Validating Simulations with Historical Data: The State Similarity Measure[1]

Joint Statistical Meetings,
Vancouver, B.C.,
July 31–August 5, 2010

Robert Marks,
Economics,
UNSW, Sydney

bobm@agsm.edu.au

Abstract:
Static estimation treats variation in the dependent data as noise, or error. In simulations using agent-based models, however — especially with dynamic responses — such variation in the simulated output may well possess valuable information from the simulation. This paper will explore previous methods of estimating simulation models (such as indirect inference, the method of simulated moments, and estimation of an auxiliary model), before examining the simulated output from an early agent-based model in marketing (Marks Midgley & Cooper 1995), and asking whether these methods or others allow the modeller to conclude with some degree of confidence that the simulated output is generated by essentially the same process that generated the historical output by measuring the degree of similarity between two sets of time-series. We introduce a new measure, the State Similarity Measure (SSM), to measure the distance beween two sets of time-series that embody dynamic responses. Given the degrees of freedom in simulation models, the SSM measure can only increase the confidence in which simulation is used and accepted.

1. Introducing the State Similarity Measure

The problem we face is to decide how well an agent-based model we have built in order to examine the rivalrous dance of oligopolistic competition (necessarily with dynamic responses) is performing, compared to historical data of the market we model. This issue is not new. Agent-based models of financial markets — exchange rate markets and markets for financial assets — have large amounts of historical data to compare their models with. Amongst other methods, they have used the "method of simulated moments" in their comparisons. We describe two papers which attempt to use this method below.

But in order to undertake this method, we need large amounts of data, which are not available for the market we have modelled — brand competition among ground coffee brands in a supermarket chain. This is one issue: given one historical realisation, and given any simulated moments we wish to derive from our model, what is the degree of confidence that the model is capturing the essential actions and reactions of the

---

1. An earlier version of this paper was presented at the workshop on Advances in Agent-Based Computational Economics, ADACE 2010, July 5–7, 2010, Center for Interdisciplinary Research (ZiF), Bielefeld University, Germany.

historical market?

Second: just what are the appropriate moments (summary statistics) to calculate for both historical data (when we have enough) and the simulations output? For the exchange markets that others (Franke 2009, Winker et al. 2007, Chen et al. 2012) have modelled, there has been much effort in analysing historical data: it is from these analyses that we have derived such characteristics of these markets as "fat-tailed distributions" and "volatility clustering"; simulated markets are a way to develop sufficient conditions for the generation of such moments (LeBaron 2006).

In this paper we, first, discuss the issue of validation of simulation models using historical data; second, the simulated method of moments, as applied to exchange markets; third, possible moments to be used to validate models of oligopolistic competition and introduce the State Similarity Measure; fourth, comparing historical data using the SSM; fifth, validating simulation outputs against historical data using the SSM; conclusion; appendices.

2. Validation of Agent-Based Models

The author has been involved in a research program to elucidate the issues involved with building agent-based models of oligopolistic competition. In particular, he and colleagues (Marks 1992, Marks Midgley & Cooper 1995, Midgley Marks & Cooper 1997, Marks 1998, Midgley Marks & Kunchamwar 2007, Marks 2007) have been involved with comparing their models of market interactions against historical records of such markets.

Such comparisons could be used for (at least) two reasons: first, to choose better parameters (or models), and, second, to measure how realistic the behaviour of a model built with parameters chosen for another reason (in this case to maximize weekly profits by brand over a period) is, compared to the historical data.

To achieve this, we have coined the phrase "model assurance" (Midgley et al. 2007) to include the twofold process of model verification (assuring ourselves that our computer model is doing what our conceptual model does — that there are no bugs in the code) and model validation (assuring ourselves that the output from the computer model is "realistic," to a degree). Of course, by their nature, models — both conceptual and computer implementations — are abstractions from reality, so some judgement must be exercised in our validation stage.

We need to distinguish choosing the model's variables (choosing the model, in effect) from asking whether the model (with variables chosen using some independent criterion) exhibits behaviour that is "close" (in some sense) to the historically observed outcomes. This distinction is important, since "validation" can sometimes mean choosing model variables in order to minimize a measure of divergence between the model's behaviour and the historical outcomes.

This distinction can only matter when the researchers have another means to determine the model variables. Midgley et al. (1997) used a learning algorithm (the Genetic Algorithm, or GA) to determine the model variables when maximizing weekly profits was the aim of the modelling exercise. Others (for instance, Hansen & Heckman 1996, and Gilli & Winker 2002) derive ("estimate"[2]) variable values indirectly: Gilli and Winker estimate the values of an agent-based model of a foreign exchange market,

and Hansen and Heckman argue that calibration of the micro foundations of macro models (of, for instance, the real business cycle) can lead to more robust models.

Another way of considering this is that we are interested techniques to test the hypothesis that the output from our agent-based model is *sufficiently* similar to historical data from the exchange markets we are interested in modelling, in order to understand the historical interactions among the players, rather than in applying techniques of variable estimation or model calibration in order to choose models or parameters.[3] We shall derive measures of similarity between model and history in order to aid the decision of what is sufficient in closeness.

In order to do this, following Fagiolo & Roventini (2012), given a set of initial conditions (including any random-number seeds), we run our model until it converges to some stable output behaviour (i.e., for at least $T > T'$ time steps). Suppose we are interested in a set $S = \{s_1, s_2, \cdots s_{n_m}\}$ of statistics to be computed on the simulated output variables. For any given run, the program will output a value for each statistic. Given the stochastic nature of the process, each run will output a distinct value for the statistics. Therefore, after having produced M independent runs, we have a distribution for each statistic containing M observations, which can be summarized by computing its moments. These moments, however, will depend on the initial conditions (parameters).[4]

Because of our use of simulated evolution (via the GA) with agent profitability as our "fitness", we are not concerned (unlike many other agent-based model researchers) with model selection through use of the simulated moments. Instead, as mentioned above, we are concerned with empirical validation of our models, using the simulated moments derived from the M runs of the model.

Chen et al. (2012) provide a good description of using the method of simulated moments to estimate the model parameters ("the indirect method", Gourieux & Monfort 1997), and discuss the earliest papers to do so, including: Winker & Gilli (2001), Gilli & Winker (2003), Winker et al. (2007), and Franke (2009). They also outline six practical issues in the use of MSM: the dimensionality of the vector of statistics, the statistics included in the vector, the set of initial parameters, the "distance function" (see below), the search algorithm used for global optimization, and the number of runs M, which is the sample size of simulated moments.[5]

---

2. Oreskes et al. (1994, p. 642) describe "calibration" of a model: "The process of tuning the model—that is, the manipulation of the independent variables to obtain a match between the observed and simulated distribution or distributions of a dependent variable or variables—is known as calibration."

3. Miller (1998) argues for testing models to destruction in the space of input parameters in order to derive a sense of the robustness of each model. Although it is important to spell out one's input parameter settings (as Winker et al. 2007 do), we are not here concerned with choosing these.

4. As Fagiolo and Roventini comment, by exploring a sufficiently large number of points in the space of initial conditions, we might get a quite deep descriptive knowledge of the behavior of the model. In a way, that is what our machine-learning algorithm achieves, while searching in the non-linear parameter space for more profitable models. See also Fagiolo et al. (2007).

3. The Method of Simulated Moments

We are interested in comparing the distributions of summary statistics (or "moments") from the model and from the empirical data (or "history"); we leave until later discussion of just what these statistics should be for our model of market interactions, whereas others have used the logarithm of returns in models of exchange.[6] Let L be the greatest lag in forming these statistics; let y_t^h be a vector of historical variables observed in period t, of arbitrary dimension; let historical observations be available over a time span $t = 1 - L, (2007) \ldots 1, \cdots T$. The summary statistics derive from n_m moment functions $m_i(.)$, defined on L-stretches of a variable y_t. Let $z_t \equiv (y_t, y_{t-1}, \cdots, y_{t-L})$. The ith empirical moment is computed as the time average

$$\hat{m}_{Ti} \equiv \frac{1}{T} \sum_{t=1}^{T} (z_t^h), i = 1, \cdots, n_m \tag{1}$$

The index T is the explicit length of the sample period. $\hat{m}$ is only an estimate of of the true unconditional mean of the real-world stochastic process, and the actual sequence $\{y_t^h\}_{t=1-L}^{T}$ is just a single realization of it.

Following Chen et al. (2012), let $\mathbf{X}$ be a set of chosen moments (statistics) derived from the historical data, $\mathbf{X} = (X_1, X_2, \cdots, X_{n_m})$, and let $\mathbf{Y}$ be the equivalent in the synthetic, simulated data, $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_{n_m})$. In addition, let D be a distance metric between $\mathbf{X}$ and $\mathbf{Y}$. The smaller the difference between them, the greater the likelihood that the model generating the synthetic moments is sufficient to describe the historical process. We use this approach below.

4. Which Moment to Use? The State Similarity Measure (SSM).

For researchers using agent-based models to build artifical stock markets, the question of which moments to use is a simple one: the standard measures of variability and clustering, including the volatility. Such researchers are attempting to explain (at least) four stylized facts of financial markets: the absence of autocorrelations, volatility clustering, long memory, and fat-tailed distributions. Frenke (2009) lists six possible moments: the mean and autovariance of the returns and absolute returns, and two moment functions that are related to the Hill estimator of absolute returns. Frenke (2009), Winker et al. (2007), and Gilli & Winker (2003) use the simulated method of moments to (indirectly) estimate the parameters and initial conditions for their models that minimise the distance between the simulated moments and the historical moments.

As mentioned above, our study differs to two ways from this: our parameters are determined to maximise our agents' weekly profits, perhaps greatly exceeding historical profits, and our agents are not buying and selling on an artificial stock market, but are competing to sell imperfect, distinct substitutes in a retail grocery oligopoly market. The first means that we are not interested in estimation of the model parameters. The second means that our moments of interest are quite different from the earlier studies.

---

5. Richiardi (2012) also discusses the methods of estimating the structural parameters of agent-based models.

6. This discussion is adapted from Franke (2009) and Chen et al. (2012).

Figure 1 shows the weekly prices and volumes sold of coffee brands in our supermarket Chain One.

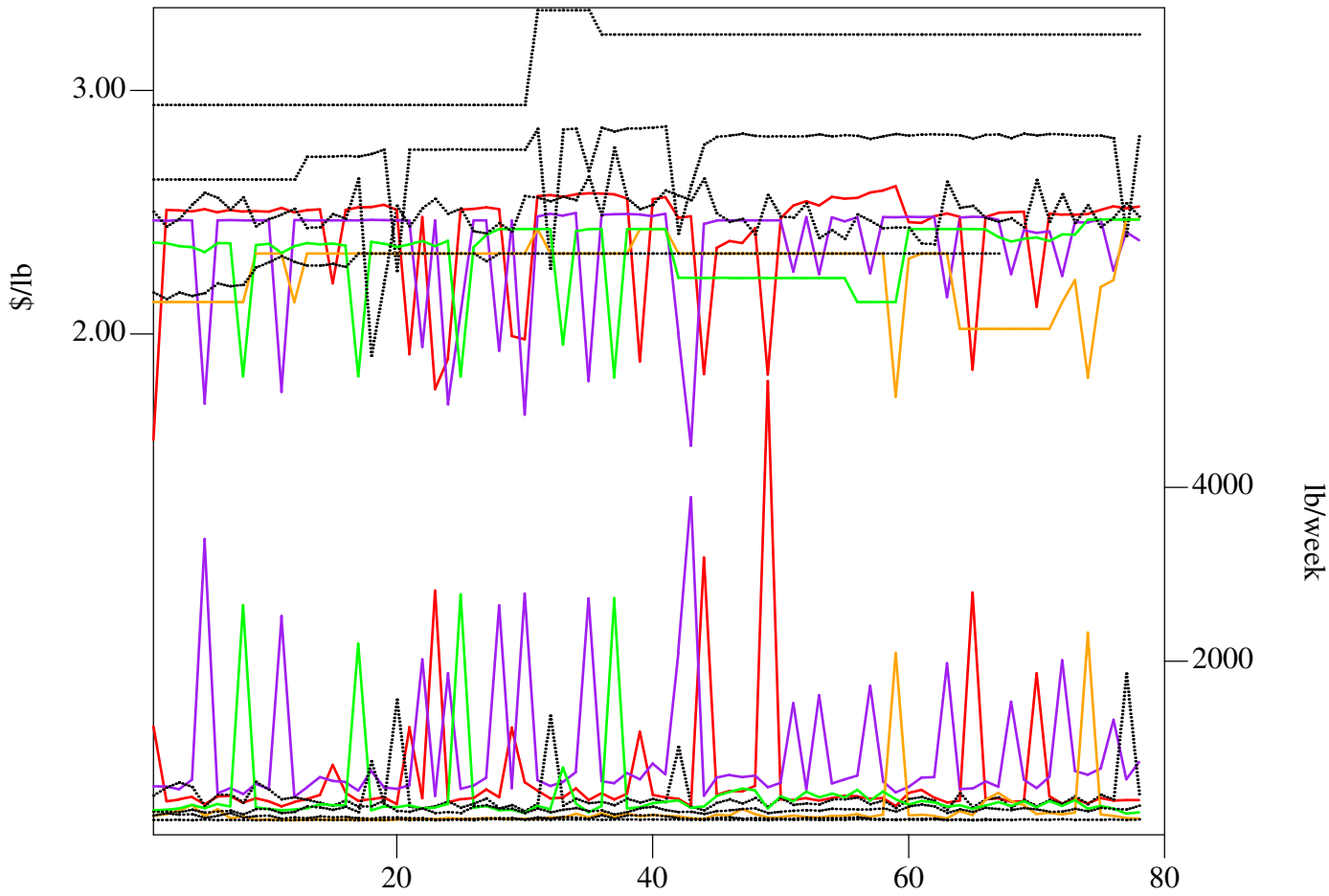

Figure 1: Weekly Sales and Prices (Source: Midgley et al. 1997)

The following stylized facts can be seen in the patterns of competition among the nine brands:

1. Much movement in the prices and volumes of four strategic brands — a rivalrous dance.
2. For these four brands, a high price (and low quantity) is punctuated by a low price (and a high volume).
3. The remaining five brands exhibit stable prices and volumes, by and large. For this reason we are abstracting away from these five brands, and focus solely on the first four.

Also notice that only one brand discounts deeply in any week and no brand discounts deeply two weeks in a row. These are restrictions placed on the brands' marketing actions by the supermarket chain. This is a moderated competition: as well as the two restrictions just noted, the brands' prices (and other marketing actions) are changed together, at midnight on Saturdays, and remain unchanged for the next seven days, which makes modelling this synchonized oligopoly easier.

We believe that these data reveal an strategic interaction among the four (coloured) brand managers, where any manager's action next week is a function of the perceived state of the market this week and perhaps in previous weeks. We simplify the state of the market in any week as the four players' prices this week. We assume that the price $P_{b,w}$ of brand b in week w is a function of the state of the market M_w at week w, where M_w in turn is the product of the weekly prices S_w of all brands over several weeks:[7]

$$P_{b,w} = f_b(M_w) = f_b(S_{w-1} \times S_{w-2} \times S_{w-3} \cdots) \tag{2}$$

Since our data represent a strategic interaction with the four evidently strategic players responding to the previous (and perhaps anticipated) actions of other strategic players, and with the supermarket's moderation of this rivalrous dance, we first consider a window of four weeks when deriving moments from the time-series data: this allows all four brands the possibility of a deep discount in the period, even if such a pattern is not always observed. It does not imply that any one brand has four weeks' memory.

We now face the curse of dimensionality: when is a rival's price change strategically significant (to be responded to) and when is it ignored? Assume that there is a threshold (and assume that it is brand-independent, for the moment). Marks (1998) develops a model of partitioning that uses the data depicted in Figure 1 to explore the revealed perceptions of the brand managers by analysing the historical data. By searching for the dichotomous partition (into "high" and "low") along the price line between each brand's highest and lowest price in the data, and by using the information measure of entropy,[8] he concludes that whether or not a rival changes its price from one week to the next is more significant than by how much the price has changed, or whether the price has changed from "high" to "low" or vice versa.

This suggests that we could coarsen the historical price data into simple dichotomous actions — whether the price changed or not — but it may be that further analysis following on from Marks (1998) would reveal that we were ignoring information that is significant to the brand managers, and so we do not partition quite as coarsely as this here.[9]

Nonetheless, partition we must, lest the high measure of dimensionality when

---

7. Incidently, in our earlier work, we used the GA to search for "better" brand-specific mappings f_b between market state M_w and brand price $P_{b,w}$.

8. Marks (1998) acknowldges that a better metric would be the profits earned as a function of the coarseness or fineness of the players' perceived partitions of prices, and exactly where the partitions occur, but this analysis has not yet been accomplished for these data.

9. Here, the time dimension of the data set is naturally discrete (the historical brands change their prices synchronously, as mentioned above), but other data sets would demand time partitioning as well.

each one-cent-per-pound change in price is strategically significant strangle us in a large number of irrelevant (from the data) states.

One way of building a simulation model is to choose actions (prices) from a restricted set of possibilities, and which are brand-specific. For instance, Midgley et al. (1997) use cluster analysis to determine the set of four (or eight) most frequent actions of each brand in the data.[10] This means, of course, that the simulated data must be restricted in their actions (or prices), unlike the historical data, which suggests that we should partition the historical data before calculating the historical moments to compare against the simulated moments in measuring the performance of the simulation model.

Having generated our simulated time-series, and partitioned our historical time-series, just what moments can we use to compare them? With the fourfold or eightfold partition of the price line, any brand's prices is not a particularly powerful moment (however summarised), but the combination of price responses among the distinct brands becomes more powerful.

With four brands and four possible price intervals, in any week there are $4^4 = 256$ possible states of the market. With a window of four weeks, when each week's state can result in any one of 256 possible combinations next week, and so on, there are $256^4 = 4,294,967,296$ (or 2^{32}) possible four-week states. Remember: implicit in this are each brand's possible reactions to the prices of its three rivals last week, which in turn are responses to the four actions of two weeks previously, which in turn ...

If we coarsen the number of possible actions by halving it to two per brand per week (Marks 1998), say "high" and "low", then the number of possible four-week states falls to $2^{16} = 65,536$. That is, each brand is in one of two possible states per week; with four brands, there are $2^4 = 16$ possibilities per week; with a four-week window, there are $16^4 = 65,536$ possibilities. On the other hand, if we refined the number of possible actions by doubling it to eight per brand per week, then the number of possible four-week states would rise to $2^{64} = 1.8 \times 10^{19}$. Figure 2 shows the time-series of Chain 1 prices (from Figure 1) of the four strategic brands after dichotomous partitioning into the two price ranges, "high" and "low". We believe that the partitioning captures the essence of the strategic interactions among the four rivalrous brands, as suggested by visual comparison of Figures 1 and 2.

With dichotomous (mid-price-range) partitioning and four-week windows, there are 65,536 possible states. With seven sets of time-series,[11] each of 78 weeks, there are $75 \times 7 = 525$ weekly historical observations (after subtracting three weeks initially). That is, the raw historical data provide a sample which equals less than 1 percent of the possible states, whereas, with simulated data, we can generate as many data as we wish.[12] Could we bootstrap (Burns Statistics 2008) the historical data to obtain more?

With sufficient historical data and simulated data, we could derive distributions

---

10. The set could have the same number of elements as the perceived strategic partition, although this identity is not necessary.

11. Corresponding to the seven supermarket chains in the historical data.

12. We shall find, however, that most possible states are never observed, in either the historical data or the simulated output.

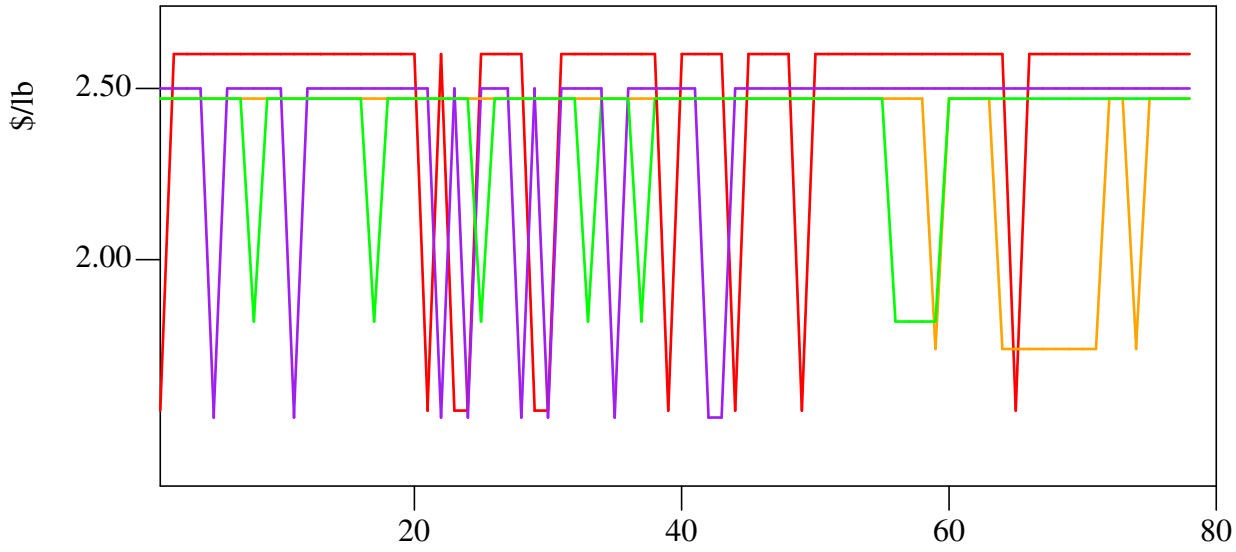

Figure 2: Partitioned Weekly Prices of the Four Chain-One Brands

of the historical and simulated states across the 65,536 possibilities. One possible measure of closeness of fit of the simulated output to the historical data could be defined as the sum of the differences D_{ad}^{sh} between the two time-series:

$$D_{ad}^{sh} = \sum_{i=0}^{i=n-1} |N_i^s - N_i^h|, \tag{3}$$

where N_i^s is the number of state i observed in the simulated data, N_i^h is the number of state i observed in the historical data, and n is the number of possible states, here 65,536. These measures could also be used to compare the behaviour observed in separate historical oligopolistic markets.

This number D_{ad}^{AB} is the distance between two time-series sets A and B. This new moment is called the *State Similarity Measure* (SSM).

5. The Historical Data

The seven historical time-series do not include all of same brands, which means that direct comparison of the strategic behaviours across the seven chains is not always possible. All seven contain Brands 1, 2, and 3, while Chains 1, 2, 3, and 7 contain Brands 1, 2, 3, 4, and 5. Chains 4, 5, and 6 do not include either of Brands 4 or 5. Various of the chains include one or more of another seven brands, although several chains have only four brands, as outlined in Table 1. The 78 weeks of all time-series cover the same historical dates, however.

	Brands											
	1	2	3	4	5	6	7	8	9	10	11	12
Chain 1	✓	✓	✓	✓	✓	✓	✓	✓				
Chain 2	✓	✓	✓	✓	✓		✓	✓		✓		
Chain 3	✓	✓	✓	✓	✓			✓	✓			
Chain 4	✓	✓	✓					✓				✓
Chain 5	✓	✓	✓			✓		✓			✓	✓
Chain 6	✓	✓	✓									✓
Chain 7	✓	✓	✓	✓	✓							✓

Table 1: The Historical Data: The Seven Chains and the Twelve Brands

(Brand 1=Folgers, 2=Maxwell House, 3=Master Blend, 4=Hills Brothers, 5=Chock Full O Nuts, 6=Yuban, 7=Chase & Sanbourne, etc.)

5.1 Comparing History: Four Chains with Four Brands

For this reason, we calculate the frequencies of four-week states of the Chains 1, 2, 3, and 7, using dichotomous, mid-point ("high" versus "low") pricing. The six absolute differences (or SSMs, from equation 3) between these four chains are given in Table 2, as measured by the differences in the frequency of each state being observed in each chain. (See Appendix 1 for the method of deriving these.) Table 2 also gives the absolute distances between each of these four chains and a single realisation of a random pattern of pricing from the four brands, where in any week each brand is equally likely to price "high" or "low".

	Chain 1	Chain 2	Chain 3	Chain 7
Chain 1	0	128	112	110
Chain 2	128	0	132	138
Chain 3	112	132	0	124
Chain 7	110	138	124	0
Random	150	150	150	150

Table 2: SSMs Between Four Chains (with Brands 1, 2, 4, 5)

How can we undestand these numbers? There are 65,536 possible states, but the historical data that we have for each chain only include 75 overlapping four-week windows, or data points. Define C as the maximum number of window states in the data. Here C = 75. At least 65,536 − 75 = 65,461 states will not be observed in the data from any one of our historical chains. (Some states may be observed multiple times.) If each state is observed the same number of times in the data from any two chains (perhaps zero times), then the difference between the two chains is zero for that state. If we observe state *k* occurring twice in one chain's data and zero times in the other's, then the difference between the two chains is two for that state. The numbers in Table 2 are the sums across all possible states of the differences between the historical data of the two chains being compared.

If the two patterns of strategic behaviour are sufficiently distinct, then none of the states oberved in one chain will be seen in the other, and vice versa. This would give a measure of $2 \times C$ (here, $2 \times 75 = 150$), as the maximum difference between two of our 75-week windows. The minimum would be zero: the strategic patterns (as measured by the coarse "high" and "low" partitions) would be identical.

None of the pairs of historical data is particularly close: the differences range from a low of 110 between Chains 1 and 7, to a high of 138 between Chains 2 and 7. We see that Chain 1 is closest to Chain 7, that Chain 2 is closest to Chain 1, that Chain 3 is closest to Chain 1, and that Chain 7 is closest to Chain 1. The differences are not, of course, additive, because of the underlying complexity of the processes and measures.

What of the time-series of random "highs" or "lows" of the four brands in any week? Choosing actions at random is equivalent to choosing $2 \times C = 75$ of 65,536 states at random (given the length of the time-series), or 0.114 percent, at random. We should not be surprised that there is virtually no overlap of states between any of our historical time-series and a time-series of randomly chosen actions for the four brands each week (see Figure 7 below).

If we use the null hypothesis that each of two sets of time-series is random, then we can set 1% and 5% one-sided confidence intervals to the SSM numbers. With four brands and $C = 75$, the largest SSM is 150. Monte Carlo simulation shows that 99% of pairs of sets of four random time series are at least 148 apart.[13] This means that, in Table 2, all six pairs of chains' time-series are significantly non-random, and the null hypothesis is rejected for each pair (except, of course, for the pairwise comparisons with a random realisation).

We show this in Figure 3, which plots the Cumulative Mass Function (CMF) of the Monte Carlo parameter bootstrap simulation against the SSMs of the six pairs. The red lines are the CMF of pairs of sets of random series (4 series, 75 observations) from 100,000 Monte Carlo parameter bootstraps.

Table 3 presents the matches between chains' states as percentages, from Table 2: a complete match is 100 percent, and no matches at all is zero percent.

5.2 Comparing History: Seven Chains with Three Brands

Table 4 shows the 21 differences between all seven chains when we focus on the strategic interactions of the three Brands 1, 2, and 3, using three-week windowing and dichotomous pricing. Table 4 also gives the absolute distances between each of these seven chains and a single realisation of a random pattern of pricing from the three brands, where in any week each brand is equally likely to price "high" or "low". With three brands each with two possible states, there are $2^3 = 8$ possible states of the market in any week. With a window of three weeks (so that each brand has the opportunity to deep discount), there are $8^3 = 512$ possible three-week states.

---

13. This number was determined by a parameter bootstrap simulation of 100,000 replications of pairs of sets of four quasi-random time-series, calculating the SSM between each pair, and examining the distribution to derive the one-sided confidence intervals. The lowest observed SSM of 140 appeared twice, that is, with a frequency of 2/100,000, or 0.002 percent. See the simulated Probability Mass Distribution (PMD) in Figure 7 below.

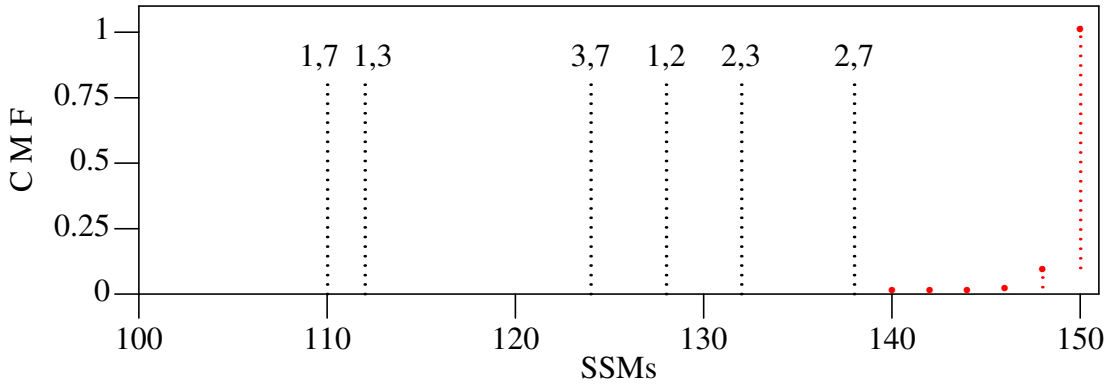

Figure 3: Four Chains, Four Brands, SSMs against Random CMF.

	Chain 1	Chain 2	Chain 3	Chain 7
Chain 1	100	14.67	25.33	26.67
Chain 2	14.67	100	12.0	8.0
Chain 3	25.33	12.0	100	17.33
Chain 7	26.67	8.0	17.33	100
Random	0	0	0	0

Table 3: Percentage Matches Between Four Chains (with Brands 1, 2, 4, 5)

	Chain						
	1	2	3	4	5	6	7
Chain 1	0	70	82	76	102	132*	74
Chain 2	70	0	82	98	90	120†	98
Chain 3	82	82	0	100	96	122†	102
Chain 4	76	98	100	0	80	128*	58
Chain 5	102	90	96	80	0	114	92
Chain 6	132*	120†	122†	128*	114	0	130*
Chain 7	74	98	102	58	92	130*	0
Random	144	136	148	144	140	146	144

Table 4: SSMs Between All Chains (with Brands 1, 2, 3)
(* : cannot reject the null of random at the 5% level)
(† : cannot reject the null of random at the 1% level)

With the three brands and three-week window of Table 4, there are 512 possible states, but the historical data only include $C = 76$ overlapping three-week windows. The greatest distance between any two chains is $2 \times C$ (here $2 \times 76 = 152$). No two pairs are that dissimilar, although the randomly derived pattern is at least 136/154 states different from any of the historical time-series.[14] The closest two chains are Chains 4 and 7, with $152 - 58 = 94$ states in common, or $94/152 = 61.84$ percent. The least similar chain to

any of the others appears to be Chain 6.

If we use the null hypothesis that each of two sets of time-series is random, then we can set 1% and 5% one-sided confidence intervals to the SSM numbers. With three brands and C = 76, the largest SSM is 152. 95% of pairs of sets of three random time-series are at least 122 apart, and 99% of pairs of sets of three random time series are at least 118 apart.[15] This means that, in Table 4, we reject the null hypothesis of random data for all chains but Chain 6, since all SSMs between Chains 1, 2, 3, 4, 5, and 7 are less than 118, so the data are significantly non-random, and the null hypothesis is rejected. For Chain 6, however, we cannot reject the null hypothesis for comparisons with Chains 1, 4 or 7 (5 percent) and with Chains 2 and 3 (1 percent); only with Chain 5 is the null rejected.

We show this in Figure 4, which plots the CMF of the MC parameter bootstrap simulation against the seven greatest SSMs of the pairs. The red lines are the CMF of pairs of sets of random series (3 series, 76 observations) from 100,000 Monte Carlo parameter bootstraps.

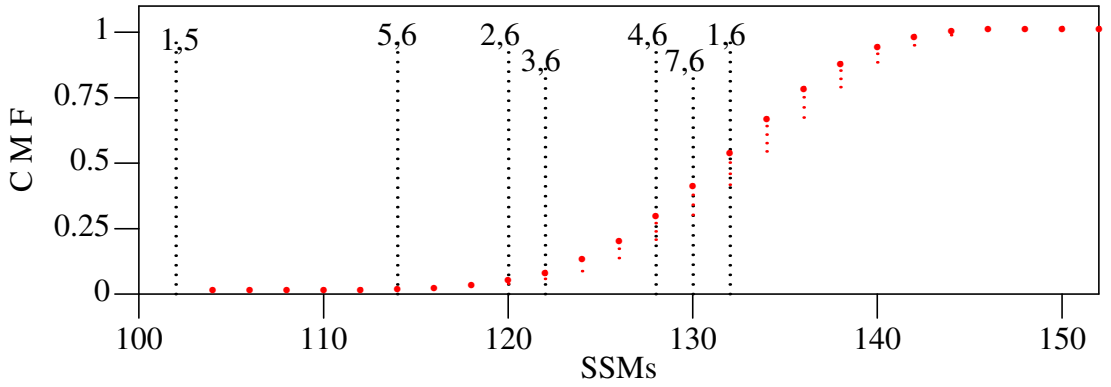

Figure 4: Seven Chains, Three Brands, SSMs against Random CMF.

6. *Simulated Data*

We use three runs from simulations undertaken for Marks et al. (1995). Each run has three interacting brands (Brands 1, 2, and 5, above), and each brand agent chooses its price from its own set of four possible prices in order to maximise its weekly profit, learning using the Genetic Algorithm. Each agent's action is determined by the state of the market in the previous week, which means 64 possible market states for each agent

---

14. With two orders of magnitude fewer possible states, there is a much greater probability that random strategic processes will share some states with historical time-series, as we see here.

15. This number was determined by a Monte Carlo bootstrap simulation of 100,000 pairs of sets of four quasi-random time-series, calculating the SSM between each pair, and examining the distribution. The lowest observed SSM of 104 appeared six times, that is, with a frequency of 6/100,000, or 0.006 percent. See the simulated Probability Mass Distribution (PMD) in Figure 8 below.

to respond to. The GA chooses the mapping from perceived state to action for each brand (with weekly profit as its "fitness").

The three runs imply three different models of the brands' interactions. Each run corresponds to a separate run of the GA search for model parameters, using weekly profits of the brands as the GA "fitness". Given the complexity of the search space and the stochastic nature of the GA, each run "breeds" a distinct model, with distinct mappings from state to brand price, and. hence different patterns of brand actions associated with each run.

Figure 5 shows a fifty-week period of simulated interactions between three brand agents (Run 26a).

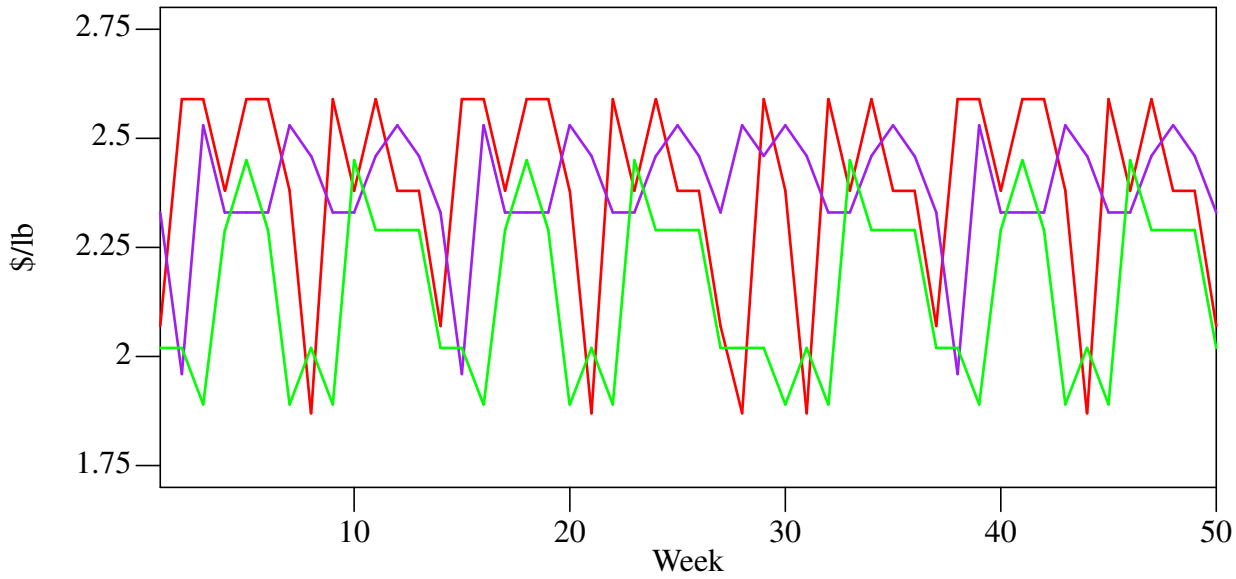

Figure 5: Example of a Simulated Oligopoly (Marks et al. 1995)

Table 5 presents the distances between historical Chain 1, and the three simulations Run 11, Run 26a, and Run 26b from Marks et al. (1995). We have truncated the historical data to 50 weeks, to match the simulated data.

	Chain 1	Run 11	Run 26a	Run 26b
Chain 1	0	82*	68	68
Run 11	82*	0	66	60
Run 26a	68	66	0	30
Run 26b	68	60	30	0

Table 5: SSMs Between Chain 1 and Three Runs (Brands 1, 2, 5)
(* : cannot reject the null at the 5% level)

With three-week windowing, the number of possible states is 50 − 2 = 48, which means that the maximum distance apart any two time-series could be is 96. From Table 5, the

three simulation runs are closer to each other than any of them is to the Chain 1 historical data, although Run 26a is only slightly more similar to Run 11 than it is to Chain 1 (66 compared to 68/96 in distance). The two runs 26a and 26b are quite close to each other: only 30/96 = 31.25 percent apart. Longer weekly data, either from longer simulation runs or from resampling the historical data (Burns Statistics 2008) would allow closer examination of the differences and similarities.

If we use the null hypothesis that each of two sets of time-series is random, then we can set 1% and 5% one-sided confidence intervals to the SSM numbers. With three brands and C = 48, the largest SSM is 96. 95% of pairs of sets of three random time-series are at least 80 apart, and 99% of pairs of sets of three random time series are at least 76 apart.[16] This means that, in Table 5, we reject the null hypothesis of random data for all pairs but Chain 1 and Run 11, since all SSMs between the five other pairs are less than 76, so the data are significantly non-random, and the null hypothesis is rejected. For the pair of Chain 1 and Run 11, however, the SSM of 82 is not significantly (5%) different from random, and the null hypothesis cannot be rejected. By construction, none of the simulated data sets is random, although they are not particularly similar, except for the pair of Run 26a and Run 26b.

We show this in Figure 6, which plots the CMF of the MC parameter bootstrap simulation against the six SSMs of the pairs. The red lines are the CMF of pairs of sets of random series (3 series, 48 observations) from 100,000 Monte Carlo parameter bootstraps.

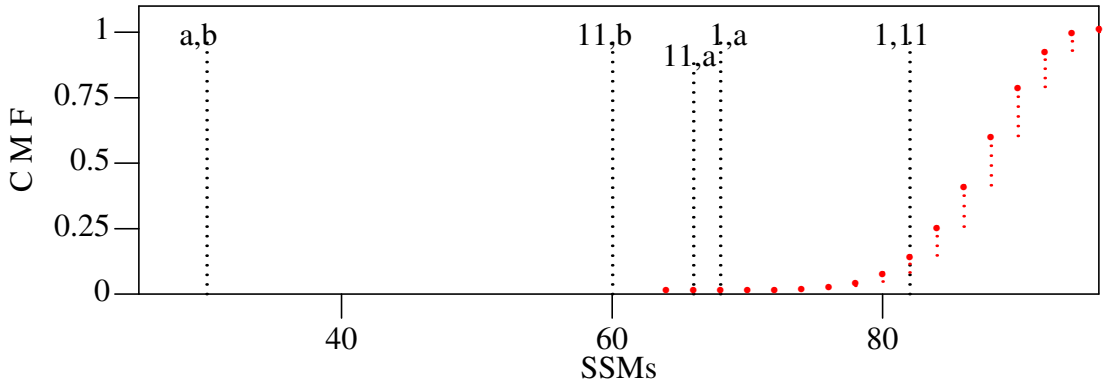

Figure 6: Chain 1 and the Runs, Three Brands, SSMs against Random CMF.

7. Conclusion

This measure, the *State Similarity Measure* (SSM), is sufficient to allow us to put a

16. This number was determined by a Monte Carlo bootstrap simulation of 100,000 pairs of sets of four quasi-random time-series, calculating the SSM between each pair, and examining the distribution. The lowest observed SSM of 64 appeared twice, that is, with a frequency of 2/100,000, or 0.002 percent. See the simulated Probability Mass Distribution (PMD) in Figure 9 below.

number on the degree of similarity between two sets of time-series which embody strategic reactions among agents, as formalised by equation (2). Such a metric is necessary for scoring the distance between any two sets, which previously was unavailable for such sets of time-series. Using simulation of sets of random time-series, we have been able to derive confidence interval SSMs as a statistical test for this measure. (We note that the PMDs suggest binomial distributions, which could be estimated from the simulated data.)

We assume that the time-series are the results of interactions among the three (or four) brand managers, where each manager's action this week is a function of all managers' actions last week and in previous weeks (equation 2). (Here we consider a window of weeks equal to the number of strategic brands.) Although we have devised the SSM for use in examining the rivalrous dance among these brands, this moment could be used to measure the similarity between two sets of time-series generated by any similar interactions.

Here, the SSM has been developed to allow us to measure the extent to which a simulation model that has been chosen on some other criterion (e.g. weekly profitability) is similar to historical sets of time-series with dynamic responses among sellers. The SSM will also allow us to measure the distance between any two such sets of time-series and so to estimate the parameters, or to help calibrate a model against history.

8. *Appendix 1: Calculating the SSM*

1. For each set, partition the time-series $\{P_{b,w}\}$ of price $P_{b,w}$ of brand b in week w into {0,1}, where 0 corresponds to "high" price (above brand b's mid-point) and 1 corresponds to "low" price to obtain $\{P_{b,w}'\}$;
2. For the set of 3- or 4-brand time-series of brands' partitioned prices $\{P_{b,w}'\}$, calculate the time-series of the state of the market each week $\{S_w\}$, where $S_w = P_{1,w}' \times P_{2,w}' \cdots$;
3. For each set, calculate the time-series of states of the 3- or 4-week moving window of partitioned prices $\{M_w\}$, from the per-week states $\{S_w\}$, where $M_w = S_{w-1} \times S_{w-2} \cdots$;
4. Count the numbers of each state observed for the set of time-series over the given time period; convey this by an $n \times 1$ vector $\boldsymbol{c}$, where $\boldsymbol{c}[s]$ = the number of observations of window state s over the period;
5. Subtract the number of observations in set A of time-series from the number observed in set B, across all n possible states; $\boldsymbol{d}^{AB} = \boldsymbol{c}^{A} - \boldsymbol{c}^{B}$;
6. Sum the absolute values of the differences across all possible states:

$$D_{ad}^{AB} = \mathbf{1}' \times |\boldsymbol{d}^{AB}|$$

This number D_{ad}^{AB} is the distance between two time-series sets A and B. This is the State Similarity Measure.

9. *Appendix 2: The Simulated Probability Mass Functions*

We present here figures of the three simulated Probability Mass Functions (PMFs) for the SSMs between two sets of random time-series, where there are three or four time-series per set, and where the number of observations varies. Each figure is the result of 100,000 independent replications. An SSM = 0 measures identity; an SSM = maximum measures zero intersections of window states. The maximum SSM is 150 for Figure 7, 152 for Figure 8, and 96 for Figure 9. Note that the distributions appear to be binomial.

Acknowledgements

I wish to thank Information Resources, Inc., for providing the historical data used in this study. I wish to thank the Zentrum für interdisziplinäre Forschung at the Universität Bielefeld for their generosity in hosting the author for the workshop Advances in Agent-Based Computational Economics, ADACE 2010, July 5–7, 2010. I wish to thank the Joint Statistical Meetings 2010 and Bill Brand for their support of my attendance. I wish to thank David Midgley for his comments, and the participants at ADACE 2010 and the Sydney Agents Group for their questions and comments.

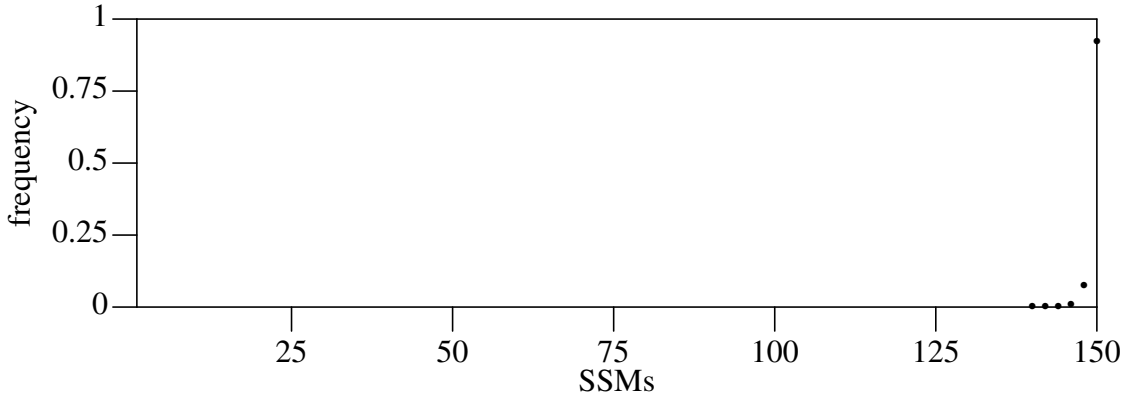

Figure 7: PMF for Two Sets of Random Series (4 series, 75 observations)

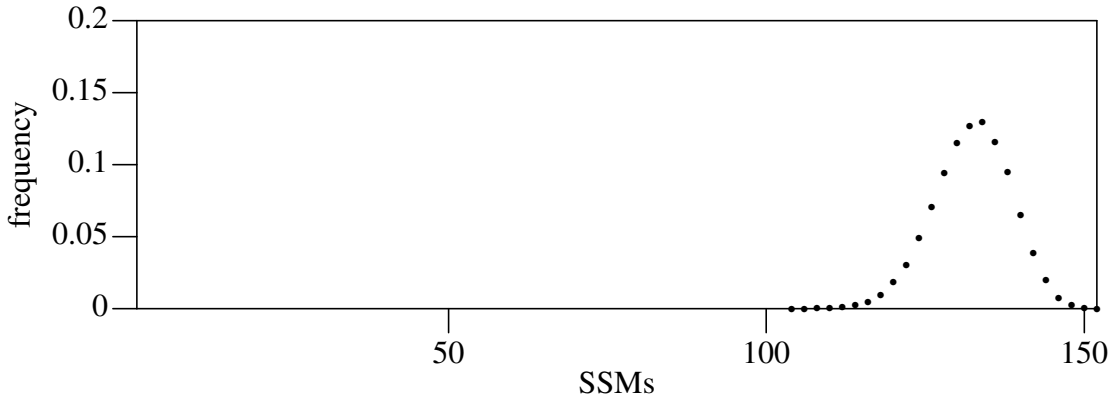

Figure 8: PMF for Two Sets of Random Series (3 series, 76 observations)

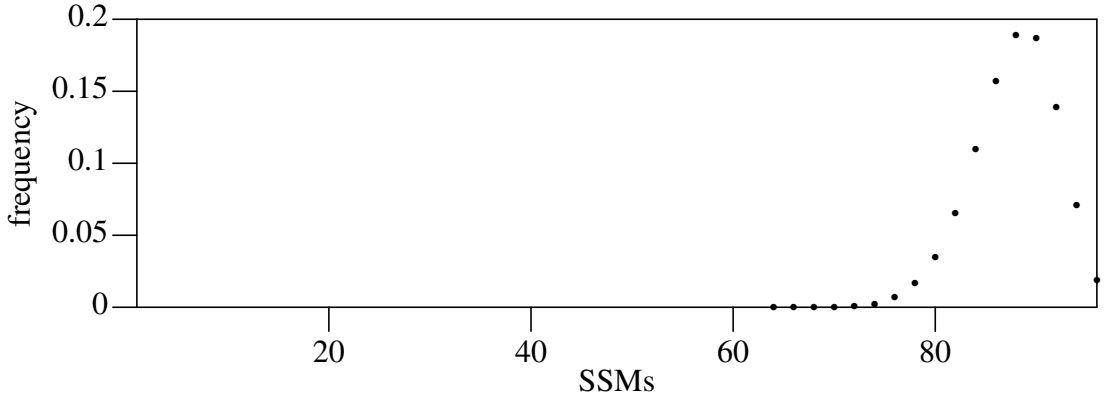

Figure 9: PMF for Two Sets of Random Series (3 series, 48 observations).

10. Bibliography

[1] Burns Statistics — Statistical Bootstrap and Other Resampling Methods. 2008 June 19. http://www.burns-stat.com/pages/Tutor/bootstrap_resampling.html

[2] Shu-Heng Chen, Chia-Ling Chang and Ye-Rong Du (2012), Agent-Based Economic Models and Econometrics, *Knowledge Engineering Review,* forthcoming.

[3] G. Fagiolo, Moneta, A. and Windrum, P. (2007), A critical guide to empirical validation of agent-based models in economics: methodologies, procedures, and open problems, *Computational Economics*, 30: 195–226.

[4] Giorgio Fagiolo and Andrea Roventini (2012), On the scientific status of economic policy: a tale of alternative paradigms, *Knowledge Engineering Review,* forthcoming.

[5] R. Franke (2009), Applying the method of simulated moments to estimate a small agent-based asset pricing model, *Journal of Empirical Finance,* 6: 804–815.

[6] M. Gilli and P. Winker (2003), A global optimization heuristic for estimating agent based models, *Computational Statistics & Data Analysis* 42: 299–312.

[7] C. Gourieroux and A. Monfort (1997), *Simulation-Based Econometric Methods*. OUP/CORE Lecture Series. Oxford University Press, Oxford.

[8] L.P. Hansen and J.J. Heckman (1996), The empirical foundations of calibration, *The Journal of Economic Perspectives*, 10(1): 87–104.

[9] B. LeBaron (2006), Agent-based computational finance, *Handbook of Computational Economics, Volume 2, Agent-Based Computational Economics*, ed. by L. Tesfatsion and K. Judd, Elsevier, Chapter 24.

[10] A. Marcet (1994), Simulation analysis of dynamic stochastic models: Applications to theory and estimation, *Advances in Econometrics. Sixth World Congress*, ed. by C.A. Sims, Volume 2, Cambridge, C.U.P., pp. 81–118.

[11] R.E. Marks (1992) Breeding hybrid strategies: optimal behaviour for oligopolists, *Journal of Evolutionary Economics*, 2: 17–38.

[12] R.E. Marks (1998) Evolved perception and behaviour in oligopolies, *Journal of Economic Dynamics and Control*, 22(8–9): 1209–1233, July.

[13] R.E. Marks (2007), Validating simulation models: a general framework and four applied examples, *Computational Economics*, 30(3): 265–290, October.

[14] R.E. Marks, D.F. Midgley, and L.G. Cooper (1995) Adaptive behavior in an oligopoly, *Evolutionary Algorithms in Management Applications*, ed. by J. Biethahn and V. Nissen, (Berlin: Springer-Verlag), pp. 225–239.
http://www.agsm.edu.au/bobm/papers/marks-midgley-cooper-95.pdf

[15] Daniel McFadden (1989), A method of simulated moments for estimation of discrete response models without numerical integration, *Econometrica*, 57(5): 995–1026.

[16] D.F. Midgley, R.E. Marks, and L.G. Cooper (1997), Breeding competitive strategies, *Management Science*, 43: 257–275.

[17] D.F. Midgley, R.E. Marks, and D. Kunchamwar D. (2007) The building and assurance of agent-based models: an example and challenge to the field, *Journal of Business Research*, Special Issue: Complexities in Markets, 60: 884–893.

[18] J. Miller (1998), Active non-linear tests (ANTs) of complex simulation models, *Management Science*, 44(6): 820–830.

[19] N. Oreskes, K. Shrader-Frechette, and K. Belitz (1994), Verification, validation, and confirmation of numerical models in the earth sciences, *Science*, 263: 641–646.

[20] M.G. Richiardi (2012), ACE: a short introduction, *Knowledge Engineering Review*, forthcoming.

[21] P. Winker, M. Gilli, and V. Jeleskovic (2007), An objective function for simulation based inference on exchange rate data, *Journal of Economic Interaction and Coordination*, 2(2): 125–145.