Original Investigation

DOI: 10.30827/

Predicting Table Tennis Tournaments: A comparison of statistical modelling techniques

Predecir torneos de tenis de mesa: una comparación de técnicas de modelización estadística

Jan Lennartz [1]; Andreas Groll [1]; Hendrik van der Wurp [1]

1 TU Dortmund University, Department of Statistics, Dortmund, Germany.

Received: 04-07-2021

Accepted: 15-02-2022

Abstract

There are two main goals of this work: 1) to compare different statistical models, which are applied to historic tournaments to find a suitable statistical model, i.e. the model with the best predictive performance, and 2) to understand which factors are important for good predictions. Every year at least one of four important recurring table tennis tournaments takes place where top players compete. Those tournaments are the World Table Tennis Championships, the Table Tennis World Cup, the Olympic Games and the ITTF World Tour. In other areas of sports, it is common to analyse major tournaments and predict future ones. This work aims to bring this aspect of analysis to the world of table tennis by evaluating recent holdings of the Men's World Cup and the Grand Finals of the Men's ITTF World Tour. The results show that it is indeed possible to apply statistical machine learning methods on table tennis tournaments for prediction with a correct classification rate of around 75% by a random forest and 74% by a penalized generalized linear logit model. Even though both models based their predictive power mainly on the official table tennis rankings and points, variables like age, playing handedness or individual strength were important factors as well.

Keywords: *Tournament analysis, random forest, statistical learning, table tennis, LASSO regression.*

Resumen

Este trabajo tiene dos objetivos principales: 1) comparar los diferentes modelos estadísticos que se aplican a torneos históricos para encontrar un modelo estadístico adecuado, es decir, el modelo con el mejor rendimiento predictivo, y 2) entender cuáles factores son importantes para una buena predicción. Cada año se celebra al menos uno de los cuatro torneos importantes y recurrentes de tenis de mesa en los que compiten los mejores jugadores. Esos torneos son el Campeonato Mundial de Tenis de Mesa, la Copa del Mundo de Tenis de Mesa, los Juegos Olímpicos, y el Circuito Mundial de Tenis de Mesa. En otras áreas del deporte, es común analizar torneos importantes y predecir los futuros. Este trabajo pretende traer ese aspecto del análisis al mundo del tenis de mesa al evaluar las competencias recientes en la Copa del Mundo y las Grandes Finales del Circuito Mundial, ambas en la categoría masculina. Los resultados demuestran que es posible aplicar métodos estadísticos de aprendizaje automático a los torneos de tenis de mesa para predecir con una tasa de clasificación correcta de alrededor del 75% a través de un bosque aleatorio y del 74% con un modelo logit lineal generalizado penalizado. Aunque ambos modelos basan su poder predictivo principalmente en las clasificaciones oficiales de tenis de mesa y puntos, las variables como la edad, la destreza en el juego o la fuerza individual también fueron factores importantes.

Palabras clave: *análisis de torneo, bosque aleatorio, aprendizaje estadístico, tenis de mesa, regresión LASSO.*

Corresponding author: Jan Lennartz, jan.lennartz@tu-dortmund.de

Cite this article as:
Lennartz, J., Groll, A., & van der Wurp, H. (2021). Predicting Table Tennis Tournaments: A comparison of statistical modelling techniques. *International Journal of Racket Sports Science*, *3*(2), 39-48.

INTRODUCTION

International sports tournaments (most of them periodically held) are usually seen as a highlight for the respective sport due to their extensive nature. Thus, they are attracting a bigger audience and gain temporary more attention than the national counterparts. As the tournaments are often scheduled a considerable time in advance, people try to predict potential outcomes and winners. This is reflected in the countless bets at bookmakers shortly before and during those tournaments, as well as in the general media activity.

With the rise of computational power and the increased popularity of machine learning tools over the last two decades, scientific research began to extend and optimize predictions of tournaments in many sports, e.g., football (Groll A. , Ley, Schauberger, & Van Eetvelde, 2019a) and tennis (Gu & Saaty, 2019). There has evolved a competition of different prediction models with lots of competitors. This also comes with an increased understanding of sports matches. More and more data become available, and once they have been analysed for prediction purposes, they can potentially improve the understanding of why a specific team or player won. Beside the general interest of sports fans, the expected continuing growth of the global sports betting market (Grand View Research, 2021) is motivating research in predicting sports results.

However, table tennis is not yet in the focus of this development. This work aims to bring that aspect of analysis to the world of table tennis by conducting recent holdings of the ITTF (International Table Tennis Federation) Men's World Cup and the Grand Finals of the Men's ITTF World Tour. There are two main goals: 1) to compare different statistical models based on historic tournaments to find the model with the best predictive performance and 2) to understand which factors are important for good predictions in table tennis. The best model can then be used to predict future tournaments. The results should give a first impression of how suitable statistical models are in the context of modelling table tennis. With the expected growth of sports betting, also table tennis bets will most likely gain more popularity. Hence, modelling table tennis matches could become a critical aspect for the betting industry.

The second main goal is first of all simply inspired by scientific curiosity. Moreover, the understanding of the models and the important features indicate what kind of data is most interesting in the context of table tennis prediction. And therefore, the most important features can hint at what sort of data could potentially improve the results.

Since this is the first time that these methods are applied in the area of international table tennis, it is also of interest of how decent the methods work in this context. We answer the question whether results of research with respect to predictions of other sports can be applied in table tennis as well. However, the authors want to state that this work only represents a first modest overview of potential statistical and machine learning models. Furthermore, the selected models were not considerably tuned. Finally, we make use of the best predictive model and analyse what it would have predicted for the 2019 ITTF Men's World Cup.

MATERIAL AND METHODS

Data

For the complete analysis the programming language R was used (R Core Team, 2019). The data set was collected via web scraping from the official web archive (ITTF Archive, 2019). The obtained data was then pre-processed, including manual updating of missing values where the information could be found elsewhere. The final full data set includes 419 matches taken from all World Cups and Grand Finals of the ITTF World Tour held between the years 2010 and 2018, see Table 1. In this manuscript, players' names are encoded as FL (F = first name, L = last name). Individual player statistics are included for each match. Additionally, there were features generated based on this data. Exemplary features are host (whether the tournament is taking part in the players' home country), age, handedness (left or right-handedness), style (attacker or defensive), grip (shake-hand grip or penhold), WTTR-position (WTTR=World Table Tennis Ranking), WTTR-points. The handedness of players was established according to which hand was used to hold the racket (Peters & Murphey, 1992).

Each row in the data set represents a match where two players (player *A* vs. player *B*) compete. For better visibility the transposed version is shown in Table 2, where each column represents a match. The result for a match can either be 1 (player *A* won) or 0 (player *A* lost, i.e., player *B* won). The indications as player *A* and *B* are assigned randomly in the beginning. The covariates then represent the differences between player *A* and player *B*. For example, the WTTR-points of player *B* are subtracted from the WTTR-points of player *A*. The variable age is corrected by the average age (*avgAge*), which is assumed to be the optimal value: *age* = | *age(B)* – *avgAge* | - | *age(A)* - *avgAge* |. This way, it takes the value 0 when e.g., player *A* is 3 years younger than the average age and player *B* is 3 years older than the average age, i.e., both have the same distance to the optimal age and are assumed to have identical age benefits. For the variable *handedness*, the encoding is 1 if player *A* is right-handed and player *B* is left-handed, -1 if player *A* is left-handed and player *B* is right-handed and 0 if both players play with the same hand. Similarly, the encoding is done for the other variables.

Furthermore, there were several dummy variables defined: A group of dummy variables is representing the continent, where players origin from. It is encoded pairwise, i.e., all possible two-pair-combinations of continents are present, and only the relevant combination for the specific match is set to 1, e.g., *EU_AF* (*Europe/Africa*), while the other combinations remain 0. A unique ID for each player was used that could feature as a strength variable where players that participated for the first time in a tournament were given a rookie ID (999999) for that tournament. For this player strength variable, all available IDs of every player that participated in the conducted tournaments are considered as dummy variables. Then, only the specific IDs for a match are set to 1 (player A) and -1 (player B). Hence, the model can identify all matches of a specific player and evaluate the player's performance. However, this is evaluated over all available matches without considering their chronological order (see also the paragraph *Cross Validation*).

Table 1.
Data set overview. In total there are 419 matches that have been used of which the majority is coming from the World Cups.

Year	World Cup Matches	Grand Final Matches	Total
2010	38	15	53
2011	38	15	53
2012	36	15	51
2013	32	15	47
2014	28	15	43
2015	28	15	43
2016	28	15	43
2017	28	15	43
2018	28	15	43
Total	**284**	**135**	**419**

Table 2.
Overview of the used data set. Note: Shown is artificial data to give an impression of the value ranges. Each column represents a match and each row a variable.

Variable Name	*Observation 1*	*Observation 2*	*Observation 3*	...
Winner_is_A	1	0	1	...
Year	2011	2012	2013	...
A_id	999999	123456	324544	...
B_id	123456	452364	999999	...
Host	0	1	0	...
Age	-3.06	4.15	-2.45	...
Hand	1	-1	0	...
Style	0	1	-1	...
Grip	0	1	-1	...
WTTR-position	-15	-66	13	...
WTTR-points	56	155	-50	...
AF_AS	0	-1	0	...
AF_EU	0	0	0	...
⋮	⋮	⋮	⋮	...
EU_LA	1	0	0	...
A_id123456	-1	1	0	...
⋮	⋮	⋮	⋮	...
A_id999999	1	0	-1	...

Statistical Models

The models of choice were a **LASSO:** *Least Absolute Shrinkage and Selection Operator* (Friedman, Hastie, & Tibshirani, 2010; Tibshirani, 1996), and a random forest (Breiman L. , 2001). The LASSO is a penalized version of a generalized linear model (Fahrmeir & Tutz, 2001; McCullagh & Nelder, 1989) in particular of a logit model. It yields a predicted win probability π_i for match i and a coefficient vector $\boldsymbol{\beta}$ corresponding to covariate effects which is well suited for interpretation purposes. The chosen logit model has the form:

$$\pi_i = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}.$$

Here, $\boldsymbol{x}_i$ is the vector of covariates for match *i*.

The fitted model is then the solution to the minimization problem

$$\hat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}}\ \mathcal{L} + \lambda \sum_{j=1}^{k} |\beta_j|,$$

where $\mathcal{L}$ is the Bernoulli likelihood (Friedman, Hastie, & Tibshirani, 2010) to which a penalty term is added, whose strength is controlled by a penalization parameter $\lambda \in [0, \infty)$. The penalty term itself is the sum over all j = 1, 2, ..., k regression coefficients.

The chosen L1-penalization allows to potentially shrink estimated coefficients to zero which effectively results in a variable selection process. Non-significant variables implicitly get removed, which yields a more stable model. The R-package `glmnet` (Friedman, Hastie, & Tibshirani, 2010) was used to fit the penalized generalized linear logit model. To find the best penalization parameter λ the implementation utilizes a cross validation via the `cv.glmnet` function.

However, the linear nature of the LASSO comes with limitations when it comes to modelling data of unknown shape. The **random forest** model (Breiman L. , 2001) is very flexible and efficient in modelling any input data. It has been shown (Schauberger & Groll, 2018) that random forests work considerably well for predicting sport results.

A random forest is based on the idea of decision trees (Breiman, Friedman, Stone, & Olshen, 1984; Theodoridis, 2015). The basic concept of a decision tree is to split the data set into chunks based on a properly chosen splitting variable. This is done subsequentially until a stopping criterion is met. The result is a tree-like chunking of the data, where the very bottom (*the leaves*) corresponds to a specifically characterized data chunk. Figure 1 shows an exemplary simplified decision tree based on solely the variables *WTTR-position* and *age*. The algorithm chooses a suitable splitting variable at each step and splits the data accordingly. The predicted probabilities then correspond to the relative frequency of won matches in this branch. This procedure can continue until a perfect separation is obtained, i.e., each observation has its own unique path in the tree. Usually this process is restricted, e.g., by pruning the tree after completion to avoid overfitting. Once the decision tree is generated, the prediction step consists of a simple evaluation of the given covariates on the tree.

A random forest is utilizing many decision trees and introduces randomization steps to decorrelate the single trees and, hence, lower the variance. The resulting ensemble of decision trees is then used with a majority voting to make predictions. This work makes use of the R-package `randomForest` (Liaw & Wiener, 2002). Even though the random forest lacks interpretability one can look at the so-called **variable importance** (Liaw & Wiener, 2002; Breiman L. , 2001) to get a rough impression of the decision process. The process of calculating the variable importance can be described exemplarily as follows: if we want to calculate the variable importance of the variable *age*, we modify the data set by permuting all age values randomly across the data set. Then, the prediction error is calculated based on this altered data set. Finally, the prediction error is compared to the original prediction error (on the non-permuted data set) with the use of the Gini-Index (Ceriani Lidia, 2012). If the error on the permutated data set substantially increased compared to the original data set, the variable importance for our variable *age* will be high. This is then done for each variable separately.

The third model is a **reference model** which solely predicts based on the current rank of the players. The predicted win probability will be 1 if the player *A* has a higher rank than player *B* and 0 otherwise.

Performance Measures

Each model yields a probability for a win of player A denoted by $\hat{\pi}$ for a given match. The true outcome is always denoted by y, which can take the values 0 (player A lost) or 1 (player A won). Based on these prediction probabilities, the performance will be measured with four different approaches.

The **classification rate** represents the proportion of correctly classified matches and is frequently used in classification problems, also in the field of sports see e.g., (Schauberger & Groll, 2018):

$$\mathcal{K} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i = \hat{y}_i), \text{where } \hat{y}_i = \begin{cases} 1, \text{ if } \hat{\pi}_i > 0.5 \\ 0, \text{ if } \hat{\pi}_i \leq 0.5 \end{cases}.$$

A measure that is capturing more information on how accurate the model predictions are is the **Bernoulli likelihood**. It represents the mean probability for the correct prediction see e.g., (Schauberger & Groll, 2018):

$$\mathcal{L} := \frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_i^*, \text{where } \hat{\pi}_i^* = \begin{cases} \hat{\pi}_i, \text{ if } y_i = 1 \\ 1-\hat{\pi}_i, \text{ if } y_i = 0 \end{cases}$$

The third measure is the **Brier Score** (Brier, 1950). This time the mean is formed over the quadratic difference between the predicted probability $\hat{\pi}$ and the actual outcome y:

$$\mathcal{B} := \frac{1}{n}\sum_{i=1}^{n}(\hat{\pi}_i - y_i)^2$$

Finally, the fourth measure is the **area under the curve (AUC)** (Fawcett, 2006; Robin, 2021). It is measured over the receiver operating characteristics (ROC) curve where the true positive rate is plotted against the false positive rate.

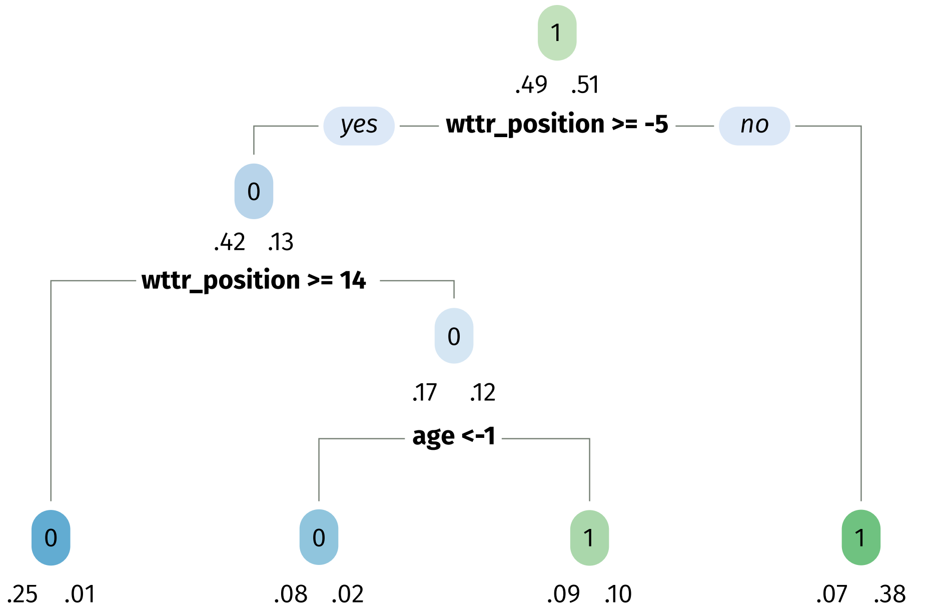

Figure 1. Overview over a simplified decision tree. Recall that the variables wttr_position (WTTR-position) and age are representing differences between the two involved players. It is visible how the probabilities for a win (1) get refined after each split.

All four measures are in a range of 0 to 1. The first two measures and the AUC are goodness-of-fit measures. Therefore, they are desired to be close to 1. The Brier Score, however, is measuring the error and thus it is desired to be close to 0.

To evaluate the performance of the prediction of the ITTF World Cup 2019 the *Tournament Rank Probability Score* (**TRPS**) (Ekstrom, Van Eetvelde, Ley, & Brefeld, 2021) is used. This score takes the whole course of tournament into account by evaluating the predictions and results on every tournament stage, i.e., group stage, round of 16 etc. The TRPS for a full tournament prediction X and the actual outcome O is defined as

$$TRPS(O,X) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{R-1}\sum_{r=1}^{R-1}(o_{rt}-x_{rt})^2.$$

Here, T represents the number of teams (or players in the present case), R is the worst possible rank and x_{rt} is the cumulative probability that player t reaches at least rank r. The lower the TRPS the better is the prediction, where the ideal prediction has a TRPS of 0. To interpret the TRPS it is advised by the authors to compare the results to a TRPS of a flat prediction, i.e., where each player has equal chances.

Cross Validation

The models are trained, and the errors are calculated via a cross-validation (CV)-type approach. This allows to train and test on the whole data set. The data set is split into a train fold of 7 years and a test fold of 1 year. This is done on all possible combinations of years. The chronological sorting is ignored, and it is assumed that the data is independent, conditioned on the covariates. The prediction on the test fold, i.e., the predicted probability $\hat{\pi}$ of a win for player A, is stored for each fold. Then, the mentioned performance measures are evaluated on the stored $\hat{\pi}$ values.

Analysis

The predictive performance analysis was based on the results of the CV and the performance measures introduced above. Furthermore, both statistical models were trained on the full data set to leverage all available information for the in-depth analysis of important factors. The coefficients of the LASSO as well as the variable importance of the random forest were then interpreted. Finally, the best performing model was used to simulate the (out of sample) ITTF World Cup 2019 and predict the outcome of the tournament.

RESULTS

Performance Comparison

The CV results with respect to different performance measures are shown in Table 3. The last row in Table 3 shows the results of a rank-based model for reference. It solely predicts based on the rank of both participating players. The player with the higher (better) rank is then assigned a 100% winning chance, i.e., if the higher ranked player is player A the prediction is 1, and 0 otherwise.

The random forest outperforms the LASSO approach with respect to every measure that was applied. However, the differences to the LASSO are rather small depending on the considered measure. The classification rate and the AUC are both similar for the different models. This might be due to the rather rough rounding nature of these two measures (probabilities are rounded to 0 or 1). For the Bernoulli likelihood and especially the Brier score, the difference is slightly more remarkable. If one compares the two statistical models (random forest and LASSO) to the ranked-based reference model it is notable that it scores better than both considered statistical models in terms of classification rate and Bernoulli likelihood. However, the Brier score and the AUC show that the random forest is on average making smaller errors in the prediction. Due to the simple prediction method of the rank-based reference model it is always making very confident predictions, which result in very high errors when the prediction is not correct. In comparison to this model, both statistical models are showing a good prediction performance. Due to the lack of comparable work in table tennis competitions it is hard to tell exactly how accurate the prediction is. In a similar work for tennis matches (Brunner & Groll, 2018) a classification rate of 0.79, a likelihood of 0.69 and a Brier score of 0.15 was achieved.

Table 3.
Results of the cross validation. Classification Rate, Bernoulli-Likelihood-Score, Brier Score and area under the curve. Models are LASSO, random forest and the ranking based reference model. Best values in bold typeset.

	CLASS	BLH	BRIER	AUC
LASSO	0.7375	0.6583	0.1844	0.8036
RF	0.7542	0.6842	**0.1781**	**0.8095**
RB	**0.7828**	**0.7828**	0.2172	0.7828

Alternative Models

Even though both statistical models from above incorporate some form of variable selection, due to the large number of covariates the quality of this selection process can not be completely guaranteed. Some variables might still show predictive power just by chance which will not yield a satisfying prediction on unseen data. In particular, the player-specific strength variable which identifies each player individually accounts for 62 variables (players) as a result of the dummy encoding. Hence, the sheer number of variables of this type could be a reason for the models to select some of them. Thus, an alternative version for both models was run with the same CV approach, where both models (random forest and LASSO) were not fed with the player's strength variable. The results are displayed in Table 4. It turns out that the runs without the players' strength variable are even slightly better for both models w.r.t. almost all performance measures. This could be indicating that there are not enough matches for each player to form a reliable strength variable. However, as we are more interested in finding certain player-specific patterns, particularly strong under- or over-performers, than in pure predictive performance, these alternative versions were not studied any further.

Table 4.
Results of cross validation for alternative models without the player's strength variable. Labels are identical to Table 3.

	CLASS	BLH	BRIER	AUC
LASSO	0.7613	0.6583	**0.1656**	**0.8324**
RF	0.7733	0.6871	0.1768	0.8072
RB	**0.7828**	**0.7828**	0.2172	0.7828

Model Interpretation

To interpret the models, they are fitted on the whole data set available. For the random forest, the variable importance is considered for interpretation, while for the LASSO approach the coefficients can be interpreted directly.

The variable importance of the random forest is shown in Figure 2. It turns out that the WTTR-points and the WTTR-position are by far the most important variables. This was to be expected as these two variables contain a lot of (similar) information and are naturally highly correlated. The higher variable importance of the WTTR-points compared to the WTTR-position could be explained by the finer scale of the WTTR-points. The third important variable is the players' age, which suggests that it contains suitable predictive power for the random forest model. The remaining variables have substantially less importance. It is notable though that out of all the available variables, the model selects especially the players' handedness, grip, and the host variable. Additionally, the matches where players from Asia play against European players and matches where Europeans play against Latin Americans show a special character. However, it is not possible to tell from the variable importance whether players from the respective continents perform below average or above average. A look at the specific player IDs reveals that the rookie variable is also considered important by the model. Therefore, there can be a considerable change in the prediction when a new player is part of the match, in contrast to when the same player has played already in a tournament before. The player whose player-specific ability ("strength variable", e.g. A_id108246) has the highest variable importance is VS (BLR). Again, due to the limitation of the measure it is not possible to tell if the player is equipped by the model with a bonus (i.e., performing above average) or with a malus (i.e., performing below average). Nevertheless, it seems that with respect to the other variables this player is standing out when it comes to predicting a match. Similarly, one can observe how much weight the model gives to the other players as seen in Figure 2. The interpretation of this, however, is not to be mistaken with a necessarily good performance of the player. The variable importance represents rather the difficulty of predicting matches when this information is missing. For example, if we would remove the information of whether VS is participating in a match or not (by randomly permuting its values), the predictive power of the model would significantly decrease. In contrast to that it would not change much if we were about to remove the information of whether BS participates in a game or not.

A more in-depth interpretation can be achieved based on the LASSO coefficients. In Table 5 the output of the full LASSO model is displayed. Noticeable is the small number of selected variables. In fact, only the WTTR-points and six player IDs have been selected by the LASSO model. A simple indicator for interpretation purposes is the sign of the coefficients. Recall: The exemplary variable A_id106884 takes the value 1 if player 106884 is player A (all matches are seen from the perspective of player A, the win probability is the probability of a win for player A). The variable A_id106884 takes the value -1 if player 106884 is the opponent (player B). If player 106884 is not part of the match the variable takes the value 0.

For the general interpretation, a positive sign for the coefficient of the players' strength variable results in a higher probability for a win of that player. Likewise, a negative sign lowers the probability of a win for this player.

Table 5 shows that the model assigns three players with a negative sign (SO, CM and AS) and two players are equipped with a positive sign (MK and MM). Because the only other variables considered are the WTTR-points, the resulting LASSO model is rather simple and easy to interpret. Since the WTTR-points have a positive sign, whenever player A has more points than player B, the estimated probability for a win is higher for player A. Roughly speaking the higher the WTTR-points difference is, the higher the win probability. The only exception occurs when one (or multiple) player(s) from the five former mentioned participate(s). Depending on the player, the model would give the player a bonus or malus. This essentially results in a correction of the WTTR-points. Those five players' performance – according to the LASSO model – deviated so far from the expected performance based on the WTTR-points that this was corrected by the model.

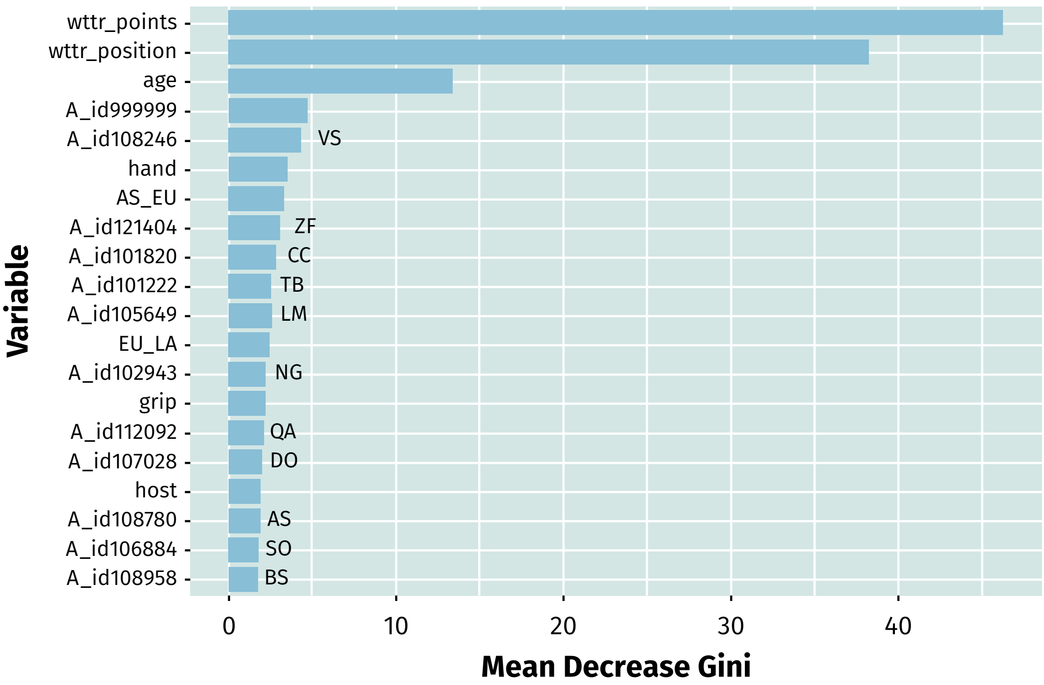

Figure 2. Variable Importance for the Random Forest Model, showing only the top 20 Variables.

Table 5.
Regression coefficients for the selected variables by the LASSO model.

Variable	Coefficient (β)	Player Name
Intercept	0.10952	-
A_id106884	-0.30410	SO
A_id105928	-0.25677	CM
A_id108780	-0.03328	AS
A_id111791	0.02068	MK
A_id105966	0.22670	MM
wttr_points	1.48755	-

PREDICTING THE WORLD CUP 2019

Based on the complete data available the random forest model was superior to the LASSO and was used to simulate the Men's World Cup 2019. Since the drawing of the groups is placed closely to the start of the tournament, the group draws were also simulated. The players YL (TPE), DH (AUT) and SG (IND) were participating in a World Cup for the first time and thus, they were equipped with the rookie variable (ID 999999). The simulation was programmed to follow the official ITTF rules (World Cup Playing System, 2019). The best 8 players (based on their WTTR-points) are bypassing the groups. Thus, only 12 players take part in the group phase. The simulation was made based on 1 million tournament courses. Each match is simulated by taking the predicted winning probabilities returned from the random forest and drawing a Bernoulli random variable with those probabilities.

The results are shown in Table 6. The model is favouring the Chinese top players ZF and LM. The third potential winner is the Japanese player TH. However, with regard to the winning probability he is far behind the former two.

The evaluation with the TRPS is done in the following way: In the simulation no match for place three was considered and thus, the six potential ranking positions were:

1. First Place
2. Second Place
3. Semi-Final
4. Round of 8
5. Round of 16
6. Preliminary Stage (Group Phase)

The TRPS was evaluated for the random forest prediction, for a flat prediction (each participant has equal chances to reach any rank position, i.e., 1/6 for each rank) and for a rank-based prediction (the predicted rank corresponds to the WTTR-Rank, i.e., probability 1 for the corresponding rank and 0 elsewhere). To be fair, the flat prediction is corrected for the first eight players which will always reach the round of 16. Thus, the predictions for those eight players are set to 0 for rank 6 and 1/5 for the other ranks. The results are shown in Table 7. The random forest has the lowest TRPS, almost 3 times smaller than the flat TRPS. The rank-based prediction is closer to the random forest, but the TRPS shows that the random forest is better suited for a whole tournament prediction.

Table 6.
Probabilities for reaching different stages at the 2019 ITTF Men's World Cup for each participant based on the random forest model.

	Player	WTTR-Rank	Round of 16	Round of 8	Semi Final	Final	World Champion	True Position
🇨🇳	ZF	2	100.0000	88.6582	74.8555	62.8887	44.3448	1
🇨🇳	LM	3	100.0000	88.0108	80.2090	67.3938	35.4300	4
🇯🇵	TH	5	100.0000	86.0425	71.5843	23.8240	9.2473	2
🇧🇷	HC	6	100.0000	70.3139	50.5881	14.9690	4.1596	5
🇩🇪	TB	7	100.0000	80.5774	29.6099	11.3912	3.6452	5
🇸🇪	MF	9	100.0000	86.7482	23.8760	6.3742	1.1941	9
🇹🇼	YL	10	100.0000	79.4553	19.4437	4.0668	0.6073	3
🇯🇵	KN	11	100.0000	84.0828	20.7018	3.8581	0.5359	5
🇧🇾	VS	21	91.8613	23.0979	5.8932	1.4128	0.2546	P
🇸🇪	KK	24	94.2586	22.0665	4.9424	1.0431	0.2172	9
🇭🇰	CW	16	95.2743	16.7852	4.2048	0.7404	0.1093	-*
🇩🇪	DO	12	96.7574	18.9190	4.5374	0.7691	0.0984	5
🇰🇷	SL	17	99.0416	15.2372	3.3489	0.5156	0.0648	9
🇫🇷	SG	20	98.3675	14.3511	3.2064	0.4210	0.0537	9
🇩🇰	JG	25	93.6055	15.3979	2.1152	0.2665	0.0325	P
🇮🇳	SG	30	20.0366	1.9689	0.1978	0.0190	0.0018	9
🇪🇬	OA	46	10.2287	1.2564	0.1237	0.0127	0.0015	P
🇺🇸	KJ	27	87.1308	5.6647	0.4403	0.0264	0.0012	9
🇦🇹	DH	40	6.9614	0.8376	0.0797	0.0057	0.0006	9
🇦🇺	HH	66	6.4763	0.5285	0.0419	0.0019	0.0002	P

Note: The true position represents the final ranking after the tournament, where 5 = round of 8, 9 = round of 16, and P = preliminary stage.
*CW was injured before the tournament and replaced by QA, who reached the round of 16.

Table 7.
Results of Tournament Rank Probability Score for the prediction of the world cup 2019. RF is the random forest, Flat refers to the flat prediction (equal chances for everyone) and RB is the rank-based reference prediction (higher ranked player wins).

	RF	***Flat***	***RB***
TRPS	**0.5478**	1.4338	0.7158

DISCUSSION

The two models compared in this work were chosen because of their previous performance in other sports (Groll, Schauberger, & Tutz, 2015; Groll A. , Ley, Schauberger, & Van Eetvelde, 2019a; Groll A. , Ley, Schauberger, Van Eetvelde, & Zeileis, 2019b; Groll, Heiner, Schauberger, & Uhrmeister, 2020). However, there is an abundance of modelling techniques that could have been used. Thus, this work can only give a first impression of how well a statistical model can perform.

In fact, as part of the initial research for this work other models were considered as well. A standard generalized linear model (Fahrmeir & Tutz, 2001) as well as a decision tree (Breiman, Friedman, Stone, & Olshen, 1984) were utilized. Since they belong to the same class of models like the LASSO and the random forest, respectively, only the latter – better performing – variants are shown here.

Additionally, instead of predicting the binary outcome of a match (win or loss), it was also considered predicting the difference in sets. This would allow to involve the notion of high wins or close matches, respectively. However, this approach yields a non-binary classification. The response variable (the difference in sets) is ordinal valued with results from the set {4, 3, 2, 1, -1, -2, -3, -4}. The results of this approach were all outperformed by the binary approach shown in this work and thus, not included here.

For comparing the statistical models to a reference, it is common to look at the bookmarker scores. Since in contrast to highly popular sports like football or tennis, for table tennis those were not freely accessible, a simple rank-based reference model was utilized. Alternatively, one could create a reference model that makes predictions based on the history of the two players and predicts according to the relative win frequency, e.g. with a history of 5-2 wins and losses for player A, the prediction would be 5/7. This would potentially result in probabilities that are not always 1 or 0, which would make it better to compare with the statistical model predictions. However, this approach was not suitable for the given data set, since it contained only very few match constellations that appeared frequently enough.

Both of the used R-functions (`glmnet` and `randomforest`) support multiple hyperparameters that can be tuned. This was not the focus of this work and thus, the standard settings have been used. For the two models it is expected that these are sufficient to compare the models appropriately, however, particularly the random forest's performance might further improve by performing a sophisticated and extensive tuning.

Data

In terms of the available data, one must state that the dataset in use is not very exhaustive. Only 419 matches are available. In future research, potentially other matches of tournaments like the Olympics or the World Table Tennis Championships could be incorporated as well. Here, it was avoided due to their slightly different tournament structure. Similarly, the available covariates are limited because of the lack of more detailed data. In other sports (e.g., football) there is a growing data pool of statistics. The amount of publicly available data in table tennis is rather small though. For example, it was not possible to get any data about betting odds for table tennis matches. This is unfortunate as these incorporate a considerable amount of information and are often regarded as a benchmark for predictions (Groll A. , Ley, Schauberger, & Van Eetvelde, 2019a). Nevertheless, more data about the players would already be of use for a better predictive performance. This data could include statistics like e.g., games played for the national team, participation in high level tournaments, number of attacks per set, points scored after serves or average match length for a specific player.

However, regarding the covariates that are present in the dataset, one can state that the most relevant information is included in the WTTR-points and WTTR-position. However, the results showed that the age also carries some predictive information. Furthermore, the models made use of individual player scores which act like a correction of the WTTR-points.

CONCLUSION

This work aimed to give a first overview over the predictive performance of two well-known modelling techniques for predicting table tennis matches. The linear regression approach, incorporated by the LASSO, yielded solid results, and allowed for a detailed interpretation of covariates. The random forest on the other hand, performed slightly better with respect to prediction and allowed also for an insight in the important covariates. Both models based their predictive power on the WTTR-points and WTTR-position which essentially represent the same information. And both models showed that other covariates like age and individual scores were also of importance. Out of the other available covariates, especially the handedness, grip and two continental combinations showed the most impact within the model.

Alternative versions of both models where the player-specific strength variable was removed yield slightly better results in terms of predictive performance. This renders our findings regarding over- and under performers inconclusive, albeit interesting. On a larger sample size, i.e. with more matches per player this variable could potentially become more reliable.

This work shows, that generally, it is possible to exploit available data for predicting table tennis matches. In the present case, the difference between the well interpretable LASSO and the more sophisticated random forest is not substantial. This allows for a detailed insight into the linear model (LASSO) without losing too much predictive power. The random forest shows a good performance in predicting table tennis matches and was able to predict the World Cup 2019 better than the rank-based reference with respect to the TRPS. This gives confidence that statistical methods for predicting table tennis matches have high potential. However, this potential will highly depend on the availability of more data regarding the table tennis sport.

REFERENCES

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Boca Raton, Florida: CRC Press.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review, 78*(1), 1-3.

Brunner, S., & Groll, A. (2018). *Modellierung und Vorhersage von Tennisspielen bei Grand Slam Turnieren*. Dortmund

Ceriani Lidia, P. V. (2012). The origins of the Gini index: extracts from Variabilitá e Mutabilitá (1912) by Corrado Gini. *The Journal of Economic Inequality, 10*(3), 421-443. https://doi.org/10.1007/s10888-011-9188-x

Ekstrøm, C. T., Van Eetvelde, H., Ley, C., & Brefeld, U. (2021). Evaluating one-shot tournament predictions. *Journal of Sports Analytics, 7*(1), 37-46. https://doi.org/10.3233/JSA-200454

Fahrmeir, L., & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2nd ed.). New York: Springer.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software, 33*(1), 1-22. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/

Grand View Research. (2021). Sports Betting Market Size, Share & Trends Analysis Report By Platform (Online, Offline), By Type (Fixed Odds Wagering, eSports Betting), By Sports Type (Football, Basketball), By Region, And Segment Forecasts, 2021 - 2028. Grand View Research. https://www.grandviewresearch.com/industry-analysis/sports-betting-market-report

Groll, A., Heiner, J., Schauberger, G., & Uhrmeister, J. (2020). Prediction of the 2019 IHF World Men's Handball Championship–A sparse Gaussian approximation model. *Journal of Sports Analytics (Preprint), 6*(3), 187-197. http://doi.org/10.3233/JSA-200384

Groll, A., Ley, C., Schauberger, G., & Van Eetvelde, H. (2019a). A hybrid random forest to predict soccer matches in international tournaments. *Journal of quantitative analysis in sports, 15*(4), 271-287. https://doi.org/10.1515/jqas-2018-0060

Groll, A., Ley, C., Schauberger, G., Van Eetvelde, H., & Zeileis, A. (2019b). Hybrid Machine Learning Forecasts for the FIFA Women's World Cup 2019. *arXiv preprint arXiv:1906.01131*. https://doi.org/10.48550/arXiv.1906.01131

Groll, A., Schauberger, G., & Tutz, G. (2015). Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: An application to the FIFA World Cup 2014. *Journal of Quantitative Analysis in Sports, 11*(2), 97-115. https://doi.org/10.1515/jqas-2014-0051

Gu, W., & Saaty, T. (2019). Predicting the Outcome of a Tennis Tournament: Based on Both Data and Judgments. *Journal of Systems Science and Systems Engineering, 28*, 317-343. https://doi.org/10.1007/s11518-018-5395-3

ITTF Archive. (2019). Retrieved from https://results.ittf.link/index.php?option=com_content&view=featured&Itemid=101

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18-22. https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). New York: Chapman & Hall.

Peters, M., & Murphey, K. (1992). Cluster analysis reveals at least three, and possibly five distinct handedness groups. *Neuropsychologia, 30*(4), 373-380. https://doi.org/10.1016/0028-3932(92)90110-8

R Core Team. (2019). *R: A language and environment for statistical computing*. (R. F. Computing, Producer). R Core Team. https://www.R-project.org/

Robin, X. (2021). pROC (R-Package). *Display and Analyze ROC Curves*. Expasy. http://expasy.org/tools/pROC/

Schauberger, G., & Groll, A. (2018). Predicting matches in international football tournaments with random forests. *Statistical Modelling, 18*(5-6), 460-482. https://doi.org/10.1177/1471082X18799934

Theodoridis, S. (2015). *Machine Learning - A Bayesian and Optimization Perspective*. Amsterdam: Elsevier Ltd.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

World Cup Playing System. (2019). Retrieved from https://ittf.cdnomega.com/eu/2019/02/