
home position 8	Numerical	Pregame	
home batter 9	Categorical	Pregame	
home position 9	Numerical	Pregame	
visiting finisher	Categorical	Postgame	Visiting finisher pitcher
home finisher	Categorical	Postgame	Home finisher pitcher

Table A.2: Bgame output fields

FIELD	FORMAT	INFERRED FROM	DESCRIPTION
game id	Categorical	Game	Game identifier
date	Categorical	Bgame	
home team	Categorical	Bgame	
visiting team	Categorical	Bgame	
inning	Numerical		
batting team	Numerical		
outs	Numerical	pitch sequence	
balls	Numerical	pitch sequence	
strikes	Numerical	pitch sequence	
pitch sequence	Categorical		Pitch sequence at batting phase
vis score	Numerical		Visiting team score
home score	Numerical		Home team score
batter	Categorical		
batter hand	Categorical	Jugador	
res batter	Categorical		Responsible batter
res batter hand	Categorical	Jugador	Responsible batter's hand
pitcher	High	Categorical	
pitcher hand	Categorical	Jugador	
res pitcher	Categorical	Responsible pitcher	
res pitcher hand	Categorical	Jugador	Responsible pitcher's hand
catcher	Categorical		
first base	Categorical		
second base	Categorical		
third base	Categorical		
shortstop	Categorical		
left field	Categorical		
center field	Categorical		
right field	Categorical		
first runner	Categorical		
second runner	Categorical		
third runner	Categorical		
event text	Categorical	event text	Event description
leadoff flag	Categorical	event text	Lead-off flag
pinchhit flag	Categorical		Pinch-hit flag
defensive position	Numerical	Bgame	Batter's defensive position
lineup position	Numerical	Bgame	Batter's position within the batting order

Table A.3: Bevent output fields

FIELD	FORMAT	INFERRED FROM	DESCRIPTION
event type	Numerical	event text	Event numerical classification (0-23)
batter event flag	Categorical	event text	
ab flag	Categorical	event text	At bat flag
hit value	Numerical	event text	Base reached by batter
sh flag	Categorical	event text	Sacrifice hit flag
sf flag	Categorical	event text	Sacrifice fly flag
outs on play	Numerical	event text	
double play flag	Categorical	event text	
triple play flag	Categorical	event text	
rbi on play	Numerical	event text	Runs batted in
wild pitch flag	Categorical	event text	
passed ball flag	Categorical	event text	
fielded by	Numerical	event text	Defending player fielding the ball
batted ball type	Categorical	event text	Batted ball classification (G/L/F/P)
bunt flag	Categorical	event text	
foul flag	Categorical	event text	
hit location	Categorical	event text	Ball fielding location
num errors	Numerical	event text	Number of errors
1st error player	Numerical	event text	
1st error type	Categorical	event text	
2nd error player	Numerical	event text	
2nd error type	Categorical	event text	
3rd error player	Numerical	event text	
3rd error type	Categorical	event text	
batter dest	Numerical	event text	Batter destiny
runner on 1st dest	Numerical	event text	Runner on 1^{st} destiny
runner on 2nd dest	Numerical	event text	Runner on 2^{nd} destiny
runner on 3rd dest	Numerical	event text	Runner on 3^{rd} destiny
play on batter	Categorical	event text	
play on runner on 1st	Categorical	event text	
play on runner on 2nd	Categorical	event text	
play on runner on 3rd	Categorical	event text	

Table A.4: Bevent output fields

FIELD	FORMAT	INFERRED FROM	DESCRIPTION
sb for runner on 1st flag	Categorical	event text	Stolen base for runner on 1^{st} flag
sb for runner on 2nd flag	Categorical	event text	Stolen base for runner on 2^{nd} flag
sb for runner on 3rd flag	Categorical	event text	Stolen base for runner on 3^{rd} flag
cs for runner on 1st flag	Categorical	event text	Caught stealing for runner on 1^{st} flag
cs for runner on 2nd flag	Categorical	event text	Caught stealing for runner on 2^{nd} flag
cs for runner on 3rd flag	Categorical	event text	Caught stealing for runner on 3^{rd} flag
po for runner on 1st flag	Categorical	event text	Put-out for runner on 1^{st} flag
po for runner on 2nd flag	Categorical	event text	Put-out for runner on 2^{nd} flag
po for runner on 3rd flag	Categorical	event text	Put-out for runner on 3^{rd} flag
responsible pitcher for runner on 1st	Categorical	event text	
responsible pitcher for runner on 2nd	Categorical	event text	
responsible pitcher for runner on 3rd	Categorical	event text	
new game flag	Categorical	event text	
end game flag	Categorical	event text	
pinchrunner on 1st	Categorical	event text	
pinchrunner on 2nd	Categorical	event text	
pinchrunner on 3rd	Categorical	event text	
runner removed for pinchrunner on 1st	Categorical	event text	
runner removed for pinchrunner on 2nd	Categorical	event text	
runner removed for pinchrunner on 3rd	Categorical	event text	
batter removed for pinchhitter	Categorical	event text	
position of batter removed for pinchhitter	Numerical	Bgame	
fielder with first putout	Numerical	event text	
fielder with second putout	Numerical	event text	
fielder with third putout	Numerical	event text	
fielder with first assist	Numerical	event text	
fielder with second assist	Numerical	event text	
fielder with third assist	Numerical	event text	
fielder with fourth assist	Numerical	event text	
fielder with fifth assist	Numerical	event text	
fielder with sixth assist	Numerical	event text	
fielder with seventh assist	Numerical	event text	
event num	Numerical		Event number

Table A.5: Bevent output fields