Simulation leagues: Analysis of competition formats

David Budden[1], Peter Wang[2], Oliver Obst[2], Mikhail Prokopenko[2]

[1] The University of Melbourne, Parkville, VIC 3010, Australia
[2] Statistical Learning, CSIRO Computational Informatics, Epping, NSW 1710, Australia
david.budden@unimelb.edu.au
{peter.wang,oliver.obst,mikhail.prokopenko}@csiro.au

Abstract. The selection of an appropriate competition format is critical for both the success and credibility of any competition, both real and simulated. In this paper, the automated parallelism offered by the RoboCupSoccer 2D simulation league is leveraged to conduct a 28,000 game round-robin between the top 8 teams from RoboCup 2012 and 2013. A proposed new competition format is found to reduce variation from the resultant statistically significant team performance rankings by 75% and 67%, when compared to the actual competition results from RoboCup 2012 and 2013 respectively. These results are statistically validated by generating 10,000 random tournaments for each of the three considered formats and comparing the respective distributions of ranking discrepancy.

1 Introduction

1.1 The RoboCup humanoid challenge

RoboCup (the "World Cup" of robot soccer) was first proposed in 1997 as a standard problem for the evaluation of theories, algorithms and architectures in areas including artificial intelligence (AI), robotics and computer vision [1]. This proposal followed the observation that traditional AI problems were increasingly unable to meet these requirements and that a new challenge was necessary to initiate the development of next-generation technologies.

The overarching RoboCup goal of developing a team of humanoid robots capable of defeating the FIFA World Cup champion team, coined the "Millennium Challenge", has proven a major factor in driving research in AI and related areas for over a decade, with a search for the term "RoboCup" in a major literature database yielding over 23,000 results. Since 1997, researchers and competitors have decomposed this ambitious pursuit into two complementary categories [2]:

– **Physical robot league:** Using physical robots to play soccer games. This category now contains many different leagues for both wheeled robots (small-sized [3] and mid-sized leagues [4]) and humanoids (standard platform league [5] and humanoid league [6]), with each focusing on different aspects of physical robot design [7], motor control and bipedal locomotion [8, 9], real-time localisation [10, 11] and computer vision [12–14].
– **Software agent league:** Using software or synthetic agents to play soccer games on an official soccer server over a network. This category contains both 2D [15–17] and 3D [18] simulation leagues.
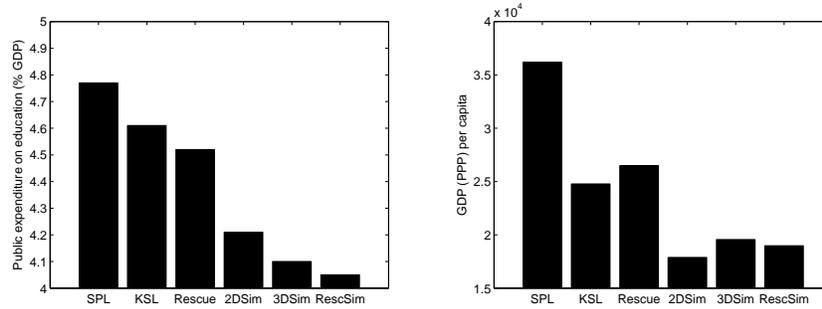
Fig. 1. PEoE (public expenditure on education as a percentage of GDP [23]) and GDP/cap (gross domestic product at purchasing power parity per capita [24]) for the home country of each participating RoboCup 2013 team, averaged over each of the six largest RoboCup leagues. Each of the three major simulation leagues (2D, 3D and rescue) exhibit significantly lower values than those requiring the purchase or development of physical robots.

The annual RoboCup competition, which attracted 2,500 participants and 40,000 spectators from 40 countries in 2013 [19], now exhibits a number of non-soccer competitions. The oldest and largest of these, RoboCupRescue, is also separated into physical and simulation leagues [20, 21].

1.2   Significance of simulation leagues

The RoboCupSoccer simulation leagues traditionally involve the largest number of international participating teams, reaching 40 in 2013 [22]. The ability to simulate soccer matches without physical robots abstracts away low-level issues such as image processing and motor breakages, allowing teams to focus on the development of complex team behaviours and strategies for a larger number of autonomous agents. The remainder of this section expands upon some specific contributions of the RoboCupSoccer simulation leagues toward the 'Millennium Challenge".

Financial inclusiveness of competing nations

The physical robots required by non-simulation leagues remain particularly expensive. By removing these costs and those associated with robot repairs and transportation, the simulation leagues allow institutes with access to less funding to actively contribute to and participate in the RoboCup initiative. To quantify this claim, Fig. 1 presents the PEoE (public expenditure on education as a percentage of GDP [23]) and GDP/cap (gross domestic product at purchasing power parity per capita [24]) for the home country of each participating RoboCup 2013 team, averaged over each of the six largest RoboCup leagues. The countries participating in the standard platform league, which requires teams to field five Aldebaran Nao humanoid robots, have the highest average PEoE and GDP/cap of any league considered. The kid-sized humanoid and rescue

leagues, each of which require the purchase or construction of physical robots, also involve countries with a high average PEoE and GDP/cap. Each of the three major simulation leagues (2D, 3D and rescue) exhibit significantly lower values, suggesting that the inclusion of simulation leagues supports financial inclusiveness within the competition.

Statistically significant analyses by automated competition parallelism

The automation of multiple parallel games makes RoboCupSoccer simulation leagues ideal platforms for analysing the complexities of complex team behaviours. Most team games and sports (both real and virtual) are characterised by rich, dynamic interactions that influence the contest outcome. As described by Vilar *et al.*, "quantitative analysis is increasingly being used in team sports to better understand performance in these stylized, delineated, complex social systems" [25]. Early examples of such quantitative analysis include *sabermetrics*, which attempts to "search for objective knowledge about baseball" by considering statistics of in-game activity [26]. A recent study by Fewell *et al.* involved the analysis of basketball games as networks, with properties including degree centrality, clustering, entropy and flow centrality calculating from measurements of ball position throughout the game [27]. This idea was extended by Vilar *et al.*, who considered the local dynamics of collective team behaviour to quantify how teams occupy sub-areas of the field as a function of ball position [25]. Recently, Cliff *et al.* presented several information-theoretic methods of quantifying dynamic interactions in football games, using the RoboCupSoccer 2D simulation league as an experimental platform [28].

The ability to automate thousands of simulation league games allows for the analysis of competition formats to determine which best approximate the true performance rankings of competing teams. The selection of an appropriate competition format is critical for both the success and credibility of any competition. Unfortunately, this choice is not straightforward: The format must minimise randomness relative to the true performance ranking of teams while keeping the number of games to a minimum, both to satisfy time constraints and retain the interest of participants and spectators alike. Furthermore, maintaining competition interest introduces a number of constraints to competition formats: As an example, multiple games between the same two opponents (the obvious method of achieving a statistically significant ranking) should be avoided.

The remainder of this paper quantifies the appropriateness of different tournament formats (a major consideration in many human sports) by determining the statistically significant performance rankings of 2012 and 2013 RoboCupSoccer 2D simulation teams. A new competition format is then proposed and verified by leveraging the automated parallelism facilitated by the 2D simulation league platform. In addition to demonstrating the utility of simulation leagues for statistical analysis of team sport outcomes given some system perturbation, it is anticipated that the adoption of the proposed format would improve the success and credibility of the RoboCupSoccer simulation leagues in future years.

2   Previous competition formats

The following two competition formats were adopted by the RoboCupSoccer 2D simulation league in 2012 and 2013:

- In 2012, a total of 20 games were played to determine the final rank of the top 8 teams. Specifically, the top 4 teams played 6 games each (3 quarterfinal round-robin, 2 semifinal and 1 final/third place playoff), and the bottom 4 teams player 4 games each.
- In 2013, a *double-elimination* system was adopted, where a team ceases to be eligible to place first upon having lost 2 games [29, 30]. A total of 16 games were played to determine the final rank of the top 8 teams. Specifically, 14 games were played in the double-elimination format (i.e. $2n - 2, n = 8$) in addition to 2 classification games.

Previously, it has been unclear whether this change in competition format improves the fairness and reproducibility of the final team rankings. In general, lack of reproducibility is due to non-transitivity of team performance (a well-known phenomena that occurs frequently in actual human team sports). This may be addressed by a round-robin competition (where all 28 possible pairs of teams play against one another), yet it is also unclear whether this increase in the number of games is guaranteed to improve ranking stability.

3   Methods of ranking team performance

Before evaluating different competition formats, it is necessary to establish a fair (i.e. statistically significant) ranking of the top 8 RoboCupSoccer 2D simulation league teams for 2012 and 2013. This was accomplished by conducting an 8-team round-robin for both years, where all 28 pairs of teams play approximately 1000 games against one another. In addition, two different schemes were considered for point calculation:

- **Continuous scheme:** Teams are ranked by sum of average points obtained against each opponent across all 1000 games.
- **Discrete scheme:** Firstly, the average score between each pair of teams (across all 1000 games) is rounded to the nearest integer (e.g. "1.9 : 1.2" is rounded to "2 : 1"). Next, points are allocated for each pairing based on these rounded results: 3 for a win, 1 for a draw and 0 for a loss. Teams are then ranked by sum of these points received against each opponent.

The final rankings generated for 2012 and 2013 RoboCupSoccer 2D simulation league teams under these two schemes are presented in Sec. 5.1. Finally, in order to formally capture the overall difference between two rankings $\mathbf{r^a}$ and $\mathbf{r^b}$, the L_1 distance is utilised:

$$d_1(\mathbf{r^a}, \mathbf{r^b}) = \|\mathbf{r^a} - \mathbf{r^b}\|_1 = \sum_{i=1}^{n} |r_i^a - r_i^b| \,, \qquad (1)$$

where i is the index of the i-th team in each ranking, $1 \leq i \leq 8$. The difference between rankings for different competition formats are presented in Sec. 5.2.

4   Proposed competition format

Sec. 3 describes two schemes under which statistically significant rankings of RoboCup-Soccer 2D simulation league teams can be achieved. However, it remains unclear whether the previously adopted competition formats are able to replicate these rankings with minimal noise for considerably fewer games, or whether a new format may achieve improved results in this regard. One possible format involves the following two steps:

– Firstly, a preliminary round-robin is conducted where 1 game is played for all 28 pairs of teams.
– Following the rankings obtained in the previous step, 4 classification games are played: The final between the top 2 teams and playoffs between third and fourth, fifth and sixth, and seventh and eighth places. It is possibly to use the best-of-three format for each of these classification games.

The 32 games required involved in this competition format could still fit readily in a 1-2 day time frame, particularly with 2 games running simultaneously as per RoboCup 2013.

5   Results

5.1   Statistically significant rankings versus previous competition formats

Following iterated round-robin and two point calculation schemes described in Sec. 3, statistically significant rankings were generated for the top 8 RoboCupSoccer 2D simulation league teams for 2012 and 2013. These results are presented below.

RoboCup 2012 results

The final round-robin results of the top 8 teams for RoboCup 2012 are presented in Table 1 and Table 2, for the continuous and discrete scoring schemes described in Sec. 3 respectively. Results are ordered according to actual performance at RoboCup 2012, $\mathbf{r^a}$.

Table 2 presents the continuous (non-rounded) scores averaged across the approximately 1000 games for each pair in the round-robin, in addition to the points allocated according to the discretisation scheme (3 for a win, 1 for a draw and 0 for a loss). The tie-breaker is the rounded goal difference (not shown), which was used only to separate first place (WrightEagle, +39 points) from second (Helios, +26 points). The final ranking corresponds exactly with that generated under the continuous scheme, as presented in Table 1.

Despite the agreement between continuous and discrete scoring schemes, it is obvious that this ranking (generated from the results of approximately 28,000 games) disagrees significantly from the actual RoboCup 2012 results. This can be quantified using the distance metric defined in (1):

$$d_1(\mathbf{r^a}, \mathbf{r^c})_{2012} = |1-2|+|2-1|+|3-5|+|4-4|+|5-6|+|6-8|+|7-3|+|8-7| = 12,$$

r^a	Team	Helios	Wright	Marlik	Gliders	GDUT	AUT	Yushan	RobOTTO	Points	Goal Diff	Rank, r^c
1	Helios		1.397	2.442	2.517	2.948	2.970	2.880	2.998	18.152	+ 26.0	2
2	Wright	1.406		2.792	2.835	2.900	2.998	2.970	2.998	18.899	+ 38.7	1
3	Marlik	0.309	0.129		1.147	2.121	2.804	0.874	2.615	9.999	+ 0.3	5
4	Gliders	0.261	0.102	1.396		1.809	2.957	0.903	2.863	10.291	+ 3.4	4
5	GDUT	0.029	0.074	0.633	0.960		2.955	0.552	2.597	7.800	- 6.0	6
6	AUT	0.007	0.001	0.107	0.026	0.024		0.003	0.209	0.377	- 39.3	8
7	Yushan	0.084	0.021	1.822	1.875	2.316	2.994		2.993	12.105	+ 6.5	3
8	RobOTTO	0.001	0.001	0.233	0.087	0.228	2.418	0.005		2.973	- 29.6	7

Table 1. Round-robin results (average goals scored) for the top 8 teams from RoboCup 2012, ordered according to their final competition rank, r^a. The final points for each team were determined by summing the average points scored against each opponent over approximately 1000 games, resulting in the round-robin with continuous point allocation scheme ranking, r^c.

	Helios	Wright	Marlik	Gliders	GDUT	AUT	Yushan	RobOTTO	Goals	Points	r^d
Helios		2.3 : 2.3	1.4 : 0.1	1.6 : 0.1	4.4 : 0.2	7.7 : 0.0	4.5 : 0.7	7.6 : 0.1	29.5 : 3.5	19	2
Wright	2.3 : 2.3		3.2 : 0.3	3.3 : 0.2	5.8 : 1.2	12.1 : 0.1	7.2 : 1.0	10.1 : 0.2	44.0 : 5.3	19	1
Marlik	0.1 : 1.4	0.3 : 3.2		0.46 : 0.56	1.4 : 0.4	2.3 : 0.1	0.7 : 1.2	2.1 : 0.2	7.4 : 7.1	10	5
Gliders	0.1 : 1.6	0.2 : 3.3	0.56 : 0.46		1.9 : 1.2	4.3 : 0.1	1.4 : 2.2	4.6 : 0.8	13.1 : 9.7	12	4
GDUT	0.2 : 4.4	1.2 : 5.8	0.4 : 1.4	1.2 : 1.9		4.0 : 0.2	2.0 : 3.9	3.4 : 0.8	12.4 : 18.4	6	6
AUT	0.0 : 7.7	0.1 : 12.1	0.1 : 2.3	0.1 : 4.3	0.2 : 4.0		0.1 : 7.1	0.7 : 3.1	1.3 : 40.6	0	8
Yushan	0.7 : 4.5	1.0 : 7.2	1.2 : 0.7	2.2 : 1.4	3.9 : 2.0	7.1 : 0.1		6.7 : 0.4	22.8 : 16.3	13	3
RobOTTO	0.1 : 7.6	0.2 : 10.1	0.2 : 2.1	0.8 : 4.6	0.8 : 3.4	3.1 : 0.7	0.4 : 6.7		5.6 : 35.2	3	7

Table 2. Round-robin results (average goals scored and discretised points allocated) for the top 8 teams from RoboCup 2012, ordered according to their final competition rank, r^a. Discretised points are determined by calculating the average number of goals scored over approximately 1000 games rounded to the nearest integer, then awarding 3 points for a win, 1 point for a draw and 0 points for a loss. The resultant round-robin with discrete point allocation scheme ranking, r^d, is equivalent to that generated under the continuous scheme.

where r^a represents the actual RoboCup 2012 rankings and r^c represents the ranking generated under continuous scoring scheme round-robin. This large difference suggests that the 2012 competition format did not succeed in capturing the true team performance ranking.

RoboCup 2013 results

The final round-robin results of the top 8 teams for RoboCup 2013 are presented in Table 3 and Table 4, for the continuous and discrete scoring schemes described in Sec. 3 respectively. Results are ordered according to actual performance at RoboCup 2013, r^a, and presented in the same format as Table 1 and Table 2 for RoboCup 2012.

Unlike RoboCup 2012, there is a slight disagreement between the rankings generated using continuous and discrete scoring schemes, with a swap between third and

r^a	Team	Wright	Helios	Yushan	Axiom	Gliders	Oxsy	AUT	Cyrus	Points	Goal Diff	Rank, r^c
1	Wright		1.877	2.470	2.880	2.397	2.901	2.991	2.792	18.308	+ 22.5	1
2	Helios	0.883		2.841	2.940	2.194	2.343	2.969	2.767	16.937	+ 14.9	2
3	Yushan	0.406	0.093		2.506	1.892	1.557	2.059	0.921	9.434	- 1.3	4
4	Axiom	0.072	0.042	0.367		0.590	0.395	1.224	1.023	3.713	- 14.5	8
5	Gliders	0.437	0.490	0.884	2.249		1.612	1.871	0.828	8.371	- 2.0	6
6	Oxsy	0.065	0.385	1.159	2.437	1.105		2.225	2.167	9.543	- 2.2	3
7	AUT	0.006	0.017	0.718	1.491	0.878	0.575		0.731	4.416	- 14.0	7
8	Cyrus	0.137	0.136	1.791	1.740	1.926	0.632	2.046		8.408	- 3.4	5

Table 3. Round-robin results (average goals scored) for the top 8 teams from RoboCup 2013, ordered according to their final competition rank, r^a. The final points for each team were determined by summing the average points scored against each opponent over approximately 1000 games, resulting in the round-robin with continuous point allocation scheme ranking, r^c.

	Wright	Helios	Yushan	Axiom	Gliders	Oxsy	AUT	Cyrus	Goals	Points	r^d
Wright		1.9 : 1.2	2.8 : 0.9	4.9 : 0.3	2.5 : 0.7	5.4 : 0.8	6.4 : 0.3	3.4 : 0.6	27.3 : 4.8	21	1
Helios	1.2 : 1.9		2.8 : 0.2	4.1 : 0.2	1.2 : 0.2	2.2 : 0.4	4.1 : 0.1	2.5 : 0.2	18.1 : 3.2	18	2
Yushan	0.9 : 2.8	0.2 : 2.8		2.7 : 0.8	1.8 : 1.1	1.4 : 1.2	1.7 : 0.8	0.9 : 1.4	9.6 : 10.9	11	3
Axiom	0.3 : 4.9	0.2 : 4.1	0.8 : 2.7		1.0 : 2.3	0.7 : 2.7	0.9 : 1.1	1.4 : 2.0	5.3 : 19.8	1	8
Gliders	0.7 : 2.5	0.2 : 1.2	1.1 : 1.8	2.3 : 1.0		1.3 : 1.0	1.8 : 1.1	0.9 : 1.7	8.3 : 10.3	7	6
Oxsy	0.8 : 5.4	0.4 : 2.2	1.2 : 1.4	2.7 : 0.7	1.0 : 1.3		2.3 : 0.8	2.2 : 1.0	10.6 : 12.8	11	4
AUT	0.3 : 6.4	0.1 : 4.1	0.8 : 1.7	1.1 : 0.9	1.1 : 1.8	0.8 : 2.3		0.8 : 1.8	5.0 : 19.0	1	7
Cyrus	0.6 : 3.4	0.2 : 2.5	1.4 : 0.9	2.0 : 1.4	1.7 : 0.9	1.0 : 2.2	1.8 : 0.8		8.7 : 12.1	10	5

Table 4. Round-robin results (average goals scored and discretised points allocated) for the top 8 teams from RoboCup 2012, ordered according to their final competition rank, r^a. Discretised points are determined by calculating the average number of goals scored over approximately 1000 games rounded to the nearest integer, then awarding 3 points for a win, 1 point for a draw and 0 points for a loss. The tie-breaker is the total of rounded goal differences (not shown). The resultant round-robin with discrete point allocation scheme ranking, r^d, is slightly different to that generated under the continuous scheme.

fourth teams. Again using the distance metric defined in (1), the difference between these rankings and the actual RoboCup 2013 results can be quantified:

$$d_1(\mathbf{r^a}, \mathbf{r^c})_{2013} = |1-1|+|2-2|+|3-4|+|4-8|+|5-6|+|6-3|+|7-7|+|8-5| = 12,$$

$$d_1(\mathbf{r^a}, \mathbf{r^d})_{2013} = |1-1|+|2-2|+|3-3|+|4-8|+|5-6|+|6-4|+|7-7|+|8-5| = 10,$$

where $\mathbf{r^a}$ represents the actual RoboCup 2013 rankings, while $\mathbf{r^c}$ and $\mathbf{r^d}$ represent the ranking generated under continuous and discrete scoring schemes of round-robins respectively. It is evident that the 2013 double-elimination format yielded as much overall divergence as the 2012 single-elimination format, but with slightly fewer individual discrepancies. It is also clear that, given very small points differences between adjacent

	Helios	Wright	Marlik	Gliders	GDUT	AUT	Yushan	RobOTTO	Points	Rank	$\mathbf{r^P}$
Helios		4 : 1	4 : 0	1 : 0	4.4 : 0.2	1 : 0	4.5 : 0.7	2 : 0	21	1	1
Wright	1 : 4		2 : 1	2 : 0	5 : 1	12.1 : 0.1	6 : 1	10.1 : 0.2	18	2	2
Marlik	0 : 4	1 : 2		1 : 0	1 : 0	2.3 : 0.1	1 : 1	2.1 : 0.2	13	3	4
Gliders	0 : 1	0 : 2	0 : 1		1.9 : 1.2	2 : 0	1.4 : 2.2	3 : 0	9	5	5
GDUT	0.2 : 4.4	1 : 5	0 : 1	1.2 : 1.9		1 : 0	3 : 2	3.4 : 0.8	9	6	6
AUT	0 : 1	0.1 : 12.1	0.1 : 2.3	0 : 2	0 : 1		0.1 : 7.1	1 : 0	3	7	8
Yushan	0.7 : 4.5	1 : 6	1 : 1	2.2 : 1.4	2 : 3	7.1 : 0.1		3 : 1	10	4	3
RobOTTO	0 : 2	0.2 : 10.1	0.2 : 2.1	0 : 3	0.8 : 3.4	0 : 1	1 : 3		0	8	7

Table 5. Combined actual and average results for the top 8 teams from RoboCup 2012, ordered according to their final competition rank. Each goal difference represents the actual (integer) game results from RoboCup 2012 where possible. As this previous format does not necessarily require all pairs of teams to play against one another, some of these results are not available: In these cases, the average (continuous-valued) scores from Table 2 were utilised. Using these results, it is possible to infer the final ranking, $\mathbf{r^P}$, for RoboCup 2012 under the competition format proposed in Sec. 4.

teams, it may be necessary to play classification games even after a statistically significant round-robin. It is therefore proposed that the format described in Sec. 4 should improve reliability of the competition outcomes.

5.2 Evaluation of proposed competition formats

In order to evaluate the proposed competition format described in Sec. 4, the actual game results from RoboCup 2012 and 2013 were used where possible. As these previous formats do not necessarily require all pairs of teams to play against one another, some of these results are not available: In these cases, the average scores from Table 2 and Table 4 were utilised for RoboCup 2012 and 2013 respectively.

Using these results, it is possible to infer final rankings for RoboCup 2012 and 2013 under the proposed competition format. These results are presented below.

RoboCup 2012 results

The combined actual and average results of top 8 teams from RoboCup 2012 are presented in Table 5, in addition to the inferred final ranking, $\mathbf{r^P}$, for RoboCup 2012 under the competition format proposed in Sec. 4. Using the distance metric defined in (1), the difference between $\mathbf{r^P}$ and the ranking generated from the 28,000 game round-robin, $\mathbf{r^c}$, can be quantified:

$$d_1(\mathbf{r^P}, \mathbf{r^c})_{2012} = |1-2|+|2-1|+|4-5|+|5-4|+|6-6|+|8-8|+|3-3|+|7-7| = 4.$$

This is a considerably smaller difference than the 12 produced under the actual RoboCup 2012 format, suggesting that the proposed format better captures the true team performance ranking. Furthermore, this result is achieved using a majority of actual game results (i.e. 18 from 28 pairs, with only 10 using the averages from Table 2).

	Wright	Helios	Yushan	Axiom	Gliders	Oxsy	AUT	Cyrus	Points	Rank	$\mathbf{r^P}$
Wright		3 : 1	2.8 : 0.9	4.9 : 0.3	2.5 : 0.7	5 : 3	6.4 : 0.3	7 : 0	21	1	1
Helios	1 : 3		2 : 0	4.1 : 0.2	2 : 0	2.2 : 0.4	4.1 : 0.1	4 : 0	18	2	2
Yushan	0.9 : 2.8	0 : 2			4 : 1	0 : 1	3 : 0	1.7 : 0.8	2 : 0	12	3	3
Axiom	0.3 : 4.9	0.2 : 4.1	1 : 4		3 : 3	1 : 6	2 : 1	1.4 : 2.0	4	7	8
Gliders	0.7 : 2.5	0 : 2	1 : 0	3 : 3		4 : 0	1.8 : 1.1	0.9 : 1.7	10	4	4
Oxsy	3 : 5	0.4 : 2.2	0 : 3	6 : 1	0 : 4		2.3 : 0.8	2.2 : 1.0	9	5	5
AUT	0 : 7	0.1 : 4.1	0.8 : 1.7	1 : 2	1.1 : 1.8	0.8 : 2.3		3 : 1	3	8	7
Cyrus	0.6 : 3.4	0 : 4	0 : 2	2.0 : 1.4	1.7 : 0.9	1.0 : 2.2	1 : 3		6	6	6

Table 6. Combined actual and average results for the top 8 teams from RoboCup 2013, ordered according to their final competition rank. Each goal difference represents the actual (integer) game results from RoboCup 2013 where possible. As this previous format does not necessarily require all pairs of teams to play against one another, some of these results are not available: In these cases, the average (continuous-valued) scores from Table 4 were utilised. Using these results, it is possible to infer the final ranking, $\mathbf{r^P}$, for RoboCup 2013 under the competition format proposed in Sec. 4.

RoboCup 2013 results

The combined actual and average results of top 8 teams from RoboCup 2013 are presented in Table 6, in addition to the inferred final ranking, $\mathbf{r^P}$, for RoboCup 2013 under the competition format proposed in Sec. 4. Again using the distance metric defined in (1), the difference between $\mathbf{r^P}$ and the ranking generated from the 28,000 game round-robin, $\mathbf{r^c}$ or $\mathbf{r^d}$, can be quantified:

$$d_1(\mathbf{r^P}, \mathbf{r^c})_{2013} = |1-1|+|2-2|+|3-4|+|8-8|+|4-6|+|5-3|+|7-7|+|6-5| = 6$$

$$d_1(\mathbf{r^P}, \mathbf{r^d})_{2013} = |1-1|+|2-2|+|3-3|+|8-8|+|4-6|+|5-4|+|7-7|+|6-5| = 4$$

Similarly to the results for 2012, these are considerably smaller differences than the 12 (or 10) produced under the actual RoboCup 2013 format, providing further evidence that the proposed format better captures the true team performance ranking. Again, this result is achieved using a majority of actual game results (i.e. 15 from 28 pairs, with only 13 using the averages from Table 4).

Statistical validation

To statistically validate that the proposed competition format is significantly more appropriate than those adopted at RoboCup 2012 and 2013, 10,000 tournaments were generated for each format by randomly sampling game results from the 28,000 game round-robin. For each tournament, the L_1 distance, $d_1(\mathbf{r^a}, \mathbf{r^b})$ (see eq. (1)), was calculated to capture the discrepancy between the tournament and true team rankings. These results are presented in Fig. 2 for the top 8 teams from RoboCup 2012 (a) and 2013 (b). In both cases, it is evident that the proposed format yields more statistically robust rankings (i.e. smaller L_1 distance) than the formats adopted in RoboCup 2012 (second best) and 2013 (worst).
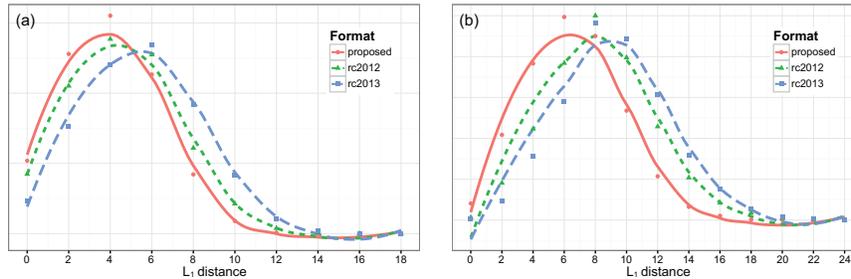
Fig. 2. Discrepancy between tournament and true team rankings, captured as an L_1 distance (see eq. (1)), for 10,000 randomly-generated tournaments structured according to the three considered formats. It is evident that the proposed format (red) yields more statistically robust rankings (i.e. smaller L_1 distance) than the formats adopted in RoboCup 2012 (second best, green) and 2013 (worst, blue), considering the top 8 teams from both RoboCup 2012 (a) and 2013 (b).

6   Conclusions

The selection of an appropriate competition format is critical for both the success and credibility of any competition. This is particularly true in the RoboCupSoccer 2D simulation league, which provides an ideal computational platform for examining different formats by facilitating automated parallel execution of a statistically significant number of games.

A 28,000 game round-robin competition was conducted between the top 8 2D simulation league teams from both RoboCup 2012 and 2013. The difference between the resultant rankings was calculated relative to the actual results of RoboCup 2012 and 2013 (12 and 12 respectively) and compared to those that would have resulted under a proposed new format (4 and 6 respectively). This suggests a significant reduction in randomness relative to true team performance rankings while only requiring the number of games to be increased to 32; a number that would still fit readily in a 1-2 day

time frame, particularly utilising the round-robin parallelism enabled by the stable 2D simulation platform.

The RoboCup "Millennium Challenge" requires robots to exhibit both physical and strategic prowess, necessitating the decomposition of the larger problem into both physical robot and simulation leagues. Although often overlooked, the simulation leagues contribute significantly to this goal, both through improving financial inclusiveness of competing nations and providing a stable platform for statistically significant analysis of team behaviour and competition format. In addition to highlighting the latter of these contributions, it is anticipated that the introduction of the proposed new format will improve the reliability of final competition rankings and consequently success and credibility of the RoboCupSoccer simulation leagues in future years.

References

1. Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I., Osawa, E.: RoboCup: The robot world cup initiative. In: Proceedings of the first international conference on autonomous agents, ACM (1997) 340–347
2. Kitano, H., Asada, M.: The robocup humanoid challenge as the millennium challenge for advanced robotics. Advanced Robotics **13**(8) (1998) 723–736
3. RoboCup Technical Committee: Laws of the RoboCup Small Size League 2013. `http://robocupssl.cpe.ku.ac.th/_media/rules:ssl-rules-2013-2.pdf` (2013)
4. RoboCup Technical Committee: Middle Size Robot League rules and regulations for 2013. `http://wiki.robocup.org/images/9/98/Msl_rules_2013.pdf` (2013)
5. RoboCup Technical Committee: RoboCup Standard Platform League (NAO) rule book. `http://www.tzi.de/spl/pub/Website/Downloads/Rules2013.pdf` (2013)
6. RoboCup Technical Committee: RoboCup Soccer Humanoid League rules and setup. `http://www.tzi.de/humanoid/pub/Website/Downloads/HumanoidLeagueRules2013-05-28.pdf` (2013)
7. Ha, I., Tamura, Y., Asama, H., Han, J., Hong, D.W.: Development of open humanoid platform DARwIn-OP. In: SICE Annual Conference (SICE), 2011 Proceedings of, IEEE (2011) 2178–2181
8. Fountain, J., Walker, J., Budden, D., Mendes, A., Chalup, S.: Motivated reinforcement learning for improved head actuation of humanoid robots. In: RoboCup 2013: Robot Soccer World Cup XVII, Springer (2013)
9. Budden, D., Walker, J., Flannery, M., Mendes, A.: Probabilistic gradient ascent with applications to bipedal robot locomotion. In: Australasian Conference on Robotics and Automation (ACRA). (2013)
10. Annable, B., Budden, D., Mendes, A.: Nubugger: A visual real-time robot debugging system. In: RoboCup 2013: Robot Soccer World Cup XVII, Springer (2013)
11. Budden, D., Prokopenko, M.: Improved particle filtering for pseudo-uniform belief distributions in robot localisation. In: RoboCup 2013: Robot Soccer World Cup XVII, Springer (2013)
12. Budden, D., Fenn, S., Walker, J., Mendes, A.: A novel approach to ball detection for humanoid robot soccer. In: Advances in Artificial Intelligence (LNAI 7691), Springer (2012)
13. Budden, D., Fenn, S., Mendes, A., Chalup, S.: Evaluation of colour models for computer vision using cluster validation techniques. In: RoboCup 2012: Robot Soccer World Cup XVI (LNAI 7500), Springer (2013)

14. Budden, D., Mendes, A.: Unsupervised recognition of salient colour for real-time image processing. In: RoboCup 2013: Robot Soccer World Cup XVII, Springer (2013)
15. Chen, M., Foroughi, E., Heintz, F., Huang, Z., Kapetanakis, S., Kostiadis, K., Kummeneje, J., Noda, I., Obst, O., Riley, P., Steffens, T., Wang, Y., Yin, X.: RoboCup Soccer Server. `http://wwfc.cs.virginia.edu/documentation/manual.pdf`
16. Prokopenko, M., Obst, O., Wang, P., Budden, D., Cliff, O.: Gliders2013: Tactical analysis with information dynamics. In: RoboCup 2013 symposium and competitions: Team description papers, Eindhoven, The Netherlands, June 2013. (2013)
17. Prokopenko, M., Wang, P., Obst, O.: Gliders2014: Dynamic tactics with voronoi diagrams. In: RoboCup 2014 symposium and competitions: Team description papers, Joo Pessoa, Brazil, July 2014. (2014)
18. RoboCup Technical Committee: RoboCup Soccer Simulation League 3D competition rules and setup. `http://homepages.herts.ac.uk/~sv08aav/RCSoccerSim3DRules2013.1.pdf` (2013)
19. Butler, K.: RoboCup 2013: Humanoid robots play soccer for world title. `http://www.upi.com/Science_News/Blog/2013/07/01/RoboCup-2013-Humanoid-robots-play-soccer-for-world-title/9671372691642/` (2013)
20. Kitano, H., Tambe, M., Stone, P., Veloso, M., Coradeschi, S., Osawa, E., Matsubara, H., Noda, I., Asada, M.: The RoboCup synthetic agent challenge 97. In: RoboCup-97: Robot Soccer World Cup I, Springer (1998) 62–73
21. Kitano, H., Tadokoro, S.: Robocup rescue: A grand challenge for multiagent and intelligent systems. AI Magazine **22**(1) (2001) 39
22. Bai, A., Chen, X., MacAlpine, P., Urieli, D., Barrett, S., Stone, P.: WrightEagle and UT Austin Villa: RoboCup 2011 Simulation League Champions. In: RoboCup 2011: Robot Soccer World Cup XV. Springer (2012) 1–12
23. World Bank: World Development Indicators 2012. `http://data.worldbank.org/data-catalog/world-development-indicators` (2012)
24. World Bank: GDP per capita, PPP (current international $). `http://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD` (2012)
25. Vilar, L., Araújo, D., Davids, K., Bar-Yam, Y.: Science of winning soccer: Emergent pattern-forming dynamics in association football. Journal of Systems Science and Complexity **26**(1) (2013) 73–84
26. Grabiner, D.: The sabermetrics manifesto. `http://seanlahman.com/baseball-archive/sabermetrics/sabermetric-manifesto/` (2004)
27. Fewell, J., Armbruster, D., Ingraham, J., Petersen, A., Waters, J.: Basketball teams as strategic networks. PloS one **7**(11) (2012) e47445
28. Cliff, O., Lizier, J., Wang, R., Wang, P., Obst, O., Prokopenko, M.: Towards quantifying interaction networks in a football match. In: RoboCup 2013: Robot Soccer World Cup XVII, Springer (2013)
29. David, H.A.: The method of paired comparisons. DTIC Document **12** (1963)
30. Edwards, C.T.: Double-elimination tournaments: Counting and calculating. The American Statistician **50**(1) (1996) 27–33