# Using AI to Correct Play-by-play Substitution Errors

Steven Wu and Tim Swartz
Simon Fraser University
Email: swa157@sfu.ca, tim@stat.sfu.ca

## 1. Introduction

The sophistication of analytics in the basketball community is at an all-time high due to the availability of spatio-temporal data in the NBA that is driving innovation. A significant proportion of the cutting edge research showcased in recent years of MIT SSAC is enabled by camera software that allows for tracking individual players and the ball. However, the cost of the camera software required to record this data is too high for virtually any basketball league other than the NBA. For those such leagues, the most common and acquirable granular data is play-by-play, which is manually recorded by scorekeepers. This method is inexpensive, as the only costs are for human labour and (optionally) software that assists the recording. For the uninitiated, play-by-play is a log that contains the details of the sequence of events (e.g.: steals, turnovers, shot attempts) that occur in a game of sport.

It was not too long ago when basketball's cutting edge analytics were derived from play-by-play data [1]. Smaller leagues that do not have the budget for tracking software, but which produce play-by-play, should be able to enjoy the same level of analytical discourse that the NBA has had. Access to these analytics improve coaching strategy, roster management, and fan engagement. However, the problem of smaller budgets is a compounding one that affects the quality of the play-by-play itself: the less money that the league has to spend on the data collection, the less reliable the data is with respect to the true events that occurred in the game.

An example of this problem is the data from the U Sports (formerly known as the Canadian Interuniversity Sport, or CIS) league, which will be the league analyzed throughout the rest of this paper. Specifically, smaller budgets affect the following factors that have a large influence on the data's resulting quality:

- software to assist in annotation: a keyboard software that maps shortcuts for event annotations presents the opportunity for mistakes made by mistyping keys
- wage: keepers are not paid like industry data entry professionals, and are usually students whose primary incentives and motivations to work are their interest in the sport
- training: there is no standardized training across the keepers for each school, leading to high variance in consistency and reliability of the keepers across schools

Carleton University's men's basketball team, winners in 12 of the past 14 years in the U Sports league [2,3] - and still hungry for any edge it can gain over its opponents – were interested in the impact of specific units of players deployed. Methods that can answer such a question, such as With Or Without You (WOWY) or adjusted plus-minus, are dependent on knowing which ten players are on the court at all times during the course of the game. It was found to be impossible to get sensible results using these proven methodologies because it relied on the play-by-play's substitution logs being accurate. An automated solution that can "clean" the play-by-play's substitutions to guarantee five players on each side of the court at all times is necessary – one that can suggest the five most likely players on the court that matches close to the reality of the game that occurred is ideal.

In this paper, we present a novel implementation of an artificial intelligence agent which uses the contextual data surrounding the substitutions to reliably infer who is actually on the court, guaranteeing five players on the court for each team at all times. Because the goal state of our AI agent is unknown (without watching film of games to verify the correctness of the recorded play-by-play substitutions) we define two performance measures that quantify the success of our agent. Using over 6000 games from the U Sports league, we discuss the results of our framework for automated play-by-play cleaning.

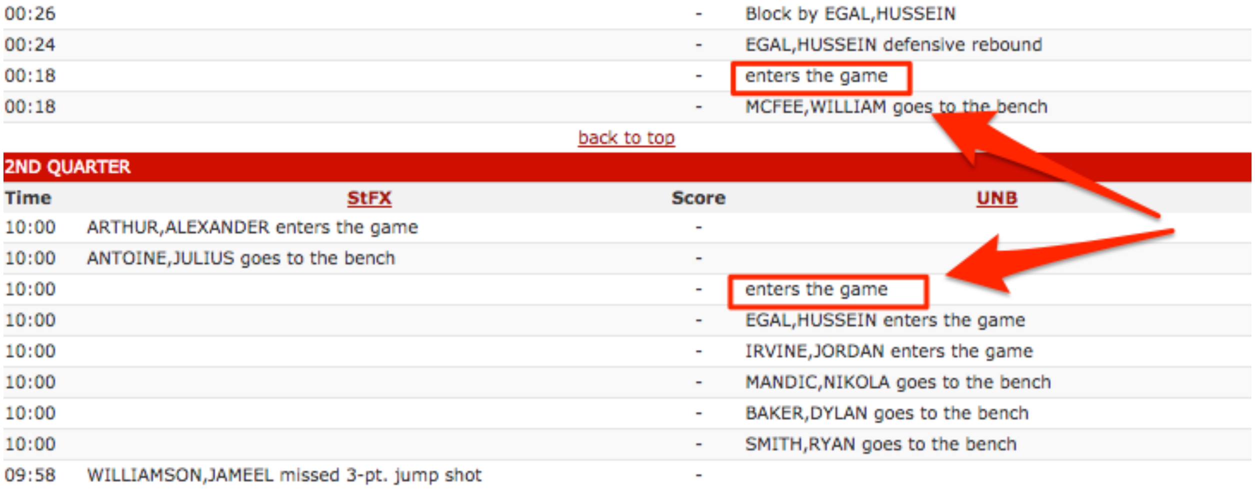
## 2. Data

U Sports, which has 40+ participating universities across Canada for both the men's and women's league, has published their game data online every year since the 2009-2010 season. The data collected for this paper is six seasons of the publicly available play-by-play and boxscore data. For every game, one group of workers are responsible for recording the play-by-play and another group of workers are responsible for recording the boxscore tallies.

The logging of substitutions in the play-by-play data is where we see the largest inconsistency in quality, with a variety of problems that occur repeatedly:

- recording an unequal number of players entering the game vs. going to the bench
- player's substitution patterns not alternating between entering the game vs. going to the bench (i.e.: a player is marked as going to the bench, then the next substitution involving him is him going to the bench again)
- recording the wrong player name, or not recording a name at all, in the substitution event
- missing substitutions (most frequently between quarters, but mid-quarter substitutions too)
- inconsistency between the recorded substitution and the events recorded before/after (e.g.: "DOE,JOHN goes to the bench" followed by "DOE,JOHN made layup")
- some games in each season have no substitutions recorded at all

The play-by-play has four columns: timestamp, away team plays, score, and home team plays. The recorded events are: goes to the bench, enters the game, foul, turnover, steal, block, defensive or offensive rebound, and made or missed 2-pt jumpshot/3-pt jumpshot/layup/dunk/free throw/tip in. An example of the play-by-play and a sample of the errors are below in Figures 1 and 2.



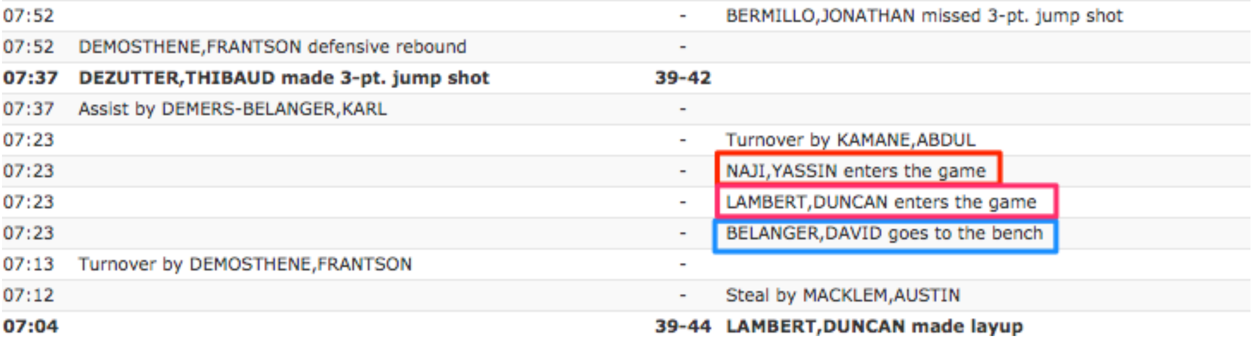
|  |  |  |  |
|---|---|---|---|
| 00:26 |  | - | Block by EGAL,HUSSEIN |
| 00:24 |  | - | EGAL,HUSSEIN defensive rebound |
| 00:18 |  | - | enters the game |
| 00:18 |  | - | MCFEE,WILLIAM goes to the bench |

back to top

2ND QUARTER

| Time | StFX | Score | UNB |
|---|---|---|---|
| 10:00 | ARTHUR,ALEXANDER enters the game | - |  |
| 10:00 | ANTOINE,JULIUS goes to the bench | - |  |
| 10:00 |  | - | enters the game |
| 10:00 |  | - | EGAL,HUSSEIN enters the game |
| 10:00 |  | - | IRVINE,JORDAN enters the game |
| 10:00 |  | - | MANDIC,NIKOLA goes to the bench |
| 10:00 |  | - | BAKER,DYLAN goes to the bench |
| 10:00 |  | - | SMITH,RYAN goes to the bench |
| 09:58 | WILLIAMSON,JAMEEL missed 3-pt. jump shot | - |  |

*Figure 1: An example of substitutions with no names. 201411121 StFX vs. UNB*



|  |  |  |  |
|---|---|---|---|
| 07:52 |  | - | BERMILLO,JONATHAN missed 3-pt. jump shot |
| 07:52 | DEMOSTHENE,FRANTSON defensive rebound | - |  |
| **07:37** | **DEZUTTER,THIBAUD made 3-pt. jump shot** | **39-42** |  |
| 07:37 | Assist by DEMERS-BELANGER,KARL | - |  |
| 07:23 |  | - | Turnover by KAMANE,ABDUL |
| 07:23 |  | - | NAJI,YASSIN enters the game |
| 07:23 |  | - | LAMBERT,DUNCAN enters the game |
| 07:23 |  | - | BELANGER,DAVID goes to the bench |
| 07:13 | Turnover by DEMOSTHENE,FRANTSON | - |  |
| 07:12 |  | - | Steal by MACKLEM,AUSTIN |
| **07:04** |  | **39-44** | **LAMBERT,DUNCAN made layup** |

*Figure 2: An example of uneven substitutions. 20160122 Laval vs. Bishops*

To aid our discussion, we will define terms that will repeatedly come up in the next sections:

- **stoppage:** the play-by-play row index where the clock is stopped and a substitution is permissible
- **substoppage:** the play-by-play row index where a stoppage in play occurs and at least one substitution was recorded in the play-by-play
- **active play:** a recorded play performed by a player that is not a substitution or technical foul

*Table 1: The frequency of errors we can objectively identify without cross-referencing the data with video*

| Season | # games | # games w/ 0 subs | Avg # sub stoppages | Avg # subs | Avg # unequal substoppages | Avg # non alternating subs | Avg # missing playername subs |
|---|---|---|---|---|---|---|---|
| **2009** | 798 | 112 | 39.71 | 115.30 | 2.87 | 55.46 | 0.000 |
| **2010** | 828 | 48 | 43.75 | 126.24 | 2.75 | 60.88 | 0.002 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **2011** | 829 | 43 | 42.45 | 122.87 | 2.70 | 59.12 | 0.008 |
| **2012** | 883 | 23 | 43.68 | 127.42 | 2.66 | 61.00 | 0.003 |
| **2013** | 911 | 1 | 44.45 | 129.34 | 2.69 | 61.29 | 0.015 |
| **2014** | 895 | 1 | 44.64 | 129.52 | 3.09 | 62.12 | 0.102 |
| **2015** | 905 | 1 | 44.30 | 130.10 | 2.30 | 62.35 | 0.107 |
| **Totals** | **6049** | **229** | **43.35** | **126.04** | **2.72** | **60.41** | **0.035** |

# 3. Automatically Cleaning Play-by-play Substitutions

## 3.1. Artificial Intelligence Techniques

In general, an agent is an entity with sensors to perceive its environment and actuators to act upon its environment. The computer representation, or model, of the environment at a given point in time is called the state. A utility-based agent uses a utility function (or heuristic) to map the current state to the utility of the state, and behave in a goal-directed manner to maximize its utility. The goal state of an artificial intelligence agent is either known (and the search is directed toward finding it) or the goal state is unknown (and the search is an exploration to try and find it, or the best possible solution if not the true solution).

Our goal state is the play-by-play which has the correct substitutions recorded to reflect the substitutions in the game that actually occurred. In our case, it is unknown – the only way to obtain it is to watch the game's film, which is not a feasible task given how many games there are in a season. In our case, we want the best approximation to the truth. Below we describe how we model the agent's environment and the heuristic functions it uses to move from state to state.

We can break our problem of cleaning a game into sub-problems of cleaning the substitutions of each period. For each period, we can further partition the play-by-play by its substoppages. Formally, denote the whole game's play-by-play as $\mathbb{P} = \cup_{j=1}^{J} P_n^j$, where $P_n^j$, is period *j*'s play-by-play containing *n* rows. For each period *j*, there is a set of substoppages $\mathbb{s}^j = \{ss_1, \dots, ss_d\}$ (for periods that do not have recorded substitutions attempting to account for transactions of players, i.e.: if $ss_1$ != 0, we add 0 to the set). Recall that each substoppage refers to the row index where substitution occurs, so we also have a substitution map for period *j*, $\mathbb{S}^j$, which maps substoppages to the substitutions observed at that substoppage.

The initial state is the play-by-play received, with the set of current players in the game $\omega = \emptyset$. The actions the agent performs are removing a recorded substitution or imputing a substitution that it believes should have been recorded, at each substoppage. For each period *j*, the agent iterates through each substoppage $ss_k \in \mathbb{s}^j$. We assign a confidence score to each recorded substitution found in $\mathbb{S}^j[ss_k]$ based on the contextual evidence and discard those substitutions which do not pass the classification threshold. If there are any correct substitutions, it updates $\omega$ by removing the players recorded as going to the bench and inserting the players recorded as entering the game.

The agent uses $P_n^j[ss_k: n]$ (where the $P[q: r]$ notation means the play-by-play from row index $q$ until row index $r$) to assign an activity score to every player and infer who the five most likely players are on the court for each side. Once all of the periods and substoppages are iterated through exactly once, the agent is finished its task.

From surveying the data and common errors, substitution plays cannot be trusted as much as the active plays. Intuitively, it is easier as a record keeper to assign the correct player to a single action involving the ball during a live play than to correctly account for up to ten players substituting for each team. Given this, we can use the active plays as contextual data at each substoppage in the game to infer what player transactions were most likely to have actually occurred. Specifically, for each substoppage, we have two problems that we need solutions for:
(1) remove the recorded substitutions that show enough evidence of being incorrect (detailed in **Section 3.2**)
(2) given the remaining substitutions, infer and impute the substitutions that should have been recorded, ensuring five unique players are on for each team (**detailed in Section 3.3**)

## 3.2. Binary classification of recorded substitutions

An important task for our agent is to accurately classify whether a recorded substitution correct or not, only using the context of the play-by-play surrounding it. A natural question from coaching staff is: how close is the resulting mutated game to the truth? Thus, it is desirable for our system to improve with examples of play-by-play that have had the substitutions annotated carefully by a separate party.

To train our classifier, we obtained five video replays of full games featuring distinct teams from the 2015-2016 men's season, which had commentators and a running score count on the video feed. For each game, we recorded which players were actually on the court for each row of the play-by-play. Knowing the true five players on the court for each side, we were able to deduce which substitutions were recorded correctly or incorrectly. Features that we believed to be predictive in whether a substitution is recorded correctly or not were collected for each substitution from the annotated games, and are detailed in the following subsection.

The resulting dataset is 312 "enters the game" substitutions and 311 "goes to the bench" substitutions. 76.1% of the "enters the game" substitutions and 76.8% of the "goes to the bench" substitutions were correct.

### 3.2.1. Extracted Features

Using domain expertise from conversing with coaches, play-by-play record keepers, and from our own knowledge of the data, for each substitution $s_i$ in $\mathbb{S}^j[ss_k]$ for substoppage $ss_k$ we extract the following features for our model:

*Table 2: Description of extracted features for every recorded substitution in our dataset*

| Notation | Name | Explanation |
|---|---|---|
| $Y_i$ | Correct/incorrect substitution | Our response variable. The value is 1 when the recorded substitution $s_i$ is correct. |
| $X_{1,i}$ | Absolute difference of # in vs. # out for team of player in $s_i$ | Int. From 0 to 5. If $\mathbb{S}^j[ss_k]$ contains an unbalanced number of players entering and leaving, then all |

| | | |
|---|---|---|
| | | substitution's likelihoods of being correct should be equally "punished" |
| $X_{2,i}$ | Total number of substitutions for team of player in $s_i$ | Int. From 1 to 10. The higher the number of substitutions that a record keeper has to track for the team on this substoppage, the more opportunity for mistakes |
| $X_{3,i}$ | More than five substitutions of the same type for team of player in $s_i$ | Boolean. Like $X_1$, a signal that the score keepers made a mistake and that $s_i$ is likely incorrect |
| $X_{4,i}$ | Is beginning of quarter | Boolean. It is not a practice consistent among scorekeepers to record the transactions occurring between periods, and even if they are recorded, they are either wrong or redundant (recording a player entering who was already last recorded as entering) |
| $X_{5,i}$ | Previous substitution is opposite | Boolean. Whether the most recent previous substitution with the player's name is opposite to the type in $s_i$ |
| $X_{6,i}$ | Next substitution is opposite | Boolean. Whether the most recent next substitution with the player's name is opposite to the type in $s_i$ |
| $X_{7,i}$ | Player appears more than once in substoppage | Boolean. Whether the player appears more than once in $\mathbb{S}^j[ss_k]$ |
| $X_{8,i}$ | Plays before ratio | Float. This is the ratio of player activity before the substoppage (in $P_n^j[ss_{k-1}: ss_k]$). The numerator is the sum of active plays seen by the player in the substitution, and the denominator is the number of total active plays by the player's team. A small adjustment of 1.0 is added to both the numerator and denominator in cases where the denominator is 0. |
| $X_{9,i}$ | Plays after ratio | Same as $X_{8,i}$, except after the substoppage (in $P_n^j[ss_k: ss_{k+1}]$). |

### 3.2.2. Model for Classification

We use a logistic regression model for our classifier for the interpretability of the coefficients, the probabilistic framework (which allows us to adjust the classification thresholds), and the strong performance relative to the other classifiers we tried.

Since the coefficients vary drastically depending on the type of substitution (intuitively, for a "enters the game" substitution we do not want to see any active plays before the substitution and we do want to see active plays afterwards – and vice versa for "goes to the bench"), we train a separate model for each substitution type using the same predictors.

Thus our model for a substitution $i$'s of type $t$ correctness is:

$$P(Y_i = 1) = \sigma(X_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \beta_5 X_{5,i} + \beta_6 X_{6,i} + \beta_7 X_{7,i} + \beta_8 X_{8,i} + \beta_9 X_{9,i}))$$

where $\sigma(X) = \frac{\exp(X)}{1+\exp\,(x)}$

### 3.2.3. Experiment Results

| Variable | Coefficient |
|---|---|
| $X_{1,i}$ | -0.774 |
| $X_{3,i}$ | -0.875 |
| $X_{4,i}$ | 1.153 |
| $X_{5,i}$ | 0.045 |
| $X_{6,i}$ | -1.702 |
| $X_{7,i}$ | -0.284 |
| $X_{8,i}$ | 0.320 |
| $X_{9,i}$ | -1.872 |
| | |
| **10 CV score** | 88.0% |

***Table 3: Coefficients and 10 CV classification score for "goes to the bench" substitutions***

| Variable | Coefficient |
|---|---|
| $X_{1,i}$ | -0.906 |
| $X_{3,i}$ | -0.568 |
| $X_{4,i}$ | 0.786 |
| $X_{5,i}$ | -0.149 |
| $X_{6,i}$ | -1.072 |
| $X_{7,i}$ | -0.182 |
| $X_{8,i}$ | -3.234 |
| $X_{9,i}$ | 0.283 |
| | |
| **10 CV score** | 89.2% |

***Table 4: Coefficients and 10 CV classification score for "enters the game" substitutions***

As expected, our coefficients indicate decreasing likelihood of a correct substitution if there is an unbalanced number of substitutions recorded, if the substoppage is at the beginning of the quarter, and if the player appears more than once in the same substoppage. Our intuition for a relationship existing between the number of substitutions recorded and the probability of a substitution being a successful recording seems to hold from the data as well.

Note that $X_{2,i}$ was excluded from the results table due to lack of examples of substoppages with more than five substitutions of either type in our limited dataset.

## 3.3. Inferring substitutions that should have been recorded

After confidently determining which substitutions are incorrect, we discard them. Applying the remaining substitutions on $\omega$, we are left with either less than, exactly, or more than five players in $\omega$. In all cases, we are interested in knowing who are the most active players from $P_n^j[ss_k\colon ss_{k+1}]$.

The "enters the game" substitutions not recorded by the record keepers are easily inferred by sudden activity of a new player in $P_n^j[ss_k\colon ss_{k+1}]$ who is not in $\omega$. Active plays, especially numerous counts, are almost sure indicators that a player is on the court. However, a lack of active plays does not necessarily mean the player is on the bench. As the period progresses, there is less opportunity to make a play. Due to variance in player skill, if the player is not a contributor (in terms of tallied/recorded events) they may play long stretches without a logged event. Thus, the classification step prior to this one is important to gain information on these situations where the evidence does not make it obvious who is on (when a missed "enters the game" occurs for a player who did not contribute much, or a missed "goes to the bench").

$$\mathrm{AH}_p^l = \sum_{k=ss_l}^{ss_{l*}} \mathbb{I}_{k,j,p} + \frac{1}{3}\sum_{k=ss_{l*}}^{ss_{l**}} \mathbb{I}_{k,j,p} \tag{1}$$

Equation (1) shows our activity heuristic (AH) for player $p$ at substoppage $ss_l$, where

- $\mathbb{I}_{k,j,p} = \begin{cases} 1 & if\ row\ k\ in\ P_n^j\ contains\ player\ p\ making\ an\ active\ play \\ 0 & else \end{cases}$
- *l+1* is set to *n* if *l* is the last substoppage
- $l* = \begin{cases} n & if\ l\ is\ the\ last\ substoppage \\ l+1 & else \end{cases}$
- $l** = \begin{cases} n & if\ l\ is\ the\ last\ substoppage \\ row\ index\ after\ l*\ where\ 5\ unique\ players\ seen & else \end{cases}$

The reason why we add evidence from $P_n^j[ss_{l*}: ss_{l**}]$ is because (a) sometimes $ss_{l*}$ is too close to $ss_l$ to gather any meaningful evidence in $P_n^j[ss_l: ss_{l*}]$ and (b) the likelihood of all five players substituting off at $ss_l$ is unlikely, and so the evidence in $P_n^j[ss_{l*}: ss_{l**}]$ is likely to contain evidence about who should sub in at $ss_l$ (the evidence is down-weighted to reflect this uncertainty).

We sort for the players who were most active by our heuristic and add the most active players to $\omega$. In the event where there are not enough correct substitutions nor active players seen to deduce who the five players should be on the court (usually at substoppages with incorrect substitutions at the end of a period), the tiebreaker for players is the boxscore minutes.

Note: since deflections and out-of-bounds events are not recorded, not all stoppages can be determined from the play-by-play. Thus, we restrict ourselves to inferring substitutions only at the substoppages in the game. We can clean games that do not have substitutions at all, by replacing substoppages with the stoppages we can infer from the play-by-play (turnovers not forced by steal, fouls, timeouts, beginning/end of period).

## 4. Results

Agent systems are evaluated on performance measures: an objective criterion for success of an agent's behavior. We define two simple performance measures that can objectively quantify the result of our agent cleaning the play-by-play.

### 4.1. Minutes criterion

In the U Sports league, a separate party from the ones responsible for recording the play-by-play is responsible for compiling the boxscore statistics. It is important to emphasize the fact that the boxscore tallies do not result from the play-by-play itself. Since it is an account of the game from another objective party, we can use the minutes tallied in the boxscore to compare the minutes that are tallied from our cleaned boxscore.

$$\mathrm{MC}_i = \frac{1}{N}\sum_{k=0}^{N} |p_k^g - p_k^b| \tag{2}$$

Equation (3) shows the minutes criterion ($\mathrm{MC}_i$) calculated for each game *i*, where

- N = number of players in game *i*
- $p_k^g$ is player *k*'s minutes tallied from the cleaned game
- $p_k^b$ is player *k*'s minutes tallied from the boxscore

## 4.2. Unknown players criterion

Though less frequently occurring, there is the possibility of an incorrect player name recorded for an active play. Particularly, this occurs when a lot of active plays are in quick succession or when inexperienced record keepers get lost behind in the action. In the situation where the evidence suggests {P1, P2, P3, P4, P5} are on the court after $ss_j$, and in $P_n^j[ss_j : ss_{j+1}]$ there is a record of P6 performing an active play, then (taking our estimate as the true lineup on the floor) there is a contradiction that must be resolved. We resolve it by replacing the player name observed with an "UNKNOWN,UNKNOWN" player. The solution which minimizes the frequency of this occurring is a better solution, since our agent uses the active plays as evidence for inferring the correct substitutions.

$$\text{UPC}_i = \sum_{k=0}^{n_i} \mathbb{U}_k \qquad (3)$$

Equation (4) shows the unknown players criterion (UPC$_i$) calculated for each game *i*, where

- n$_i$ = number of play-by-play rows in game *i*
- $\mathbb{I}_k = \begin{cases} 1 & if\ the\ play\ in\ row\ k\ contains\ the\ \text{"UNKNOWN,UNKNOWN"}\ string \\ 0 & else \end{cases}$

## 4.3. Discussion of Results

*Table 4: Performance measures of algorithm for every season for men's and women's games*

| Season | Avg # unknowns per game | Avg # active plays per game | Avg seconds discrepancy per player | Avg # of players observed playing per game |
|---|---|---|---|---|
| **2009** | 2.96 | 376.60 | 192.92 | 20.20 |
| **2010** | 2.92 | 380.47 | 194.50 | 20.21 |
| **2011** | 3.12 | 377.52 | 194.87 | 20.33 |
| **2012** | 3.14 | 374.45 | 194.51 | 20.38 |
| **2013** | 3.23 | 375.46 | 199.10 | 20.51 |
| **2014** | 2.98 | 374.11 | 197.46 | 20.33 |
| **2015** | 3.01 | 377.12 | 197.27 | 20.28 |

| | | | | |
|---|---|---|---|---|
| **Total** | **3.08** | **376.33** | **197.30** | **20.33** |

Looking at the average number of "UNKNOWN,UNKNOWN" instances and active plays per game, we are unable to attribute a player that we estimate to be on the floor to an active play less than 1% of the time. The algorithm shows consistent performance across seasons with a very small sample of training data (relative to the number of games that have occurred).

It is worth noting that the play-by-play's timestamps are in MM:SS format, however the boxscores are only in MM format. Thus, even the correct play-by-play will have some discrepancy in minutes obtained from the log compared to minutes obtained from the boxscore.

## 5. Conclusions

In this work, we explored the effectiveness of an automated single agent framework that can clean play-by-play showing a variety of inconsistencies in recorded substitutions. The solution can improve with more data when it is fed examples of manually cleaned play-by-play. To the best of our knowledge, this is the first type of automated solution that solves the problems that result from human recorded basketball play-by-play. We define two performance measures and show that the agent, with a small amount of initial training data and simple heuristic functions, is objectively successful – with the average absolute difference between minutes extracted from the play-by-play and from the boxscore being approximately three minutes per player. For our specific application, the U Sports league, the analysis that can be derived from the cleaned play-by-play provide access to historical and current statistics beyond the boxscore, such as adjusted plus-minus and WOWY, to coaches and avid fans. For coaching staff, these metrics can inform strategy, decision making and roster management in a similar fashion to how counterparts in the NBA community took advantage of play-by-play derived metrics in the past decade. For media and fans, it introduces and amplifies the growing analytical discourse that our game is seeing. As a league that recently rebranded in 2016 to appeal to a wider audience, as well as to spread stories of young Canadian university athletes [4], this is a cost-effective method that can help accomplish both of its stated goals.

This approach can be extended to any other lower revenue leagues which suffer the same problems of possessing play-by-play containing manual errors which dramatically affect the results of metric calculations. Play-by-play is an important data medium, particularly for leagues that cannot afford the infrastructure for video tracking data. We believe this work is an important step in raising the awareness and the standard of analytics for many basketball leagues around the world.

## References

[1] Sill, J. (2013) *Improved NBA Adjusted +/- Using Regularization and Out-of-Sample Testing*. Paper presented at MITSSAC, 2010.
[2] Past Champions – CIS. http://en.cis-sic.ca/championships/mbkb/past_champs.
[3] Conn, J. R. (2014, March 3). The Canadian College Basketball Dynasty You've Never Heard Of. *Grantland*. Retrieved from http://grantland.com.
[4] Shoalts, D. (2016, October 10). CIS rebrands as U Sports, aims to bring student stories to Canadians. Retrieved from http://theglobeandmail.com.