A DATA DRIVEN MACHINE LEARNING APPROACH TO DISCOVERING RULES OF PRICE BEHAVIOR IN A FINANCIAL MARKET SIMULATION

Roger M. Stein
Department of Information Systems
Leonard N. Stern School of Business
New York University
44 West 4th Street, Suite 9-181
New York, NY 10012-1126
(212) 998-0800
fax: (212) 995-4228
rstein@stern.nyu.edu

Moody's Investors Service
99 Church Street
New York, NY 10007
(212) 553-4928

August 1997

Working Paper Series
Stern #IS-97-17

Abstract

The field of agent-based simulation of financial markets has grown considerably in the last decade. However, the interpretation of simulation results has received far less attention. Typically, the results of a large number of simulations are reduced to one or two summary statistics, such as sample moments. While such summarization is useful, it overlooks a vast amount of additional information that might be gleaned by examining patterns of behavior that emerge at lower levels. In this paper we propose an approach to interpreting simulation results that involves the use of so-called data mining techniques to identify the rules of behavior that govern an underlying system. We demonstrate the approach by using data from a single run of an order market simulation to derive rules about the behavior of prices in that simulation.

1. Introduction*

Recent advances in micro-computer technology and programming languages have made agent-based simulation techniques a widely accessible approach to testing hypothesis about complex systems. This activity has been particularly notable in research in the area of financial markets (e.g.: Arthur. 1994; Boehme, 1994; Weber, 1994; Schwartz and Weber, 1997). Simulation has proven itself to be a valuable tool for understanding systems in which many agents, governed by simple rules, interact to create complex, and often locally unpredictable, behavior.

Twenty-five years ago, for example, Cohen, James and March (1972) described the behavior of autonomous interacting agents:

> Though the specifications are quite simple, their interaction is extremely complex so that investigation of the probable behavior of the system fully characterized by [the model] and previous specifications requires computer simulation. (p. 16)

While they concede that "No real system can be fully characterized [by simulation]," (p. 16), they support the method. In a more recent survey, Starbuck (1983) offers that:

> ...[S]imulation offers deductive capabilities that, in principle, can extend well beyond those of algebraic analyses. Because computers can accommodate very complex assumptions about multitude variables, simulators...can discover the consequences of many, nonlinear, discontinuous interacting assumptions that no one knows how to analyze algebraically. Simulators can make assumptions they believe to be realistic, even if the assumptions are not mathematically tractable...Although computer simulation is no panacea, its significant capabilities make it the only effective methodology for some research tasks, and the best methodology for others. (p. 156 - 159)

But the method is not without its drawbacks. Some researchers object to simulation methods on the grounds that they result from arbitrary assumptions about an environment and lead to unclear relationships between inputs and outputs. Starbuck (1983), for example, warns that:

> Simulators have to specify activity sequences and sufficient assumptions even when they lack information about them...Very large models are too large to validate in detail... Many simulations are hard for their creators to understand...Relations that are too complex to analyze algebraically tend to remain stubbornly incomprehensible by alternative means (p. 156-7)

While the concerns about prior assumptions are not without merit, they are also not unique to simulation approaches. With respect to the validity of assumptions, as Clemons and Weber (1996) point out, most statistical methods require researchers to make (sometimes questionable) assumptions about distributional properties of errors, model structure, etc., a fact which is often ignored in the literature. While simulation does rely on assumptions and researcher interpretation, these concerns are common to many research methods and should not be used to rule out simulation *a priori*.

However, with respect to the concerns about interpretation of outputs, counter-arguments are less clear cut. Antagonists often complain that simulations generate so much data that simulation researchers must resort to "reading tea leaves" in order to make sense of it. In fact, simulators often reduce large numbers of simulations to one or two summary statistics, usually means and variances (or their non-parameteric counterparts). While such summarization is useful, it overlooks a vast amount of potentially useful information that could be used to draw conclusions about the deeper structure of the process being simulated.

In this paper we propose a method of data-mining to aid in the interpretation of simulation output. The method is based on the "evolution" and refinement of rules that predict well the price movements within the simulation. We achieve this through the use of a genetic algorithm designed specifically for this purpose, however any robust rule-discovery algorithm (ex: CART, CHAID, etc.) could be used. To demonstrate the viability of this approach, we apply it to a small sample of data obtained from a single run of a simulation designed to emulate the behavior of agents in an order-driven financial market.

---

* The author wishes to thank Bruce Weber and Vasant Dhar of the Department of Information Systems at NYU for useful discussions. All errors are, however, those of the author.

The remainder of this paper proceeds as follows. Section 2 describes the rule representation we use. Section 3 discusses the a method, based on the use of genetic algorithms, for discovering good rules. Section 4 discusses the application of this technique to the simulation output and presents a sample of the results. Section 5 suggests alternative applications of the technique to simulation interpretation and simulation refinement. Section 6 presents concluding comments.

2. Defining what is meant by "good" rules

For the purposes of this paper we are interested in determining whether certain patterns in price movement persist in the simulation data. On way of doing this is to determine if systematic trading rules can be discovered for predicting the likely price of an asset one trade into the future.

Thus, for example, given the following data, t:

TIME	BID	ASK	BIDASK	HI	LO	LAST	LastChg
0.033	24.625	25.25	0.625	25.375	24.625	24.625	-0.005
			.				
			.				
			.				
0.062	25	25.25	0.250	25.375	24.625	25	0.015

we wish to predict the change in price at t+1.

We consider patterns of the form:

```
(0.125 <= BIDASK <=  0.75) & (0.0125 <= LASTCHG <=  0.0250)
```

which can be interpreted as follows:

- the BIDASK spread is between 0.125 and 0.75

  (AND)

- the LAST CHANGE in traded price is between 0.0125 and 0.0250.

We can see that the first data record would not match the pattern (LastChg is out of the condition's permitted range), but the second data record would.

By examining the data, we can determine whether the population of future price changes is statistically different for those data vectors that match the pattern than it would be for the general population.

Good patterns are those that make a difference in the distribution, and bad patterns are ones that do not significantly filter the population. Patterns can contain any number of conditions.

A rule, therefore, becomes a good pattern and the change it implies in the distribution of price outcomes. For example, a rule using the above pattern might be:

IF

- the BIDASK spread is between 0.125 and 0.75

  (AND)

- the LAST CHANGE in traded price is between 0.0125 and 0.0250.

THEN the next trade is likely to increase the price.

3. The rule discovery approach: Genetic rule refinement

We use genetic algorithms (GAs) to refine rules the price behavior in the simulated market. GAs have their basis in the biological metaphor of survival of the fittest. GAs have been found to be useful for finding good solutions for a wide variety of optimization problems, including classes of problems that were previously computationally

prohibitive (Holland, 1970/1992; Davis, 1991; Goldberg, 1989). In this case the optimization problem is one of finding high quality rules based on the simulation data. The quality of a rule is judged by the level of statistical significance in the difference between the sub-population or trades selected by the rule and the population at large. The approach we use is similar to an approach used by Packard (1990) with several important enhancements.

A genetic algorithm attempts to solve a problem by creating a range (called a population) of possible solutions. These take the form of strings describing a particular pattern of data. Each member of the population (an individual pattern) is then ranked in terms of its fitness. Fitness is an assessment of how well a particular individual rule predicts price movements.

Individual rules are then matched randomly with other rules in the population such in a way that those with higher fitness are more likely to be selected. The results of this "mating" form the offspring that make up the population of the next generation and the process can be repeated with this new population.

During the mating process two operations take place: mutation and crossover. Mutation involves changing the value of one portion of a rule. Crossover involves the exchange of portions of a rule between two individuals.

By mutating and crossing over, the GA is, in effect, experimenting with new rules solutions while preserving potentially valuable interim results (Holland, 1970/1991, Davis, 1991; Goldberg, 1989;) If an experiment (crossover or mutation) fails (that is, produces a relatively poor predicting rule), then the rule will, in all likelihood, be dropped from the population within a few generations due to its inferior fitness.

On the other hand, if the experiment is successful, then these new interim results can be passed on to the future generations for further refinement. Thus the more promising areas of a solution space are explored, and lower payoff areas are examined in a more cursory manner. The genetic algorithm paradigm allows the search of potentially huge problem spaces in a parallel and efficient manner (Go
Goldberg, 1989).

4. A sample application to data from a simulated order driven market

We applied the methodology described in Section 3 to a data set generated by the ISMARTS (Weber, 1996) simulation environment. The particular simulation we used was of an order driven market in which a percentage of the traders informed.

We ran the algorithms several times while adjusting parameters to obtain a wider variety of rules. Examples of the rules are shown in Table 1, below. (Data values are normalized for presentation).

Table 1: Examples of discovered rules

	Num. of trades matching	*English rule*	*Raw rule*	*Direction (mean % chg)*	*Confindence (SD % chg)*
1	40	**IF** • **Ask price is moderately high** • **the Bid/Ask spread is average for the price level** • **the last price traded was a little better than average** **THEN it is pretty likely that *the next trade will bring the price down a little***	(-0.21 <= ASK <= 1.30) & (-0.59 <= BAPCT <= -0.02) & (0.87 <= LAST <= 1.59)	-0.155	0.382
2	11	**IF** • **the Bid/Ask spread is narrow for the price level** • **the last trade was somewhat higher than the Hi/Lo midpoint** **THEN it is highly likely that *the next trade will not move the price much***	(-2.90 <= BAPCT <= -0.10) & (0.88 <= LSTHL <= 1.29)	-0.021	0.187
3	70	**IF** • **the last price traded was close to the bid** **THEN it is more likely that the next trade will *bring the price down* somewhat**	(-3.60<= LSTBA <= -0.65)	-0.352	0.741
4	28	**IF** • **the last price traded was closer to the ask** • **the last trade brought the price down** **THEN it is more likely that the next trade will *bring the price up* somewhat**	(0.89<= LSTBA <= 3.99) & (-2.27 <= LastChg <= 0.19)	0.261	0.688
7	120	**IF** • **the Bid/Ask spread is not extremely wide for the price level** • **the last trade changed the price positively** **THEN volatility will *increase* slightly on the next trade**	(-1.63<= BAPCT <= 1.47) & (-0.65 <= LastChg <= 5.70)	0.033	1.049

5. Alternative Applications

The purpose of this paper is not to debate the degree to which the rules in Table 1 imply a deeper structure in this particular simulation. Rather it is to demonstrate that data-mining techniques, like the one described in this paper, can be used to examine the output of simulation research.

Having said this, how might we use these discovered rules to understand simulations better?

Firstly, a the rules themselves give insight into the dynamics of the market that is being simulated. For example, although Rule 4 gives evidence of the efficient market hypothesis (EMH), Rule 3 gives some evidence against the strong form of the EMH and for momentum trading. In addition, rules discovered on one run of a simulation can be compared to another run of the simulation, using the same assumptions, to determine how robust the rules are and thereby determine whether the patterns discovered are actually characteristic of the trading environment or whether they are transient effects. Furthermore, discovered rules from one set of simulation assumptions can be compared to discovered rules from a simulation with different assumptions to determine the impact of changing the assumptions.

An alternative use could involve determining the degree to which known rules could be recovered from simulation data. For example, if a particular agent is programmatically an "informed" trader, can this be discovered via the data mining technique? If so, the simulation environment could be used as a test-bed for developing fraud detection systems to be applied to real markets. In addition, finding such known patterns would give confidence in the techniques.

Finally, data-mining techniques can be used to determine how realistic a particular simulation actually is. If a particular set of patterns, say, in price movements, is discovered in a simulation, one would expect to see similar patterns in the actual process being simulated. Real market data could be analyzed using the same problem formulation as was used for analyzing the simulation. If similar patterns were discovered, this would lend support to the model.

If, in contrast, patterns similar to the simulation patterns were not found, this could point to deficiencies in the simulation model. To the extent that more prevalent patterns were found in the real data, these provide insight into lapses in the model design.

6. Conclusions

Simulation researchers are often plagued with the problem having too much data on their simulations. It is ironic that many researchers in social sciences criticize simulation research for this while simultaneously bemoaning their own lack of observable field data.

We have shown how an application of data mining techniques to simulation data can partially combat the data volume problem and yield interesting insight into the dynamics of the simulated process.

In fairness some of the rules discovered in our example were not terribly surprising. For example, Rule 2 basically says that in a tight market where there is no negative "news" and no terribly positive news, the price won't move very much. But even this can be useful if the objective of the data mining were validation of a specific simulation.

Other sample rules were more interesting in that they seemed to support the idea of momentum trading in the simulated markets and gave some insight into the degree to which this impacts market prices (as shown by the expected change in prices). If momentum were a component of the simulation design (it was), then this shows that the data mining technique is able to recover known dynamics. If the behavior were not part of the system, such a rule might show the emergence of such behavior in a benign market.

In the end, data mining techniques offer an alternative method for simulation analysis that is uniquely suited to the problem of discovering interesting patterns among large data sets. These techniques can be useful compliments to the more traditional population level statistics used to summarize simulation data.

References

1. Arthur, Brian, J. Holland, and R. Palmer, "Adaptive Behavior in the Stock Market," *Santa Fe Working Paper Series*, 1994.

2. Boehme, Frank, "Adaptive Coordination Mechanisms in the Economic Modeling of Money Markets," in *Many Agent Simulation and Artificial Life*, E. Hillebrand and J. Stender, Eds., IOS Press, Amsterdam, 1994.

3. Clemons, Eric, and B. Weber, "Competition Between Stock Exchanges and New Trading Systems: A Demonstration of Adverse Selection and an Exchange Response, Working Paper, *Information Systems Working Paper Series*, Stern School of Business, 1996.

4. Cohen, Michael D., J. G. March, and J. P. Olsen, "A Garbage Can Model of Organizational Choice," *Administrative Science Quarterly*, Vol. 17, 1972.

5. Goldberg, David. E., *Genetic Algorithms in Search, Optimization, & Machine Learning*, Addison-Wesley, 1989.

6. Holland, John. H., *Adaptation in Natural and Artificial Systems*, MIT Press, 1970 / 1995.

7. Packard, N. H., "A Genetic Learning Algorithm for the Analysis of Complex Data," *Complex Systems*, Vol. 4. No. 5, 1990.

8. Starbuck, William H., "Computer Simulation of Human Behavior," *Behavioral Science*, Vol. 28, 1983.

9. Schwartz, Robert., and Weber B., "Combining Quote-Driven and Order-Driven Trading Systems, *HICSS-30*, 1997.

10. Weber, Bruce W., "Assessing Alternative Market Structures Using Simulation Modeling," in *Global Equity Markets: Technological Competitive, and Regulatory Challenges*, R. A. Schwartz (ed.), Irwin Professional, 1994.