A Statistical Analysis of the Trading Agent Competition 2001

Pier Luca Lanzi and Alessandro Strada
Politecnico di Milano, Dipartimento di Elettronica e Informazione,
Artificial Intelligence and Robotics Laboratory

We present a statistical analysis on the data from the Second Trading Agent Competition (TAC01). Our goal is to study the effectiveness of this competition as a testbed for the evaluation of trading strategies in artificial markets. Our results suggest that in TAC01 no agent performed *significantly* better than all the others, from a pure statistical viewpoint. Instead, it reveals groups of agents which performed *significantly* better than others. Thus, our results suggest that although this competition may not give a *quantitative* evaluation of the agents' trading strategies, it can still provide some useful *qualitative* evaluation.

Categories and Subject Descriptors: J.7 [**Computer Applications**]: Computer in other systems

General Terms: Economics

Additional Key Words and Phrases: Agent, Auction, E-Commerce

1. INTRODUCTION

The Trading Agent Competition (TAC) is an annual event which brings together people involved in the study of electronic marketplaces and in the development of autonomous software agents for on-line trading. The competition challenges artificial trading agents with difficult issues regarding bidding strategies, market prediction, and resource allocation [Wellman et al. 2002]. TAC considers an artificial market scenario in which artificial travel agents have to satisfy a set of customer requests for short trips to an hypothetic town. Each customer specifies constraints on flights, on hotel accommodations, and eventually on entertainments. Each travel agent (i.e., software agent) must satisfy the requests of a group of customers.

In this paper, we present a statistical analysis of the data from the Second Trading Agent Competition (TAC01) [Wellman et al. 2001], held October 14^{th} in Tampa, Florida, where we participated with our agent, polimi_bot. The analysis is aimed at studying the effectiveness of this type of competition as a testbed to evaluate trading strategies in artificial markets. The analysis presented here suggests that in general during TAC01 no agent performed *significantly* better than all the others,

Address: P. L. Lanzi and A. Strada, Politecnico di Milano, Dipartimento di Elettronica e Informazione, P.zza L. Da Vinci 32, Milano, I-20133 Italy.

from a pure statistical viewpoint. Therefore it supports the organizers decision not to declare any *winner*. However, our analysis reveals groups of agents which performed *significantly* better than others. Therefore, we argue that, although these competitions may not give a *quantitative* evaluation of the agents' trading strategies, they can still provide some useful *qualitative* evaluations.

2. THE TRADING AGENT COMPETITION

The TAC01 competition was structured into four main phases: (i) the qualifying round, (ii) the seeding round, (iii) the semifinals, and (iv) the finals. The initial qualifying round selected sixteen agents for the semifinals. The next seeding round served to divide these sixteen agents into two groups of eight agents each, which formed the two heats of the semifinals. The four best performing agents from each semifinal pool, competed into the final. During the qualifying and seeding rounds, about 1750 games were played; on average, in these two phases each agent played more than 500 games. During the semifinals, 11 games were played in each heat, and during the finals, only 24 games were played; the agents involved into the last two rounds played at most 35 games. Note that during the qualifying games, and during the seeding rounds, the organizers allowed agent modifications.

Each round consisted of a set of *games*; each game involved eight trading agents competing one against the others. Each agent represents a travel agent which acts on the behalf of eight customers; each customer wants to travel from the hypothetic *TACTown* to Tampa and back, during a notional five day period. Customers are characterized by random preferences for best arrival date, best departure date, an hotel room reservation value, and reservation values for each of the three types of entertainment events requested. The (artificial) travel agent assembles travel packages buying the three types of goods (*flights*, *hotel rooms*, *entertainment tickets*) in separate on-line auctions, with different rules, which run for 12 minutes. Flights are sold with a *Single Seller Auction*, hotel rooms are sold through *16^{th} price English Auctions* which is a combination of English and Vickrey auctions. Tickets for entertainments are sold through a *Continuous Double Auction* (CDA). There are obvious interdependencies. For instance, customers need an hotel room for every night between arrival and departure; in addition they can attend entertainment events only during that interval. The agents' goal is to maximize customers' satisfaction. At the end of a game, the final score for an agent is computed as the difference between the sum of the individual customers' utilities and the agent's net expenditure. The agents' score in each round is determined as a weighted sum of the scores in each game. In particular, later scores were weighted more heavily, to encourage early experimentations, while rewarding agents with stabler performance at the end of the round see [Wellman et al. 2001] for further details on TAC01.

TAC01 has seen the participation of 28 teams, from eight nations, three continents. Most agents implemented strategies based on heuristics or solvers; some were based on parametric algorithms. More details on the agents can be found in [Wellman et al. 2001]; [Wellman et al. 2002] illustrates the competition in details and compare the strategies of two finalists, namely livingagents and ATTac-2001; [Strada 2002] illustrates our agent polimi_bot. The whole TAC competition ran on the Michigan Internet AuctionBot platform [Wurman 1998].

3. THE STATISTICAL ANALYSIS OF TAC01

To analyze the results of TAC01 we consider the raw scores of each game in the *qualifying* and *seeding* rounds. Semifinals and finals are left out since the number of games played in each round is limited and does not allow statistically reliable results. Note that we use raw data *not* the weighted scores used to draw up the results at the end of each round. In fact weights would alter the scores average and variance, making data infeasible for statistical analysis. We also do not take into account the worst ten scores for each agent. A preliminary analysis showed that such lower scores are usually outliers, probably originated from bugs in agents' code rather than from flaws in agents' strategies. To compare the agents' performances we employ a two steps process. In the first step, we perform an *One-way Analysis of Variance* (1-way ANOVA [Glantz and Slinker 2001]) on the raw data of each round, to find out *whether* there are significant differences among the agents' performances. In the second step, we consider the rounds where ANOVA found significant differences, and we apply appropriate *Multiple Comparison Procedures* (also known as *post hoc tests* [Glantz and Slinker 2001]) to find *which* agents perform significantly different from the others. These statistical procedures require that the populations, from which the scores are extracted, are normally distributed and that they have the same variance. We applied the *Kolmogorov-Smirnov* test to check whether populations have normal distributions, and the *Levene* test to check whether populations have same variance. As we expected none of the above requirements is met. A further analysis showed that the tails of the distributions are asymmetric. This is due to the scoring policy employed in the competition: agents' scores have an upper bound but there is no lower bound to agents' loss [Wellman et al. 2001]. Nonetheless, since the *kurtosis* is positive in both the populations we can still apply ANOVA which is more conservative when the *kurtosis* test is positive [Glantz and Slinker 2001]. In addition, although the populations distributions are skewed, the tests are still robust given that the sample contains enough cases.

The Qualifying Round. Table I shows the results of ANOVA applied to the score from the qualifying round; the p-value is nearly zero, therefore in the qualifying round there are agents which performs significantly different from others. We apply a set of *post hoc tests* to find the groups of agents with similar performance.

First, we apply the Scheffé test on the qualifying round scores; results are depicted in Table II(a). The agent position in the table reflects the descending ordering of the agents' scores in the round (e.g., SouthamptonTAC was the first in the round while Retsina was the fifth). In the triangular table: a gray cell means that the agent on the corresponding row has about the same performances as the agent on the corresponding column (i.e., the p-value is less than 0.01); a white cell means that the performance of the agent on the row is worst than that of agent on the column; a light gray cell means that the result is only marginally significant (i.e., the p-value is between 0.01 and 0.10). Note that the agent in position 19, LongAgent, is an outlier. The agent entered the competition later and remained only for a few games, the results for this agent are not significant, because it played too few games. Scheffé is the most conservative test, in fact the gray area is very wide. The result depicted in Table II(a) shows a wide grey area which includes the first 17 agents, and a wide white area which includes the remaining agents. This

suggests that the first 17 agents, have similar performances, while the remaining agents below position 18 perform significantly worse than the first seventeen agents. This finding is confirmed by the Scheffé homogeneity test depicted in Table V(a) where each column labeled with an "s" represents a cluster of agents with similar performance; at the bottom line of the table there is the p-value, the highest the p-value the more homogeneous the group is. The test finds ten clusters. As can be noted there is a clear cut between the clusters 9 and 10 which include the top 17 agents, and the other eight groups which include the remaining 12 agents.

Second, we apply the Tukey test, see Table II(b). Tukey is less conservative than Scheffé, in fact the gray area is smaller. Even this test suggests that the top seventeen agents have better performances. In particular, the top seventeen agents may be grouped in three homogeneous non-disjoint subgroups: the first subgroup includes the top six agents which are included in a small uniform grey area; the second subgroup includes agents from 2 to 14, in this case the area is less uniform since the performance of some agents is *marginally* different than that of others (e.g., see agents 2 and 14); the third subgroup includes the agents from 7 to 17 which define an almost uniform grey area (e.g., agents 7 and 17 perform significantly different). Note however, that the performance differences among these three subgroups are smaller than the difference between the top seventeen agents and the bottom twelve agents. The same conclusions can be drawn from the Tukey homogeneity test (Table V(a)); here the clusters of agents are represented by the columns labeled with a "T." As can be noted, Tukey finds more clusters (12 instead of 10) but basically evidences the same partition of the 17 top performing agents.

Finally, we apply the SNK homogeneity test; the results are depicted in Table V(a), where the columns labeled with an "S" represent clusters of agents with similar performances. SNK is less conservative than the previous tests. In fact it is the only test among those we applied which can isolate a group consisting of *one* top scoring agent ($SouthamptonTAC$), see Table V(a), column 13. Apart from this, also SNK separates the top seventeen agents from the others.

The Seeding Round. Table III shows the results of ANOVA applied to the raw data from the seeding round. As in the previous case, since the p-value is nearly zero, we can claim that there are some significant differences among agents' performances in this round. First we apply the Scheffé test. The results are depicted in Table IV(a); the agents positions in the table correspond to the descending order of the agents' scores in the round (e.g., whitebear was the second in the round). Note that three agents (Attac-2000, dummy_buyer, and bang) were added to the sixteen selected from the qualifying round for experimental purposes. The results show that the first thirteen agents have similar performances, as demonstrated by the wide grey area; while the agents below position 14 perform significantly worst than the first thirteen agents. Among the top scoring agents, we can find two more homogeneous non-disjoint subgroups: a group including the top eleven agents, and the group including the agents from position 4 to 13. The Scheffé homogeneity test in Table V(b) on the columns labeled with an "s", confirms these conclusions: the groups including the top thirteen agents are wide, and there is a clear-cut division between the top thirteen agents and the other agents. The Tukey test (Table IV(b)) also points out this separation between the top thirteen agents with better performances and the

remaining agents. Within the first thirteen agents we can isolate three homogeneous non-disjoint groups: the first includes the top eight agents, the second includes places agents from 4 to 11, and the third includes agents from 6 to 13. Note again that the differences in performance among these three subgroups are smaller than the difference between the top thirteen agents and the remaining agents. This results is confirmed by the Tukey homogeneity test, see the columns labeled with "T" in Table V(b). Because SNK is less conservative, it finds smaller groups of agents (see the columns labeled with a "T" in Table V(b)). However, even SNK highlights the performance difference between the top thirteen agents.

The Semifinals and the Final. The TAC01 final rounds were held October 14^{th} in Tampa, Florida. The eight agents which participated to the final round, from the top scoring agent to the least scoring, were: livingagents, ATTac, whitebear, Urlaub01, Retsina, SouthamptonTAC, CaiserSose, and TacsMan. SouthamptonTAC crashed in one game during the final. Due to the limited number of games played, 35, most statistical procedures on these data are not reliable [Glantz and Slinker 2001]. Applying ANOVA[1] on the data comprising *all the three rounds* (i.e., semi-finals *and* final) we find a marginal difference among the top scoring seven agents and TacsMan. The application of the usual post hoc procedures results in few large clusters which include the majority of the top scoring agents. For instance, Scheffé put all the agents in the same cluster. But the statistical significance of such clusters is poor. These results again support the decision of the organizers not to indicate a winner out of the competition.

4. CONCLUSIONS

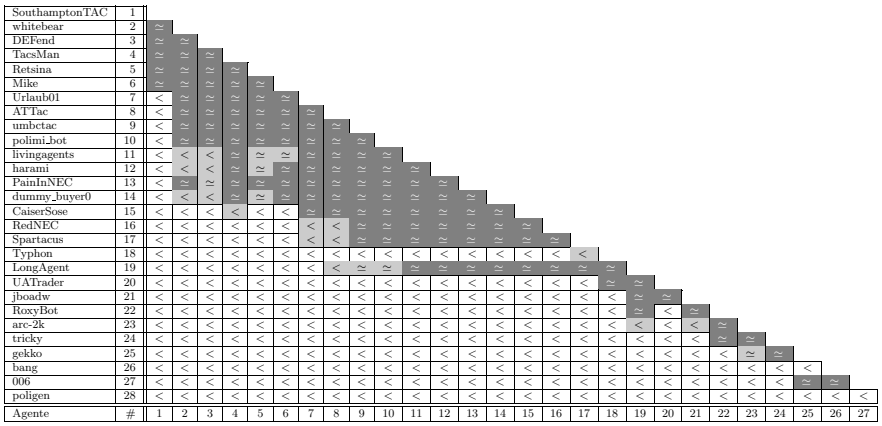
The analysis of the data from TAC01 shows that no agent performed *significantly* better than the others, from a statistical viewpoint. Nevertheless all the tests we applied, even the most conservative ones, find groups of agents that performed significantly better than others. This suggests that in TAC01 there is not a real "winner", as also stated by the organizers. Michael Wellman wrote to the TAC entrants, that the TAC Team would have not declared any "winner", letting observers draw their own conclusions from the raw data. The difficulties in determining a "winner" are mainly due to the difficult challenges that the competition poses and to the high variance of the scores. In particular, high score variance is probably caused both from the competition rules and from errors in the agents' implementations. As a result, TAC01 turns out to be too "noisy" in general and therefore not completely effective in highlighting a single outstanding strategy. Many solutions could reduce this *background noise*, for instance, the agents might play more games, and the environment might be more controlled not allowing modifications to the agents' code. Yet, our analysis suggests that this type of competition can be useful to obtain interesting *qualitative* information by identifying clusters of agents with similar performance.

[1]Note that assumptions required by ANOVA do not hold in these data so the result is not *statistically* reliable.

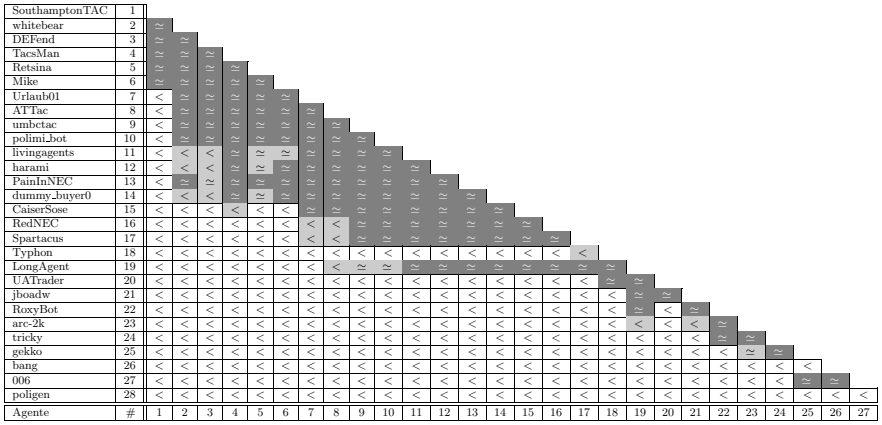
Table I. ANOVA: Qualifying Round.

	SS	df	MS	F	p-value
Between-Groups	13053123999.118	27	483449037.004	195.196	.000
Within-Groups	17035033771.426	6878	2476742.334		
Total	30088157770.545	6905			

Table II. Multiple comparison procedures on Qualifying Round. Pairwise tests. A $\simeq$ in a gray cell means that the row agent has about the same performances of column agent; a $<$ in a white cell means that the row agent has worse performances than column agent; a $\simeq$ or a $<$ in a light gray cell means that the result is only marginally significant (the p-value is between 0.01 and 0.10).



(a) Scheffé test on Qualifying Round.



(b) Tukey test on Qualifying Round.

Table III. ANOVA: Seeding Round.

	SS	df	MS	F	p-value
Between-Groups	4846493486.426	18	269249638.135	208.560	.000
Within-Groups	7449024944.695	5770	1290992.191		
Total	12295518431.121	5788			

Table IV. Multiple comparison procedures on Seeding Round. Pairwise tests. A $\simeq$ in a gray cell means that the row agent has about the same performances of column agent; a $<$ in a white cell means that the row agent has worse performances than column agent; a $\simeq$ or a $<$ in a light gray cell means that the result is only marginally significant (the p-value is between 0.01 and 0.10).

SouthamptonTAC	1																	
whitebear	2	$\simeq$																
Urlaub01	3	$\simeq$	$\simeq$															
livingagents	4	$\simeq$	$\simeq$	$\simeq$														
TacsMan	5	$\simeq$	$\simeq$	$\simeq$	$\simeq$													
CaiserSose	6	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$												
polimi_bot	7	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$											
umbctac	8	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$										
RoxyBot	9	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$									
ATTac	10	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$								
Retsina	11	$<$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$							
PainInNEC	12	$<$	$<$	$<$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$						
ATTac2000	13	$<$	$<$	$<$	$<$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$					
harami	14	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$\simeq$				
dummy_buyer0	15	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$\simeq$			
jboadw	16	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$\simeq$		
006	17	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	
bang	18	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$\simeq$	$\simeq$
arc-2k	19	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$
Agente	#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

(a) Scheffé test on Seeding Round.

SouthamptonTAC	1																	
whitebear	2	$\simeq$																
Urlaub01	3	$\simeq$	$\simeq$															
livingagents	4	$\simeq$	$\simeq$	$\simeq$														
TacsMan	5	$\simeq$	$\simeq$	$\simeq$	$\simeq$													
CaiserSose	6	$<$	$\simeq$	$\simeq$	$\simeq$	$\simeq$												
polimi_bot	7	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$											
umbctac	8	$<$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$										
RoxyBot	9	$<$	$<$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$									
ATTac	10	$<$	$<$	$<$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$								
Retsina	11	$<$	$<$	$<$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$							
PainInNEC	12	$<$	$<$	$<$	$<$	$<$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$	$\simeq$						
ATTac2000	13	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$\simeq$	$\simeq$	$\simeq$	$\simeq$					
harami	14	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$				
dummy_buyer0	15	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$			
jboadw	16	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$		
006	17	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	
bang	18	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$\simeq$
arc-2k	19	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$	$<$
Agente	#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

(b) Tukey test on Seeding Round.

Table V. Multiple comparison procedures on Seeding Round. Homogeneous tests. Each column shows a homogeneous group for each test (s Scheffé, T Tukey, S SNK). The highest is the p-value in the last line the more homogeneous is the group.

	Homogeneous groups																																		
	1			2			3			4			5			6			7			8			9			10			11		12		13
Agent	s	T	S	s	T	S	s	T	S	s	T	S	s	T	S	s	T	S	s	T	S	s	T	S	s	T	S	s	T	S	T	S	T	S	S
SouthamptonTAC																												1					1		1
DEFend																									3			3			3		3	3	
whitebear																									2			2			2		2	2	
Retsina																									5			5			5	5	5	5	
Mike																									6			6	6		6	6	6	6	
TacsMan																									4			4	4		4	4	4	4	
Urlaub01																									7	7		7	7	7	7	7	7	7	
ATTac																									8	8		8	8	8	8	8		8	
polimi_bot																						10			10	10	10	10	10	10	10	10		10	
umbctac																						9			9	9	9	9	9	9	9	9		9	
PainInNEC																						13			13	13	13	13	13	13	13	13		13	
harami																			12			12			12	12	12	12	12	12	12	12			
dummy_buyer0																			14			14			14	14	14	14	14	14	14	14			
livingagents																			11			11			11	11	11	11	11	11	11	11			
CaiserSose																			15			15			15	15	15		15	15					
RedNEC																			16			16	16		16	16	16								
Spartacus																			17			17	17		17	17	17								
Typhon																18			18	18		18	18	18											
LongAgent																19			19	19	19			19											
UATrader													20			20	20			20	20														
jboadw										21			21	21		21	21	21																	
RoxyBot										22	22		22	22	22			22																	
arc-2k							23	23		23	23	23	23	23	23																				
tricky				24			24	24		24	24	24																							
gekko				25	25		25	25	25	25																									
006				27	27	27	27																												
bang	26			26	26	26																													
poligen	28	28	28																																
p-value	.34	1.0	1.0	.26	.07	.25	.32	.35	1.0	.07	.82	.31	.07	.21	.17	.12	.36	.07	.08	.35	.08	.08	.06	.20	.66	.11	.17	.17	.06	.13	.10	.06	.06	.07	1.0

(a) Homogeneous tests on Qualifying Round.

	Homogeneous groups																																	
	1			2			3			4			5			6			7			8			9			10			11		12	13
Agent	s	T	S	s	T	S	s	T	S	s	T	S	s	T	S	s	T	S	s	T	S	s	T	S	s	T	S	s	T	S	T	S	S	S
SouthamptonTAC																												1			1			1
whitebear																									2			2	2		2			2
Urlaub01																									3	3		3	3		3		3	3
livingagents																						4	4		4	4		4	4		4	4	4	4
TacsMan																			5			5	5		5	5		5	5	5	5	5	5	5
polimi_bot																			7	7		7	7		7	7	7	7	7	7	7	7	7	
umbctac																			8	8		8	8		8	8	8	8	8	8		8	8	
CaiserSose																			6	6		6	6	6	6	6	6	6	6	6		6	6	
RoxyBot																	9		9	9		9	9	9	9	9	9	9		9		9		
ATTac																	10		10	10		10	10	10	10		10	10		10				
Retsina																	11		11	11		11	11	11	11		11							
PainInNEC																	12		12	12	12	12		12										
ATTac2000																13	13		13		13													
harami													14	14		14		14																
dummy_buyer0										15	15		15		15																			
jboadw							16	16		16		16																						
bang				18	18		18		18																									
006				17	17	17																												
arc-2k	19	19	19																															
p-value	1.0	1.0	1.0	1.0	.64	1.0	.30	1.0	1.0	.76	1.0	1.0	.42	1.0	1.0	.41	.13	1.0	.05	.11	.27	.22	.06	.05	.14	.08	.35	.07	.06	.08	.06	.09	.07	.10

(b) Homogeneous tests on Seeding Round.

REFERENCES

Glantz, S. A. and Slinker, B. K. 2001. *Primer of Applied Regression & Analysis of Variance*, 2nd ed. McGraw Hill, New York.

Strada, A. 2002. Autonomous agents for electronic auctions. M.S. thesis, Politecnico di Milano. Dipartimento di Elettronica e Informazione. (In Italian).

Wellman, M. P., Greenwald, A., Stone, P., and Wurman, P. R. 2001. The trading agent competition 2001. `http://auction2.eecs.umich.edu`.

Wellman, M. P., Greenwald, A., Stone, P., and Wurman, P. R. 2002. The 2001 trading agent competition. Submitted to the Fourteenth Innovative Applications of Artificial Intelligence Conference(IAAI-2002) available at `http://www.cs.cmu.edu/~pstone/`.

Wurman, P. R. 1998. The Michigan Internet AuctionBot: A configurable auction server for human and software agents. In *Proceedings of 2nd International Conference on Autonomous Agents*. Minneapolis, 301–308.