The Influence of Strategic Potential of a Tennis Game on Effort: Understanding the Best Efforts Clause with `oktennis`

Omkar A. Katta*

December 18, 2018

1 Question

1.1 Motivation

Spurred by Nick Kyrgios's intentional loss against Mischa Zverev, tennis authorities and fans have shed a brighter spotlight on the rules and practices of tennis to identify and correct match-fixing (Clarey, 2016). In the Official Rulebook written by the Association of Tennis Players (ATP), Section VII Citation 4h(i) reads: "A player shall use his best efforts during the match when competing in a tournament" (ATP, 2018, p. 177). Otherwise known as the "Best Efforts Clause," this rule was established to discourage match-fixing and to avoid establishing the reputation of tennis as a money-making scheme. Ultimately, this rule enforces the belief that a match should not be intentionally lost.

The renewed emphasis on the Best Efforts Clause also begs the question of whether sets, games, or points can be intentionally lost. This is a complicated issue for the reason that the structure of tennis matches makes it possible for a player to win the match but lose more points or games than his opponent. In particular, if the goal is to win a match, then the players maximize the number of sets they win. Hence, players need not try to win every point or game if it is inconsequential to the outcome of the set or match.

To consider this in greater detail, we must consider the reasons that a player might lose a game on purpose. Assuming that there is a finite amount of energy that each player can exert during a match, each player must choose how to allocate that energy. Perhaps there is a benefit to saving energy and losing a game now to become better equipped to win a game in the future, i.e., "lose the battle to win the war." Players might also intentionally lose a game to give a false sense of confidence to his opponent. Apart from psychological warfare, another motivation could be to get to a changeover more quickly. A changeover is a two-minute window that usually occurs after every odd-numbered game and is when the players switch sides of the tennis court. During this time, players get to sit down, stay hydrated, and rework their strategy. Often times, players talk with their medical teams to ensure they have not been injured from being in such mentally and physically draining situations.

So, there are clearly some strategic benefits to losing a game in tennis. In particular, intentionally losing a game saves energy for future games and future matches. Identifying which games are strategically inconsequential to the outcome of a match will help determine whether players truly follow the Best Efforts Clause.

* I am grateful for the mentorship of Professor Francis DiTraglia and for the assistance given by Alejandro Sanchez Becerra, Barry Plunkett, Jonathan Sanchez Becerra, and the rest of the ECON224 class.

1.2 Methodology

We aim to provide a way to understand the importance of a game by determining its a priori strategic influence on the exertion of efforts' of players.

First, we focus our attention to the games in men's grand-slam matches where players must make a choice in the exertion of effort. Hence, we disregard any matches with tiebreakers; it would be a waste for players to throw the game of a tiebreaker since both players have worked hard to get to the point where they are one game away from winning a set. Hence, in tiebreakers, both players would choose to exert full effort, which defeats the purpose of our project.

In order to describe the gameplay, we define a **lagging player** to be the player who is losing the current set by at least four games. His opponent is defined as the **leading player**. We shall control for their abilities by taking a look at the differences in their rank, their age, their dominant hand, and the difference in the percentage of service games won.

Next, we define strategic potential of a game to be a binary variable that takes on the value of 1 if it is strategic to exert full effort in winning a game and 0 otherwise. We consider a very loose definition of strategic potential by assigning 1 to the games in which the opponent has already won 2 sets and 0 otherwise. If the opponent has already won 2 sets, then if the lagging player's goal is to win the match, he would prevent the leading player from winning any more games because the current set is the last set the leading player needs in order to win. Hence, the lagging player should exert full effort to win the game, which is why the strategic potential is defined to be 1. Notice that this definition of strategic potential is from the point of view of the lagging player. It answers the question: "is it strategically important for the lagging player to exert full effort?"

We would like our response variable to measure the exertion of effort, which is operationalized by the outcome of the game. This binary variable will take on the value of 1 if the lagging player wins the game and 0 otherwise. If the lagging player wins the game, then that means he has exerted enough effort to win the game despite the fact that he is losing the set. Note that another obvious choice of proxy is the outcome of the set. However, there are no instances in our data sample of a lagging player winning the set. Thus, the outcome of the game is a better indicator of the lagging players' effort levels.

There are two nontrivial interpretations to the relationship between the strategic potential and the outcome of the game. If the relationship is positive, that means that the lagging player recognizes the need to exert full effort when there is a chance that the current set is the final set of the match. In other words, when the pressure of losing the match is greatest, the lagging player exerts more effort in winning the game. This result demonstrates that players follow the Best Efforts Clause.

If the relationship is negative, then this means that the lagging player exerts more effort when he knows he has a better chance of winning the match. That is, if the leading player has already won 2 sets, the deficit in the number of games won by the lagging player is too great for him to overcome, thus resulting in less effort. However, if the leading player has not yet won 2 sets, then there is a chance for the lagging player to win the set and ultimately the match. This interpretation demonstrates that match-fixing could be problem in tennis because the lagging player might be violating the Best Efforts Clause.

Physical ability is an important factor in player-versus-player sports, and tennis is no exception. Thus, we must control for the different abilities of the lagging and leading players. We must also control for the history of the game. Players who are on a losing or winning streak are more likely to lose or win the following point or game, respectively. We crudely attempt to capture this correlative phenomenon by taking into account how far the lagging player is behind in the number of games in the set and the number of games in the overall match.

2 Literature Review

Due to the hierarchical structure of tennis (i.e., points are nested in games, which in turn are nested in sets that belong to a match), there was a focus on determining the probability of a player winning a match given the probability of a few basic parameters such as the probability of his winning a point on his serve (Croucher, 1986; Riddle, 1988; Newton and Keller, 2005). Unfortunately, during the inception of these models, these parameters were not measured due to the lack of technology needed to quantify them. The introduction of IBM's Hawk-Eye has filled this technological void. By tracking the ball in real-time during matches, Hawk-Eye collects data on the location where and the speed with which the ball hits the court (Gray, 2015, p. 28).

An interesting offshoot of this research is the point-wise analysis of relative importance, i.e. which points were more important than others. Disregarding the gameplay of a deuce, there are sixteen different possible combinations because each player has an opportunity to win 4 points (denoted by: 0, 15, 30, and 40). Thus, it is not too difficult to understand the importance of each of these points.

A relatively unexplored extension is the game-wise analysis of relative importance, i.e., which games are more important in determining the final outcome of the match. This is more difficult than the point-wise estimates because there are many more different possible combinations of scores on the set level. Suppose the player wins the set with six games; then there are a total of thirty different scores that could have arisen because the winning player wins six games and the losing player wins at most four games. Since there is no upper limit to the number of games in a set, there are in fact more combinations than thirty. Thus, determining the importance of the games to a set or a match is more difficult. However, the benefit to understanding such a concept is that there is greater stability in the models. The macro-perspective offered by game-wise analysis as opposed to the micro-perspective of the point-wise analysis is subject to less measurement error in the players' abilities (both physical and mental) because the gameplay is subject to less variability in a single game as opposed to a single rally (Croucher, 1986).

The advent of Hawk-Eye makes it possible now to understand point-wise and game-wise importance measurements. However, match-fixing obscures the data. To identify the validity of the models, it must be understood whether players violate the Best Efforts Clause. If it is, then the measurements made by Hawk-Eye is skewed by the bias and private motivations of the players. This paper adds to the literature by determining whether players follow the Best Efforts Clause by analyzing their effort in relation with the strategic potential of a particular game.

3 Data Description

The raw data used in this project come from two sources. Because an intensive data-cleaning process transformed the raw data, we refer the readers to the sources themselves for a greater insight into the production and overviews of the data.

3.1 Jeff Sackmann's Point by Point Records

Jeff Sackmann[1] has collected point-by-point records of tennis matches for many years across different tournaments and draws. In this paper, we consider grand-slam matches in ATP tournaments that do not have any tiebreakers to ensure consistency in the rules of the game.

[1] https://github.com/JeffSackmann

For an explanation of Sackmann's data, we refer the reader to the documentation of `pbp_raw.df`, which is in the package `oktennis` (Katta, 2018). As it stands in the original data set, the data are reported in terms of `server1` and `server2`, which we shall sometimes refer to as `player1` and `player2`. The allocation of the two players to these titles is arbitrary, so instead, we shall characterize these players as lagging and leading, which are defined in the Methodology section beginning on page 2.

3.2 ATP's Player Statistics

The Association of Tennis Professionals (ATP) has up-to-date information about the statistical characteristics of every major player who participates in Grand Slam Tournaments (*Official Site of Men's Professional Tennis — ATP World Tour — Tennis*). Below is a description for an abridged set of variables that we include in our analysis. For a complete list of the statistics available from the ATP website, we refer the reader to the documentation of the datasets available in the `oktennis` package, namely `players_stats` (Katta, 2018).

`age`	numerical variable representing age of player
`rank`	numerical variable representing rank of player
`hand`	binary variable that takes on the value of 1 if right-handed or 0 if left-handed
`serv_game_won`	numerical variable representing the percentage of service games won

Table 1: Abdridged Set of Variables in `players_stats`

In all these variables of interest, we consider the difference between the variable with respect to the lagging player and the variable with respect to the leading player. For instance, if the lagging player is 40 years old and the leading player is 22 years old, then the difference is written as 18.

Age serves as an indicator of physical performance and mental aptitude. It is reasonable to expect that older players are not as agile as younger players, but they are more experienced. Data about the ranks of the players also capture this information. ATP-recorded ranks are determined by the successes of the players during the past year. Because players' successes or failures occur in streaks (the success of a player builds on past successes as a morale boost), the ranks of the lagging and leading player is important to consider. We include the difference in `hand` because a game between two players of the same dominant hand follows a different strategy than a game between two players whose dominant hands are not the same. Lastly, the difference in `serv_game_won` intends to capture the skill level of each player when he has the serve. It is clear that the player serving has a big advantage over the returner because the server has the opportunity for obtaining aces, for achieving higher speeds on the ball, and for choosing ball placement without worrying about returning the ball. However, if the `serv_game_won` is similar for the two players, then the service advantage is not so pronounced Hence, we include this difference in our analysis.

3.3 Cleaned Data

The dataset `tennis_data` from the package `oktennis`[2] (Katta, 2018) combines a transformed version of Sackmann's dataset as well as player characteristics from the ATP website. The dimensions of this data are given by 17571 rows and 39 columns, where each row represents a game in which there is a lagging and leading player. The columns represent variables of interest, but due to the

[2]To access this package in R, please use the following command: `devtools::install_github("kattaoa/oktennis")`

many number of variables, we refer the reader to the documentation file of the data set for the full details.

Many of these variables are player-specific in that they only relate to the leading player or the lagging player. To coalesce the two in a uniform way, differences were computed such that the leading player's value is subtracted from the lagging player's value, avoiding the problem of multicollinearity. Below is a table with the resulting variables we consider (apart from the player characteristics described above).

`lagging_serve`	denotes whether lagging player has service
`rally_length`	length of rally
`diff_game_set`	difference in the number of games won by each player in the set

Table 2: Abdridged Set of Variables in `tennis_data`

As discussed above, the server has an advantage over the returner. Thus, we consider `lagging_serve` in our regression. `rally_length` serves as an indicator for the fatigue exerted in a game. It is reasonable to assume that longer rallies occur at the beginning of a match when both players have more energy. However, the more tired they become as the match transpires, the rallies become shorter. While this variable measures the fatigue of both players, `diff_game_set` measures the fatigue of the individual player by the same rationale.

3.4 Class Imbalance and Possible Solutions

A cursory glance at `tennis_data` reveals a problem of class imbalance. Consider the null model, which classifies all the observations in the most commonly occuring class. In this case, the null model will predict that the lagging player will lose every time. Because the percentage of games in our sample for which the lagging player wins is 4.55%, This null model will misclassify 4.55% of the observations in our sample. This measure of accuracy may appear impressive, but it is only because of the imbalance in the number of observations in each class of `lagging_game`. The table below presents the sizes of the samples based on the treatment and control groups to highlight the issue of class imbalance[3].

		STRATEGIC		
		0	1	Total
`lagging_game`	0	15079	1693	16772
	1	764	35	799
	Total	15843	1728	17571

Table 3: Class Imbalance

The easiest solution is to gather more data. This solution will only work if the class imbalance is only a feature of the raw data and not of the population. Another solution would be to inflate or deflate subsamples generated by the two classes of `lagging_player` so that the size and quality of the two classes are more comparable. For instance, we might want to consider matches in which the players are comparable to each other. In particular, we might be interested in similarly ranked players. Lastly, rather than use the error rate to measure the performance of models, it may be worthwhile using other measures such as AUC.

[3]Further analysis of the data reveals that no lagging player wins the set.

4 Statistical Methods

Controlling for the players' statistics, we use a logistic regression (James et al., 2013) to determine the probability that a player wins a game in which he is lagging in order to identify the relationship between the strategic potential of a game, operationalized by the variable `STRATEGIC`, and the lagging player's efforts, operationalized by the variable `lagging_game`.

The results of our full model is presented below in Table 4 on page 7. Interestingly, there are many statistically insignificant coefficients in the model contrary to intuition, which suggests a better model selection. Hence, we perform lasso regression to send some of these coefficients to zero, using cross-validation to select the tuning parameter (0.00109) to minimize the estimated test error rate (0.049). All this was done with the `stats::glm()` and `stats::glmnet()` functions (Friedman, Hastie, and Tibshirani, 2010). In order to ensure our treatment variable was not sent to zero by the lasso procedure, we perform lasso on all predictors except `STRATEGIC`, and then we add it to the model. Notice that ultimately, `STRATEGIC` is statistically significant, even though including it in the regression resulted in one of the other features becoming statistically insignificant.

5 Results

Running this regression reveals a negative and statistically significant relationship between the response and treatment variable. As discussed above in the Methodology section, this result lends itself to the possibility that the lagging player might be violating the Best Efforts Clause.

Notice that a statistically significant and positive set-wise difference of 4 games witnesses an increased predicted likelihood in the lagging player winning the game as opposed to the statistically insignificant set-wise difference of 5 games. These observations add credibility to the idea that the lagging player would exert more effort when he knows he has a better chance of catching up to the leading player.

There are many complications with this result. First and foremost, the problem of class imbalance poses an issue with model selection. Regardless of our choice in the tuning parameter, including variables in our regression would not behave better than the null model due to the class imbalance. This is best seen in Figure 1 on page 8. Notice how the MSE as a function of the tuning parameter is predominantly constant in a neighborhood around the minimum MSE. The error rate of our model is greater than that of the null model, which further underscores the issues of class imbalanace. Other observations highlighting the issue of class imbalance is the large confidence intervals and the large number of seemingly relevant features being statistically insignificant. Secondly, the effort exerted by the lagging player is perhaps related to the effort exerted by the leading player. It might in fact be the case that the lagging player does exert more effort when the leading player has already won two sets. However, the leading player might exert more effort too because he is close to winning the match. Thus, the outcome of the game from the perspective of the lagging player is not an accurate indicator of his effort. Furthermore, our definition of `STRATEGIC` might be too course to capture any meaningful relationship with the effort of the lagging player.

Future extensions of this paper would benefit from addressing these complications to validate the violation of the Best Efforts Clause. Understanding this will help determine the conditions under which a player might intentionally lose a game or a point, and whether it is a strategic choice for ultimately winning the match or if the lagging player gave up on the match itself.

Table 4: Logistic Regression

	lagging_game	
	Full Model	Lasso Model
	(1)	(2)
Constant	−22.458 (595.701)	−21.979 (227.117)
Lagging Service	2.644*** (0.119)	2.561*** (0.116)
Difference in Number of Games Won in Set == 4	0.325 (643.819)	
Difference in Number of Games Won in Set == 5	17.907 (595.701)	
Rally Length	−0.031** (0.016)	
Difference in Age	0.005 (0.007)	
Different Hand Dominance	0.045 (0.087)	
Difference in Rank	0.0001 (0.0001)	
Difference in Percentage of Service Games Won	0.226 (0.493)	
Difference in Number of Sets Won	−0.146* (0.082)	
Difference in Total Games Won (Low)	0.287 (0.281)	
Difference in Total Games Won (Middle)	0.331 (0.234)	
Difference in Total Games Won (Really High)	0.850 (1.116)	
Difference in Total Games Won (Really Low)	0.058 (0.531)	
Difference in Number of Games Won in Set == 4		17.620 (227.117)
STRATEGIC	−0.716*** (0.212)	−0.574*** (0.180)
Observations	17,571	17,571

Note: *p<0.1; **p<0.05; ***p<0.01
Differences calculated by subtracting lagging by leading players' statistics.

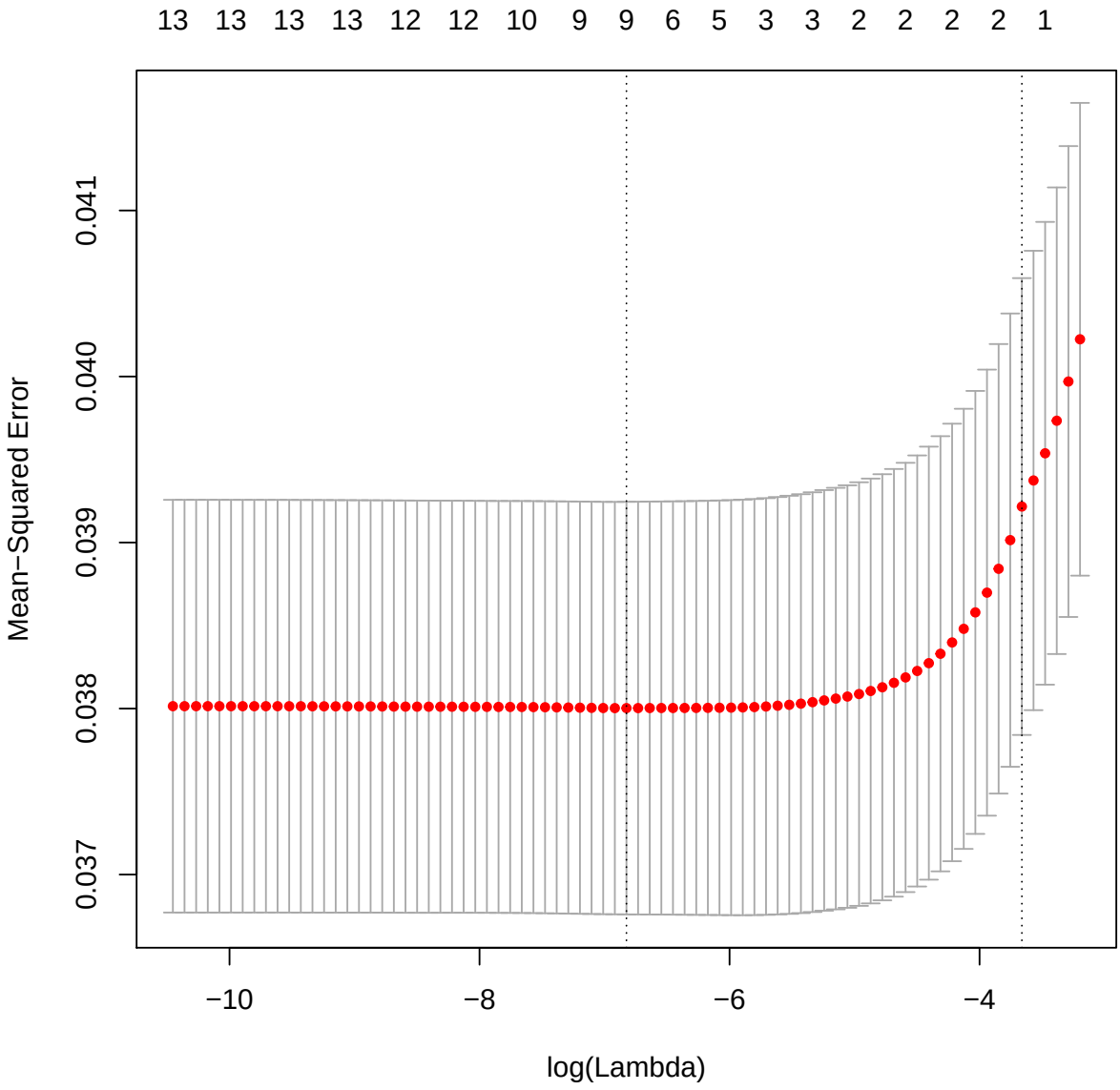

Figure 1: MSE with respect to Tuning Parameter of Lasso Regression

6 Coda: Brief Description of the Structure of Tennis Matches

Tennis is a sport in which two racket-wielding players, who are on either side of a tennis court, hit a ball over the net and do their best to return the ball to their opponent. As simple as this sounds, the structure of a tennis match creates interesting gameplay and strategies that ultimately makes tennis both a physically-taxing and mentally-taxing sport.

A tennis match is divided into sets, which is in turn divided into games. The game is won by the player who is the first to score four points, which are typically measured on the scale 15-30-40-win. The player to win at least six games in a set by a margin of two wins the set. Furthermore, the player who wins the match is the first player to win three sets (the outcome of the match is decided by a best-of-five-sets format). To illustrate the structure of the game, consider a match whose score is given by: (6-3, 5-7, 6-4, 8-6). In this format, each pair of numbers describes the outcome of a set, with the first coordinate describe how many games a player has won. For instance, in set 1, player 1 won 6 games while player 2 won 3 games. Hence, player 1 wins that set. In set 2, player 1 won 5 games, but player 2 won 7 games. Notice that player 2 wins the set because he wins at least six games and wins two more games than his opponent. The following two sets are won by player 1. Therefore, player 1 wins three sets whereas player 2 wins one set. This means that player 1 wins the entire match. Keep in mind that these rules and structures vary across different tournaments. For the purposes of this paper, we will only consider Grand Slam Tournaments, which are fairly consistent in their rules and structure.

References

Angrist, Joshua D and Jörn-Steffen Pischke (2014). *Mastering'metrics: The path from cause to effect*. Princeton University Press.

ATP (2018). *2018 ATP Official Rulebook*.

Clarey, Christopher (2016). *Nick Kyrgios Gives Up, and Tennis Gives Him an Easy Out*. URL: https://www.nytimes.com/2016/10/14/sports/nick-kyrgios-gives-up-and-tennis-gives-him-an-easy-out.html.

Croucher, John S (1986). "The conditional probability of winning games of tennis". In: *Research Quarterly for Exercise and Sport* 57.1, pp. 23–26.

Dowle, Matt and Arun Srinivasan (2018). *data.table: Extension of 'data.frame'*. R package version 1.11.0. URL: https://CRAN.R-project.org/package=data.table.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1, pp. 1–22. URL: http://www.jstatsoft.org/v33/i01/.

Gray, Chris (2015). "Game, set and stats". In: *Significance* 12.1, pp. 28–31.

Henry, Lionel and Hadley Wickham (2017). *purrr: Functional Programming Tools*. R package version 0.2.4. URL: https://CRAN.R-project.org/package=purrr.

Hlavac, Marek (2018). *stargazer: Well-Formatted Regression and Summary Statistics Tables*. R package version 5.2.2. Central European Labour Studies Institute (CELSI). Bratislava, Slovakia. URL: https://CRAN.R-project.org/package=stargazer.

James, Gareth et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Katta, Omkar (2018). *oktennis: Tennis Datasets*. R package version 0.0.0.9000.

Leeper, Thomas J. (2018). *margins: Marginal Effects for Model Objects*. R package version 0.3.23.

Mller, Kirill and Hadley Wickham (2018). *tibble: Simple Data Frames*. R package version 1.4.2. URL: https://CRAN.R-project.org/package=tibble.

Newton, Paul K and Joseph B Keller (2005). "Probability of winning at tennis I. Theory and data". In: *Studies in applied Mathematics* 114.3, pp. 241–269.

O'donoghue, Peter G (2001). "The most important points in grand slam singles tennis". In: *Research quarterly for exercise and sport* 72.2, pp. 125–131.

Official Site of Men's Professional Tennis — ATP World Tour — Tennis. URL: https://www.atpworldtour.com/.

Riddle, Lawrence H (1988). "Probability models for tennis scoring systems". In: *Applied Statistics*, pp. 63–75.

Varshney, K (2014). "An analysis of losing unimportant points in tennis". In: *KDD Workshop on LargeScale Sports Analytics*.

Walker, Alexander (2018). *openxlsx: Read, Write and Edit XLSX Files*. R package version 4.1.0. URL: https://CRAN.R-project.org/package=openxlsx.

Wickham, Hadley (2018a). *forcats: Tools for Working with Categorical Variables (Factors)*. R package version 0.3.0. URL: https://CRAN.R-project.org/package=forcats.

— (2015). *R packages: organize, test, document, and share your code*. " O'Reilly Media, Inc.".

— (2016). *rvest: Easily Harvest (Scrape) Web Pages*. R package version 0.3.2. URL: https://CRAN.R-project.org/package=rvest.

— (2018b). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.3.0. URL: https://CRAN.R-project.org/package=stringr.

— (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1. URL: https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley and Lionel Henry (2018). *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.8.1. URL: https://CRAN.R-project.org/package=tidyr.

Wickham, Hadley, James Hester, and Jeroen Ooms (2018). *xml2: Parse XML*. R package version 1.2.0. URL: https://CRAN.R-project.org/package=xml2.

Wickham, Hadley, Jim Hester, and Romain Francois (2017). *readr: Read Rectangular Text Data*. R package version 1.1.1. URL: https://CRAN.R-project.org/package=readr.

Wickham, Hadley et al. (2018a). *dplyr: A Grammar of Data Manipulation*. R package version 0.7.5. URL: https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley et al. (2018b). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.0.0. URL: https://CRAN.R-project.org/package=ggplot2.

Xie, Yihui (2018). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.20. URL: https://CRAN.R-project.org/package=knitr.