INTERNATIONAL TRANSACTIONS IN OPERATIONAL RESEARCH

INTERNATIONAL TRANSACTIONS IN OPERATIONAL RESEARCH

Intl. Trans. in Op. Res. 21 (2014) 399–414
DOI: 10.1111/itor.12068

Mathematical programming as a tool for virtual soccer coaches: a case study of a fantasy sport game

F. Bonomo[a,b], G. Durán[b,c,d] and J. Marenco[a,e]

[a]*Departamento de Computación, FCEN, UBA, Argentina*
[b]*CONICET, Argentina*
[c]*Departamento de Matemática and Instituto de Cálculo, FCEN, UBA, Argentina*
[d]*Departamento de Ingeniería Industrial, FCFM, Universidad de Chile, Chile*
[e]*Instituto de Ciencias, Universidad Nacional de General Sarmiento, Argentina*
E-mail: fbonomo@dc.uba.ar [Bonomo]; gduran@dii.uchile.cl [Durán]; jmarenco@ungs.edu.ar [Marenco]

Received 1 April 2013; received in revised form 11 November 2013; accepted 23 November 2013

Abstract

General awareness of the potential of mathematics to support sports decision making has been raised by the movie "Moneyball" through its portrayal of a real professional American baseball club that used mathematical tools to improve team performance. This article addresses the same challenge using as a test case a fantasy sport game organized by an Argentinian newspaper. Two mathematical programming models are presented that act as virtual coaches that choose a virtual team lineup for each round of the real Argentinian soccer league. The *a priori* design creates a competitive team for the game, while the *a posteriori* model determines what would have been its optimal lineup once the season is over and all the results are known. The *a priori* model was entered in the fantasy game, achieving results that positioned it among the highest scoring participants. Further development of such models would provide useful tools for supporting decision making by coaches or managers of real sports.

Keywords: fantasy sport game; mathematical programming; soccer; sports analytics

1. Introduction

Gran DT is a fantasy soccer game created and run by a major Argentinian newspaper for the first division of the real Argentinian professional soccer league. The objective of the game is to build the best possible virtual team by combining real players from the division's various clubs so as to accumulate the highest team point score. The virtual teams win or lose points depending on the weekly performance of the players, which is measured using both objective statistics (goals scored, shutouts, yellow and red cards shown) and subjective ones defined by the newspaper (individual player performance in each match, man of the match awards). The game rules require that each

© 2013 The Authors.
International Transactions in Operational Research © 2013 International Federation of Operational Research Societies
Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main St, Malden, MA02148, USA.

team satisfies a series of restrictions (budget, number of players per position, number of players per club). The virtual teams are dynamic in the sense that participants can make changes after each round of matches to improve their lineups.

Gran DT was first played during a few tournaments in the 1990s, then relaunched in August 2008, and has continued without interruption through all first division tournaments held since that date. Participation in the game has been massive, never falling below one million competitors and reaching a peak of almost two million (close to 5% of the Argentinian population) in the first tournament of 2009.

The inspiration for Gran DT was "Fantacalcio" (http://www.fantacalcio.kataweb.it/), the fantasy game of Italy's Serie A professional soccer league. Created by Riccardo Albini (http://www.fantagazzetta.com/esclusive-fg/come-inventai-il-fantacalcio-intervista-esclusiva-a-riccardo-albini-inventore-del-fantacalcio-165805) in the late 1980s, Fantacalcio is also run by an important national newspaper. Albini himself was inspired by the fantasy game of the North American Major League Baseball (MLB; http://mlb.mlb.com/mlb/fantasy/). The rules of the Italian game are very similar to those that were eventually adopted by Gran DT in Argentina.

Currently, there are various other fantasy sport games around the world such as English soccer's fantasy premier league (http://fantasy.premierleague.com/) and the NBA fantasy basketball game in the United States (http://www.nba.com/fantasy/). Similar games for virtual soccer have also become very popular (http://www.hattrick.org/; http://www.xperteleven.com/). Particularly interesting, however, is the growing use of fantasy games in recent years to improve the teaching of mathematics and stimulate student motivation in the subject at all education levels (Beliën et al., 2011; http://www.fantasysportsmath.com/).

This article presents two mathematical programming models that act as virtual soccer coaches supporting Gran DT competitors choices of virtual team lineups for each round. One of them was designed *a posteriori* and is referred to as descriptive because it addresses the problem of what would have been the optimal team round by round over the course of the tournament if the results had been known in advance. It is thus able to identify an ideal set of teams that would have had to be devised in order to obtain the highest possible total points while satisfying the game constraints.

The second model was designed *a priori* and is called prescriptive because it proposes changes to the team lineup round by round that are intended to improve team robustness as the tournament progresses without knowledge of future results. It does this by using existing information on player performance in earlier tournaments and past matches in the current one as well as data on key characteristics of the upcoming round.

The prescriptive model was tested by entering it in the fantasy game as a virtual participant. It routinely finished the tournament in the top 4% of all competitors, on one occasion ending up among the top one-tenth of 1%. If the six tournaments our model has played in so far are considered as a single tournament, the model ranks among the best two-tenths of 1%.

The literature in the field known as "sports analytics" (SA), which covers mathematical and computational developments for solving sports-related problems, has grown considerably in recent years. Prominent journals in operations research, applied mathematics, statistics, management, and economics have published numerous articles on SA (Coleman, 2012). In the case of mathematical programming for sports, the area of greatest development is the definition of league season schedules for various sports and competition formats, a subfield of SA known as "sports scheduling" (SS). Excellent surveys of the state of the art in SS, with analyses of a

© 2013 The Authors.
International Transactions in Operational Research © 2013 International Federation of Operational Research Societies

number of open problems, are found in Kendall et al. (2010) and Ribeiro (2012). A review of the main instances for various sports that have been studied in the literature appears in Nurmi et al. (2010).

Although software packages for assisting coaches of different sports in compiling, storing, and consulting data have been commercially available for some years, to the best of our knowledge there are no optimization or mathematical programming algorithm applications of the sort proposed in the present work that provide support for real or virtual coaches. The one that comes closest is perhaps Fantarobot (http://fantacalcio.repubblica.it/index.php?page=faq&ck_fantacalcio#16), an optional functionality on the Fantacalcio webpage that reorders the squad of starting and substitute players chosen by the participant into a "best" team, but as far as we are aware, its recommended selection is based only on the current average performance of each player. As regards to the use of mathematical techniques for supporting decision making in real sports, the best-known example is in American baseball (Lewis, 2003), where the applications employed are primarily statistical tools. An interesting use of computer support for football coaches is presented in Sierksma (2006).

The remainder of this paper is organized into five sections. Section 2 contains a detailed description of the Gran DT fantasy soccer game. Section 3 introduces the descriptive linear programming model that determines what the ideal teams would be over the course of the tournament if they could be devised for each round with advance knowledge of the entire tournament results. Section 4 sets out the prescriptive model that is built around an integer linear programming model that attempts to maximize the team score in each upcoming round based on its predictions of the point score each first division player will obtain. Section 5 contains the results obtained by the two models in different tournaments plus analyses of the impact of certain modifications to elements of the models. Finally, Section 6 presents our conclusions and some possibilities for future research.

2. Description of the game

The Gran DT fantasy game begins in the fourth round of the Argentinian first division's closing tournament, which despite its name is played in the first half of the year, and starts again in the fifth round of the opening tournament, played in the second half of the year. Since in each half-year tournament the 20 first division clubs play 19 rounds in a round-robin format, Gran DT consists of 15 or 16 rounds.

After signing up for the game under his or her Argentinian National Identity number, each competitor then creates a virtual team (hereafter simply "team") of 11 starting and four substitute players for a total of 15, chosen from about 500 players in the first division. The 11 can play in any one of three different formations: one goalkeeper, four defenders, four midfielders, and two forwards (1-4-4-2, the sole formation allowed in the early editions of Gran DT); one goalkeeper, four defenders, three midfielders, and three forwards (1-4-3-3); or one goalkeeper, three defenders, four midfielders, and three forwards (1-3-4-3). The four substitutes consist of a goalkeeper, a defender, a midfielder, and a forward, and can play only if one of the starting players in the same position either does not play or plays less than 20 minutes and is therefore awarded no points by the newspaper for his performance.

© 2013 The Authors.
International Transactions in Operational Research © 2013 International Federation of Operational Research Societies

Each player chosen for a team is assigned a symbolic monetary value of anywhere from 300,000 pesos for those who are just starting in the first division to more than 10 million for the top players in previous tournaments. The total value of the team must not exceed the budget limit of 65 million pesos (this figure has varied over the years between 60 and 70 million pesos). No more than three players can be chosen from any one first division club.

Each starting player wins or loses points for each round according to subjective criteria (the score awarded to him by the newspaper, whether he was man of the match) and objective criteria (goals scored, whether or not he was cautioned, whether or not he was expelled). The point score awarded by the newspaper on each criterion is an integer ranging from 1 to 10. Scoring a goal in open play wins 10 extra points for a goalkeeper, 9 points for a defender, 6 points for a midfielder, and 4 points for a forward. A penalty kick goal is worth 3 points regardless of the scorer's position. The man of the match as decided by the newspaper is awarded 4 extra points. For a clean sheet (shutout), the goalkeeper is awarded 3 extra points and the defenders 2 extra points. The goalkeeper loses 1 point for each goal conceded and wins 4 extra points for stopping a penalty kick (4 points are also deducted from the player taking the kick who failed to score). A yellow card (caution) costs a player 2 points and a red card (expulsion) costs him 4. A player who is on the field less than 20 minutes of a match is deemed not to have played that match and his actions are not counted, whether they are goals, yellow or red cards, etc. If he is a starter in the fantasy team, his place is taken by the substitute in the same position. If more than one starting player in the same position does not participate in a given round, their team plays with less than 11 players due to the rule limiting substitutes to one per position.

Starting players and substitutes can be switched from one round to the next as often as desired. Up to four transfers, in which a current player on the team is replaced with a new one, are allowed per round as long as the basic restrictions regarding formation and budget are met (this rule has changed since the first editions of the game, when only three such transfers were permitted). Finally, the player formation can be changed as often as desired as long as the transfers remain within the three permitted variations. For example, a game competitor who choses the 1-4-3-3 formation for a given round could adopt the 1-4-4-2 formation for the next one, thus replacing a starting forward with a starting midfielder.

3. The descriptive model

The *a posteriori* or descriptive model is run for a given tournament once that tournament is over. It determines what the optimal team configurations would have been with the complete tournament results already known. In other words, it uses as data the points obtained by each player in each round (or a given set of rounds if used for a partial analysis).

The model thus builds what we call the "perfect team." Beginning with an initial lineup in the first round of the game, it indicates round by round what changes should be made between starters and substitutes and which new players should be incorporated in order to obtain the highest possible final team point total while satisfying the game's restrictions.

The model is formulated as an integer linear program in which the objective function maximizes the team's point total while the constraints ensure the solution meets the game restrictions on team selection, permitted transfers, and the budget. It is described formally in the next subsection.

© 2013 The Authors.
International Transactions in Operational Research © 2013 International Federation of Operational Research Societies

3.1. The perfect team: mathematical formulation

We begin the formulation of the model by letting E be the set of first division clubs, J the set of first division players, and P the player positions on the field. In addition, we define F as the set of rounds and $F' = F\backslash\{\min(F)\}$ as the set of all rounds except the first one, this latter subset is needed because of the fact that changes to the team only begin with the second round of the game. A number of parameters representing various data items are defined for each player and each position as follows:

- For each player $j \in J$, the parameter $\text{team}_j \in E$ is the club player j belongs to; the parameter $\text{position}_j \in P$ is the position of player j on the field; the parameter $\text{price}_j \in \mathbb{R}_+$ is the price (value) of the player; and finally, $\text{points}_{jk} \in \mathbb{Z}$ specifies, for each player $j \in J$ and each round $k \in F$, the points player j obtained in round k.
- For each position $p \in P$, the parameters, $\max_p$ and $\min_p$, specify the minimum and maximum number of players in that position as determined by the three allowable formations.

For each player $j \in J$ and each round $k \in F$ we introduce the binary variables x_{jk}, which are equal to 1 if and only if player j in round k is a starter, and y_{jk}, which are equal to 1 if and only if player j in round k is a substitute. Finally, for $j \in J$ and $k \in F'$ we define the binary variable z_{jk} such that $z_{jk} = 1$ if and only if j is included on the team in round k. With these definitions we now set out the perfect team model as follows:

$$\max \sum_{j,k \in J \times F} \text{points}_{jk} x_{jk}$$

$$x_{jk} + y_{jk} \leq 1 \quad \forall\, j,k \in J \times F \tag{1}$$

$$\sum_{j \in J} x_{jk} = 11 \quad \forall\, k \in F \tag{2}$$

$$\sum_{j \in J:\text{position}_j = p} y_{jk} = 1 \quad \forall\, p,k \in P \times F \tag{3}$$

$$\sum_{j \in J} \text{price}_j (x_{jk} + y_{jk}) \leq 65,000,000 \quad \forall\, k \in F \tag{4}$$

$$\sum_{j \in J:\text{team}_j = e} (x_{jk} + y_{jk}) \leq 3 \quad \forall\, e,k \in E \times F \tag{5}$$

$$\sum_{j \in J:\text{position}_j = p} x_{jk} \geq \min_p \quad \forall\, p,k \in P \times F \tag{6}$$

$$\sum_{j \in J:\text{position}_j = p} x_{jk} \leq \max_p \quad \forall\, p,k \in P \times F \tag{7}$$

$$x_{jk} + y_{jk} - x_{j,k-1} - y_{j,k-1} \leq z_{jk} \quad \forall\, j,k \in J \times F' \tag{8}$$

$$x_{jk} + y_{jk} \geq z_{jk} \quad \forall\, j,k \in J \times F' \tag{9}$$

© 2013 The Authors.
International Transactions in Operational Research © 2013 International Federation of Operational Research Societies

$$1 - (x_{j,k-1} + y_{j,k-1}) \geq z_{jk} \quad \forall\, j,k \in J \times F' \tag{10}$$

$$\sum_{j \in J} z_{jk} \leq 4 \quad \forall\, k \in F' \tag{11}$$

$$x_{jk}, y_{ik} \in \{0,1\} \quad \forall\, j,k \in J \times F \tag{12}$$

$$z_{jk} \in \{0,1\} \quad \forall\, j,k \in J \times F' \tag{13}$$

The objective function maximizes the point total obtained by the starting players over the course of the tournament (note that since starting and substitute players can be interchanged as often as desired, the substitute players can be ignored in the objective function without loss of generality). Constraints (1) specify that in any round, each player is either a starter, a substitute, or not on the team. Constraints (2) impose that the team has exactly 11 starting players. Constraints (3) mandate exactly one substitute per position, thus making up the four substitutes required under the game rules. Constraints (4) set the budget limits and constraints (5) the maximum number of players who can be selected from each first division club. Constraints (6) and (7) establish the minimum and maximum quantities of players in each position (taken together, these restrictions and the one imposing exactly 11 starting players per team guarantee that one of the three permitted formations is selected).

Constraints (8), (9), and (10) relate variables x and y to z in such a manner that $z_{jk} = 1$ if and only if player j is added to the team (as a starting or substitute player) at round k. This definition allows constraints (11) to limit the number of players added to the team at each round (i.e., transfers) to four. Finally, constraints (12) and (13) define the nature of the variables.

This model has certain characteristics that are worthy of comment. Constraints (1–7) are knapsack-type restrictions, an extensively studied structure that is easily handled by commercial integer programming solvers. The imposition of a maximum number of transfers (constraints (8–11)), together with the introduction of the z transfer variables, complicates the computational solution process. This, as well as the fact that the instances to be solved are relatively sizeable (some 2000 variables and 3500 constraints), means that solution times may run to several hours, especially if we increase the number of game rounds (on this point, see Section 5.5).

4. The prescriptive model

The *a priori* or prescriptive model poses a greater challenge since in this case it must find good teams without knowledge of the players' future performance. The initial step in this process is to construct an index for each player that generates a prediction of the point score he will obtain in the next round. The model itself is applied in two very similar versions. The first version identifies the initial team by maximizing the global team index (i.e., the sum of the individual players' indexes) for the first round of the game while satisfying the game restrictions. Section 5.3 will consider the different impacts of identifying the initial team using a myopic and (relatively) nonmyopic procedure. The

© 2013 The Authors.
International Transactions in Operational Research © 2013 International Federation of Operational Research Societies

second version is applied starting with the second round of the game to determine the round-to-round changes and transfers that will maximize the global team index in each round while satisfying all the restrictions.

The entire process is similar to the descriptive model but with two notable differences: first, instead of the players' actual point score, the prescriptive model uses prediction indexes; and second, the global index is maximized round by round whereas the descriptive model maximizes the global score for the whole tournament since the players' point totals for every round are known in advance.

The challenge in the prescriptive case is thus to build an index that can produce a reasonable representation of what will actually happen. After some initial testing, we concluded that a player's point average in recent rounds was not by itself a good predictor of the points he would earn in the next round because it took no account of key match characteristics such as the rival club to be played, the match's home or away status, the current performance, or situation of his club, etc.

We therefore decided to construct each player's index as his point average in the rounds already played in the current tournament but weighted by three factors: the home or away status of his club's next round match (the weights applied were 1.05 for home games, 0.95 for away games), the league table position of his club's next round rival (1 to 1.05 if in the bottom five of the table, 0.95 to 1 if in the top five), and the current performance or situation of the player or his club (up to 5% more if on a scoring or winning streak, respectively; up to 5% less if on a scoreless or winless streak, or tired after, for example, a recent league or international match). For players who participated in either or both of the previous two tournaments, their corresponding averages are incorporated as though they referred to two additional rounds in the current tournament; for players who participated in neither of them, no corresponding points are assigned and the additional rounds are not included in their averages.

One last consideration included in the indexes is the "starting lineup" factor, which is 1 for those players assumed to be starters in the next round (based on coaches' announcements, press reports, or information posted on the game website), and 0 for all the other players. In exceptional cases, a value of 1 can also be assigned to a player who, although announced as a substitute, typically plays more than 20 minutes in every match. This factor is incorporated as an attempt to ensure that the team's 11 starting players will indeed be playing in the coming round. Though it is a crucial piece of information, it cannot always be known *a priori* due to the game rule stipulating that all lineup changes for a given round must be made no later than half an hour before kickoff in that round's first scheduled match. Since first division rounds typically spread the various matches across the "weekend" (Friday to Monday), there will be Fridays when the definitive starting lineup of some clubs playing later in the weekend will not have been announced yet. In such cases, the substitute players take on considerable importance and having "good" substitutes is thus advisable, though it may still be prudent not to spend much of the budget on them since in most cases they will not be used.

With all this in mind, after testing various values a weight factor of 0.1 was applied to substitutes in the objective function. Note that if a player has played more than 20 minutes in less than k rounds of the current tournament, where k (usually set to 3) includes the additional rounds representing participation in the previous two tournaments, he is not considered to be a candidate for the starting lineup even though he might be in it in the next round. This way, the model will not be tempted to choose a player who has performed well but has yet to establish himself as a starter. Also, with $k = 3$ a top player who is injured early in the season and who played in the previous two tournaments

© 2013 The Authors.
International Transactions in Operational Research © 2013 International Federation of Operational Research Societies

will again be "eligible" just one round after returning from his injury. The two versions of the prescriptive model will be developed in turn in the next two subsections.

4.1. The initial team

Once the player indexes have been calculated, the prescriptive model generates an initial team that will attempt to score the highest possible number of points in the first round of the game. As with the descriptive model, we let E be the set of first division clubs, J the set of first division players, and P the player positions on the field. The parameters are also the same as those in the descriptive model except for the $\text{index}_j \in \mathbb{R}$ parameter associated with each player $j \in J$, which represents the index discussed above. With the foregoing definitions, we set out the prescriptive model as follows:

$$\max \sum_{j \in J} (\text{index}_j x_j + 0.1\ \text{index}_j y_j)$$

$$x_j + y_j \leq 1 \quad \forall\, j \in J \tag{14}$$

$$\sum_{j \in J} \text{price}_j (x_j + y_j) \leq 65,000,000 \tag{15}$$

$$\sum_{j \in J:\text{team}_j = e} (x_j + y_j) \leq 3 \quad \forall\, e \in E \tag{16}$$

$$\sum_{j \in J:\text{position}_j = p} x_j \leq \max_p \quad \forall\, p \in P \tag{17}$$

$$\sum_{j \in J:\text{position}_j = p} x_j \geq \min_p \quad \forall\, p \in P \tag{18}$$

$$\sum_{j \in J} x_j = 11 \tag{19}$$

$$\sum_{j \in J:\text{position}_j = p} y_j = 1 \quad \forall\, p \in P \tag{20}$$

$$x_j, y_j \in \{0, 1\} \quad \forall\, j \in J \tag{21}$$

The objective function attempts to maximize the global team index for the first round of the game, using a weighting factor of 0.1 for the substitute players. Constraints (14) impose that each player be either a starter, a substitute or not on the team, while constraints (15)–(20) set the conditions the team must satisfy to stay within the game rules. Finally, constraints (21) specify the nature of the variables.

© 2013 The Authors.
International Transactions in Operational Research © 2013 International Federation of Operational Research Societies

4.2. Changes and transfers

Beginning with the second round of the game, the team can be updated for each round by interchanging starting and substitute players and making up to four transfers as permitted by the game rules. Each player's index is also updated with the data from the last round and the characteristics of the next one as described above (home or away status, current performance or situation of player or his club, next round rival's league table position, and the starting lineup factor).

The updating is performed by the changes and transfers version of the prescriptive model. This formulation contains the same sets, parameters, and variables as the initial team version presented in the previous subsection, the sole difference being that it includes the following constraint for a subset $A \subseteq J$ representing the current team to which changes and transfers can be made:

$$\sum_{j \in A} (x_j + y_j) \geq 11. \tag{22}$$

This added restriction ensures that the new team has at least 11 players who were present in the previous team. It thus implements the rule limiting the number of transfers to four, given that the total number of players on a team including starters and substitutes is 15. Both of the prescriptive models just described have about 1000 variables and 500 constraints, and are solved to optimality in a couple of seconds.

5. Results

In the first and second subsections below, we report the results obtained by the descriptive and prescriptive models, respectively. The rest of this section will be of particular interest to Gran DT players. In the third subsection, we present certain results obtained with the descriptive model that are then used to analyze how the initial team of the prescriptive model can be improved. In the fourth subsection, we apply the descriptive model to analyze the effect on scores of using the different permitted formations plus one that is not permitted and thus determine their relative performance. Finally, in the fifth subsection, we examine the scalability of the model, that is, how solution times vary with the number of game rounds. The data for the 2009 and 2010 tournaments, which were the basis for the computations in this section, can be found in http://mate.dm.uba.ar/~gduran/docs/test/.

5.1. Descriptive model results

The results obtained by the descriptive model in the four 2009 and 2010 tournaments are set forth in Table 1 together with their respective solution times. In column 3 we have added the value for the Optimal Fixed Team, a lineup obtained using a simplification of the descriptive model. This team, which in each case remains constant throughout the tournament, consists only of the 11 starting

© 2013 The Authors.
International Transactions in Operational Research © 2013 International Federation of Operational Research Societies

Table 1
Results of descriptive model for the 2009 and 2010 tournaments

Tournament	GDT winner	Optimum fixed team	Perfect team	Solution time
Cl. 09	1279	1318	1990	26 seconds
Op. 09	1375	1336	2173	9 minutes
Cl. 10	1227	1232	2027	20 minutes
Op. 10	1394	1412	2168	21 minutes

players who maximize the final point total while satisfying the game rules. The points obtained by the actual Gran DT game winner are shown in column 2 of the table.

As can be seen, the optimum fixed team with the original starting players was more or less tied with the human winners of the game, winning by a small number of points in three of the tournaments and losing by a few in the remaining one. These scores demonstrate that the winners played very well considering they could not have advance knowledge of the results, which are known only *a posteriori*. Clearly, good gamers are able to take effective advantage of the dynamic nature of the game's team updating process to improve their teams from round to round and thus compete at the same level as the team with a fixed lineup over the entire tournament but full knowledge of the future results.

The perfect team model, on the other hand, did much better than any human competitor, with point totals 50–70% higher than the game winner depending on the tournament. The big difference was due fundamentally to the fact that this model captures players who perform well sporadically, something not even the most expert Gran DT gamers are normally able to achieve. The point total differences between the closing and opening tournaments were due to the fact that in the latter the game starts one round earlier. These results were published in the newspaper running the game on various occasions (http://edant.clarin.com/diario/2009/07/09/deportes/d-01955551.htm; http://edant.clarin.com/diario/2009/12/20/deportes/d-02104838.htm; http://edant.clarin.com/diario/2010/05/20/deportes/d-02197713.htm).

Also given in Table 1 are the solution times for the perfect team. In every case, the problem was solved to optimality in periods ranging from 26 seconds to 21 minutes. The perfect team model contained approximately 2000 variables and 3500 constraints. The experiments were run using the CPLEX 12.2 optimizer on a PC with 2 GB of RAM and a 1.6 GHz processor.

5.2. Prescriptive model results

We analyzed the performance of the prescriptive model for the 2010 closing and opening tournaments. The global team index predictions generated by the model for the team it chose in each round are set forth in Tables 2 and 3 alongside the point totals actually obtained. As expected, since the model chooses the players who have performed best as of the last round and their individual performances are quite variable, the predictions tend to be higher than the actual values.

To check this behavior, we created a team of randomly chosen players for the 2010 closing tournament, the only additional condition being that in each round all 11 starting players must play, that is, that the starting lineup factor for each of them be equal to 1. This team won 899 points,

© 2013 The Authors.
International Transactions in Operational Research © 2013 International Federation of Operational Research Societies

Table 2
Results of the prescriptive model for the 2010 closing tournament

Round	Predicted points	Actual points
5	97.84	48
6	90.47	76
7	90.06	59
8	86.40	63
9	86.84	67
10	95.64	77
11	87.05	81
12	91.27	81
13	90.12	88
14	94.81	69
15	86.51	69
16	90.31	73
17	91.01	72
18	91.72	59
19	83.68	88
Total	1353.73	1070

Table 3
Results of the prescriptive model for the 2010 opening tournament

Round	Predicted points	Actual points
4	89.85	91
5	101.77	77
6	92.64	105
7	97.07	95
8	93.9	65
9	89.6	83
10	102.44	82
11	95.85	97
12	95.01	84
13	89.25	88
14	100.34	55
15	101.16	90
16	95.58	80
17	96.73	78
18	98.03	67
19	92.56	85
Total	1531.78	1322

whereas the predicted total was 934 points, a difference of less than 4%. The prediction was thus much closer to the true figure than was the case with the prescriptive model results for the 2010 closing and opening tournaments, where the differences were 21% and 14%, respectively.

In the Gran DT game for the 2010 closing tournament there were 1,442,682 competitors and the winner obtained 1227 points. Our model came in 13,547th place with 1070 points, positioning it within the top 1% of all participants. The random team, with 899 points, ended up in 498,726th

© 2013 The Authors.
International Transactions in Operational Research © 2013 International Federation of Operational Research Societies

Table 4
Results of our model for all the tournaments

Tournament	Position of our model	Number of competitors
Cl. 10	13,547	1,442,682
Op. 10	643	1,445,531
Cl. 11	34,788	1,358,982
Op. 11	41,246	1,368,522
Cl. 12	94,191	1,037,240
Op. 12	15,046	1,222,433
Global	530	343,017

place. Since a randomly chosen team would finish around the middle of the league table, this latter result suggests the number of active teams (by "active" we mean teams that are updated from round to round) would have been approximately one million. A nonactive team will typically finish the tournament participating with less than 11 players in each round, given that it does not replace those who are injured or dropped from the starting lineup as the tournament progresses. The estimate of one million active participants agrees with that of the game organizers, who have observed that two-thirds of the teams update from round to round. If we consider only active teams, our model's performance would place it among the top 1.5% of competitors in the game.

In the 2010 opening tournament, there were 1,445,531 competitors and the winner obtained 1394 points. Our model came in 643rd place with 1322 points, positioning it in the top 0.1% of all participants even if only active teams are considered. This was by far the model's best performance of the six tournaments it was entered in.

In both of the 2010 tournaments, the predicted point totals exceeded even those of the winner, but, as was noted earlier, it was not to be expected that the model's actual results would be similar to its predictions.

One possible reason for the model's tendency to perform better in the opening tournament is that the closing tournament and the Libertadores Cup tournament are both played in the first half of the year. Since the best teams play in the Libertadores Cup (the main South American competition), so do the best players, which means they often play tired or simply do not play at all in league matches. For the same reason, club coaches often do not decide their starting lineups until hours before the matches. These factors complicate the selection of a virtual team due to the uncertainty they generate and thus may affect the final results.

The prescriptive model was tested by entering it into six tournaments of the fantasy game competition. The relative performance of the model in each of the six tournaments is detailed in Table 4 together with a global result considering all six as though they were a single tournament. In the global calculation the model came in 530th place out of 343,017 competitors who participated in all six tournaments, positioning it among the top 0.2%.

5.3. Improving the initial team

As we saw in the previous subsection, the prescriptive model generates the initial team by maximizing the team's global index in the first round of the game. In this subsection, we attempt to determine

© 2013 The Authors.
International Transactions in Operational Research © 2013 International Federation of Operational Research Societies

Table 5
Results of the perfect team varying the initial team for the 2009 and 2010 tournaments

Tournament	Perfect team	Perfect team (myopic initial team)	Perfect team (nonmyopic initial team)
Cl. 09	1990	1950	1982
Op. 09	2173	2144	2171
Cl. 10	2027	2002	2013
Op. 10	2168	2150	2156

whether a nonmyopic approach would result in better initial teams. To do this, we perform a number of tests on two different initial teams using the perfect team of the descriptive model and the known results for the four 2009 and 2010 tournaments as data. In specific terms, for each tournament we compare the perfect team score to: (a) the best score obtained by the perfect team with the initial team defined myopically as the best team possible for the first round; and (b) the best score obtained by the perfect team with the initial team defined nonmyopically as the best team possible based not only on the first round but also on the two subsequent ones. In the latter case the results are derived by first determining the perfect team for the three first rounds only (as though they constituted an entire tournament), then taking the first-round team from this calculation and using it in running the model to determine the perfect team for the entire tournament.

The results of the comparisons are given in Table 5. As can be appreciated, the perfect team with a less myopic initial team generated scores that were less than 1% below those of the global perfect team, clearly closer to the latter than the perfect team with a myopic initial team whose scores were up to 2% below. These outcomes suggest that when using the prescriptive model both to generate the initial team and make round-by-round changes and transfers, it may be beneficial to consider more than one subsequent round in order to take advantage of known future data on match characteristics such as home/away status or the rival club to be played. A possible drawback of this approach, however, is that the prescriptive model does not use real scores but rather score prediction indexes, which may be increasingly inaccurate the later is the round considered.

5.4. How do the player formations compare?

Another analysis of interest to Gran DT players is determining which of the permitted formations (1-4-4-2, 1-4-3-3, 1-3-4-3) give the best results. In what follows, we compare these three plus one additional formation (1-3-5-2) not currently allowed by the game rules. Our test procedure uses the descriptive model to calculate the perfect teams for each formation in the four 2009 and 2010 tournaments, assuming each of the permitted formations held constant throughout, and then compares their performance with perfect teams that use a formation defined dynamically as the tournaments progress.

The outcomes are displayed in Table 6. Note that the perfect team for the 2009 closing tournament scored 1990 points because the old rule permitting only the 1-4-4-2 formation was then still in effect. The table shows the result that the perfect team would have obtained with the rule beginning in the 2009 opening which allows three different dynamic formations. The mean scores for each formation

© 2013 The Authors.
International Transactions in Operational Research © 2013 International Federation of Operational Research Societies

Table 6
Perfect team results with different player formations

Tournament	Perfect team	1-4-4-2	1-4-3-3	1-3-4-3	1-3-5-2
Cl. 09	2116	1990	2002	2015	1995
Op. 09	2173	2128	2155	2161	2126
Cl. 10	2027	1988	2015	2005	1968
Op. 10	2168	2143	2137	2147	2141
Mean	2121	2062.25	2077.25	2082	2057.25

Table 7
Solution time (in seconds) to obtain the perfect team with the descriptive model, by number of rounds

Tournament	15 rounds	16 rounds	17 rounds	18 rounds	19 rounds
Cl. 09	26	1129	1185	4669	11,697
Op. 09	–	557	1572	368	3803
Cl. 10	1196	2342	4077	1853	5949
Op. 10	–	1272	2413	1245	1065

indicate that the best ones were 1-4-3-3 and 1-3-4-3, that is, with three forwards rather than just two. This result is not unexpected considering that forwards usually have the highest point totals. The nonpermitted formation (1-3-5-2) performed more poorly even than 1-4-4-2. The point total of the formation with the highest mean score (1-3-4-3) was about 2% below that of the perfect team with a dynamic formation modifiable at every round.

5.5. Model scalability

Yet another interesting question is the scalability of the descriptive model in terms of solution times. Recall that the real Argentinian football tournaments have 19 rounds, whereas the Gran DT game starts at the fourth round of the closing tournament, for a total of 16 rounds, and at the fifth round of the opening tournament, for a total of 15 rounds. To test the model's scalability from the fantasy game to the real tournaments, we compared its running times to calculate the perfect team for the four 2009 and 2010 tournaments starting with the fifth, fourth and each of the earlier rounds—in other words, for tournaments with 15, 16, 17, 18, and 19 rounds.

The solution times obtained are summarized in Table 7. They show that in the most difficult instance, the 2009 closing tournament, if the model began from the first round (implying a total of 19 rounds), the perfect team would have been obtained in 11,697 seconds, or a little more than three hours. Also notable is that the times did not always increase with the number of rounds.

6. Conclusions and future research

This article investigated the contribution mathematical programming can make to the design of a virtual sports coach. The analysis took the form of a case study of an established fantasy sport

© 2013 The Authors.
International Transactions in Operational Research © 2013 International Federation of Operational Research Societies

game based on the two annual tournaments of Argentina's first division soccer league. Each edition of the game, organized by a major local newspaper, has attracted more than a million participants.

Two mathematical programming models were presented, one designed *a priori* and referred to as "prescriptive," the other formulated *a posteriori* and denoted "descriptive." They were applied to the identification of optimal or good teams for the fantasy game. The descriptive model was able to identify the ideal teams that would have obtained the highest possible point total while satisfying all of the constraints imposed by the game rules. It was also used to analyze different ways of defining the initial team of the prescriptive model, the different player formations permitted by the game, and the scalability of the models in terms of number of rounds and solution times.

The prescriptive model used historical data and the characteristics of the next match round to create a competitive team that was then tested by entering it into six tournaments of the fantasy game competition. The results obtained by the model positioned it in the top 0.1% of the game participants in one of the tournaments (opening 2010), in the top 4% in four of the tournaments, and in the top 10% in the remaining tournament. Indeed, if the six tournaments are considered as one, our virtual competitor placed within the top 0.2% (considering only those gamers who participated in all six tournaments). We can confidently expect that the longer is the tournament, the better our statistical and optimization tools will function.

In our case, the indications generated by the prescriptive model were followed 100% of the time, but it could also be used as a complementary support tool by an expert game competitor. For example, it can propose the k best sets of transfers for the team in each round and the expert could then choose among them. This would be done by running the model k times, in each case prohibiting the optimal solutions previously generated, or by deciding the inclusion or exclusion of a given player outside the model and then running the changes and transfers version of the model (see Section 4.2) to determine the rest of the modifications.

As regards to future research, a number of ideas for improving our virtual gamer model could be pursued. One would be to form a "high-risk" team with players whose point scores from round-to-round display high variance but good, although not the best, indexes as defined by the model. This might lead to worse results in general, but very good results in a certain number of cases.

Another idea would be to find the best players for each round without attempting to predict their point scores given their high variability. This could be done by implementing sophisticated statistical models using historical data to determine which variables (home or away match, rival club played against, stadium characteristics, referee, etc.) most accurately predict who will be the best players in a given round. These variables would provide a clear indication of the players that should be chosen for the team.

Yet another possible topic for further analysis is the apparent similarities between the problem analyzed in this study and the classic problem in finance of selecting a stock portfolio that maximizes investor income. Some models used to predict share behavior, such as the capital asset pricing model (CAPM) (Lintner, 1965; Sharpe, 1964) based on Markowitz's (1952) portfolio theory, could perhaps be usefully applied to the determination of robust virtual soccer coach models.

Finally, an especially interesting challenge would be to determine how tools such as those developed in this paper might provide valuable support for coaches in real sports. The combination of sports and mathematics for this purpose was well portrayed in the Bennett Miller film "Moneyball," based on a book by Michael Lewis (2003) and starring Brad Pitt and Jonah Hill. The movie tells the

© 2013 The Authors.
International Transactions in Operational Research © 2013 International Federation of Operational Research Societies

true story of an American baseball club manager who radically changed the team's strategy after incorporating mathematical techniques into his decision making, with excellent results.

Acknowledgments

The authors would like to dedicate this article to Carlos Prieto, a wonderful human being who departed this world not long ago and who was an enthusiastic promoter of this work. The authors would particularly like to thank Javier Romero and Jorge Blanco, organizers of the Gran DT game at the *Clarín* newspaper in Buenos Aires, for their constant help in bringing this paper to fruition. They are also grateful to Andrés Farall, Leonardo Faigenbom, Leonel Spett, and Pablo Groisman for the many discussions that contributed to this project; to Mario Guajardo, who carefully reviewed this paper and made numerous suggestions for its improvement; to Sebastián Ceria, Kenneth Rivkin, and Gustavo Braier for their comments; and finally to both anonymous referees and the associate editor for their extremely useful observations. This study was partially funded by project nos. ANPCyT PICT-2012-1324 (Argentina), CONICET PIP 112-200901-00178 (Argentina), and UBACyT 20020100100980 (Argentina), and by the Complex Engineering Systems Institute (ISCI, Chile). The second author was partly financed by project no. FONDECyT 1110797 (Chile).

References

Beliën, J., Goossens, D., Van Reeth, D., De Boeck, L., 2011. Using mixed integer programming to win a cycling game. *INFORMS Transactions on Education* 11, 3, 93–99.

Coleman, B.J., 2012. Identifying the "Players" in sports analytics research. *Interfaces* 42, 2, 109–118.

Kendall, G., Knust, S., Ribeiro, C.C., Urrutia, S., 2010. Scheduling in sports: an annotated bibliography. *Computers & Operations Research* 37, 1, 1–19.

Lewis, M., 2003. *Moneyball: The Art of Winning an Unfair Game*. W.W. Norton, New York.

Lintner, J., 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47, 1, 13–37.

Markowitz, H.M., 1952. Portfolio selection. *The Journal of Finance* 7, 1, 77–91.

Nurmi, K., Goossens, D., Bartsch, T., Bonomo, F., Briskorn, D., Durán, G., Kyngas, J., Marenco, J., Ribeiro, C.C., Spieksma, F., Urrutia, S., Wolf-Yadlin, R., 2010. A framework for scheduling professional sports leagues. In Ao, S.-I., Katagir, H., Xu, L., Chan, A.H.-S. (eds) *IAENG Transactions on Engineering Technologies*, Vol. 5. American Institute of Physics, College Park, MD, pp. 14–28.

Ribeiro, C.C., 2012. Sports scheduling: problems and applications. *International Transactions in Operational Research* 19, 201–226.

Sharpe, W.F., 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. *The Journal of Finance* 19, 3, 425–442.

Sierksma, G., 2006. Computer support for coaching and scouting in football. In Moritz, E.F. and Haake, S. (eds) *The Engineering of Sport 6, Developments for Innovation*, Vol. 3. Springer, New York, pp. 215–219.

© 2013 The Authors.
International Transactions in Operational Research © 2013 International Federation of Operational Research Societies