Mueller, Steffen Q.

Working Paper

Pre- and within-season attendance forecasting in Major League Baseball: A random forest approach

Hamburg Contemporary Economic Discussions, No. 65

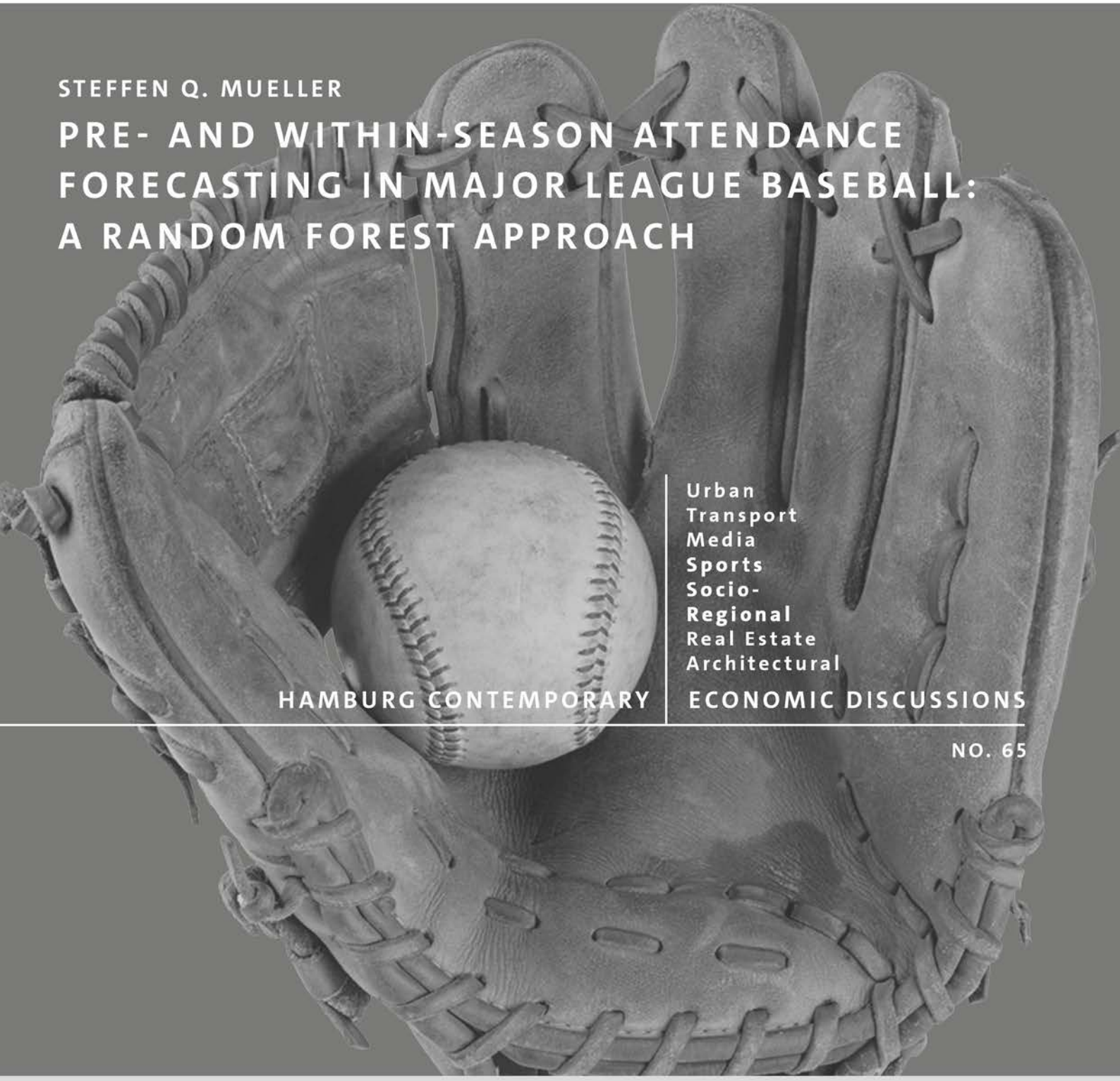
Provided in Cooperation with:
University of Hamburg, Chair for Economic Policy

Suggested Citation: Mueller, Steffen Q. (2018) : Pre- and within-season attendance forecasting in Major League Baseball: A random forest approach, Hamburg Contemporary Economic Discussions, No. 65, ISBN 978-3-942820-45-5, University of Hamburg, Faculty of Business, Economics and Social Sciences, Chair for Economic Policy, Hamburg

This Version is available at:
https://hdl.handle.net/10419/200698

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Faculty of Business,
Economics and Social Sciences
Chair for Economic Policy



STEFFEN Q. MUELLER

PRE- AND WITHIN-SEASON ATTENDANCE FORECASTING IN MAJOR LEAGUE BASEBALL: A RANDOM FOREST APPROACH

Urban
Transport
Media
Sports
Socio-
Regional
Real Estate
Architectural

HAMBURG CONTEMPORARY | ECONOMIC DISCUSSIONS

NO. 65

Hamburg Contemporary Economic Discussions
University of Hamburg
Faculty of Business, Economics and Social Sciences
Chair for Economic Policy
Von-Melle-Park 5
20146 Hamburg | Germany
Tel +49 40 42838 - 4622
Fax +49 40 42838 - 6251
https://www.wiso.uni-hamburg.de/en/fachbereich-vwl/professuren/maennig/home.html
Editor: Wolfgang Maennig

Steffen Q. Mueller
University of Hamburg
Faculty of Business, Economics and Social Sciences
Chair for Economic Policy
Von-Melle-Park 5
20146 Hamburg | Germany
Tel +49 40 42838 - 5297
Fax +49 40 42838 - 6251
Steffen.Mueller@uni-hamburg.de

Photo Cover: Nor Gal/Shutterstock.com

ISSN 1865 - 2441 (Print)
ISSN 1865 - 7133 (Online)

ISBN 978-3-942820-44-8 (Print)
ISBN 978-3-942820-45-5 (Online)

Steffen Q. Mueller

Pre- and within-season attendance forecasting in Major League Baseball: A random forest approach

Abstract: This study explores the forecasting of Major League Baseball game ticket sales and identifies important attendance predictors by means of random forests that are grown from classification and regression trees (CART) and conditional inference trees. Unlike previous studies that predict sport demand, I consider different forecasting horizons and only use information that is publicly accessible in advance of a game or season. Models are trained using data from 2013 to 2014 to make predictions for the 2015 regular season. The static within-season approach is complemented by a dynamic month-ahead forecasting strategy. Out-of-sample performance is evaluated for individual teams and tested against least-squares regression and a naive lagged attendance forecast. My empirical results show high variation in team-specific prediction accuracy with respect to both models and forecasting horizons. Linear and tree-ensemble models, on average, do not vary substantially in predictive accuracy; however, OLS regression fails to account for various team-specific peculiarities.

Keywords: Attendance, Major League Baseball, Random forest, Conditional forest, Sport demand, Sports forecasting, Ticket sales, Variable importance
JEL: C44, C53, Z2
Version: June 2018

1 Introduction

According to sport franchises, predicting sport demand in advance of a season is necessary for ticket pricing, and forecasting short-run fluctuations in attendance is important for staffing (Kleps, 2014). However, existent studies on predicting sport demand do not consider multiple forecasting horizons and mainly use linear and normal censored regression methods (e.g. Beckman et al., 2012; J. Borland & Macdonald, 2003; Denaux et

al., 2011; Lemke et al., 2010).[1] In contrast, this paper investigates tree-based ensemble methods for both pre- and within-season attendance forecasting.

In this study, I forecast ticket sales by means of random forest regressions for all 29 US Major League Baseball (MLB) teams for the regular 2015 season using data from 2013 to 2014 for model training. Precisely, I predict individual game attendance and identify important predictors by random forests that are grown from classification and regression trees (CART) (Breiman, 2001) and conditional inference trees (Strobl et al., 2007, 2008). To this extent, I distinguish between two sets of predictors. The first set includes variables that are known in advance of a season (e.g. game and promotion schedule), while the second set is extended to include variables that are observed as a season progresses (e.g. lagged attendance and team performance). Similar to McHale & Morton (2011), who forecast tennis match results, I complement my static predictions by introducing a dynamic month-ahead forecasting strategy in which the training data and models are iteratively updated on a monthly basis.

The random forest (RF) ensemble technique is a state-of-the-art machine learning algorithm that has been shown to yield accurate predictions in a wide range of regression and classification tasks (e.g. Lessmann et al., 2010; Lessmann & Voß, 2017; Nedellec et al., 2014; Swartz et al., 2017). RF automatically accounts for complex non-linear dependencies between considered predictors and the dependent variable (Hastie et al., 2009). This ability makes RF a promising tool in attendance forecasting, since there are many variables that are likely to impact fans' preferences in various and interdependent ways. As an example, fans want to experience an exciting game and, at the same time, want their home team to win, which is not necessarily the same objective and may interact

[1] It is common practice in the sport demand literature to use attendance and ticket sales as proxies for sport demand (J. Borland & Macdonald, 2003). Furthermore, the officially reported attendance figures are the total number of sold tickets per game, not the number of fans that were present at a game. In this paper, the terms sport demand, ticket sales, and attendance are used interchangeably.

with additional factors such as game importance, fan rivalries, and media coverage (Forrest et al., 2005).

This study makes several important contributions. First, it introduces a novel strategy for both pre- and within-season attendance forecasting by exploring the predictive capabilities of static and dynamic random forest approaches. Second, out-of-sample performance is evaluated for individual teams and tested against least-squares regression and a naive lagged attendance forecast. Third, I restrict the set of considered predictors exclusively to measures that are observable and publicly accessible before a season starts or a game is played. Fourth, I provide a robust assessment of variables' impact on predictive accuracy by comparing permutation importance measures that are derived from the random and conditional forest predictions.

The remainder of the paper is organized as follows: Section 2 discusses aspects in predicting attendance, and Section 3 describes the data that are employed in this study. Section 4 briefly reviews the methodologies of RF and CF regression. Section 5 shows the results for both the static pre- and within-season approach and the dynamic short-run forecasting strategy. Section 6 presents the conclusions of the paper.

2 Predicting game attendance and determinants of demand

Fans decide to attend a game based on not only economic variables of demand theory such as income and ticket price but also specific sport and game characteristics, e.g. competitive balance and outcome uncertainty (e.g. Dennis Coates et al., 2014; Forrest & Simmons, 2002). The list of potentially relevant predictors is extensive. Among others, additional attendance drivers are the day and time of a game, promotions, weather conditions, newly constructed stadiums, and city and population characteristics (e.g. Denaux et al., 2011; Feddersen et al., 2006; Winfree et al., 2004).

Studies on predicting season or game attendance usually focus on single sports, e.g. soccer (Villa et al., 2011), basketball (Zhang et al., 1995), ice hockey (D. Coates & Humphreys,

2012), Australian football (Jeff Borland & Lye, 1992), U.S. football (Welki & Zlatoper, 1999), and MLB (Lemke et al., 2010). However, most articles on sport demand attempt to explain in-sample attendance variation and use information on exogenous variables that is not strictly observable or publicly accessible in advance of a game (J. Borland & Macdonald, 2003). Examples include average season ticket prices, team payroll, game-day temperature, and macroeconomic variables on various geographical levels (e.g. Beckman et al., 2012; Lemke et al., 2010; Tainsky & Winfree, 2010; Villa et al., 2011; Winfree et al., 2004). I found only two studies that predict stadium attendance without relying on information that is not accessible before a game has started and both use artificial neural network models to forecast short-run soccer match attendance rates (Şahin & Erol, 2017; Strnad et al., 2017).

Frequently employed models in predicting attendance are linear regression methods such as OLS, and censored-normal regression models since stadium capacity limits game attendance (Beckman et al., 2012; Denaux et al., 2011; Lemke et al., 2010). A commonly applied variable transformation is the natural logarithm of game attendance, and some studies consider interaction terms between certain predictors, e.g. squared stadium age (Tainsky & Winfree, 2010). Conversely, RF is a data-driven method that accounts for the impact of higher-order interactions and non-linear dependencies without the need for pre-specification (Hastie et al., 2009).

To the best of my knowledge, only one article has been published in a peer-reviewed journal that also applies tree-based methods to analyze sport demand. King (2017) predicts individual NBA game attendance by CART RF. In contrast to my study, King (2017) employs a static forecast without considering multiple forecasting horizons and includes information on predictors that is not accessible at the time of model training and prediction. Furthermore, tree-based ensemble methods have already been applied in MLB research. Mills & Salaga (2011) and Freiman (2010) predict the election of hitters and pitchers into the National Hall of Fame by the Baseball Writers' Association by RF classification and Swartz et al. (2017) estimate pitch quality by RF regression.

3 Data description

The variables that are employed in this study are all publicly accessible in advance of a season or the night before game-day. My data sources are retrosheet.org (game-log data), MLB.com (promotions), seamheads.com (information on stadiums), covers.com (betting odds), darksky.net (weather API), and Beckman et al. (2012) and Lemke et al., (2010) (team rivalries).[2]

The original data sample covers all 7290 games that were played over the course of the 2013, 2014 and 2015 regular seasons. Since I include lagged attendance as a predictor in my analysis, I drop the corresponding 90 first home games. Furthermore, I only consider US teams in this study and, thus, drop the remaining 240 home games that were hosted by the Toronto Blue Jays. After additional minor adjustments that are common in the sport economics literature, the final data sample includes observations on 6852 games: 4571 records from the 2013 and 2014 seasons as a training set and 2281 records from the 2015 season as a hold-out test set. Concise descriptions of the data cleaning process, variable specifications, and descriptive statistics are provided in the Appendix. The 38 predictor variables that are employed in this study are summarized in Table 1.

[2] http://www.retrosheet.org, https://www.mlb.com, http://www.seamheads.com, https://www.covers.com, https://darksky.net.

Table 1 Description of pre- and within-season predictor variables

Variables observed in advance of a season
5 variables related to the date and time a game is scheduled
6 variables related to stadium, city, and team characteristics
5 variables related to team rivalries and specific match characteristics
6 variables related to teams' former season success
4 variables related to game promotions
Variables observed as a season progresses
Lagged home team game attendance
Home team's winning probability (calculated from betting odds)
4 variables related to relative team performance
5 variables related to weather conditions (day before game-day)
Season (only included in the dynamic forecast)

Notes: This study includes 38 predictor variables: 12 numerical and 26 categorical variables with a total of 98 levels (see Section 2 in the Appendix).

Although weather conditions can be expected to have an impact on game attendance, they are often not considered in empirical research or only refer to the temperature that is measured at the beginning of a game (e.g. Kappe et al., 2014; Lemke et al., 2010). In contrast, I include several measures that account for the weather conditions of the day before a game is played. However, there are numerous potential attendance factors that are not considered in this study. For example, one may include information on fans' preferences that is derived from social-media activities.

4 Methodology

4.1 Random forest regression

The RF technique is an ensemble method that combines multiple de-correlated decision tree predictors on the basis of various sub-sets of a data sample (Hastie et al., 2009). The original RF approach averages the predictions that are generated from many unpruned single CART trees that are fitted to random draws of the training data with replacement, which is referred to as bootstrap aggregation ('bagging') (Breiman, 1994, 2001; Breiman et al., 1984). An RF is grown from B bootstrap samples that each include individual observations multiples times, while some observations are not included (approximately

one third). The observations that are not included in the data that are used to fit a tree are called out-of-bag (OOB) observations. In contrast to bagging, the RF procedure imposes an additional form of randomness by only considering a random subset of M predictors for the respective candidate variable any time a node is split in the tree building process. As a result, RF generates more diverse trees by allowing splitting rules on variables at early stages of a single tree that would otherwise be neglected (Breiman, 2001).

A convenient feature of bagged models is that they allow hyper-parameters to be determined in a way that is similar to cross-validation. Precisely, we can evaluate model performance by predicting the outcome for an observation i using each of the single trees in which this observation was not included in the training process, i.e. in which this observation was OOB. This evaluation yields approximately $B/3$ predictions for the ith observation. The RF OOB prediction for the ith observation is simply the average of those $B/3$ predictions (or majority vote for classification). Using the OOB estimates for model tuning is less computationally demanding than cross-validation since no additional models (forests) must be trained to test a set of parameters (Lessmann et al., 2010).

The importance of each predictor in the RF tree building process can be assessed via different measures of variable importance. The arguably most-advanced RF measure is computed by calculating the difference in prediction accuracy that results from randomly permuting a predictor variable using the observations that are recorded in the OOB data (Strobl et al., 2007). The reasoning is intuitive: Let us assume that the difference in the prediction accuracy on the OOB records is substantially affected by whether we include a predictor X_j or not, i.e. X_j is a strong predictor. Then, it is reasonable to assume that assigning a different value to X_j increases the resulting prediction error. Hence, permuting a variable over its values that were recorded in the OOB data enables one to mimic the exclusion of the predictor and calculate the resulting mean difference in MSE on the OOB data (Breiman, 2001; Strobl et al., 2007).

4.2 Conditional random forest regression

While the RF permutation importance measure covers both the individual impact of the assessed predictor and complex higher-order interactions with other predictors, it is biased in favor of numerical over categorical variables and similarly favors categorical variables with many levels (e.g. Archer & Kimes, 2008; Strobl et al., 2007). Precisely, Strobl et al. (2007) show that the RF inhibits a variable selection bias that emerges from CART and an additional bias that is induced by bootstrap sampling. As an alternative to CART, Strobl et al. (2007) propose using conditional inference trees as base learners. The main difference with RF is that the conditional forest (CF) aggregation scheme of the single-tree predictors within a forest involves averaging observation weights that are extracted from each of the trees, not simply averaging the predictions directly (Strobl et al., 2007, 2008).

In a later study, Strobl et al. (2008) find that the CF approach in Strobl et al. (2007) still favors correlated predictors in the tree building process; this bias is induced by the unconditional variable importance permutation scheme of CF. To account for this bias, Strobl et al. (2008) suggest conditionally permuting predictor variables to correlated ones, which they refer to as conditional permutation importance. However, there is no general consensus on how to interpret the importance measures when predictors are correlated and, more importantly, it is unclear how those correlations effectively impact CF importance measures (e.g. Nicodemus et al., 2010).

5 Implementation and results

I use *R* (R Core Team, 2017) and the packages *randomForest* (Liaw & Wiener, 2002), *party* (Hothorn et al., 2015), and *lattice* (Sarkar, 2008) for the main computations and graphics in this paper.

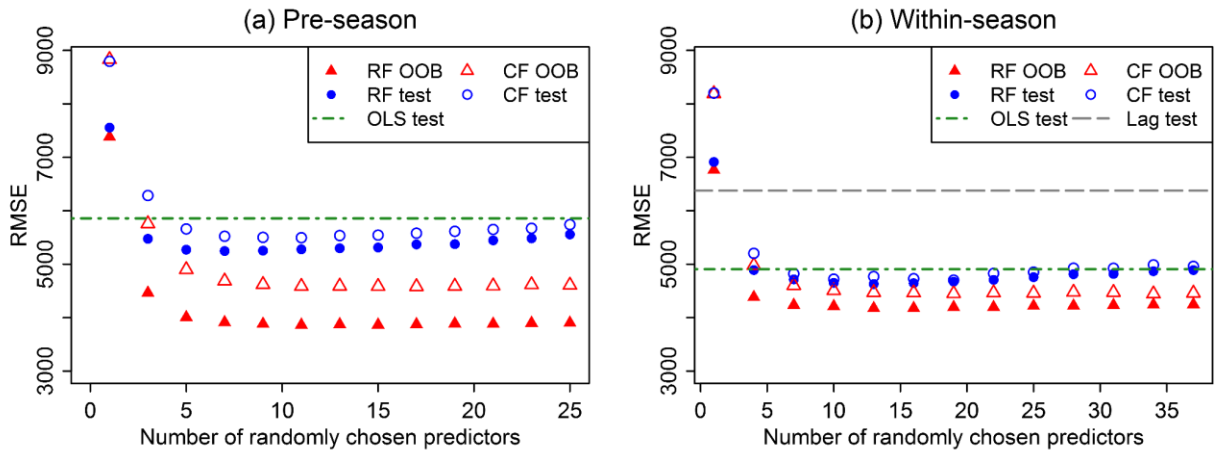
5.1 Static pre- and within-season forecasts

This study employs 37 within-season predictor variables, which are denoted as X^{ws}, in the static forecasting approach: 26 pre-season variables X^{ps} and 11 short-run variables X^{sr}. I forecast individual game attendance y_i by random forest regressions based on CART (RF) and conditional inference trees (CF), a standard OLS regression model, and a naive forecast that equals the lagged home-team game attendance (Lag).

5.1.1 Model performance evaluation

In this paper, I follow the suggestion of Hastie et al. (2009) and exclusively grow trees to their maximal depths. This procedure simplifies parameter tuning and, with respect to this study, requires a justifiable increase in computational cost. Moreover, I quickly observed that the predictive performance for both pre- and within-season models is not very sensitive to the number of included trees per forest. For example, using the suggested default value (one third) for the number of randomly chosen predictors in the tree building process, the RMSE on the OOB and test data for both the RF and CF approaches stabilizes after averaging the prediction results of less than 100 trees (see Section 4 of the Appendix). However, in the further analyses, I train RF and CF models on the basis of $B = 500$ trees to ensure stable estimates of variable importance measures (Liaw & Wiener, 2002). Figure 1 shows the OOB and test set RMSEs for the pre-season [within-season] forecasts that are generated by the RF, CF, OLS, and lagged attendance models as functions of the number $M_{ps} = \{1, 3, \dots, 25\}$ $[M_{ws} = \{1, 4, \dots, 37\}]$ of randomly considered predictors at each split in the tree building process.

Figure 1 Model performance evaluation: RMSE by number of randomly chosen predictors



Notes: Out-of-sample (test) and out-of-bag (OOB) MLB game attendance RMSEs by random forest regressions based on CART (RF) and conditional inference trees (CF), OLS regression, and a simple lagged home-team attendance forecast (Lag). Maximal complex RF and CF are trained using $B = 500$ trees.

The RF yields the most accurate results but only slightly outperforms OLS and CF regressions for both the pre- (a) and within-season (b) forecasting horizons. The RF and CF performances on the OOB records appear to be relatively stable after the inclusion of 10 randomly chosen predictors at each split. However, the suggested default value of one third for the number of randomly considered predictors M appears to approximately minimize RMSE for the RF and CF approaches for both the pre-season (a) and within-season (b) test data. For (a), the corresponding results for the RF [CF] model with $M_{ps} = 7$ yield minimum RMSEs of 3912 [4686] on the OOB data and 5241 [5522] on the test data. For (b), the corresponding results for RF [CF] with $M_{ws} = 12$ yield a minimal RMSE of 4205 [4478] on the OOB data and 4634 [4743] on the test data. The OLS model RMSEs are 5858 (a) and 4908 (b), while the simple home-team-specific lagged attendance forecast (Lag) achieves an RMSE of 6377 (b). Hence, the differences in prediction accuracy among RF, CF, and OLS are stronger for the pre-season forecast, but do not vary substantially when trained with the additional information that is provided by the within-season variables.

5.1.2 Team-specific results

Based on the performance evaluation in Section 5.1.1, I use $B = 500$ trees and set the number of randomly chosen predictors to $M_{ps} = 7$ and $M_{ws} = 12$ for all static pre- and

within-season RF and CF models. Table 2 shows the corresponding out-of-sample RMSEs for each home team for the static pre- and within-season predictions, together with the attendance summary statistics for the regular seasons from 2013 to 2015.

Table 2 Out-of-sample MLB game attendance prediction accuracy by home team

	Seasons 2013 to 2015			Out-of-sample RMSE for the 2015 season						
	Attendance summary			Random forest			OLS			Lag att
Team name	Abr.	Mean	SD	Pre	Within	Diff	Pre	Within	Diff	Within
All	-	30283	9609	5241	4634	607	5858	4908	950	6377
Arizona Diamond Backs	ARI	25639	7029	4959	4393	566	4820	4531	289	8737
Atlanta Braves	ATL	28313	8735	7063	6273	790	7465	5634	1831	8799
Baltimore Orioles	BAL	29671	9277	6861	6257	604	6421	6391	30	10230
Boston Red Sox	BOS	35694	2039	2048	1818	230	3381	3349	32	1532
Chicago Cubs	CHC	33869	4373	5023	4296	727	5657	3850	1807	3025
Chicago White Sox	CHW	21369	6265	4532	4300	232	4812	4275	537	6762
Cincinnati Reds	CIN	30366	7543	5047	4848	199	5886	5132	754	7797
Cleveland Indians	CLE	18386	6580	3850	3809	41	3815	3538	277	5702
Colorado Rockies	COL	32965	6746	5428	4572	856	4596	4035	561	6653
Detroit Tigers	DET	35781	5219	3430	3024	406	4868	3657	1211	4671
Houston Astros	HOU	22626	6579	6907	4833	2074	6805	3892	2913	6361
Kansas City Royals	KCR	26297	8176	5002	5280	-278	10447	8510	1937	5298
Los Angeles Angels	LAA	37479	4053	4139	3949	190	3400	3285	115	5282
Los Angeles Dodgers	LAD	46377	5100	3789	3905	-116	4023	4237	-214	5215
Miami Marlins	MIA	20676	4200	3880	3871	9	4684	4373	311	5349
Milwaukee Brewers	MIL	32223	6316	4794	4328	466	4403	4164	239	6700
Minnesota Twins	MIN	28560	5231	4652	4272	380	5020	4123	897	5641
New York Mets	NYM	28208	6260	7262	6562	700	6905	5418	1487	7341
New York Yankees	NYY	40892	4723	4748	4368	380	4737	4696	41	5193
Oakland Athletics	OAK	22944	7516	4851	4805	46	5089	4914	175	7779
Philadelphia Phillies	PHI	30047	7172	8955	6358	2597	9622	6531	3091	5148
Pittsburgh Pirates	PIT	29614	7926	4369	3900	469	5199	4463	736	6470
San Diego Padres	SDP	27886	8042	6121	5790	331	8077	6780	1297	8881
Seattle Mariners	SEA	24522	8773	7061	6345	716	7683	6650	1033	8717
San Francisco Giants	SFG	41583	618	427	410	17	3960	3685	275	406
St. Louis Cardinals	STL	42851	2464	2069	1874	195	3589	3193	396	2223
Tampa Bay Rays	TBR	17069	5895	4746	4283	463	4400	3696	704	5214
Texas Rangers	TEX	34169	6701	7355	5728	1627	7717	6554	1163	7002
Washington Nationals	WAS	32254	5339	3939	3848	91	4338	3930	408	5827
R2	-	-	-	0.697	0.763	0.066	0.630	0.737	0.107	0.551

Notes: Summary statistics and static pre-and within-season forecasts for US home-team game attendance.

With respect to average team results, the RF model performs only slightly better than the OLS model for both pre-season (PS) and within-season (WS) predictions. For the pre-season (PS), the RF model yields an RMSE of 5241 and within-season (WS) an RMSE of 4634. The naive home-team-specific lagged attendance (LAG) model performs worst, with an average RMSE of 6377. The CF results do not differ substantially from the RF results and, therefore, are included in the

Appendix. The RMSEs for the OLS regression are 5858 (PS) and 4908 (WS), and 6377 (WS) for the LAG predictions.

Most importantly, the results in Table 2 reveal the high variation in prediction accuracy within and across teams, models, and forecasting horizon. A good example for team-specific peculiarities are the San Francisco Giants (SFG). With respect to team-specific differences in the PS and WS results for the RF approach, the corresponding RMSEs are only 427 (PS) and 410 (WS) sold tickets per game. In contrast, with RMSEs of 3960 (PS) and 3685 (WS), the OLS model is not able to account for the unique peculiarities that are associated with SFG. Precisely, SFG's standard deviation in ticket sales per game across the 2013, 2014 and 2015 seasons is as low as 618 and the average game attendance is 41583. SFG's stadium capacity is reported as 41915 over all seasons, which implies that practically every game was almost sold out. As a result, the lagged attendance model (LAG) achieves a corresponding RMSE of 406 sold tickets per game. Similarly, the LAG is also more accurate than RF and OLS are for BOS, CHC, and PHI, and more accurate than OLS is for STL and KCR.

There are also instances in which the OLS model performs best with respect to the PS forecasting horizon (ARI, BAL, CLE, COL, HOU, LAA, MIL, NYM, NYY, and TBR) and the WS predictions (ATL, CHW, CLE, COL, HOU, LAA, MIL, NYM, MIN, and TBR). However, the corresponding differences between OLS and RF are small for the teams for which the OLS performs better, e.g. the maximal difference for PS is 832 (COL) and for WS 1144 (NYM). In contrast, the maximal difference in RMSEs for the teams for which the RF outperforms the OLS model are 5445 for PS (KCR) and 3275 for WS (KCR). Moreover, there are substantial differences with respect to the improvement in prediction accuracy that is obtained by including short-run information in the WS framework. For the RF approach, the difference between the PS and WS RMSEs is negative for KCR and LAD, and for the OLS model for LAD. Lastly, the highest RMSEs for the RF model are 8955 for PS (PHI) and 6562 for WS (NYM), and for the PS and WS OLS model 10447 (KCR) and 8510 (KCR), respectively. However, I train and evaluate the RF and CF approaches over all teams. To improve prediction accuracy, we may simply train and optimize models with respect to each team individually.

5.1.3 Variable importance analysis

Table 3 shows the ranking of predictors according to their impact on the forest building process in terms of the permutation importance measures of the RF (scaled mean decrease in MSE) (Breiman, 2001), the unconditional CF (Strobl et al., 2007), and the conditional CF (CCF) (Strobl et al., 2008) approaches. Precisely, I present the RF, CF, and CCF rankings for the ten most important predictors of the RF pre-season and within-season models (lowest rank corresponds to highest impact).

Table 3 Comparison of random and conditional forest variable importance rankings

	Pre-season			Within-season		
Variable	RF	CF	CCF	RF	CF	CCF
Variables observed before a season starts						
Weekday	1	3	1	1	3	1
Home team (HT) indicator	2	1	2	2	1	15
HT season game number	3	9	10	4	14	30
Distance between stadiums	4	14	5	10	23	34
Month	5	7	4	13	9	36
Visiting team (VT) indicator	6	13	3	16	16	18
Fireworks promotion	7	12	16	7	7	17
Stadium capacity	8	2	6	8	4	23
Day or night game	9	15	17	14	13	16
VT is Division Series Winner	10	18	13	22	22	32
Variables observed before a game starts						
HT lagged attendance				3	2	2
Maximum temperature				5	20	11
HT games behind				6	17	10
Minimum temperature				9	18	21
HT winning percentage				12	10	6
VT winning percentage				15	19	3
VT games behind				17	26	7
HT implied winning probability				20	24	5
Relative humidity				21	27	8
Weather conditions				30	33	26

Notes: Permutation importance rankings (lowest rank corresponds to highest impact) are derived from the OOB estimates for the 2013 and 2014 regular seasons from the random and conditional forest regressions for US home-team-specific MLB game attendance (see Section 5.1.2). $P_{ps} = 26$ $[P_{ws} = 37]$ included predictors for the pre-season [within-season] forecast. Ranking of predictor relevance in the forest building process according to permutation variable importance is performed using the (scaled) CART random forest (RF) and conditional forest (CF) measures and the conditional permutation importance CF measure (CCF).

For the pre-season forecast, the rankings in terms of RF, CF, and CCF largely appear to identify the same predictors as being of relatively high importance in the tree and forest building process. Weekday and home team (HT) effects are consistently ranked among

the three most important predictors. However, there are differences among the considered ranking approaches: The CF measure ranks ballpark capacity as the second most important variable, and the third rank in the CCF approach is assigned to the dummy variable that indicates whether a game is played against BOS, CHC, or NYY (VT).

The largest differences in rankings are observed for the distance between stadiums, VT, and fireworks promotions. Conversely, an HT's season game number is a numeric variable that seems to be artificially preferred in the RF. With respect to highly correlated variables, in contrast to the RF and CCF, the CF assigns the first and second ranks to stadium capacity and the HT indicator.

The differences among the RF, CF, and CCF rankings appear to be more severe for the within-season forecast. First, the rank for lagged HT attendance is between two and three for RF, CF, and CCF. However, CCF results in a vastly different ranking compared to all other PS and WS rankings for an HT's indicator and the number of games, distance, and month. This result is unexpected and seems unreasonable since the WS forecast is not substantially better than the PS forecast is for the RF or the CF model (see Table 2). Although the RF and CF results suggest that the additional WS information (e.g. relative team rankings) has no substantial impact on the predictive accuracy, CCF still ranks many of the additional WS variables among the ten most important predictors. Lastly, the differences between the RF and CF ranking are small. However, similar to the PS rankings, the RF appears to rank specific numeric and continuous variables relatively higher. However, one should be careful when interpreting these results since the stadium capacities and distances between competing teams' ballparks do not vary substantially within seasons, and there is high correlation between current- and previous-season success and individual teams (Tainsky & Winfree, 2010).

Lastly, similar to Lessmann et al. (2010), I assess the observed significance in the differences among the variable importance rankings that are produced by CF, RF, and CCF by computing the corresponding ranking correlation coefficients by means of Kendall's tau. The WS CCF ranking is reported to be statistically significantly different from all other

PS and WS rankings at a minimum p-value of 0.311. All other combinations of differences in importance rankings across models for both PS and WS are similar. The precise results are included in the Appendix.

5.2 Dynamic within-season forecast

In the dynamic within-season forecast, I iteratively update the training data after each month. The models are trained using the updated training set to make predictions for the games of the next consecutive month. Then, I repeat this procedure for all months of the 2015 regular season. Moreover, in contrast to the static within-season approach, I include an additional categorical variable that indexes the corresponding season, which results in $X^{dyn} = 38$ predictors for the dynamic forecasting strategy. Unlike the static predictions, this approach allows one to account for seasonal differences in preferences for game attendance. Table 4 shows the out-of-sample results for the dynamic RF, CF, OLS, and lagged attendance month-ahead predictions. As in the static forecasting approach, I train the RF and CF models using $B = 500$ trees and $M_{ws} = 12$ randomly chosen predictors.

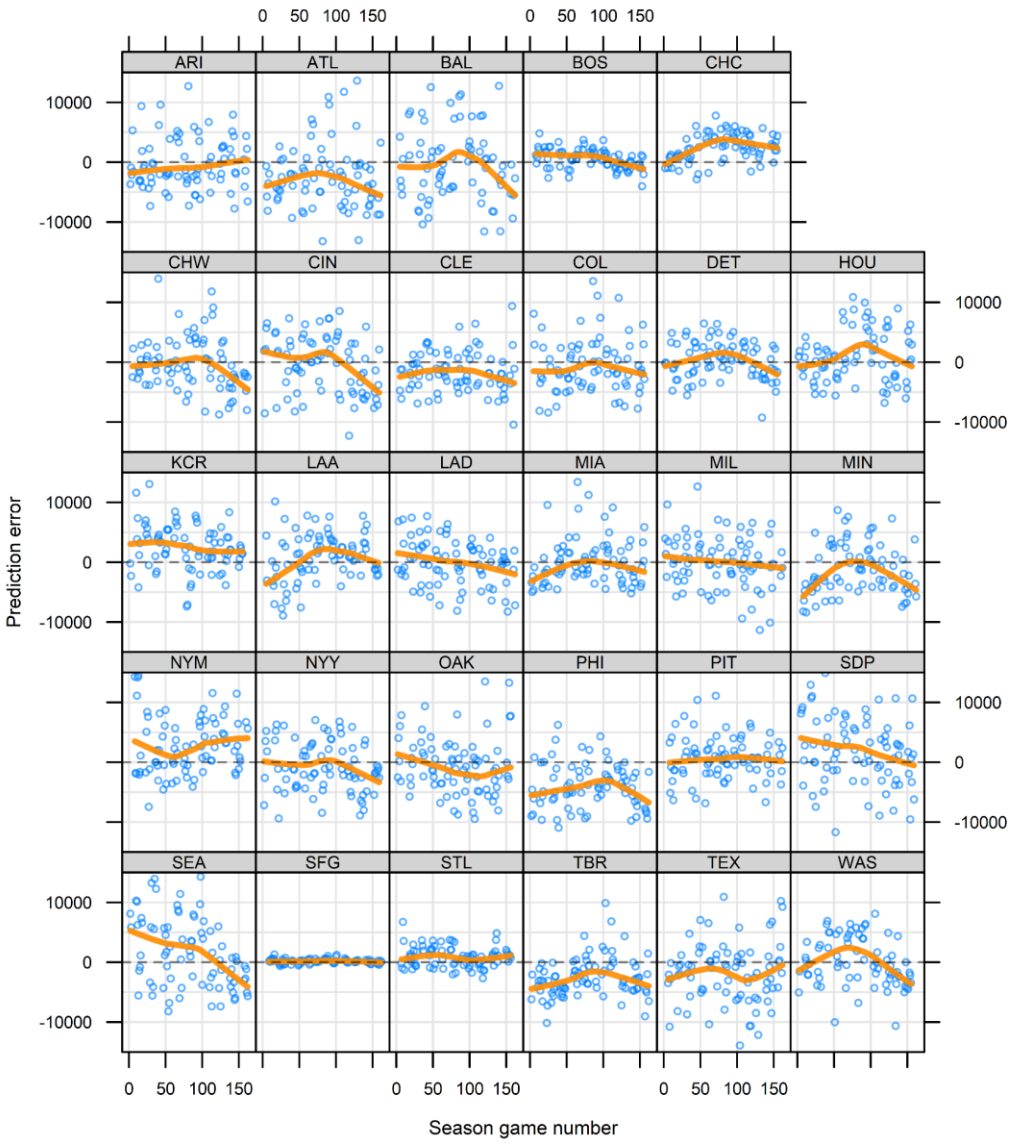
Table 4 Dynamic within-season attendance forecasts

		Out-of-sample month-ahead prediction accuracy					
Model		Apr	May	Jun	Jul	Aug	Sep
Random forest	RMSE	4608	4481	3994	4373	4424	4405
	R2	0.79	0.78	0.80	0.76	0.76	0.80
Conditional forest	RMSE	4780	4655	4198	4589	4523	4519
	R2	0.77	0.76	0.78	0.73	0.75	0.80
OLS	RMSE	4540	4481	4411	4432	4419	4427
	R2	0.74	0.75	0.78	0.78	0.79	0.78
Lagged attendance	RMSE	7863	6587	5846	5983	6505	5694
	R2	0.39	0.52	0.57	0.54	0.48	0.67

Notes: Training data and models are iteratively updated after each month of the MLB 2015 season.

The dynamic forecasting approach produces only slightly more accurate predictions than those of the static approach. For example, the difference between the RMSE of the static WS RF approach and the average monthly RMSE of the dynamic WS RF approach is only 299 tickets per game. The RF model performs only marginally better than the OLS model does, and both explain, on average, 78% of the variation in attendance, while the monthly average for the CF model is 76%. For the July and August game attendance predictions, the OLS model is even more accurate than the RF and CF models are. Furthermore, the lagged attendance prediction results indicate that variation in game-to-game attendance at the beginning of the 2015 season is relatively high (April), but relatively low during the end (September). Instead of the team-specific RMSE, I show the aggregated monthly prediction errors by season game number and team for the dynamic RF model in Figure 2.

Figure 2 Dynamic random forest attendance forecast



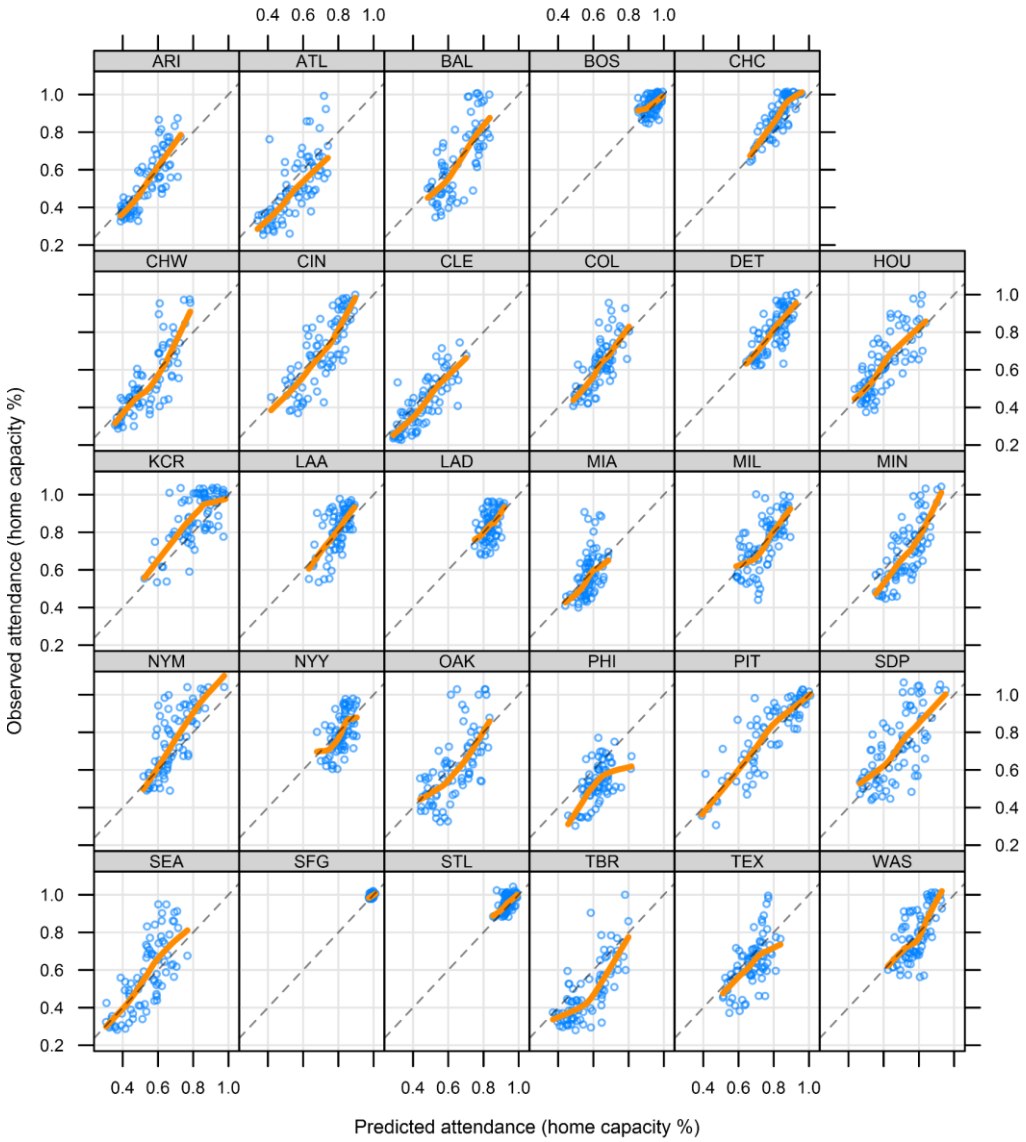
Notes: Dynamic random forest attendance forecasts for the 2015 MLB regular season by aggregated out-of-sample month-ahead prediction error for US home-team-specific game attendance. Orange lines correspond to LOESS smoothing curves.

As in the static pre- and within-season forecasts, the dynamic approach shows high heterogeneity in the predictive accuracy for game attendance across home teams. In contrast to the team-specific results that are presented in Section 4.1, Figure 1 shows for which games the RF forecasts over- and underestimate attendance. A casual inspection reveals a bell-shaped LOESS curve, especially for BAL, HOU, LAA, MIN, PHI, and WAS. However, there are also teams that show more linear and approximately unbiased curves, e.g. MIL, PIT, and SFG. Furthermore, SEA appears to be an interesting case in

which the variance of prediction accuracy decreases as the season progresses. The lowest prediction errors are obtained for BOS, SFG, and STL. Examples of high variation in predictive accuracy in terms of magnitude and direction are produced for ATL, BAL, SDP, and SEA.

However, Figure 1 does not account for the large differences in team-specific attendance rates. To complete my analysis, I show the differences in observed and predicted game attendance relative to stadium capacity in Figure 2.

Figure 3 Dynamic random forest attendance rate forecast



Notes: Dynamic random forest attendance forecast for the MLB 2015 regular season by aggregated out-of-sample month-ahead prediction error for US home-team-specific game attendance rates and stadium capacities. Orange lines correspond to LOESS smoothing curves. Dashed lines indicate a perfect attendance rate forecast.

The results in Figure 3 reveal that there is high variation not only in team-specific prediction accuracy of absolute attendance but also in attendance relative to stadium capacity. The general pattern seems reasonable since the games of teams with consistently high attendance rates throughout the season are well predicted, e.g. the games of BOS, SFG, and STL. In contrast, ticket sales for teams that face a greater variation in game attendance are predicted less well, e.g. ATL, SDP, and SEA. Moreover, with respect to

teams with low attendance rates throughout the season, RF appears to overestimate attendance for TBR and PHI and underestimate attendance for, e.g. CIN, KCR, and NYM.

6 Conclusions

The vast majority of studies that predict stadium attendance have employed linear and censored regression models, do not consider multiple forecasting horizons, and use information on variables that is nonexistent or not publicly accessible in advance of a game or season. In contrast, this study explores the predictive capabilities of RF and CF regressions for pre- and within-season attendance forecasting without relying on such information. In addition to static predictions, I propose a dynamic month-ahead forecasting strategy in which the training data are iteratively updated on a monthly basis. In an example of forecasting game ticket sales and identifying important attendance predictors in MLB, I find that prediction accuracy and within-season information gain can highly depend on team-specific characteristics. My empirical results show that OLS regression, on average, performs only slightly worse than RF does. However, OLS fails to account for the peculiarities of a small number of teams. Consequently, this study shows that data-driven methods are promising tools in sports demand forecasting since relevant attendance factors are likely to impact fans' preferences across teams in different and interdependent ways.

References

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics and Data Analysis*, *52*(4), 2249–2260.

Beckman, E. M., Cai, W., Esrock, R. M., & Lemke, R. J. (2012). Explaining Game-to-Game Ticket Sales for Major League Baseball Games Over Time. *Journal of Sports Economics*, *13*(5), 536–553.

Borland, J., & Lye, J. (1992). Attendance at Australian Rules football: A panel study. *Applied Economics*, *24*(9), 1053–1058.

Borland, J., & Macdonald, R. (2003). Demand for Sport. *Oxford Review of Economic Policy*, *19*(4), 478–502.

Breiman, L. (1994). *Bagging predictors: Technical Report No. 421. Machine Learning*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. *Wadsworth International Group*.

Coates, D., & Humphreys, B. R. (2012). Game Attendance and Outcome Uncertainty in the National Hockey League. *Journal of Sports Economics*, *13*(4), 364–377.

Coates, D., Humphreys, B. R., & Zhou, L. (2014). Reference-dependent preferences, loss aversion, and live game attendance. *Economic Inquiry*, *52*(3), 959–973.

Denaux, Z. S., Denaux, D. A., & Yalcin, Y. (2011). Factors Affecting Attendance of Major League Baseball: Revisited. *Atlantic Economic Journal*, *39*(2), 117–127.

Feddersen, A., Maennig, W., & Borcherding, M. (2006). The Novelty Effect of the New Football Stadia: The Case of Germany. *International Journal of Sport Finance*, *1*(3), 174–188.

Forrest, D., & Simmons, R. (2002). Outcome uncertainty and attendance demand in sport: The case of English soccer. *Journal of the Royal Statistical Society Series D: The Statistician*, *51*(2), 229–241.

Forrest, D., Simmons, R., & Buraimo, B. (2005). Outcome uncertainty and the couch potato audience. *Scottish Journal of Political Economy*.

Freiman, M. H. (2010). Using Random Forests and Simulated Annealing to Predict Probabilities of Election to the Baseball Hall of Fame. *Journal of Quantitative Analysis in Sports*, *6*(2).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). New York: Springer.

Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. (2015). party: A Laboratory for Recursive Partytioning. *R Package Version 0.9-0*, (1994), 37.

Kappe, E., Stadler Blank, A., & DeSarbo, W. S. (2014). A general multiple distributed lag framework for estimating the dynamic effects of promotions. *Management Science*,

60(6), 1489–1510.

King, B. E. (2017). Predicting National Basketball Association Game Attendance Using Random Forests. *Journal of Computer Science and Information Technology*, *5*(1), 1–14.

Kleps, K. (2014). Indians are forecasting, studying park attendance. Retrieved January 1, 2017, from http://www.crainscleveland.com/article/20140727/SUB1/307279990/indians-are-forecasting-studying-park-attendance.

Lemke, R. J., Leonard, M., & Tlhokwane, K. (2010). Estimating Attendance at Major League Baseball Games for the 2007 Season. *Journal of Sports Economics*, *11*(3), 316–348.

Lessmann, S., Sung, M. C., & Johnson, J. E. V. (2010). Alternative methods of predicting competitive events: An application in horserace betting markets. *International Journal of Forecasting*, *26*(3), 518–536.

Lessmann, S., & Voß, S. (2017). Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy. *International Journal of Forecasting*, *33*(4), 864–877.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(December), 18–22.

McHale, I., & Morton, A. (2011). A Bradley-Terry type model for forecasting tennis match results. *International Journal of Forecasting*, *27*(2), 619–630.

Mills, B. M., & Salaga, S. (2011). Using Tree Ensembles to Analyze National Baseball Hall of Fame Voting Patterns: An Application to Discrimination in BBWAA Voting. *Journal of Quantitative Analysis in Sports*, *7*(4).

Nedellec, R., Cugliari, J., & Goude, Y. (2014). GEFCom2012: Electric load forecasting and backcasting with semi-parametric models. *International Journal of Forecasting*, *30*(2), 375–381.

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, *11*.

R Core Team. (2017). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*.

Şahin, M., & Erol, R. (2017). A Comparative Study of Neural Networks and ANFIS for Forecasting Attendance Rate of Soccer Games. *Mathematical and Computational Applications*, *22*(4), 43.

Sarkar, D. (2008). Lattice multivariate data visualization with R. *Use R!*, xvii, 265 p.

Strnad, D., Nerat, A., & Kohek, Š. (2017). Neural network models for group behavior prediction: a case of soccer match attendance. *Neural Computing and Applications*, *28*(2), 287–300.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*, 1–11.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(25).

Swartz, P., Grosskopf, M., Bingham, D., & Swartz, T. B. (2017). The Quality of Pitches in Major League Baseball. *The American Statistician*, *To appear*, 1–18.

Tainsky, S., & Winfree, J. A. (2010). Short-Run Demand and Uncertainty of Outcome in Major League Baseball. *Review of Industrial Organization*, *37*(3), 197–214.

Villa, G., Molina, I., & Fried, R. (2011). Modeling attendance at Spanish professional football league. *Journal of Applied Statistics*, *38*(6), 1189–1206.

Welki, A. M., & Zlatoper, T. J. (1999). U.S. Professional Football Game-Day Attendance. *Atlantic Economic Journal*, *27*(3), 285–298.

Winfree, J. A., McCluskey, J. J., Mittelhammer, R. C., & Fort, R. (2004). Location and attendance in major league baseball. *Applied Economics*, *36*(19), 2117–2124.

Zhang, J. J., Pease, D. G., Hui, S. C., & Michaud, T. J. (1995). Variables affecting the spectator decision to attend NBA games. *Sport Marketing Quarterly*, *4*(4), 29–39.

Appendix

1 Introduction

This Appendix provides additional information on the data that are used in this study, empirical specifications, and descriptive statistics, complementing the main analysis by providing robustness verification and detailed results for variable importance rankings and the dynamic within-season forecasting approach. Although this Appendix includes text and results from the main paper for clarity, it is not meant to stand alone.

2 Data and context

Major League Baseball (MLB) is divided into the American League and the National League, which are each divided into three divisions: East, Central, and West. Since 2013, each League has consisted of fifteen teams. There are 29 US teams and one Canadian team, which are equally distributed among the six divisions. The regular season is played from April to September and includes 2430 officially scheduled games in total.[3] In this study, the corresponding 162 games per team and season include 20 inter-league games, 66 inter-division games, and 76 intra-division games.

2.1 Sources and empirical specifications

The data that are used in this study are collected from various sources and only cover variables that are observed and publicly accessible before a season starts (pre-season) and before a game is played (within-season). Most of the variables are obtained from retrosheet.org[4] (game-log data), MLB.com[5] (promotions), seamheads.com[6] (information

[3] There are a few games that are scheduled at the end of March or at the beginning of October. In addition, very few games are usually cancelled at the end of a season, e.g. due to bad weather conditions. However, games are only cancelled if the game does not affect team rankings.

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

[5] https://www.mlb.com

[6] http://www.seamheads.com

on stadiums), covers.com[7] (betting odds), and Lemke et al. (2010) and Beckman et al. (2012) (team rivalries). The geographical regions of the historic weather data are specified with respect to a ballparks' longitude and latitude coordinates. The precise measurements refer to the day before a game is played and are obtained using Dark Sky's weather API[8].

The data sample covers all 7290 games that were played over the course of the 2013, 2014 and 2015 regular seasons. Since lagged attendance is included as a predictor in this analysis, I drop the corresponding 90 first home games. Furthermore, I only consider US teams in this study and, thus, drop the remaining 240 home games that were hosted by the Toronto Blue Jays. In addition to those adjustments, I follow a standard practice in the sport economics literature and discard all 106 rescheduled games from the data sample.[9] Those games are usually rescheduled due to bad weather conditions or other extreme events and sometimes the same games are rescheduled more than once. Lastly, there are two observations with missing attendance numbers for unknown reasons, which are dropped as well. However, I calculate all relevant variables using the whole data sample before I discard any observations, e.g. I include all observations in calculating a team's winning percentage and games behind. In this context, in 2015, three home-team games of the Baltimore Orioles against the Texas Rangers were rescheduled to be played in Arlington. I defined those games as home-team games that were played by the Rangers. As a matter of course, I exclude the two observations with missing attendance data before computing the lagged home-team-specific game attendance. The final data sample includes observations on 6852 games: 4571 records for the training set (2013 and 2014 seasons) and 2281 records for the hold-out test set (2015 season).

Moreover, there are 11 games in the data sample that were not finished during their officially scheduled day. Instead, they were extended and finished one or two days after the scheduled game day. Unfortunately, no attendance numbers are available for the

[7] https://www.covers.com

[8] https://darksky.net.

[9] The average model accuracy only decreases marginally when rescheduled games are included. The main reason I discard rescheduled games is to provide an approach that does not rely on data that are not observable or publicly accessible in advance of a season.

games that were extended and finished on another day or the second game day of a double-header. As is common practice by the MLB Association and in the sports economics literature, I treat the outcomes of extended games as if they had been realized during the first official game day and set the second-day game attendance equal to the first-day game attendance for double-headers, instead of discarding those observations (e.g. Lemke et al., 2010). In the remaining sample, there are two games that were played in March and 53 games that were played in October. I specify those games as if they had been played in April and September, respectively. The predictor variables that are employed in this study are described in Table A.1.

Table A.1 Pre- and within-season predictor variable descriptions

Predictor	Description	Levels
Variables observed before a season starts (#26)		
HT.id	Home-team identification number: ARI, ATL, BAL, BOS, CHC, … , WAS	29
HT.NoG[a]	HT's number of games within seasons	Numeric
WDay[a]	Weekday: Mon, Tue, Wed, Thu, Fri, Sat, Sun	7
Month[a]	Month: Apr., May, Jun., Jul., Aug, Sep.	6
Night[a]	Night: No, Yes	2
PHoliday	Public holiday: No, Yes (Labor Day, 4th, or Memorial Day)	2
CTeams	Number of teams in HT's city or county: One, two	2
Capacity[b]	Stadium capacity	Numeric
SType[b]	Stadium type: Open, retractable roof, dome	3
SBuild[b]	Stadium age: 1-5 years, 6-10 years, +10 years	3
ILGame[c]	Interleague game: No, Yes	2
DivGame[c]	Division game: No, Yes	2
DRgame[d]	Division rivalry game: No, Yes	2
ILRGame[d]	Interleague rivalry game: No, Yes	2
VTeam	Visiting team (VT): Other, BOS, CHC, NYY	4
HT.WSW[c]	HT is last season's World Series winner: No, Yes	2
VT.WSW[c]	VT is last season's World Series winner: No, Yes	2
HT.LCSW[c]	HT is last season's league championship series winner: No, Yes	2
VT.LCSW[c]	VT is last season's league championship series winner: No, Yes	2
HT.DSW[c]	HT is last season's league division series winner: No, Yes	2
VT.DSW[c]	VT is last season's league division series winner: No, Yes	2
Distance[b]	Distance between HT's and VT's stadiums (in miles)	Numeric
FWorks[c]	Fireworks promotion: No, Yes	2
BHeads[c]	Bobblehead promotion: No, Yes	2
OPromo[c]	Other promotion or giveaway: No, Yes	2
DHeader[a]	Game is played as a double-header: No, first game, second game	3
Variables observed as a season progresses (#12)		
Lag.GAttend[a]	Lagged HT-specific game attendance	Numeric
HT.Wprob[e]	HT's winning probability (calculated from betting odds)	Numeric
HT.GB	Games behind between HT and its division-leading team	Numeric
VT.GB	Games behind between VT and its division-leading team	Numeric
HT.Wper	HT's winning percentage (within-season)	Numeric
VT.Wper	VT's winning percentage (within-season)	Numeric
Humidity[f]	Relative humidity during the game before game day (day before)	Numeric
TempMax[f]	Maximum temperature (day before)	Numeric
TempMin[f]	Minimum temperature (day before)	Numeric
Weather[f]	Clear, partly cloudy, cloudy, wind, fog, rain, snow (day before)	6
Precip[f]	Precipitation: No, Yes (day before)	2
Season	Season year: 2013, 2014, 2015 (only included in the dynamic forecast)	3

Notes: The data sample covers 6852 games from the 2013, 2014, and 2015 MLB regular seasons for all 29 US teams. Each HT's first game of the season and rescheduled games are not included. The HT's winning probability is calculated from betting odds. Game log data are obtained from [a] Retrosheet.com. Additional data sources: [b] Seamheads.com, [c] MBL.com, [d] Lemke et al. (2010), Beckman et al. (2012), [e] covers.com [f] and dark-sky.net (API).

While most variable descriptions are self-explanatory, in the following, I discuss further details with respect to their empirical specifications and corresponding implications. The categorical variable that accounts for home-team-specific effects also captures dependencies with respect to city characteristics such as market size, income, and demographic structure variables (Tainsky & Winfree, 2010). Similarly, team-specific ticket prices do not vary substantially from season to season and home-team effects also account for differences in ticket prices across teams (Beckman et al., 2012). The distance between ball parks is defined with respect to their longitude and latitude coordinates as the geodetic ellipsoidal distance using Vincenty's (1975) equations. In addition to a dummy variable for fireworks during a game, I include two additional distinctive but not mutually exclusive promotion categories: Bobblehead promotions are found to have a significant effect on attendance in MLB (Kappe et al., 2014; Siegfried & Eisenberg, 1980); therefore, I include a dummy variable to account for their impact. An additional dummy variable captures all other promotions, e.g. kids' days, autograph signing events, and free T-shirts or other giveaways.

A home team's implied winning probability is calculated from the historic betting odds (money line) that are taken from covers.com. A negative money line ($Ml < -100$) results in an implied winning probability (WP) of greater 50%, which is calculated as $WP = (Ml/(M - 100))$. A positive money line ($Ml > 100$) results in a WP that is smaller 50%, which is calculated as $WP = (Ml/(Ml + 100))$. However, although betting odds are commonly used to approximate a home team's winning probability, they are not equivalent to the winning probability and betting odds may inhibit several biases (Coates & Humphreys, 2012; Forrest & Simmons, 2002; Tainsky & Winfree, 2010). Moreover, I assign the same home-team winning probabilities for second games of included double-headers as were retrieved and computed for the corresponding first games. However, the vast majority of double-headers in the data are the result of rescheduled games, which I do not include in this analysis.

All variables that account for relative within-season team performance are computed such that they include the outcome of the last game that was scheduled on the day before game day, e.g. games behind (GB). GB is a popular measure that accounts for the differences in relative team success between a leading team L and another team i at

time t. I define games behind with respect to a team's assigned division d and compute it as $B_{i,t,d} = \left(\left(\sum_{t=1}^{t} Win_{L,t,d} - \sum_{t=1}^{t} Win_{i,t,d}\right) + \left(\sum_{t=1}^{t} Loss_{i,t,d} - \sum_{t=1}^{t} Loss_{L,t,d}\right)\right)/2$. It follows that a leading team's GB equals zero. However, the leading team is defined in terms of the highest (positive) difference between wins ($\sum_{t=1}^{t} Win_{L,t}$) and losses ($\sum_{t=1}^{t} Loss_{L,t}$) at time t. Hence, GB does not take into account the number of remaining games in the season and several teams of the same division can show a GB of zero at the same time.

Home-team division and interleague rivalry data are taken from Beckman et al. (2012) and Lemke et al. (2010). Assignment of division rivals is not constrained to be symmetric and, furthermore, I make two adjustments due to changes in teams' assigned divisions over time. CIN changed their division in 2008 and HOU their league and division in 2013. Both teams have no assigned division rivals in Lemke et al. (2010) and HOU and TEX are still interleague rivals in 2012. Therefore, I define HOU's former interleague rival TEX as their division rival and vice versa. Following the MLB attendance literature, I also include a categorical variable that accounts for games against BOS, CHC, or NYY (e.g. Beckman et al., 2012; Lemke et al., 2010). The precise division and interleague rivalry mapping is presented in Table A.2.

Table A.2 MLB home-team names and team rivalries

Team	Home-team name	Division rivals	Interleague rival
ARI	Arizona Diamond Backs	COL	-
ATL	Atlanta Braves	NYM, MIA	-
BAL	Baltimore Orioles	NYY, BOS	WAS
BOS	Boston Red Sox	NYY	-
CHC	Chicago Cubs	MIL, STL	CHW
CHW	Chicago White Sox	CLE, DET	CHC
CIN	Cincinnati Reds	-	CLE
CLE	Cleveland Indians	DET	CIN
COL	Colorado Rockies	ARI	-
DET	Detroit Tigers	CLE	-
HOU	Houston Astros	TEX	-
KCR	Kansas City Royals	-	STL
LAA	Los Angeles Angels	OAK	LAD
LAD	Los Angeles Dodgers	SFG	LAA
MIA	Miami Marlins	ATL	TBR
MIL	Milwaukee Brewers	CHC	MIN
MIN	Minnesota Twins	CLE	MIL
NYM	New York Mets	ATL, PHI	NYY
NYY	New York Yankees	BOS	NYM
OAK	Oakland Athletics	LAA	SFG
PHI	Philadelphia Phillies	NYM	-
PIT	Pittsburgh Pirates	-	-
SDP	San Diego Padres	-	-
SEA	Seattle Mariners	-	-
SFG	San Francisco Giants	LAD	OAK
STL	St. Louis Cardinals	CHC	KCR
TBR	Tampa Bay Rays	-	MIA
TEX	Texas Rangers	HOU	-
WAS	Washington Nationals	-	BAL

Notes: The home-team division and interleague rivalry data are obtained from Lemke et al. (2010) and Beckman et al. (2012). Assignment of division rivals is not constrained to be symmetric and I made two adjustments due to changes in teams' assigned divisions over time. CIN changed their division in 2008 and HOU their league and division in 2013. Both teams have no assigned division rivals in Lemke et al. (2010). HOU and TEX are still interleague rivals in 2012. Therefore, I define HOU's former interleague rival TEX as their division rival and vice versa.

2.2 Descriptive statistics

This section shows a list of the included predictor variables, their precise encodings, and the corresponding summary statistics in Table A.3.

Table A.3 Variable specifications and descriptive statistics

Variable	Value	Description	Mean	St. Dev	Min	Max
Dependent variable						
GAttend[a]	-	Game attendance (as ticket sales)	30283	9609	8701	53509
Variables observed before a season starts						
HT.NoG	-	HT's number of games (within season)	82	46	2	163
Weekday[a]	1	Monday	0.10	0.30	0	1
	2	Tuesday	0.15	0.36	0	1
	3	Wednesday	0.16	0.36	0	1
	4	Thursday	0.11	0.31	0	1
	5	Friday	0.16	0.37	0	1
	6	Saturday	0.16	0.37	0	1
	7	Sunday	0.16	0.37	0	1
Month[a]	1	March / April	0.14	0.35	0	1
	2	May	0.18	0.38	0	1
	3	June	0.17	0.38	0	1
	4	July	0.16	0.37	0	1
	5	August	0.18	0.38	0	1
	6	September / October	0.17	0.38	0	1
Night[a]	1	During the night	0.68	0.47	0	1
Pholiday	1	Labor Day / 4th of July / Memorial Day	0.02	0.13	0	1
CTeams	1	1 Team in HT's City/County	0.86	0.34	0	1
	2	2+ Teams in HT's City/County	0.14	0.34	0	1
Capacity[b]	-	Stadium capacity	42980	5037	31042	55500
SType[b]	1	Open stadium	0.79	0.41	0	1
	2	Dome	0.04	0.18	0	1
	3	Retractable roof	0.18	0.38	0	1
SBuild[b]	1	Stadium is 0-5 years old	0.15	0.36	0	1
	2	Stadium is 6-10 years old	0.77	0.42	0	1
	3	Stadium is 10+ years old	0.08	0.27	0	1
ILGame[c]	1	Interleague game	0.12	0.33	0	1
DivGame[c]	1	Division game	0.47	0.50	0	1
DRgame[d]	1	Division rival game	0.11	0.31	0	1
ILRGame[d]	1	Interleague rival game	0.02	0.13	0	1
VTeam	0	Other VT	0.91	0.29	0	1
	1	VT is BOS	0.03	0.17	0	1
	2	VT is CHC	0.03	0.18	0	1
	3	VT is NYY	0.03	0.17	0	1
HT.WSW[c]	1	HT is last season's WS winner	0.03	0.18	0	1
VT.WSW[c]	1	VT is last season's WS winner	0.03	0.18	0	1
HT.LCSW[c]	1	HT is last season's LCS winner	0.07	0.25	0	1
VT.LCSW[c]	1	VT is last season's LCS winner	0.07	0.25	0	1

Variable	Value	Description	Mean	St. Dev	Min	Max
HT.DSW[c]	1	HT is last season's DS winner	0.14	0.34	0	1
VT.DSW[c]	1	VT is last season's DS winner	0.13	0.34	0	1
Distance[b]	-	Between stadiums (in miles)	995	698	7	2732
FWorks[c]	1	Fireworks promotion	0.08	0.27	0	1
BHeads[c]	1	Bobblehead promotion	0.05	0.22	0	1
OPromo[c]	1	Other promotion	0.66	0.47	0	1
DHeader[a]	0	Regular game	0.01	0.11	0	1
	1	First game of a double-header	0.01	0.11	0	1
	2	Second game of a double-header	0.00	0.01	0	1
Variables observed as a season progresses						
Lag.GAttend[a]	-	Lagged HT's game attendance	30405	9672	8701	53518
HT.Wprob[e]	-	HT's winning probability	0.55	0.08	0.252	0.780
HT.GB	-	HT games behind	6.91	7.16	0	44
VT.GB	-	VT games behind	6.85	7.14	0	43
HT.Wper	-	HT's winning percentage	0.50	0.09	0	1
VT.Wper	-	VT's winning percentage	0.50	0.09	0	1
Humidity[f]	-	Humidity	0.67	0.14	0.07	0.95
TempMax[f]	-	Maximal measured temperature	24.782	6.523	-1.867	44.439
TempMin[f]	-	Minimal measured temperature	16.501	6.130	-11.000	32.711
Weather[f]	1	Clear day	0.46	0.50	0	1
	2	Cloudy day	0.31	0.46	0	1
	3	Snowy day	0.00	0.04	0	1
	4	Rainy day	0.16	0.37	0	1
	5	Windy day	0.06	0.23	0	1
	6	Foggy day	0.01	0.09	0	1
Precip[f]	1	Precipitation	0.23	0.42	0	1
Season[f]	1	2013	0.33	0.47	0	1
	2	2014	0.33	0.47	0	1
	3	2015	0.33	0.47	0	1

Notes: The data sample covers 6852 games from the 2013, 2014, and 2015 MLB regular seasons for all 29 US teams. First home team (HT)-specific season games and rescheduled games are not included (see Section 2.1 for a detailed description of the data cleaning process). Each HT's implied winning probability is calculated from betting odds. Data sources: aRetrosheet.org, bSeahmheads.com, cMBL.com, dLemke et al. (2010), Beckman et al. (2012), ecovers.com, and fdarksky.net (API).

2.3 Variable importance ranking and predictor correlations

To compare and assess the observed significance of the differences in the variable importance rankings that are produced by CF, RF, and CCF for the static pre- and within-season forecasts, I follow Lessmann et al. (2010) and compute the corresponding ranking correlation coefficients by means of Kendall's τ. Table A.4 shows all correlation coefficients and the associated p-values for the inter- and intra-season comparisons of the RF, CF, and CCF rankings.

Table A.4 Correlation between variable importance rankings by Kendall's tau.

(a) Intra-season model correlation

			Within-season
Ranking	RF	CF	CCF
RF	1	0.582***	0.083
		(0.000)	(0.567)
CF	0.471***	1	0.145
	(0.001)		(0.311)
CCF	0.551***	0.440***	1
	(0.000)	(0.002)	
	Pre-season		

(b) Inter-season model correlation

Ranking	Within RF	Within CF	Within CCF
Pre RF	0.797***	0.465***	0.015
	(0.000)	(0.001)	(0.930)
Pre CF	0.526***	0.871***	0.102
	(0.000)	(0.000)	(0.481)
Pre CCF	0.422***	0.397***	-0.114
	(0.003)	(0.005)	(0.428)

Notes: Kendall's rank correlation coefficient (Kendall's τ) for variable importance ranking is derived from OOB estimates of random forest and conditional random forest regressions for US home-team-specific MLB game attendance for 4571 games of the regular 2013 and 2014 seasons as a training set. Maximal complex forests are trained using $B = 500$ trees for $M_{ps} = 7$ [$M_{ws} = 12$] randomly chosen predictors at each node of the $P_{ps} = 26$ [$P_{ws} = 37$] included predictors for the pre-season [within-season] model (Hothorn et al., 2015; Liaw & Wiener, 2002). The dynamic month-ahead approach includes an additional categorical variable that accounts for seasonal differences in game attendance. The results show the rankings of predictors' relevance in the forest building process for the permutation importance measures of the biased RF (scaled mean decrease in MSE), the CF (Strobl et al., 2007), and the conditional CF (CCF) approaches (Strobl et al., 2008). *** $p < 0.01$.

The WS CCF ranking is reported to be statistically significantly different from all other PS and WS rankings at a minimum p-value of 0.311. All other combinations of differences in importance rankings across models for both PS and WS are not significantly different from each other. Moreover, I note that for the inter-season rank comparison, only the rankings of the PS variables are compared to the relative ranks of the 27 variables in the WS rankings.

The main text only shows the ten most important pre- and within-season predictors for the static forecasting approach. The complete variable importance rankings for the static and dynamic RF and CF permutation importance measures are reported in Table A.5.

Table A.5 Random forest and conditional random forest variable importance rankings.

		Pre		Within		Month-ahead											
						Apr		May		Jun		Jul		Aug		Sep	
#	Variable	RF	CF	RF	CF	RF	CF	RF	CF	RF	CF	RF	CF	RF	CF	RF	CF
1	WDay	1	3	1	3	1	3	1	3	1	3	1	3	1	3	1	3
2	HT.id	2	1	2	1	2	1	2	1	2	2	2	2	2	2	2	2
3	HT.NoG	3	9	4	14	4	14	4	13	4	13	4	14	4	13	4	13
4	Distance	4	14	10	23	11	23	10	24	11	24	8	24	12	23	8	23
5	Month	5	7	13	9	12	10	8	9	8	11	11	9	7	10	12	10
6	VTeam	6	13	16	16	15	19	16	19	14	18	14	15	15	18	13	18
7	FWorks	7	12	7	7	5	8	6	7	5	8	6	7	9	7	10	7
8	Capacity	8	2	8	4	8	4	9	4	9	4	12	4	10	4	9	4
9	Night	9	15	14	13	16	13	12	14	13	14	13	13	13	12	15	11
10	VT.DSW	10	18	22	22	25	24	25	23	26	23	25	25	25	25	26	29
11	OPromo	11	10	18	12	19	11	18	12	19	12	19	12	19	14	19	12
12	BHeads	12	16	11	15	13	15	13	15	15	15	15	16	14	15	14	17
13	SBuild	13	6	23	8	23	7	23	8	25	7	23	8	23	8	23	8
14	DivGame	14	19	32	30	30	31	31	31	36	32	30	30	32	30	29	30
15	HT.DSW	15	5	19	5	20	5	22	5	21	6	20	6	21	6	21	5
16	SType	16	4	24	6	24	6	24	6	23	5	24	5	24	5	24	6
17	ILGame	17	20	27	28	32	27	35	29	34	29	31	28	29	29	33	28
18	VT.LCSW	18	24	31	35	28	36	32	35	28	35	29	36	34	35	31	33
19	CTeams	19	11	25	21	27	22	28	22	27	20	26	21	27	21	27	21
20	DRGame	20	22	29	32	34	32	34	34	32	37	34	33	36	34	34	35
21	ILRGame	21	21	26	29	26	30	26	32	24	30	27	32	26	32	25	31
22	Pholiday	22	23	34	34	37	34	30	33	33	33	35	35	31	37	35	36
23	HT.LCSW	23	8	28	11	29	12	29	11	29	9	28	11	28	11	28	14
24	VT.WSW	24	25	33	36	31	37	33	37	31	36	32	34	33	31	32	37
25	HT.WSW	25	17	35	25	33	25	36	25	35	26	37	23	37	24	37	24
26	DHeader	26	26	37	37	38	38	38	38	38	38	38	38	38	38	38	38
27	lag.GAttend	-	-	3	2	3	2	3	2	3	1	3	1	3	1	3	1
28	TempMax	-	-	5	20	10	20	5	20	7	21	10	20	6	20	5	20
29	HT.GB	-	-	6	17	6	16	14	16	12	17	7	18	11	16	11	15
30	TempMin	-	-	9	18	7	17	7	17	6	19	9	19	8	19	7	19
31	HT.Wper	-	-	12	10	9	9	11	10	10	10	5	10	5	9	6	9
32	VT.Wper	-	-	15	19	14	21	15	21	16	22	18	22	16	22	16	22
33	VT.GB	-	-	17	26	18	29	17	27	18	27	17	29	17	28	17	27
34	HT.Wprob	-	-	20	24	17	28	20	28	22	25	21	26	20	26	22	26
35	Humidity	-	-	21	27	22	26	21	26	20	28	22	27	22	27	20	25
36	Weather	-	-	30	33	35	35	27	30	30	31	33	31	30	33	30	34
37	Precip	-	-	36	31	36	33	37	36	37	34	36	37	35	36	36	32
38	Season	-	-	-	-	21	18	19	18	17	16	16	17	18	17	18	16

Notes: Variable importance rankings are derived from OOB estimates of RF and CF regressions for US home-team-specific MLB game attendance for 4571 games of the regular 2013 and 2014 seasons as a training set. Maximal complex forests are trained using $B = 500$ trees for $M_{ps} = 7$ [$M_{ws} = 12$] randomly chosen predictors at each node of the $P_{ps} = 26$ [$P_{ws} = 37$] included predictors for the pre-season [within-season] model (Hothorn et al., 2015; Liaw & Wiener, 2002). The dynamic month-ahead approach includes an additional categorical variable that accounts for seasonal differences in game attendance. The results show the rankings of predictors' relevance in the forest building process for the permutation importance measures of the RF (scaled mean decrease in MSE) and the CF approaches (Strobl et al., 2007).

Lastly, Table A.6 shows the linear correlations between all numeric predictor variables that are employed in this study.

Table A.6 Correlations between numeric predictor variables and game attendance.

Variables	GAttend	Lag.GAttend	Distance	HT.Wprob	HT.GB	VT.GB	HT.Wper	VT.Wper	HT.NoG	Humidity	TempMax	TempMin	Precip
GAttend	1.000												
Lag.GAttend	0.773	1.000											
Distance	-0.050	-0.042	1.000										
HT.Wprob	0.181	0.188	0.005	1.000									
HT.GB	-0.235	-0.251	-0.016	-0.400	1.000								
VT.GB	-0.008	-0.006	-0.021	0.405	0.135	1.000							
HT.Wper	0.264	0.268	-0.011	0.390	-0.580	0.069	1.000						
VT.Wper	0.032	0.007	-0.013	-0.372	0.084	-0.566	-0.285	1.000					
HT.NoG	0.062	0.027	-0.021	0.004	0.449	0.444	-0.017	0.009	1.000				
Humidity	-0.029	-0.040	-0.011	0.030	0.031	0.066	0.034	-0.005	0.141	1.000			
TempMax	-0.021	-0.029	0.033	-0.033	0.179	0.159	-0.036	0.003	0.396	-0.216	1.000		
TempMin	-0.075	-0.090	0.071	-0.039	0.215	0.174	-0.076	0.007	0.442	0.081	0.853	1.000	
Precip	-0.088	-0.103	-0.078	-0.013	-0.013	-0.036	-0.003	0.020	-0.071	0.405	-0.093	0.039	1.000

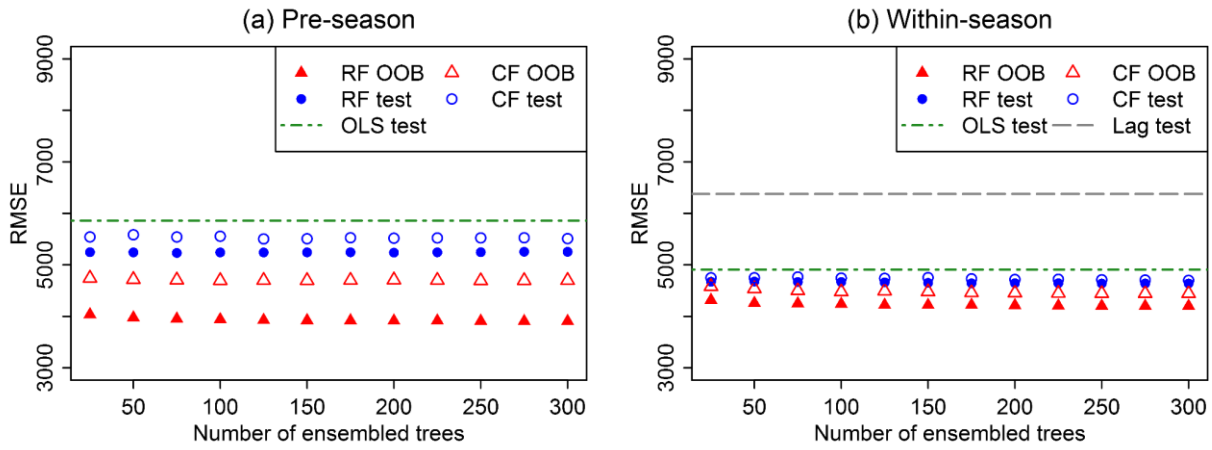
Notes: Correlations between the 13 numeric variables that are employed in this study (see Table A1). Data are based on 6852 individual MLB games from the 2013, 2014, and 2015 regular seasons.

3 Model performance evaluation

A popular approach in machine learning model tuning is a systematic grid-search over specific hyper-parameters (Hamza & Larocque, 2005; Lessmann et al., 2010). However, I quickly observed that the predictive performances of both the RF and CF approaches for the pre- and within-season models are not very sensitive to the number of trees per forest. The model performance evaluation in terms of the number of randomly considered predictors at each split is presented in the main paper in Section 3.1.

Figure A.1 shows the predictive accuracy in terms of RMSE on the OOB and test samples for the RF and CF regressions, together with the OLS and naive home-team-specific lagged attendance forecasts (Lag) as a benchmark. The number of randomly chosen predictors in the tree building process, which is denoted as M, is set to the suggested default value (one third) and RMSE is reported as the number of trees per forest $B = \{25, 50, \dots, 300\}$. The corresponding results show that RF yields the most accurate results and RF and CF outperform the OLS model for both the pre-season (a) and the within–season forecasts (b). The OLS model yields RMSEs of 5858 (a) and 4908 (b), while the naive HT-specific lagged attendance forecast (Lag) results in an RMSE of 6377 (b).

Figure A.1 Model performance evaluation: RMSE by number of ensembled trees



Notes: Out-of-sample MLB attendance predictions by the CART (RF), conditional inference random forests (CF), OLS, and lagged attendance (Lag) models for 2281 games of the regular season of 2015 are used as a test set and 6852 games of the regular 2013 and 2014 seasons as a training set. OOB refers to a forest's predictive performance on the out-of-bag (OOB) data. Maximal complex RF and CF are trained using $\mathrm{M_{ps}} = 7$ $[M_{ws} = 12]$ randomly chosen predictors at each node. The pre-season [within-season] model includes $P_{ps} = 26$ $[P_{ws} = 37]$ predictors and I grow $B = \{25, 50, \ldots, 300\}$ trees per forest.

The RMSEs for both the RF and CF approaches on the OOB and test data stabilize after averaging the prediction results of 50 trees. For (a), the RF (CF) yields minimum RMSEs of 3906 (4693) on the OOB data and 5231 (5503) on the test data. For (b), the RF (CF) yields minimum RMSEs of 4201 (4444) on the OOB data and 4638 (4670) on the test data. While the differences in prediction accuracy across models are stronger for the pre-season forecast, they do not vary substantially when trained with the additional information that is provided by the within-season variables. Moreover, the RF appears to be more affected by issues that are associated with overfitting to the training data, thereby resulting in a low RMSE on the OOB data in (a), which is adjusted based on the additional within-season information in (b). In contrast, the differences in prediction accuracy between the OOB and test data are smaller for the CF approach.

4 Pre- and within-season random forest predictions

Section 5 shows the team-specific RF and CF results that are omitted in the main text. Table A.7 shows the resulting prediction accuracy for the dynamic within-season RF approach and Table A.8 shows the static and dynamic CF forecasting results.

Table A.7 Random forest predictions and attendance summary statistics by month and team

	Season 2015			Out-of-sample monthly step-ahead RMSE					
	Attendance summary			Random forest					
HT	N	Mean	SD	Apr	May	Jun	Jul	Aug	Sep
All	2281	30197	9515	4608	4481	3994	4373	4424	4405
ARI	80	25389	7504	3989	4668	3255	5353	3454	4724
ATL	79	24748	8298	4498	3436	4567	6544	7265	6987
BAL	72	30001	8746	5140	6597	4041	6077	5991	6726
BOS	79	35572	1709	1987	2047	1736	1503	1114	1374
CHC	76	36467	4210	1523	2583	4459	4428	3219	2776
CHW	78	21687	7391	2233	4910	3423	3847	5253	4182
CIN	77	29568	7778	5264	4513	3562	4531	5893	4700
CLE	77	17573	5783	3570	3144	2848	3583	3880	5090
COL	76	31341	6719	4559	4461	3017	6042	5076	4063
DET	79	33576	4648	2660	3221	3202	2512	2103	3592
HOU	80	26373	6622	1843	2253	5263	6259	4724	4233
KCR	79	33422	5061	5614	5690	4878	4694	2880	2765
LAA	79	37092	5085	4954	4834	4302	3448	2020	3175
LAD	80	46391	4242	4564	3834	3838	3930	3106	3755
MIA	80	21441	4439	3269	3704	4649	4807	3347	3330
MIL	80	31207	5795	3693	4518	3932	3275	4238	5207
MIN	79	27173	6134	5391	4280	4581	3030	3602	4433
NYM	79	31447	7151	8807	4828	2549	4608	6022	4651
NYY	79	39814	4983	5265	3884	3640	4102	4009	4419
OAK	80	21651	6461	3985	3586	4015	4636	5869	5694
PHI	77	23189	4564	5911	6177	5303	4107	3128	6591
PIT	78	30744	7163	4216	4643	4031	2705	3350	3566
SDP	79	30287	7795	7213	6464	4590	5047	5465	5221
SEA	80	26846	8984	7167	8214	6446	6564	3958	3750
SFG	80	41673	387	602	331	343	627	237	230
STL	79	43380	1957	2450	1912	1979	1366	1190	1855
TBR	83	15133	4940	3426	5335	2779	3984	4043	4203
TEX	80	30537	6412	5281	4147	4593	5088	7459	5235
WAS	77	32453	5351	2232	4120	4375	4571	3144	3884
R2	-	-	-	0.790	0.778	0.798	0.755	0.761	0.804
Monthly season 2015 attendance summary									
N	6852	6852	0	287	408	386	362	402	436
Mean	-	-	-	28154	29639	30748	33006	31158	28355
SD	-	-	-	10055	9519	8879	8842	9057	9945

Notes: The out-of-sample month-ahead prediction accuracies for US home-team-specific MLB game attendance for 2281 games of the regular season 2015 are used as a test set. The 4571 games of the regular 2013 and 2014 seasons are used as a training set that is updated after each month. Maximal complex random and conditional forests are trained using $B = 500$ trees for M_{ws} =12 randomly chosen predictors at each node of the $P_{ws} = 38$ included predictors for the dynamic within-season forecast (Hothorn et al., 2015; Liaw & Wiener, 2002).

Table A.8 Static and dynamic conditional random forest predictions by month and team.

Conditional random forest out-of-sample RMSEs									
	Static forecast			Dynamic monthly step-ahead forecast					
HT	Pre	Within	Diff	Apr	May	Jun	Jul	Aug	Sep
All	5523	4705	818	4780	4655	4198	4589	4523	4519
ARI	4793	4523	270	4183	4790	3258	5622	3747	4710
ATL	7516	6436	1080	5968	3892	4852	6932	7638	7005
BAL	6589	6213	376	4858	6897	4393	5987	6292	6938
BOS	1886	1411	475	1566	1545	1548	1548	1055	1409
CHC	5116	3619	1497	2264	2409	4444	4162	3347	2853
CHW	5176	4603	573	3180	4997	3761	3834	5682	4467
CIN	5400	4988	412	5878	4692	3564	4588	5597	4869
CLE	4922	4016	906	4138	3422	3032	3956	4518	5650
COL	5874	4809	1065	4581	4857	2928	6195	5056	4246
DET	3493	3208	285	3934	3645	3262	2468	2119	3545
HOU	7059	4770	2289	1603	2176	5324	6475	4774	4616
KCR	4976	5480	-504	5746	6655	5407	5544	3808	3414
LAA	4438	4079	359	4792	5277	4564	3647	2292	3389
LAD	4238	3886	352	4567	3946	4061	4032	3076	3747
MIA	4051	3887	164	3037	4018	4625	5177	3395	3130
MIL	4343	4600	-257	4003	4757	4058	3573	4305	5577
MIN	5699	4434	1265	5788	4507	4569	3362	3873	4367
NYM	7536	6387	1149	8156	4981	2709	4826	5680	5185
NYY	4720	4251	469	5232	4161	3709	4207	3979	4100
OAK	5135	5096	39	4862	3113	4194	5179	6257	6163
PHI	10259	6400	3859	6185	6283	5494	4683	3235	6112
PIT	4684	4196	488	5217	4878	4498	3624	3591	3914
SDP	6533	5890	643	6659	6660	4528	5455	5600	5501
SEA	7317	6580	737	6595	8240	7054	6535	3942	3912
SFG	772	380	392	375	307	378	565	292	355
STL	2205	1697	508	2132	1641	2016	1413	1259	1469
TBR	4625	4095	530	3227	5046	3303	4070	4254	3800
TEX	7356	5868	1488	6357	4691	5205	4908	6616	5394
WAS	4260	4047	213	2849	4016	4790	4907	3616	3843
R2	0.663	0.755	-0.092	0.773	0.76	0.776	0.73	0.75	0.793
Monthly season 2015 attendance summary									
N	-	-	-	287	408	386	362	402	436
Mean	-	-	-	28154	29639	30748	33006	31158	28355
SD	-	-	-	10055	9519	8879	8842	9057	9945

Notes: The out-of-sample month-ahead prediction accuracies for US home-team-specific MLB game attendance for 2281 games of the regular season 2015 are used as a test set. The 4571 games of the regular 2013 and 2014 seasons are used as a training set that is updated after each month. Maximal complex forests are trained using $B = 500$ trees for $M_{ps} = 7$ [$M_{ws} = 12$] randomly chosen predictors at each node of the $P_{ps} = 26$ [$P_{ws} = 37$] included predictors for the pre-season [within-season] model (Hothorn et al., 2015). The dynamic month-ahead approach includes an additional categorical variable that accounts for seasonal differences in game attendance.

References

Beckman, E. M., Cai, W., Esrock, R. M., & Lemke, R. J. (2012). Explaining Game-to-Game Ticket Sales for Major League Baseball Games Over Time. *Journal of Sports Economics*, *13*(5), 536–553.

Coates, D., & Humphreys, B. R. (2012). Game Attendance and Outcome Uncertainty in the National Hockey League. *Journal of Sports Economics*, *13*(4), 364–377.

Forrest, D., & Simmons, R. (2002). Outcome uncertainty and attendance demand in sport: The case of English soccer. *Journal of the Royal Statistical Society Series D: The Statistician*, *51*(2), 229–241.

Hamza, M., & Larocque, D. (2005). An empirical comparison of ensemble methods based on classification trees. *Journal of Statistical Computation and Simulation*, *75*(8), 629–643.

Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. (2015). party: A Laboratory for Recursive Partytioning. *R Package Version 0.9-0*, (1994), 37.

Kappe, E., Stadler Blank, A., & DeSarbo, W. S. (2014). A general multiple distributed lag framework for estimating the dynamic effects of promotions. *Management Science*, *60*(6), 1489–1510.

Lemke, R. J., Leonard, M., & Tlhokwane, K. (2010). Estimating Attendance at Major League Baseball Games for the 2007 Season. *Journal of Sports Economics*, *11*(3), 316–348.

Lessmann, S., Sung, M. C., & Johnson, J. E. V. (2010). Alternative methods of predicting competitive events: An application in horserace betting markets. *International Journal of Forecasting*, *26*(3), 518–536.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(December), 18–22.

Siegfried, J. J., & Eisenberg, J. D. (1980). The demand for minor league baseball. *Atlantic Economic Journal*, *8*(2), 59–69.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*, 1–11.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable

importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(25).

Tainsky, S., & Winfree, J. A. (2010). Short-Run Demand and Uncertainty of Outcome in Major League Baseball. *Review of Industrial Organization*, *37*(3), 197–214.

Vincenty, T. (1975). Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review*, *23*(176), 88–93.

Hamburg Contemporary Economic Discussions

(Download: https://www.wiso.uni-hamburg.de/en/fachbereich-vwl/professuren/maennig/research/hceds.html)

65 MUELLER, S. Q.: Pre- and within-season attendance forecasting in Major League Baseball: A random forest approach, 2018.

64 KRUSE, F. K. / MAENNIG, W.: Suspension by choice – determinants and asymmetries, 2018.

63 GROTHE, H. / MAENNIG, W.: A 100-million-dollar fine for Russia's doping policy? A billion-dollar penalty would be more correct! Millionenstrafe für Russlands Doping-Politik? Eine Milliarden-Strafe wäre richtiger! 2017.

62 MAENNIG, W., / SATTARHOFF, C. / STAHLECKER, P.: Interpretation und mögliche Ursachen statistisch insignikanter Testergebnisse - eine Fallstudie zu den Beschäftigungseffekten der Fußball-Weltmeisterschaft 2006, 2017.

61 KRUSE, F. K. / MAENNIG, W.: The future development of world records, 2017.

60 MAENNIG, W.: Governance in Sports Organizations, 2017.

59 AHLFELDT, G. M. / MAENNIG, W. / FELIX J. RICHTER: Zoning in reunified Berlin, 2017.

58 MAENNIG, W.: Major Sports Events: Economic Impact, 2017.

57 MAENNIG, W.: Public Referenda and Public Opinion on Olympic Games, 2017.

56 MAENNIG, W. / WELLBROCK, C.: Rio 2016: Sozioökonomische Projektion des Olympischen Medaillenrankings, 2016.

55 MAENNIG, W. / VIERHAUS, C.: Which countries bid for the Olympic Games? Economic, political, and social factors and chances of winning, 2016.

54 AHLFELDT, G. M. / MAENNIG, W. / STEENBECK, M.: Après nous le déluge? Direct democracy and intergenerational conflicts in aging societies, 2016.

53 LANGER, V. C. E.: Good news about news shocks, 2015.

52 LANGER, V. C. E. / MAENNIG, W. / RICHTER, F. J.: News Shocks in the Data: Olympic Games and their Macroeconomic Effects – Reply, 2015.

51 MAENNIG, W.: Ensuring Good Governance and Preventing Corruption in the Planning of Major Sporting Events – Open Issues, 2015.

50 MAENNIG, W. / VIERHAUS, C.: Who Wins Olympic Bids? 2015 (3rd version).

Hamburg Contemporary Economic Discussions

(Download: https://www.wiso.uni-hamburg.de/en/fachbereich-vwl/professuren/maennig/research/hceds.html)

49 AHLFELDT, G. M. / MAENNIG, W. / RICHTER, F.: Urban Renewal after the Berlin Wall, 2013.

48 BRANDT, S. / MAENNIG, W. / RICHTER, F.: Do Places of Worship Affect Housing Prices? Evidence from Germany, 2013.

47 ARAGÃO, T. / MAENNIG, W.: Mega Sporting Events, Real Estate, and Urban Social Economics – The Case of Brazil 2014/2016, 2013.

46 MAENNIG, W. / STEENBECK, M. / WILHELM, M.: Rhythms and Cycles in Happiness, 2013.

45 RICHTER, F. / STEENBECK, M. / WILHELM, M.: The Fukushima Accident and Policy Implications: Notes on Public Perception in Germany, 2014 (2nd version).

44 MAENNIG, W.: London 2012 – das Ende des Mythos vom erfolgreichen Sportsoldaten, 2012.

43 MAENNIG, W. / WELLBROCK, C.: London 2012 – Medal Projection – Medaillenvorausberechnung, 2012.

42 MAENNIG, W. / RICHTER, F.: Exports and Olympic Games: Is there a Signal Effect? 2012.

41 MAENNIG, W. / WILHELM, M.: Becoming (Un)employed and Life Satisfaction: Asymmetric Effects and Potential Omitted Variable Bias in Empirical Happiness Studies, 2011.

40 MAENNIG, W.: Monument Protection and Zoning in Germany: Regulations and Public Support from an International Perspective, 2011.

39 BRANDT, S. / MAENNIG, W.: Perceived Externalities of Cell Phone Base Stations – The Case of Property Prices in Hamburg, Germany, 2011.

38 MAENNIG, W. / STOBERNACK, M.: Do Men Slow Down Faster than Women? 2010.

37 DU PLESSIS, S. A. / MAENNIG, W.: The 2010 World Cup High-frequency Data Economics: Effects on International Awareness and (Self-defeating) Tourism, 2010.

36 BISCHOFF, O.: Explaining Regional Variation in Equilibrium Real Estate Prices and Income, 2010.

Hamburg Contemporary Economic Discussions

(Download: https://www.wiso.uni-hamburg.de/en/fachbereich-vwl/professuren/maennig/research/hceds.html)

35 FEDDERSEN, A. / MAENNIG, W.: Mega-Events and Sectoral Employment: The Case of the 1996 Olympic Games, 2010.

34 FISCHER, J.A.V. / SOUSA-POZA, A.: The Impact of Institutions on Firms Rejuvenation Policies: Early Retirement with Severance Pay versus Simple Lay-Off. A Cross-European Analysis, 2010.

33 FEDDERSEN, A. / MAENNIG, W.: Sectoral Labor Market Effects of the 2006 FIFA World Cup, 2010.

32 AHLFELDT, G.: Blessing or Curse? Appreciation, Amenities, and Resistance around the Berlin "Mediaspree", 2010.

31 FALCH, T. / FISCHER, J.A.V.: Public Sector Decentralization and School Performance: International Evidence, 2010.

30 AHLFELDT, G. / MAENNIG, W. / ÖLSCHLÄGER, M.: Lifestyles and Preferences for (Public) Goods: Professional Football in Munich, 2009.

29 FEDDERSEN, A. / JACOBSEN, S. / MAENNIG, W.: Sports Heroes and Mass Sports Participation – The (Double) Paradox of the "German Tennis Boom", 2009.

28 AHLFELDT, G. / MAENNIG, W. / OSTERHEIDER, T.: Regional and Sectoral Effects of a Common Monetary Policy: Evidence from Euro Referenda in Denmark and Sweden, 2009.

27 BJØRNSKOV, C. / DREHER, A. / FISCHER, J.A.V. / SCHNELLENBACH, J.: On the Relation Between Income Inequality and Happiness: Do Fairness Perceptions Matter? 2009.

26 AHLFELDT, G. / MAENNIG, W.: Impact of Non-Smoking Ordinances on Hospitality Revenues: The Case of Germany, 2009.

25 FEDDERSEN, A. / MAENNIG, W.: Wage and Employment Effects of the Olympic Games in Atlanta 1996 Reconsidered, 2009.

24 AHLFELDT, G. / FRANKE, B. / MAENNIG, W.: Terrorism and the Regional and Religious Risk Perception of Foreigners: The Case of German Tourists, 2009.

Hamburg Contemporary Economic Discussions

(Download: https://www.wiso.uni-hamburg.de/en/fachbereich-vwl/professuren/maennig/research/hceds.html)

23 AHLFELDT, G. / WENDLAND, N.: Fifty Years of Urban Accessibility: The Impact of Urban Railway Network on the Land Gradient in Industrializing Berlin, 2008.

22 AHLFELDT, G. / FEDDERSEN, A.: Determinants of Spatial Weights in Spatial Wage Equations: A Sensitivity Analysis, 2008.

21 MAENNIG, W. / ALLMERS, S.: South Africa 2010: Economic Scope and Limits, 2008.

20 MAENNIG, W. / WELLBROCK, C.-M.: Sozio-ökonomische Schätzungen Olympischer Medaillengewinne: Analyse-, Prognose- und Benchmark-möglichkeiten, 2008.

19 AHLFELDT, G.: The Train has Left the Station: Real Estate Price Effects of Mainline Realignment in Berlin, 2008.

18 MAENNIG, W. / PORSCHE, M.: The Feel-good Effect at Mega Sport Events – Recommendations for Public and Private Administration Informed by the Experience of the FIFA World Cup 2006, 2008.

17 AHLFELDT, G. / MAENNIG, W.: Monumental Protection: Internal and External Price Effects, 2008.

16 FEDDERSEN, A. / GRÖTZINGER, A. / MAENNIG, W.: New Stadia and Regional Economic Development – Evidence from FIFA World Cup 2006 Stadia, 2008.

15 AHLFELDT, G. / FEDDERSEN, A.: Geography of a Sports Metropolis, 2007.

14 FEDDERSEN, A. / MAENNIG, W.: Arenas vs. Multifunctional Stadia – Which Do Spectators Prefer? 2007.

13 AHLFELDT, G.: A New Central Station for a Unified City: Predicting Impact on Property Prices for Urban Railway Network Extension, 2007.

12 AHLFELDT, G.: If Alonso was Right: Accessibility as Determinant for Attractiveness of Urban Location, 2007.

11 AHLFELDT, G., MAENNIG, W.: Assessing External Effects of City Airports: Land Values in Berlin, 2007.

10 MAENNIG, W.: One Year Later: A Re-Appraisal of the Economics of the 2006 Soccer World Cup, 2007.

Hamburg Contemporary Economic Discussions

(Download: https://www.wiso.uni-hamburg.de/en/fachbereich-vwl/professuren/maennig/research/hceds.html)

09 HAGN, F. / MAENNIG, W.: Employment Effects of the World Cup 1974 in Germany.

08 HAGN, F. / MAENNIG W.: Labour Market Effects of the 2006 Soccer World Cup in Germany, 2007.

07 JASMAND, S. / MAENNIG, W.: Regional Income and Employment Effects of the 1972 Munich Olympic Summer Games, 2007.

06 DUST, L. / MAENNIG, W.: Shrinking and Growing Metropolitan Areas – Asymmetric Real Estate Price Reactions? The Case of German Single-family Houses, 2007.

05 HEYNE, M. / MAENNIG, W. / SUESSMUTH, B.: Mega-sporting Events as Experience Goods, 2007.

04 DU PLESSIS, S. / MAENNIG, W.: World Cup 2010: South African Economic Perspectives and Policy Challenges Informed by the Experience of Germany 2006, 2007.

03 AHLFELDT, G. / MAENNIG, W.: The Impact of Sports Arenas on Land Values: Evidence from Berlin, 2007.

02 FEDDERSEN, A. / MAENNIG, W. / ZIMMERMANN, P.: How to Win the Olympic Games – The Empirics of Key Success Factors of Olympic Bids, 2007.

01 AHLFELDT, G. / MAENNIG, W.: The Role of Architecture on Urban Revitalization: The Case of "Olympic Arenas" in Berlin-Prenzlauer Berg, 2007.

04/2006 MAENNIG, W. / SCHWARTHOFF, F.: Stadium Architecture and Regional Economic Development: International Experience and the Plans of Durban, October 2006.

03/2006 FEDDERSEN, A. / VÖPEL, H.: Staatliche Hilfen für Profifußballclubs in finanziellen Notlagen? – Die Kommunen im Konflikt zwischen Imageeffekten und Moral-Hazard-Problemen, September 2006.

02/2006 FEDDERSEN, A.: Measuring Between-season Competitive Balance with Markov Chains, July 2006.

Hamburg Contemporary Economic Discussions

(Download: https://www.wiso.uni-hamburg.de/en/fachbereich-vwl/professuren/maennig/research/hceds.html)

01/2006 FEDDERSEN, A.: Economic Consequences of the UEFA Champions League for National Championships – The Case of Germany, May 2006.

04/2005 BUETTNER, N. / MAENNIG, W. / MENSSEN, M.: Zur Ableitung einfacher Multiplikatoren für die Planung von Infrastrukturkosten anhand der Aufwendungen für Sportstätten – eine Untersuchung anhand der Fußball-WM 2006, May 2005.

03/2005 SIEVERS, T.: A Vector-based Approach to Modeling Knowledge in Economics, February 2005.

02/2005 SIEVERS, T.: Information-driven Clustering – An Alternative to the Knowledge Spillover Story, February 2005.

01/2005 FEDDERSEN, A. / MAENNIG, W.: Trends in Competitive Balance: Is there Evidence for Growing Imbalance in Professional Sport Leagues? January 2005.

Hamburg

Contemporary Economic Discussions

HCED NO. 65

ISSN 1865-2441 (PRINT)
ISSN 1865-7133 (ONLINE)

ISBN 978-3-942820-44-8(PRINT)
ISBN 978-3-942820-45-5 (ONLINE)