International Journal of Computer Science in Sport

Volume 13/2014/Edition 2

TABLE OF CONTENTS

Editorial

Arnold Baca

Department of Biomechanics, Kinesiology and Applied Computer Science,

ZSU, University of Vienna

Dear readers:

Welcome to the winter 2014 issue of the **International Journal of Computer Science in Sport (IJCSS).**

The issue contains three research papers and two project reports.

Julien Henriet designed a CBR-based system which considers previous training sessions to design new ones. The system was applied to aikido and was evaluated by experts.

Stephen J. Robinson developed a method to determine a baseball player's worth in context of his team's lineup using a refinement of the 'wins above replacement' method.

Peter O'Donoghue compared the accuracy of different regression models of the FIFA 2014 World Cup to each other using different sets of variables. The models which did not satisfy the assumptions of linear regression were more accurate than the models which satisfied the assumptions.

In their project report **Peter Buzzacott et al.** introduce a new model of head-up displays for recreational divers which can be fitted to the regulator hose outside the mask lens. The displayed information can help preventing rapid ascent injuries.

The project report by **Robert Strange** and **Lior Shamir** illustrates how offensive playing patterns in NFL games can be predicted using Support Vector Machines and a Weighted Nearest Distance method.

If you have any questions, comments, suggestions and points of criticism, please send them to me.

Best wishes for 2015!

Arnold Baca, Editor in Chief
University of Vienna, arnold.baca@univie.ac.at

Collaborative CBR-based Agents in the Preparation of Varied Training Lessons

Julien Henriet

Université de Franche-Comté, Chrono-Environnement UMR 6249 CNRS, FEMTO-ST DISC UMR 6174 CNRS, 16 Route de Gray, 25000 Besançon, France

Abstract

Case-Based Reasoning (CBR) is widely used as a means of intelligent tutoring and e-learning systems. Indeed, course lessons are elaborated by analogy: this kind of system produces sets of exercises with respect to student level and class objective. Nevertheless, CBR systems always result in the same solution to a given problem description, whereas teaching requires that monotony be broken in order to maintain student motivation and attention. This is particularly true for sports where trainers must propose different exercises to practise the same skills for many weeks. We designed a CBR-based system that takes into account any previous lessons offered and designs new ones so as to vary the exercises each time: reference to prior lessons helps to avoid giving the same lesson twice. In addition, the system is based on collaborative agents, each taking into account the exercises proposed by others so that each activity is proposed only once during a lesson. Five qualified sports trainers tested and evaluated the ability of this system as a means to design varied aikido training lessons and proved that our system is capable of creating classroom activities that are diverse, changing, pertinent and consistent.

KEYWORDS: CASE-BASED REASONING, MULTI-AGENT SYSTEM, COLLABORATION, EDUCATION SYSTEM, SPORTS TRAINING.

Introduction

One of the major challenges in teaching is to maintain student motivation. Repetition of the same exercises may lead to monotonous and boring lessons. In contrast, originality and exercise diversity will challenge students and maintain their interest, even if the same aspect is practised during many class sessions. This is particularly true in sports where trainers must propose varied exercises while having to train for the same skill over a block of weeks. Most of the tools provided by computer science, particularly from Artificial Intelligence (AI), would nevertheless produce exactly the same exercises and lessons for training in a single given skill. In this particular domain, repetitive activities are a drawback, yet lesson planning is a process based on adaptation of past experiences. This paper presents a MultiAgent System (MAS) that uses Case-Based Reasoning (CBR) systems to provide lessons with widely differing progressions. CBR is a problem solving method that adapts the solutions from similar known problems in order to solve new problems (Kolodner J. , 1993). It stores cases called source cases that are composed of two parts, problem and solution. Problems that occur and must be solved are called target cases. CBR describes a target case problem, retrieves the source cases whose problem parts most resemble those of the target case, reuses (adapts) the solutions of these most similar source cases, revises the adapted source case and capitalises this new experience, storing it in the system's knowledge base. CBR is widely employed in Intelligent

Tutoring Systems (ITS) and e-learning systems (Graesser, Conley, & Olney, 2012). It is actually well-suited to this kind of system (Jamsandekar & Patil, 2013), as well as other tools from AI-like multiagent systems (Rishi, Govil, & Sinha, 2007), Artificial Neural Network (Baylari & Montazer, 2009) and Genetic Algorithm (GA) (Tan, Shen, & Wang, 2012). J. L. Kolodner (Kolodner, Cox, & Gonzales-Calero, 2005) distinguished between two types of CBR-inspired approaches to education: Goal-Based Scenarios (Schank, Fano, Bell, & Jona, 1994) where learners achieve missions in simulated worlds, and Learning By Design (Kolodner, Owensby, & Guzdial, 2004) in which learners design and build working devices to obtain feedback, thus confronting themselves with the real world. O.P. Rishi et al. designed an ITS based on agents and a CBR system (Rishi, Govil, & Sinha, Distributed case-based reasoning for intelligent tutoring system: An agent based student modeling paradigm, 2007) in which a Personal Agent is responsible for determining student level. A Teaching Agent then determines the education strategy with the help of CBR regarding the description of the student level transmitted. Finally, a Course Agent provides and revises the lessons and exercises corresponding to the strategy proposed by the system with the help of a tutor. The CBR and GA based e-learning system proposed by Huang et al. also provides lessons taking into account the curriculum and the incorrect response patterns of a pre-test given to the learner (Huang, Huang, & Chen, 2007). A. Baylari and Gh. A. Montazer focused on the adaptation of tests to obtain a personalised estimation of a student's level (Baylari & Montazer, 2009). They used an Artificial Neural Network to correlate the student's answers to the tests and the exercises proposed by the teachers. Nevertheless, these approaches are insufficient for our system which seeks to change the proposed exercises. The systems in these approaches always correlate the same exercises to a single objective and learner level/experience.

Our application domain requires that a variety of solutions be proposed for a given problem, taking into account solutions previously proposed. Indeed, when an athletic trainer wants to prepare an athlete, he/she must take care to maintain the latter's motivation. Thus, the proposed training sessions and exercises must be varied, since the same exercises practised time after time would be boring. Finally, our system is based on agents: a lesson has one objective which is divided into sub-objectives and each agent is in charge of one sub-objective. This introduces another difficulty since an exercise must not be proposed more than once in the same lesson. Thus, solutions proposed by all the agents must be chosen collectively, taking into account the training history (the previous lessons proposed to the athlete during the season) as well as the solutions proposed by all other agents. Our application is therefore a MAS based on collaborative agents. E. Plaza and L. McGinty presented different policies suited to different cases of distributed CBR systems (Plaza & Mc Ginty, 2005). Nevertheless, these policies are well-suited to determining which solution is the best, considering a set of concurrent proposed solutions. Each agent of our system must provide solutions to different problems and which are drawn from a set of common exercises, and this without proposing exercises already chosen by the other agents. In the next section, we present the architecture of the distributed system we have designed. Its implementation and performance are then presented and analysed.

Methods

In the first part of this section, we detail the lesson structure of our distributed system. The distributed architecture and the data flows are examined in the second part. Finally, in the third part, we present how a lesson is designed theoretically, with an example to illustrate.

Lesson structure

To coach an athlete for participation in a specific competitive event, the trainer must make him/her improve different skills beforehand. The trainer divides the course into cycles of three to seven weeks, with a minimum of three lessons per week. During each cycle, the trainer emphasises one particular skill. Thus, over many lessons, the trainer sets the same objective (Matveev, 1965), (Issurin, 2010), (Garcia-Pallares, Garcia-Fernandez, Sanchez-Medina, & Izquierdo, 2010), (Ronnestad, et al., 2012). In sports, an objective is usually expressed as follows: "the athlete(s) must become capable of doing something". This objective may be technical, tactical, physical or sometimes even psychological. The objective is then divided into different sub-objectives which are elementary and that permit the athlete to reach the main objective. Many lessons are usually necessary in order to work and attain all the sub-objectives of a single objective. In addition, one sub-objective may be shared by two or more objectives. Going a step further, the same relations and constraints exist between sub-objectives and exercises: to reach one sub-objective, the trainer will choose and prepare different exercises from among a set of known ones, and a given exercise may help to reach many other sub-objectives. For our application, we have taken the example of training in aikido, a Japanese martial art. Table 1 presents an example of a support document elaborated by aikido trainers. As presented in this table, the trainer chose an objective for the lesson in function of the students' level and abilities. The objective is then divided into sub-objectives; aikido techniques are proposed for practice in order to reach each sub-objective. Students receive instructions with each technique.

Table 1. Example of document prepared by aikido trainers for their lessons.

Objective: Using a grip		Duration: 30 min.	
Sub-objective	aikido technique	Duration	Instructions
Breaking the partner's posture	Ryotedori Tenchinage	10 min	- Place yourself behind your partner; - Use both of your hands to place your partner's shoulders behind his/her heels.
	Katatedori Kokyuhoo	10 min	- Your gripped hand goes to the floor; - Synchronise yourself with your partner; - Your partner must go for the hand he/she wants to grab.
Relaxing despite a grip	Katateryotedori Kokyunage	10 min	- Even if your wrist is strongly grabbed, all the other parts of your body can move (shoulders, hips, feet).

In this example, the chosen objective 'using a grip' is divided into two sub-objectives. The first is practised through two aikido techniques that help the students to understand how to break their partner's posture and to train themselves to do this. Finally, a third technique invites the students to move all the parts of their body that the partner is not grabbing. Thus, aikido is based on martial techniques that will emphasise a particular point and therefore allow one to understand, practise and reach each sub-objective listed above. An exercise consists of practising one technique following the trainer's instructions. A technique will require different skills, so trainer can propose the same technique for different sub-objectives. The instructions will then stress one particular aspect of the technique regarding the sub-objective to be reached. Consequently, our application must (1) propose pertinent sub-objectives in function of the objective decided by the trainer and the level the students have reached in their previous lessons, (2) provide aikido techniques with instructions and durations as regards each sub-objective and the abilities students have already acquired, (3) ensure that each technique is not proposed more than once during a given lesson and (4) require all the lessons in the same cycle (to reach a single objective) to be different. In addition, the trainer must (5) determine the

importance and the time to be spent on each sub-objective and aikido technique, (6) quantify the degree of assimilation by students for each objective and sub-objective, and (7) modify some parts of the proposed lessons at will. Thus, we have created two types of CBR, and our system deals with four types of case:

- CBR that proposes sub-objectives regarding objectives:
  - Source cases noted $s=(O, \cup\{(SO, D^{S}_{SO,O})\})$ where objective O is the problem part of s, and the set of sub-objectives SO with the duration SO must be practised during the season (noted $D^{S}_{SO,O}$) in order to reach O, and is the solution part of s noted $\cup\{(SO, D^{S}_{SO,O})\}$;
  - Target cases noted $t=(O, \cup\{(SO, D^{T}_{SO,O})\})$;
- CBRs that propose exercises regarding sub-objectives:
  - Source cases noted $\sigma=(SO, \cup\{(EX, INSTR_{EX,SO}, CD^{\sigma}_{EX,SO}, RD^{\sigma}_{EX,SO})\})$: SO is the problem part of σ, and its solution part is composed of EX which is the proposed exercise, $INSTR_{EX,SO}$ which is the set of instructions to give to the students when EX is proposed in order to reach SO, $CD^{\sigma}_{EX,SO}$ which is the duration EX must be practised when proposed in regards of SO (CD stands for Constant Duration), and $RD^{\sigma}_{EX,SO}$ which is the duration EX must be practised during the season in order to reach SO (RD stands for Remaining Duration);
  - Target cases noted $\tau=(SO, \cup\{(EX, INSTR_{EX,SO}, CD^{\tau}_{EX,SO}, RD^{\tau}_{EX,SO})\})$.

Actually, durations $D^{S}_{SO,O}$ and $RD^{\sigma}_{EX,SO}$ give the priorities of each sub-objective and exercise, priorities which increase with the time available. The initial durations are given by the trainer, but further study will base these values on pre-tests as proposed in different approaches (Tan, Shen, & Wang, 2012), (Huang, Huang, & Chen, 2007). The system must be initialised each year, at the beginning of the season. Further investigation will focus on its initialisation process.

System architecture and communication model

MAS constitute a paradigm designed to handle distributed systems. They are the product of AI research and reflect its limits: a single AI representing the behaviour of a unique entity cannot deal with collective behaviour. Thus the idea of intelligence distribution emerges and so one can speak of Distributed Artificial Intelligence (DAI). In a MAS, an agent is a physical or abstract entity with some specific characteristics: a perception of its environment (including itself and the other agents), a capability to act (upon itself or upon the environment) and an autonomy in its decisions and actions. To design a MAS is not only to design the different agents but the environment too. As explained in the previous section, the choice of the sub-objectives regarding an objective is an autonomous process, as well as the determination of the exercises regarding a sub-objective, the other exercises chosen and their priority level. Each process is based on specific rules and reasoning. In addition, it must interact with the other processes and take their choices into account. Thus, each process must be autonomous, make decisions, infer changes as to the choices made by the others, be aware of its environment, communicate and interact with the others. Consequently, we can call them agents. As shown in Figure 1, the system is composed of four types of agent: the trainer agent, the student agents, the Objective Agent (OA) which is responsible for choosing the sub-objectives regarding an objective requested by the trainer, and the exercise agents. Each of these agents is responsible for proposing the exercises the best suited to one sub-objective. A single sub-objective is given by the OA or by another Exercise Agent (EA) to each EA. Each EA must also take into account the choices made by the other EAs: each exercise may be done only once during the

entire lesson. Thus, the choices of the other EAs are shared and a decision policy is designed and tested.

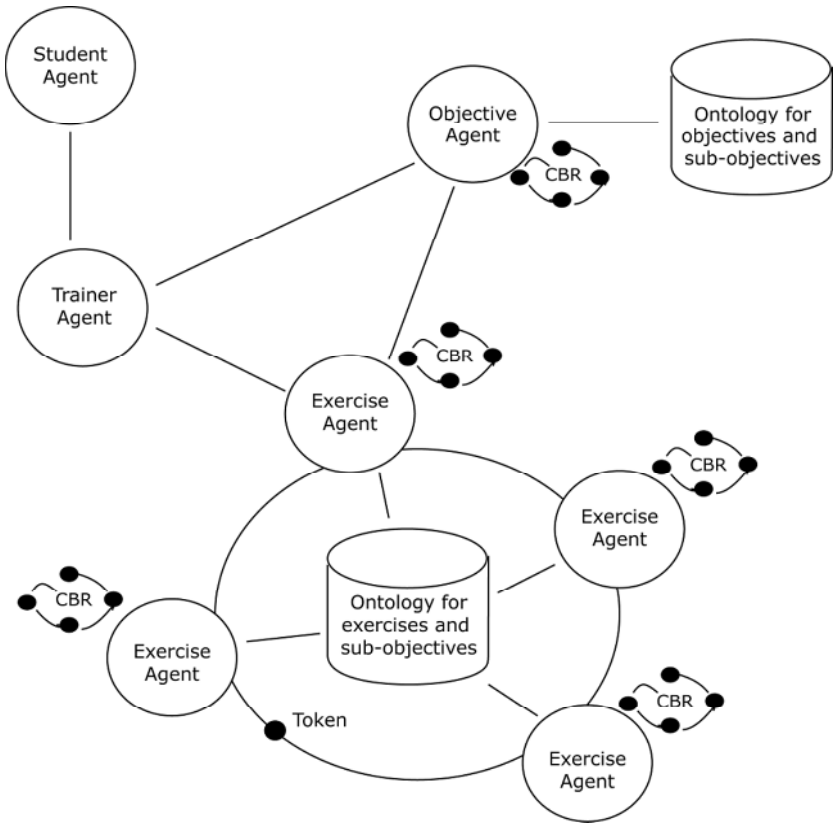

Figure 1. Platform communication model

To share the memory among the EAs, we have chosen the Pilgrim protocol which is an efficient and secured protocol for concurrent, cooperative and collaborative works with shared memory (Garcia, Guyennet, Henriet, & Lapayre, 2006). The agents are dispatched over a ring and a token is exchanged. The originality in this protocol is the fact that the token is carrying each stored modification. Each agent has a copy of the shared memory (i.e. the set of exercises proposed by the other agents). In this protocol, the token is a structured entity that is transmitted from agent to agent and dispatched over a logical ring in an order determined at the beginning of the cooperative work. When an agent wants to modify the shared memory, it must place a reservation flag above the token if there is not yet another flag above it. The token continues its course over the ring and when the agent again receives the token, it is allowed to re-write the modifications. Thus, all the other agents receive these modifications during the next token revolution, after which all the agents have exactly the same version of the shared memory even if two or more agents want to modify it at the same time. The decision policy is implemented over each EA that is able to modify its set of proposed exercises, if one or more of its exercises are identical to any of those found in another set.

Determination of sub-objectives

Once the lesson objective is chosen by the trainer, and after having analysed any additional student needs, the objective agent is responsible for choosing the set of sub-objectives and their duration in function of those which have been previously met and also according to the students' degree of assimilation. This choice is made by the OA following the CBR approach.

Case model

For the OA, a case is a set comprised of two parts, a problem and a solution. As previously presented, each problem part is composed of an objective O, and the solution part of a set of (SO, $D_{SO,O}$) where SO is a sub-objective. The durations $D_{SO,O}$ are updated during the capitalisation process taking into account the remarks made by the trainer during the revision phase. Since the students' levels of expertise rise, we can consider that durations decrease and we thus speak of "remaining durations".

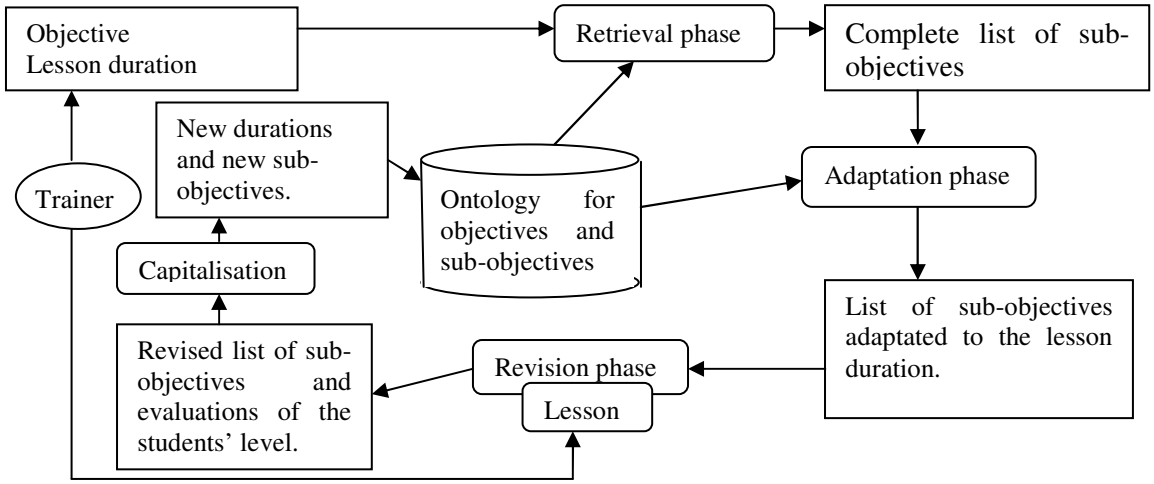

Figure 2. Process of determination of sub-objectives

Retrieval phase

As shown in Figure 2, the retrieval phase is responsible for retrieving and sorting all the possible sub-objectives linked to a single objective. Thus, the retrieval phase consists of retrieving the source case for which the chosen objective is the problem part.

Adaptation phase

As presented in Figure 2, the adaptation phase consists of computing the durations of each sub-objective. Firstly, the adaptation module sorts the set of sub-objectives with their durations so that the sub-objective with the maximum duration comes first and the one with the minimum comes last. Then, the proposed durations of t are calculated. For this computation, the trainer must give the desired number of sub-objectives, the total number of sub-objectives associated with one objective and the total duration of the course respectively noted N_{SO}, $Card(Sol_S)$ and D. The adaptation consists of giving the same duration to each desired sub-objective.

Thus, $t = (O, \bigcup_{n\in\{1..Card(Sols)\}}\{(SO_n, D^{T,n}{}_{SOn,O})\})$ where $D^{T,n}{}_{SOn,O}=\min(D^{S,n}{}_{SOn,O}, (D/N_{SO}))\ \forall\ n \in \{1..N_{SO}\}$, then filling the possible remaining time with sub-objectives of lower priority:

$\forall n\in\{(N_{SO}+1)..Card(Sol_S)\}$, if $(\Sigma_{i=1}^{n-1} D^{T,n}{}_{SOn,O})<D)$

then $D^{T,n}{}_{SOn,O}=\min(D^{S,n}{}_{SOn,O}, (D-\Sigma_{i=1}^{n-1} D^{T,n}{}_{SOn,O}))$,

otherwise $(SO_n, D^{T,n}{}_{SOn,O})$ is removed from the solution part of t.

Revision phase

The revision phase begins just after the adaptation, before the lesson, and finishes afterwards with remarks and evaluation of the students' acquired level of capacity. Consequently, the revision phase is one of the most important in our system after the beginning of the lesson. As

reported in Figure 2, the trainer modifies (adds, updates, removes) the sub-objectives and durations, and following the lesson, the trainer evaluates the skills mastered by the students. After having modified the solution part of *t* and before the lesson begins, each element of *t* is transmitted to one EA that will have to associate the corresponding exercises. After the lesson, each sub-objective duration is modified: for each element of the solution part of *t*, $D^{T,n}_{SOn,O}=D^{T,n}_{SOn,O}$x($level_{SOn}$/10) where $level_{SOn}$ is the evaluation from 0 to 10 of the students' level for the sub-objective SO_n. Thus, the remaining time associated with each sub-objective will decrease slowly as long as students do not reach the required level. In contrast, this remaining duration will decrease quickly once they have reached the expected level.

Capitalisation phase

The first task consists of retrieving the source case *s* for which the selected objective *O* is the problem part, and of adding the sub-objectives that do not yet exist (the ones that may have been added during the revision phase). The system then subtracts the durations of *t* from all the durations of the source cases: for all *SO* in the solution part of *t*, for all *s* for which *SO* appears in the solution part (even if its objective *Os* is not the selected objective *O*), $D^{S}_{SO,Os}$=min(0 , ($D^{S}_{SO,Os}$-$D^{T}_{SO,O}$)).

Example of how lesson sub-objectives are selected

This part presents the method through an example of how an aikido training lesson is planned.

Table 2. Two source cases stored in the system.

Source case	Objective	Sub-objective	Duration (minutes)
1	Using a grip	Breaking the partner's posture	100
		Relaxing despite a grip	100
		Making the partner lose balance	90
		Pivoting around a grip	90
2	Breaking a grip	Breaking a single grip	100
		Relaxing despite a grip	80

Table 3. Chosen sub-objectives and their evaluation by the trainer.

Sub-objective	Duration (minutes)	Trainer's evaluation (/10)
Breaking the partner's posture	30	7
Relaxing despite a grip	30	3
Making the partner lose balance	30	4

Table 4. Cases stored in the ontology objective/sub-objective.

Source case	Objective	Sub-objective	Duration (minutes)
1	Using a grip	Breaking the partner's posture	100-(30x7/10)=79
		Relaxing despite a grip	100-(30x3/10)=91
		Making the partner lose balance	90-(30x4/10)=78
		Pivoting around a grip	90
2	Breaking a grip	Breaking a single grip	100
		Relaxing despite a grip	80-(30x3/10)=71

Table 2 presents two source cases. Assuming, the trainer chooses the objective *O* ='Using a grip' as the main point of a lesson *D* = 90 minutes long and with a minimum of N_{SO} = 3 different sub-objectives. The trainer transmits these parameters to the OA. Obviously, the OA remains case 1. The adaptation process then sorts the sub-objectives according to their durations and designates the solution presented in Table 3, allocating *D* = 90/3 = 30 to each selected sub-objective duration. Hence, assuming the trainer has not modified the list of sub-objectives, the examples of the trainer's evaluations, made during the revision process just

after the course, are reported in the last column of Table 3. Consequently, after capitalisation, the new durations will be as reported in Table 4. Even if a sub-objective is associated with another objective, its duration is diminished for both. This is the case for the last associated sub-objective of source case 2. As shown in this table, the less assimilated sub-objectives ("Relaxing despite a grip" and "Pivoting around a grip") become the most immediate ones. We also note that, as required for the system specification, if the same objective ("Using a grip") is selected again, the less assimilated sub-objective with other sub-objectives will be selected ("Relaxing despite a grip", "Pivoting around a grip", "Breaking the partner's posture"). Thus, as required, the proposed solutions will change even if the same objective is requested again later.

Selection of exercises

This sub-section presents how the exercises are chosen regarding the selected sub-objectives.

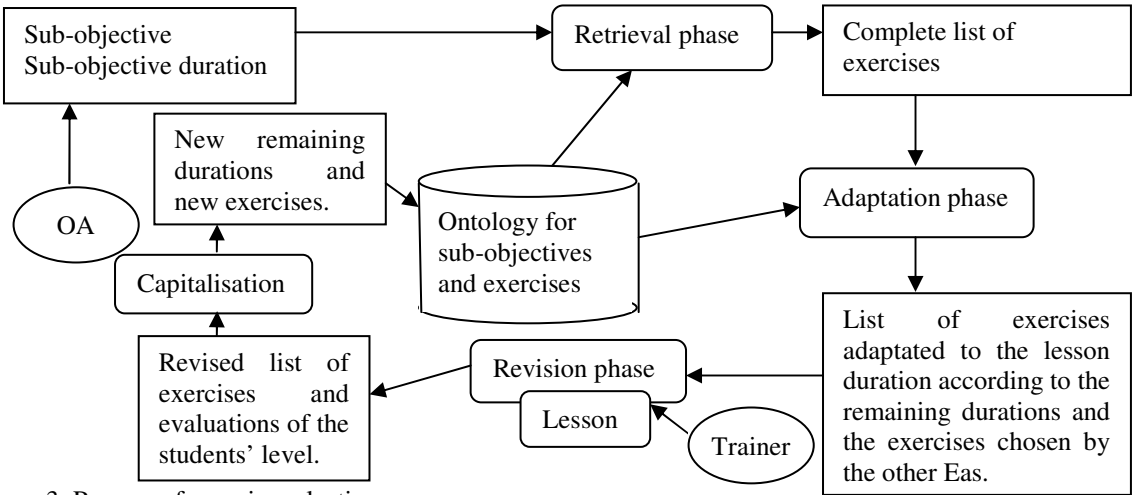

Figure 3. Process of exercise selection

Case model

For this part of the system the problem part of a case is composed of a sub-objective while the solution part contains a set of exercises with a specified duration to be spent on them. Though the durations for exercises are constant, for sub-objectives the total remaining duration required for each exercise must be specified and will decrease from one lesson to the next. Consequently, as presented above, each source case $\sigma=(SO, \bigcup\{(EX, INSTR_{EX,SO}, CD^{\sigma}_{EX,SO}, RD^{\sigma}_{EX,SO})\})$ contains the exercises possible regarding SO. Assuming $Card(Sol_{\sigma})$ is the number of exercises of the solution part of σ, the target case τ_i taken into account by the EA EA_i is noted $\tau_i=(SO_i , \bigcup_{n\in\{1..Card(Sol\sigma)\}}\{(EX_n , INSTR_{EXn,SOi} , CD^{\tau i}_{EXn,SOi} , RD^{\tau i}_{EXn,SOi})\})$. The different steps of the exercise selection process are reported in Figure 3.

Retrieval phase

EA_1 is the EA initiated by the trainer. Hence, its role also consists of initiating as many EAs as required, and creating the token. Each EA_i then retrieves the source case corresponding to SO_i.

Adaptation phase

The adaptation phase requires the EAs to communicate with each other in order to associate exercises with sub-objectives sharing their solutions according to the Pilgrim protocol (Garcia, Guyennet, Henriet, & Lapayre, 2006). The number of exercises selected depends on the

duration required. In addition, we assume that the exercises with the highest remaining durations are the ones the most important to practice, or must be practiced regularly to reach a sub-objective. For this reason, the exercises are decreasingly ordered according to their remaining duration and each one is added until the sum of constant durations reaches the required level for sub-objective SO_i, i.e. $(\Sigma_n CD^{\tau i}_{EXn,SOi}) \geq D^{T,i}_{SOi,O}$. Before placing its list of exercises on the token, EA_i verifies whether each exercise has been selected with a higher or equal duration by another EA. If so, EA_i removes its exercise and replaces it with the next unselected one. EA_1 acknowledges when all the sub-objectives have been provided with exercises and acknowledges again when the list is complete. Then, when EA_1 receives the token again, it verifies whether any modification has been made by the other EAs. The token continues to travel around the ring until EA_1 receives it with the complete list acknowledged and with no further modification. During the process of selection of SOs, the trainer may have created a new sub-objective which will not be associated with any exercise. If so, the EA may confirm that no exercise was found for this SO by using the exercise named *LackOfExercise* with the duration required. Also, the EA may not have enough remaining exercises for the duration period and can also use *LackOfExercise* for this.

Revision phase

The OA performs a revision phase, but the trainer can also create, modify, replace or remove exercises, their durations and their instructions before the lesson. After the course, the trainer evaluates student levels. The same computations as those applied to *SO* determination are applied to the exercise durations of the ontology:

$\forall i$, $\forall n \in \{1..N^{EX}_i\}$, $CD^{\tau i}_{EXn,SOi} = CD^{\tau i}_{EXn,SOi}$ x $(level_{EXn}/10)$.

Capitalisation phase

Table 5. Cases stored in the ontology sub-objective/exercise.

Sub-objective	Aikido technique	Constant duration (min.)	Remaining duration (min.)
Breaking the partner's posture	Ryotedori Tenchinage	10	50
Breaking the partner's posture	Katatedori Kokyuhoo	10	50
Breaking the partner's posture	Katateryotedori Kokyuhoo	10	40
Breaking the partner's posture	Aihamikatatedori Ikyo	10	40
Breaking the partner's posture	Aihamikatatedori Iriminage	10	30
Relaxing despite a grip	Katateryotedori Kokyunage	10	50
Relaxing despite a grip	Katateryotedori Kokyuhoo	10	50
Relaxing despite a grip	Katateryotedori Udekimenage	10	40
Making the partner lose balance	Katatedori Kokyunage	10	50
Making the partner lose balance	Aihaminkatatedori Ikyo	10	50
Making the partner lose balance	Ushiroryotedori Kokyunage	10	50

Table 6. Exercises initially selected by EA_1, EA_2 and EA_3.

Agent	Aikido technique	Constant duration (min.)	Remaining duration (min.)
EA_1	Ryotedori Tenchinage	10	50
EA_1	Katatedori Kokyuhoo	10	50
EA_1	Katateryotedori Kokyuhoo	10	40
EA_2	Katateryotedori Kokyunage	10	50
EA_2	Katateryotedori Kokyuhoo	10	50
EA_2	Katateryotedori Udekimenage	10	40
EA_3	Katatedori Kokyunage	10	50
EA_3	Aihamikatatedori Ikyo	10	50
EA_3	Ushiroryotedori Kokyunage	10	50

Table 7. Modified list of exercises of EA_1.

Agent	Aikido technique	Constant duration (min.)	Remaining duration (min.)
EA_1	Ryotedori Tenchinage	10	50
EA_1	Katatedori Kokyuhoo	10	50
EA_1	Aihamikatatedori Iriminage	10	30

During the capitalisation process, any new exercises possible are added, as previously, to the corresponding case. Then, the same kind of computation as before is applied to each exercise, whatever the associated SO: $\forall SO$, $\forall EX$, $RD^{\sigma}_{EX,SO}=\min(0 , (RD^{\sigma}_{EX,SO} - CD^{\tau}_{EX,SO}))$.

Table 8. Exercises finally transmited to the trainer.

Sub-objective	Aikido technique	Duration (min.)
Breaking the partner's posture	Ryotedori Tenchinage	10
Breaking the partner's posture	Katatedori Kokyuhoo	10
Breaking the partner's posture	Aihamikatatedori Iriminage	10
Relaxing despite a grip	Katateryotedori Kokyunage	10
Relaxing despite a grip	Katateryotedori Kokyuhoo	10
Relaxing despite a grip	Katateryotedori Udekimenage	10
Making the partner lose balance	Katatedori Kokyunage	10
Making the partner lose balance	Aihamikatatedori Ikyo	10
Making the partner lose balance	Ushiroryotedori Kokyunage	10

Table 9. Cases stored in the ontology sub-objective/exercise after revision.

Sub-objective	Aikido technique	Constant duration (min.)	Trainer evaluation (/10)	Remaining duration (min.)
Breaking...	Ryotedori Tenchinage	10	8	50 – 8 = 42
Breaking...	Katatedori Kokyuhoo	10	6	50 – 6 = 44
Breaking...	Katateryotedori Kokyuhoo	10		40 – 4 = 36
Breaking...	Aihaminkatatedori Ikyo	10		40 – 5 = 35
Breaking...	Aihamikatatedori Iriminage	10	7	30 – 7 = 23
Relaxing...	Katateryotedori Kokyunage	10	3	50 – 3 = 47
Relaxing...	Katateryotedori Kokyuhoo	10	4	50 – 4 = 46
Relaxing...	Katatryotedori Udekimenage	10	2	40 – 2 = 38
Making...	Katatedori Kokyunage	10	4	50 – 4 = 46
Making...	Aihamikatatedori Ikyo	10	5	50 – 5 = 45
Making...	Ushiroryotedori Kokyunage	10	3	50 – 3 = 47

Example of exercise selection for one sub-objective

In this subsection we consider the example presented in the previous section and the selected sub-objectives with durations reported in Table 3. Table 5 presents certain sets of aikido techniques, called by their Japanese name, that correspond to each of the selected sub-

objectives. For greater clarity we have not reported the instructions for each exercise in the example, assuming they were of no help in understanding how the distributed system works. Thus, three EAs are called by the OA: EA_1 is the corresponding agent that must deal with the sub-objective 'Breaking the partner's posture'. It places the two remaining sub-objectives on the token which it transmits to EA_2, in charge of the sub-objective 'Relaxing despite a grip'. EA_2 then transmits the token with the last sub-objective to EA_3 which will have to manage 'Making the partner lose balance'. By the time the token returns to EA_1, the latter will have selected the techniques and durations reported in Table 6. The agent places them on the token and transmits it to EA_2 which, upon receiving the token, has already selected first the techniques reported in Table 6. It is worth noting that EA_1 and EA_2 have both chosen 'Katateryotedori Kokyuhoo', but since the remaining duration of this technique for the EA_2 sub-objective (50) is greater than the remaining duration for the EA_1 sub-objective (40), EA_2 places it and EA_1 will be the agent that must replace it. The token is then transmitted to EA_3. Before the token arrives, EA_3 has selected the techniques reported in Table 6. None of the selected techniques has been used by the other agents, thus EA_3 can place them on the token and transmit them to EA_1. When EA_1 receives the token, it remarks that it must change one of the techniques that it has proposed. Thus, it selects the immediately preceding and most available technique instead and adds a complete list of acknowledgements to the token before transmitting it to EA_2. The techniques finally proposed by EA_1 are presented in Table 7. Since each technique figures only once, the lesson presented in Table 8 is transmitted to the trainer. After the lesson, the trainer must evaluate the assimilation of each technique proposed. Assuming the evaluations as reported in Table 9, the new remaining durations are computed for each exercise and stored in the ontology by EA_1.

Results

In this section we present the application's performances and how they were evaluated. The system was evaluated for both of the criteria for which it was designed: its ability to propose varied courses (sub-objectives and exercises), and to propose pertinent courses (sub-objectives and exercises). Hence, five qualified aikido trainers evaluated 15 lessons proposed by the implemented system with the same objective ('Use a grip'). They previously entered 6 corresponding sub-objectives and their initial durations. They also entered 5 techniques per sub-objective aimed at the sub-objectives previously entered, along with their durations. They simulated the skill level of a group of students to whom the course was proposed. We then analyse in greater detail the contents of the lessons of one of the trainers.

Figure 4 shows the scores given to the lessons proposed by our application to each trainer. These scores are based on two criteria: (1) the pertinence of the chosen sub-objectives regarding the lessons, along with previously given scores, and (2) the pertinence of the exercises (techniques) determined by the system. We also asked the trainers to take into account the variety of the sub-objectives and exercises. Ten points were given to very satisfying lessons, whereas poorly satisfying courses were given a score of 0. Most of the lessons have scores from 5 to 9 points. The mean scores given to the courses by the trainers vary from 5.7 to 7.7 points. Most of the trainers gave lower scores (4 points) to the third lesson which they felt to be too similar to the two previous lessons. After any too-repetitive a lesson, the system always proposed a very different one, which was much appreciated by the trainers. This is particularly highlighted by the scores given by the third trainer who did not initialise the system with enough exercises and sub-objectives and was increasingly disappointed by the courses proposed from lesson 9 to lesson 15. The trainers were also disturbed by the order of

presentation of the exercises and sub-objectives. Actually, there is an internal order of presentation of the sub-objectives and techniques for each lesson that the system does not yet take into account. This will be the subject of a future study.

Figure 4. Scores given by five aikido trainers to 15 lessons proposed by the system

For each lesson we have listed the sub-objectives (Table 10) and techniques (Table 11) proposed by the system to Trainer 1. All of the sub-objectives appear with about the same frequency (7 or 8 times each). The sub-objectives most often chosen for the first three lessons (’Break partner’s posture’ and ’Do Irimi despite a grip’) were the ones with the longest durations. This table clearly shows there is a variation and turn-over in the chosen sub-objectives. Nevertheless, for the first 6 lessons, this trainer gave the same score to all of the sub-objectives of each lesson. The use of different scores for the sub-objectives in the following lessons implied that these should be mixed and varied.

Table 10. Sub-objectives chosen by the system.

Objective: Use a grip						
Lesson number	Break partner’s posture	Do Irimi despite a grip	Move despite a grip	Pivot despite a grip	Relax despite a grip	Make lose balance
1	Chosen	Chosen	Chosen			
2	Chosen	Chosen		Chosen		
3	Chosen	Chosen	Chosen			
4				Chosen	Chosen	Chosen
5				Chosen	Chosen	Chosen
6	Chosen	Chosen	Chosen			
7				Chosen	Chosen	Chosen
8	Chosen	Chosen	Chosen			
9			Chosen		Chosen	Chosen
10		Chosen		Chosen		Chosen
11	Chosen		Chosen		Chosen	
12				Chosen	Chosen	Chosen
13	Chosen	Chosen	Chosen			
14				Chosen	Chosen	Chosen
15		Chosen	Chosen			Chosen

Table 11. Number of times each technique was chosen.

Aikido technique	Number of times of appearance (lesson numbers)	Shannon index	Aikido technique	Number of times of appearance (lesson numbers)	Shannon index
Aihamikatatedori Ikyo	7 (1-5;8;13)	0.85	Katateryotedori Kokyuhoo	8 (1-5;7;9;12)	0.98
Ryotedori Ikyo	7 (1;3;6;8;9;11;15)	0.85	Katateryotedori Udekimenage	5 (2;4;5;7;12)	0.61
Katatedori Ikyo	3 (6;10;15)	0.37	Katadorimenushi Kokyunage	5 (4;7;9;12;14)	0.61
Katatedori Kokyuhoo	7 (1-3;6;8;11;13)	0.85	Maeryokatadori Iriminage	4 (3;8;13;15)	0.49
			Maeryokatadori Kokyunage	1 (15)	0.12
Katadori Shihonage	4 (4;10;12;14)	0.49	Katadorimenushi Kotegaeshi	5 (5;7;10;12;15)	0.61
Katadori Kokyunage	7 (1-3;6;8;10;15)	0.85	Katateryotedori Shihonage	2 (11;14)	0.24
Katateryotedori Ikyo	3 (2;9;14)	0.37	Aihamikatatedori Iriminage	4 (1;6;10;13)	0.49
Ushiroryotedori Ikyo	4 (6;11;13;14)	0.49	Katateryotedori Kokyunage	7 (1;4;5;7;9;11;12)	0.85
Yokomenushi Ikyo	4 (6;8;11;13)	0.49	Katateryotedori Kotegaeshi	2 (11;14)	0.24
Ryotedori Kotegaeshi	7 (1;3;6;8;9;13;15)	0.85	Ushiroryotedori Iriminage	5 (5;8;9;11;13)	0.61
Katadori Iriminage	3 (5;10;12)	0.37	Ushiroryotedori Kokyunage	6 (4;7;9;10;12;15)	0.73
Ryotedori Tenchinage	7 (1-3;6;8;11;13)	0.85	Ushirokatatedorikubishime Sankyo	3 (7;12;14)	0.37
Katatedori Kokyunage	9 (2-5; 7; 9; 10; 14; 15)	1.10	Katatedori Udekimenage	6 (2;4;5;7;10;14)	0.73

We also discussed with the trainers the fact that some lessons mainly proposed the same sets of sub-objectives. They did not consider this to always be a drawback since a student must practise again and again to reach an important sub-objective. Some sub-objectives are thus repeated in the beginning of the training cycle; progressively, as soon as possible, other sub-objectives are wisely introduced and practised more and more at the end of the cycle. The trainers considered this to be satisfying for the selection of sub-objectives, but thought that it would have been a real problem if the selected sets of exercises had been nearly identical from one lesson to another. However, all the sets of exercises proposed by our application were different. Table 11 shows the number of times each exercise (aikido technique) was chosen and the lesson numbers in which they appeared. As shown in this table, all of the techniques were selected once or more, which reveals a good turn-over of exercises from one lesson to another. Techniques chosen once, twice or three times were those with the lowest scores and that appeared once in the sub-objectives with the lowest duration. According to the aikido teachers, some techniques are more fundamental than others to each sub-objective. The teacher gave a higher score to these techniques and they were actually chosen 6, 7, 8 or 9 times: 'Katatedori Kokyunage' to help students learn the way to make a partner lose his/her balance and how to pivot round a grip. 'Katateryotedori Kokyuhoo' is of great help in learning how to break a partner's posture and also how to move and relax despite a grip, etc. All techniques highly useful in learning multiple abilities are in this category, followed, to varying degrees, by those associated with just one sub-objective. In addition, in order to quantify the diversity of the solutions proposed by our application, we computed the Shannon index (Shannon, 1948) of each selected exercise regarding the lessons in which it has been proposed. Considering that the application delivered 15 lessons mixing 27 aikido techniques, and assuming N_i is the number of times of appearance of the aikido technique i, the Shannon index H' is computed according to the equation: $H' = -\Sigma_{(i=1)}^{15}(N_i/27) \times ln(N_i/27)$. These results are reported in Table

11. The indices vary from 0.12 to 1.10. Actually, the aikido trainers confirmed that those associated with the less important entropies were not often used every season whereas the techniques associated to the greatest values were quasi-systematically used for different reasons inherent in this martial art. The mean and median Shannon entropy is equal to 0.61 and quantifies the uncertainty associated with the prediction of retrieving each technique in one of the generated lessons.

Discussion

The results presented above prove that by taking into account previous experiences, the system and our method are capable of proposing pertinent and varied lessons. An important part of our method resides in the agents' capacity in real time to take into account remarks made by users during the revision processes and also to adapt their own solutions to those concurrently provided by other agents of the distributed system. This was possible only due to the use of a concurrency management protocol designed to share real-time memory and concurrent actions. Indeed, drawing attention to the solutions would have introduced concurrent and inconsistent data and lessons since there is no global schedule in collaborative and distributed systems. This particular aspect of distributed CBR systems will be examined through further study. One limit to our system lies in the fact that it must be initialised each year at the beginning of the season or whenever there is no remaining duration left. At that time, the trainer must produce all of the initial values. Additional investigation will propose a process based on experience in order to compute initial values. Most of the other approaches use pre-tests (Huang, Huang, & Chen, 2007), (Tan, Shen, & Wang, 2012) that evaluate students' levels. Another limit to our system is found in the way creations of sub-objectives are managed: In a future investigation we will focus on the way an EA might propose exercises for a newly created sub-objective in function of the exercises proposed for similar sub-objectives. At the very least, the trainers must take time to again sort through the sub-objectives and to examine the order of the exercises proposed by the system: after having selected them, it may be of interest to take into account another parameter, closer to the nature of the exercises in order to better classify them. Finally, like other approaches (Cordier, Fuchs, & Mille, 2006), (Craw, Wiratunga, & Rowe, 2006), (Lieber, 2007), (Dufour-Lussier, Le Ber, Lieber, & Nauer, 2013), and (Henriet, Leni, Laurent, & Salomon, 2014), ours establishes a link between adaptation and capitalisation of revisions. Indeed, we have examined a way to use the remarks of users made during the revision phase in order to enhance the accuracy of the adaptation process of CBR-systems.

Conclusion

This paper presents a multi-agent system that can generate and personalise lessons. The design is based on agents that use CBR systems to propose varied courses. It emphasises the importance of the revision process and its consequence on the adaptation phase. During this revision process, the user furnishes information as to the level that students have attained while the system stores the proposed lesson in order to enhance the adaptation of future lessons. Though AI systems learn to always produce the same solutions to the same problems, our system is capable of proposing a variety of solutions. It is capable of proposing sports training lessons that are diverse, changing, pertinent and consistent, enabling students to practise the same given skill over many weeks. Ours is a distributed system that shares memory. A collaboration protocol based on a token ring has been adapted and used, as well as a policy capable of merging and arbitrating between concurrent solutions proposed by the collaborative agents during their adaptation process. In a future study we will focus on the order in which the

exercises and sub-objectives appear in lessons and the enhancement of the initialisation process since the pertinence of the solutions proposed depends heavily on this order of appearence and the initial values stored. Introducing processes such as the pre-testing of students may improve the quality of the adapted solutions generated.

References

Baylari, A., & Montazer, G. A. (2009). Design a personalized e-learning system based on item response theory and artificial neural network approach. *Expert Systems With Applications 36*, 8013-8021.

Cordier, A., Fuchs, B., & Mille, A. (2006). Engineering and learning of adaptation knowledge and Case-Based Reasoning. *Springer-Verlag*, 303-317.

Craw, S., Wiratunga, N., & Rowe, R. C. (2006). Learning adaptation knowledge to improve case-based reasoning. *Artificial Intelligence 170*, 1175-1192.

Dufour-Lussier, V., Le Ber, F., Lieber, J., & Nauer, E. (2013). Automatic case acquisition from texts for process-oriented case-based reasoning. *Information Systems 40*, 153-167.

Garcia, E., Guyennet, H., Henriet, J., & Lapayre, J. C. (2006). Towards an optimistic management of concurrency: A probabilistic study of the Pilgrim protocol. (W. e. Shen, Hrsg.) *Computer Supported Cooperative Work in Design 3865*, 51-60.

Garcia-Pallares, J., Garcia-Fernandez, M., Sanchez-Medina, L., & Izquierdo, M. (2010). Performance changes in world-class kayakers following two different training periodization models. *European Journal of Applied Physiology 110*, 99-107.

Graesser, J. L., Conley, M. W., & Olney, A. (2012). Intelligent tutoring systems. (K. R. Harris, S. Graham, & T. Urdan, Hrsg.) *The APA Educational Psychology Handbook 3*, 451-473.

Henriet, J., Leni, P. E., Laurent, R., & Salomon, M. (2014). Case-based reasoning adaptation of numerical representations of human organs by interpolation. *Expert Systems With Applications 40*, 260-266.

Huang, M. U., Huang, H. S., & Chen, M. Y. (2007). Constructing a personalized e-learning system based on genetic algorithm and case-based reasoning approach. *Expert Systems With Applications 33*, 551-564.

Issurin, V. (2010). New horizons for the methodology and psychology of training periodization. *Sports medicine 40* (3), 189-206.

Jamsandekar, P. P., & Patil, M. K. (2013). Online learning - CBR approach. *International Journal of Research in Computer Science and Information Technology 1* (A), 111-113.

Kolodner, J. L. (1993). Case-Based Reasoning. *Morgan Kaufmann Publishers* .

Kolodner, J. L., Cox, M. T., & Gonzales-Calero, P. A. (2005). Case-based reasoning-inspired approaches to education. *The Knowledge Engineering Review 20*, 299-303.

Kolodner, J. L., Owensby, J. N., & Guzdial, M. (2004). Case-based learning aids. (D. H. Jonassen, Hrsg.) *Handbook of Research for Educational Communications and Technology*, 829-861.

Lieber, J. (2007). Application of the Revision Theory to Adaptation in Case-Based Reasonning: the Conservative Adaptation. *7th International Conference on Case-Based Reasoning ICCBR 2007 4626,* 239-253.

Matveev, L. P. (1965). Periodization of sports training. *Fizkultura i Sport* .

Plaza, E., & Mc Ginty, L. (2005). Distributed case-based reasoning.

Rishi, O. P., Govil, R., & Sinha, M. (2007). Agent based student modeling in distributed CBR based intelligent tutoring system. *Proceeding of the World Congress on Engineering and Computer Science WCECS 2007*, 473-477.

Rishi, O. P., Govil, R., & Sinha, M. (2007). Distributed case-based reasoning for intelligent tutoring system: An agent based student modeling paradigm. *World Academy of Science, Engineering and Technology 5*, 273-276.

Ronnestad, B. R., Ellefsen, S., Nygaard, H., Zacharoff, E. E., Vikmoen, O., Hansen, J., et al. (2012). Effects of 12 weeks of block periodization on performance and performance indices in well-trained cyclists. *Scandinavian Journal of Medicine and Science in Sports 24*(2), 327-335.

Schank, R. C., Fano, A., Bell, B., & Jona, M. (1994). The design of goal-based scenarios. *Journal of the Learning Sciences 3* (4), 305-346.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal* , 379–423 and 623–656.

Tan, X. H., Shen, R. M., & Wang, Y. (2012). Personalized course generation and evolution based on genetic algorithms. *Journal of Zhejiang University-SCIENCE C (Computers and Electronics) 13*(12), 909-917.

Extracting Individual Offensive Production from Baseball Run Distributions

Stephen J. Robinson

Department of Chemistry and Physics, Belmont University, Nashville, TN, USA

Abstract

Advanced statistical measures are increasingly popular in sports, especially baseball. These measures are used for reasons from contract negotiations to roster changes to award determinations. One of the most popular methods of determining a player's worth in baseball uses WAR: wins above replacement. The methods described herein work to fine-tune the goals of the offensive component of WAR by evaluating players in the context of their own team's lineup. This is done by extracting a player's worth from his team's simulated run distribution using win probabilities.

KEY WORDS: BASEBALL, OPTIMIZATION, WEIBULL, DISTRIBUTIONS

Introduction

Since the inception of Major League Baseball's (MLB) Most Valuable Player (MVP) award in 1931, there has been uncertainty around the definition of "most valuable," namely with regard *to whom a player is most valuable* (e.g., his team, his sport, fans, his team's owner, etc.) and, if such an entity were decided upon, *what criteria would be used to decide said value* (e.g., offensive or defensive statistics, leadership ability, ability to sell tickets and merchandise, etc.). Throughout the history of the award, the general trend has been to give the award to a player on an above-average team who could also be regarded as one of the best offensive players in the league. The historical edge for players on good teams indicates that voters have viewed wins and not just raw individual statistics as valuable. Meanwhile, pitchers have been largely ignored because they are eligible to receive the Cy Young award, and defense has been overlooked because of the difficulty of statistical analysis (Jensen, Shirley, & Wyner 2009).

The debate over the 2012 American League MVP award brought to light a growing division in opinions regarding the analysis of player performance. Awarded for the first time since 1967, Miguel Cabrera of the Detroit Tigers won the American League Triple Crown by having the highest totals in batting average, home runs, and runs batted in. For many, these statistics are the standard measures of player performance. However, the past decade has witnessed an explosion of computer-generated statistics that claim to objectively value a player's overall worth (and hence, salary) through offensive and defensive considerations. The most notable of these statistics is wins above replacement (WAR), which calculates how many wins a player is estimated to have contributed to his team per year compared to a minor-league replacement player at the same defensive position. In 2012, Cabrera had an exceptional WAR of 7.2 (baseball-reference.com) but paled in comparison to the Los Angeles Angels' Mike Trout's WAR of 10.8, one of the highest in modern baseball history. Cabrera ultimately won the MVP award, but the result was debated by media and fans alike.

Despite WAR's breadth (Ormiston 2012), relative objectivity (Catania 2013), and popularity

(Eder 2013), it suffers from two main flaws. First, its three most popular creators—Baseball Reference, Baseball Prospectus, and FanGraphs—do not concur on the details of its calculation. Due to small disagreements on the importance of certain statistical measurements and coefficients, these three, for example, gave Mike Trout a 2012 WAR of 10.8, 9.0, and 10.0, respectively. The difficulty of quantitatively measuring "value" may never be fully overcome, and the complexity and volume of data needed to calculate WAR have made it a paradoxically simple yet incomprehensible statistic to the average fan.

Second, despite its name, using WAR to measure the number of wins a player contributes to his team is slightly misguided. It is an excellent tool to compare two players' abilities or a player's trade value or salary, but apart from adjustments to playing in certain stadiums, it is computed in a vacuum as if a player were not on a team. Of course, WAR is popular for this very reason, and it rightly ignores statistics that depend directly on one's teammates such as runs scored and runs batted in. However, truly understanding how many wins a player contributes to his team requires analyzing his statistics *among his teammates' statistics*. For example, a good base stealer may have a different true value on a team with power hitters than on a team with singles hitters, but would have the same WAR on either team. Alternatively, if a team had a pitching staff that never allowed a run, *every* offensive player would have virtually no effect on the number of wins for the team; that is, a team of high-WAR Hall-of-Famers would win the same number of games (all of them) as a team of low-WAR minor leaguers. This is an absurd case, of course, but it highlights WAR's shortcoming in doing what its name says it does. There are certainly strong correlations between players' WARs on a team and its wins, but similar correlations can be made with several simpler statistics (Petti 2011).

This paper addresses both of these drawbacks by presenting a simple-to-understand method of determining a baseball player's offensive worth to his individual team. It is not intended to be a replacement of WAR, but an alternate view that considers how a player's abilities complement his teammates'.

Methods

To see how a player affects his team offensively, it is crucial to evaluate all possible methods of his contribution. For example, a walk followed by a double is more productive than a double followed by a walk, so, in the batting order, placing a power hitter after a player who often walks is a better strategy than vice versa. In this way, *both* players increase their win value through lineup optimization. Much work (e.g., Hirotsu 2011; Bukiet, Harold, & Palacios 1997; Freeze 1973) has gone into lineup optimization, but for professional teams, the gain—about one win per season (Tango, Lichtman, & Dolphin 2007)—is generally not worth the effort. However, using lineup optimization, not to increase team wins, but to evaluate individual professional players, has been overlooked. By comparing a team's run output when a player is present in a lineup vs. (1) absent from or (2) replaced by another player in a lineup, it is a straightforward task to determine how much said player is worth to his team. For example, if a team averages 4.5 runs per game (RPG) with a given player in an optimized lineup (i.e., the one that produces the most RPG), but only 4.3 without him, he is worth 0.2 offensive RPG to his team. Using probability distribution functions, those 0.2 runs can be directly related to an expected increase in wins per season: the crux of this study.

There are many strategies for lineup optimization, but for the purposes of this article, the method used is relatively unimportant; here, a baseball game simulation algorithm is employed, as detailed by Robinson (2013). This algorithm uses a player's raw statistics to

calculate the probabilities of several batter outcomes that do not strongly depend on one's teammates' abilities: singles (1B), doubles (2B), triples (3B), home runs (HR), strikeouts (SO), walks (BB), times grounded into a double play (GDP), and times hit by a pitch (HBP), as well as runner outcomes: stolen bases (SB) and times caught stealing (CS). Next, it simulates a large set of baseball games with a given lineup and then cycles through all possible lineups to determine the optimal lineup—the one that scores the most runs on average. With n players, there are $n!$ possible lineups; for the eight- and nine-man lineups discussed below, this corresponds to about 4.0×10^4 and 3.6×10^5 possibilities.

Just as a good manager would optimize his lineup to account for roster changes, we need to optimize lineups both with and without a player to compare them accurately. For this study, the 2012 Chicago White Sox were chosen as a sample team due to the consistency of their lineup; they had nine players with at least 500 plate appearances, comprising 83% of their team's plate appearances. Their statistics appear in Table 1.

Table 1. The statistics used to simulate 2012 Chicago White Sox baseball games.

name	G	PA	1B	2B	3B	HR	SO	BB	HBP	GDP	SB	CS
Beckham	151	582	83	24	0	16	89	40	7	10	5	4
De Aza	131	585	103	29	6	9	109	47	9	1	26	12
Dunn	151	649	50	19	0	41	222	105	1	8	2	1
Konerko	144	598	111	22	0	26	83	56	7	16	0	0
Pierzynski	135	520	84	18	4	27	78	28	8	8	0	0
Ramirez	158	621	120	24	4	9	77	16	4	15	20	7
Rios	157	640	114	37	8	25	92	26	4	18	23	6
Viciedo	147	543	85	18	1	25	120	28	6	18	0	2
Youkilis	122	509	67	15	2	19	108	51	17	10	0	0

Naturally, the randomness of a single game simulation can be removed with large numbers of simulations. To determine an appropriate number of games to simulate, the algorithm was tested on a team of identical average MLB players, as determined by averaging all of the non-pitcher statistics for 2012. The results are shown in Figure 1. Simulating 5×10^4 games per lineup (1.8×10^{10} in total for all configurations of a nine-man lineup) substantially reduces random error, so this number was chosen as a compromise between accuracy and computing time.

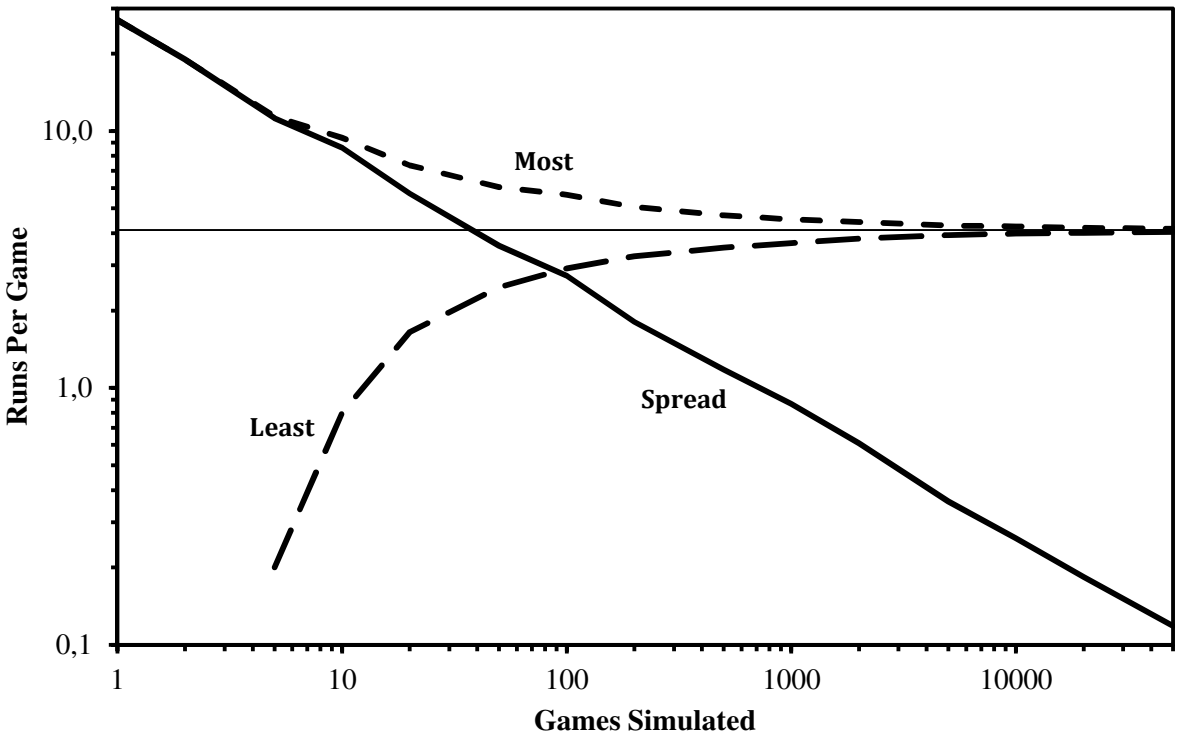

Figure 1. For a team of average players, the most and least runs scored per 27 outs (9 complete innings) as a function of the number of games simulated; both converge to 4.11 RPG. "Spread" is the range between the most and least average RPG.

The actual average number of RPG in MLB in 2012 was 4.32, while Figure 1 settles on the smaller number of 4.11 ± 0.009. The algorithm purposely accounts for neither errors nor run-adding situational strategies such as sacrifice bunts, pinch hits, and pinch runners. These are important for the outcome of a game, but reveal very little about a player's offensive talent. Sacrifice hits and sacrifice flies constituted only 1.5% of plate appearances in 2012, but certainly provided more than that share of runs due to their very nature. In addition, any adjustment to the algorithm to force the RPGs to match (e.g., adding a certain likelihood of bunting) would affect all players equally and is unnecessary to determine comparative player value. Finally, there is no reason to assume that a team of average players would perform the same as the average major-league team with two different averages being represented.

To determine a player's value from different perspectives, three sets of lineup simulations are performed, each with its own advantage:

(1) Remove a player from an optimized nine-man lineup to produce an optimized eight-man lineup. This is obviously not allowed in the rules of baseball, but with continuous batting orders, has no bearing on how the game is played. When compared to the simulated nine-man lineup, it determines the player's offensive contribution to his team.

(2) Replace a player with an average MLB player. This gives a measure of performance closest to the offensive component of WAR, albeit with a comparison to the average player rather than a minor-league replacement; the two would differ by a constant.

(3) Put an individual on a team with eight other average MLB players and compare the results to those of a team comprised of only average players. This is useful to compare

to the results from the first simulation type to see how a player's teammates affect his offensive worth.

Note that none of these simulation methods necessarily keeps any one player in the same position in the lineup; with different rosters, each player has a different collaborative strength and will move around accordingly.

After running these simulations, we need to convert a player's run-addition to a win-addition; thus, understanding how major-league runs are scored is vital. Figure 2 shows the actual distribution of runs over MLB's 2430 games in 2012 alongside simulated games (i.e., the "Model") using average players.

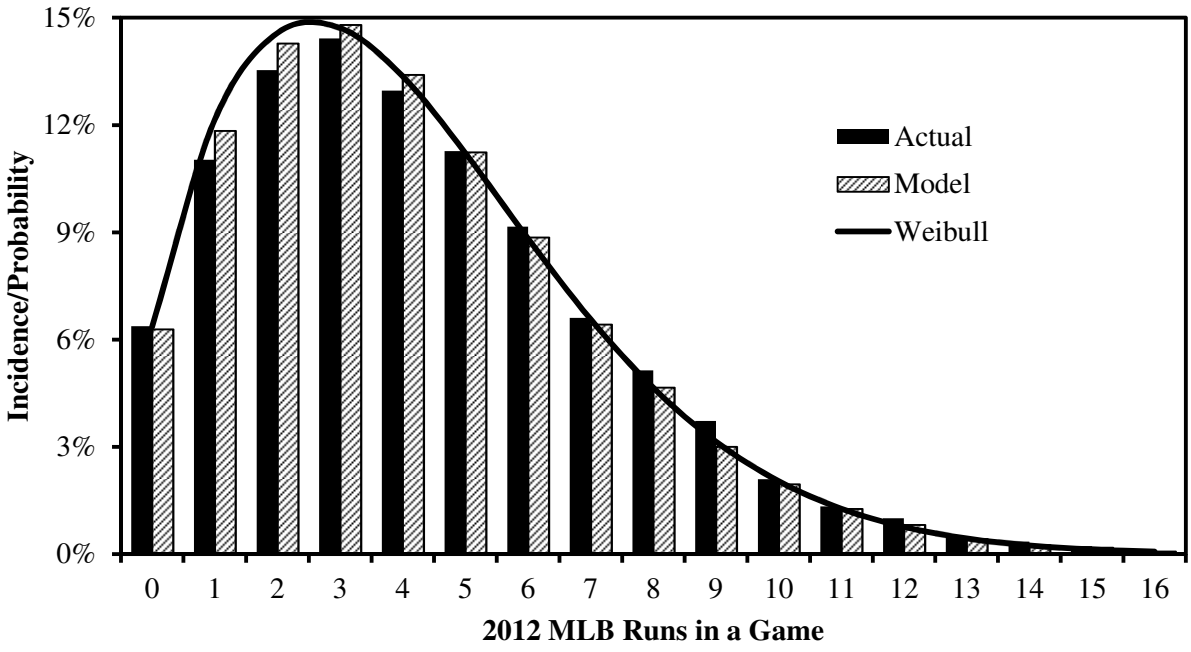

Figure 2. The distribution of RPG in MLB in 2012 as compared to the simulation's results. A Weibull distribution is fit to the model's data. Data points above 16 were left off the graph.

The Weibull distribution is an extremely versatile function that allows a good fit as shown in Figure 2 and quantified in Table 2. Its three-parameter version is expressed as

$$f(x;\alpha,\beta,\gamma) = \begin{cases} \frac{\gamma}{\alpha}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-\left(\frac{x-\beta}{\alpha}\right)^{\gamma}} & x \geq \beta \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

γ is known as the shape parameter and α is the scale parameter. β, the location parameter, simply shifts the function along the x-axis. As pointed out by Miller (2007), the discreteness of baseball runs is best accounted for by setting β = –0.5, and the best fits are shown in Table 2. p is computed from Pearson's chi-square test and represents the probability that differences between the Weibull and actual or modeled distributions are due to chance alone and not due to a fundamental error in fitting; that is, the smaller 1 – p is, the better the fit, so that 10–1 and below represents at least 90% confidence in the fit. For example, there is a p = 99.977% chance that the differences between the modeled average MLB offense and the Weibull distribution are due to chance alone, giving 1 – p = 2.3 × 10–4. This is effectively equivalent to saying there is a 0.023% chance that the Weibull distribution does not adequately explain the data.

Table 2.The mean μ, standard error of the mean SE, standard deviation σ, and best-fit parameters for Weibull run distributions with $\beta = -0.5$. α was calculated from the data set's mean (Miller 2007) and the best fit was determined by varying γ to maximize the p-value between the applicable data set and Weibull distribution.

	μ	SE	σ	α	γ	$1-p$
actual MLB	4.32	0.061	3.00	5.40	1.67	8×10^{-2}
optimized model MLB	4.17	0.013	2.90	5.23	1.68	2×10^{-4}
actual White Sox offense	4.62	0.255	3.25	5.72	1.63	9×10^{-2}
optimized model White Sox offense	4.48	0.013	2.97	5.60	1.76	6×10^{-9}
actual White Sox defense	4.17	0.234	2.98	5.22	1.61	2×10^{-2}

With this probability distribution of runs scored (RS), we can then compare it to one of runs against (RA) to compute the probability of one team beating another. Miller (2007) showed that, using the Weibull distribution, the probability of a team winning a game (i.e., its expected winning percentage) if it scores RS RPG and allows RA RPG is

$$P_{\text{win}} = \frac{(\text{RS} - \beta)^{\gamma}}{(\text{RS} - \beta)^{\gamma} + (\text{RA} - \beta)^{\gamma}} \qquad (2)$$

This is popularly known as the Pythagorean expectation, and was first developed by James (1983) with $\beta = 0$ and $\gamma = 2$. Many different values for γ have been suggested over the years, each dependent on the data set, time period, and fitting method used. Equation (2) uses only one value of γ; here, since comparisons are made using only simulated data, we use the modeled MLB fit ($\beta = -0.5$, $\gamma = 1.68$). (As shown in Figure 3, the best fit of Equation (2) to 2012 MLB actual winning percentages is $\beta = -0.5$, $\gamma = 2.22$. However, there is much more flexibility with γ to acquire a good fit than is commonly assumed; $1 \leq \gamma \leq 3$ produces results within the 95% confidence interval for $-0.5 \leq \beta \leq 0$.)

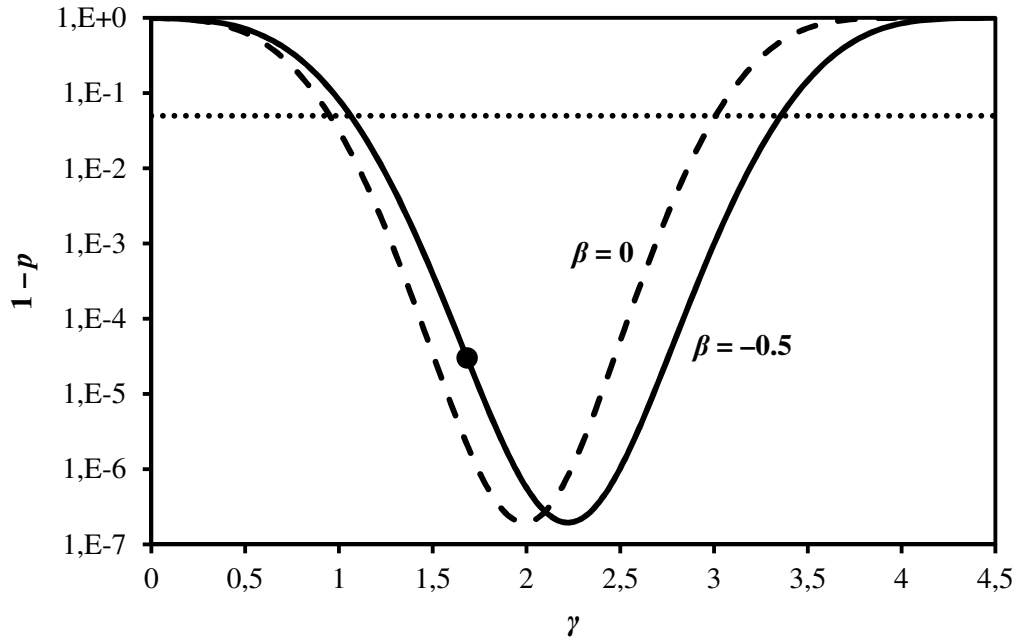

Figure 3. A measure of confidence levels (the difference between the p-value and 1) for the Pythagorean expectation (Equation 2) of 2012 MLB games as a function of the exponent; the best fits are at the bottom of the graph. The horizontal dotted line and all points below it represent at least 95% confidence that the differences in predicted and actual winning percentages of MLB teams is due to chance alone and not an error in the model. The best-fit exponent to the simulated run distribution ($\beta = -0.5$, $\gamma = 1.68$) is shown as a dot.

Now, armed with γ and RA, we can use a team's individual offensive statistics to determine the probability of winning with a certain player in or out of the lineup, thereby determining his offensive worth.

Results

It should be noted here that the difference between simulated and actual RS, when compared to actual RA, will cause a small difference in expected winning percentage. As shown in Equation (2), the winning percentage is a nonlinear function of runs, but this problem can be solved by—on a team-by-team basis—adding a constant to the simulated RS to predict a player's contribution more accurately. Practically, this can be thought of as a team's RPG that have little to do with individual player performance, such as errors by the other team. If not accounted for, this will produce a small error in win probability.

For example, suppose an actual team scores 4.3 RPG and allows 4.1 RPG. Over 162 games, they would, using $\gamma = 1.68$, expect to win 84.2 games. If a simulation predicts that a team will score only 4.2 RPG, the win expectation drops to 82.6. This is an important difference for a team; for an individual player, the effect is much less pronounced. Suppose that removing a player from the lineup drops the simulated RPG to 4.0. The team would then expect to win 79.3 games, meaning that he contributes 82.6 – 79.3 = 3.3 offensive wins per season. If one accounts for the 0.1 RPG difference mentioned above, his team would score 4.1 RPG for 81.0 wins. Then, we would say that he contributes 84.2 – 81.0 = 3.2 offensive wins per season. This 3.3 – 3.2 = 0.1 difference is well within the tolerance of any popular version of WAR, but to maintain the most accuracy, the difference between simulated and actual runs will be accounted for.

The actual 2012 Chicago White Sox averaged 4.62 RPG, gave up 4.17 RPG, and won 85 games. Equation (2) predicts a winning percentage of .542 (87.9 wins). (Recall that we optimized γ for run distributions, not winning percentages, and that, as mentioned, many small random factors can contribute to a win or loss). Meanwhile, the nine-man optimized simulation calculates 4.48 RPG, a 0.133-RPG difference from the actual mean. Simulation (1) then predicts that removing, for example, Paul Konerko from the lineup to produce an eight-man optimized lineup yields a 4.33-RPG team. Including the difference of 0.133, this eight-man team would average 4.46-RPG. The predicted winning percentage of this team with the same defense is then .525. Over a 162-game season, Paul Konerko played 144 games, so he was worth $144 \times (.542 - .525) = 2.50$ wins to his team. Notice that this number is not compared to any "replacement" player, as is done with WAR; rather, it is specific to his team, assuming optimized lineups with him present or absent.

Simulation (2), on the other hand, replaces Konerko with an average MLB player to see how the lineup fares. Using an identical analysis, this shows Konerko to be worth 2.80 wins. The fact that it would be better to remove Konerko from the lineup than replace him with an average player shows that an average player is below average on the White Sox; or, put another way, the White Sox were an above-average team, as confirmed by their winning percentage. Simulation (3) places Konerko in a lineup with eight other average players. A team of average players, as shown in Table 2, would score and allow a simulated 4.17 RPG. (This number is coincidentally the same as the actual White Sox's defensive RPG; they are unrelated.) That is, without Konerko, they would have a winning percentage of .500. Simulation (3) shows that the RPG increases to 4.36 with Konerko, which is adjusted to 4.52 to coincide with the MLB average of 4.32. This predicts a .517 winning percentage, giving Konerko $144 \times (.517 - .500) = 2.38$ wins. The fact that Konerko is worth slightly more to his actual team than he is to a team of average players is most likely an indication that his skillset fits well with his teammates'.

WAR is comprised of three components: batting, fielding, and baserunning. For comparison's sake, as this discussion revolves around batting and base-stealing only, the results herein are shown alongside weighted runs above average (wRAA) and weighted stolen base runs (wSB), the batting and base-stealing components of WAR: ten of either equals one win and represents a measure of performance compared to the league average. (wRAA is computed directly from weighted on-base average (wOBA), a measure of how, historically, different baseball plays—e.g., walks and singles—contribute to runs.) These comparisons are compiled into Table 3 and shown graphically in Figure 4. Keeping in mind that all lineups with and without a player are optimized to produce maximum RPG, positive win values for players are interpreted as follows:

- Simulation (1): A lineup produces fewer wins if the player is removed. This would be the preferred method to determine a player's trade value as it is a measure of how he complements a team. This value is team-dependent.
- Simulation (2) and wRAA + wSB: The team produces fewer wins if an average MLB player replaces the given player. By subtracting the MLB average win contribution, these methods could be used to decide whether to call a minor-league player up to replace said player. The simulation is team-dependent while wRAA + wSB is team-independent.
- Simulation (3): A team of average MLB players wins more games when the player is inserted in the lineup. This simulation could be useful to determine salaries with regard to MLB average salaries. This value is team-independent.

For Simulations (2) and (3), a positive team average of win values is an indication that the

team has above-average talent and/or complementarity (and is probably an above-average team). For Simulation (1), a positive team average indicates that the players play well together (the whole is greater than the sum of its parts), but it can also be a result of the fact that better players also bat more frequently, creating a nonlinear effect of player removal. (However, only the starting 9 of the 27 White Sox batters were analyzed; the remaining 18 had mostly negative wRAA + wSB values, which would, as expected, bring the actual team average closer to zero.) The fact that the White Sox's Simulations (1)–(3) follow the same general trend is a reminder that complementary teams are usually also good teams. Finally, because of these different baselines, when determining player values, one may find it more instructive to directly compare two players' win values rather than to compare an individual player's win value to zero.

Table 3. The offensive wins contributed by players on the 2012 Chicago White Sox using the four methods described in the text. R^2 gives the correlation between the simulation result and wRAA + wSB (fangraphs.com).

Name	Sim. (1)	Sim. (2)	Sim. (3)	(wRAA + wSB)/10
Beckham	–1.14	–0.64	–1.33	–1.07
De Aza	0.13	0.58	0.16	0.81
Dunn	1.77	2.02	1.47	1.55
Konerko	2.50	2.80	2.38	2.60
Pierzynski	1.72	1.81	1.45	1.46
Ramirez	–1.94	–1.27	–1.98	–1.57
Rios	1.95	2.13	1.60	2.54
Viciedo	0.22	0.32	–0.07	0.14
Youkilis	0.66	0.86	0.50	0.61
R^2	0.948	0.959	0.958	

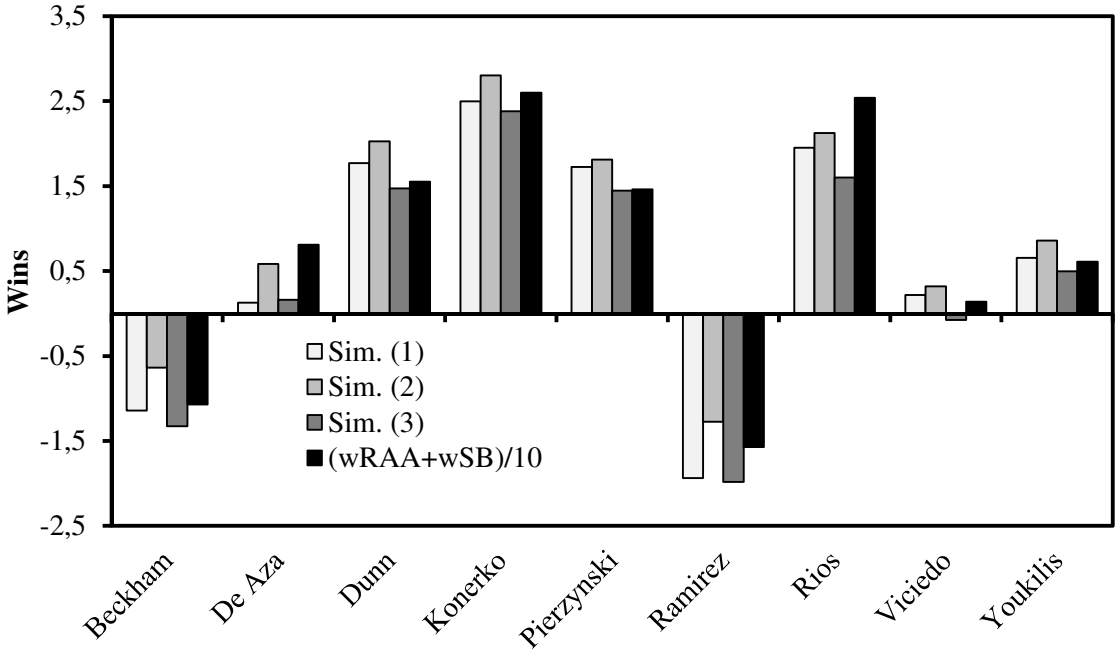

Figure 4. Comparison of win contributions of 2012 Chicago White Sox players using the three simulation methods and wRAA + wSB. Through Equation (2), the standard error of 0.013 about the RPG means in Table 2 produces a standard error of approximately 0.17 for the win values in each simulation. wRAA and wSB are not published with uncertainties, but there is consensus that it is a significant fraction of a win for their composite statistic WAR (Cameron 2013).

Discussion

Table 3 and Figure 4 confirm that these simulation methods are consistent with wRAA + wSB. However, on an individual basis, the simulation results differ from these measures by as much 0.94. (As an example, this analysis provides evidence that previous estimations have overvalued the contributions of De Aza and Rios to the 2012 White Sox.) This is central to the purpose of this study: that is, the simulations, in effect, have the same goal as wRAA + wSB, but rather than using the same set of coefficients for every player, the methods presented herein determine the wins generated by a player in the context of his own and his team's statistics. For instance, wRAA (and thus, WAR) assumes that a double carries the same weight for generating runs in all situations; as mentioned earlier, this is certainly not true when the double precedes-vs.-follows a walk. Thus, we believe this method of simulation is a more accurate representation of an individual player's offensive win contribution in the context of his team than wRAA + wSB.

In addition to baseball, the methodology used herein is general to many team sports. Its main advantage is in the use of simulations in contrast to simply interpreting historical data. First, 5×10^4 games is equivalent to more than 300 entire seasons for a baseball team. This allows us to reduce predictive statistical uncertainty greatly when compared to other statistics such as WAR, which is generally calculated over a single season or a relatively short career. Second, the ability of a real team to test the effects of lineup changes is miniscule compared to our ability to simulate *all* of a team's potential lineups. It would take over 2200 years for a baseball team to test all 3.6×10^5 possible lineups for a nine-man roster. Even changing the lineup daily for an entire season would test only 0.04% of those lineups.

Such advantages translate easily to other sports, especially with the advent of player-tracking

technology that can quantify player ability and complementarity with teammates (Tamir & Oz 2008). Specifically, this method could assist alternate lineup optimization strategies researched in other sports such as cricket (Bhattacharjee & Saikia 2014), basketball (Bhandari et al. 1997), and association football (Boon & Sierksma 2003) to produce data on individual player worth. For example, unlike baseball, a basketball player's worth greatly depends on his short-term stamina (Hoffman et al. 1996). If Player A can play 40 competitive minutes per game, he is worth more than Player B of equal talent who tires after 30 minutes. A team's general manager might find it more cost-effective to supplement Player B with Player C than to sign Player A at a high salary if the team's performance is unaffected. He could find and trade for the right Player C *for his specific team* through game simulations, even if he currently plays on another team.

Beyond sports, simulation-based optimization has been used for decades for all sorts of intractable problems (Deng 2007). The main contribution herein is to bring to light the possibility of using such algorithms to determine the worth of the individual parts. For example, simulating the optimal path of a widget through an assembly line could reveal hidden information about the effectiveness of each tool in that line, leading to better decision-making and cost-effectiveness.

Conclusions

We have developed a new method to determine an individual baseball player's worth to his team (assuming worth is measured in wins). This method was created with the intention of an intuitive approach, summed up as follows:

1. Simulate a run distribution for a team using its players' raw statistics.
2. Fit this run distribution with a Weibull probability distribution.
3. Compare this run distribution to the runs allowed by the team to develop a baseline winning percentage.
4. Remove or replace players in the lineup, simulate, and measure the team's change in offensive performance.
5. Use the new data to develop a new winning percentage.
6. The difference between this winning percentage and the baseline winning percentage determines the number of wins that a player can be expected to contribute to his team.

We have shown that this method produces results similar to the popular statistics wRAA and wSB, while, at the same time, is flexible enough to be fine-tuned for each individual player on each team. Although the details are beyond the scope of this paper, this methodology could be used to improve the efficiency of any process that can be simulated as the networking of quantifiable individual parts.

References

Bhandari, I., Colet, E., Parker, J., Pines, Z., Pratap, R., & Ramanujam, K. (1997). Advanced Scout: Data Mining and Knowledge Discovery in NBA Data. Data Mining and Knowledge Discovery *1*(1), 121–125.

Bhattacharjee, D. & Saikia, H. (2014). On Performance Measurement of Cricketers and Selecting an Optimum Balanced Team. *International Journal of Performance Analysis in Sport*, *14*(1), 262–275.

Boon, B. H. & Sierksma, G. (2003). Team formation: Matching quality supply and quality demand. *European Journal of Operational Research, 148*(2), 277–292.

Bukiet, B., Harold, E. R. & Palacios, J. L. (1997). A Markov Chain Approach to Baseball. *Operations Research, 45*(1), 14–23.

Cameron, D. WAR: Imperfect but Useful Even in Small Samples. http://www.fangraphs.com/blogs/war-imperfect-but-useful-even-in-small-samples/ (April 29, 2013).

Deng, G. (2007). *Simulation-Based Optimization* (Dissertation). University of Wisconsin-Madison, Madison, Wisconsin.

Catania, J. Making the Case for WAR as Baseball's Most Perfect Statistic. http://bleacherreport.com/articles/1642919-making-the-case-for-war-as-baseballs-most-perfect-statistic (May 18, 2013).

Eder, S. Era of Modern Baseball Stats Brings WAR to Booth. http://www.nytimes.com/2013/04/02/sports/baseball/baseball-broadcasts-introduce-advanced-statistics-but-with-caution.html (April 1, 2013).

Freeze, R. A. (1973). An Analysis of Baseball Batting Order by Monte Carlo Simulation. *Operations Research, 22*(4), 728–735.

Hirotsu, N. (2011). Reconsideration of the Best Batting Order in Baseball: Is the Order to Maximize the Expected Number of Runs Really the Best? *Journal of Quantitative Analysis in Sports, 7*(2), 13.

Hoffman, J. R., Tenenbaum, G.; Maresh, C. M., & Kraemer, W. J. (1996). Relationship between Athletic Performance Tests and Playing Time in Elite College Basketball Players. *The Journal of Strength & Conditioning Research, 10*(2), 67–71.

James, B. (1983) *The Bill James Baseball Abstract*. New York: Ballantine Books.

Jensen, S. T., Shirley, K. E., & Wyner, A. J. (2009). Bayesball: A Bayesian Hierarchical Model for Evaluating Fielding in Major League Baseball. *The Annals of Applied Statistics, 3*(2), 491–520.

Miller, S. J. (2007). A Derivation of the Pythagorean Won-Loss Formula in Baseball. *Chance, 20*(1), 40–48.

Ormiston, R. (2012) Attendance Effects of Star Pitchers in Major League Baseball. *Journal of Sports Economics*. doi: 10.1177/1527002512461155.

Petti, B. What Hitting Metrics Correlate Year-to-Year? http://www.beyondtheboxscore.com/2011/9/1/2393318/what-hitting-metrics-are-consistent-year-to-year (September 1, 2011).

Robinson, S. J. (2013) Optimizing Youth Baseball Batting Orders. *International Journal of Computer Science in Sport, 12*(1), 18–32.

Tamir, M. & Oz, G. Real-Time Objects Tracking and Motion Capture in Sports Events. US Patent 20080192116 A1 (August 14, 2008).

Tango, T. M., Lichtman, M. G., & Dolphin, A. E. (2007). *The Book: Playing the Percentages in Baseball*. Washington, D.C.: Potomac Books, Inc.

Factors Influencing the Accuracy of Predictions of the 2014 FIFA World Cup

Peter O'Donoghue

Cardiff School of Sport, Cardiff Metropolitan University, Cyncoed Campus, Cardiff, Wales, CF23 6XD, UK.

Abstract

The purpose of this paper was to compare the accuracy of different simulation models of the 2014 FIFA World Cup. There were 12 (2 x 3 x 2) models altogether (2 data sets of previous matches, 3 sets of variables and models where the data either satisfied the assumptions of linear regression or not). One set of previous data consisted of 440 matches from all international tournaments played since the 2006 FIFA World Cup. The second data set was a subset of 96 of these 440 matches that were from inter-continental tournaments. There were three predictor variables used; difference in FIFA World ranking points (PD) between the 2 teams contesting a match, difference in distance travelled to tournaments (DD) and difference in recovery days from previous matches (RD). The goal difference (GD) between the two teams contesting a match was modelled using (a) PD only, (b) PD and DD and (c) PD, DD and RD. Six models were produced without any changes in the data meaning they violated the modelling assumptions. The other 6 models were constructed using data that had been transformed and outliers being removed in order to satisfy the modelling assumptions. The standard deviation of residual values was used to add random variation about expected results within the simulator. The models that satisfied the assumptions of linear regression were not as accurate at predicting the outcomes of the 2014 World Cup matches as models where the assumptions were violated. Models based on 2 or 3 variables were more accurate than models based on PD alone. Finally, models based on the complete set of previous tournament data were more accurate than those based on the subset of data from inter-continental tournaments.

KEYWORDS**:** LINEAR REGRESSION, HOME ADVANTAGE, SOCCER.

Introduction

The ability to predict the outcomes of international soccer matches is limited. In the 2010 FIFA World Cup in South Africa, 50.8% of matches were won by the higher ranked team according to FIFA's World rankings, 27.0% were draws and the remaining 22.2% were upsets (this does not include extra time and penalty shoot outs). The author has carried out a series of studies since 2002 to evaluate the accuracy of predictive models of international soccer matches and rugby union matches. The main lesson from these studies is that statistical models, such as linear regression, logistic regression and discriminant function analysis, only produce single expected results for matches. When one considers the residual values in linear regression analyses (differences between actual observed results and expected results), there is considerable variability about the expected results. Simulation packages can use the known distributions of residual values to include random variation about expected results within

predictions. Simulation packages can simulate thousands of tournaments, accumulating outcome statistics for matches and progression statistics for the different teams within tournaments. The advantage of this is that each team's chance of winning, drawing and losing a match can be represented as well as their overall chances of winning the tournament. The main problem with models that fail to represent random chance is they predict a single result for each match based on the expected result. For example, one team may be expected to score 0.6 more goals than the opposing team. This would then be counted as a win because 0.6 rounds up to a goal difference of +1. However, when the distribution of residual values is considered with an expected goal difference of 0.6, we might have a 52% chance of a win, a 21% chance of a draw and a 27% chance of an upset (assuming a standard deviation of 1.8 for residual values). When the expected result of 0.6 goals is counted as a win, the probability of drawing or losing is considered to be 0 because the decision has failed to consider random variability in the goal difference variable. The author's previous studies of international soccer and rugby tournaments have revealed that simulation models are more successful at predicting the outcomes of international tournaments. The probability of a team progressing to the final is a combination of conditional probabilities of winning matches at the preceding stages (O'Donoghue et al., 2004). This is much better represented by simulation models than prediction methods that predict single results at each stage.

Since 2003, a sub-theme of the author's series of studies has been the comparison of the accuracy of models where the data used satisfy the assumptions of the modelling technique with the accuracy of models where the assumptions are violated. Violating the assumptions typically uses untransformed variables when producing a model based on previous match data as well as when making predictions of future matches. In order to satisfy the assumptions of modelling techniques, it is usually necessary to transform variables using square root or natural logarithms (Nevill, 2000). It is also necessary to remove matches where the values for any variables are statistical outliers. The previous investigations have produced conflicting conclusions about the accuracy of methods satisfying and violating the assumptions of the modelling techniques used. The studies of the Euro 2008 soccer tournament (O'Donoghue, 2009) and the 2011 Rugby World Cup (O'Donoghue, 2012) showed that the models based on data satisfying the assumptions were more accurate than those where the assumptions were violated. However, predictions where the data used violated the assumptions of the modelling technique were more accurate in the studies of the 2003 Rugby World Cup (O'Donoghue and Williams, 2004), the Euro 2004 soccer tournament (O'Donoghue, 2005), the 2006 (O'Donoghue, 2006) and 2010 FIFA World Cups (O'Donoghue, 2010).

Given that simulation models have been more accurate than neural network models and single predictions made using statistical analyses, the scope of the current investigation is restricted to simulation models. The particular simulation techniques of interest to the current investigation are those with underlying linear regression based models of expected match outcomes. There are some criticisms that can be made about previous research into the accuracy of simulation models. Some factors that may influence the success of such models have not been investigated. Such factors include the number of predictor variables used and differing data sets used to construct the underlying regression models. Previous attempts to predict the outcomes of international soccer matches have used FIFA World ranking points, distance between competing countries and the host country(ies) of tournaments and recovery days from previous matches. Typically, only ranking points is a significant predictor of the goal difference between two teams playing a match. However, given the previous successes of methods using all three variables, it is worth comparing models where one, two or all three of these variables are included.

A further criticism of previous research is that performance in inter-continental tournaments (such as the FIFA World Cups) have been predicted using models created by analysing data from a combination of inter-continental tournaments as well as continental tournaments (such as Copa America, the African Cup of Nations and the Euro 2004 and 2008 tournaments). One of the issues in using continental tournaments is that there are more of these than inter-continental tournaments. This could lead to some matches in inter-continental tournaments being considered as statistical outliers with respect to distance travelled by teams involved. This is because teams travel further to inter-continental tournaments than to continental tournaments. Satisfying the assumptions of linear regression may involve removal of such outliers which are actually the type of match we are trying to predict in the current study. Therefore, another purpose of the current investigation is to compare models constructed by analysing matches from previous inter-continental tournaments only with models constructed by analysing matches from all previous international tournaments.

As has already been mentioned, there is still conflicting evidence as to whether satisfying the assumptions of linear regression is effective. Therefore, this study will compare models where the necessary variable transformations have been made so that data satisfy the assumptions with models where such transformations have not been made. The assumptions of linear regression are:

- There should be 20 cases for each predictor variable (Ntoumanis, 2001, p120).
- Predictor variables must not be highly correlated with each other (Tabachnick and Fidell, 1996, p131-139).
- Predictor variables and the dependent variable should be free of outliers, especially extreme values (Fallowfield et al., 2005, p180).
- Residual values must be normally distributed (Newell et al., 2010, p140).
- Residual values must be independent from the predicted values for the dependent variable (Newell et al., 2010, p140).
- Residual values must be independent of the order of case occurrence (Newell et al., 2010, p140).

The current investigation compares the accuracy of 12 predictive models (2 data sets x 3 selections of variables x 2 types of model depending on whether the data have satisfied the assumptions or not).

Methods

Data Sources

The 12 models were created using data from previous international tournaments played since 2006 when FIFA changed their World Ranking method. All matches were considered with respect to the higher ranked team of the two teams contesting the match. The dependent variable was the difference between the higher and lower ranked teams' goals scored in the match excluding extra time and penalty shoot outs. There were three predictor variables used. These were:

- PD: Difference between the FIFA World ranking points between the two teams involved in a match.
- DD: Difference in distance traveled to the tournament by the two teams. For each team involved in a match, the distance travelled was estimated by the giant circle distance

between the capital city of the country and the capital city of the host nation. This was obtained from an internet based distance calculator (Indonesia, 2006). The higher ranked team's distance is denoted D_H while the lower ranked team's distance is dented as D_L.

- RD: The difference in the recovery days from the previous matches within the tournament played by the two teams.

There were a total of 440 matches included in the data set from all international tournaments played from the 2007 Copa de Oro de la CONCACAF (Confederation of North, Central American and Caribbean Association Football) Cup to the 2014 African Nations Cup. A subset of 96 of these matches from inter-continental tournaments was used to produce 6 of the predictive models while the other 6 models were created using the full data set. The inter-continental tournaments were the 2010 FIFA World Cup and the 2009 and 2013 Confederations Cups.

An exploratory regression analysis revealed that PD was the only significant predictor of goal difference ($p < 0.001$) with DD ($p = 0.121$) and RD ($p = 0.723$) being excluded where a stepwise variable entry method used a criteria of $p < 0.05$ for inclusion of variables. It was, therefore, decided not to produce models using all 7 combinations of one, two or all three of these variables. Instead, the models were based on three collections of variables:

- PD only
- PD and DD
- PD, DD and RD

Regression Analysis

Each of the 12 simulation models used an underlying linear regression model of goal difference, GD, based on one, two or all three predictor variables. The standard deviation of the residual values was used within the simulator to ensure random variation about the expected results reflected previous soccer matches. Six of the models were created without removing outliers or transforming variables; in each case the data violated one or more assumptions of multiple linear regression. Three of these models used the complete set of 440 previous cases while the other three models used the subset of 96 inter-continental tournament matches. SPSS Version 20.0 (SPSS: An IBM Company, Amarouk, NY) was used to perform the regression analyses, noting the regression coefficients and the standard deviation of the residual values which were saved.

The complete set of 440 matches and the sub-set of 96 inter-continental tournament matches were considered separately when making transformations and removing outliers to ensure the data satisfied the assumptions of multiple linear regression. There were sufficient previous matches for the number of predictor variables used and there were no high correlations between the predictor variables. The residual values were found to be independent of predicted values and were found to be independent of chronological match order. However, there were outliers in the three variables and residual values were not normally distributed. It was not possible to remove outliers for the RD value because the inter-quartile range was 0 (more than three quarters of the values were 0) meaning that all values less than or greater than 0 were outliers. Even when all matches with RD values of less than -1 or greater than +1 were removed, the inter-quartile range for RD was still 0. It was, therefore, decided not to exclude any matches on the basis of RD value. When the complete data set of 440 matches was considered, there were more outliers for DD than PD. Therefore, outliers were initially removed for DD followed by the removal of any remaining outliers in PD before rechecking

the distribution of DD. As outliers in DD were removed, the inter-quartile range of DD decreased introducing further outliers that were not outliers before. The process of removing outliers in DD went through 5 iterations with the matches removed being from Confederations Cups, the 2010 World Cup and matches of the AFC (Asian Football Confederation) Cup where teams had to travel further to tournaments than teams travelled in other tournaments. Once all of the outliers in DD were removed, two iterations of outlier removal were required to remove outliers in PD. These matches included Confederations Cup matches where highly ranked teams played low ranked teams, such as Tahiti, Iraq and New Zealand as well as a World Cup match between Brazil and North Korea. There were also outliers in the dependent GD variable requiring matches with GD values less than -3 or greater than +4 to be removed. This reduced the data set to 398 matches. Exploratory regression analysis of these cases revealed that the residuals were not normally distributed and, therefore, transformation possibilities had to be considered. As has already been mentioned, no transformation could have been made to RD other than removing it altogether. Therefore, it remained unaltered despite technically containing outliers. The following two transformations achieved normalisation of residual values in the linear regression models where they were used:

- PD was replaced by $PD^{0.5}$
- DD was replaced by $(D_H^{(2/3)} – D_L^{(2/3)})$

The subset of 96 matches from inter-continental tournaments was analysed to produce three further models that satisfied the assumptions of linear regression. This data set contained more outliers for PD than DD and, therefore, the outliers for PD were removed first. This required a single pass of the data removing 6 matches between the highest ranked teams (Brazil or Spain) and low ranked teams (New Zealand, Tahiti, North Korea, Iraq and South Africa twice). Two remaining outliers in DD were removed (Mexico v South Africa and Brazil v Japan). It was also necessary to remove outliers in the dependent GD variable meaning that the only matches included in this data set had GD values of between -2 and +3 inclusive. The 80 matches that remained, once the outliers in PD, DD and GD were removed, satisfied the remaining assumptions of linear regression without any transformation of the variables being required.

The Models

Table 1 summarises the 12 underlying models that were used by the simulation package. The regression coefficients show that having a greater number of ranking points than the opponent increased GD, while travelling further than the opponent decreased GD. The effect of recovery days was different in models created using previous data from all tournaments and models created using data from just inter-continental tournaments. When previous inter-continental tournaments were analysed, having one more recovery day from the previous match than the opponent was found to increase GD by 0.295 or 0.363 on average. When all previous tournaments were analysed, having one more recovery day from the previous match than the opponent was found to decrease GD. However, the additional recovery day here only resulted in an additional goal for the opponents in one match in 20 on average.

Table 1. Underlying regression models for goal difference (GD) in terms of ranking points difference (PD), distance travelled distance (DD) and recovery days difference (RD) used by the simulation package.

Data Source and Variables	Violating Assumptions	Satisfying Assumptions
<u>All tournaments</u>		
Rank	-0.0299 + 0.00249 PD Residual SD: 1.721	$-0.356 + 0.0620\ PD^{0.5}$ Residual SD: 1.460
Rank and Distance	-0.0353 + 0.00253 PD – 0.0000354 DD Residual SD: 1.716	$-0.357 + 0.0621\ PD^{0.5} - 0.000386(D_H^{2/3} - D_L^{2/3})$ Residual SD: 1.459
Rank, Distance and Recovery Days	-0.0377 + 0.00254 PD – 0.0000363 DD – 0.0548 RD Residual SD: 1.715	$-0.361 + 0.0622\ PD^{0.5} - 0.000409(D_H^{2/3} - D_L^{2/3}) - 0.0494$ RD Residual SD: 1.459
<u>Intercontinental tournaments only</u>		
Rank	-0.0460 + 0.00228 PD Residual SD: 1.967	-0.0622 + 0.00122 PD Residual SD: 1.320
Rank and Distance	-0.190 + 0.00249 PD – 0.000106 DD Residual SD: 1.883	-0.554 + 0.00123 PD – 0.00000929 DD Residual SD: 1.320
Rank, Distance and Recovery Days	-0.163 + 0.00241 PD – 0.0000985 DD + 0.363 RD Residual SD: 1.875	-0.0647 + 0.00120 PD – 0.000000166 DD + 0.295 RD Residual SD: 1.312

Simulation

A simulation package was developed in Matlab version 7.0.1 (Mathworks Inc., Natick, MA) to simulate the 2014 FIFA World Cup 20,000 times, accumulating progression statistics for each team. The simulator was run 12 times using each of the 12 underlying regression models and the standard deviation of the residuals for the dependent GD variable. The simulator was initialised with information about the teams' FIFA World ranking points (www.fifa.org, accessed 6th June 2014), distances from each country's capital city to Brasilia and details of the match schedule for the tournament to allow differences in recovery days to be used. The simulation of a match worked by determining the expected value for GD using the given regression model. A random number between 0 and 1 was then generated and used to look up a normal distribution curve with a mean value equal to the expected GD and a standard deviation being the standard deviation of the residuals in the predicted GD values from in the data used to create the model. The random number dictated the area of the normal distribution curve to the left of the simulated GD value. In pool matches, simulated GD values greater than 0.5 were rounded up to indicate a win for the higher ranked team, values less than -0.5 were rounded down to indicate an upset, with values of between -0.5 and 0.5 being counted as draws. In knock out matches, one team has to be eliminated so GD values of greater than or equal to 0 were used to represent a win for the higher ranked team in the match and values of less than 0 were counted as upsets.

The progression statistics accumulated for each team included the percentage of simulated World Cups where they finished first or second in their pool, won their second round match, quarter-final, semi-final, third place play-off and the final.

Evaluation Scheme

The 2014 FIFA World Cup consisted of 48 pool matches and 16 knock out matches. The evaluation method awards a maximum of 1 mark for each match. The fraction of a mark awarded depends on the proportion of simulated World Cups where a given model predicted the correct result. For example, consider the opening match between Brazil and Croatia which the first of the 12 models predicted to be a win for Brazil 58.1% of simulated tournaments, a win for Croatia in 21.9% of simulated tournaments and a draw in 20.1% of simulated tournaments (see Table 2). Brazil won this match and so this predictive model was awarded 0.581 points for the match. Given the 52% of wins, 21% of draws and 27% of losses that occurred in the 2010 FIFA World Cup, one would expect 18.6 results out of 48 to be predicted correctly by chance (48 x (52x52/100 + 21x21/100 + 27x27/100) / 100).

For the knock out stages, a mark was allocated for each of the 8 quarter-final places, the 4 semi-final places, the 2 final places, the third placed team and the tournament winner. No additional marks were awarded for predicting teams to be in the 16 second round places because this would double count performances in the pool matches. The first predictive model had Brazil reaching the quarter-finals in 46.5% of simulated tournaments, the semi-finals in 27.5% of predicted tournaments, the final in 16.1% of predicted tournaments and winning 8.1% of the predicted tournaments. Therefore, because Brazil came fourth in the actual 2014 FIFA World Cup, this predictive model was awarded 0.465 + 0.275 = 0.740 marks for predicting Brazil's performances in the knock out stages. Altogether, the maximum possible mark for a prediction is 64 but realistically this is unachievable because it would require 100% of simulated tournaments to predict the actual results of all 64 matches of the World Cup.

The Predictions

Tables 2 and 3 show the percentage of simulated tournaments where each model predicted wins, draws and losses in each pool match. Table 2 shows the predictions made by models created using all tournament data and Table 3 shows the predictions made by models created using only inter-continental tournament data. Figure 1 shows the modal prediction for the knock out stages of the World Cup. It should be noted, however, that no model gave any single team more than a 25.7% chance of winning the World Cup. The predictions as well as the evaluation method were sent to the General Editor of the International Journal of Computer Science in Sport before the World Cup commenced. The results and discussion of the paper were completed after the 2014 FIFA World Cup had completed.

Table 2. Pool match predictions for models based on all tournament data.

Pool	Team 1	Team 2	1 Variable						2 Variables						3 Variables					
			Violating Assumptions			Satisfying Assumptions			Violating Assumptions			Satisfying Assumptions			Violating Assumptions			Satisfying Assumptions		
			W	D	L	W	D	L	W	D	L	W	D	L	W	D	L	W	D	L
A	Brazil	Croatia	58.1	20.1	21.9	58.4	22.9	18.7	65.3	18.1	16.7	62.9	21.1	16.0	64.7	18.4	16.9	63.7	21.0	15.3
A	Mexico	Cameroon	56.7	20.6	22.8	56.8	25.0	18.2	56.9	20.1	23	57.2	24.1	18.7	57.3	20.0	22.6	56.7	24.6	18.7
A	Brazil	Mexico	59.0	20.1	20.9	59.0	23.3	17.7	64.2	19.2	16.6	63.2	20.9	15.9	63.0	19.2	17.9	60.9	23.6	15.4
A	Croatia	Cameroon	58.6	20.3	21.2	58.0	23.0	19.0	56.2	21.4	22.4	57.1	24.1	18.8	53.7	22.2	24.1	56.7	23.9	19.3
A	Brazil	Cameroon	75.5	15.0	9.5	70.3	19.0	10.7	80.5	11.8	7.7	73.6	17.0	9.4	79.7	12.9	7.4	72.7	17.7	9.6
A	Croatia	Mexico	38.5	23.3	38.2	33.8	27.8	38.4	36.8	23.4	39.8	33.7	27.5	38.9	38.1	22.6	39.3	35.3	27.1	37.6
B	Spain	Holland	67.4	17.2	15.5	64.4	22.4	13.2	68.5	17.2	14.3	64.9	21.1	13.9	67.6	17.2	15.2	65.3	21.0	13.6
B	Chile	Australia	66.0	18.4	15.6	65.0	20.2	14.8	75.3	14.0	10.7	68.5	19.0	12.5	75.2	14.5	10.3	68.2	20.6	11.2
B	Spain	Chile	63.8	19.2	17.0	63.3	21.1	15.6	61.2	19.8	1.09	61.4	22.0	16.6	61.3	18.5	20.2	61	22.6	16.4
B	Holland	Australia	64.2	19.2	16.5	63.3	21.1	15.6	68.1	17.3	14.6	64.2	21.1	14.6	68.4	17.2	14.4	64.8	20.7	14.6
B	Spain	Australia	86.5	9.2	4.4	77.4	15.2	7.4	89.5	8.2	2.4	79.5	14.0	6.6	89.5	8.0	2.5	79.0	14.4	6.6
B	Holland	Chile	36.2	24.1	39.7	34.9	27.4	37.7	31.8	23.5	44.7	33.5	27.1	39.4	31.2	23.3	45.5	32.5	27.2	40.4
C	Columbia	Greece	41.9	22.7	35.4	40.9	27.1	32.0	45.7	23.0	31.3	42.7	27.0	30.3	46.1	22.6	31.4	42.7	27.1	30.2
C	Ivory C	Japan	48.3	23.4	28.3	49.0	26.3	24.8	59.4	20.2	20.3	54.1	23.9	22.0	59.0	20.1	20.9	53.8	24.5	21.7
C	Columbia	Ivory C	56.9	20.6	22.5	58.6	22.7	18.7	58.1	21.5	20.4	58.2	23.2	18.6	57.5	22.0	20.5	58.4	23.2	18.4
C	Greece	Japan	63.2	19.1	17.6	62.3	22.3	15.5	69.1	17.4	13.5	64.2	20.9	14.9	69.4	17.0	13.6	63.6	22.6	13.7
C	Columbia	Japan	67.3	18.3	14.5	64.9	21.6	13.5	78.1	13.8	8.1	68.6	19.6	11.8	77.0	14.1	8.8	68.9	19.4	11.7
C	Greece	Ivory C	53.3	21.1	25.7	54.0	23.7	22.3	49.4	22.2	28.4	52.3	24.8	23.0	48.8	22.7	28.6	52.1	25.5	22.5
D	Uruguay	C Rica	59.6	20.2	20.1	60.3	23.1	16.6	62.9	19.2	17.9	61.5	21.7	16.8	62.7	19.6	17.8	61.6	22.0	16.5
D	England	Italy	37.6	23.6	38.8	38.9	27.0	34.0	37.2	23.4	39.5	38.8	27.8	33.4	37.6	24.4	38.0	39.1	27.3	33.5
D	Uruguay	England	40.9	23.5	35.7	39.1	27.3	33.6	45.3	23.7	31.0	41.8	27.3	30.9	45.9	23.4	30.7	42.1	26.9	31.0
D	C Rica	Italy	21.6	20.0	58.4	18.7	23.3	58.0	23.2	21.7	55.1	18.9	24.4	56.7	24.3	21.4	54.4	19.5	23.3	57.1
D	Uruguay	Italy	39.6	23.3	37.1	36.9	27.3	35.8	45.1	22.9	32.0	39.8	27.4	32.9	44.3	23.1	32.6	39.3	27.2	33.5
D	C Rica	England	22.4	20.1	57.5	18.2	22.9	58.9	24.3	21.9	53.8	19.6	24.4	56.0	25.9	22.1	52.0	20.5	24.7	54.8
E	Switzerland	Ecuador	58.7	20.5	20.7	59.1	23.3	17.6	54.7	21.2	24.2	57.0	24.2	18.8	54.0	22.2	23.8	56.7	23.5	19.8
E	France	Honduras	47.6	23.7	28.7	48.4	27.0	24.6	44.7	23.0	32.3	47.9	26.5	25.6	44.9	23.2	31.9	47.6	26.5	25.9
E	Switzerland	France	52.9	21.2	26.0	53.1	25.4	21.5	52.3	21.1	26.6	54.3	24.7	21.1	52.3	21.1	26.6	53.0	24.0	23.0
E	Ecuador	Honduras	40.9	22.9	36.2	40.1	27.0	32.9	41.8	23.9	34.3	40.4	27.1	32.5	41.2	24.8	34.0	39.2	27.7	33.1
E	Switzerland	Honduras	61.8	19.3	18.9	61.1	22.3	16.6	59.5	20.2	20.4	59.8	23.6	16.6	59.1	20.5	20.4	59.4	23.6	17.0

E	Ecuador	France	32.7	23.5	43.8	27.9	27.3	44.8	35.4	23.5	41.0	29.6	27.1	43.3	35.5	23.6	41.0	29.2	27.1	43.7
F	Argentina	Bosnia	55.3	20.7	24.0	56.3	23.9	19.8	62.0	19.9	18.1	58.7	23.2	18.1	61.2	20.0	18.8	58.7	24.1	17.2
F	Iran	Nigeria	38.1	22.6	39.3	28.2	27.4	44.4	32.8	22.9	44.3	26.5	27.3	46.2	32.8	23.4	43.8	26.9	27.5	45.6
F	Argentina	Iran	68.7	16.8	14.5	66.1	20.1	13.7	74.8	15.6	9.6	69.5	19.2	11.4	74.0	15.1	10.9	68.0	19.3	12.7
F	Bosnia	Nigeria	51.1	21.6	27.3	53.6	24.1	22.3	47.8	23.4	28.8	52.0	24.2	23.9	46.6	23.4	30.0	50.5	25.2	24.3
F	Argentina	Nigeria	68.8	17.1	14.1	65.7	21.0	13.3	71.6	16.1	12.3	67.4	19.8	12.8	72.3	15.2	12.5	67.0	19.9	13.1
F	Bosnia	Iran	51.9	22.1	26.0	52.7	24.5	22.8	53.7	21.3	25.0	53.9	24.5	21.6	54.5	20.9	24.6	53.5	24.6	21.9
G	Germany	Portugal	44.2	22.9	32.9	43.9	27.0	29.1	42.2	22.9	34.9	44.3	26.2	29.4	41.9	23.3	34.9	42.8	27.3	29.9
G	Ghana	USA	21.4	20.9	57.7	19.0	22.9	58.1	22.9	21.3	55.8	18.6	24.8	56.7	22.3	20.8	56.9	18.7	24.3	57.0
G	Germany	Ghana	71.3	16.4	12.4	68.5	19.2	12.3	68.6	17.8	13.6	66.2	21.3	12.5	69.7	16.8	13.5	66.3	20.4	13.3
G	Portugal	USA	45.7	23.3	31.0	48.1	26.2	25.7	46.0	22.7	31.2	47.1	27.2	25.7	46.3	22.8	30.9	47.1	26.6	26.3
G	Germany	USA	53.3	22.3	24.4	53.8	25.2	21.0	51.0	21.9	27.1	53.6	24.9	21.5	49.4	22.8	27.7	52.6	24.9	22.5
G	Portugal	Ghana	66.3	18.2	15.5	64.6	20.9	14.5	64.9	18.4	16.7	63.2	21.8	15.0	66.0	18.1	15.9	64.5	22.0	13.5
H	Belgium	Algeria	49.7	22.6	27.7	51.6	25.2	23.2	48.3	23.7	28.0	51.1	25.0	23.9	48.6	23.4	28.1	50.9	25.1	23.9
H	Russia	S Korea	58.3	19.8	21.9	57.9	23.4	18.8	62.8	19.0	18.2	60.3	23.2	16.5	63.3	18.9	17.8	60.1	22.9	16.9
H	Belgium	Russia	48.0	23.5	28.5	49.0	26.8	24.2	50.1	22.0	27.9	50.1	25.2	24.7	50.5	21.8	27.7	49.8	25.2	25.0
H	Algeria	S Korea	55.8	21.1	23.1	56.4	23.7	19.9	63.4	18.7	17.9	58.5	23.8	17.7	63.3	19.3	17.3	59.1	23.4	17.5
H	Belgium	S Korea	68.0	16.9	15.1	65.5	20.6	13.9	74.2	15.0	10.8	68.0	19.3	12.8	73.8	15.4	10.8	68.4	18.9	12.7
H	Algeria	Russia	36.7	23.1	40.2	36.8	26.9	36.2	39.9	23.4	36.7	36.7	27.3	36.0	39.9	23.3	36.8	37.1	27.3	35.7

Table 3. Pool match predictions for models based on inter-continental tournament data.

Pool	Team 1	Team 2	1 Variable						2 Variables						3 Variables					
			Violating Assumptions			Satisfying Assumptions			Violating Assumptions			Satisfying Assumptions			Violating Assumptions			Satisfying Assumptions		
			W	D	L	W	D	L	W	D	L	W	D	L	W	D	L	W	D	L
A	Brazil	Croatia	54.9	19.2	25.9	49.0	28.2	22.8	73.3	14.1	12.6	52.0	27.2	20.8	72.0	15.5	12.5	49.3	28.6	22.2
A	Mexico	Cameroon	53.9	18.9	27.2	48.2	29.3	22.5	51.3	20.3	28.5	48.4	28.8	22.8	51.7	20.1	28.3	48.0	28.8	23.2
A	Brazil	Mexico	56.6	18.0	25.4	49.2	28.8	22.0	69.0	16.2	14.8	52.5	27.1	20.4	74.7	13.4	11.9	59.0	25.4	15.6
A	Croatia	Cameroon	55.0	19.2	25.8	48.7	28.6	22.8	48.0	20.2	31.8	48.2	28.9	22.9	56.6	19.0	24.3	58.6	25.7	15.7
A	Brazil	Cameroon	70.4	15.4	14.2	61.2	24.6	14.2	82.0	11.1	6.9	64.2	23.6	12.2	86.6	9.1	4.2	70.5	19.8	9.7
A	Croatia	Mexico	39.8	20.2	40.0	36.5	30.5	33.0	31.2	21.0	47.8	36.3	30.6	33.2	24.8	20.0	55.2	29.6	29.9	40.5
B	Spain	Holland	62.2	17.3	20.5	55.9	26.1	18.1	64.9	17.3	17.8	56.1	26.4	17.5	65.2	17.8	17.0	55.3	26.7	18.0
B	Chile	Australia	62.5	17.1	20.4	56.2	25.1	18.7	82.1	11.4	6.5	57.7	25.4	16.9	81.3	11.2	7.5	54.9	26.3	18.7
B	Spain	Chile	59.7	18.6	21.8	54.2	26.7	19.1	49.1	20.5	30.4	52.2	27.0	20.8	49.3	20.8	29.9	54.0	26.7	19.3
B	Holland	Australia	60.4	18.5	21.1	54.2	26.1	19.7	70.1	14.9	15.0	55.1	26.3	18.6	69.2	16.3	14.5	54.3	26.6	19.2
B	Spain	Australia	80.0	11.4	8.6	70.8	20.5	8.7	90.2	6.3	3.5	73.2	18.2	8.6	89.5	7.1	3.5	71.6	19.0	9.4
B	Holland	Chile	38.8	20.1	41.1	31.8	30.8	37.5	27.6	21.0	51.4	30.8	28.9	40.4	28.5	19.7	51.7	32.0	29.3	38.8
C	Columbia	Greece	42.0	21.2	36.8	38.8	30.3	30.9	52.8	20.1	27.1	41.7	29.7	28.6	52.1	20.2	27.7	38.6	30.8	30.6
C	Ivory C	Japan	46.7	20.8	32.5	42.8	30.3	26.9	72.1	15.4	12.4	47.2	28.2	24.6	71.8	14.8	13.4	43.0	30.4	26.6
C	Columbia	Ivory C	54.1	19.1	26.8	49.0	28.2	22.7	56.9	19.2	23.9	48.5	28.1	23.3	56.1	18.8	25.0	48.6	28.6	22.8
C	Greece	Japan	58.9	18.6	22.6	52.6	28.3	19.1	75.2	14.4	10.5	54.9	25.8	19.2	74.6	14.0	11.4	53.5	26.8	19.7
C	Columbia	Japan	62.7	17.2	20.2	55.4	27.1	17.5	86.6	8.8	4.5	58.8	25.3	15.9	85.2	9.2	5.6	56.2	26.0	17.8
C	Greece	Ivory C	51.2	19.5	29.3	46.6	29.3	24.2	38.6	22.4	39.0	44.8	29.5	25.7	39.4	20.9	39.7	46.2	29.1	24.7

D	Uruguay	C Rica	57.4	18.2	24.4	51.1	28.3	20.6	62.9	17.6	19.5	52.6	27.4	20.1	61.2	18.2	20.6	51.2	28.3	20.5
D	England	Italy	39.5	20.0	40.4	32.5	30.2	37.3	41.8	22.6	35.6	33.1	29.8	37.1	42.8	20.5	36.7	32.9	30.1	37.0
D	Uruguay	England	40.4	21.2	38.4	39.0	29.5	31.6	53.0	20.5	26.5	40.5	29.8	29.7	52.9	20.0	27.1	39.7	29.9	30.4
D	C Rica	Italy	25.6	19.2	55.2	22.7	28.0	49.4	33.7	21.8	44.5	23.5	28.2	48.3	34.3	21.2	44.5	22.6	28.2	49.2
D	Uruguay	Italy	40.7	20.6	38.7	37.9	30.4	31.7	52.7	20.6	26.7	40.6	28.8	30.6	60.3	17.9	21.9	46.5	29.9	23.7
D	C Rica	England	26.8	19.0	54.2	23.2	28.2	48.6	34.8	21.1	44.0	23.6	28.6	47.8	27.1	21.4	51.6	16.0	25.1	58.9
E	Switzerland	Ecuador	56.4	17.7	25.9	50.2	28.3	21.4	43.1	21.5	35.5	47.9	27.9	24.1	42.7	22.2	35.1	49.3	28.7	22.0
E	France	Honduras	47.1	20.4	32.6	43.3	29.7	27.0	37.7	21.3	41.0	41.6	30.5	27.9	38.1	22.1	39.9	43.0	30.0	27.0
E	Switzerland	France	49.6	19.2	31.2	44.7	29.7	25.6	46.9	21.3	31.8	45.1	29.0	26.0	48.0	20.0	32.0	45.1	29.0	25.9
E	Ecuador	Honduras	42.6	20.2	37.2	39.6	30.1	30.4	42.2	21.2	36.6	38.5	30.1	31.5	41.7	22.9	35.3	38.6	29.6	31.8
E	Switzerland	Honduras	58.2	18.7	23.1	52.1	27.0	21.0	50.5	20.1	29.5	51.2	28.0	20.8	49.8	21.7	28.6	51.9	27.4	20.7
E	Ecuador	France	34.7	20.8	44.5	28.6	30.1	41.3	46.8	21.7	31.5	30.6	29.7	39.8	47.0	21.4	31.6	28.9	30.3	40.8
F	Argentina	Bosnia	53.1	19.2	27.7	48.4	28.0	23.6	67.5	16.2	16.3	49.6	28.8	21.6	66.3	17.4	16.3	47.8	28.2	24.0
F	Iran	Nigeria	39.1	20.2	40.8	36.3	30.5	33.2	24.7	19.2	56.1	35.2	30.1	34.7	24.5	20.3	55.1	37.2	30.3	32.6
F	Argentina	Iran	65.0	15.9	19.1	57.0	25.4	17.6	81.3	11.4	7.3	59.1	25.6	15.3	85.4	9.2	5.5	65.1	22.2	12.7
F	Bosnia	Nigeria	50.3	19.5	30.2	44.8	29.3	25.9	39.9	20.9	39.2	44.4	29.0	26.6	49.0	20.7	30.3	54.3	26.6	19.1
F	Argentina	Nigeria	64.2	17.1	18.7	57.1	25.4	17.6	70.8	15.6	13.6	57.8	25.5	16.7	70.9	15.2	13.9	56.0	26.1	17.9
F	Bosnia	Iran	49.5	19.4	31.1	44.8	29.8	25.4	52.4	20.3	27.3	45.6	29.8	24.6	52.8	20.4	26.8	44.6	29.9	25.5
G	Germany	Portugal	43.2	20.5	36.3	40.9	29.2	30.0	35.8	22.1	42.1	40.2	29.0	30.7	36.9	21.2	42.0	40.4	30.8	28.8
G	Ghana	USA	27.1	19.4	53.6	22.9	28.2	48.9	28.9	19.6	51.5	23.0	29.5	47.5	28.5	20.5	51.1	22.4	28.0	49.7
G	Germany	Ghana	66.8	16.1	17.1	59.6	25.1	15.3	57.5	19.5	23.1	57.5	25.5	17.0	59.7	18.1	22.2	59.5	24.9	15.6
G	Portugal	USA	45.7	20.5	33.9	41.3	30.8	27.8	41.7	22.6	35.8	41.8	30.8	27.4	41.9	21.9	36.2	42.4	30.4	27.2
G	Germany	USA	50.9	20.5	28.7	45.7	29.8	24.4	42.4	20.8	36.7	46.4	28.1	25.5	51.1	21.2	27.7	56.2	26.3	17.5
G	Portugal	Ghana	61.6	17.9	20.5	55.4	26.1	18.5	58.0	19.6	22.4	53.4	27.5	19.1	50.5	21.2	28.3	44.8	29.7	25.6
H	Belgium	Algeria	49.1	20.2	30.7	44.1	29.9	26.0	44.6	21.9	33.5	43.9	29.1	27.0	44.1	22.2	33.7	44.2	29.9	25.9
H	Russia	S Korea	54.8	19.3	25.9	48.6	28.5	22.9	67.4	17.3	15.2	52.2	27.1	20.7	67.1	16.0	16.9	48.9	28.2	22.9
H	Belgium	Russia	47.1	20.3	32.6	42.9	30.6	26.6	49.6	20.6	29.7	43.7	30.0	26.3	49.3	20.7	30.0	43.3	29.8	26.9
H	Algeria	S Korea	54.5	18.9	26.6	48.2	28.3	23.5	72.6	15.3	12.1	50.9	27.7	21.4	71.3	14.8	13.9	48.4	28.1	23.6
H	Belgium	S Korea	63.3	17.0	19.7	56.2	26.1	17.6	80.0	11.6	8.4	59.5	24.9	15.6	77.8	12.2	9.9	56.3	25.9	17.8
H	Algeria	Russia	38.7	20.2	41.2	32.2	30.2	37.6	48.5	20.6	30.9	33.3	30.1	36.6	48.3	20.2	31.6	32.1	30.4	37.5

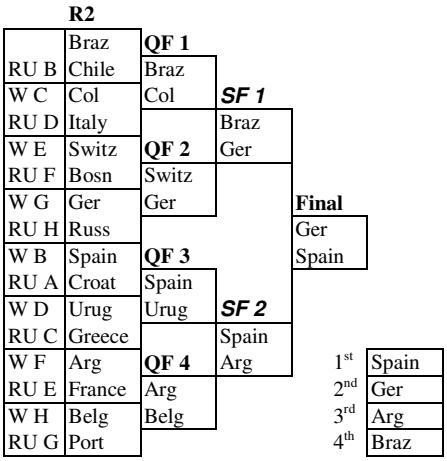

Figure 1(a) All tournaments-1 Variable-Violating Assumptions.

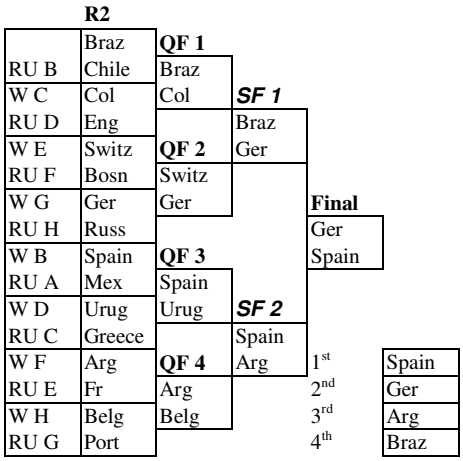

Figure 1(b) All tournaments-1 Variable-Satisfying Assumptions.

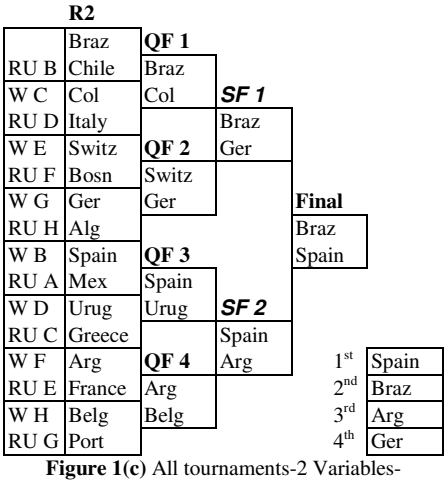

Figure 1(c) All tournaments-2 Variables-Violating Assumptions.

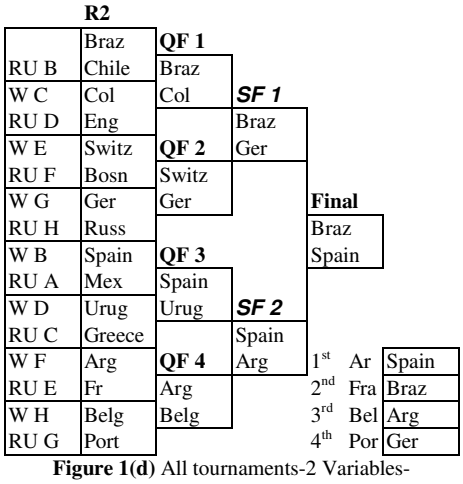

Figure 1(d) All tournaments-2 Variables-Satisfying Assumptions.

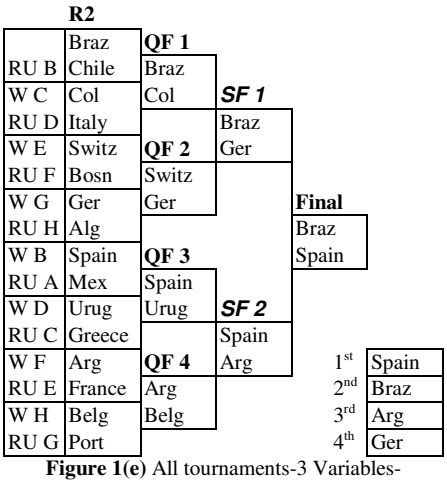

Figure 1(e) All tournaments-3 Variables-Violating Assumptions.

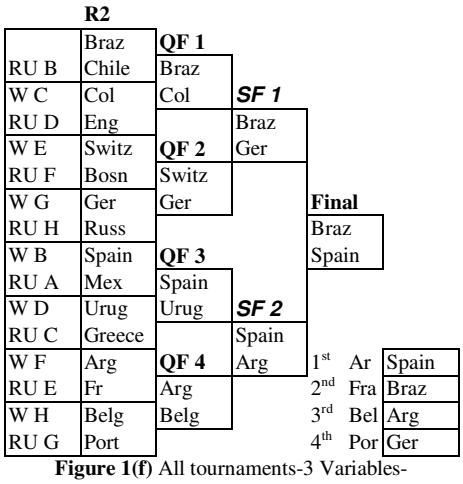

Figure 1(f) All tournaments-3 Variables-Satisfying Assumptions.

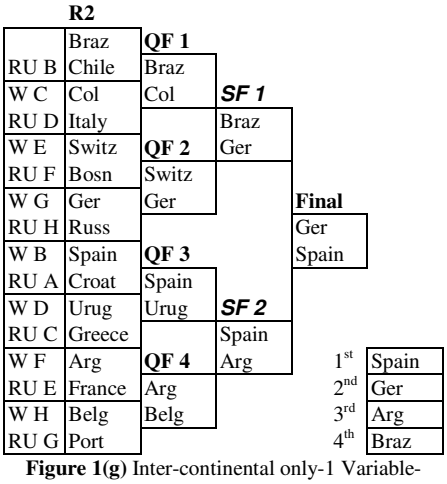

Figure 1(g) Inter-continental only-1 Variable-Violating Assumptions.

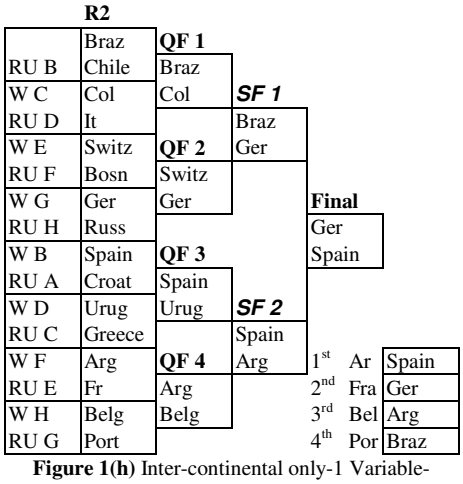

Figure 1(h) Inter-continental only-1 Variable-Satisfying Assumptions.

Figure 1. Modal predictions for knockout stages (continued on next page).

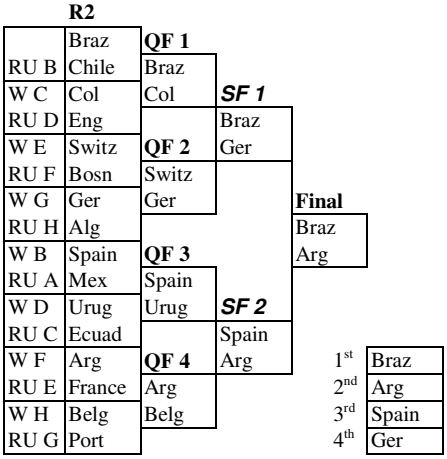

Figure 1(i) Inter-continental only-2 Variables-Violating Assumptions.

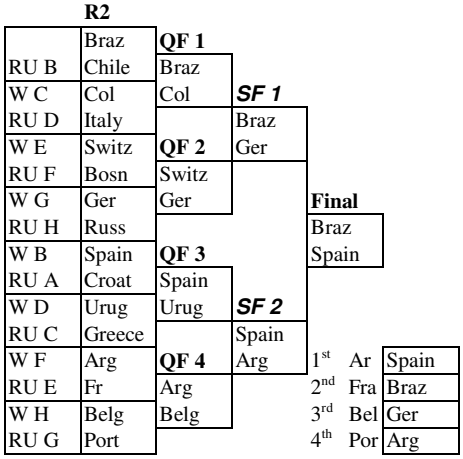

Figure 1(j) Inter-continental only-2 Variables-Satisfying Assumptions.

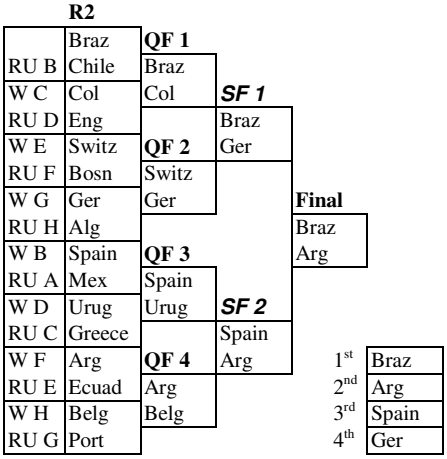

Figure 1(k) Inter-continental only-3 Variables-Violating Assumptions.

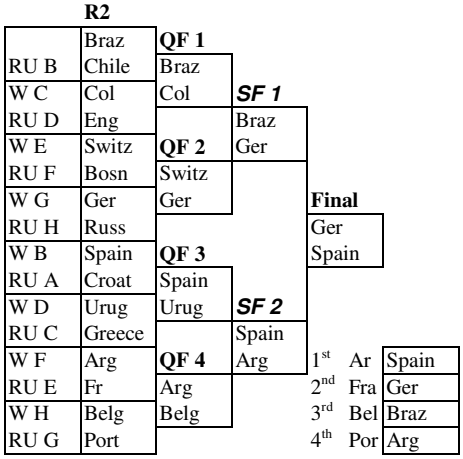

Figure 1(l) Inter-continental only-3 Variables-Satisfying Assumptions.

Figure 1. Modal predictions for knockout stages (continued from Previous page).

Results

Table 4 shows the percentage of simulated tournaments where pool matches were correctly predicted by each of the 12 models. Table 5 shows the percentage of simulated World Cups where teams were correctly predicted to reach different rounds of the knockout tournament. Table 6 shows the marks awarded using the evaluation scheme for pool matches, knockout matches and for the overall tournament. All 12 models predicted a greater number of pool matches correctly than the 18.6 out of 48 expected by random chance. The correctness of predictions of the knockout structure was more difficult because the matches within the knockout stages were unknown prior to the tournament commencing.

Table 4. Percentage of simulated tournaments where the correct result for pool matches was predicted.

Pool	Team 1	Team 2	Result	1 Variable				2 Variables				3 Variables			
				Violating Assumptions		Satisfying Assumptions		Violating Assumptions		Satisfying Assumptions		Violating Assumptions		Satisfying Assumptions	
				All	Inter-C	All	Inter-C	All	Inter-C	All	Inter-C	All	Inter-C	All	Inter-C
A	Brazil	Croatia	Win	58.1	54.9	58.4	49.0	65.3	73.3	62.9	52.0	64.7	72.0	63.7	49.3
A	Mexico	Cameroon	Win	56.7	53.9	56.8	48.2	56.9	51.3	57.2	48.4	57.3	51.7	56.7	48.0
A	Brazil	Mexico	Draw	20.1	18.0	23.3	28.8	19.2	16.2	20.9	27.1	19.2	13.4	23.6	25.4
A	Croatia	Cameroon	Win	58.6	55.0	58.0	48.7	56.2	48.0	57.1	48.2	53.7	56.6	56.7	58.6
A	Brazil	Cameroon	Win	75.5	70.4	70.3	61.2	80.5	82.0	73.6	64.2	79.7	86.6	72.7	70.5
A	Croatia	Mexico	Lose	38.2	40.0	38.4	33.0	39.8	47.8	38.9	33.2	39.3	55.2	37.6	40.5
B	Spain	Holland	Lose	15.5	20.5	13.2	18.1	14.3	17.8	13.9	17.5	15.2	17.0	13.6	18.0
B	Chile	Australia	Win	66.0	62.5	65.0	56.2	75.3	82.1	68.5	57.7	75.2	81.3	68.2	54.9
B	Spain	Chile	Lose	17.0	21.8	15.6	19.1	19.0	30.4	16.6	20.8	20.2	29.9	16.4	19.3
B	Holland	Australia	Win	64.2	60.4	63.3	54.2	68.1	70.1	64.2	55.1	68.4	69.2	64.8	54.3
B	Spain	Australia	Win	86.5	80.0	77.4	70.8	89.5	90.2	79.5	73.2	89.5	89.5	79.0	71.6
B	Holland	Chile	Win	36.2	38.8	34.9	31.8	31.8	27.6	33.5	30.8	31.2	28.5	32.5	32.0
C	Columbia	Greece	Win	41.9	42.0	40.9	38.8	45.7	52.8	42.7	41.7	46.1	52.1	42.7	38.6
C	Ivory C	Japan	Win	48.3	46.7	49.0	42.8	59.4	72.1	54.1	47.2	59.0	71.8	53.8	43.0
C	Columbia	Ivory C	Win	56.9	54.1	58.6	49.0	58.1	56.9	58.2	48.5	57.5	56.1	58.4	48.6
C	Greece	Japan	Draw	19.1	18.6	22.3	28.3	17.4	14.4	20.9	25.8	17.0	14.0	22.6	26.8
C	Columbia	Japan	Win	67.3	62.7	64.9	55.4	78.1	86.6	68.6	58.8	77.0	85.2	68.9	56.2
C	Greece	Ivory C	Win	53.3	51.2	54.0	46.6	49.4	38.6	52.3	44.8	48.8	39.4	52.1	46.2
D	Uruguay	C Rica	Lose	20.1	24.4	16.6	20.6	17.9	19.5	16.8	20.1	17.8	20.6	16.5	20.5
D	England	Italy	Lose	38.8	40.4	34.0	37.3	39.5	35.6	33.4	37.1	38.0	36.7	33.5	37.0
D	Uruguay	England	Win	40.9	40.4	39.1	39.0	45.3	53.0	41.8	40.5	45.9	52.9	42.1	39.7
D	C Rica	Italy	Win	21.6	25.6	18.7	22.7	23.2	33.7	18.9	23.5	24.3	34.3	19.5	22.6
D	Uruguay	Italy	Win	39.6	40.7	36.9	37.9	45.1	52.7	39.8	40.6	44.3	60.3	39.3	46.5
D	C Rica	England	Draw	20.1	19.0	22.9	28.2	21.9	21.1	24.4	28.6	22.1	21.4	24.7	25.1
E	Switzerland	Ecuador	Win	58.7	56.4	59.1	50.2	54.7	43.1	57.0	47.9	54.0	42.7	56.7	49.3
E	France	Honduras	Win	47.6	47.1	48.4	43.3	44.7	37.7	47.9	41.6	44.9	38.1	47.6	43.0
E	Switzerland	France	Lose	26.0	31.2	21.5	25.6	26.6	31.8	21.1	26.0	26.6	32.0	23.0	25.9
E	Ecuador	Honduras	Win	40.9	42.6	40.1	39.6	41.8	42.2	40.4	38.5	41.2	41.7	39.2	38.6
E	Switzerland	Honduras	Win	61.8	58.2	61.1	52.1	59.5	50.5	59.8	51.2	59.1	49.8	59.4	51.9
E	Ecuador	France	Draw	23.5	20.8	27.3	30.1	23.5	21.7	27.1	29.7	23.6	21.4	27.1	30.3
F	Argentina	Bosnia	Win	55.3	53.1	56.3	48.4	62.0	67.5	58.7	49.6	61.2	66.3	58.7	47.8
F	Iran	Nigeria	Draw	22.6	20.2	27.4	30.5	22.9	19.2	27.3	30.1	23.4	20.3	27.5	30.3
F	Argentina	Iran	Win	68.7	65.0	66.1	57.0	74.8	81.3	69.5	59.1	74.0	85.4	68.0	65.1
F	Bosnia	Nigeria	Lose	27.3	30.2	22.3	25.9	28.8	39.2	23.9	26.6	30.0	30.3	24.3	19.1
F	Argentina	Nigeria	Win	68.8	64.2	65.7	57.1	71.6	70.8	67.4	57.8	72.3	70.9	67.0	56.0
F	Bosnia	Iran	Win	51.9	49.5	52.7	44.8	53.7	52.4	53.9	45.6	54.5	52.8	53.5	44.6
G	Germany	Portugal	Win	44.2	43.2	43.9	40.9	42.2	35.8	44.3	40.2	41.9	36.9	42.8	40.4
G	Ghana	USA	Lose	57.7	53.6	58.1	48.9	55.8	51.5	56.7	47.5	56.9	51.1	57.0	49.7
G	Germany	Ghana	Draw	16.4	16.1	19.2	25.1	17.8	19.5	21.3	25.5	16.8	18.1	20.4	24.9

G	Portugal	USA	Draw	23.3	20.5	26.2	30.8	22.7	22.6	27.2	30.8	22.8	21.9	26.6	30.4
G	Germany	USA	Win	53.3	50.9	53.8	45.7	51.0	42.4	53.6	46.4	49.4	51.1	52.6	56.2
G	Portugal	Ghana	Win	66.3	61.6	64.6	55.4	64.9	58.0	63.2	53.4	66.0	50.5	64.5	44.8
H	Belgium	Algeria	Win	49.7	49.1	51.6	44.1	48.3	44.6	51.1	43.9	48.6	44.1	50.9	44.2
H	Russia	S Korea	Draw	19.8	19.3	23.4	28.5	19.0	17.3	23.2	27.1	18.9	16.0	22.9	28.2
H	Belgium	Russia	Win	48.0	47.1	49.0	42.9	50.1	49.6	50.1	43.7	50.5	49.3	49.8	43.3
H	Algeria	S Korea	Win	55.8	54.5	56.4	48.2	63.4	72.6	58.5	50.9	63.3	71.3	59.1	48.4
H	Belgium	S Korea	Win	68.0	63.3	65.5	56.2	74.2	80.0	68.0	59.5	73.8	77.8	68.4	56.3
H	Algeria	Russia	Draw	23.1	20.2	26.9	30.2	23.4	20.6	27.3	30.1	23.3	20.2	27.3	30.4

Table 5. Percentage of simulated tournaments where correct teams were predicted to reach different stages of knockout tournament.

Knockout stage condition	1 Variable				2 Variables				3 Variables			
	Violating Assumptions		Satisfying Assumptions		Violating Assumptions		Satisfying Assumptions		Violating Assumptions		Satisfying Assumptions	
	All	Inter-C	All	Inter-C	All	Inter-C	All	Inter-C	All	Inter-C	All	Inter-C
Brazil reach Quarter Finals	46.5	42.5	47.4	41.7	53.4	62.6	53.0	45.3	53.5	63.0	52.6	45.6
Columbia reach Quarter Finals	39.7	37.2	41.7	37.5	44.9	48.7	43.8	38.8	45.4	47.2	44.9	37.6
Germany reach Quarter Finals	53.8	46.5	55.6	47.4	51.1	37.7	55.4	46	50.7	42.1	53.9	50.3
France reach Quarter Finals	24.6	25.6	25.1	27.1	21.7	16.7	23.8	25.4	22.1	16.8	24.0	26.5
Holland reach Quarter Finals	21.9	23.6	22.1	23.7	17.3	12.9	19.1	22.1	17.0	13.2	18.9	22.8
Costa Rica reach Quarter Finals	8.0	10.7	6.2	11.2	8.0	14.2	6.7	11.0	8.5	12.5	6.7	9.0
Argentina reach Quarter Finals	53.4	48.2	54.5	47.3	60.7	63.8	58.1	49.5	60.4	64.4	57.8	49.9
Belgium reach Quarter Finals	33.8	33.9	33.7	33.4	34.7	33.8	33.9	33.5	34.5	34.4	34.0	33.5
Brazil reach Semi Finals	27.5	24.6	29.5	24.9	34.3	42.3	34.4	27.9	34.0	42.3	34.1	26.8
Germany reach Semi Finals	35.9	29.6	37.5	30.6	33.3	20.5	36.7	29.3	32.3	23.9	36.0	32.6
Holland reach Semi Finals	9.7	11.2	9.8	11.1	7.4	5.0	8.2	10.0	7.0	5.4	7.8	10.3
Argentina reach Semi Finals	29.4	26.6	30.0	25.3	36.5	43.0	33.7	28.0	36.3	42.9	33.5	26.6
Germany reach Final	20.5	16.7	21.9	17.2	17.9	8.8	20.5	16.0	16.9	10.6	19.9	18.8
Argentina reach Final	14.0	13.5	14.4	12.7	18.2	23.5	16.4	14.7	18.3	24.0	16.5	13.9
Holland finish 3rd	2.4	2.8	2.5	2.6	1.8	1.1	1.9	2.3	1.6	1.4	1.9	2.4
Germany win the tournament	11.7	9.6	12.7	9.9	9.6	3.9	11.6	9.6	9.1	5.0	11.3	11.4

Table 6. Percentage correctness score for each prediction for different stages of the tournament.

Stage	1 Variable				2 Variables				3 Variables			
	Violating Assumptions		Satisfying Assumptions		Violating Assumptions		Satisfying Assumptions		Violating Assumptions		Satisfying Assumptions	
	All	Inter-C	All	Inter-C	All	Inter-C	All	Inter-C	All	Inter-C	All	Inter-C
Pool	21.4	20.8	21.2	19.7	22.1	22.5	21.6	19.9	22.1	22.6	21.5	19.9
Knockout	4.3	4.0	4.4	4.0	4.5	4.4	4.6	4.1	4.5	4.5	4.5	4.2
All	25.7	24.8	25.6	23.7	26.7	26.8	26.1	24.0	26.6	27.0	26.1	24.1

Figure 2 shows that the methods based on data where the modelling assumptions were violated earned higher evaluation scores than when the assumptions were satisfied. The evaluation score was also higher when previous data from all matches was used to produce the models than when data from just inter-continental tournaments was used. Despite difference in World ranking points (PD) being the only significant predictor of goal difference within matches, models restricted to this variable earned 1 evaluation point less than those based on 2 or 3 variables. The mean of 25.95 ranking points achieved by models using all three predictor variables was only marginally greater than the 25.90 when two of the variables (PD and DD) were used.

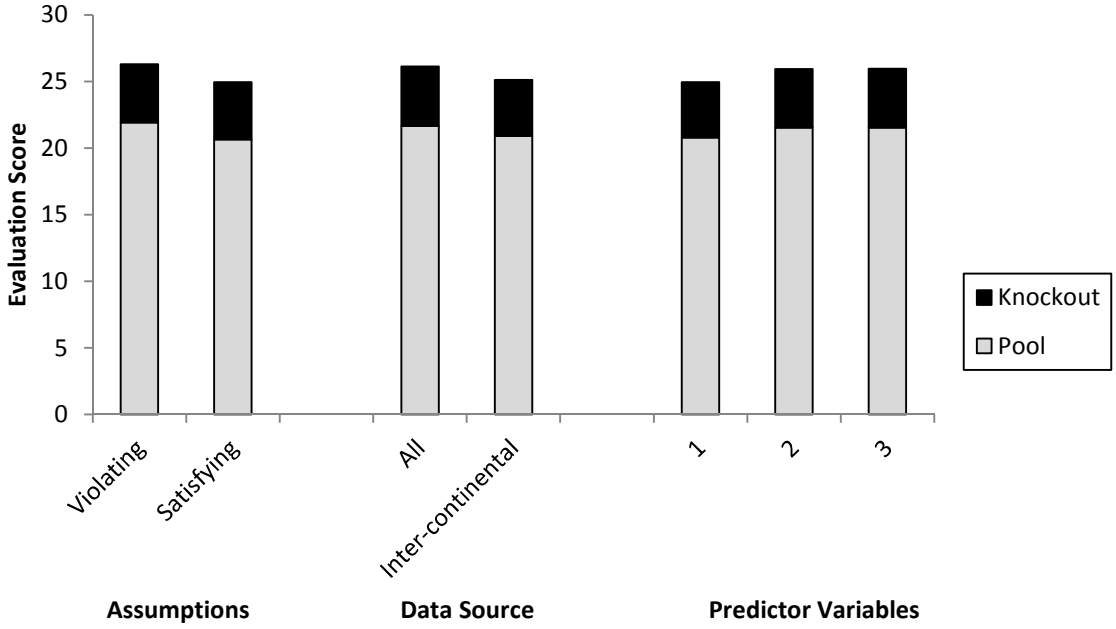

Figure 2. Evaluation scores for different sets of models.

Discussion

Predicting the outcomes of international soccer matches remains a difficult task as is evidenced by the evaluation scores being below 50% of the marks available for all 12 models. The pool stage matches have three different outcomes with the most likely outcome of some matches having a lower probability of occurring than 0.5. While the actual tournament winners, Germany, were predicted to be semi-finalists by all 12 models, no more than 13% of simulated tournaments were won by Germany according to any of the models. The FIFA World Cup remains a wide open tournament with many close matches and some knockout stage matches being decided after extra time and penalty shoot outs. There are many similarities between the modal predictions from the models with all 12 including the same 8 quarter-finalists and the same 4 semi-finalists. This is because the FIFA World ranking points of teams is the dominant variable within the models. Brazil were the winners and Argentina were the runners up in the modal simulated tournaments derived from the two models with the highest evaluation scores. The main difference between these two models and the remaining models (all of which had Spain winning more simulated tournaments than any other team) is the weighting they placed on distance travelled. These two models violated some modelling assumptions and were created using data from previous inter-continental tournaments only. Table 1 shows that the

weighting given to distance travelled in these two models was 3 times that in the corresponding models created using previous data from all international tournaments. The negative association of distance travelled with goal difference within matches meant that these two models favoured teams from South and Central America more than other models.

The difference in the evaluation scores between the models based on data satisfying and violating the assumptions is very little. The models where the assumptions were violated by the data used to create them were slightly more accurate. This makes it difficult to justify the effort of transforming data in order to satisfy the assumptions of multiple linear regression. This agrees with the conclusions of previous predictive modelling studies (O'Donoghue and Williams, 2004; O'Donoghue, 2005, 2006, 2010). The differences between these two sets of models can be explained in terms of the variability in predicted outcomes. In order to satisfy the assumptions of multiple linear regression, outliers had to be removed in the dependent goal difference variable (Fallowfield et al., 2005, p180). This excluded high scoring matches from the data used to create the models and artificially reduced the spread of goal differences that were predicted. The excluded matches were real matches and the goal difference values did not result from measurement error. Tables 2 and 3 present models where the assumptions were violated beside the corresponding models where the assumptions were satisfied. These tables show fewer drawn matches within the simulations where the assumptions were violated than when they were satisfied. This is the case for all 48 pool matches in each of the 6 pairs of models. Furthermore, Table 1 shows that the residual values for each model violating the assumptions had a higher variability (SD) than the corresponding model where the assumptions were satisfied. The variables were not transformed using methods recommended by Nevill (2000) within the models that violated the assumptions. This meant that the negative impact of distance travelled was considered to be linear within these models as opposed to curvilinear (taken to the power of 2/3 in two of the models where the assumptions were satisfied). This amplified the chances the simulator gave to teams from North and Central America meaning that methods where the assumptions were violated gained more evaluation points from Argentina's, Brazil's, Columbia's, Uruguay's and Costa Rica's progress in the actual World Cup than other models.

The models based on 2 or all 3 predictor variables achieved higher evaluation scores than those based on just FIFA World ranking points. The very slight difference in evaluation score between models based on 2 and 3 predictor variables suggests that recovery days has little impact on the accuracy of prediction with the higher evaluation scores of these models being attributed to distance travelled. This supports the evidence that home advantage is still important in soccer and agrees with existing research concluding the prevalence of home advantage in sport (Corneya and Carron, 1992; Nevill et al., 1996, 2002; Pollard and Pollard, 2005; Pollard and Gómez, 2009; Goumas, 2014). Spain winning the 2010 World Cup in South Africa and Germany winning the 2014 World Cup in Brazil are anecdotal evidence that the impact of home advantage is diminishing. However, the success of teams from South and Central America in the 2014 World Cup was reflected in the evaluation scores of models that included distance travelled as a predictor variable.

While the 96 matches from previous inter-continental tournaments may be more representative of the FIFA World Cup than Continental tournaments, the models based on all tournament data achieved higher evaluation scores than those based only on inter-continental tournament data. This is possibly explained by the greater volume of matches in the wider data set being more representative of international soccer than the inter-continental data which ultimately came from just 3 tournaments.

The models can be used to investigate the effect of different factors on results. For example, there has been speculation about the effect of playing in the Amazonia Arena in Manaus on teams' performances in the next matches. There were 4 pool matches played in this stadium. Seven of the 8 teams involved played subsequent matches; Honduras's match in the Amazonia Arena was their final match of the tournament. The next match of the other 7 teams can be considered in terms of goal difference and expected goal difference according to the models. The difference between these two variables indicates how much better a team did than expected. As an example, the expected results from the most successful model were used (all 3 variables, the previous data used to create the model coming from inter-continental tournaments only and violating the assumptions of multiple linear regression). Only one of the 7 teams that played in the Amazonia Arena did better than expected in the next match; Portugal. The seven teams had a goal difference that was 0.384 less than the expected goal difference in their first match after playing in the Amazonia Arena. This example is limited because of actual goal difference results being integer numbers while predicted goal difference results are real numbers. However, where factors are investigated using a greater number of matches, individual matches will have less impact on the mean observed goal difference making such investigations more meaningful.

In conclusion, the current investigation has provided evidence that models based on more than one predictor variable are more accurate than those based on a single variable and that predictive accuracy is increased by using a larger data set of previous cases to create the models. The results also suggest that removing outliers from variables and transforming variables so that data satisfy the modelling assumptions does not lead to improved accuracy of match outcome prediction in soccer.

References

Courneya, K.S. and Carron, A.V. (1992). The home advantage in sports competitions: a literature review. *Journal of Sport and Exercise Psychology, 14*, 13-27.

Fallowfield, J.L., Hale, B.J. and Wilkinson, D.M. (2005). *Using statistics in sport and exercise science research*. Chichester, UK: Lotus Publishing.

Goumas, C. (2014). Tyranny of distance: Home advantage and travel in international club football. *International Journal of Performance analysis in Sport, 14*, 1-13.

Indonesia (2014). www.indo.com/distance, accessed June 2014.

Nevill, A. (2000). Just how confident are you when publishing the results of your research?. *Journal of Sports Sciences, 18*, 569-570.

Nevill, A.M., Newell, S.M. and Gale, S. (1996). Factors associated with home advantage in English and Scottish soccer matches. *Journal of Sports Sciences, 14*, 181-186.

Nevill, A.M., Balmer, N.J. and Williams, A.M. (2002). Can crowd reactions influence decisions in favour of the home side?. In W. Spinks, T. Reilly, and A. Murphy (Eds.), *Science and Football IV* (pp. 308-319). London: Routledge.

Newell, J., Aitchison, T. and Grant, S. (2010). *Statistics for sport and exercise science: a practical approach.* Harlow, UK: Prentice Hall.

Ntoumanis, N. (2001). *A step-by-step guide to SPSS for sport and exercise studies.* London: Routledge.

O'Donoghue, P.G. (2005). Evaluation of Computer-based Predictions of the Euro2004 Soccer Tournament, 5th International Symposium of Computer Science in Sport, Hvar, Croatia, 25th-28th May, Book of abstracts, pp.36.

O'Donoghue, P.G. (2006). The effectiveness of satisfying the assumptions of predictive modeling techniques: an exercise in predicting the FIFA World Cup 2006. *International Journal of Computing Science in Sport, 5*(2), 5-16.

O'Donoghue, P.G. (2009). Predictions of the 2007 Rugby World Cup and Euro 2008, 3rd International Workshop of the International Society of Performance Analysis of Sport, Lincoln, UK, 6th-7th April 2009.

O'Donoghue, P.G. (2010). The effectiveness of satisfying the assumptions of predictive modelling techniques: an exercise in predicting the FIFA World Cup 2010. *International Journal of Computer Science in Sport, 9*(3), 15-27.

O'Donoghue, P.G. (2012). The Assumptions Strike Back! A comparison of prediction models for the 2011 Rugby World Cup. *International Journal of computer Science in Sport, 11*(2), 29-40.

O'Donoghue, P.G. and Williams, J. (2004). An evaluation of human and computer-based predictions of the 2003 rugby union world cup. *International Journal of Computer Science in Sport, 3*(1), 5-22.

O'Donoghue, P.G., Dubitzky, W., Lopes, P., Berrar, D., Lagan, K., Hassan, D., Bairner, A. and Darby, P., (2004). An Evaluation of quantitative and qualitative methods of predicting the 2002 FIFA World Cup. *Journal of Sports Sciences, 22*, 513-514.

Pollard, R. and Gómez, M. A. (2009). Home advantage in football in South-West Europe: Long-term trends, regional variation, and team differences. *European Journal of Sport Science, 9*, 341-352.

Pollard, R. and Pollard G. (2005). Long-term trends in home advantage in professional team sports in North America and England (1876–2003). *Journal of Sports Sciences, 23*, 337–350.

Tabachnick, B.G. and Fidell, L.S. (1996). *Using multivariate statistics, 3rd Edition.* New York: Harper Collins.

A New Model of Head-Up Display Dive Computer Addressing Safety-Critical Rate of Ascent and Returning Gas Pressure - A Pilot Trial

Peter Buzzacott[1,2], Andreas Schuster[3], Amir Gerges[4], Walter Hemelryck[5] , Kate Lambrechts[1], Dennis Madden[6], Virginie Papadopoulou[5,7], Yurii Tkachenko[8], Aleksandra Mazur[1], Frauke Tillmans[1], Miroslav Rozloznik[1] Qiong Wang[1] Andreas Møllerløkken[9], François Guerrero[1] & Arne Sieber[3]

[1]*Laboratoire Optimisation des Régulations Physiologiques (ORPhy),Université de Bretagne Occidentale*

[2]*School of Sports Science, Exercise and Health, University of Western Australia*

[3]*Seabear GmbH*

[4]*Divers Alert Network Europe*

[5]*Haute Ecole Paul Henri Spaak*

[6]*School of Medicine, University of Split*

[7]*Department of Bioengineering, Imperial College London*

[8]*Medical University of Gdansk*

[9]*Department of Circulation and Medical Imaging, Norwegian University of Science and Technology*

Abstract

Head up displays (HUD) are beneficial in diving situations when the diver uses both hands for an activity, e.g. photography, scientific work, operating a diver propulsion vehicle or during diver training. They remove the need to locate a submersible pressure gauge or remember to look at a personal dive computer. A new model of HUD, one that can easily be retrospectively fitted to a recreational diver's regulator hose outside the mask lens, has been developed. A pilot study of 93 open circuit recreational dives was conducted over one week in Croatia, to assess the HUD-user interface. An electronic survey was developed and completed twice after 16 dives. Mean maximum depth was 23 m and mean total dive time 38 mins. 34 dives (37%) were made with the HUD and 59 made with traditional submersible pressure gauges. There was good test-retest agreement (kappa score=0.9) between repeated surveys. The HUD was relatively easy to attach and could be operated without the necessity of reading the user manual. The HUD has two potential mechanisms for preventing rapid ascent injuries. Firstly, displaying an ascent rate warning directly in the divers' field of vision and, secondly, by reducing the likelihood of an out-of-gas situation.

KEYWORDS: ASCENT WARNING, COMPUTERS – DIVING, INJURY PREVENTION, LOW AIR, RECREATIONAL DIVING, RISK FACTORS

INTRODUCTION

Established risk factors for recreational diving injuries include running out of gas and/or rapid ascent (Buzzacott, Pikora, Rosenberg, & Heyworth, 2012; Buzzacott, Rosenberg, Heyworth, & Pikora, 2011). Reasons for either of these events to occur were suggested by a panel of diving experts (Buzzacott, Rosenberg, & Pikora, 2009). Failing to monitor the submersible pressure gauge (SPG) was the most likely reason given for running out of gas while recreational diving (Buzzacott et al., 2009). Among 1032 recreational dives in Western Australia, 183 ended with less than 50 bar remaining in the cylinder (Buzzacott et al., 2011). Concurring with expert opinion, surprise at the remaining air pressure was significantly associated with returning with less than 50 bar in reserve (Buzzacott et al., 2011). Not only is the last 50 bar shaded in red on popular SPGs but it is also the minimum pressure at which manufacturers must ensure each SPG displays pressure within a tolerance specified by European standard EN250:2000.

Rate of ascent is now also more commonly displayed on personal dive computers (Buzzacott et al., 2012) either as a vertical column graph or numerically. Regardless, displayed pressure or rate of ascent will not assist recreational divers avoid related injuries if divers fail to physically look at the appropriate display. The inability of commercial and military divers to read diving information displays during conditions of very low visibility, coupled with the advantages of hand-free operation, has led to the development of Head-Up Display (HUD) dive computers. Head up displays are also beneficial in other diving situations such as when the diver uses both hands for a certain activity, e.g. underwater photography, scientific work, operating a diver propulsion vehicle (DPV) or during diver training such as controlled emergency ascent practice.

The CompuMask ® and DataMask ® HUDs (American Underwater Products, San Leandro, CA) are recreational diving computers which are fully integrated into a traditional diving mask based on an LCD display together with an optical system (Sieber, Kuch, Enoksson, & Stoianova-Sieber, 2012). They feature a single colour LCD screen and wireless tank pressure readout. Potential disadvantages of this system include the limited range of face profiles each mask is suited to and that the life of the dive computer is tied to the life of the mask. The US Navy Experimental Diving Unit has developed a variety of head up displays (Gallagher, 1999) however the units are costly and available only as military equipment.

An alternative model of HUD, one that can be retrospectively fitted to a recreational diver's regulator hose outside the mask lens, has been under development for military rebreather divers, as previously reported (Koss & Sieber, 2011a, 2011b; Sieber, Schuster, Reif, Madden, & Enoksson, 2013). It was originally designed to be located inside a full face mask, which required the installation of a port on the side window of the full face mask visor (Sieber et al., 2012). Installing such a port however is a major change of a full face mask, thus a new CE certification would be required. As a consequence, the device was redesigned to be located outside of the visor. A further adaptation now allows it to be attached to a second stage regulator.

A typical HUD for diving is mounted in close vicinity (5 to 10 cm) from the diver's eye. A person with normal eyesight cannot focus on such short distances, thus an optical system has to be introduced. In its simplest form, it consists of a single convex lens placed between the screen and the eye. Attachment of the HUD to either full face mask or second stage regulator is shown in Figure 1.The core component is a tiny colour OLED screen with a diagonal of 24 mm. A two lens system consisting of a plano-convex lens and a bi-concave lens creates a virtual image of the OLED display with a size of 20x30cm at a distance of approximately 1m.

The plane side of the convex lens is in contact with water.

Figure 1: The HUD fitted to a full-face mask and to a second stage regulator

The HUD screen shows depth, decompression information including remaining no decompression time, decompression ceiling and time to surface, tilt-compensated compass heading, cylinder pressure reading and ascent and descent rate graphs (Figure 2). The decompression algorithm is a pure unmodified Bühlmann ZH-L16C for either air or nitrox diving, with diver selectable gradient factors (Bühlmann, 1995). The tank pressure sensor housing includes a 350 bar ceramic pressure sensor and a rechargeable Li Ion battery with a capacity of 500 mAh.

Figure 2: Typical display as seen by the diver. In this case maximum depth 39.5m, current depth 25.1m, dive time 20:50min, heading SE 145°, battery OK (green), remaining no decompression time 15min, cylinder pressure 62bar, ascent rate warning (red >10 m/min) and temperature 19°C.

The aim of this pilot study was to assess the utility of this portable HUD among a group of recreational divers..

Methods

As part of the Marie Curie Initial Training Network Phypode program (Buzzacott, 2013) a group of 12 early career researchers and postdoctoral fellows met with the manufacturer of the new HUD in Labin, Croatia. Prior to the experiments the HUD was tested and CE certified against EN13319. The tank pressure sensor was tested according to the relevant parts of EN250 by DEKRA, (Essen, Germany) a notified body for personal protective equipment. The team had two days to familiarise themselves with the equipment during test dives. Each diver

wearing a HUD and/or a full face mask was accompanied by at least one dive buddy not wearing test equipment. During this familiarization period a user survey was drafted to assess display readability, clarity of information, ease of use and comfort. After each test dive the survey questionnaire was assessed for face and content validity and refined as needed. To assess the intuitiveness of the HUD no instruction manuals were provided. On day three the survey was pilot-tested and any last, minor revisions were made as needed, to ensure they were understood by divers from non-English speaking backgrounds.

The revised version was then transformed into an electronic format and uploaded to the internet survey site SurveyMonkey (Finley, 1999) which provides a popular internet-based electronic survey program. Four divers then tested two HUD units during a dive at night. Thereafter divers either wore the HUD or did not, and repeated dives were made during the day, from shore or from a commercial dive charter boat, with full face masks or regular dive masks and independent second stage regulators. The intention was to test the HUD in a wide variety of conditions including night/day, poor/good visibility, from shore or in 'blue water' from a boat.

Immediately following each dive the diver had dive details logged on the day's manifest and those wearing the HUD then also anonymously completed the electronic survey online. One hour later each diver repeated the online survey. On each occasion the diver nominated a fictional nickname so that the two surveys could be matched. No personal information was collected.

This training exercise conformed with the approved work plan of the Phypode Project and, as a research training exercise, no Human Research Ethics Committee approval was required.

Analysis

A Cohen's kappa coefficient(k) was calculated to assess agreement between the first and second time each person completed the post-dive survey (Smeeton, 1985). The kappa-statistic measure is a value between -1 and 1, with 0 corresponding to the value expected by chance and 1 perfect agreement. Interpretation of the values (suggested by Landis and Koch) are given as: below 0.00 – Poor, 0.00–0.20 – Slight, 0.21–0.40 – Fair, 0.41–0.60 – Moderate, 0.61–0.80 – Substantial and 0.81–1.00 – Almost perfect (Landis & Koch, 1977a, 1977b).

Results

Ninety-three open circuit recreational dives were made to assess the utility of the HUD, with mean dive time of 38 mins and mean maximum depth of 23 m. Eleven (12%) were made from a boat and 82 (88%) from shore, 15 of those (16%) by divers wearing a full face mask. 34 dives (37%) were made with the HUD, the remainder were made by accompanying buddies not wearing test equipment. All dives finished with at least 50 bar remaining in the cylinder, as per local regulations, which prevented the less-than-50 bar warning from being observed underwater but two divers simulated this by closing the dive cylinder's pillar valve and inhaling, to lower the pressure in the first stage below 50 bar. Lifetime diving experience (number of post-certification open water dives) of the divers ranged from 4~1500.

Matched pairs of post-dive surveys were collected from divers making the last 16 HUD-wearing dives, after the survey questionnaire had been pilot-tested on the third day. One subject failed to complete the re-test survey on the same day as the initial post-dive survey so this dive was excluded from further statistical analysis. Mean agreement (k) between the

remaining 15 pairs of tests was 0.9, implying high reliability. A summary of the first responses of the group is presented in Table 1.

Table 1: Post-dive questionnaire responses by divers wearing the HUD

	First Response			Second Response		
Survey question	Yes	No	N/a	Yes	No	N/a
Was the HUD bright enough?	15	0	-	15	0	-
Was maximum depth clear to you?	10	5	-	10	5	-
Was your current depth clear to you?	15	0	-	15	0	-
Was the compass clear to you?	15	0	-	15	0	-
Were the compass bearings clear to you?	15	0	-	12	3	-
Was tank pressure clear to you?	13	2	-	13	2	-
Was the remaining no decompression time clear to you?	13	2	-	14	1	-
Was ascent rate and bar graph clear to you?	6	9	-	6	9	-
Was the water temperature clear to you?	10	5	-	11	4	-
Was the HUD convenient to set up?	13	2	-	11	3	1
During the entry into the water, did the HUD stay in place?	6	9	-	5	9	1
During the dive, did HUD stay in place?	9	6	-	8	6	1
Did the HUD restrict your vision?	4	11	-	3	12	-
Did you notice the Low on Air warning?	2	13	-	3	12	-
Did you notice a rapid ascent warning?	4	11	-	1	11	3
If both of your hands were occupied during the dive were you still able to read the display?	14	1	-	15	0	1
Was the HUD comfortable to wear?	14	1	-	14	1	-
Did the HUD run out of battery?	0	15	-	0	15	-

From the comment section it was identified that the ascent rate warning (a coloured column graph) was not displayed clearly enough to be immediately obvious. However with repetitive use, the displayed information became clearer as the wearer became more familiar with the HUD. Concurring with the responses in Table 1, divers wearing ordinary masks and regulators reported experiencing more movement of the HUD during dives compared with full face mask divers, especially while entering the water or clearing the mask. It was suggested that a safety-stop countdown at the end of each dive might be added to the dive computer's software.

Discussion

Despite the obvious pilot-trial nature of this study, to our knowledge this is the first report on the utility of a retrofitted HUD among recreational divers. HUD masks with in-built dive computers have been marketed to recreational divers in recent years but this is the first recreational dive computer of its kind that can be retrospectively fitted to either a full-face mask or else to an ordinary scuba second-stage regulator and then viewed through an existing recreational diving mask.

Given the developmental nature of the models tested there were some improvements suggested, including a more obvious ascent-rate warning and greater stability of the display during entry to the water. Despite these, we propose this HUD has the potential to reduce the likelihood of unintentionally running low on gas, in particular because an obvious warning is displayed (a flashing screen) at 50 bar. At this early stage this is speculative however and a human trial comparing returning gas pressures between HUD and SPG users may help determine what effect the HUD has upon running low on gas. In one anecdotal report an inexperienced diver was offered the loan of a DPV in a confined shallow bay. After five minutes the diver reported having only 140 bar left in his dive cylinder and after a further five minutes of circling the small bay he reported having only 50 bar left. At this time it was observed that his 'octopus' (reserve second stage regulator) was trailing behind him, releasing the contents of his cylinder while he was being propelled through the water. This highlighted an advantage of the HUD over an SPG while operating a DPV.

Overall the HUD added substantially to the total mass of the second stage regulator held in the mouth which reduced the comfort of divers, especially those with the least experience, who reported diving usually only while on vacation. There was high agreement (14/15 divers, 93%) that it would require more experience diving with the HUD before comfort levels returned to normal. The majority of dives were made with regular masks and second-stage regulators while the HUD was originally designed for use with full-face masks. Once underwater, however, the HUD is essentially neutrally buoyant.

Conclusion

The effect a HUD with an obvious ascent-rate warning might have upon actual ascent rate during recreational dives cannot be predicted from this pilot trial alone. Ascending rapidly has been found associated with losing buoyancy control among West Australian recreational divers though which is causal requires further research. Previous research established that running out of gas was significantly more likely to implicate diving injuries when associated with a rapid ascent (Acott, 1994). Therefore, the HUD described in this study has two potential mechanisms focussed upon prevention of rapid ascent injuries. Firstly, by displaying an ascent rate warning directly in the divers field of vision and, secondly, by potentially reducing the likelihood of any rapid ascent occurring in conjunction with an out-of-gas situation. Further

research and development will continue to address the occurrence of these known risk factors for diving injuries among recreational divers.

Acknowledgements/Conflict of Interest

The authors thank the S.P. Marina dive centre for logistic support during the week-long study. This research was supported by the European Commission under the FP7-PEOPLE-2010-ITN program (grant agreement n° 264816). In June 2012 Peter Buzzacott was unconditionally given a DG-05 dive computer by Hollis (San Leandro, CA, USA), a company that manufactures dive computers. One of the test divers in the pilot trial survey (A. Schuster) is an employee of SeaBear GmbH, the company that manufactures the HUD. No other conflicts of interest are declared.

References

Acott, C. (1994). Diving incidents - Errors divers make. *Procedings of Safe Limits: An international dive symposium. Cairns, Australia, October 21-23, 1994* (pp. 25-38). Cairns, QLD: Division of Workplace Safety.

Bühlmann, A. A. (1995). *Tauchmedizin*. Berlin: Springer-Verlag.

Buzzacott, P. (2013). Phypode Fellows update. *Diving and Hyperbaric Medicine*, *43*(3), 178-179.

Buzzacott, P., Pikora, T., Rosenberg, M., & Heyworth, J. (2012). Rapid ascent and buoyancy problems in Western Australia. *Diving and Hyperbaric Medicine*, *42*(1), 30-35.

Buzzacott, P., Rosenberg, M., Heyworth, J., & Pikora, T. (2011). Risk factors for running low on gas in recreational divers in Western Australia. *Diving and Hyperbaric Medicine*, *41*(2), 85-89.

Buzzacott, P., Rosenberg, M., & Pikora, T. (2009). Using a Delphi technique to rank potential causes of scuba diving incidents. *Diving and Hyperbaric Medicine*, *39*(1), 29-32.

Finley, R. (1999). *SurveyMonkey*. Retrieved June 16, 2013 from www.surveymonkey.com

Gallagher, D. G. (1999). Development of miniture, head-mounted, virtual image displays for navy divers. *Proceeding of the Oceans '99 MTS / IEEE Conference: Riding the crest into the 21st century, Seattle, Washington, September 13-16, 1999* (pp. 1098-1104). New York, NY: Institute of Electrical and Electronics Engineers.

Koss, B., & Sieber, A. (2011a). Development of a graphical head-up display (HUD) for rebreather diving. *International Journal of the Society for Underwater Technology*, *29*(4), 1-6.

Koss, B., & Sieber, A. (2011b). Head mounted display for diving computer platform. *Journal of Display Technology*, *7*(4), 193-199.

Landis, J. R., & Koch, G. G. (1977a). An application of hiererchical kappa type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, *33*, 363-374.

Landis, J. R., & Koch, G. G. (1977b). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Sieber, A., Kuch, B., Enoksson, P., & Stoianova-Sieber, M. (2012). Development of a head-up displayed diving computer capability for full face masks. *International Journal of the Society of Underwater Technology*, *30*(4), 195-199.

Sieber, A., Schuster, A., Reif, S., Madden, D., & Enoksson, P. (2013). Head-up display system for closed circuit rebreathers with antimagnetic wireless data transmission. *Marine Technology Society Journal*, *47*(6), 42-51.

Smeeton, N. C. (1985). Early history of the kappa statistic. *Biometrics, 41*, 795.

Prediction of American Football Plays Using Pattern Recognition

Robert Strange & Lior Shamir

Lawrence Technological University

21000 W Ten Mile Rd., Southfield, MI 48075, USA.

Abstract

American football is played primarily as a sequence of two types of plays – runs and passes, determined before the play by the team that has the possession of the ball. All other plays are much less frequent, and considered "special plays". In this study we show that the type of play can be predicted using pattern recognition methods with accuracy higher than random chance based on several indicators that reflect the status of the game such as the down, time left in the game, score difference, etc. These values were used to predict the next play by using Support Vector Machine and Weighted Nearest Distance classification schemes. Experiments with data from all plays in 11 National Football League (NFL) seasons show that the ability to predict the next offensive play can be as high as 74% for an entire season of a single team.

KEYWORDS: AMERICAN FOOTBALL, NFL, MACHINE LEARNING, PATTERN RECOGNITION

Introduction

American football is currently the most popular sport in North America. The popularity of the game can be reflected by the TV rating of the National Football League championship finals – the "Super Bowl" – which is North America's most viewed television broadcast. According to Forbes, the annual revenue of the National Football League alone is $9.5B, and the combined value of the NFL teams is over $37B. These numbers exclude college and high school American football, as well as other American football leagues inside and outside North America.

As a highly technical game, personal accolades and team records within each game and season are inevitable. Forecasting and effective analysis of players, plays, and team efficacy under different game conditions would benefit team owners, head coaches, and defensive and offensive coordinators. The ability to predict what the opposing team will do under certain game conditions can provide an edge over the competition during the game, and a tool to simulate and prepare the defense for a match.

The American football "passing premium" is defined as the existence of a balance between the number of passing and running plays, even though there is a greater expected return in passing plays (Alamar, 2007). That is, a passing play is expected to earn on average more yardage than a running play, while also having more risk for turnovers or change of possession due to inability of the team to advance ten yards in order to keep possession of the ball. Rockerbie (2008) produced an optimal model of balance between passing and running plays, and tested it

using the actual share of running plays in 2006 NFL season. The results also confirmed the tradeoff between yardage return and risk (Rockerbie, 2008).

The ability to make accurate forecasting and analysis of sequence of plays is of high value in sports (Rudelsdorfer, 2014). For example, Boulier and Stekler (2003) predicted the outcome of American football games based on power scores, and showed experimental results using the outcome of the NFL seasons of 1994-2000 and power scores provided by the *New York Times*, as well as some naïve models, the betting market, and the opinions of sports editors. Results showed that the betting market is the best predictor, followed by the predictions based on power scores (Boulier and Stekler, 2003).

In this study we show that individual American football plays can be predicted based on several numerical values that reflect the status of the game. The ability to predict the play can be used to gain an advantage during the game, as well as to better prepare for the game or improve game simulations.

Methods

Data

The raw data used in this study was taken from advancedfootballanalytics.com. The data include information about all plays in the regular season, such that each play is one raw in the table. Each raw includes a short text description about the play, from which the information used in this study was extracted. Kicking plays and penalties were ignored in this study. Interceptions were considered as an attempt of a pass play. The information extracted for each play is described in Table 1.

Table 1. The variables used as indicators for each play.

Variable	Units	Range	Type
The time until half	seconds	0-3600	integer
The time left in the game	seconds	0-1800	integer
Down	count	1-4	integer
Distance to a first down	yards	1-10	integer
Quarter of play	count	1-4	integer
Scoring difference	count	Unlimited	integer
Yard line (distance to the goal)	yards	1-100	integer
Offensive team score	count	Unlimited	integer
Defensive team score	count	Unlimited	integer

Time left to the end of the quarter was not used because unlike other sports such as basketball, the second quarter of each half starts exactly from where the previous quarter ended.

Each play also had the ground truth, which is whether the play started from the position

described by the variables in Table 1 was a running or a passing play. The plays were separated to seasons, and the plays of each season was separated to teams, such that the offensive plays of each of the 32 NFL teams in each of the 11 seasons (2002-2012) provided the dataset.

Pattern recognition

Two methods were used to predict the play based on the status of the game: Weighted Nearest Distance (WND) and Support Vector Machine (SVM). WND (Shamir, 2008; Shamir et al., 2008) is a pattern recognition method designed for large feature spaces. WND works by first computing Fisher discriminant scores (Bishop, 2006) for all features, as defined by

$$W_f = \frac{\sum_{c=1}^{N}\left(\overline{T_f} - \overline{T_{f,c}}\right)^2}{\sum_{c=1}^{N}\sigma_{f,c}^2} \cdot \frac{N}{N-1}$$

where W_f is the Fisher discriminant score, N is the total number of classes, T_f is the mean of the values of feature f in the entire dataset, $T_{f,c}$ is the mean of the values of feature f in the class c, and $\alpha^2_{f,c}$ is the variance of feature f among all samples of class c. The distance between a test sample x and a class c is then measured by:

$$d(x,c) = \frac{\sum_{t \in T_c}\left[\sum_{f=1}^{|x|} W_f^2\left(x_f - t_f\right)^2\right]^p}{|T_c|}$$

where T is the training set, T_c is the training set of class c, t is a feature vector from T_c, |x| is the length of the feature vector x, x_f is the value of image feature f, W_f is the Fisher discriminant score of feature f, $|T_c|$ is the number of training samples of class c, d(x,c) is the computed distance from a given sample x to class c, and p is the exponent, which is set to −5 as thoroughly discussed with experimental results in (Orlov et al., 2008). More information about the WND method is provided in (Shamir, 2008; Shamir et al., 2008).

The experiments were performed using 700 plays, 350 passing plays and 350 running plays. Three hundred plays from each type were randomly allocated for training, and the remaining plays were used for testing. Each experiment was repeated 20 times such that in each run different plays were randomly allocated to training and test sets.

The second classification method used was Support Vector Machine (Vapnik, 1995). SVM is a supervised learning method for binary classification that divides a high-dimensional feature space to determine the hyperplanes with the maximal margin between samples of different classes.

A linear Support Vector Machine can be defined as an optimization problem under constraints as described in the equation:

$$\min\|w\|$$

Under the constraint

$$y_i(w \cdot x - b) \geq 1$$

where *i* is the index of the sample, y_i is the class of the sample {-1,1}, *x* is the feature vector of sample *i*, and *w* is the normal vector to the hyperplane. The problem can be solved by replacing $\|w\|$ with $0.5*\|w\|^2$, and with the Lagrange multiplier α producing the equation

$$\arg \min_{w} \max_{\alpha \geq 0} \left\{ \frac{1}{2} \|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i (w \cdot x - b) - 1] \right\}$$

Once the hyperplane is determined using the training data, a new sample point can be classified based on the area of the feature space they are in, as divided by the hyperplane. More information about SVM is provided in (Vapnik, 1995). In this study we used the SVM^{light} (Joachims, 2008) implementation of the SVM algorithm.

Results

The ability to predict the type of play based on the status of the game was measured by the number of plays that the method predicted correctly, divided by the total number of attempts to predict the type of play. Since the number of passing plays in the dataset is equal to the number of running plays, random guessing of the next play would provide 50% of accuracy. The experimental results indicate that in many teams the next offensive play can be predicted based on the status of the game in accuracy higher than random guessing. For instance, in the season of 2002 the offensive plays of the Baltimore Ravens were predicted in ~69% of the cases, while the plays of the Chicago Bears could not be predicted with accuracy higher than random guessing of 50%. While the play prediction accuracy changed between teams, it did not change substantially between seasons. In all seasons the average play prediction accuracy (of all 32 teams) was consistent, and ranged between ~59% to ~60%.

SVM provided slightly better accuracy in predicting the next play. The highest prediction accuracy of ~74% was observed in the offensive plays made by the Detroit Lions in the season of 2011 and Saint Louis Rams in 2004, but six other teams also had play prediction accuracy of over 70% (Arizona in 2010 and 2009, Colts in 2009, Eagles in 2005, Vikings in 2005, and Raiders in 2005). Figure 1 shows the number of teams that their offensive plays were predicted with accuracy higher than different thresholds. Since the data cover 11 seasons, and there are 32 different teams in the NFL, the total number of teams is 352.

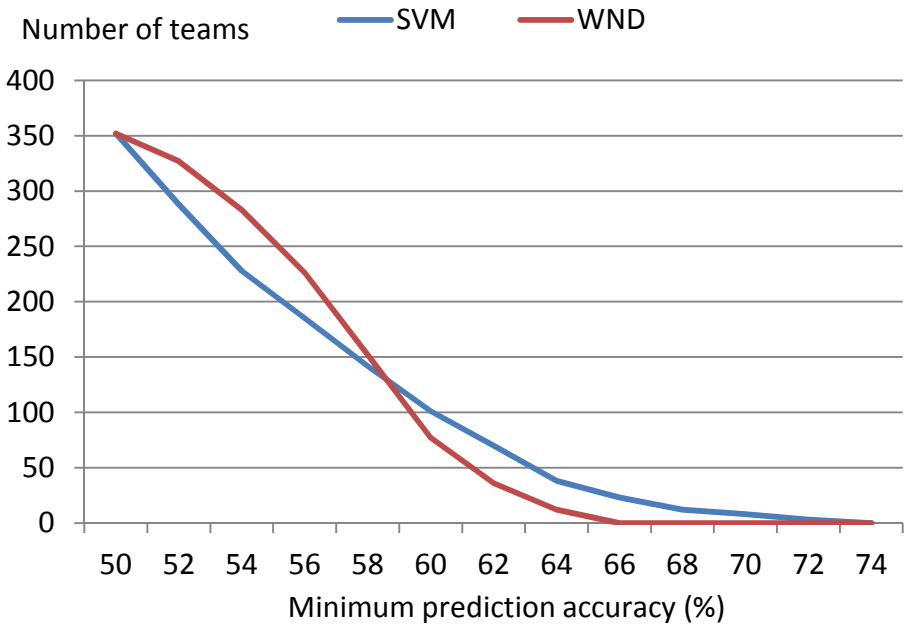

Figure 1. The number of teams that their offensive plays can be predicted with accuracy above a certain threshold.

As the graph shows, WND and SVM are about equally informative for predicting the next play in an American football play, but overall SVM performed better than WND. While in some teams the selection of offensive plays is less systematic and cannot be predicted with accuracy beyond random guessing, in some cases the offensive plays over a certain season can be predicted with accuracy higher than 70%.

Conclusion

In this study we showed that the next play in an American football game can be predicted with accuracy above random guessing by analyzing the status of the game. Clearly, the type of play cannot be predicted with perfect accuracy, as the selection of the play is a human decision, and therefore depends on intuition in addition to the complex analysis of the game status. Also, the analysis does not take into consideration injuries, as well as weather conditions such as wind or rain that might also affect the decision regarding the offensive play. Future research will focus on the analysis of the predictability of offensive plays and its association to characteristics of the offense such as the identity of the offensive coordinators, offense productivity, quarterback ranking, etc.

References

Alamar D. W. (2007). The passing premium puzzle. *Journal of Quantitative Analysis in Sports*, *2*(4), 5.

Bishop C. M. (2006). *Pattern recognition and machine learning*. Springer, Berlin.

Boulier B. L., & Stekler H.O. (2003). Predicting the outcomes of National Football League games. *International Journal of Forecasting*, *19*(2), 257-270.

Joachims T. (2008). SVM^{light} – support vector machine, Cornell University.

Rockerbie D. W. (2008). The passing premium puzzle revisited. *Journal of Quantitative Analysis in Sports 4*(2).

Rudelsdorfer, P., Schrapf, N., Possegger, H., Mauthner, T., Bischof, H., & Tilp, M. (2014). A novel method for the analysis of sequential actions in team handball. *International Journal of Computer Science in Sport, 13*(2).

Shamir L., Orlov N., Eckley D. Mark, Macura T., Johnston J., & Goldberg I.G. (2008). WC – an open source utility for biological image analysis. *Source Code for Biology and Medicine*, *3*(1), 13.

Shamir, L. (2008). Evaluation of face datasets as tools for assessing the performance of face recognition methods. *International Journal of Computer Vision*, *79*(3), 225-230

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, Berlin.