When are they coming? Understanding and forecasting the timeline of arrivals at the FC Barcelona stadium on match days

Feliu Serra-Burriel[1] · Pedro Delicado[2,3] · Fernando M. Cucchietti[4] · Eduardo Graells-Garrido[5] · Alex Gil[6] · Imanol Eguskiza[6]

Abstract

Futbol Club Barcelona operates the largest stadium in Europe (with a seating capacity of almost one hundred thousand people) and manages recurring sports events. These are influenced by multiple conditions (time and day of the week, weather, adversary) and affect city dynamics—e.g., peak demand for related services like public transport and stores. We study fine grain audience entrances at the stadium segregated by visitor type and gate to gain insights and predict the arrival behavior of future games, with a direct impact on the organizational performance and productivity of the business. We can forecast the timeline of arrivals at gate level 72 h prior to kickoff, facilitating operational and organizational decision-making by anticipating potential agglomerations and audience behavior. Furthermore, we can identify patterns for different types of visitors and understand how relevant factors affect them. These findings directly impact commercial and business interests and can alter operational logistics, venue management, and safety.

Keywords Functional data analysis · Count functional data · Generalized functional principal components analysis · Function-on-scalar regression · Functional R-square

1 Introduction

Sports venues around the world gather tens of thousands of people for short scheduled events. In this regard, maximizing stadium attendance is crucial for professional sports and businesses (Schreyer & Ansari, 2021). An underused stadium may lead to inefficient staffing, lower merchandising sales, and lower volumes of other potential sources of revenue. Therefore, stadium attendance is critical for a sport club's external stakeholders (McDonald, 2010), as well as its market value and stock returns, since stadium attendance is a proxy indicator of reputation (Gimet & Montchaud, 2016)—with a direct

---

Extended author information available on the last page of the article

impact on club branding (Richelieu, 2014)—and presents a positive correlation with other sources of revenue (Késenne, 2014).

Stadium attendance has been proven to yield important effects on a team's performance, both in terms of game perception by attendees and player's performance on the field. On the one hand, stadium attendance enhances crowd chanting and improves stadium atmospheres, benefiting both attendees (Morrow, 1999), and home teams (Forrest et al., 2005). On the other hand, low occupancy rates negatively affect the perceived game quality of customers (Oh et al., 2017).

The focus of this research is to model a stadium's attendance during match days and identify visitor arrival patterns. To do so, we build a set of tools combining data and advanced methodologies to help on the organizational and operational needs of our use case, the Camp Nou stadium. Apart from business applications, we aim at improving the mobility and security of the stadium. These potential improvements range from identifying peak agglomeration times and zones, to detecting inefficient or underused gates.

Football attendance demand has been studied previously in the literature (Buraimo, 2014; Schreyer & Ansari, 2021), as well as attendance prediction in football (Reade, 2007; Şahin & Erol, 2018) or other sports (García & Rodríguez, 2009), estimating overall turnouts based on factors such as weather (Cairns, 1984), city size (Walker, 1986), outcome uncertainty (Forrest & Simmons, 2002), home advantage (Forrest et al., 2005), the effect of superstar players (Lawson et al., 2008; Jane, 2016; Humphreys & Johnson, 2020), rivalry (Tyler et al., 2017), team loyalty (Reade, 2020), days of the week (Goller & Krumer, 2020), as well as broader scope reviews of the economics of football (Dobson et al., 2001). Moreover, the estimation of price elasticity and the effect on the demand for football have also been studied (García & Rodríguez, 2002).

However, there is a gap in previous literature regarding the timeline of stadium arrivals. Count data or disaggregated arrivals at stadiums are valuable sources of information for clubs to understand how these are affected by different conditions and how different types of events differ. Previous studies use aggregated data, providing little information on the dynamics of visitor arrivals during match days (Schreyer & Ansari, 2021). Furthermore, few researchers have delved into the information provided by separating match day ticket holders and season ticket holders (Allan & Roy, 2008). To the best of our knowledge, there is no research using this kind of disaggregated data to model and predict attendance time series. The collection of this data in a systematic manner, for long periods to obtain enough observations, requires sophisticated data infrastructures and complex organizational management. However, these are not out of reach for top clubs in most professional sports.

To prepare the operations and the commercial offer required for these kinds of events, the club needs to estimate the number of tickets that will be sold and the number of member-reserved tickets that will be available, 72-48 hours prior to the game. Using historical and present-day data, the club forecasts the demand and supply, always ensuring the dimensions of the required logistics. For this purpose, the use of disaggregated stadium attendance data enables the understanding of how people arrive at the stadium, acknowledging the importance of efficient resource allocation, to optimize the maximization of profits on merchandising and other sources of revenue. In addition, it can directly impact on stadium management, providing a data-driven way of allocating resources and pinpointing bottlenecks and potential improvements in stadium mobility. We believe that other types of venues worldwide could benefit from this research, including non-sports venues and businesses.

Potentially effective models and tools to learn from data and make valuable predictions are emerging in many fields. However, knowing and understanding when these tools can be trusted is of the utmost importance. These systems need to satisfy crucial properties, such as reliability and trustworthiness. The models created in this line of research produce reliable and interpretable estimations, allowing us to predict what would have happened if a different scenario or type of event had taken place.

Futbol Club Barcelona (FCB) has collected data on the arrivals at each entrance to the stadium for each particular match over many years. Here we focus on seasons 2016–2017, 2017–2018, 2018–2019 and 2019–2020, since most of the opponent teams characteristics and factors affecting stadium attendance change rapidly. In addition, broadcasting rights change every few years (Scelles et al., 2020), influencing game turnouts (Cox, 2012), and the times at which matches are held. Forecasting attendance allows planning the operational resources required for a given match. Previous research has focused on this forecast (Reade, 2007; Şahin & Erol, 2018), as well as in other sports (García & Rodríguez, 2009), but only on overall attendances to stadiums – at most, aggregating separated visitors season ticket holders, and pay-at-the-gate home and visiting team supporters (Allan & Roy, 2008), looking at how different factors (e.g., live broadcasts) affect overall attendances. Exploiting the nature of this data can provide us with both operational and commercial insights towards improving visitors' experience and optimizing the club's coordination of services. For example, by analyzing stadium utilization, as well as the crowdedness of the different stadium gates, we can determine the most critical places and at which times these might tend to present more extensive queues or other problems.

The data we are interested in consists of discrete non-negative counts of people per minute for each entrance at the stadium. Hereafter in this study, we refer to entrances as gates, within which there can be a different number of turnstiles. The modelization of attendee arrivals is usually tackled with regression methods for count data, where the response variables are non-negative integers (Cameron & Trivedi, 2013). This has been typically modeled using Poisson, geometric, or negative binomial distributions (Coxe et al., 2009; Heinen, 2003). The above-mentioned models fall under the generalized linear models family, where the count data is assumed to follow one of the aforementioned discrete distributions. However, these methodologies only apply to fit the total attendance at discrete time periods separately, and not the number of arrivals over time, as continuous-time curves do. Here we use a different approach. Function-on-scalar regression models (Reiss et al., 2010; Goldsmith et al., 2015) provide powerful tools to explore and understand previous data and predict curves of attendance as a function of time, relative to the beginning of the event, for future events at the stadium. This, combined with adequate management of operational requirements, can enhance stadium capabilities and directly impact the organizational performance and productivity of the business (Atkin & Brooks, 2021).

The specific research questions we want to tackle in this research are the following:

- Are the arrivals between distinct visitors different? If so, how?
- What are the different kinds of match days at the venue?
- How do different factors affect stadium arrivals?
- Can we estimate the busyness of each gate, in order to better allocate staff and potentially homogenize it in future events?

Our results show that it is informative to use this data for modeling and forecasting future match day events, segregating between season member holders and tourists. Different types of games present diverse patterns of arrivals, and these patterns may be instructive for

future events. These models provide timely and reliable forecasts with confidence intervals for the arrivals at the stadium and the possibility of forecasting other quantities of interest for the business. When dealing with scenarios involving tens of thousands of people, decision-makers can find data-driven solutions by discerning patterns from the different types of matches. For example, this information can improve visitors' experience at large venues, efficiently allocating staff, spotting most frequented areas from data, and recommending times and entries for different kinds of visitors in order to reduce lines and improve flows of people within the stadium.

Facilities that gather tens of thousands of people during short time spans, such as the Camp Nou stadium, influence the city dynamics. All of the modeling done in this research can be used to understand and optimize the logistics of a city. Camp Nou's highly complicated logistics and managerial requirements need to be studied and considered at the time of planning city mobility (Meta et al., 2021).

2 Data and methods

The data, provided by FCB, is taken from 108 games played during the seasons 2016–2017, 2017–2018, 2018–2019 and 2019–2020 at the Camp Nou stadium. Entries were restricted at the end of the 2019–2020 season because of the COVID-19 pandemic, and thus irregular games, played after the national lockdown in March 2020 were excluded from this study, as these occurred with small irregular amounts of visitors. Out of the 108 matches, there were two Spanish Supercup and four Trofeu Gamper games, which were not included in season member passes and we excluded them from the rest of the modelization. In addition, three games from the Spanish Cup were also excluded from this study as these were identified as irregular games, leaving us with ninety-nine games to analyze.

From these 99 matches, we used three primary sources of data: (i) time arrivals and destinations at the Camp Nou stadium, registered by the sensors and turnstiles at the entrances, (ii) tickets sold for these seasons, and (iii) regular season passes for members of FC Barcelona club members.

The stadium has a total seating capacity of 99, 354, with more than 100 entrance gates and 6 floors. There are 292 inlets or sections in the stadium, with different amounts of inlets on each floor. VIP areas were excluded from this study. See Fig. 1 for further details.

Every individual entering the stadium was registered, and the times of entries were recorded, permanently preserving visitors' privacy by anonymizing buyer IDs. In addition, for each entry, we can know if it is a club member or not and the entry gate and destination section allocated to that individual. The data was anonymized and no personal data was stored or processed. The high quality of the data and the level of detail is presented in Fig. 2, showing a sample game with the attendance curves on the different gates of the stadium.

Figure 3 shows data corresponding to a sample gate (gate P−49) aggregated over the studied period. The arrows pointing at inlets represent the destination of the people that entering through that gate, and the colors represent the proportion of people with each inlet assigned as the destination.

It is crucial to notice the way FCB sells tickets and season member passes. Every year, the club has a fixed number of season member passes, and the remaining seats (typically the minority) are sold throughout the year as regular visitor tickets. There are different types of season member passes. The most common ones include attendance to regular-season
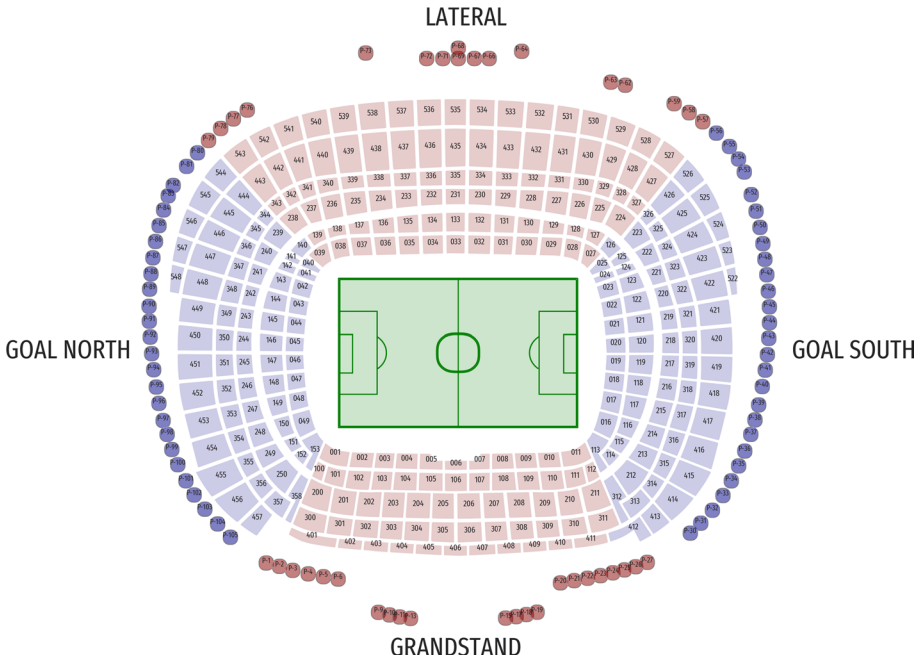
Fig. 1 *Camp Nou stadium schema* Gates are represented as circles, with their respective names in black (as P-# of gate). The other polygons represent the different inlets of the Camp Nou as numbers, where the first number always indicates the floor of the inlet, ranging from 0 to 5. The colors in the inlets or sections represent the main sections of the stadium (bottom seats in red of the plot are the grandstand, right and left seats in blue show the goal south and goal north sections of the stadium respectively, and the top seats in red show the other lateral section)
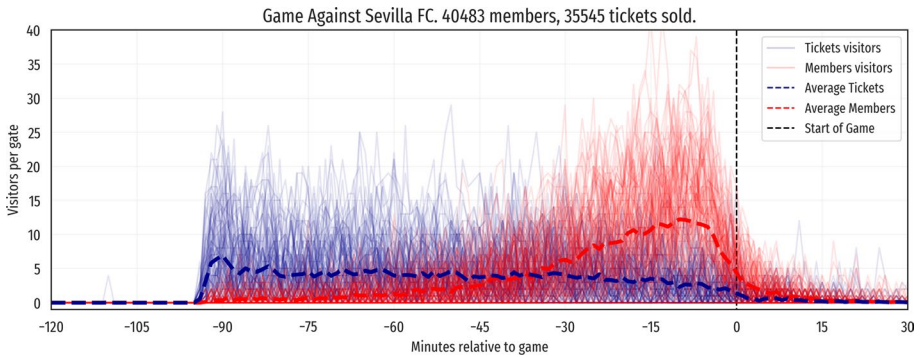
Fig. 2 *Sample game arrivals per gate against Sevilla FC.* Member arrivals and visitors that bought tickets are separated (red and blue, respectively). Game kickoff is indicated in the X-axis as 0. The dashed lines, described as "Average" in the legend (tickets and members), show the pointwise average of attendance across all gates

games played in the Spanish League (LaLiga), as well as Spanish Cup (Copa del Rey) matches. Other season passes also include European UEFA Champions League matches. Other types of matches are not included in these passes, e.g., Trofeu Gamper, Supercups,
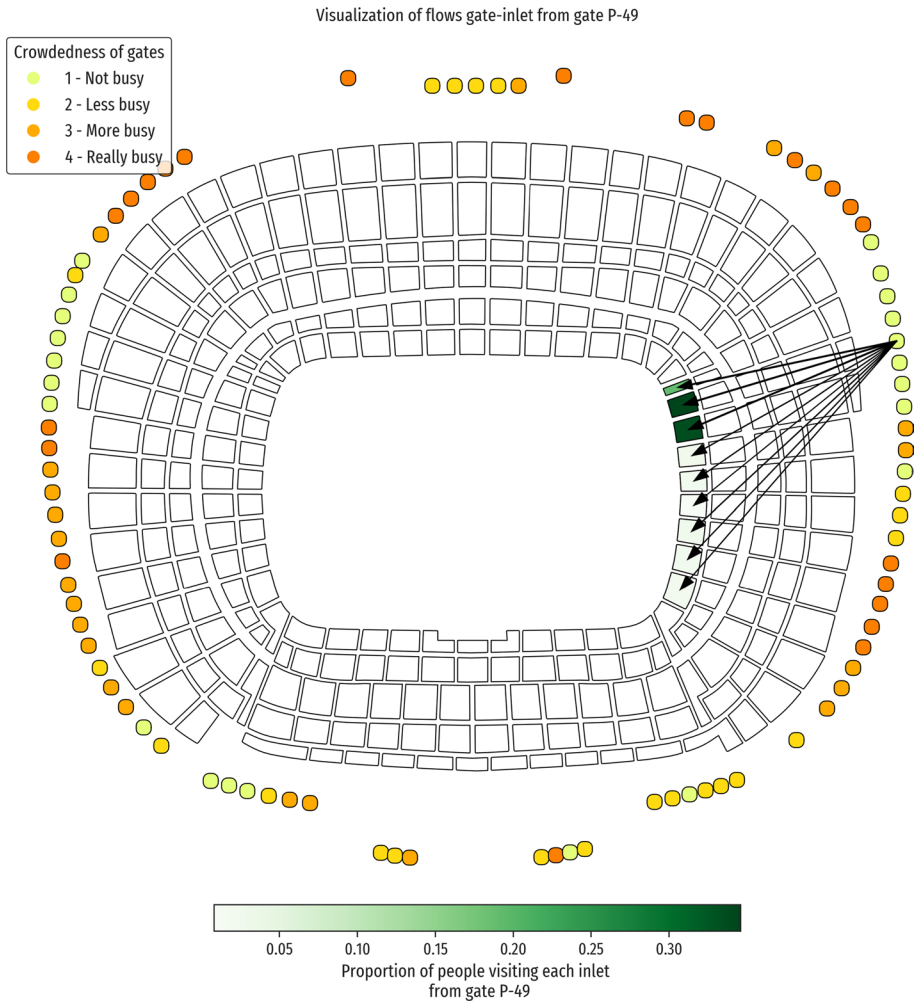
Fig. 3 *Camp Nou stadium schema of entries of gate P-49.* This shows where the visitors that enter this particular gate have their seats allocated as a percentage of the total people that entered that gate. The colors on the gates indicate the crowdedness, classified into four groups based on the quartiles of their busyness

etc. Whenever a season pass holder does not attend a game, they can notify the club in advance so that the seat can be vacated for that specific game, and purchased by another visitor. In this way, the club shares profits with the pass holder in a regulated and incentivized manner. Thus, season pass holders can have an outsize influence on the tickets, which are usually sold during the last few days before each game.

In addition to the above-mentioned datasets, we gathered additional data describing the context and characteristics of each particular game, such as type of competition, rival, day and time of the game, etc. These variables are described in Table 1.

Our quantity of interest is the number of people entering the stadium at any given time, either aggregated or disaggregated by type of visitor or by gate. To forecast attendance, two main types of covariates can be used. Firstly, fixed covariates are usually the characteristics

Table 1 Table with variables included in the models and their description

Fixed variables	
Weekends	Binary variable indicating whether a game was played on aweekend or not
Hour of the day	Binary variable indicating whether a game was played at 7PM or before, or after 7PM
Derby	Binary variable indicating whether the game was played against a historical rival: Real Madrid, RCD Espanyol or Girona CF
Competition	Categorical variable indicating the name of the competition of the game: LaLiga, Copa del Rey, Champions League
Cluster	UEFA Ten-year club coefficients rankings, clustered into 4 groups based on the quantile that the club belongs to. Clubs that are not included in the FIFA ranking were set at group 4
Rain	Binary variable indicating if a game is played under raining conditions or not
Playoff Champions	Binary variable indicating whether a game is a Champions League playoff match (round of 16, quarter-finals, or semifinals)
Second half of season	Binary variable indicating whether the game is played in the second-half of the season
Timely variables	
Tickets sold 3 days before	Number of tickets sold 3 days before the start of the game
Avg price 3 days before	Average price of tickets sold the week before 3 days before the game
Accreditations 3 days before	Number of accreditations given 3 days before the game
Seats freed 3 days before	Number of seats freed 3 days before the game by season member holders

of any given game, such as the day of the week, the time of the day, or whether the game is against a direct rival or a derby. These factors are usually known beforehand (e.g., at least 1 or 2 weeks ahead) and can be used to predict the attendance for any given match with much anticipation. Secondly, there are the covariates that change over time, such as the number of seats available for a given game, or the number of tickets sold τ days before the game. Table 1 lists the variables used in this study. In the following subsections we explain how to represent this data and the methodologies used to study how different characteristics and conditions affect the number of people arriving at the stadium.

2.1 Construction and description of functional data using curves of attendance

We represent the attendance data as functions of time, with the beginning of the games set as time 0. Let $Y(t)$ be the number of people that arrive at the stadium for any given match at time t, where $t \in \{t_0, \ldots, t_f\}$. In this study, t is discretized in minutes, with $t_0 = -120$ and $t_f = 30$, since most of the people ($> 99\%$ of visitors) enter the stadium between those times relative to the matches. We assume that $Y(t)$ is a random variable with distribution $Poisson(\lambda(t))$, so its expected value is $\mathbb{E}(Y(t)) = \lambda(t) > 0$, and its probability function is

$$Pr\left(Y(t) = y; \lambda(t)\right) = e^{-\lambda(t)} \frac{\lambda(t)^y}{y!}$$

where y is a non-negative integer. We assume that $Y(t)$ and $Y(s)$ are independent for $t \neq s$. The way we model the similarity between $Y(t)$ and $Y(s)$ when t and s are close to each other is by the continuity of the expectation function $\lambda(t)$, that is assumed to be a smooth function of t.

Since we are considering n games, we have $Y_i(t) \sim Poisson(\lambda_i(t))$, $i = 1, \ldots, n$, that we assume to be independent. Let y_{it} be the observed value of $Y_i(t)$ for $t = t_0, \ldots, t_f$. In order to obtain a smoothed estimation of $\lambda_i(t)$, $t \in [t_0, t_f]$, we fit a non-parametric Poisson regression model with time t as the only explanatory variable. Let $\tilde{\lambda}_i(t)$ be this estimation. Poisson models are fitted using the funcion `gam` from the R library `mgcv` (Wood, 2017) with link Poisson.

We consider $\{\tilde{\lambda}_i(t), i = 1, \ldots, n\}$, as a functional dataset (Ramsay and Silverman, 2005), with functions depending on time $t \in [t_0, t_f]$. Observe that we are dealing with constrained functions, since $\tilde{\lambda}_i$ can only take non-negative values.

Given that many of the functional data analysis methods are designed to work with non-restricted functions, it is instructive to transform the functions $\tilde{\lambda}_i(t)$ to obtain a non-restricted set of functional data. The standard transformation applied to the parameter λ of a Poisson distribution is the logarithmic transformation [canonical transformation in generalized linear models, GLM (McCullagh & Nelder, 2019)]. However, $\log(\tilde{\lambda}_i(t))$ results in very large negative values of t between -120 and -90 in many cases, because arrivals are very close to 0 in this early time interval. These large negative values might ruin the later analysis. Thus, in order to avoid posterior distortions, we will use this functional dataset from now on:

$$\tilde{\varphi}_i(t) = \log(\tilde{\lambda}_i(t) + 1), i = 1, \ldots, n; t \in [t_0, t_f]. \tag{1}$$

Further statistical analyses, both descriptive and modelization, are performed using $\tilde{\varphi}_i(t)$. Basic summary statistics like average and standard deviation curves are computed point-wise. However, depth based summary curves could also be used. See Fraiman and Muniz (2001), Febrero-Bande and Fuente (2012) for a complete definition and implementation of depth based descriptive measurements for functional data, such as the functional trimmed mean or the functional trimmed variance. Lastly, for visualization purposes, the results obtained will be transformed back to the original scale of attendance using the inverse transformation $e^\varphi - 1$.

When we focus the interest onto arrivals disaggregated by gate, we work with functions $\tilde{\lambda}_{ik}(t)$ for gate k and match i, and the corresponding transformations $\tilde{\varphi}_{ik}(t)$ are defined analogous to equation (1).

The descriptive statistics we performed on the functional dataset defined in equation (1) are, first, a Generalized Functional Principal Components Analysis (GFPCA) (Ramsay & Silverman, 2005; Goldsmith et al., 2015), which is analogous to Principal Component Analysis and enables for pattern recognition through dimensionality reduction on functional data; and second, hierarchical clustering based on the weighted L_2 distance between functions (Jacques & Preda, 2014; Febrero-Bande & Fuente, 2012).

2.2 Function-on-scalar regression

Once the functional dataset is constructed and explored, we want to model the outcomes of interest over the covariates or characteristics of each match, using function-on-scalar regressions (Reiss et al., 2010; Goldsmith et al., 2015). Let $x_{i1}, \ldots, x_{ip}$ be the observations of p explanatory variables for game $i = 1, \ldots, n$. Then, we fit a function-on-scalar regression as follows:

$$\tilde{\varphi}_i(t) = \beta_0(t) + \sum_{j=1}^{p} \beta_j(t)x_{ij} + \varepsilon_i(t), i = 1, \ldots, n.$$

Observe that this model has functional coefficients depending on t, where $\beta_0(t)$ is the functional intercept and $\beta_j(t)$ indicates the contribution over time of the j-th explanatory variable. We estimate the coefficient functions using the **pffr** function from the *refund* package (Goldsmith et al., 2020). Once we have the estimated coefficient functions $\hat{\beta}_j(t)$, we can obtain what we call the fitted values of this model, namely $\hat{\varphi}$:

$$\hat{\varphi}_i(t) = \hat{\beta}_0(t) + \sum_{j=1}^{p} \hat{\beta}_j(t)x_{ij}, i = 1, \ldots, n \tag{2}$$

and the corresponding fitted values of $\tilde{\lambda}_i(t)$ are

$$\hat{\lambda}_i(t) = e^{\hat{\varphi}_i(t)} - 1 = e^{\hat{\beta}_0(t)} \prod_{j=1}^{p} e^{\hat{\beta}_j(t)x_{ij}} - 1. \tag{3}$$

Now, we further investigate the disaggregated arrivals at a gate-level. For this purpose, we consider two different approaches using function-on-scalar regressions: on the one hand, including all gates in a single model, which uses information from all gates at each game, and on the other hand, considering a collection of gate-specific models.

First, in the single model, equation (2) is modified to include the different K gates of the stadium, as well as a factor $\gamma(t)$, with K levels, indicating the specific gate:

$$\hat{\varphi}_{ik}(t) = \hat{\beta}_0(t) + \sum_{j=1}^{p} \hat{\beta}_j(t)x_{ij} + \sum_{h=1}^{K} \hat{\gamma}_h(t)I(h = k); i = 1, \ldots, n; k = 1, \ldots, K, \tag{4}$$

where the subscript i indicates a particular match, j denotes the covariate, and k the gate, while the rest of the notation remains the same as in equation (2).

The other way to model this problem is to have K separate models, similar to the equation (2), where each model is fitted using only the n observations corresponding to that particular gate. This modelization strategy will have more freedom at the time of fitting the curves, and thus is likely to yield an improved goodness-of-fit, at the expense of increased complexity when interpreting results.

To evaluate these regressions, we use the adjusted coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \sum_{t=t_0}^{t_f} \left(\tilde{\varphi}_i(t) - \hat{\varphi}_i(t) \right)^2}{\sum_{i=1}^{n} \sum_{t=t_0}^{t_f} \left(\tilde{\varphi}_i(t) - \overline{\overline{\varphi}} \right)^2}, \tag{5}$$

where $\overline{\overline{\varphi}} = \left(\sum_{i=1}^{n} \sum_{t_0}^{t_f} \tilde{\varphi}_i(t) \right) / \left(n(t_f - t_0) \right)$ is the global mean, obtained from the standard output of *refund* package (Goldsmith et al., 2020), and our own weighted functional version of it,

$$R^2_{\text{funct}} = 1 - \frac{\sum_{i=1}^{n} \int_{t_0}^{t_f} \left(\tilde{\varphi}_i(t) - \hat{\varphi}_i(t) \right)^2 w(t)dt}{\sum_{i=1}^{n} \int_{t_0}^{t_f} \left(\tilde{\varphi}_i(t) - \overline{\tilde{\varphi}}_i(t) \right)^2 w(t)dt}, \tag{6}$$

where the weights are

$$w(t) = \frac{(1/n) \sum_{i=1}^{n} \tilde{\lambda}(t)}{\int_{t_0}^{t_f} (1/n) \sum_{i=1}^{n} \tilde{\lambda}(t) dt}.$$

Both measures are within the interval [0, 1] and they provide an intuitive scale of the accuracy of our models.

2.3 Synthetic matches and counterfactual predictions

To further validate our models and demonstrate their applications, we create *synthetic matches*, where we fix the value of all variables, but change one that we wish to explore. For example, if we want to know what would have happened if a specific game was played at the same time but on a weekend day, or within a different competition. These comparisons reveal helpful information, allowing us to evaluate the models and predict hypothetical scenarios that might appear in the future (e.g., due to uncontrolled conditions like rain), or even to assist in planning.

This method also allows us to estimate the confidence intervals for each of the predictions, assuming normal standard errors, which allows us to know the uncertainty in forecasts and perform e.g., stochastic optimization of scenarios.

3 Results

3.1 Descriptive statistics of raw data

The arrivals are distinct for every match, and each particular game has a different set of characteristics. However, by studying previous games we can gain interesting insights on the different types of arrivals. A strongly distinguishing feature is if the game is included in the season pass or not, as this means that all attendees bought a ticket for that particular game. For those games not comprised in season passes (e.g., Trofeu Gamper, Supercup matches, or friendly matches), people tend to arrive much earlier and there is no apparent peak right before kickoff. Figure 4 shows attendance curves for the different types of games.

The first quantity of interest in this research is the number of visitors that attended each particular match studied. Figure 5 shows an attendance summary visualization over time for the four seasons studied. Two main patterns are present: first, the cluster variable representing the quality of the opponent team seems a relevant feature for anticipating attendance, as mentioned in previous literature (Serrano et al., 2015); and second, the competition of the matches seems relevant (this is also related to the quality of the opponent team), as well as when games are played, since games played on the second half of the season appear to have better turnouts (Champions League matches played on the second half of the season are Playoff games, and thus are more appealing to spectators, besides being played against stronger opponents). To further analyze the patterns that we observe in Fig. 5, we build linear regressions in the next subsection.
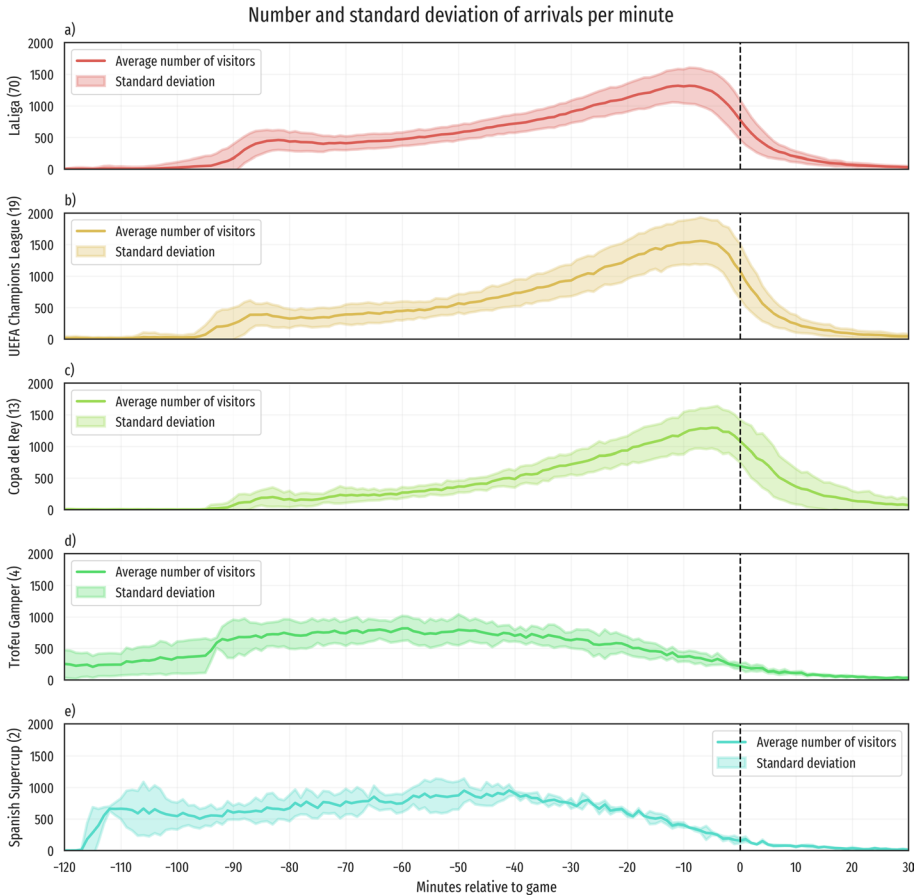
Fig. 4 *Pointwise average number and pointwise standard deviation of arrivals per minute segregated by type of game.* The types of game are shown in the different subplots: **a** shows LaLiga games, **b** shows UEFA Champions League games, **c** shows Spanish Cup games, **d** shows Trofeu Gamper friendly game, and **e** shows Spanish Supercup games. The number of matches averaged in each sub-figure is shown in the brackets of the Y-axis label

3.2 Forecasting overall stadium attendances

For completeness, we create a linear regression model that predicts the overall attendance, which also allows us to learn which characteristics are the most important for predicting attendance in general. We fit a linear model instead of a Poisson regression (even though the overall attendance is a count variable) because the response variable takes values between 42.500 and 97.989, and a Poisson distribution with expected value of this magnitude is very close to a normal distribution.

The estimated coefficients from the linear regression are shown in Table 2 and we interpret them as follows. The intercept is approximately equal to 75 thousand people, a value close to the average of attendance in the dataset (as shown in Fig. 5). Champions League matches have a small positive coefficient, compared to LaLiga matches, which is the reference class. The Spanish Cup matches have a larger negative coefficient, but it is only
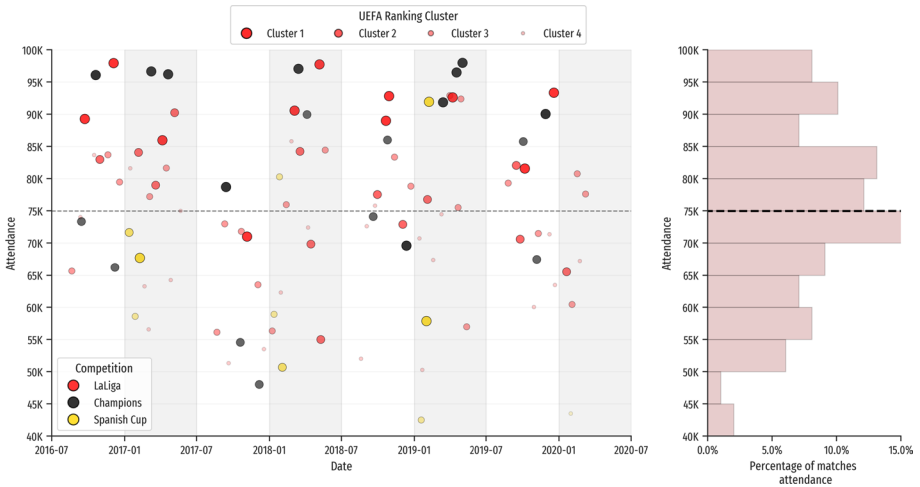
Fig. 5 *Visualization of the attendace at Camp Nou.* This visualization shows the characteristics of the games played during the four seasons studied. In the left-hand side plot, the color of the marker indicates the competition, and the size and transparency indicate in which cluster of the UEFA Ranking the opposing team belongs to. The black dashed lines in the middle of both figures represent the average attendance on the dataset. The right-hand side plot shows a histogram of the distribution of all the matches' attendance

Table 2 Results of the Linear regression model used to estimate attendance and interpret covariates

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	74789.3584	6230.1625	12.00	0.0000***
Champions league	360.3079	3586.5653	0.10	0.9202
Spanish cup	−7069.1617	4249.1506	−1.66	0.1000*
Playoff champions	4361.0049	4586.2280	0.95	0.3445
Second half season	2202.7743	1851.8348	1.19	0.2377
Rain	−9062.8118	3232.4070	−2.80	0.0063**
Weekends	5019.8692	2787.7713	1.80	0.0754 .
Hour	1696.2765	1879.0621	0.90	0.3693
Tickets sold 3 days before	0.8678	0.1572	5.52	0.0000***
Avg. prices 3 days before	4.5478	21.7386	0.21	0.8348
Seats freed 3 days before	−1.6212	0.2580	−6.28	0.0000***
Derbi Madrid	5586.0334	5735.9984	0.97	0.3330
Derbi Espanyol	11935.1441	4272.6281	2.79	0.0065**
Derbi Girona	6319.5168	5650.7301	1.12	0.2667
Cluster 1	10648.0606	3236.5168	3.29	0.0015**
Cluster 2	4562.3504	2510.8563	1.82	0.0729 .
Cluster 3	3658.0240	2335.0403	1.57	0.1211

Multiple R-squared: 0.7672; Adjusted R-squared: 0.7218; LOOCV R-squared: 0.6336

meaningful at a 0.1 significance level. Champions League playoff games add up to the coefficient for Champions League games. However, as literature points out (Serrano et al., 2015), attendance is highly related to the quality of the opposing team, and Champions

League playoff games are usually against strong opponents. The coefficient indicating matches played on the second half of the season (e.g., games played between January and July) shows that more people attend matches that are close to the end of the season, as these matches usually have a higher repercussion regarding the outcome of the competition. The coefficient for presence of rain is the largest negative coefficient, showing that approximately nine thousand fewer people show up on average at rainy games. The variable indicating games played on weekends has a positive coefficient (these matches are all LaLiga games), and matches played after 8PM show a small positive coefficient. We also see that tickets sold show a positive coefficient so that for each ticket sold, 0.9 more people show up on average. However, this coefficient is probably also related with the average price and seats freed up. The coefficient for prices of the tickets is also positive, as expensive games are also positively correlated with the quality of the opposing team, and negatively correlated with the number of seats freed up. The vacated seats have a negative coefficient, meaning that the greater the number of freed up seats, the less people tend to come – possibly reflecting that the match is less appealing to season pass holders. The three derby coefficients show that more people attend these games, as these have a special importance to the audience. Lastly, the Cluster variable from the UEFA Ranking 10–year club coefficients has the largest positive value, with the first cluster showing a 10.5K coefficient and being the only strongly significant one out of these coefficients, followed by Cluster 2 with a 4.5K positive coefficient, and lastly Cluster 3 which has a slightly positive non-significant coefficient. Cluster 4 was used as the reference class and thus the coefficient is 0.

There are two other main quantities of interest for the business. The first one is the number of seats freed on the last days after this estimation. This is important as it determines the supply of tickets that can be sold. The second one is the number of people that will neither assist nor release their seat. This is primarily relevant for organizational purposes. Figure 6 shows the boxplots of the residuals from these regressions. The variables included in the models are the same as in Table 2. We observe that the most predictable variable using linear regressions is the number of seats freed, followed by the variable indicating the amount of people that neither freed their seats nor attend the match, and lastly, the attendance.
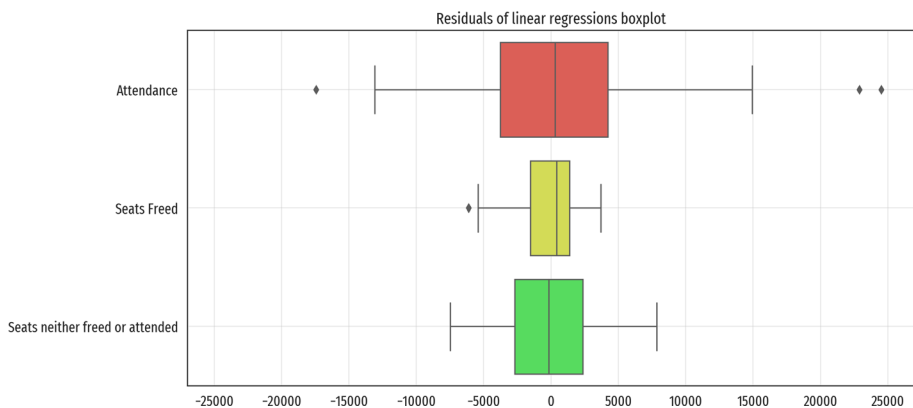

Fig. 6 *Boxplots of residuals from Linear Regressions,* using variables from Table 1 predicting quantities of interest three days before

Our findings show similar results to previous literature, where the most important feature seems to be the *ex-ante* quality or reputation of the opposing team (Czarnitzki & Stadtmann, 2002; García & Rodríguez, 2002). It is important to remark that our findings are limited by the number of observations in the dataset, and further research on these coefficients and their interactions should be carried out with caution. Weather also affects stadium attendance. As shown in previous literature, greater rainfall on match day correlates with lower attendance (Cairns, 1984). However, our results indicate that rain seems to affect the attendance of season pass holders particularly. Other studies evaluate the competition quality (Borland & Macdonald, 2003). We find that while competition is indeed a relevant factor, it is also highly connected to the opposing team's quality, and the importance of it is complicated to determine. Our results also indicate that scheduling impacts attendance demand, but there is a need for further research on scheduling conflicts, which may produce potential demand reductions (Simmons, 2006). Moreover, rivalry also seems to have an effect, as previous literature has pointed out (Tyler et al., 2017). Lastly, the price of tickets affects attendance, with the number of tickets being limited to the number of seats freed up by season pass holders showing opposite coefficients. Previous research on sport attendance demand indicates that the price of a large proportion of the tickets is inelastic (Coates & Humphreys, 2007). Dynamic ticket prices also change the paradigm of ticket pricing, as these prices can be adapted to the changing demand for football match attendance (Kobritz & Palmer, 2010). In this study we use an average of the prices sold during the last week before the time of the estimation. Our estimates show that larger prices are highly correlated with larger interest from fans, and thus the higher demand for attending those matches. Hence, these coefficients should be interpreted with caution, as there might be spurious correlations between those variables.

3.3 Construction and description of functional data

Next, we construct the functional data using the non-parametric Poisson regression model mentioned in the methods section (as in equation (1)). Figure 7 shows the results, where we observe that the number of arrivals over time by type of visitor is distinct, showing that visitors that purchased a ticket arrive earlier on average at the game, whereas people holding season passes arrive closer to the start of the game. Figure 7a shows that visitors buying tickets tend to arrive before visitors with member passes. In fact, the global average of raw individual arriving times for visitors buying tickets is 42.1 minutes before the game, whereas this average for visitors with member passes is 19.5 minutes before the game. Regarding the times with peak number of visitors entering the stadium, Fig. 7 b) shows the arrival of people buying tickets peak 32 minutes before the game with a peak number of 354 people per minute, whereas visitors with season passes peak 5 minutes before the game with a maximum number of 1041 people per minute.

Once the data is functionally represented, we can construct the Generalized Functional Principal Component Analysis (GFPCA), revealing interesting patterns in the data. GFPCA shows the principal modes of variation of $\tilde{\varphi}_i(t)$ around the mean function $\bar{\tilde{\varphi}}(t)$. Attendance data from the 99 games studied shows distinct patterns depending on how we disaggregate it. The first column of Fig. 8 (panels a, d, g and j) shows the first four functional principal components and the mean function for the aggregated data (transformed back to the original scale of attendance using the inverse transformation $e^{\varphi} - 1$), where both ticket visitors and people with season member passes are added. The first principal component (panel a) explains 48.6% of the variance in the data, differentiating between
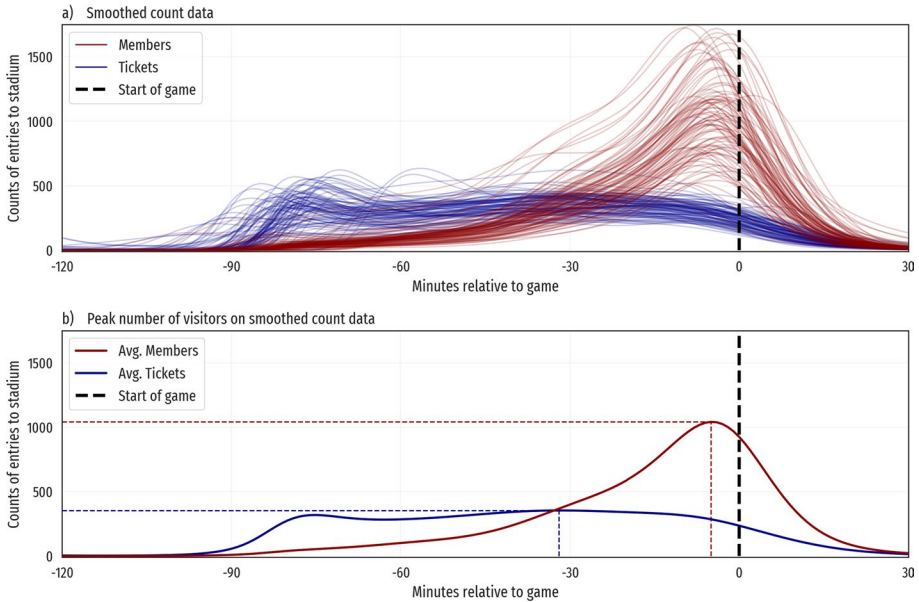
Fig. 7 *Camp Nou stadium arrivals over 99 games, smoothed with a GAM as a function of time with a Poisson family.* This shows the differences between arrivals by people with season passes and purchased tickets. In **a**, each blue line represents the number of people that arrived at the stadium and bought a ticket for a different game, and each red line represents the arrivals by the season member holders at the stadium over time, relative to the start of the game. In **b**, the blue and red lines represent the pointwise average of this smoothed data, and the time when they are highest

games with large attendance (specially with large ticket holder attendance, which are usually tourists who frequently arrive over gate overture, approximately 90 minutes prior to kickoff), from less popular games with fewer ticket visitors. The second principal component (panel d) explains a 23.4% of the variance, distinguishing games to which people arrive later (meaning that there are less tourists and more season pass holders, the last ones arriving mostly right before kickoff) from games in which the opposite happens. The third principal component (panel g) explains a much lower proportion of the variance, with only a 10.8% explained, and it differentiates between games in which there are more visitors overall arriving earlier (e.g., Champions League games or games against direct title competitors) from others with global lower attendance and later arrivals. The fourth functional principal component (panel j) only explains an 8.5% of the variance, separating games in which visitors arrive either very early or right before kickoff (meaning that there are two peaks, the first one upon gate overture, and the second one when season pass holders show up), from those games at which these peaks are not detected so clearly.

As there seems to be a large difference between club members and non-members, it is instructive to proceed to perform the same FPCA analysis separating the arrivals between members of the club and non-members. The second column of Fig. 8 (panels b, e, h and k) shows the mean arrivals of season pass holders and the respective modes of variation around the mean function. The first component explains a 49.5%, showing games where attendance by members or season pass holders arrive before the start of the game. The pattern shows games with attendees arriving closer to the last 60 minutes prior to kickoff. The
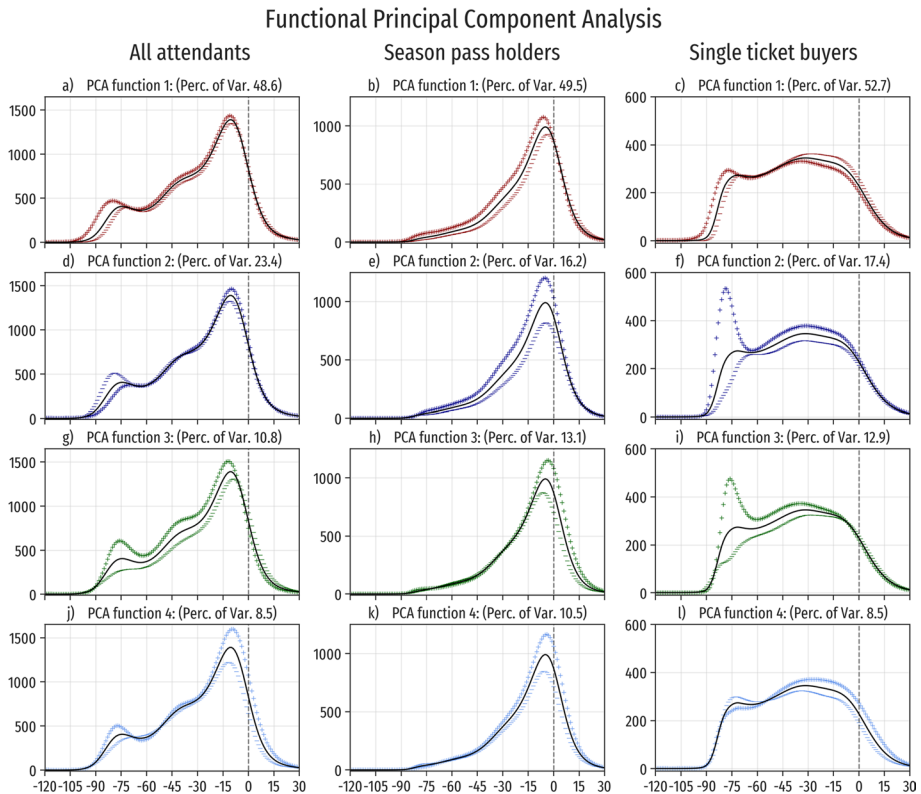
Fig. 8 *First four Functional Principal components of overall attendance, season pass holders and ticket holders.* The first four principal components of the FPCA for all attendees are shown in **a**, **d**, **g**, **j**, FPCA for members is shown in **b**, **e**, **h**, **k**, and FPCA for tickets in **c**, **f**, **i**, **l**

second functional principal component explains a 16.2% of the variance, showing games to which people arrive closer to kickoff. The last two components only explain a 13.1% and 10.5% respectively, showing similar patterns, in which people arrive mostly between 20 minutes before and after kickoff.

Lastly, the same analysis applied to non-member visitors shows a very different average distribution of attendance over time (see Fig. 8c, f, i and l). There are two main attendance peaks: the first one is concentrated around 80 minutes before games start, while the second one is a flatter peak located around 30 minutes before kickoff. The variability from game to game for the first peak is larger than for the second one. The first and fourth principal components (explaining 52.7% and 8.5% of variance, respectively) account for transferring attendants from the first peak to the second one. Variability along the first principal component implies changes from earlier arrivals (larger first peak and lower second peak) to later arrivals, while variability in the fourth component implies larger differences in arrivals closer to kickoff. The second and third components (explaining 17.4% and 12.9% of the variance, respectively) differentiate between games with larger ticket buyers' attendance, from those with lower attendance. In the second principal direction, we mainly observe differences in both arrival peaks, while in the third principal direction the differences are more important in arrival times between peaks.

3.4 Explaining overall attendance curves

Function-on-scalar regression models allow us to analyze the arrivals and explain how different matches' characteristics affect stadium attendance curves. The covariates used in this analysis are the same already listed in Table 1, with three differences: firstly, numerical variables have been standardized to have zero mean and unit standard deviation; secondly, in order to avoid overfitting, derby binary variables have been removed, as these contain too few observations falling under the less common category (less than five games for each binary variable); and thirdly, the variable indicating matches played during the second half of the season was removed.

To study how match factors or characteristics influence members and non-member attendance over time, we construct three regressions: one aggregating all the attendance, and two separating count data between season pass holders and purchased tickets. The models show a strong goodness-of-fit measured with the R^2 from equation (5), with the overall model (including all attendants) explaining 92.1% of the variance, and separate models on locals and ticket holders explaining 95.8% and 91.0% of the variance, respectively. However, they show less explanatory power according to the weighted R^2_{func} from equation (6), with a 36.0% of the variance explained on the overall model, in addition to a 57.7% and 26.1% of the variance explained for members and ticket holders, respectively. Both accuracy measures show that season pass attendees are more predictable than visitors buying spare tickets. However, there are opportunities to improve these models, collecting data on new matches and/or adding new covariates.

Figure 9 shows the time-varying intercepts of the two later models, showing how visitors with season member passes tend to arrive at minutes closer to the start of the game, whereas people that bought tickets arrive much earlier before kickoff.

It is important to remark that these coefficients are what the predictions of both models would look like with the standardized numerical covariates x_{ij} set to 0 and the categorical variables set to their modal class. The effects of estimated coefficient functions on the intercepts from the models are shown in the figures featured in the "Appendix" (See Figs. 14, 15, 16, 17). These show how estimates of arrivals vary depending on the characteristics of games, adversaries, types of game according to the competition, and numeric variables three days before games. It is worth noting that functional coefficients change over time, and can swap signs within the studied time interval, increasing attendance for some time periods (when the functional coefficients are positive), but decreasing attendance for others
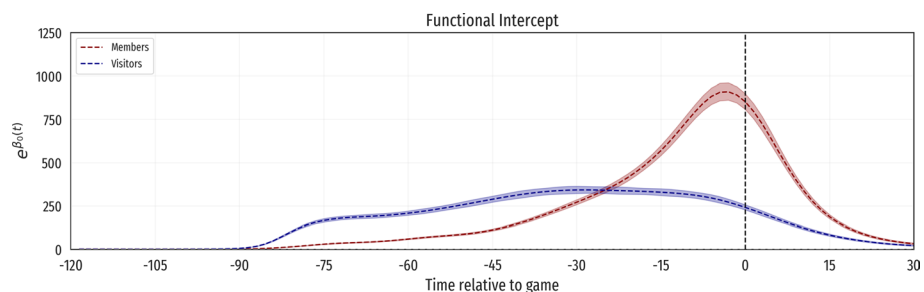
Fig. 9 *Functional intercepts of the models fit to the Camp Nou stadium arrivals over 99 games.* This shows the differences between arrivals by people with season passes and purchased tickets. The confidence intervals are pointwise 95% confidence bands

(when they are negative). Some factors yield previously known results, but others seem to provide new information relevant to the business. For example, Fig. 16 shows that for UEFA Champions League games, visitors buying tickets tend to arrive closer to the opening of the gates, while season pass holders arrive closer to kickoff. However, for Playoff games on the Champions League, season member holders tend to arrive before the start of the match, showing up earlier before the start of the game. After combining these functional coefficients, we can estimate the curve for a new set of characteristics, as we show in Fig. 10 for a regular LaLiga season match, with and without rain.

We can use this modelization to approximate different types of match attendance curves. We have modeled the arrivals in 3 groups: ticket buyers, season pass holders, and all attendees. Observe that the sum of the first two groups is not equal to the third, because the latter group also includes arrivals by accreditation, which have not been modeled separately because they represent less than 10% of the turnouts. To depict this technique, Fig. 11 shows three examples, with three distinct types of games: the average LaLiga game, an average UEFA Champions League game and the average Spanish Cup game. By *average game* we mean that we set the explanatory variables to their mean (if they are continuous) or mode (if they are categorical). The main advantages of this method are that we can estimate attendance curves for different scenarios and sets of characteristics for each particular game, as well as for different groups of visitors. In addition, we can predict what the attendance curves will look like for future games, and thus anticipate future visitors' behaviors.

3.5 Exploring and predicting arrivals by gate

Arrivals data contains information on each particular gate where each visitor entering the stadium is recorded. To exploit the nature of the data, we construct the functional data using the same non-parametric Poisson regression model, but this time segregating at a gate-level. Then, we proceed to obtain descriptive information about each gate, such as the mean function of each gate, computed using the *fda.usc* package in R (Febrero-Bande & Fuente, 2012). This reveals information on when each gate is more crowded and which are the busiest sections of the stadium.

In order to further determine which are the busiest sections of the stadium, we cluster the mean functions, using a hierarchical functional clustering methodology on the smoothed functional data, as mentioned in the previous section. To determine the number
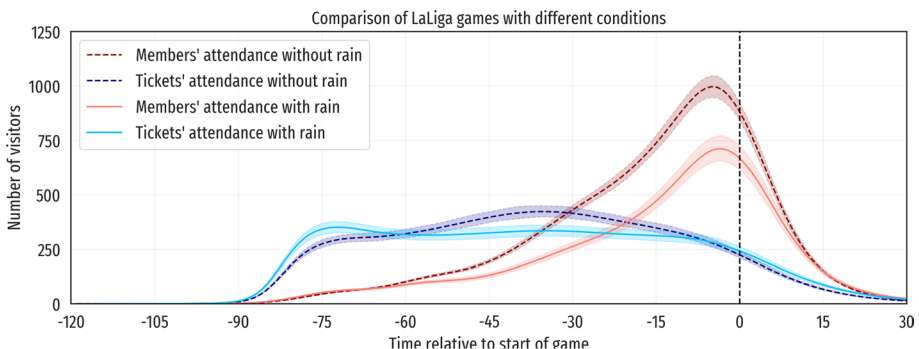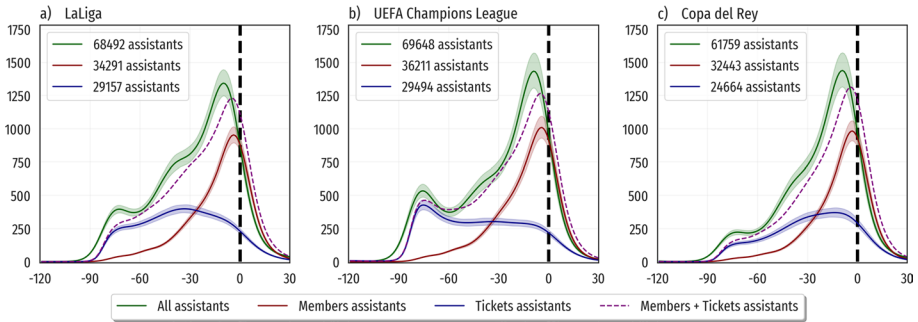

Fig. 10 *Comparison of average LaLiga games with and without rain*

Fig. 11 *Curves of attendance for three types of games, estimated three days before the game, using three different models.* The three models are: all assistants (green lines), only ticket buyers (blue lines) and season members (red lines). The numerical variables in the model are all set to the average values (e.g., number of tickets sold 3 days before, numbers of seats freed up 3 days before, the average ticket price logarithm, and the number of accreditations given out 3 days before). In **a** shows a regular season game, played on a weekend after 8PM, without rain, against a team in Cluster 2 of the UEFA ranking, estimated 3 days before. In **b** displays a game in the Champions League playoff game, with the rest of the variables set equal to the values used to compute Figure **a**. In **c** depicts a game played on the Spanish Cup, with the rest of the variables set equal to the values used to compute Figures **a** and **b**. Notice that we are plotting the results from three distinct models, and that the prediction for the global model is not the sum of the other two models (shown in purple lines for reference), since coefficients on each model are different (Color figure online)

of clusters we use the Calinski-Harabasz metric (Caliński & Harabasz, 1974), resulting in 5 clusters. Figure 12 shows the results of the hierarchical clustering. The clusters show distinct patterns of arrivals at the gates, separating the most frequented gates of the stadium from the less frequented ones. This provides information, not only to the event organizers for planning the allocation of gates to the different stadium inlets, but also other insights that could be used to understand peak demands for related services around the stadium area.

Whenever a ticket is sold, it always has a gate and an inlet and sit assigned to it. However, the data shows in which gates people enter the stadium, and their expected inlet. In case there are relevant differences between the assigned gate at the time of the sale, and the actual gate they choose to enter through, that can evidence factors not considered at the time of planning or assigning seats. These could be considered and potentially improved in the future.

The differences in curves of arrivals show that it can be interesting to model the arrivals by gate in various ways. Overall and segregated attendance curves are interesting for many reasons, as explained in the introduction of this work. Nevertheless, we are also interested in predicting arrivals at a gate level. Once the data is functionally represented, we can also regress this data using function-on-scalar regressions. We want to predict attendance curves for each particular gate on different matches. To do so, we perform two different modelizations, a multilevel model including all observations and gates at the same time (see Eq. 4), and gate-level models, fitting a separate model with all the matches at each gate, as explained in equation (2). The two different types of models show a tradeoff between the information provided by means of multilevel modeling, which includes all the gates in the same model, and the flexibility provided by fitting a separate model to each gate.

In the first model, the number of observations is equal to the number of gates times the number of matches played. In that case, we functionally represent each gate and game as a separate function, and use these observations to fit a unique model. This model contains
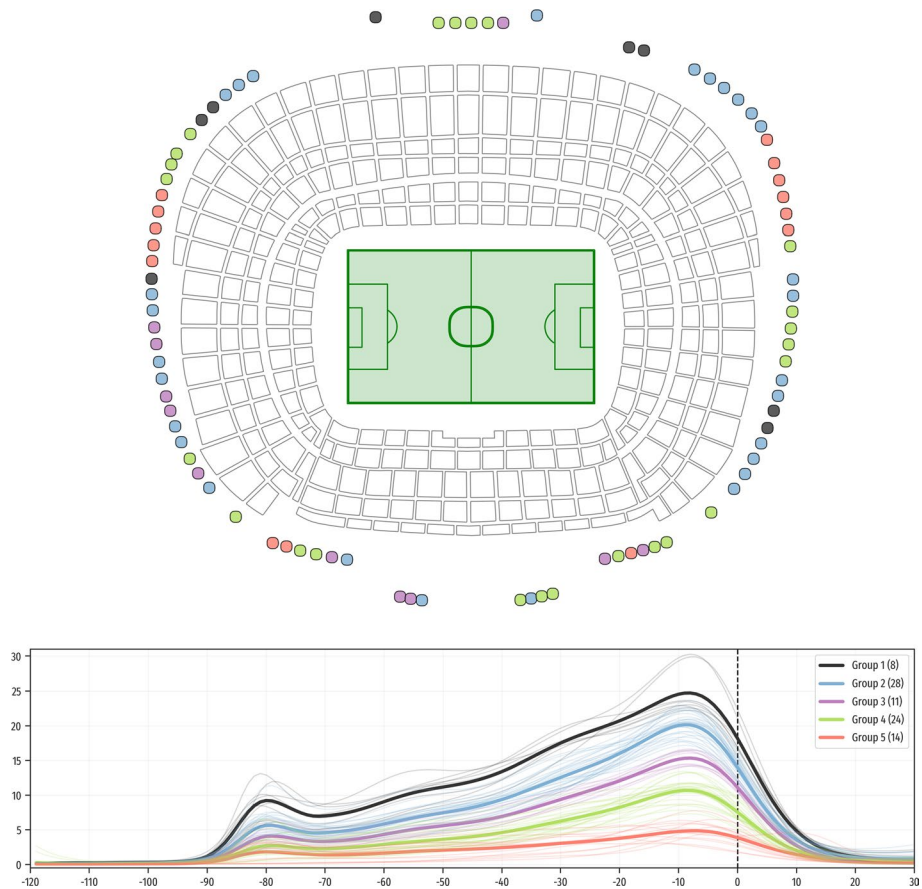
Fig. 12 *Camp Nou schema with colors on gates representing to which cluster they belong.* The thin curves of attendance on the bottom Figure shows matches pointwise average of the smoothed values for each gate, respecting the colors from the subfigure on the top. The thick lines are the pointwise average curves for each cluster

k gate functional coefficients, $\gamma_k(t)$ for $k = 1, \ldots, K$. The rest of functional coefficients are computed using the same covariates from the overall model, similarly to the ones shown in the "Appendix" on Figs. 14, 15, 16, 17. This model explains an 88.9% of the variance (using R^2 from equation (5)) and a 74.5% on the weighted functional R^2 from equation (6), showing different patterns for different gates. The advantage of this model is that it allows to observe how different factors affect all the gates at the same time.

On the other hand, the second way of modeling this problem is to make an ensemble of models, where each gate is fitted separately, performing a fit of the **pffr** function (as in equation (3)) separately for all the games modeled as in equation (2), with each game representing one observation per each of the K models. The models have more freedom to fit each specific gate separately, missing the opportunity to take advantage from the multilevel

information from close/similar gates. Theses models show that the variance explained on all the models fitted is $R^2 = 91.5\%$ and $R^2_{\text{funct}} = 81.6\%$, showing an improvement over the single model. However, these models are harder to understand as a whole, and we lose the multilevel information from the other model.

The separate gate modelization shown above provides insights on how predictable the different arrivals at the stadium are. Figure 13 shows that different gates and regions are more or less predictable. Bottom gates of entry are less predictable than the rest, as these are made from the tribune (see Fig. 1). This result is probably linked to the crowdedness of the gates, as shown in Fig. 3. The variables used to regress the arrivals at these gates are generic (these variables are the same ones explained in Table 1), and could be further adapted with more domain knowledge of the specific characteristics of each gate. Nevertheless, Fig. 13 allows us to understand which gates behaviors are more predictable.

4 Discussion and future research

A profiling of the financial performance of the highest revenue-generating clubs in the world (Deloitte, 2021) shows that match day profits for major professional football clubs in Europe used to represent about 15% of their turnover. Likewise, this profiling indicates that there was a substantial drop in match day revenue (17% or 257M €) in the season
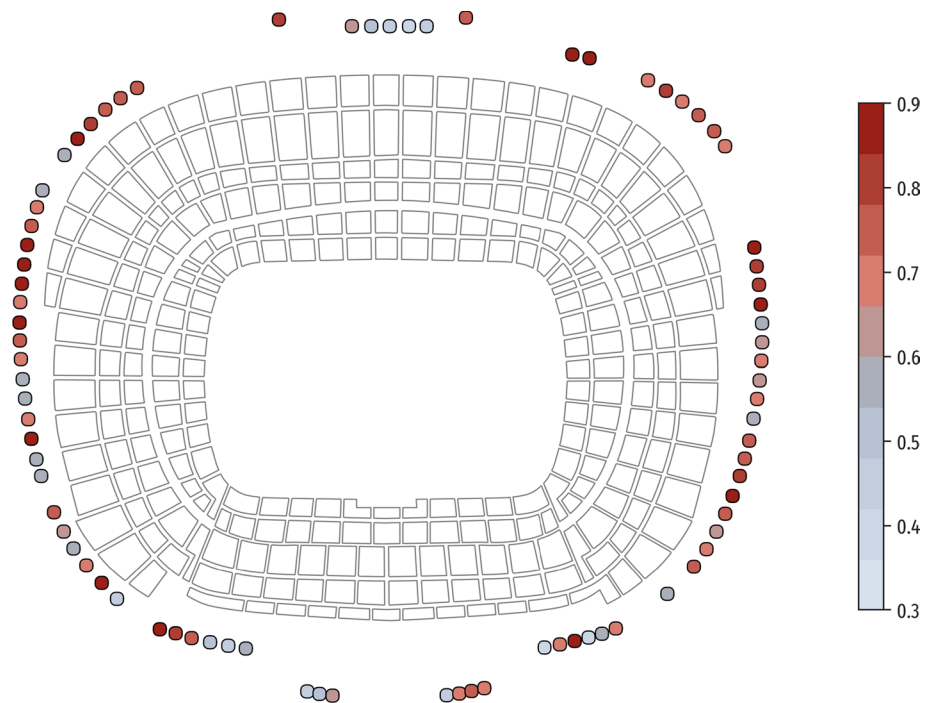
Fig. 13 *Plot of the weighted functional R-squareds by gate on the gate-wise models*

2020–2021 due to the pandemic, as matches were first postponed and then either canceled or resumed behind closed doors. Thus, there is an apparent need of upgrading the mobility of large facilities. After resuming economic activity, businesses that depend on large venues need to improve their safety and security measures, as well as plan the needs for future scenarios in advance, and upgrade event security protocols.

One new way that organizations have found to manage large venues is to recommend arrival times instead of the previous non-scheduling of arrivals. This helps to avoid overcrowded zones, and unnecessary proximity between attendees. Hence, the next natural step for this research is to design and optimize a tool for the creation of visitor arrival schedules, taking into account their preferences (e.g., tourists usually do not mind arriving earlier prior to kickoff). In order to avoid large agglomerations at the gates and improve entry smoothness, it may be helpful to incentivize season pass holders to arrive earlier at their assigned gates.

The models shown here could be improved in several ways. Firstly, additional game characteristics or with regard to the stadium surroundings could be included. Secondly, functional data could be aligned to match the peaks of people arriving across different games. That would probably yield better estimates of the functional coefficients. Thirdly, the data size could be increased and other types of events could be used to expand the capabilities of the models. The natural difference in behavior between season pass holders and ticket buyers might suggest segregating strategies that optimize not only safety and security, but also business interests, as it would be natural to expect tourists to spend more money during their visit than stadium regulars.

The results from individual gate modeling indicate that the multilevel models explain less variability, probably due to the lack of spatial dependence between gates in the model, aside from the restriction on the number of parameters, whereas the gate-level model provides better fits and larger accuracy, at the cost of having many different functional coefficients for each of the covariates. Thus, these are harder to interpret or obtain useful information from, other than the forecasts. Future research can expand the data used in these models and explore new ways in which functional regression models could better include the spatial dependence of arrivals, modelling these including the spatial-temporal nature of the data.

5 Conclusions

In this paper we have studied the arrivals of visitors at one of the largest sports venues in Europe. We use data from previous seasons to model and forecast new attendance curves for helping club decision-making processes, providing hints for a better management of its operational requirements and understanding of match days or event days.

Our research shows that we can effectively model the arrivals at the stadium with count data and function-on-scalar regressions, showing the differences between ticket buyers, and season pass holders. The use of disaggregated stadium attendance data allows us to understand how people arrive at the stadium, acknowledging the importance of an efficient

resource allocation, in order to optimize the maximization of profits on merchandising and other sources of revenue.

In addition, our results show that attendance can be effectively characterized, and foreseeing future events can potentially prevent undesired scenarios and adapt the necessities for each particular match day. The fan experience could also be improved and adapted to the desires of the different types of visitors and their behavior – offering access through special gates, or placing discounts on items at select places and times to manage and exploit the presence of potential buyers.

The forecasting potential of the models can be used to foresee attendance scenarios (as well as their uncertainty) and assist on logistic decision-making decisions for the team schedule, as well as with regard to financial planning.

To the best of our knowledge, no previous study has had access to this fine-grained dataset, nor has it focused on the differences between the match day arrivals between visitors and locals using disaggregated data, and how these differences can expand the opportunities, and personalize attendee experiences at large venues. The methodology we have presented is applicable in many different scenarios and venues (E.g. concerts, events, sport events, etc.), providing insights to optimize operational requirements, and we expect it can help organizations and institutions improve their management of large events. In addition, a compelling expansion of this approach could involve facilitating the measurement of urban and transportation dynamics, such as the spatial distribution of individuals arriving at venues or neighborhoods, enabling the anticipation of high concentrations of people in specific regions within short time periods.

Appendix: Additional figures

See Figs. 14, 15, 16, 17.
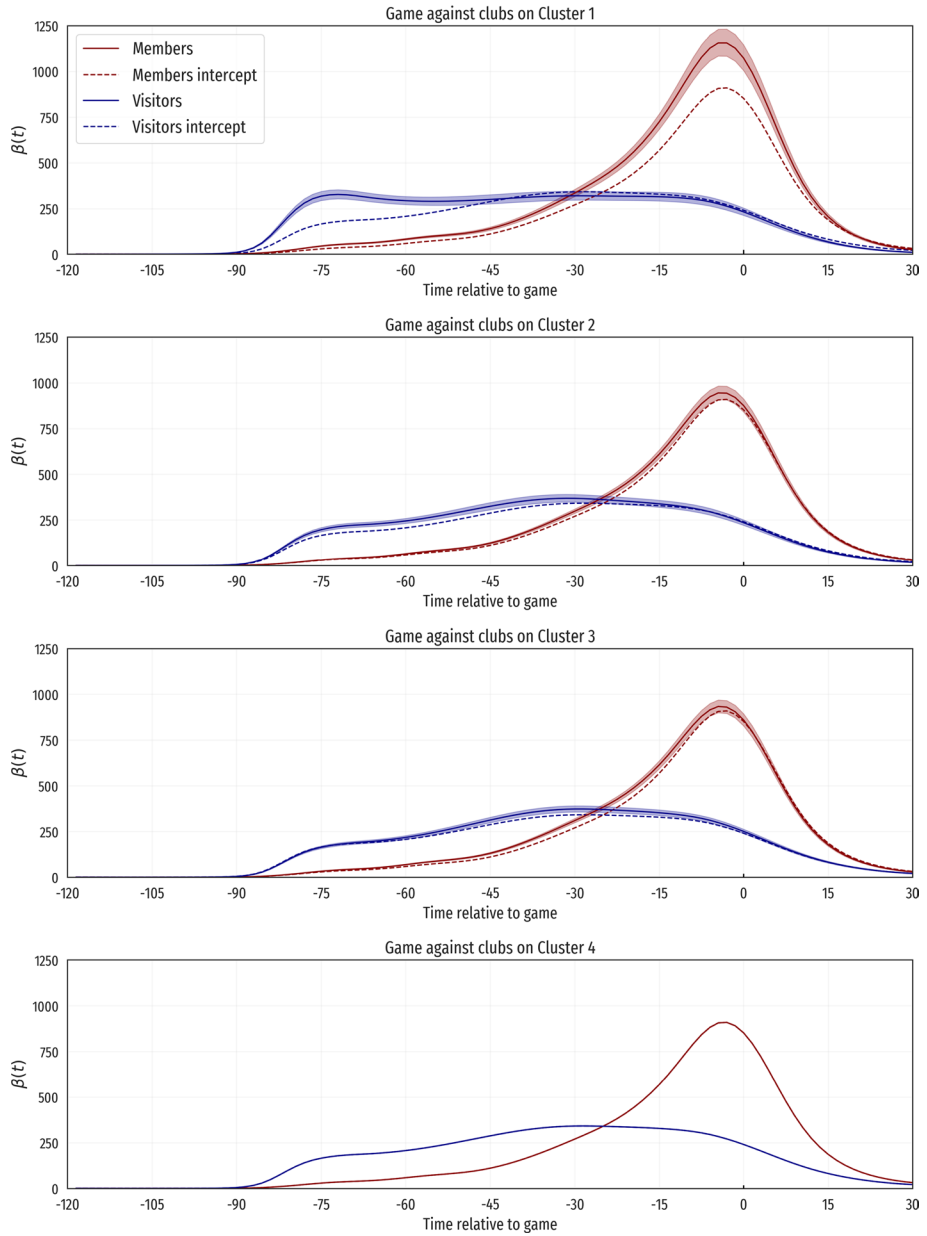
Fig. 14 *Effect of the coefficients of the UEFA Ranking clusters.* Cluster 4 is the reference class and thus has no effect over the attendance curve
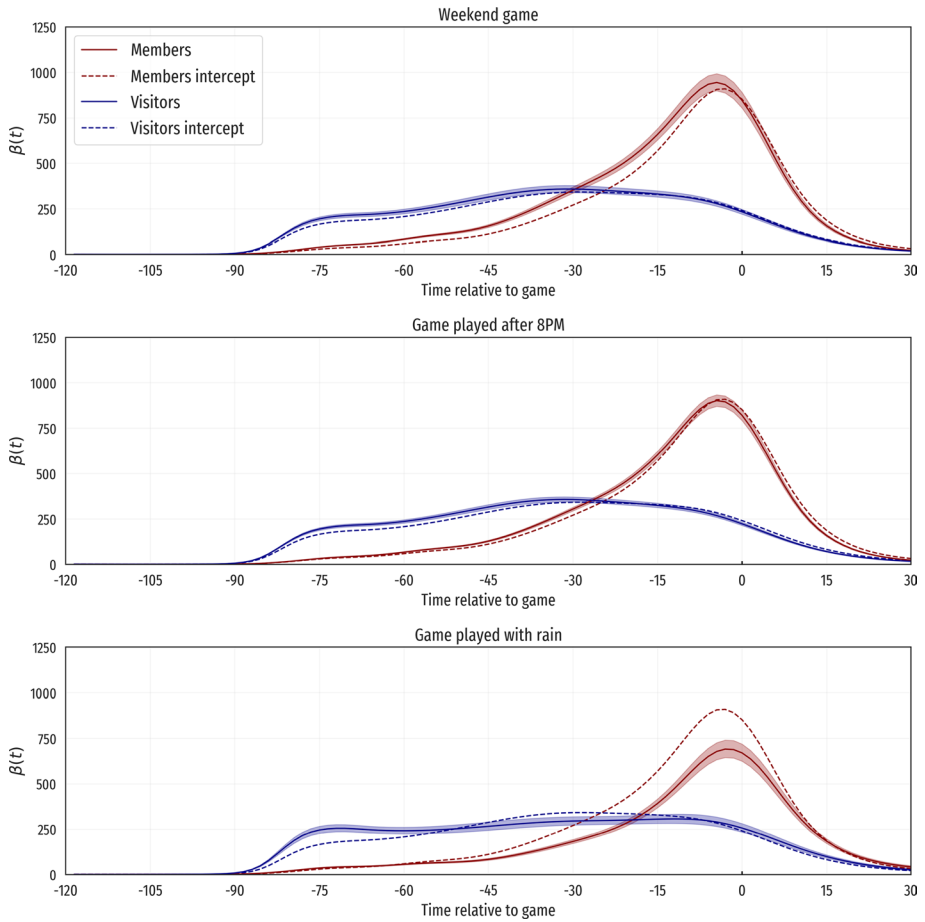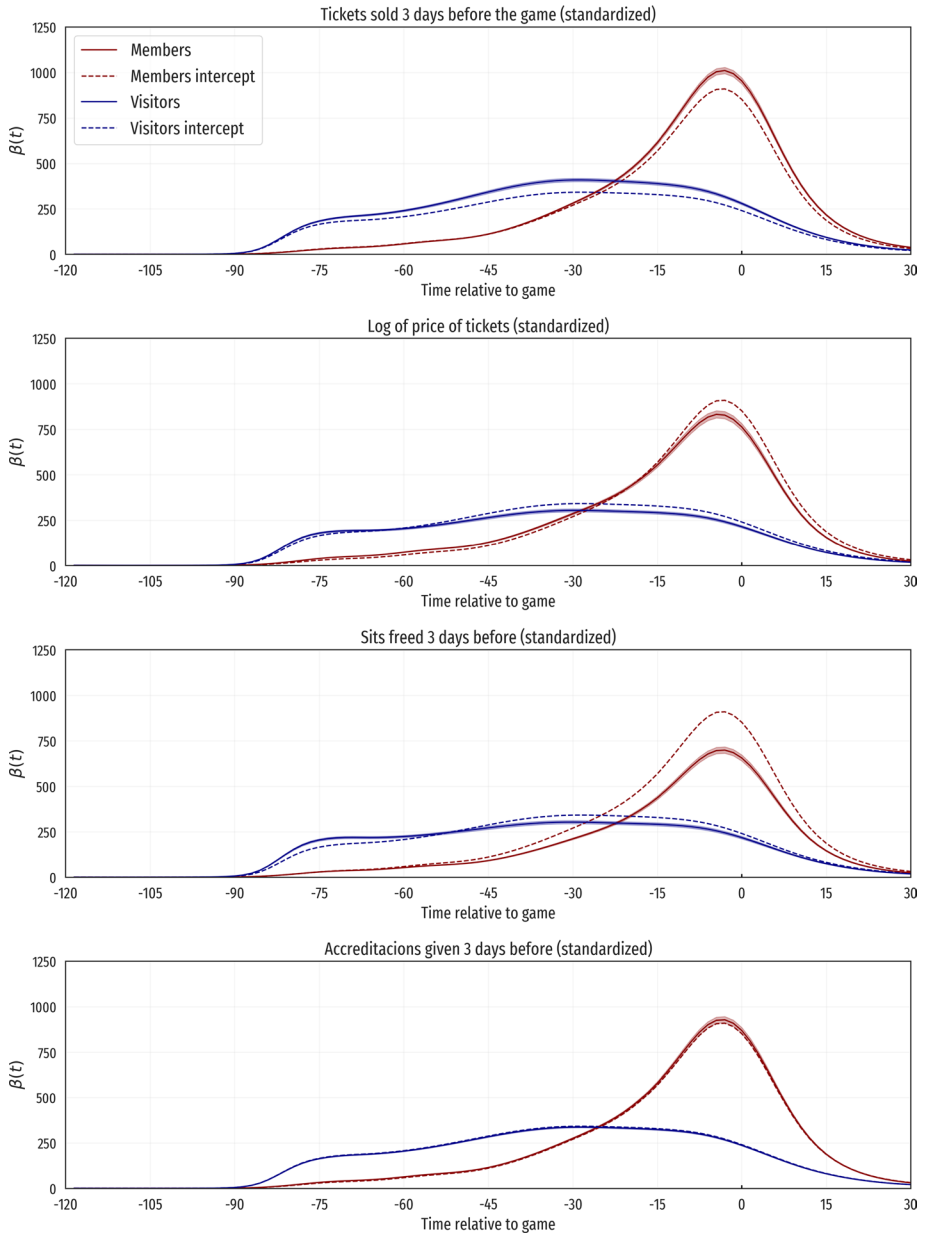
Fig. 15 *Effect of the coefficients corresponding to the characteristics of the game.*
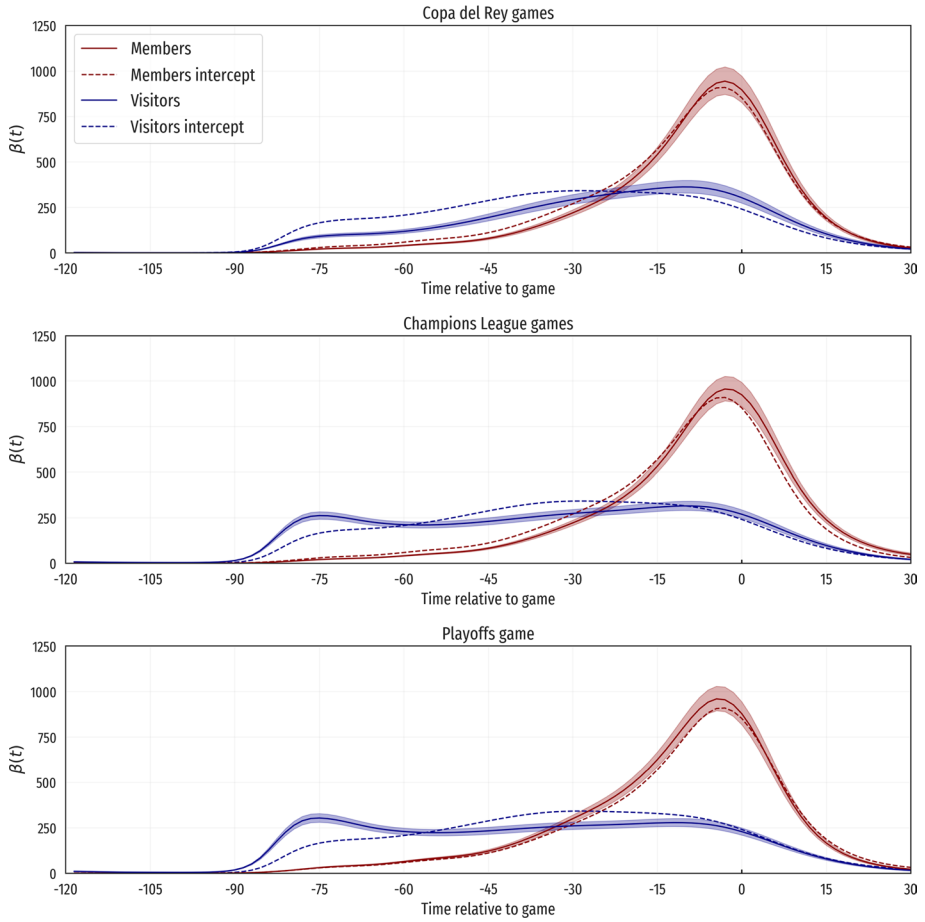
Fig. 16 *Effect of the coefficients corresponding to the type of game.* LaLiga games are used as the reference category

Fig. 17 *Effect of the coefficients corresponding to numerical variables.* In each figure, the continuous curve is the predicted arrival pattern for a hypothetical match where the value of the corresponding explanatory variable is one standard deviation over the mean, and the other variables are at their mean or reference category

Data availability The data used in this project are of private domain and cannot be open sourced.

Code availability The code used in this project is of private domain and cannot be open sourced.

Declarations

Conflict of interest Alex Gil, and Imanol Eguskiza are employees of FC Barcelona. However, the authors declare that they have not had any pressure that could have influenced the work reported in this paper.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Allan, G., & Roy, G. (2008). Does television crowd out spectators?: New evidence from the Scottish premier league. *Journal of Sports Economics, 9*(6), 592–605.

Atkin, B., & Brooks, A. (2021). *Total facility management*. John Wiley & Sons.

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

Buraimo, B. (2014). *Spectator demand and attendances in english league football*. In Handbook on the economics of professional football, Edward Elgar Publishing.

Cairns, J. A. (1984). Effect of weather on football attendances. *Weather (London); (United Kingdom)*, 39(3).

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods, 3*(1), 1–27.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (Vol. 53). Cambridge university press.

Coates, D., & Humphreys, B. R. (2007). Ticket prices, concessions and attendance at professional sporting events. *International Journal of Sport Finance, 2*(3), 161.

Cox, A. (2012). Live broadcasting, gate revenue, and football club performance: Some evidence. *International Journal of the Economics of Business, 19*(1), 75–98.

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment, 91*(2), 121–136.

Czarnitzki, D., & Stadtmann, G. (2002). Uncertainty of outcome versus reputation : Empirical evidence for the First German Football Division. *Empirical Economics, 112*, 101–112.

David Tyler, B., Morehead, C. A., Cobbs, J., & Deschriver, T. D. (2017). What is rivalry ? Old and new approaches to specifying rivalry in demand estimations of spectator sports. *Sport Marketing Quarterly, 26*(4), 204–222.

Deloitte Sports Business Group. Testing times Football Money League, January 2021.

Dobson, S., Goddard, J. A., & Dobson, S. (2001). *The economics of football* (Vol. 10). Cambridge University Press.

Febrero-Bande, M., & de la Fuente, M. O. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software, 51*(4), 1–28.

Forrest, D., Beaumont, J., Goddard, J., & Simmons, R. (2005). Home advantage and the debate about competitive balance in professional sports leagues. *Journal of Sports Sciences, 23*(4), 439–445.

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.

Fraiman, R., & Muniz, G. (2001). Trimmed means for functional data. *Test, 10*, 419–440.

García, J., & Rodríguez, P. (2002). The determinants of football match attendance revisited: Empirical evidence from the Spanish football league. *Journal of Sports Economics, 3*(1), 18–38.

García, J., & Rodríguez, P. (2009). Sports attendance: A survey of the literature 1973–2007. *Rivista di Diritto e di Economia dello Sport, 5*(2), 112–151.

Gimet, C., & Montchaud, S. (2016). What drives european football clubs' stock returns and volatility? *International Journal of the Economics of Business, 23*(3), 351–390.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., & Reiss, P. T. *refund: Regression with Functional Data*, 2020. R package version 0.1-23.

Goldsmith, J., Zipunnikov, V., & Schrack, J. (2015). Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics, 71*(2), 344–353.

Goller, D., & Krumer, A. (2020). Let's meet as usual: Do games played on non-frequent days differ? Evidence from top European soccer leagues. *European Journal of Operational Research, 286*(2), 740–754.

Heinen, A. (2003). Modelling time series count data: An autoregressive conditional Poisson model. Technical report, SSRN, 2003. Suggested citation: Heinen, Andréas, modelling time series count data: An autoregressive conditional Poisson Model (July 1, 2003). Available at SSRN: https://ssrn.com/abstract=1117187 or http://dx.doi.org/10.2139/ssrn.1117187.

Humphreys, B. R., & Johnson, C. (2020). The effect of superstars on game attendance: Evidence from the NBA. *Journal of Sports Economics, 21*(2), 152–175.

Jacques, J., & Preda, C. (2014). Functional data clustering: A survey. *Advances in Data Analysis and Classification, 8*(3), 231–255.

Jane, W.-J. (2016). The effect of star quality on attendance demand: The case of the national basketball association. *Journal of Sports Economics, 17*(4), 396–417.

Késenne, S. (2014). *The economic theory of professional team sports: An analytical treatment*. Edward Elgar Publishing.

Kobritz, J., & Palmer, S. (2010). Dynamic pricing: The next frontier in the evolution of ticket pricing in sports. *International Handbook of Academic Research and Teaching, 4*(9), 138.

Lawson, R. A., Sheehan, K. & Frank Stephenson, E. (2008). Vend it like beckham: David Beckham's effect on mls ticket sales. *International Journal of Sport Finance*, 3(4).

McCullagh, P., & Nelder, J. A. (2019). *Generalized linear models*. Routledge.

McDonald, H. (2010). The factors influencing churn rates among season ticket holders: An empirical analysis. *Journal of Sport Management, 24*, 676–701. 11.

Meta, I., Serra-Burriel, F., Carrasco-Jiménez, J. C., Cucchietti, F. M., Diví-Cuesta, C., Calatrava, C. G., García, D., Graells-Garrido, E., Navarro, G., Làzaro, Q., Reyes, P., Navarro-Mateu, D., Julian, A. G. & Martínez, I. E. (2021). The Camp Nou stadium as a testbed for city physiology: A modular framework for urban digital twins. Complexity.

Morrow, S. (1999). *The new business of football: Accountability and finance in football*. Springer.

Oh, T., Sung, H., & Kwon, K. D. (2017). Effect of the stadium occupancy rate on perceived game quality and visit intention. *International Journal of Sports Marketing and Sponsorship, 18*(2), 166–179.

Ramsay, J., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). Springer.

Reade, J.J. (2020). *Football attendance over the centuries*. Technical report, Henley Business School, Reading University.

Reade, J.J. (2007). Modelling and forecasting football attendances. *Oxonomics, 2*(1–2), 27–32.

Reiss, P. T., Huang, L ., & Mennes, M. (2010). Fast function-on-scalar regression with penalized basis expansions. *The International Journal of Biostatistics*, 6(1).

Richelieu, A. (2014). Strategic management of the brand in the world of sport. *Journal of Brand Strategy, 2*(4), 403–415.

Şahin, M., & Erol, R. (2018). Prediction of attendance demand in european football games: Comparison of anfis, fuzzy logic, and ANN. *Computational intelligence and neuroscience*, 2018.

Scelles, N., Dermit-Richard, N., & Haynes, R. (2020). What drives sports TV rights? A comparative analysis of their evolution in English and French men's football first divisions, 1980–2020. *Soccer and Society, 21*(5), 491–509.

Schreyer, D., & Ansari, P. (2021). Stadium attendance demand research: A scoping review. *Journal of Sports Economics, 23*(6), 749–788.

Serrano, R., García-Bernal, J., Fernández-Olmos, M., & Espitia-Escuer, M. A. (2015). Expected quality in European football attendance: Market value and uncertainty reconsidered. *Applied Economics Letters, 22*(13), 1051–1054.

Simmons, R. (2006). The demand for spectator sports. *Handbook on the economics of sport*, pp. 77–89.

Walker, B. (1986). The demand for professional league football and the success of football league teams: Some city size effects. *Urban Studies, 23*(3), 209–219.

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. CRC Press.

Authors and Affiliations

Feliu Serra-Burriel[1] · Pedro Delicado[2,3] · Fernando M. Cucchietti[4] · Eduardo Graells-Garrido[5] · Alex Gil[6] · Imanol Eguskiza[6]

✉ Feliu Serra-Burriel
   feliuserraburriel@gmail.com

✉ Pedro Delicado
   pedro.delicado@upc.edu

   Fernando M. Cucchietti
   fernando.cucchietti@bsc.es

   Eduardo Graells-Garrido
   egraells@dcc.uchile.cl

   Alex Gil
   lexgil@gmail.com

   Imanol Eguskiza
   imanol.eguskiza@fcbarcelona.cat

[1] Hemav Technology SL, Fontsanta 46, 08970 Sant Joan Despí, Barcelona, Spain

[2] Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Jordi Girona 31, 08034 Barcelona, Spain

[3] Institut de Matemàtiques de la UPC-BarcelonaTech (IMTech), Universitat Politècnica de Catalunya, Jordi Girona 31, 08034 Barcelona, Spain

[4] Barcelona Supercomputing Center, Jordi Girona 29, 08034 Barcelona, Spain

[5] Department of Computer Science, University of Chile, Santiago, Chile

[6] FC Barcelona, Av. Arístides Maillol, s/n, 08028 Barcelona, Spain